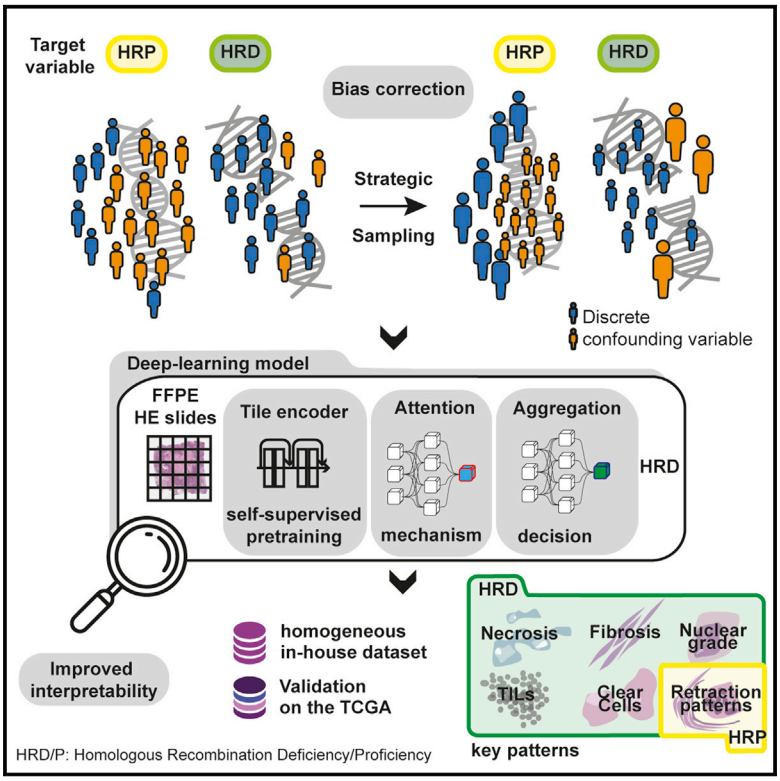
Cell Reports
Medicine

Article

Deep learning identifies morphological patterns of homologous recombination deficiency in luminal breast cancers from whole slide images

Graphical abstract



Highlights

- Homologous recombination deficiency is predictable from H&E slides with high accuracy
- Biases in computational pathology data can be alleviated by strategic sampling
- We present a method to identify morphological patterns of complex phenotypes
- We identified five HRD- and two HRP-related morphological patterns

Authors

Tristan Lazard, Guillaume Bataillon, Peter Naylor, ..., Etienne Decencière, Thomas Walter, Anne Vincent-Salomon

Correspondence

thomas.walter@mines-paristech.fr (T.W.), anne.salomon@curie.fr (A.V.-S.)

In brief

Deep-learning models predict homologous recombination deficiency (HRD) from H&E-stained pathology slides. Dataset biases, either biological or technical, can be alleviated by strategic sampling. Interpretation of the predictive models reveals several morphological patterns related to HRD and opens new hypotheses about its phenotypic impact.

Lazard et al., 2022, Cell Reports Medicine 3, 100872
December 20, 2022
https://doi.org/10.1016/j.xcrm.2022.100872

CellPress

Cell Reports Medicine

CellPress
OPEN ACCESS

Article

Deep learning identifies morphological patterns of homologous recombination deficiency in luminal breast cancers from whole slide images

Tristan Lazard,[1,2,3,13] Guillaume Bataillon,[2,3,4,11,13] Peter Naylor,[1,2,3,12] Tatiana Popova,[5] François-Clément Bidard,[6,7] Dominique Stoppa-Lyonnet,[5,8] Marc-Henri Stern,[4,5] Etienne Decencière,[9] Thomas Walter,[1,2,3,13,*] and Anne Vincent-Salomon[4,10,13,14,*]

[1]Center for Computational Biology (CBIO), Mines Paris, PSL University, 60 Boulevard Saint Michel, 75006 Paris, France
[2]Institut Curie, PSL University, 75005 Paris, France
[3]INSERM U900, 75005 Paris, France
[4]Diagnostic and Theranostic Medicine Division, Institut Curie, PSL University, Paris, France
[5]INSERM U830, DNA Repair and Uveal Melanoma (DRUM), Equipe Labellisée par la Ligue Nationale Contre le Cancer, Institut Curie, PSL Research University, 75005 Paris, France
[6]Department of Medical Oncology, Institut Curie, Université de Versailles Saint-Quentin, Saint-Cloud, France
[7]INSERM CIC-BT 1428, Institut Curie, Paris, France
[8]Université Paris Cité, 75006 Paris, France
[9]Center for Mathematical Morphology (CMM), Mines Paris, PSL University, 77300 Fontainebleau, France
[10]INSERM U934, CNRS UMR 3215, Paris, France
[11]Present address: Department of Pathology, University Cancer Institute of Toulouse-Oncopole, Toulouse, France
[12]Present address: RIKEN AIP, Kyoto, Japan
[13]These authors contributed equally
[14]Lead contact
*Correspondence: thomas.walter@mines-paristech.fr (T.W.), anne.salomon@curie.fr (A.V.-S.)
https://doi.org/10.1016/j.xcrm.2022.100872

SUMMARY

Homologous recombination DNA-repair deficiency (HRD) is becoming a well-recognized marker of platinum salt and polyADP-ribose polymerase inhibitor chemotherapies in ovarian and breast cancers. While large-scale screening for HRD using genomic markers is logistically and economically challenging, stained tissue slides are routinely acquired in clinical practice. With the objectives of providing a robust deep-learning method for HRD prediction from tissue slides and identifying related morphological phenotypes, we first show that digital pathology workflows are sensitive to potential biases in the training set, then we propose a method to overcome the influence of these biases, and we develop an interpretation method capable of identifying complex phenotypes. Application to our carefully curated in-house dataset allows us to predict HRD with high accuracy (area under the receiver-operator characteristics curve 0.86) and to identify morphological phenotypes related to HRD. In particular, the presence of laminated fibrosis and clear tumor cells associated with HRD open new hypotheses regarding its phenotypic impact.

INTRODUCTION

Worldwide, 2.1 million women are newly diagnosed per year with breast cancer (BC), which is a leading cause of cancer-related death. Improvement of metastatic BC treatment is therefore of highest priority. BC is a heterogeneous disease with four major molecular classes (luminal A and B, HER2 enriched, and triple-negative breast cancer [TNBC]) benefiting from different therapeutic approaches. If early BC patients have an overall survival of 70%–80%, metastatic disease is incurable with a short duration of survival.[1] Homologous recombination (HR) is a major and high-fidelity repair pathway of DNA double-strand breaks. Its deficiency, HRD, results in high genomic instability[2] and occurs through diverse mechanisms, including germline or acquired somatic mutations in DNA-repair genes, most frequently *BRCA1*, *BRCA2*, or *PALB2*, or through epigenetic alterations of *BRCA1* or *RAD51C*. Importantly, HRD leads to high sensitivity to poly-ADP-ribose polymerase inhibitors (PARPi) *in vitro*,[3,4] a treatment that has been shown to improve metastatic BC progression-free survival.[5,6] HRDs induced by *BRCA1* and *BRCA2* mutations are known predictive markers for response to PARPi[2,6] and platinum salt,[7] and somatic HRD has been more recently recognized as a predictive marker for PARPi in ovarian cancer[2] and BC.[8]

Several methods have been developed to detect HRD, including genomic instability profiling, mutational signatures, or integrating structural and mutational signatures.[9–13] Today, HRD is diagnosed in clinical practice by DNA-repair gene sequencing, germinal in BCs and somatic in ovarian cancers,

respectively. For ovarian cancers, HRD is also assessed by genomic instability tests such as the HRD MyChoice CDx test (Myriad Genetics).

The majority of hereditary *BRCA1* cancers are TNBC and up to 60%–69% of sporadic TNBCs harbor a genomic profile of HRD.[8,9,14] In contrast, the majority of hereditary *BRCA2* cancers are luminal,[15] and HRD also exists in sporadic luminal B[8,16] or in HER2 tumors.[17,18] Of note, germline or sporadic alterations of *BRCA* harbor indistinguishable genomic alterations in triple-negative or luminal tumors.[16,19] Also, the recent results of the Olympia trial emphasize the need for an efficient method of screening for *BRCA1* and *BRCA2* mutations across all BC phenotypes.[6]

In this context, it seems appropriate to systematically screen for HRD induced by *BRCA1* and *BRCA2* mutations not only for TNBC (18% of all BCs), but also for luminal B tumors (35% of all BCs). This, however, would represent a real challenge in clinical practice, both economically and logistically. To overcome these challenges, we hypothesized that HRD might be predictable from its phenotypic consequences visible in stained tissue slides acquired in clinical practice. On the other hand, no specific routinely assessed phenotype has been reported to indicate the presence of HRD. For this reason, we set out to predict HRD from whole slide images (WSIs) by deep learning and to identify the underlying morphological patterns.

