scientific reports

OPEN

Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer

Alona Levy-Jurgenson[1]✉, Xavier Tekpli[2,3], Vessela N. Kristensen[2,3,4] & Zohar Yakhini[1,5]✉

Digital analysis of pathology whole-slide images is fast becoming a game changer in cancer diagnosis and treatment. Specifically, deep learning methods have shown great potential to support pathology analysis, with recent studies identifying molecular traits that were not previously recognized in pathology H&E whole-slide images. Simultaneous to these developments, it is becoming increasingly evident that tumor heterogeneity is an important determinant of cancer prognosis and susceptibility to treatment, and should therefore play a role in the evolving practices of matching treatment protocols to patients. State of the art diagnostic procedures, however, do not provide automated methods for characterizing and/or quantifying tumor heterogeneity, certainly not in a spatial context. Further, existing methods for analyzing pathology whole-slide images from bulk measurements require many training samples and complex pipelines. Our work addresses these two challenges. First, we train deep learning models to spatially resolve bulk mRNA and miRNA expression levels on pathology whole-slide images (WSIs). Our models reach up to 0.95 AUC on held-out test sets from two cancer cohorts using a simple training pipeline and a small number of training samples. Using the inferred gene expression levels, we further develop a method to spatially characterize tumor heterogeneity. Specifically, we produce tumor molecular cartographies and heterogeneity maps of WSIs and formulate a heterogeneity index (HTI) that quantifies the level of heterogeneity within these maps. Applying our methods to breast and lung cancer slides, we show a significant statistical link between heterogeneity and survival. Our methods potentially open a new and accessible approach to investigating tumor heterogeneity and other spatial molecular properties and their link to clinical characteristics, including treatment susceptibility and survival.

Digital pathology, including the automated computer-vision analysis of pathology whole-slide images (WSIs), is fast becoming a game changer in cancer diagnosis and treatment. Deep learning methods have been studied extensively in this context, and were recently shown to be efficient for certain tasks, such as detecting metastases[1–3], immune cells[4–7], mitosis[8,9] and tissue type[4,10,11] as well as for offering clinicians additional insights[12–16]. These achievements led researchers to more recently explore whether such methods could go a step further, and identify molecular traits that are not known to be associated with cell/tissue morphology, such as mutations[17,18], copy-number alterations[18,19], gene expression[18–20] and hormone receptor status[15,19,21].

Alongside these technological advances, the importance of tumor heterogeneity is being increasingly recognized as a major feature associated with resistance to treatments and as a determinant of prognosis[21–25]. Polyclonality and tumor subclones—sub-populations of tumor cells that differ in molecular characteristics such as mutations, copy number aberrations and gene expression profiles—are a hallmark of cancer and may affect

[1]Department of Computer Science, Technion - Israel Institute of Technology, Haifa 32000, Israel. [2]Department of Medical Genetics, Institute of Clinical Medicine, University of Oslo and Oslo University Hospital, Oslo, Norway. [3]Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital, 0310 Oslo, Norway. [4]Division of Medicine, Department of Clinical Molecular Biology and Laboratory Science (EpiGen), Akershus University Hospital, Lørenskog, Norway. [5]Interdisciplinary Center, Arazi School of Computer Science, Herzliya 4610101, Israel. ✉email: levy.alona@gmail.com; zohar.yakhini@gmail.com

treatment outcome and disease progression[7,26–30]. Bulk measurements, which offer high cell-coverage, have the potential to characterize tumor heterogeneity. A key downside, however, is that bulk measurements lack spatial context. Recent work showed that clonal estimation from single-region sampling is less accurate than that obtained by multi-region sampling and that single-region clonal composition estimates vary greatly between techniques[31]. To better capture the variation between tumor regions, a method for characterizing intra-tumor heterogeneity in the spatial context of immunohistochemically (IHC) stained slides was developed[32,33]. Using this method, the spatial heterogeneity of MKI67 was identified as an independent predictor of overall survival in breast cancer patients[32]. Realizing the importance of spatial context, new technologies for spatial transcriptomics have begun to emerge and are being increasingly used by the scientific community alongside other methods to spatially resolve molecular measurements[34–38]. Recently, spatial transcriptomics data collected from 23 breast cancer patients was used to train a deep neural network to predict spatial variation in gene expression[34]. In a cohort of 41 gastric cancer patients, an association between heterogeneity and survival was discovered using a genome-wide single-nucleotide variation array to estimate the number of clones[21]. This was followed by fluorescence in situ hybridization (FISH) to derive clonal locations for 3 samples across 4 target regions. Others combined single cell RNA sequencing (scRNA-seq) with spatial transcriptomics to map and characterize the different cell populations in heterogeneous pancreatic tumors[39]. In multiple myeloma, spatial organization was inferred from scRNA-seq data, using a clustering-based approach, to characterize immunological alterations occurring in the tumor microenvironment during disease progression[40]. Similarly, a computational approach was used to infer spatial position probabilities for individual cells from scRNA-seq data, enabling the spatial reconstruction of single-cell gene expression[41]. Recently, single-cell pathology subgroups were spatially resolved using mass cytometry imaging, covering an average of 2, 246 cells per image across 381 images, to characterize clonal populations in breast cancer[24]. One of their findings associated a specific single-cell pathology subtype that comprised multiple epithelial cell communities with poor survival. In a similar setting, spatially-derived statistics from single-cell data were shown to improve prognostic predictions[23]. These findings emphasize the importance of characterizing tumor heterogeneity in a spatial context. However, single-cell and spatial transcriptomics techniques are expensive and are limited in cell coverage compared to WSIs. While WSIs hold the potential to spatially resolve bulk molecular measurements using complete cell coverage[16,18–20,42], existing pipelines often require multiple modelling steps, large training sets and expert intervention. Furthermore, existing methods for characterizing spatial tumor heterogeneity from WSIs are focused on IHC slides[32,33]. In contrast, inferring tumor heterogeneity directly from H&E WSIs would alleviate the dependency on IHC availability, and would enable the consideration of many molecular traits at once.

In this paper, we present an automated method to both visualize and quantify tumor heterogeneity in the spatial context of WSIs using a simple pipeline, a small number of bulk-labeled training slides and no expert intervention. Applying our approach, we discover a significant link between tumor heterogeneity and survival outcome in breast and lung cancer. Briefly, we train deep neural networks to provide molecular cartographies of mRNA and miRNA expression from WSIs. We use a simple training-inclusion criteria to potentially reduce noise and facilitate model convergence speed and performance. We then use the inferred cartographies to produce heterogeneity maps and to quantify the level of heterogeneity within each WSI using our heterogeneity index (HTI). Applying our methods to breast and lung cancer slides, we show a significant link between heterogeneity and survival. An overview of our method is shown in Fig. 1.

Our main contributions are: (1) A simple pipeline that maps from WSIs to gene expression levels, reaching up to 0.95 AUC on held-out random test sets. Our pipeline uses a single model architecture and requires only a small number of bulk-labeled training slides (with no expert annotations); (2) A method for constructing heterogeneity maps from the inferred gene-expression maps; (3) A heterogeneity index (HTI) that quantifies the level of heterogeneity within the WSI based on the spatial co-location of molecular traits; (4) A statistically significant link between tumor heterogeneity and survival outcome in two cancer cohorts. Our code is available online (see Methods).