Deep learning has revolutionized biomedical image analysis and in particular digital pathology. Traditionally, the majority of methods developed in this field were dedicated to computer-aided diagnosis, whereby the objective is to partially automatize human interpretation of slides in order to help pathologists in their diagnostic task, e.g., the detection of mitoses[20] or the identification of metastatic axillary lymph nodes.[21,22] Beyond the automatization of manual inspection, deep learning has also been successfully applied to prediction of patient variables, such as outcome,[23] and molecular features, such as gene mutations,[24,25] expression levels,[26] or genetic signatures.[24,27] However, one of the major drawbacks of deep-learning algorithms is their black-box character: because deep learning relies on automatically generated rather than predefined features with a clear biological interpretation, it is difficult to know how a decision was made. This has two major consequences: first, it is difficult to identify potential confounders, i.e., variables that correlate with the output because of the composition of the dataset and that are predicted instead of the intended output variable. Second, even in the absence of statistical artifacts, understanding how the decision was generated in the first place can point to interesting mechanistic hypotheses and to patterns in the image that have so far been overlooked.

One way to overcome the latter problem is to use hand-crafted biologically meaningful features.[27] This, however, requires an extraordinary effort in terms of annotation. Here, we take a conceptually different approach. Instead of working in a pan-cancer setting on a large number of signatures, we concentrate on one single medically highly relevant signature in one cancer type in a controlled dataset, where we can investigate and correct for potential biases. To understand how the deep-learning decision is generated and which morphological patterns are related to the output variable, we propose a visualization technique that overcomes limitations of current approaches in the presence of complex phenotypes. This paves the way to "machine teaching," i.e., a data-driven approach to identify phenotypic patterns related to genomic signatures that is capable of pointing to new mechanistic hypotheses.

In this study, we present an image-based approach to predict HR status from WSIs stained with hematoxylin and eosin (H&E) using deep learning from a large retrospective series of luminal and triple-negative breast carcinomas with a genomically defined HR status from a single cancer center. Furthermore, we identify the morphological patterns associated with HRD. For this we have to tackle two important methodological challenges: the identification and correction of biases in the training data and the identification of morphological patterns linked to the output variable in the presence of complex pleiotropic phenotypes. Application of these methods to our curated dataset allows us to predict HRD with high accuracy and allows the discovery of decisive, previously unknown morphological patterns related to HRD, leading to new hypotheses on disease-relevant genotype-phenotype relationships.

RESULTS

A deep-learning architecture to predict HRD from whole slide images

We scanned the most representative H&E-stained tissue section of the surgical resection specimens of BC from 714 patients with known HR status. The series was composed of 309 homologous recombination proficient (HRP) tumors and 406 HRD tumors (Table S4).

Because of their enormous size, analysis of WSIs typically relies on the multiple instance learning (MIL) paradigm.[28–31] MIL techniques only require slide-level annotations and share the overall architecture (Figure 1), consisting of four main steps: tiling and encoding, tile scoring, aggregation, and decision.

The WSI is divided into tile images (dimensions: 224 × 224 pixels) arranged in a grid. Background tiles are removed and tissue tiles are encoded into a feature vector. Instead of using representations trained on natural image databases and unlike most studies in this domain, we used the self-supervised technique momentum contrast (MoCo;[32] see STAR Methods). This method consists in training a neural network (NN) to recognize images after transformations, such as geometric transformations, noise addition, and color changes. By choosing the type and strength of transformations, we can impose invariance classes, i.e., variations in the input that do not result in significantly different representations. After tile encoding, the feature vector of each tile is then mapped to an attention score by an NN. The slide representation is obtained by the sum of the individual tile representations, weighted by the learned attention scores.[28] Finally, the slide representation is classified by the decision module (Figure 1). We optimized hyperparameters by a systematic random search strategy (see STAR Methods). For hyperparameter setting and performance estimation, we used nested 5-fold cross-validation, which allowed us to obtain realistic performance estimations. All reported performance results are averaged over five independent test folds (see STAR Methods).

CellPress
OPEN ACCESS

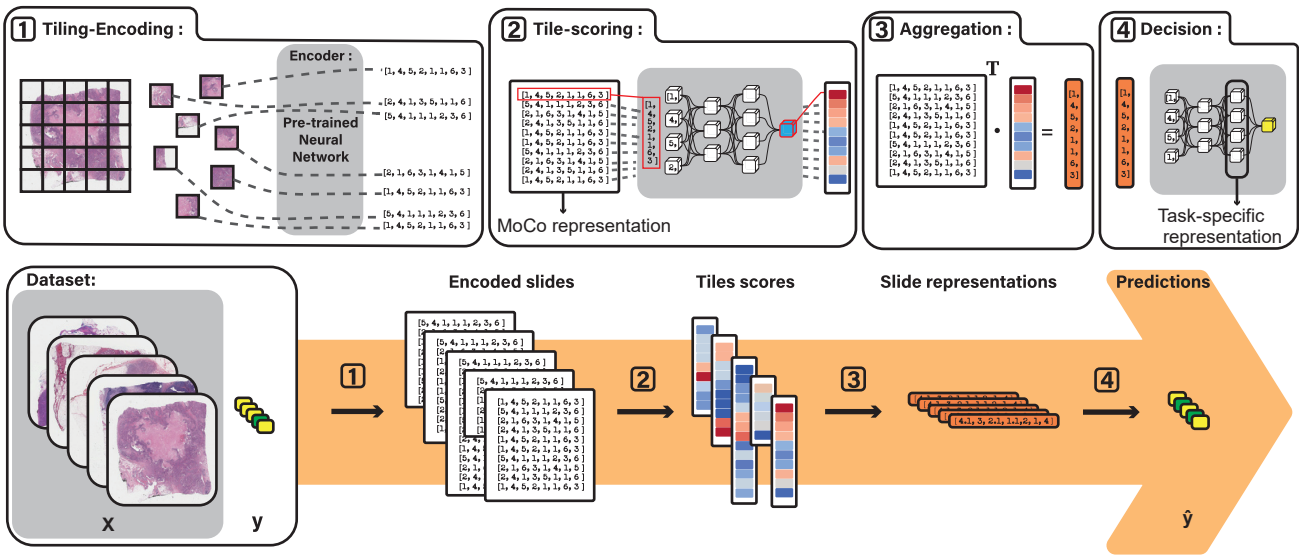


Figure 1. From WSI to prediction

Four major components are used in this end-to-end pipeline. First, the WSIs (x) are tiled, the tissue parts are automatically selected, and the resulting tiles are embedded into a low-dimensional space (block 1). The embedded tiles are then scored through the attention module (2). An aggregation module outputs the slide-level vector representative (3) that is finally fed to a decision module (4), which outputs the final prediction. When training, the binary cross-entropy loss between the ground truth y and the prediction $\hat{y}$ is computed and back-propagated to update the parameters of the modules. Both the decision module and the attention module are multilayer perceptrons, the encoder is a ResNet18, and the aggregation module consists of a weighted sum of the tiles, the weights being the attention scores.

HRD prediction with correction for potential biases

Prediction results obtained without bias correction

We applied this method to predict HRD from the WSI in The Cancer Genome Atlas (TCGA) cohort and obtained results (area under the receiver-operator characteristics curve [AUC] = 0.71, Figure 2C) in line with previous reports.[27,33–35] While TCGA is an invaluable resource for pan-cancer studies in genomics and histopathology, it is often seen rather as a starting point whose results need to be corroborated by other cohorts.[36] Furthermore, TCGA contains images from many centers around the world with potentially different sample preparation and image-acquisition protocols. While this technical variability might reflect to some degree what could be expected in clinical practice for multiple institutions, we hypothesized that to prove the predictability of HRD independently of potential technical and biological biases, as well as in an in-depth study of morphological patterns related to HRD, it might be advantageous to work on a more homogeneous dataset where we can carefully control for potential technical and biological confounders. We thus turned to our in-house dataset, hereafter referred to as the Curie dataset (see STAR Methods), with data from 714 patients.

We trained an NN to predict HRD on this carefully curated dataset, and we observed a prediction performance largely superior to the best reported to date, trained and tested on TCGA (AUC = 0.88, Figure 2C).

Identification and correction of biases

As the cohort was generated over 25 years, two experimental variables representing changes in experimental protocols have been identified as potential confounders (c_1 corresponding to the fixation protocol and c_2 to the impregnation protocol, see STAR Methods).

To measure the confounding effects of these variables on the model predictions, we developed a bias score (see STAR Methods). This score is close to zero in the unbiased case and increases with increasing bias. We found that model predictions were indeed biased by these two confounders (Figure 2A).

We then devised a sampling strategy that mitigates biasing during training. Bias mitigation is an increasingly important line of research in machine learning. For instance, it is a well-known problem in training predictive models for functional MRI data, where the age of the patient has been shown to be an important confounder.[37] While several techniques for bias mitigation exist,[38–41] a recent comparison[42] indicates that strategic sampling is the method of choice if the distribution is not too imbalanced. Strategic sampling aims at ensuring that irrespective of the composition of the training set, each batch presented to the NN is composed of roughly the same number of samples for each value combination of output and confounding variable. Correcting for c_1 and c_2 resulted in a 4-fold reduction of the bias score in comparison with the uncorrected model and a slightly lower accuracy (AUC = 0.86, Figure 2C). These results are corroborated using the bias-amplification (BA) measure, a metric widely used in the machine learning fairness literature:[39,42] on the in-house dataset, correcting for c_1 and c_2 lowers the BA from −0.02 to −0.05; on TCGA dataset, the subtype correction lowers the BA from −0.06 to −0.15.