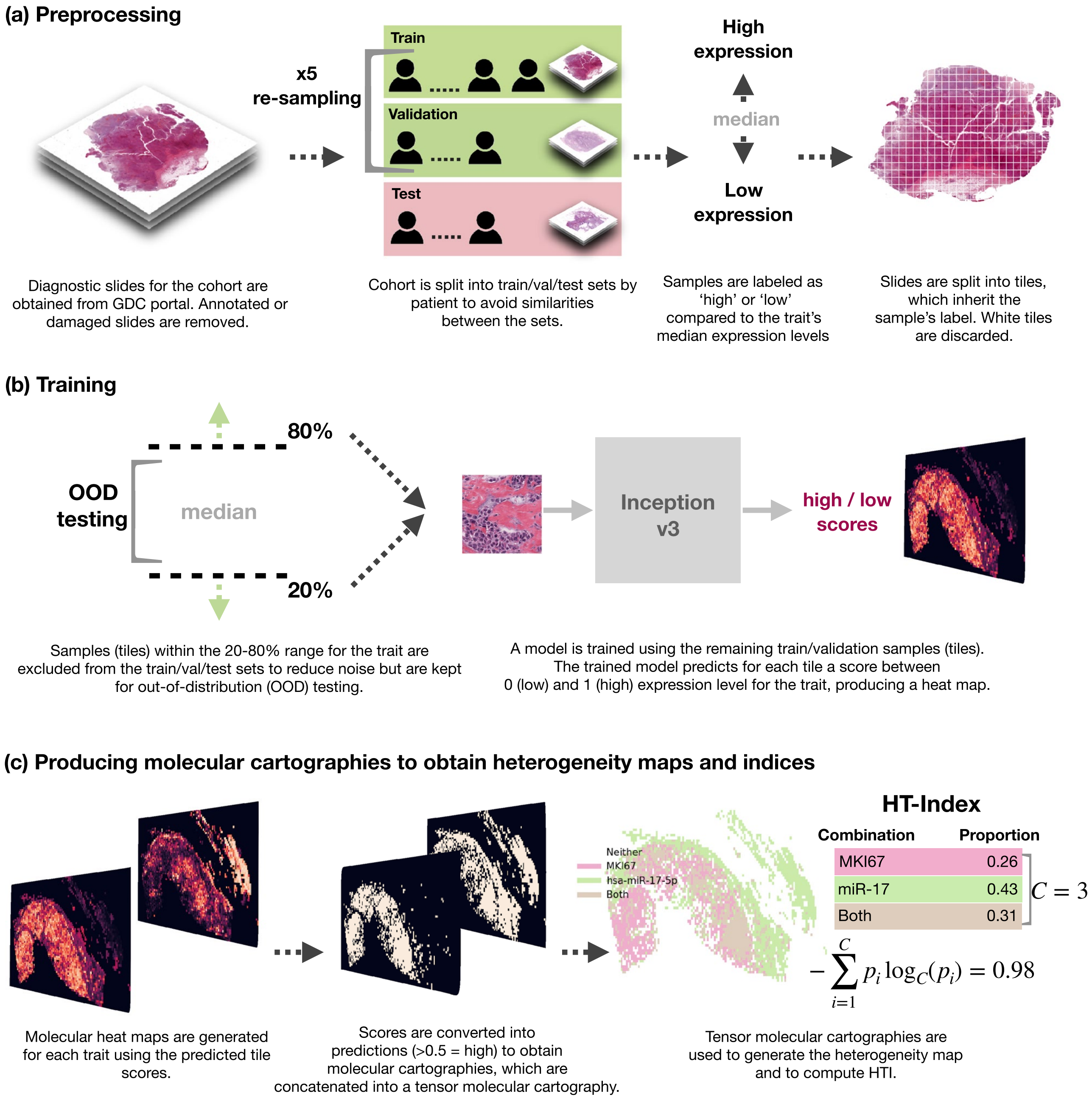
Results

Training models to identify gene expression on pathology whole-slide images.

An overview of our methods is shown in Fig. 1. We begin by obtaining formalin-fixed, paraffin-embedded (FFPE) hematoxylin and eosin (H&E) slides for TCGA BRCA and LUAD samples and matching mRNA and miRNA expression data. We discard any damaged, heavily stained or annotated slides as they can cause the model to focus on artifacts, including clues from pathologists' markings. This results in 761 slides for breast and 469 for lung. We obtain matching normalized mRNA and miRNA expression data for a total of 10 molecular traits for breast and 5 for lung, as listed in Table 1. Breast mRNAs were chosen from the PAM50 genes[43] and lung mRNAs and all miRNAs (breast and lung) were based on the literature (see Methods for further details).

Each cohort is processed, trained and evaluated separately (Fig. 1a, b). We begin by randomly assigning subjects into a held-out test set (10%). We then randomly split the remaining subjects into train (80%) and validation (10%) sets 5 times (bootstrap sampling) to obtain 5 different and randomly selected train-validation sets. Note that these are further reduced in size by a simple inclusion criteria described below. We split on subjects rather than slides so that slides from the same subject are assigned to the same set to avoid similarities between test and train/validation. Importantly, the split is performed before any molecular trait is processed to ensure that all traits see the same split. We then assign each slide into one of the sets according to its subject's association and proceed to label the slides and prepare our training set.

Per trait, we label each slide based on its sample's expression percentile for that trait. We then split each slide into non-overlapping tiles of 512 × 512 at ×20 zoom, resulting in hundreds to thousands of tiles per slide (depending on the slide's size). Following previous methods[17], each tile inherits its labels from its parent slide.



Diagnostic slides for the cohort are obtained from GDC portal. Annotated or damaged slides are removed.

Cohort is split into train/val/test sets by patient to avoid similarities between the sets.

Samples are labeled as 'high' or 'low' compared to the trait's median expression levels

Slides are split into tiles, which inherit the sample's label. White tiles are discarded.

Samples (tiles) within the 20-80% range for the trait are excluded from the train/val/test sets to reduce noise but are kept for out-of-distribution (OOD) testing.

A model is trained using the remaining train/validation samples (tiles). The trained model predicts for each tile a score between 0 (low) and 1 (high) expression level for the trait, producing a heat map.

Molecular heat maps are generated for each trait using the predicted tile scores.

Scores are converted into predictions (>0.5 = high) to obtain molecular cartographies, which are concatenated into a tensor molecular cartography.

Tensor molecular cartographies are used to generate the heterogeneity map and to compute HTI.

Figure 1. Overview of our methods: (**a**) prepossessing, (**b**) model training per trait and (**c**) producing tensor molecular cartographies (multi-layer heatmaps) per slide from which heterogeneity maps and indices are derived. This figure was generated using GIMP[45].

We label each slide (and tiles) based on its percentile expression level: high (1) for those above median low (0) for those below. This raises the concern that not all tiles are representative of its slide label (e.g. a tile representing a healthy portion of a tumor sample), leading to noisy labeling. To potentially reduce noise and facilitate faster model convergence, we set aside complete slides (all corresponding tiles) with expression levels between the 20th and 80th percentiles and use them for out-of-distribution (OOD) evaluation (Fig. 1b). These samples do not participate in the model's train/validation/test sets. This inclusion criteria enables us to efficiently train on a small number of slides: at most 243 slides for BRCA and at most 150 slides for LUAD, depending on the trait (see Methods). We proceed to use the labeled tiles to train our models.

We train each trait separately and repeat this process 5 times (once for each of the random train/validation splits of the non-test samples). We use the Inception v3 classifier[44] as a single end-to-end model that predicts tile scores. This is the only model architecture used to obtain predictions. Additionally, no expert intervention is included at any step.

Once trained, we obtain the trait's validation performances for all 5 runs by measuring the slide-level AUC for each of the five models (we evaluate using slide-level AUC following previous research[17,46]). To do so, we first transition from predictions on tiles to predictions on slides by computing the slide's percent positively classified

		Per-slide AUC			
	Trait	**Test 5-run average (range)**	**Test ensemble**	**OOD-near ensemble**	**OOD-all ensemble**
TCGA-BRCA	miR-17-5p	**0.83 (0.79–0.88)**	**0.87**	**0.62**	0.56
	MKI67	**0.74 (0.53–0.87)**	**0.87**	**0.72**	**0.63**
	FOXA1	**0.7 (0.6–0.76)**	**0.74**	**0.61**	0.56
	MYC	**0.62 (0.58–0.67)**	**0.63**	0.59	0.58
	miR-29a-3p	**0.62 (0.5–0.69)**	**0.67**	**0.65**	0.55
	ESR1	0.58 (0.55–0.63)	0.56	**0.75**	0.57
	CD24	0.57 (0.47–0.67)	0.56	0.57	0.47
	FOXC1	0.53 (0.46–0.57)	0.52	**0.68**	**0.61**
	ERBB2	0.5 (0.46–0.57)	0.53	0.52	0.5
	EGFR	0.5 (0.35–0.67)	0.42	0.57	0.57
TCGA-LUAD	miR-17-5p	**0.85 (0.72–0.95)**	**0.95**	0.43	0.54
	KRAS	**0.64 (0.33–0.9)**	**0.73**	0.56	0.52
	CD274 (PD-L1)	**0.61 (0.57–0.63)**	**0.6**	**0.64**	0.51
	miR-21-5p	0.55 (0.47–0.62)	0.54	**0.75**	**0.63**
	EGFR	0.44 (0.22–0.78)	0.39	0.48	0.46

Table 1. Results for breast and lung cohorts for test and OOD sets. Slides are scored using the percent positively classified tiles. Each of the 5 trained models is evaluated on the held-out test set and the two OOD sets separately. Average AUC results and range (brackets) are shown on the left, followed by ensemble AUC results for test, OOD-near and OOD-all. In bold are AUCs of at least 0.6.

tiles[17]. We then compute the slide-level AUC for each model using these scores as the slide-level prediction. The final model for each of the 5 rounds uses the weights that had the best slide-level AUC on that round's validation set.

Besides providing a better evaluation of our models' performance, a key advantage of training multiple models on different train/validation splits is the option to combine them into an ensemble model[47]. We develop such an ensemble model for each trait by combining its top three models, as measured on their respective validation set, and taking their majority vote. For example, if the last three models from the 5 train/validation repeats yielded the best performances (each on their respective validation set) and their predictions for a given tile are: 0, 0, 1 then the ensemble will predict 0 for that tile by majority vote.

For each trait, we evaluate both its 5 models and its ensemble model on the trait's held-out test set (test slides with expression levels within the 0–20% and 80–100% percentiles for that trait) as well as on its OOD sets (slides within the 20–80% percentiles). Neither of these test sets participated in the trait's model development (train/validation) in any of the 5 rounds.

Achieving high prediction performance for identifying gene expression on pathology whole-slide images.

We obtain high test performance rates across multiple traits. The number of slides in the test set varies across traits, depending on cohort (TCGA-LUAD is roughly half the size of TCGA-BRCA) and on label (expression) availability, and is therefore described per trait in brackets below. In Table 1, we observe that the average test performance measured across all 5 train/validation rounds (first column) obtains high AUCs for over half the traits. Most prominent are miR-17-5p (N = 33), MKI67 (N = 36) and FOXA1 (N = 31) in breast as well as miR-17-5p (N = 13) and KRAS (N = 17) in lung. Especially noteworthy is miR-17-5p as it performed exceptionally well in both cohorts (for which model development and testing is completely separate), further suggesting that mRNA and miRNA expression can be detected from tissue morphology and that this may be applicable for other cancer types.

The use of ensemble models further improves the performance in nearly all traits. miR-17-5p improves from 0.83 to 0.87 AUC in breast and from 0.85 to 0.95 AUC in lung. We observe similar effects for MKI67 and FOXA1 in breast as well as for KRAS in lung. In Fig. 2a (breast) and d (lung) we can see the test-set distribution of the ensembles' slide scores for the ground-truth labels across the different traits. We clearly observe that the images with bulk labels above median have higher slide-level predictions. This is especially true for miR-17-5p (in both cohorts), MKI67, FOXA1, MYC, miR-29a-3p and ESR1 in breast as well as KRAS and CD274 (PDL-1) in lung. We also computed the correlation of our predicted levels to the actual percentiles (Supplementary S1). For example, in breast miR-17 we observe Spearman correlations of 0.63 (FDR *p* value $8e^{-04}$).

To the best of our knowledge, these are the first models to automatically detect miRNA expression levels on H&E whole-slide images. Importantly, our method achieves these results using only a small number of training samples, a single model architecture and no expert knowledge. Model convergence took less than 12 h (worst-case) on a single server with 8 low-cost GPUs (Tesla K80).

Performance for out-of-distribution (OOD) samples.

On top of the potential noise reduction, an added advantage in setting aside the OOD slides is the ability to further challenge our models on a large number