In addition to these technical confounders, we identified the molecular subtype of the tumor to be a potential biological confounder. Successful correction of this biological confounder in TCGA (Figure 2B) led, however, to a dramatic drop in performance (AUC = 0.63). This result suggests that NN trained on the entire BC subset of TCGA for HRD prediction without

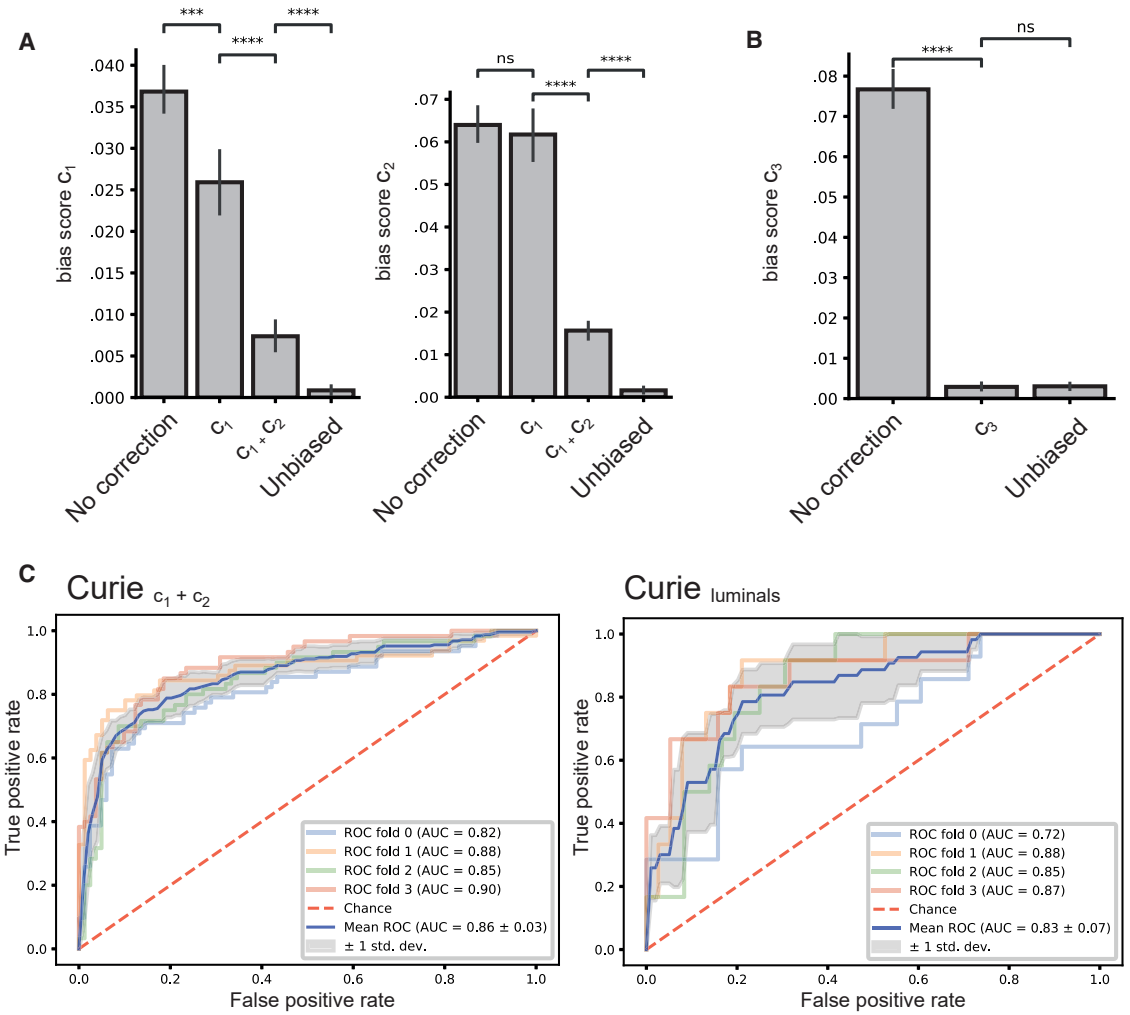


Figure 2. Bias corrections and prediction performances

(A and B) Estimation of the bias score of two technical confounders (c_1, c_2) and one biological confounder (c_3) for the Curie dataset (A) and the bias score of the confounder c_3 for TCGA dataset (B) for different correction strategies. A Mann-Whitney-Wilcoxon test, two-sided with Bonferroni correction, is performed for each pair of correction strategies. As detailed in STAR Methods, for each correction strategy a series of 30 unbiased subtest sets are sampled on which the model's bias is evaluated. Error bars indicate standard deviations over the subtest sets. The significance test is performed on this distribution of 30 estimations. The bias score of a model is the average of this distribution. ns, not significant ($p > 0.05$); *$p < 0.05$, **$p < 0.01$, ***$p < 1 \times 10^{-3}$, ****$p < 1 \times 10^{-4}$.

(C) Receiver-operating characteristic curves. The name of each model indicates the origin of its training set. Indices indicate the correction applied through strategic sampling (Curie_{c_1} has been debiased with respect to c_1). $\text{Curie}_{\text{luminals}}$ corresponds to the model trained on a subset containing only luminal tumors.

stratification or bias correction might actually predict to a large extent the molecular subtype, which is also a predictable variable (AUC = 0.89). This shows that the molecular subtype is indeed a biological confounder. In our in-house dataset, we decided to build a subtype-specific NN that specifically predicts HRD for luminal BC instead of applying bias mitigation. The reason for this decision was 3-fold: first, we argued that a dataset focusing on only one molecular subtype was more likely to reveal the underlying patterns exclusively related to HRD; second, HRD prediction in luminal BC is of particular importance for clinical practice, as very few morphological patterns are known to be related to HRD in luminal BC, the most frequent BC phenotype; and third, the relatively low number of TNBCs in our dataset made strategic sampling on three confounding variables challenging. Therefore, we composed a dataset containing only luminal BC and setting both technical confounders, leading us to keep 251 BC WSIs (188 HRD tumors and 63 HRP tumors). We obtained a good, albeit slightly lower performance of this bias-corrected NN (AUC = 0.83; Figure 2C and Table 1). The trained model carefully freed from both technical and biological biases and validated with respect to cross-dataset performance (Table S3) was then used for the identification of morphological patterns described in the next section. We additionally performed benchmarking experiments to evaluate the influence of the tile encoder network and the MIL algorithm on the classification performances (Tables S1 and S2).

Table 1. Classification performances

	AUC		B_{Acc}	
	Mean	SD	Mean	SD
$TCGA_{raw}$	0.71	0.10	0.59	0.08
TCGA c_3	0.63	0.08	0.54	0.02
$Curie_{raw}$	0.88	0.03	0.81	0.02
Curie c_1 + c_2	0.86	0.03	0.78	0.04
$Curie_{luminals}$	0.83	0.07	0.72	0.06

Summary of performance metrics. Mean and standard deviation (SD) are computed over the five test sets of the cross-validation. The name of each model indicates the origin of its training set. Indices indicate the correction applied through strategic sampling ($Curie_{c_1}$ has been debiased with respect to c_1). $Curie_{luminals}$ corresponds to the model trained on a subset containing only luminal tumors. We provide an in-depth benchmark of the algorithm in Tables S1 and S2 and cross-dataset experiments in Table S3. AUC, area under the (receiver-operating characteristics) curve; B_{Acc}, balanced accuracy.

Visualization reveals HRD-specific tissue patterns

Visualization of attention scores can be misleading

To understand which phenotypic patterns are related to HRD on the WSI, we turned to visualization techniques for NNs. The used MIL framework is equipped with an inherent visualization mechanism: the second module of the algorithm, the tile-scoring module, is in fact an attention module that assigns to each tile an attention score that determines how much a given tile will contribute to the slide representation (and thus to the decision). Attention scores are often used for visualization in the field of digital pathology,[22,43–45] in the form of either heatmaps to localize the origin of the relevant signals or galleries of tiles of interest (tiles with highest attention scores). However, attention scores do not per se extract the tiles that are related to a certain output variable; they simply reflect that the tile has been taken into consideration in the decision. In particular, in the case of genetic signatures, where we would expect that the output variables can be related to several morphological patterns, analyzing only the attention scores might thus be limited. Figure S2 illustrates the results obtained by attention-based explanation: while we observe one specific cluster for HRP, most attended tiles seem to be present in both HRD and HRP slides. A possible explanation is that the HRD/HRP decision might be related to the frequency of certain tissue phenotypes rather than to their mere presence.

The decision-based visualization technique provides a global explanation of the model

Given these limitations, we propose a visualization protocol that allows us to extract the tiles that are directly associated with a particular slide-level label. As the slide representation is the weighted sum of the tile representations, we applied the decision module, specifically trained to classify slide representations between HRD and HRP, to the individual tile representations. This gives us a score for each tile that can be interpreted as the (tile) probability of being HRD or HRP (see STAR Methods for details). Selecting the tiles with the highest posterior probability for HRD and HRP, respectively, and projecting the tile representations of this selection to a low-dimensional space leads to the emergence of distinct clusters corresponding to different tumor tissue patterns with a clear relation to HRD or HRP and therefore providing a morphological map of HRD (Figure 3).