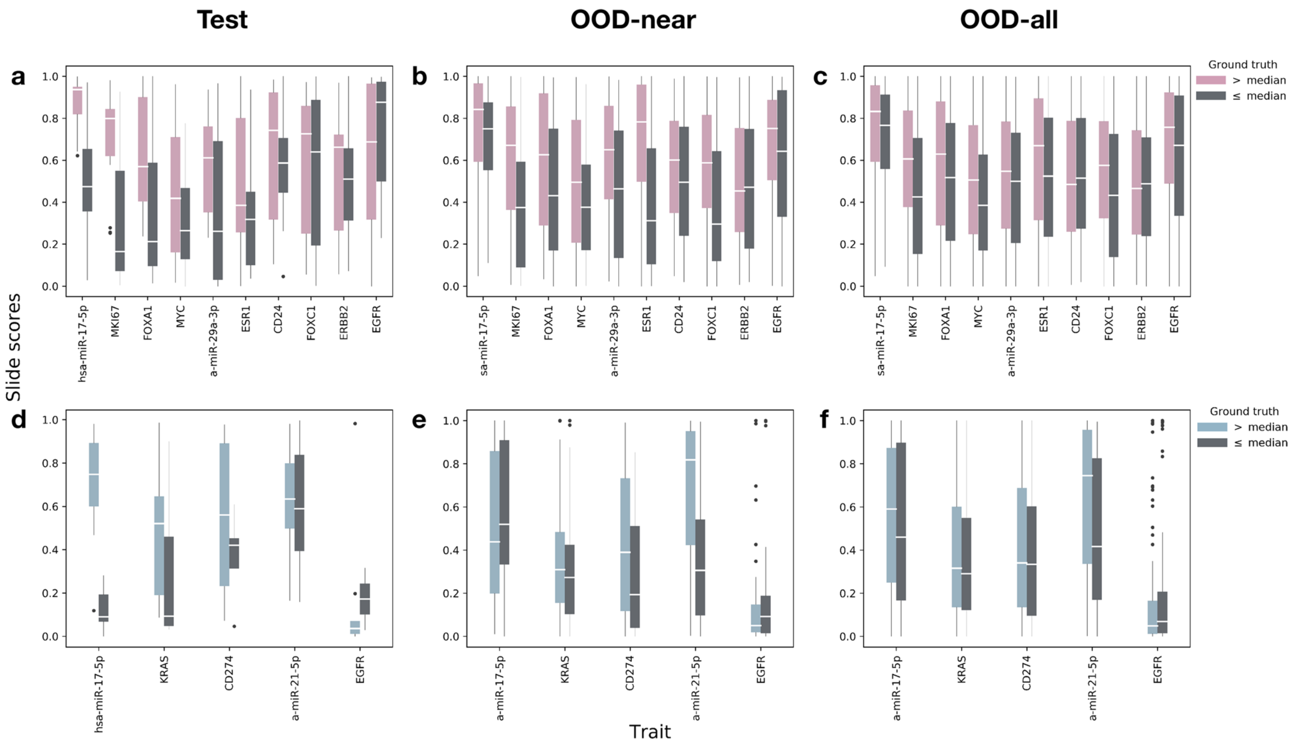


Figure 2. WSI score distribution for test, OOD-near and OOD-all sets compared to the actual bulk-measured values (lower values in gray). Scores are computed using the ensemble model for each trait (percent positive tiles by majority vote of the best three). Middle white lines reflect the median slide score. Whiskers extend to $1.5 \times IQR$. (**a**)–(**c**) breast test, OOD-near and OOD-all respectively; (**d**)–(**f**) lung test, OOD-near and OOD-all (respectively). In the test set, especially worth noting are miR-17-5p in both cohorts, MKI67, FOXA1, MYC, mir-29a-3p and ESR1 in breast as well as KRAS and CD274 (PD-L1) in lung. In the OOD sets most notable are MKI67, FOXA1, MYC, ESR1 and FOXC1 in breast as well as miR-17-5p and miR-21-5p in lung.

of out-of-distribution samples. This is important, especially since certain traits had a relatively small held-out test set. We use the ensemble models for each trait. In Table 1 we report two types of AUC results on the OOD set: one for the next decile tier after the test set's percentiles, i.e. 0.2–0.3 and 0.7–0.8, which we designate as near-distribution (OOD-near) and another for the full OOD set (OOD-all). As before, the number of slides available in the two OOD sets depends on the cohort and the intersection between slide and label availability for the trait. In the OOD-near set, breast has between 140 and 158 and lung has between 79 and 103 slides (roughly 20% from each cohort's slides). The OOD-all set has between 423 and 454 slides for breast, and between 256 and 283 for lung (roughly 60% of each cohort's slides—the remainder after using the top and bottom 20% for model development). We are able to test our models on such large test sets since, by design, none of these slides were included in the trait's train/validation/test sets. Our results demonstrate that the models extend to out-of-distribution samples relatively well. This can also be observed in Fig. 2 b, c (breast) and e, f (lung) which depict the distribution of scores per label for the OOD-near and OOD-all sets. This is especially evident in the breast cancer cohort. For example, the score distributions for MKI67 for the "above median" (pink) class tend to be higher than those for the "below median" (gray) class. As before, we also computed the correlation of our OOD prediction levels to the actual percentiles (Supplementary S1). The overall better breast results are likely due to the larger amount of data available for training, emphasizing the importance of collecting larger data sets.

Tensor molecular cartographies of gene expression as a window to tumor heterogeneity. After confirming the performance of our models, we proceed towards our goal of analyzing tumor heterogeneity (Fig. 1c). We do so by producing, for each slide, multiple molecular cartographies, each representing a single molecular trait. We then combine the molecular cartographies into a tensor molecular cartography (multi-layer heatmaps) to obtain a richer spatial representation of the molecular traits detected within the slide. Creating a molecular cartography for a single trait is straight forward: we simply spatially arrange the binary predictions of a slide's tiles in a matrix so that the position of a tile's score in the matrix corresponds to its position in the pre-tiled slide, as illustrated in (Fig. 1c left). Once we obtain several molecular cartographies for a single slide, we stack them into a tensor molecular cartography (Fig. 1c middle) that now represents the slide across several traits.

We use the tensor molecular cartography to visualize heterogeneity—this will later serve us to confirm our method for quantifying heterogeneity. As shown in Fig. 1c (right), we do so by identifying which tiles are positive for each possible combination of traits and assign different colors to each combination. For example, using two traits, A and B, we obtain a tensor of depth 2 (one molecular cartography per trait) and each tile is colored as

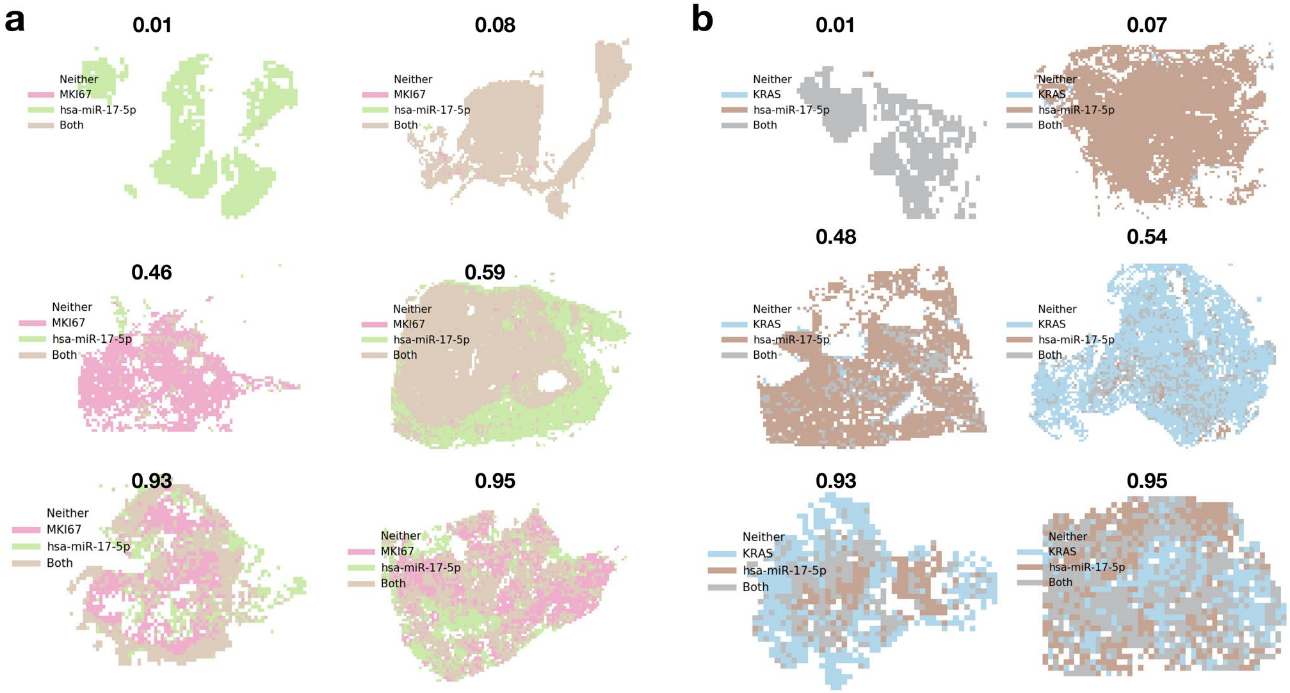


Figure 3. Heterogeneity maps and corresponding HTIs. (**a**) Breast cancer cohort with traits MKI67 (pink) and miR-17 (green). Brown indicates that both traits manifest (both models predicted positive). (**b**) Lung cancer cohort with traits KRAS (blue) and miR-17 (brown). Gray indicates both traits manifest. Corresponding HTIs appear directly above. Rows appear in increasing HTI order from top to bottom (top 0–0.1, middle 0.4–0.6, bottom 0.9–1).

one of three options: A (only), B (only) and Both. Figure 3 shows examples for such heterogeneity maps, along with their level of heterogeneity, as obtained using HTI described below. Such molecular maps of pathological images could be used by pathologists, in addition to routinely stained diagnostic markers, to further classify and identify cancer subtypes, potentially leading to better informed clinical decisions.