Two expert pathologists labeled these clusters. The HRD signal relied on several clusters: HRD tumors present a high tumor cell density, with a high nucleus/cytoplasm ratio and conspicuous nucleoli. They also show regions of hemorrhagic suffusion associated with necrotic tissue. In the stroma, the HRD signal revealed the presence of striking laminated fibrosis and, as expected, high content of tumor-infiltrating lymphocytes (TILs). Lastly, one large cluster contained a continuum of several phenotypes, namely adipose tissue intermingled with scattered and clear tumor cells, histiocytes, and plasma cells. In contrast, the HRP signal was mostly carried by one cluster characterized by low tumor cell density, the cells being moderately atypical, and tumor cell nests separated from the stroma by clear spaces. Notably, it included a few invasive lobular carcinomas (all of the tiles per cluster are available in Figures S6–S8).

Validation of the morphological patterns

Some of these patterns, namely high-grade and TIL, had been previously associated with phenotypic hallmarks of HRD in TNBCs.[48] To validate these results in the luminal BC cohort, TIL density and nuclear grade were evaluated for each luminal tumor of the in-house dataset by an expert pathologist. As predicted by our algorithm, TILs and nuclear grade were positively associated with the HR status of the tumor in the luminal subset (mean TIL HRD, 29; mean TIL HRP, 17; t test p value, 0.017; mean nuclear grade HRD, 2.7; mean nuclear grade HRP, 2.3; χ^2 p value, 1.2×10^{-6}). Moreover, a logistic regression trained on the components of the grade (architecture grade, atypia grade, and mitosis grade) and on the TIL count estimation has an average AUC of 0.76 (5-fold cross-validation).

To further validate the association of these morphological patterns with HRD, we turned to the independent TCGA cohort. Despite the modest prediction accuracy after bias correction, we found that a NN trained on TCGA-extracted morphological patterns strikingly similar to those obtained from our in-house dataset (Figure S3), with the exception of cluster 4 (Figure 3). Regarding HRP, we were able to validate all patterns related to HRP, but artifact classes were also identified, which is unsurprising given the limited slide quality and heterogeneity of TCGA dataset and may explain the poor classification performance.

To test the subtype specificity of the morphological patterns, we trained a network on the small TNBC subset of TCGA (129 slides). While classification performances remain poor (AUC = 0.62), because of the small size and large heterogeneity of the dataset, the extracted patterns explaining the predictions are in line with the literature (Figure S4), suggesting that HRD for TNBC is characterized by high content of TILs and necrosis, while the retraction figures are still an explanation of the HRP signal. This result further confirms the specificity of our extracted morphological patterns and suggests that there are indeed HRD-related morphological patterns specific to the luminal subtype.

Our NN works with different internal representations. While the tile representations provided by MoCo permit the emergence of phenotypic similarity clusters (Figure 3), internal representations closer to the decision module encode information relevant for

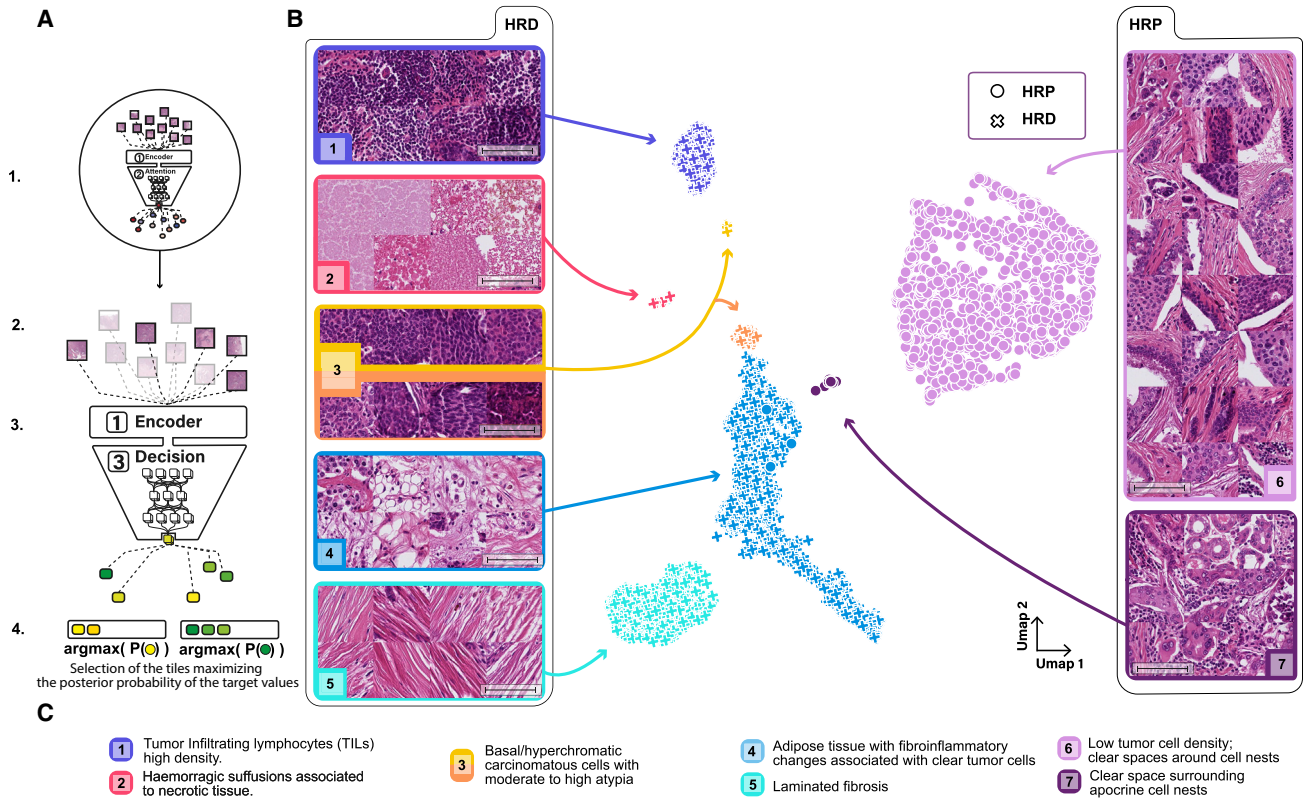


Figure 3. Decision-based visualization

(A) Mechanism of the decision-based visualization. 1: each tile in the whole dataset is scored by the attention module. 2: per slides, the 300 best scoring tiles are selected as candidate tiles. 3: the selected tiles are presented to the decision module, and the logit of the probability of each of these tiles being HRD or HRP (yellow or green) is kept. 4: finally, the K tiles with maximal probability for either HRD/HRP are selected.

(B) Morphological map of the HR status in the luminal BC cohort. Each dot is the uniform manifold approximation and projection (UMAP) of a tile extracted by the decision-based visualization method. Crosses (circles) are tiles with high HRD (HRP) logit. Each cluster has been linked to a morphological phenotype by two expert pathologists. We identified six different morphological phenotypes associated with the HRD and two associated with the HRP. The exhibited tiles have been randomly sampled among each cluster. 228 slides contributed to the HRP clusters and 232 to the HRD cluster. In total, 249 among 251 slides contributed to the whole figure. The same protocol has been applied to the public datasets TCGA breast invasive carcinoma (BRCA), TCGA BRCA-TNBC, and TCGA ovarian cancer (see Figures S3, S4, and S5, respectively). Scale bars, 100 μm.

(C) Pathological interpretation of the clusters presented in (B).

HRD. The representation in the penultimate layer can therefore be interpreted as encoding "HRD-ness" of the tiles. Figure 4 illustrates a low-dimensional representation of this HRD-ness for the same tiles as those present in Figure 3, where point color represents the HRD score (tile probability to be classified as HRD). From there, we extracted two tile trajectories going from low HRD-ness to high HRD-ness. The magenta trajectory illustrates the successive visual changes corresponding to an increase in tumor cells or inflammatory cell density (from low-density tiles to high-density tiles with large nuclei, nuclear atypia, and infiltrative lymphocytes). The blue trajectory shows, conversely, a decrease in tumor cell density replaced successively by an inflammatory reaction and apoptotic cells, loose fibrosis, and hemorrhagic suffusion associated with necrosis. These different trajectories illustrate the manifestations of HRD and show the pleiotropic character of the induced phenotypes. Moreover, the highlighted gradation of these phenotypes opens the path to a possible reading grid of WSIs for pathologists.