Quantifying tumor heterogeneity from tensor molecular cartographies.

As intra-tumor heterogeneity has been proposed as an obstacle to effective treatment and cancer eradication, we propose an approach to compute and quantify the level of heterogeneity in a given image from its tensor molecular cartography. We used a variation of Shannon's entropy, commonly used to measure diversity and heterogeneity in various settings[24,48]. Formally, we compute:

$$HTI = -\sum_{i=1}^{C} p_i \log_C(p_i)$$

where C is the maximum number of non-empty trait combinations that may be observed on a slide and equals $2^{|traits|} - 1$ (the number of subsets excluding the empty set), and p_i is the proportion of tiles for which exactly all models in combination i provided a positive prediction.

For example, given a tensor molecular cartography for two traits A and B (e.g. FOXA1 and MKI67), $C = 3$ (3 possible non-empty trait combinations: A (only), B (only) and Both). If the slide is homogeneous with nearly all of its tiles falling into one of these 3 options (say Both), then $p_{Both} \approx 1, p_A \approx 0, p_B \approx 0$ (each of the single-trait molecular cartographies is nearly all 1s), resulting in an HTI of 0. If, however the slide is heterogeneous with 1/3rd of the tiles falling into each option then: $p_{Both} = 1/3, p_A = 1/3, p_B = 1/3$ and we obtain an HTI of 1. The logarithm base C guarantees that HTI $\in [0, 1]$. If A and B are two molecular traits (e.g. FOXA1 and MKI67), a high HTI would reflect there may be two subclones whereas a low HTI would reflect single clonal dominance.

We note that HTI can be applied to any set of binary matrices (or vectors) of identical shape, each of which describes the presence of a single trait. As such, it can be used in other settings involving the localization of clinically relevant traits.

Figure 3 depicts several heterogeneity maps and their associated HTIs.

A statistically significant link between tumor heterogeneity and survival.

We sought to understand whether tumor heterogeneity is linked to survival outcome and whether the link can be inferred through the spatial analysis of pathology whole-slide images. To do so, we combine the test and OOD slides and split them into high and low heterogeneity groups based on their HTIs. Specifically, for a given cohort, we first generate tensor molecular cartographies per slide using the ensemble models of the top two performing molecular

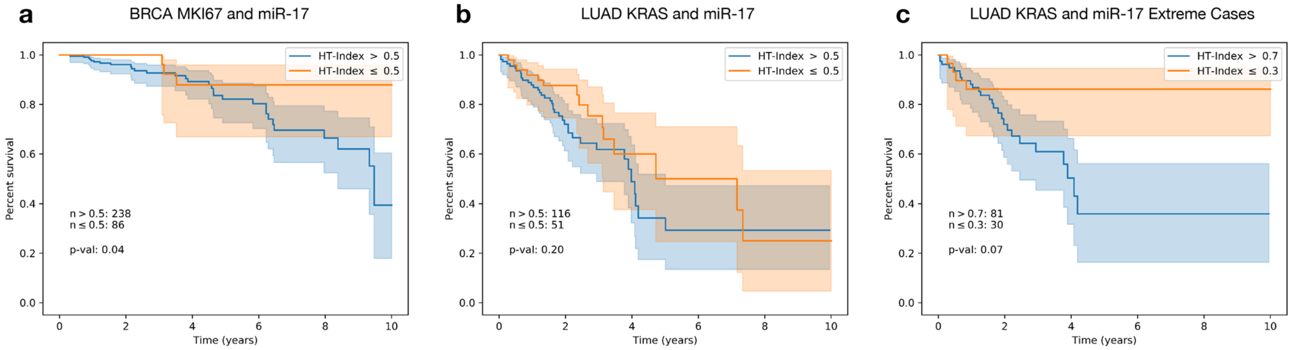


Figure 4. Survival analysis with respect to HTI derived from two traits in breast and lung cancer. For each trait combination in (**a**) and (**b**), slides were split into high and low HTI (> 0.5 and ≤ 0.5 or blue and orange respectively). In (**c**) slides were split into > 0.7 and ≤ 0.3 HTI. Each survival curve is shown with a 95% confidence interval.

traits for that cohort (from Table 1). We then compute HTI for each slide and split the slides into two groups: > 0.5 and ≤ 0.5 (high and low heterogeneity respectively). We perform survival analysis on these groups using Mantel's log-rank test and Kaplan-Meier curves. We use only slides in the combined set of OOD and test slides. Since OOD slides are trait-dependent, we keep only those slides that were designated as OOD for both traits.

Figure 4 describes the survival analysis results for each cohort using the top two traits by test performance from Table 1 in each (MKI67 and miR-17 for breast and miR-17 and KRAS for lung). In breast, survival is significantly worse for subjects exhibiting high heterogeneity (blue) compared to those with low heterogeneity (orange), with a log-rank *p* value of 0.04 (Fig. 4a). In lung, we observe significant differences in survival between higher and lower HTIs when heterogeneity differences are more distinct (> 0.7 vs. ≤ 0.3), with a log-rank *p* value of 0.07 (Fig. 4c).

Discussion

This work offers a method for analyzing tumor heterogeneity from the rich spatial data available in H&E WSIs. Using deep learning we created high resolution maps of multiple mRNA and miRNA expression levels within a whole-slide image and combined these maps into a tensor molecular cartography. We then used the tensor cartographies to spatially visualize and quantify tumor heterogeneity in the form of heterogeneity maps and HTI scores. While other methods, such as single-cell profiling and spatial transcriptomics, can also be used to infer heterogeneity, they are expensive, may lack sufficient spatial context and only cover a few thousand cells.

We applied our method to both breast and lung cancer pathology whole-slide images (H&E). We trained several models per trait and tested each of these models on both a randomly sampled held-out test set and two large out-of-distribution sets, containing hundreds of WSIs that the models have never before encountered. Test results show that several mRNA and miRNA can be identified and localized automatically within whole-slide images with high AUCs. Furthermore, our results demonstrate this can be achieved using only a small number of training slides, a single model architecture and no expert intervention. Especially notable are our results for miR-17, which obtained high AUCs (up to 0.95) in both lung and breast. This is interesting in light of indications that miR-17 is over-expressed across many cancer types[49].

Using our models, we generated a tensor molecular cartography for each slide, enabling us to both visualize the distribution of traits, in the form of heterogeneity maps, and to compute HTI. By representing each patient with their HTI and performing survival analysis, we showed that high tumor heterogeneity is significantly linked to poor survival, especially in breast cancer. We stress that this link cannot be identified through obvious means by directly using expression levels (Fig. 5) or through the PAM50[43] breast cancer types, which are not associated with HTI (Fig. 6). This analysis highlights the potential clinical value in producing tensor molecular cartographies and heterogeneity maps from H&E WSIs.

Our methods open a window to further analyses. For example, from a molecular biology perspective, generating heterogeneity maps of miRNAs from the same family may be interesting in light of recent findings showing they are context (e.g. tissue) dependent[50]. Similarly, miRNA/mRNA relationships[51,52] may be analyzed from a spatial perspective. From a technical aspect, it may be interesting to explore whether transfer learning between molecular traits or cohorts is possible and to what extent. Also a multilabel approach may be possible, although it may require larger datasets and careful label-balancing to obtain satisfactory results across all traits.

As heterogeneity plays an increasingly key role in cancer treatment, providing researchers and practitioners with a solution to view the distribution of clinically relevant traits that are not currently visible on WSIs may be of great value. Since H&E slides are a standard component of routine diagnostic protocols, a natural solution may be offered by the fast and automated digital mapping of the molecular landscape within H&E slides. One such solution is offered through our simple pipeline for producing tensor molecular cartographies and heterogeneity maps from WSIs.

As our approach requires a single model architecture, uses a small number of training slides and does not call for expert annotations, it can apply to many more traits and in broader contexts. As a future direction we hope to combine our approach with single-cell data and spatially resolved transcriptomics data to obtain finer resolution

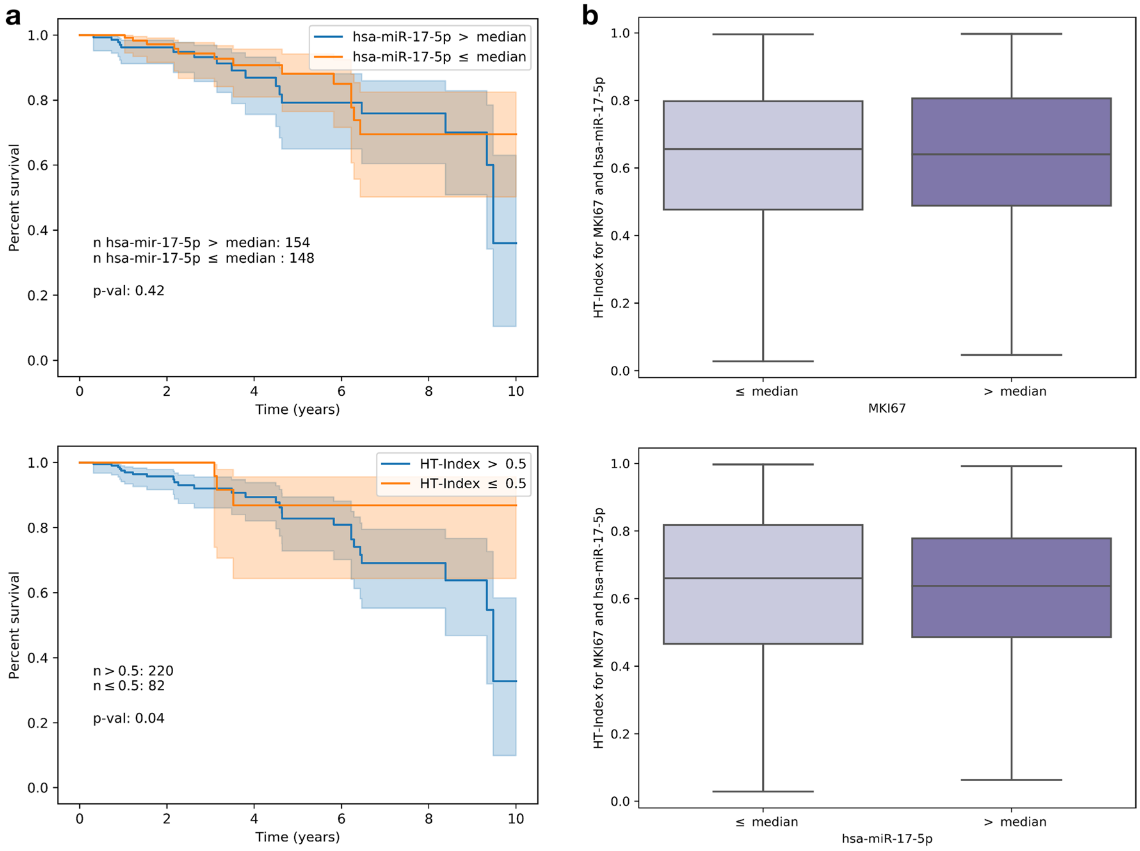


Figure 5. Left panel (**a**) Survival analysis using bulk-measured expression levels for miR-17 (top) compared to our approach (bottom) when applied to the same patient basis available for both. Right panel (**b**) distribution of HTI for above and below median ground-truth expression levels for MKI67 (top) and miR-17 (bottom).

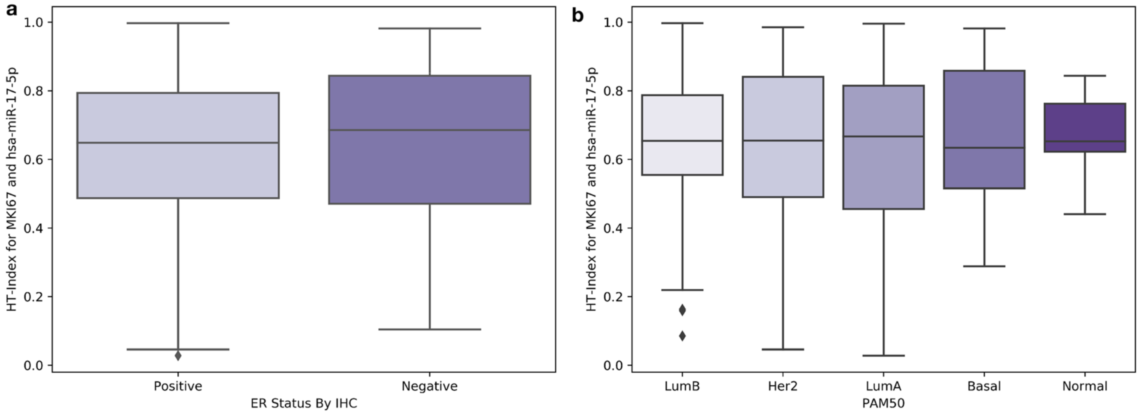


Figure 6. Distribution of HTI per ER status (**a**) and PAM50 type (**b**) in the data used for survival analysis.

mapping, improving the relevance of both H&E and transcriptomics. Molecular cartography from H&E potentially enables a new approach to investigating tumor heterogeneity and other spatial molecular properties and their link to clinical characteristics. An interesting aspect of such endeavors will be to link spatial properties to treatment susceptibility and precision care.

Methods

Data. All whole-slide images are available online at the GDC repository (https://portal.gdc.cancer.gov/repository), by selecting *Diagnostic Slide* under *Experimental Strategy* for the relevant project (e.g. TCGA-BRCA). Matching expression levels were obtained from the GDC's website at: https://gdc.cancer.gov/about-data/publications/pancanatlas (see RNA and miRNA). Matching survival data were obtained from cBioportal at: http://

www.cbioportal.org/datasets from the following files: "Breast Invasive Carcinoma (TCGA, Firehose Legacy)" and "Lung Adenocarcinoma (TCGA, Firehose Legacy)".

Preprocessing. Damaged or marked slides were removed, leaving 761 slides for BRCA and 469 slides for LUAD. Remaining slides were split into non-overlapping 512 × 512 tiles at a magnification level of 20×. Tiles with more than 50% background were removed.