DISCUSSION

In this study, we set out to predict the HR status in BC from H&E-stained WSIs and to analyze the phenotypic patterns related to HRD. The prediction of HRD is an important challenge in clinical practice. The use of PARPi for BC patients was initiated for metastatic TNBC patients with germline mutations of *BRCA1* or *BRCA2*. However, *BRCA2*, as well as *PALB2* and a minority of *BRCA1* cancer patients, develop luminal tumors. The necessity of predicting HRD is therefore not limited to TNBC but extends also to luminal BC. On the other hand, luminal BCs represent a far more frequent group than TNBC. For this reason, systematic screening of HR gene alterations for luminal cancers will be problematic and, in many countries, even unfeasible due to both economic and logistic issues. Therefore, preselection of patients with a high probability of being HR deficient by analysis of WSIs is a cost-efficient strategy that has so far only been hampered by the lack of knowledge about HRD-specific

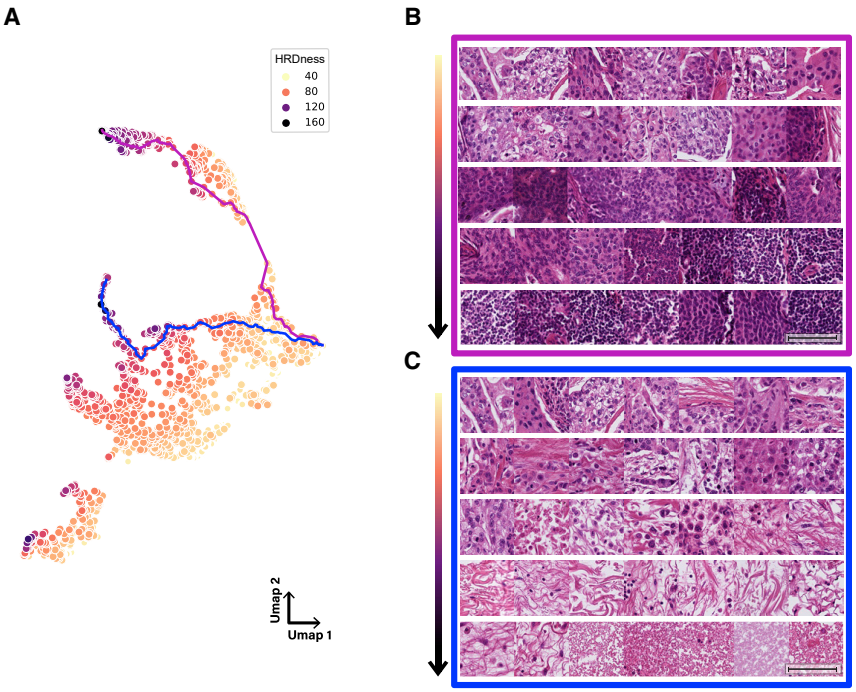


Figure 4. Illustration of two phenotypic HRD-ness trajectories

(A) UMAP projection of the HR status-specific representation of the meaningful tiles relative to the HRD. HRD-ness is the score given to each tile by the HRD output neuron. Two tile trajectories have been extracted (blue and magenta) starting from the same low HRD-ness region, each leading to a different high HRD-ness region.

(B and C) Tiles sampled along each of the trajectories. These are ordered from low HRD-ness to high HRD-ness and read from left to right and from top to bottom. Scale bars, 100 μm. (B) Magenta trajectory, toward densely cellular tumors or inflammatory cells. (C) Blue trajectory, toward fibroinflammatory tumor changes and hemorrhagic suffusions.

morphological patterns in luminals. Indeed, only high grades and to a lower extent pushing margins have previously been reported to be associated with HRD. In this context, the identification of HRD from WSIs by deep learning and the identification of related morphological patterns could both facilitate the preselection of BCs for molecular determination of HRD, which is particularly important for luminal cancers.

TCGA provides a precious dataset from which to train models for the prediction of genetic signatures from H&E data.[24,27] While we obtained promising results for the prediction of HRD on TCGA dataset in line with previous reports, we found that this result was partly due to the fact that the molecular subtype acts as a biological confounder. This was particularly problematic, as we wanted to investigate the morphological signature of HRD. Of note, the existence of biological and technical confounders is presumably not limited to HRD prediction but may concern many genetic signatures. The use of carefully curated datasets where technical and biological confounders can be controlled for is, thus, an important step in investigating the predictability of genetic signatures as well as the identification of their morphological counterparts.

In most cases, such in-house datasets also contain technical and biological biases due to the long period during which the dataset is acquired. This motivated us to propose a method to mitigate bias in computational pathology workflows, based on strategic sampling. Such strategies are already used in other fields of medical imaging but have so far, to the best of our knowledge, not been used in computational pathology. We have shown that this approach can successfully mitigate or even eliminate bias. In a larger perspective, it is essential to investigate potential confounding variables in the dataset when applying deep-learning-based methods for the prediction of slide-level variables. Biased datasets can lead to false expectations and misinterpretation. For this reason, we expect proper treatment of such variables to become a standard in the field.

While bias correction on TCGA led to a drop in AUC to 0.63, we found that HRD was predictable in our in-house dataset of 251 luminal BC patients with an AUC of 0.83. While homogeneous datasets do not reflect the variability between centers and thus limit direct applicability of the trained networks, they allow for controlled feasibility studies, which now need to be complemented by multicenter studies. In addition, we will validate this algorithm in a prospective neoadjuvant clinical trial for which patients' HRD status will be assessed with the MyChoice CDx test (Myriad Genetics).

Homogeneous datasets are well suited for the identification of underlying phenotypic patterns, even in cases where no or few such patterns are known a priori, such as in the case for HRD. To identify a phenotypic signature related to an output variable (here HRD), either we can use biologically meaningful encodings, also known as human interpretable features (HIF), and infer the most relevant features by analyzing the weights in the predictive model,[27] or we can turn to network introspection. The HIF approach relies on detailed and exhaustive annotations of a large number of WSIs, for instance,[27] leverage annotations provided by hundreds of pathologists consisting of hundreds of thousands of manual cell and tissue classifications. Here, we provide a new network introspection scheme relying on the powerful MoCo encodings, trained without supervision directly on histopathology data, and a decision-based tile selection that allows us to automatically cluster tiles and to relate these clusters to the output variable. Interestingly, while our approach confirms the recently published finding that necrosis is a hallmark of HRD[27] and identifies morphological features common to HRD in TNBC and luminal BC, such as necrosis, high density in TILs, and high nuclear anisokaryosis,[46] it also points to more specific patterns that have so far been overlooked. For instance, we found tiles enriched in carcinomatous cells with clear cytoplasm, suggesting activation of specific metabolic processes in these cells. Moreover, we found intratumoral laminated

fibrosis as an HRD-related pattern. Also, we were able to validate most of these patterns on TCGA. This leads to the hypothesis that cancer-associated fibroblasts (CAFs) within the stroma of HRD luminal tumors may play a role in the viability and fate of tumor cells. Furthermore, the presence of adipose tissue within the tumor suggests first, a different tumor cell density and second, a specific balance between CAFs and adipocytes in the context of a luminal HRD tumor. The molecular mechanisms achieving these patterns remain to be determined by *in vitro* models.

Similar to what we have shown here with respect to HRD, the visualization framework we have developed is versatile and can in principle be applied in the context of other genetic signatures. In particular, our visualization scheme overcomes the limitations of the thus far predominating technique of visualizing attention scores alone. Indeed, attention scores were used previously to identify tumor regions under weak supervision. However, if the output variable depends on the quantity of several morphological patterns in contrast to the presence/absence of a single tissue phenotype, attention scores might not provide a suitable tile selection and visualization tool and might thus be ill suited to investigate the underlying morphological phenotypes. Because the algorithm is fully automated, using the MIL algorithm and the proposed visualization method can constitute a useful tool for the discovery of morphological features related to the predicted genetic signatures. This has the potential to generate new biological hypotheses about the phenotypic impact of these genetic disorders. To maximize the benefit for the scientific community, we release the code to train MIL models on WSIs and create morphological maps as well as tile trajectories publicly and free of charge, and provide detailed documentation.

Altogether, this study provides new and versatile tools for the prediction and phenotypic dissection of genetic signatures from histopathology data. Application to luminal BCs allowed us to show that HRD is predictable from WSIs and to shed light on the phenotypic consequences of HRD. These tools have the potential to impact BC patient care.

Limitations of the study

Our study involves a homogeneous, carefully controlled cohort that allowed us to train a network for HRD prediction with high accuracy and correction for technical and biological confounders. We could thus convincingly show that HRD is predictable from WSIs. However, the study was not designed for the demonstration of clinical applicability. To use HRD prediction in clinical practice, we will need to validate the workflow on larger, multicenter cohorts.

Furthermore, we have identified morphological patterns related to HRD. While our validation results obtained from TCGA suggest that the method works robustly and that these patterns are truly linked to HRD, we will need to validate these findings in a larger independent cohort. In addition, the development and demonstration of a mechanistic model explaining these morphological phenotypes will be a challenging and exciting perspective. Finally, it will be important to further explore the variability of the morphological patterns in different cancer types.

At a methodological level, we have proposed strategic sampling as a method to mitigate biases in digital pathology datasets. While we were able to show that this method is highly effective, it must be noted that it is limited by the number of variables we can correct for as well as by the class imbalance it can handle. In some cases, stratification might therefore be preferable. Furthermore, we have proposed a method to improve the interpretability of the MIL approach for HRD prediction. However, it is still difficult to precisely understand how the identified tiles impact the prediction. For instance, the method does not give information on a potential hierarchical relation between the morphological clusters. Also, the current strategy does not allow us to assess whether the tiles of a given cluster influence the decision by their proportion on the slide, their mere presence, or the simultaneous presence of tiles from other clusters. A promising methodological perspective is therefore the improvement of these visualization techniques.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - In-house dataset (Institut Curie)
  - Public dataset (TCGA)
  - Architecture and optimization parameters
  - Strategic sampling
  - Bias score
  - Learning MoCo representations
  - Visualization methods
  - Computation resources
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Technical biases in the Curie dataset
  - Manual validation of the morphological patterns
  - Bias metric significance test

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrm.2022.100872.

ACKNOWLEDGMENTS

The authors thank Hélène Guenon, Saida Sahiri, Martial Caly, and Laure Annette for their help in retrieving the H&E slides and their technical expertise. The authors thank AstraZeneca for the funding of technical time essential for the preparation of the material for the pathology case series. G.B. was supported by a Fondation Curie grant. T.L. was supported by a Q-Life PhD fellowship (Q-life ANR-17-CONV-0005). Furthermore, this work was supported by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

AUTHOR CONTRIBUTIONS

A.V.-S. and G.B. initiated the project. A.V.-S., G.B., F.-C.B., and D.S.-L. generated the patient cohort. G.B. reviewed all the slides. T.P. and M.-H.S.

performed the genomic analyses. T.L., E.D., and T.W. designed the AI and statistical methods. T.L. and P.N. developed the software. T.L. performed the analysis under the supervision of T.W. and E.D. A.V.S. and G.B. interpreted the morphological patterns. A.V.-S., G.B., T.L., E.D., and T.W. discussed methods, results, and design choices. T.L. prepared the figures. T.L., T.W., and A.V.-S. wrote the manuscript and its revisions.

DECLARATION OF INTERESTS

A.V.-S. is a member of the IBEX scientific advisory board. A.V.-S. has received a grant from AstraZeneca to support the technical work to prepare the series of breast cancers analyzed in this series. The authors have filed the patent with PCT application number PCT/EP2022/071130.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: August 10, 2021
Revised: January 4, 2022
Accepted: November 22, 2022
Published: December 13, 2022

REFERENCES

1. Deluche, E., Antoine, A., Bachelot, T., Lardy-Cleaud, A., Dieras, V., Brain, E., Debled, M., Jacot, W., Mouret-Reynier, M.A., Goncalves, A., et al. (2020). Contemporary outcomes of metastatic breast cancer among 22, 000 women from the multicentre ESME cohort 2008–2016. Eur. J. Cancer *129*, 60–70. https://doi.org/10.1016/j.ejca.2020.01.016.
2. Miller, R.E., Leary, A., Scott, C.L., Serra, V., Lord, C.J., Bowtell, D., Chang, D.K., Garsed, D.W., Jonkers, J., Ledermann, J.A., et al. (2020). ESMO recommendations on predictive biomarker testing for homologous recombination deficiency and PARP inhibitor benefit in ovarian cancer. Ann. Oncol. *31*, 1606–1622. https://doi.org/10.1016/j.annonc.2020.08.2102.
3. Bryant, H.E., Schultz, N., Thomas, H.D., Parker, K.M., Flower, D., Lopez, E., Kyle, S., Meuth, M., Curtin, N.J., and Helleday, T. (2005). Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. Nature *434*, 913–917. https://doi.org/10.1038/nature03443.
4. Farmer, H., McCabe, N., Lord, C.J., Tutt, A.N.J., Johnson, D.A., Richardson, T.B., Santarosa, M., Dillon, K.J., Hickson, I., Knights, C., et al. (2005). Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. Nature *434*, 917–921. https://doi.org/10.1038/nature03445.
5. Tung, N.M., Robson, M.E., Ventz, S., Santa-Maria, C.A., Nanda, R., Marcom, P.K., Shah, P.D., Ballinger, T.J., Yang, E.S., Vinayak, S., et al. (2020). Tbcrc 048: phase II study of olaparib for metastatic breast cancer and mutations in homologous recombination-related genes. J. Clin. Oncol. *38*, 4274–4282. https://doi.org/10.1200/JCO.20.02151.
6. Tutt, A.N.J., Garber, J.E., Kaufman, B., Viale, G., Fumagalli, D., Rastogi, P., Gelber, R.D., de Azambuja, E., Fielding, A., Balmaña, J., et al. (2021). Adjuvant olaparib for patients with *BRCA1* - or *BRCA2* -mutated breast cancer. N. Engl. J. Med. *384*, 2394–2405. https://doi.org/10.1056/NEJMoa2105215.
7. Tutt, A., Tovey, H., Cheang, M.C.U., Kernaghan, S., Kilburn, L., Gazinska, P., Owen, J., Abraham, J., Barrett, S., Barrett-Lee, P., et al. (2018). Carboplatin in BRCA1/2-mutated and triple-negative breast cancer BRCAness subgroups: the TNT Trial. Nat. Med. *24*, 628–637. https://doi.org/10.1038/s41591-018-0009-7.
8. Chopra, N., Tovey, H., Pearson, A., Cutts, R., Toms, C., Proszek, P., Hubank, M., Dowsett, M., Dodson, A., Daley, F., et al. (2020). Homologous recombination DNA repair deficiency and PARP inhibition activity in primary triple negative breast cancer. Nat. Commun. *11*, 2662. https://doi.org/10.1038/s41467-020-16142-7.
9. Popova, T., Manié, E., Rieunier, G., Caux-Moncoutier, V., Tirapo, C., Dubois, T., Delattre, O., Sigal-Zafrani, B., Bollet, M., Longy, M., et al. (2012). Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with *BRCA1/2* inactivation. Cancer Res. *72*, 5454–5462. https://doi.org/10.1158/0008-5472.CAN-12-1470.
10. Birkbak, N.J., Wang, Z.C., Kim, J.-Y., Eklund, A.C., Li, Q., Tian, R., Bowman-Colin, C., Li, Y., Greene-Colozzi, A., Iglehart, J.D., et al. (2012). Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. Cancer Discov. *2*, 366–375. https://doi.org/10.1158/2159-8290.CD-11-0206.
11. Abkevich, V., Timms, K.M., Hennessy, B.T., Potter, J., Carey, M.S., Meyer, L.A., Smith-McCune, K., Broaddus, R., Lu, K.H., Chen, J., et al. (2012). Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. Br. J. Cancer *107*, 1776–1782. https://doi.org/10.1038/bjc.2012.451.
12. Polak, P., Kim, J., Braunstein, L.Z., Karlic, R., Haradhavala, N.J., Tiao, G., Rosebrock, D., Livitz, D., Kübler, K., Mouw, K.W., et al. (2017). A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. Nat. Genet. *49*, 1476–1486. https://doi.org/10.1038/ng.3934.
13. Davies, H., Glodzik, D., Morganella, S., Yates, L.R., Staaf, J., Zou, X., Ramakrishna, M., Martin, S., Boyault, S., Sieuwerts, A.M., et al. (2017). HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. Nat. Med. *23*, 517–525. https://doi.org/10.1038/nm.4292.
14. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.L., et al. (2013). Signatures of mutational processes in human cancer. Nature *500*, 415–421. https://doi.org/10.1038/nature12477.
15. Lakhani, S.R., Van De Vijver, M.J., Jacquemier, J., Anderson, T.J., Osin, P.P., McGuffog, L., and Easton, D.F. (2002). The pathology of familial breast cancer: predictive value of immunohistochemical markers estrogen receptor, progesterone receptor, HER-2, and p53 in patients with mutations in BRCA1 and BRCA2. J. Clin. Oncol. *20*, 2310–2318. https://doi.org/10.1200/JCO.2002.09.023.
16. Manié, E., Popova, T., Battistella, A., Tarabeux, J., Caux-Moncoutier, V., Golmard, L., Smith, N.K., Mueller, C.R., Mariani, O., Sigal-Zafrani, B., et al. (2016). Genomic hallmarks of homologous recombination deficiency in invasive breast carcinomas. Int. J. Cancer *138*, 891–900. https://doi.org/10.1002/ijc.29829.
17. Ferrari, A., Vincent-Salomon, A., Pivot, X., Sertier, A.-S., Thomas, E., Tonon, L., Boyault, S., Mulugeta, E., Treilleux, I., MacGrogan, G., et al. (2016). A whole-genome sequence and transcriptome perspective on HER2-positive breast cancers. Nat. Commun. *7*, 12222. https://doi.org/10.1038/ncomms12222.
18. Turner, N.C. (2017). Signatures of DNA-repair deficiencies in breast cancer. N. Engl. J. Med. *377*, 2490–2492. https://doi.org/10.1056/NEJMcibr1710161.
19. Holstege, H., Horlings, H.M., Velds, A., Langerød, A., Børresen-Dale, A.L., van de Vijver, M.J., Nederlof, P.M., and Jonkers, J. (2010). BRCA1-mutated and basal-like breast cancers have similar aCGH profiles and a high incidence of protein truncating TP53 mutations. BMC Cancer *10*, 654. https://doi.org/10.1186/1471-2407-10-654.
20. Veta, M., Diest, P.J.V., Willems, S.M., Wang, H., Madabhushi, A., Cruz-roa, A., Gonzalez, F., Larsen, A.B.L., Vestergaard, J.S., Dahl, A.B., et al. (2014). Assessment of algorithms for mitosis detection in breast cancer histopathology images. Med. Image Anal., 1–23.
21. Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J.A.W.M., the CAMELYON16 Consortium; Hermsen, M., Manson, Q.F., Balkenhol, M., et al. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA *318*, 2199–2210. https://doi.org/10.1001/jama.2017.14585.
22. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., and Fuchs, T.J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. *25*, 1301–1309. https://doi.org/10.1038/s41591-019-0508-1.

23. Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D.A., Barnholtz-Sloan, J.S., Velázquez Vega, J.E., Brat, D.J., and Cooper, L.A.D. (2018). Predicting cancer outcomes from histology and genomics using convolutional networks. Proc. Natl. Acad. Sci. USA *115*, E2970–E2979. https://doi.org/10.1073/pnas.1717139115.

24. Kather, J.N., Heij, L.R., Grabsch, H.I., Loeffler, C., Echle, A., Muti, H.S., Krause, J., Niehues, J.M., Sommer, K.A.J., Bankhead, P., et al. (2020). Pan-cancer image-based detection of clinically actionable genetic alterations. Nat. Cancer *1*, 789–799. https://doi.org/10.1038/s43018-020-0087-6.

25. Coudray, N., Ocampo, P.S., 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. https://doi.org/10.1038/s41591-018-0177-5.

26. Schmauch, B., Romagnoni, A., Pronier, E., Saillard, C., Maillé, P., Calderaro, J., Kamoun, A., Sefta, M., Toldo, S., Zaslavskiy, M., et al. (2020). A deep learning model to predict RNA-Seq expression of tumours from whole slide images. Nat. Commun. *11*, 3877. https://doi.org/10.1038/s41467-020-17678-4.

27. Diao, J.A., Wang, J.K., Chui, W.F., Mountain, V., Gullapally, S.C., Srinivasan, R., Mitchell, R.N., Glass, B., Hoffman, S., Rao, S.K., et al. (2021). Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. Nat. Commun. *12*, 1613. https://doi.org/10.1038/s41467-021-21896-9.

28. Ilse, M., Tomczak, J.M., and Welling, M. (2018). Attention-based deep multiple instance learning. Preprint at arXiv, 180204712 Cs Stat. https://doi.org/10.48550/arXiv.1802.04712.

29. Amores, J. (2013). Multiple instance classification: Review, taxonomy and comparative study. Artif. Intell. *201*, 81–105. https://doi.org/10.1016/j.artint.2013.06.003.

30. Maron, O., and Lozano-Pérez, T. (1998). A framework for multiple-instance learning. In Advances in Neural Information Processing Systems (NeurIPS), M.I. Jordan, M.J. Kearns, and S.A. Solla, eds. (MIT Press), pp. 570–576.

31. Courtiol, P., Tramel, E.W., Sanselme, M., and Wainrib, G. (2017). Classification and disease localization in histopathology using only global labels: a weakly supervised approach. CoRR *1–13*.

32. He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. Preprint at arXiv. 191105722 Cs. https://doi.org/10.48550/arXiv.1911.05722.

33. Valieris, R., Amaro, L., Osório, C.A.B.d.T., Bueno, A.P., Rosales Mitrowsky, R.A., Carraro, D.M., Nunes, D.N., Dias-Neto, E., and Silva, I.T.d. (2020). Deep learning predicts underlying features on pathology images with therapeutic relevance for breast and gastric cancer. Cancers *12*, 3687. https://doi.org/10.3390/cancers12123687.

34. Kather, J.N., Pearson, A.T., Halama, N., Jäger, D., Krause, J., Loosen, S.H., Marx, A., Boor, P., Tacke, F., Neumann, U.P., et al. (2019). Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat. Med. *25*, 1054–1056. https://doi.org/10.1038/s41591-019-0462-y.

35. Schirris, Y., Gavves, E., Nederlof, I., Horlings, H.M., and Teuwen, J. (2021). DeepSMILE: self-supervised heterogeneity-aware multiple instance learning for DNA damage response defect classification directly from H&E whole-slide images. Preprint at arXiv, 2107.09405. https://doi.org/10.48550/arXiv.2107.09405.

36. Kleppe, A., Skrede, O.-J., De Raedt, S., Liestøl, K., Kerr, D.J., and Danielsen, H.E. (2021). Designing deep learning studies in cancer diagnostics. Nat. Rev. Cancer *21*, 199–211. https://doi.org/10.1038/s41568-020-00327-9.

37. Varoquaux, G., Raamana, P.R., Engemann, D.A., Hoyos-Idrobo, A., Schwartz, Y., and Thirion, B. (2017). Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. Neuroimage *145*, 166–179. https://doi.org/10.1016/j.neuroimage.2016.10.038.

38. Zhao, Q., Adeli, E., and Pohl, K.M. (2020). Training confounder-free deep learning models for medical applications. Nat. Commun. *11*, 6010. https://doi.org/10.1038/s41467-020-19784-9.

39. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2017). Men also like shopping: reducing gender bias amplification using corpus-level constraints. Preprint at arXiv 170709457 Cs Stat. https://doi.org/10.48550/arXiv.1707.09457.

40. Adeli, E., Zhao, Q., Pfefferbaum, A., Sullivan, E.V., Fei-Fei, L., Niebles, J.C., and Pohl, K.M. (2020). Representation learning with statistical independence to mitigate bias. Preprint at arXiv. 191003676 Cs. https://doi.org/10.48550/arXiv.1910.03676.

41. Wang, T., Zhao, J., Yatskar, M., Chang, K.-W., and Ordonez, V. (2019). Balanced datasets are not enough: estimating and mitigating gender bias in deep image representations. Preprint at arXiv. 181108489 Cs. https://doi.org/10.48550/arXiv.1811.08489.

42. Wang, Z., Qinami, K., Karakozis, I.C., Genova, K., Nair, P., Hata, K., and Russakovsky, O. (2020). Towards fairness in visual recognition: effective strategies for bias mitigation. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE), pp. 8916–8925. https://doi.org/10.1109/CVPR42600.2020.00894.

43. Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Chen, R.J., Barbieri, M., and Mahmood, F. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. Nat. Biomed. Eng. *5*, 555–570. https://doi.org/10.1038/s41551-020-00682-w.

44. Dehaene, O., Camara, A., Moindrot, O., de Lavergne, A., and Courtiol, P. (2020). Self-supervision closes the gap between weak and strong supervision in histology. Preprint at arXiv, 201203583 Cs Eess. https://doi.org/10.48550/arXiv.2012.03583.

45. Courtiol, P., Maussion, C., Moarii, M., Pronier, E., Pilcer, S., Sefta, M., Manceron, P., Toldo, S., Zaslavskiy, M., Le Stang, N., et al. (2019). Deep learning-based classification of mesothelioma improves prediction of patient outcome. Nat. Med. *25*, 1519–1525. https://doi.org/10.1038/s41591-019-0583-3.

46. Rakha, E.A., El-Sayed, M.E., Reis-Filho, J., and Ellis, I.O. (2009). Pathobiological aspects of basal-like breast cancer. Breast Cancer Res. Treat. *113*, 411–422. https://doi.org/10.1007/s10549-008-9952-1.

47. Cancer Genome Atlas Research Network; Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The cancer genome atlas pan-cancer analysis project. Nat. Genet. *45*, 1113–1120. https://doi.org/10.1038/ng.2764.

48. Ioffe, S., and Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. Preprint at arXiv. 150203167 Cs. https://doi.org/10.48550/arXiv.1502.03167.

49. Kingma, D.P., and Ba, J. (2014). Adam: a method for stochastic optimization. Preprint at arXiv. 14126980 Cs. https://doi.org/10.48550/arXiv.1412.6980.

50. Chen, X., Fan, H., Girshick, R., and He, K. (2020). Improved baselines with momentum contrastive learning. Preprint at arXiv. 200304297 Cs. https://doi.org/10.48550/arXiv.2003.04297.

51. Ruifrok, A.C. Quantification of Histochemical Staining by Color Deconvolution. 21.

52. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

53. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: machine learning in Python. Mach. Learn. Res.PYTHON, 6.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
TCGA	https://portal.gdc.cancer.gov/	N/A
In-house Dataset	Curie Hospital, Paris	N/A
Model predictions	https://data.mendeley.com/datasets/w999vnkdzn/1	https://doi.org/10.17632/w999vnkdzn.1
Software and algorithms		
wsi_mil	https://github.com/trislaz/wsi_mil	https://zenodo.org/badge/latestdoi/373855800
scikit-learn	https://github.com/scikit-learn/scikit-learn	RRID: SCR_019053
openslide-python	https://github.com/openslide/openslide-python	N/A
MoCo	https://github.com/facebookresearch/moco	N/A
SciPy	https://scipy.org/	RRID: SCR_008058

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to the lead contact, Anne Salomon (anne.salomon@curie.fr).

Materials availability

This study did not generate new unique reagents.

Data and code availability

All of the TCGA[47] dataset is available at https://portal.gdc.cancer.gov/.

The in-house dataset consists of confidential medical data not open to the public.