Training. All of our models were developed in TensorFlow[53] using the Inception v3 classifier[44] with the last layer modified to one output. All models are trained from random initialization, following previous work showing improved performance by fully training Inception v3[17]. Each of the 5 train/validation rounds is trained on mini-batches of labeled tiles from the training set using the Adam optimizer[54]. Models were evaluated on the labeled validation tiles every 1/16th epoch (full pass on all training tiles) to avoid overfitting caused by tile similarities between mini-batches (each slide contains hundreds to thousands of tiles, many of which are likely to be similar to one another). No data augmentations were performed, except for random horizontal and vertical flips of the training tiles to further reduce overfitting. Learning rate started at 0.001 and was decayed when performance on the validation set plateaued for 10 steps, with early stopping after 30 steps of no validation improvement. Each final trained model used the weights that performed best on its validation set.

To potentially reduce noise and facilitate faster model convergence, we set aside complete slides (all corresponding tiles) with expression levels between the 20th and 80th percentiles. These were used for out-of-distribution evaluation and did not participate in the model's train/validation/test sets. This enabled us to efficiently train on a small number of slides. For example, if labels are available for all 761 slides, taking 80% for training out of which 40% are not OOD, leaves us with $761 \times 0.8 \times 0.4 = 243$ training slides for BRCA and $469 \times 0.8 \times 0.4 = 150$ for LUAD.

Each model was trained on a single Tesla K80 machine with 8 GPUs and took at most 12 hours until convergence on the validation set (when the aforementioned early stopping was invoked). Mini-batch size per GPU replica was 18, for a total of $18 \times 8 = 144$ tiles per training step.

mRNA and miRNA selection. **mRNAs:** For breast, we sorted PAM50[43] genes by their expression variance in our dataset and selected from the top (highest variance). CD24 is not in PAM50, but is one of the highest expression-varying mRNAs in the cohort and is over-expressed in many cancers[55]. For lung, we based our selection on previous research, e.g.[17,56–58].

miRNAs: Selection of miRs is based on previous work identifying miR-17 and miR-29 among the top up-regulated and down-regulated miRNAs (respectively) in breast cancer[52]. Others have also associated miR-17[59–61] and miR-29[62,63] with breast cancer. miR-17 was also chosen for lung since the miR-17 family was shown to be universally over-expressed in many cancers, including lung[49], and has been directly associated with lung cancer[64–66] as has miR-21[67,68].

Survival analysis. Each patient is represented by a single HTI to perform survival analysis. Where patients were associated with more than one whole-slide image (e.g. a patient with diagnostic slides DX1 and DX2), the DX1 slide was used to determine HTI. Analysis was performed using the Lifelines package for Python (https://lifelines.readthedocs.io/en/latest/).

Code availability. The code used for this work is publicly available under: https://github.com/alonalj/PathoMCH.

Received: 31 August 2020; Accepted: 14 October 2020
Published online: 02 November 2020

References

1. Campanella, G. *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1301–1309 (2019).
2. Liu, Y. *et al.* Detecting cancer metastases on gigapixel pathology images. arXiv preprint arXiv:1703.02442 (2017).
3. Hartman, D. J., Van Der Laak, J. A., Gurcan, M. N. & Pantanowitz, L. Value of public challenges for the development of pathology deep learning algorithms. *J. Pathol. Inform.* **11**, 666 (2020).
4. Swiderska-Chadaj, Z. *et al.* Learning to detect lymphocytes in immunohistochemistry with deep learning. *Med. Image Anal.* **58**, 101547 (2019).
5. Narayanan, P. L. *et al.* Unmasking the tissue microecology of ductal carcinoma in situ with deep learning. *BioRxiv* **812735**, 666 (2019).
6. Klauschen, F. *et al.* Scoring of tumor-infiltrating lymphocytes: From visual estimation to machine learning. In *Seminars in Cancer Biology*, vol. 52, 151–157 (Elsevier, Amsterdam, 2018).
7. Saltz, J. *et al.* Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Rep.* **23**, 181–193 (2018).
8. Balkenhol, M. C. *et al.* Deep learning assisted mitotic counting for breast cancer. *Lab. Investig.* **99**, 1596–1606 (2019).
9. Saha, M., Chakraborty, C. & Racoceanu, D. Efficient deep learning model for mitosis detection using breast histopathology images. *Comput. Med. Imaging Graph.* **64**, 29–40 (2018).
10. Hermsen, M. *et al.* Deep learning-based histopathologic assessment of kidney tissue. *J. Am. Soc. Nephrol.* **30**, 1968–1979 (2019).
11. Bulten, W. *et al.* Epithelium segmentation using deep learning in h&e-stained prostate specimens with immunohistochemistry as reference standard. *Sci. Rep.* **9**, 1–10 (2019).

12. Hägele, M. *et al.* Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. *Sci. Rep.* **10**, 1–12 (2020).
13. Yamamoto, Y. *et al.* Automated acquisition of explainable knowledge from unannotated histopathology images. *Nat. Commun.* **10**, 1–9 (2019).
14. Zhang, Z. *et al.* Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nat. Mach. Intell.* **1**, 236–245 (2019).
15. Seegerer, P. *et al.* Interpretable deep neural network to predict estrogen receptor status from haematoxylin-eosin images. In *Artificial Intelligence and Machine Learning for Digital Pathology*, 16–37 (Springer, Berlin, 2020).
16. AbdulJabbar, K. *et al.* Geospatial immune variability illuminates differential evolution of lung adenocarcinoma. *Nat. Med.* https://doi.org/10.1038/s41591-020-0900-x (2020).
17. 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).
18. Fu, Y. *et al.* Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat. Cancer* **1**, 800–810 (2020).
19. Kather, J. N. *et al.* Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat. Cancer* **1**, 789–799 (2020).
20. Schmauch, B. *et al.* A deep learning model to predict RNA-seq expression of tumours from whole slide images. *Nat. Commun.* **11**, 1–15 (2020).
21. Chao, J. *et al.* Association between spatial heterogeneity within nonmetastatic gastroesophageal adenocarcinomas and survival. *JAMA Netw. Open* **3**, e203652–e203652 (2020).
22. Oesper, L., Mahmoody, A. & Raphael, B. J. THetA: Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol* **14**, R80 (2013).
23. Ali, H. R. *et al.* Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer. *Nat. Cancer* **1**, 163–175 (2020).
24. Jackson, H. W. *et al.* The single-cell pathology landscape of breast cancer. *Nature* **578**, 615–620 (2020).
25. Prasetyanti, P. R. & Medema, J. P. Intra-tumor heterogeneity from a cancer stem cell perspective. *Mol. Cancer* **16**, 41 (2017).
26. Chang, H. J. *et al.* Discordant human epidermal growth factor receptor 2 and hormone receptor status in primary and metastatic breast cancer and response to trastuzumab. *Jpn. J. Clin. Oncol.* **41**, 593–599 (2011).
27. Bedard, P. L., Hansen, A. R., Ratain, M. J. & Siu, L. L. Tumour heterogeneity in the clinic. *Nature* **501**, 355–364 (2013).
28. Scheffler, M. *et al.* Spatial tumor heterogeneity in lung cancer with acquired epidermal growth factor receptor-tyrosine kinase inhibitor resistance: targeting high-level MET-amplification and EGFR T790M mutation occurring at different sites in the same patient. *J. Thorac. Oncol.* **10**, e40–e43 (2015).
29. de Bruin, E. C. *et al.* Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346**, 251–256 (2014).
30. Schwarz, R. F. *et al.* Spatial and temporal heterogeneity in high-grade serous ovarian cancer: A phylogenetic analysis. *PLoS Med.* **12**, e1001789 (2015).
31. Bhandari, V. *et al.* The inter and intra-tumoural heterogeneity of subclonal reconstruction. *bioRxiv* 418780 (2019).
32. Laurinavicius, A. *et al.* Bimodality of intratumor ki67 expression is an independent prognostic factor of overall survival in patients with invasive breast carcinoma. *Virchows Archiv* **468**, 493–502 (2016).
33. Plancoulaine, B. *et al.* A methodology for comprehensive breast cancer ki67 labeling index with intra-tumor heterogeneity appraisal based on hexagonal tiling of digital image analysis data. *Virchows Archiv* **467**, 711–722 (2015).
34. He, B. *et al.* Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat. Biomed. Eng.* **666**, 1–8 (2020).
35. Chen, W.-T. *et al.* Spatial transcriptomics and in situ sequencing to study Alzheimer's disease. *Cell* **182**, 976–991 (2020).
36. Li, G. & Neuert, G. Multiplex RNA single molecule fish of inducible MRNAS in single yeast cells. *Sci. Data* **6**, 1–9 (2019).
37. Hoang, M. *et al.* In situ rna expression profiling of 1600+ immuno-oncology targets in ffpe tissue using nanostring geomx^{TM} digital spatial profiler (2019).
38. Burgess, D. J. Spatial transcriptomics coming of age. *Nat. Rev. Genet.* **20**, 317 (2019).
39. Moncada, R. *et al.* Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat. Biotechnol.* **38**, 333–342 (2020).
40. Zavidij, O. *et al.* Single-cell RNA sequencing reveals compromised immune microenvironment in precursor stages of multiple myeloma. *Nat. Cancer* **1**, 493–506 (2020).
41. Nitzan, M., Karaiskos, N., Friedman, N. & Rajewsky, N. Gene expression cartography. *Nature* **576**, 132–137 (2019).
42. Heindl, A. *et al.* Relevance of spatial heterogeneity of immune infiltration for predicting risk of recurrence after endocrine therapy of er+ breast cancer. *J. Natl. Cancer Inst.* **110**, 166–175 (2018).
43. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160 (2009).
44. 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).
45. GIMP: GNU Image Manipulation Program Version 2.10. https://www.gimp.org/. Accessed 19 May 2020.
46. Yu, K.-H. *et al.* Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat. Commun.* **7**, 12474 (2016).
47. Rokach, L. Ensemble-based classifiers. *Artif. Intell. Rev.* **33**, 1–39 (2010).
48. Chao, A. & Shen, T.-J. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environ. Ecol. Stat.* **10**, 429–443 (2003).
49. Navon, R. *et al.* Novel rank-based statistical methods reveal microRNAs with differential expression in multiple cancer types. *PLoS ONE* **4**, e8003 (2009).
50. Dhawan, A., Scott, J. G., Harris, A. L. & Buffa, F. M. Pan-cancer characterisation of microRNA across cancer hallmarks reveals microRNA-mediated downregulation of tumour suppressors. *Nat. Commun.* **9**, 1–13 (2018).
51. Miles, G. D., Seiler, M., Rodriguez, L., Rajagopal, G. & Bhanot, G. Identifying microRNA/mRNA dysregulations in ovarian cancer. *BMC Res. Notes* **5**, 164 (2012).
52. Enerly, E. *et al.* miRNA-mRNA integrated analysis reveals roles for miRNAs in primary breast tumors. *PLoS ONE* **6**, e16915 (2011).
53. Abadi, M. *et al.* Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, (265–283) (2016).
54. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).
55. Fang, X., Zheng, P., Tang, J. & Liu, Y. CD24: From A to Z. *Cell. Mol. Immunol.* **7**, 100–103 (2010).
56. Ladanyi, M. & Pao, W. Lung adenocarcinoma: Guiding EGFR-targeted therapy and beyond. *Mod. Pathol.* **21**, S16–S22 (2008).
57. Prior, I. A., Lewis, P. D. & Mattos, C. A comprehensive survey of Ras mutations in cancer. *Cancer Res.* **72**, 2457–2467 (2012).
58. Budczies, J. *et al.* Pan-cancer analysis of copy number changes in programmed death-ligand 1 (PD-L1, CD274)-associations with gene expression, mutational load, and survival. *Genes Chromosom. Cancer* **55**, 626–639 (2016).
59. Hossain, A., Kuo, M. T. & Saunders, G. F. miR-17-5p regulates breast cancer cell proliferation by inhibiting translation of AIB1 mRNA. *Mol. Cell. Biol.* **26**, 8191–8201 (2006).
60. Li, H., Bian, C., Liao, L., Li, J. & Zhao, R. C. mir-17-5p promotes human breast cancer cell migration and invasion through suppression of HBP1. *Breast Cancer Res. Treat.* **126**, 565–575 (2011).

61. Calvano Filho, C. M. C. *et al.* Triple-negative and luminal A breast tumors: Differential expression of miR-18a-5p, miR-17-5p, and miR-20a-5p. *Tumor Biol.* **35**, 7733–7741 (2014).
62. Cittelly, D. M. *et al.* Progestin suppression of mir-29 potentiates dedifferentiation of breast cancer cells via KLF4. *Oncogene* **32**, 2555–2564 (2013).
63. Wu, Z., Huang, X., Huang, X., Zou, Q. & Guo, Y. The inhibitory role of miR-29 in growth of breast cancer cells. *J. Exp. Clin. Cancer Res.* **32**, 98 (2013).
64. Hayashita, Y. *et al.* A polycistronic microRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation. *Cancer Res.* **65**, 9628–9632 (2005).
65. Zhang, B., Chen, M., Jiang, N., Shi, K. & Qian, R. A regulatory circuit of circ-MTO1/miR-17/QKI-5 inhibits the proliferation of lung adenocarcinoma. *Cancer Biol. Ther.* **20**, 1127–1135 (2019).
66. Cho, W. C. MicroRNAs as therapeutic targets for lung cancer. *Expert Opin. Therap. Targets* **14**, 1005–1008 (2010).
67. Zheng, W. *et al.* MicroRNA-21: A promising biomarker for the prognosis and diagnosis of non-small cell lung cancer. *Oncol. Lett.* **16**, 2777–2782 (2018).
68. Peng, Z. *et al.* Identification of microRNAs as potential biomarkers for lung adenocarcinoma using integrating genomics analysis. *Oncotarget* **8**, 64143 (2017).

Acknowledgements

Cloud computation for this project was partially funded by Google Cloud Platform. We thank the Technion Computer Science Faculty for generously supporting ALJ. This project received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 847912. We thank the Yakhini Research Group for important discussions and input. We thank Doron Lipson for critical reading and important suggestions. We thank Øystein Garred for helpful discussions and for clarifying important pathology related aspects.

Author contributions

A.L.J. and Z.Y. designed the study and developed the methods. A.L.J. performed all data analysis. X.T. and V.K. performed follow up analysis. X.T. curated data and prepared it for analysis. Z.Y. supervised the study. A.L.J. and Z.Y. wrote the manuscript with contributions from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-75708-z.

Correspondence and requests for materials should be addressed to A.L.-J. or Z.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.