All original code has been deposited at github (https://github.com/trislaz/wsi_mil) and is publicly available as of the date of publication. DOI is available in the key resources table.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

In-house dataset (Institut Curie)

We retrospectively retrieved a series of 715 patients with HE slides of surgical resections specimens of untreated breast cancer and a genomically known HR status (Table S4). The series is composed of 309 Homologous Recombination Proficient tumors (HRP) and 406 Homologous Recombination Deficient tumors (HRD). The HRD status was either identified by the presence of a germline *BRCA1/2* (*gBRCA1/2*) mutation or assessed by LST genomic signature according to Popova et al.[9] for the sporadic triple-negative and luminal cancers.

All patients have been treated and followed at the Institut Curie between 1995 and 2020. The patient agreed for the use of tumor samples from their surgical resection specimens for research according to the law. Ethical approval from the Institutional Review Board (Institut Curie breast cancer study group N°DATA190031) was obtained for the use of all specimens. Clinical data have been retrieved from the Institut Curie electronic medical records and saved using Research electronic data capture (REDCap) tools hosted at the Institut Curie.

Public dataset (TCGA)

This public dataset is composed of 815 WSI of breast cancer fixed in formalin (FFPE) and stained in H&E. They are available at https://portal.gdc.cancer.gov/. Low-resolution WSI, WSI containing artifacts such as large pen marks, tissue-folds and blurred WSI were removed. The final dataset encompasses 673 WSIs. The HR status of the corresponding tumors was obtained using the LST genomic signature.

Architecture and optimization parameters

Hyperparameters have been set thanks to a random search evaluated through 5-fold nested cross-validation. The benchmark task is the prediction of the molecular class of the TCGA WSIs. Both the decision module and the tile-scoring module are multi-layer

perceptrons with batch normalization[48] after each hidden layer. The decision module has 3 hidden layers of 512 neurons, the tile-scoring module has 1 hidden layer of 256 neurons.

Dropout has been fixed at 0.4, the optimizer is ADAM[49] with a learning rate of 3e-3. A batch consists of 16 samples of WSI. A sample of WSI corresponds to a uniform sampling of 300 of its composing tiles. In fact, we observed that this uniform subsampling of the WSIs regularized training as well as diminishes its computational workload. Finally, training is performed during 200 epochs.

Training and performance evaluation are done in a 5-fold nested cross-validation framework.

Each dataset is split into 5 independent folds. For each of these folds, a validation set is randomly sampled in the complementary 4/5th. A model is trained on the remaining dataset (= 4/5 * 4/5 th of the total dataset). This process is repeated 10 times for each test fold, then the 3 best models are selected according to their validation performances, ensembled and finally tested on their test set. This process of model selection and ensembling drives itself a net improvement of the performances (see Figure S1).

Each test and validation set preserves the stratification of the whole dataset with respect to the target variable as well as the confounding variables in case we correct for them. The final performance estimation of the model is the performance averaged over the 5 test performances. During inference time, all the tiles of each WSI are processed.

Strategic sampling

Strategic sampling is used both for balancing the training dataset with respect to the output variable ($T(X) \in \{t_1, t_2, \ldots, t_m\}$) and to correct for biases ($B(X) \in \{b_1, b_2, \ldots, b_n\}$).

If X is a given WSI sampled from the dataset, then $T(X)$ and $B(X)$ are respectively the target value and the bias value of X. We note $|t_k|$ the total number of slides in the dataset labeled with t_k, and $|b_i|$ the total number of slides for which the bias variable takes the value i. $|t_k \& b_i|$ is the total number of slides with label value t_k and bias value b_i.

For achieving both balancing with respect to the output and correcting for biases, we sample the WSIs X in each batch in a distribution P under which

$P(T(X) = t_k) = P(T(X) = t_{k'})$ for all $k \neq k'$.

And.

$P(\{T(X) = t_k\} \cap \{B(X) = b_i\}) = P(\{T(X) = t_{k'}\} \cap \{B(X) = b_i\})$ for all i and $k \neq k'$.

That is, we sample the slide X depending on its target and bias value with probability:

$P(X \mid \{T(X) = t_k\} \cap \{B(X) = b_i\}) \propto \frac{|b_i|}{|t_k \,\&\, b_i|}$ for each $i \leq n$, $k \leq n$

Strategic sampling is performed on the fly when building the batches.

When correcting for several confounders simultaneously, $B \in \{b_1, b_2, \ldots, b_{n_1}\}$ and $C \in \{c_1, c_2, \ldots, c_{n_2}\}$, we simply correct for a new confounder variable that takes values in all combinations of b_i and c_j.

Bias score

We introduce the following notation: for a WSI X_D, sampled in a dataset D under the distribution P_D, $T(X_D)$ is the label of X_D and $B(X_D)$ is the candidate confounder value of X_D (for instance *bouin*).

We want to measure the bias of a predictive algorithm m that outputs, for each X_D, a prediction $m(X_D)$. We moreover define the accuracy Acc_m of m as: $Acc_m = \boldsymbol{E}\,(1_{\{\boldsymbol{m}(\boldsymbol{X_D}) = \boldsymbol{T}(\boldsymbol{X_D})\}})$

The mutual information $MI(B(X_D), m(X_D))$ between $B(X_D)$ and $m(X_D)$ measures the mutual dependence between B and m and highlights the bias of a model.

The idea of the bias score is to compute how far away the predictions of a model are from a perfectly unbiased case.

To simulate this perfectly unbiased case, we subsample (with strategic sampling) a dataset D_i such that $MI(B(X_{D_i}), T(X_{D_i})) = 0$, i.e. such that the target variable and the confounder variable are statistically independent in this dataset.

If m is unbiased, then we should observe that $MI(B(X_{D_i}), m(X_{D_i})) = 0$ too.

In contrast, the more m is biased, the more $MI(B(X_{D_i}), m(X_{D_i})) \geq 0$ will be far away from 0.

In order to obtain a more accurate estimation of the bias score, we iterate this measure over several unbiased datasets $\{D_i\}_{i \leq 30}$. The bias score $BS(B, m)$ is then the average of $MI(B(X_{D_i}), m(X_{D_i}))$ over i.

Because by construction, $BS(B, m)$ is non-negative, we build an unbiased reference m^* such that $P(\mathrm{m}^*(\mathrm{X}) = T(X)) = Acc_m$, and compute its bias-score as a reference value.

Learning MoCo representations

For learning MoCo-v2[50] representation we used the MoCo repository available at https://github.com/facebookresearch/moco.

We randomly used the following transformations: Gaussian blur, crop and resize, color jitter, grayscale, horizontal and vertical symmetries, and a color augmentation in the Hematoxylin and Eosin specific space.[51]

The training dataset is composed of 5.3e6 images of size 224x 224 pixels, or half the Curie dataset at magnification 20x (0.46 μm.px)

We used a Resnet18 and trained it from scratch for 60 epochs on 4 GPU Nvidia Tesla V100 SXM2 32 Go.

We used the SGD optimizer with a momentum of 0.9, a weight decay of 1e-4, a learning rate of 3e-3 and a batch size of 512. We used a cosine scheduler with a warm restart on the learning rate.

Visualization methods

The model used to extract the visualizations has been trained on the luminal subset of the Curie dataset (251 WSI). To benefit from the biggest dataset possible, the model has been trained on the whole dataset, without using early stopping nor testing, during 200 epochs.

To generate the attention-based visualization, the highest ranked tile with respect to the attention score is extracted, for each WSI. The selected tiles are then labeled according to the label of their WSI of origin.

Concerning the decision-based visualization, for each WSI the 300 highest ranked tiles with respect to the attention score are selected. Among this pool of tiles, the 2000 highest ranking tiles with respect to the logit of the posterior probability for HRD and HRP are selected. In order to promote diversity in the extracted images, no more than 20 tiles per slide can be selected.

Computation resources

All computations have been done on the GENCI HPC cluster of Jean-Zay.

QUANTIFICATION AND STATISTICAL ANALYSIS

Technical biases in the Curie dataset

Both technical confounders are related to technical protocols that were modified over time with an unbalanced representation between the HRD and HRP cohorts:

- c_2 corresponds to a change of fixative agent. $c_2 \in \{Bouin, AFA\}$
- c_1 corresponds to a change of impregnation technique. $c_1 \in \{Ethanol, Ethylene\}$.

We performed the exact Fisher test to test for a correlation between:

1. HRD - C1(impregnation): test-statistic 12; p value 3.9e-30
2. HRD - C2 (fixation): test-statistic 31; p value 2.8e-78

Showing the statistical relationship between both confounders and our target variable, the HR status.
Fisher test was performed with the scipy package.[52]

Manual validation of the morphological patterns

The t-test and the Xi2 test performed respectively to test the difference of TILs count and nuclear grade between HRD and HRP tumors were done with the scipy package.

The logistic regression used to predict HRD from the grade and TILs count was implemented with scikit-learn[53] package, with a parameter C = 10, all other parameters set to their default values.

Bias metric significance test

The Mann-Whitney-Wilcoxon test two-sided with Bonferroni correction appearing in the legend of Figure 2 has been performed using the scipy package. The two compared distribution correspond to the mutual information measure iterated over the 30 sub-datasets, as described in the bias score method subsection.