nature portfolio

https://doi.org/10.1038/s41588-021-00911-1

Supplementary information

A single-cell and spatially resolved atlas of human breast cancers

Supplementary Note

Materials and Methods

Tissue dissociation

Samples collected in this study (Supplementary Table 1) were analyzed from fresh surgical resections and cryopreserved tissue as previously described[1]. Tumors were mechanically and enzymatically dissociated using Human Tumor Dissociation Kit (Miltenyi Biotec), following the manufacturer's protocol. For cryopreserved tissue, tumor tissues were thawed and washed twice with RPMI 1640 prior to dissociation. Following incubation at 37^{o}C for 30 to 60 min, the sample was resuspended in RPMI 1640 and filtered through MACS® SmartStrainers (70 µM; Miltenyi Biotec). The resulting single cell suspension was centrifuged at 300 × g for 5 min. For fresh tissue processing, red blood cells were lysed with Lysing Buffer (Becton Dickinson) for 5 min and the resulting suspension was centrifuged at 300 × g for 5 min. Where viability was < 80%, viability enrichment was performed using the EasySep Dead Cell Removal (Annexin V) Kit (StemCell Technologies) as per manufacturer's protocol. Dissociated cells were resuspended in a final solution of PBS with 10% fetal calf serum (FCS) solution prior to loading on the 10X Chromium platform.

Data processing, cell cluster annotation and data integration

Raw bcl files were demultiplexed and mapped to the reference genome GRCh38 using the Cell Ranger Single Cell v2.0 software (10X Genomics). For individual samples, the EmptyDrops method from the DropletUtils package (v1.2.2)[2] was applied to filter the raw unique molecular identifiers (UMIs) count matrix for real barcodes from ambient background RNA cells. An additional cutoff was applied, filtering for cells with a gene and UMI count greater than 200 and 250, respectively. All cells with a mitochondrial UMI count percentage greater than 20% were removed. We used the Seurat v3.0.0 method[3] in *R* (v3.5.0) for data normalisation, dimensionality reduction and clustering using default parameters. Cell clusters were annotated using the Garnett method[4] (v0.1.4) using the default recommended parameters, with a classifier derived from an array of cell signatures for breast epithelial subsets from Lim *et al.* (2009)[5], and immune and stromal cell types from the XCell database[6], including T-cells, B-cells, plasmablasts, monocyte/macrophages, endothelial, fibroblast and perivascular cell signatures.

Data integration was performed using Seurat v3.0.0 using default parameters[3]. A total of 2000 features for anchoring (*FindIntegrationAnchors* step) and 30 dimensions for alignment (*IntegrateData* step) were used. For reclustering immune and mesenchymal lineages, a total of 5000 features were used for anchoring (*FindIntegrationAnchors* step), with a total of 30, 20, and 10 Principal Components were used for clustering T-cells, Myeloid cells and B-cells, respectively. The default resolution of 0.8 was used (*FindNeighbors* and *FindClusters* step). For clustering without batch correction steps, we merged all individual dataset together (*merge* function) performed clustering steps (*RunPCA*, *FindNeighbors* and *FindClusters* steps) using the “RNA” assay with a total of 100 principal components.

Identifying neoplastic from normal breast cancer epithelial cells

CNV signal for individual cells was estimated using the inferCNV method with a 100 gene sliding window. Genes with a mean count of less than 0.1 across all cells were filtered out prior to analysis, and signal was denoised using a dynamic threshold of 1.3 standard deviations from the mean. Immune and endothelial cells were used to define the reference cell inferred copy-number profiles. Epithelial cells were used for the observations. Epithelial cells were classified into normal (non-neoplastic), neoplastic or unassigned using a similar method to that previously described by Neftel *et al.*[7]. Briefly, inferred changes at each genomic loci were scaled (between -1 and +1) and the mean of the squares of these values were used to define a genomic instability score for each cell. In each individual tumor, the top 5% of cells with the highest genomic instability scores were used to create an average CNV profile. Each cell was then correlated to this profile. Cells were plotted with respect to both their genomic instability and correlation scores. Partitioning around medoids (PAM) clustering was performed using the 'pamk' function in the *R* package 'cluster' (v2.0.7-1) to choose the optimum value for k (between 2-4) using silhouette scores, and the 'pam' function to apply the clustering. Thresholds defining normal and neoplastic cells were set at 2 cluster standard deviations to the left and 1.5 standard deviations below the first cancer cluster means. For tumors where PAM could not define more than 1 cluster, the thresholds were set at 1 standard deviation to the left and 1.25 standard deviations below the cluster means. This method was used to identify 27,506 neoplastic and 6084 normal cells in all tumors, the remaining 3208 cells were classed as unassigned (Extended Data Fig. 1g and Supplementary Fig. 1). Only tumours with at least 200 epithelial cells were used for this neoplastic cell classification step.

Calling PAM50 on pseudo-bulks and matching bulk RNA-Seq

We constructed “pseudo-bulk” expression profiles for each tumor, where all the reads from all cells of a given tumor were added together, and then mapped as one sample. The resulting pseudo-bulk matrix thus constructed was named “Allcells-Pseudobulk” and was subsequently processed similarly to any bulk RNA-Seq sample (i.e. upper quartile normalized-log transformed) for calling molecular subtypes using the PAM50 method[8]. An important consideration made before PAM50 subtyping is to adjust a new sample set relative to the PAM50 training set according to their ER and HER2 status as detailed by Zhao *et al.*[9]. Thus, after ER/HER2 group-based adjustments, and then applying the PAM50 centroid predictor to the pseudo-bulk data, the methodology identified 7 of 20 Basal-like (CID3963, CID4465, CID4495, CID44971, CID4513, CID4515, CID4523), 4 of

20 HER2E (CID3921, CID4066, CID44991, CID45171), 5 of 20 LumA (CID3941, CID4067, CID4290A, CID4463, CID4530N), 3 of 20 LumB (CID3948, CID4461, CID4535) and 1 of 20 as Normal-like (CID4471).

We performed whole-transcriptome RNA-Seq using Ribosomal Depletion on 18 matching tumor samples from our single-cell dataset. RNA was extracted from diagnostic FFPE blocks using the High Pure RNA Paraffin Kit (Roche #03 270 289 001). The Sequence alignment was done using Salmon[10]. We then called PAM50 on each bulk tumor using Zhao *et al.*[9] normalization and then the PAM50 centroid predictor (Supplementary Table 3).

Calling intrinsic subtype on scRNA-Seq using scSubtype

To design and validate a new subtyping tool specific for scRNA-Seq data, we first divided our tumor samples into training and testing sets. The training dataset was defined by identifying tumors with unambiguous molecular subtypes. Here, we identified robust training set samples using two subtyping approaches: (i) PAM50 subtyping of the *Allcells-Pseudobulk* datasets (described above); and (ii) hierarchical clustering of the *Allcells-Pseudobulk* data with the 1,100 tumors in the TCGA breast cancer RNA-Seq dataset using ~2000 genes from an intrinsic breast cancer genelist[8]. We first identified tumors that shared the same “concordant” subtype from both *Allcells-Pseudobulk* PAM50 calls and TCGA hierarchical clustering based subtype classifications (Supplementary Table 3). Next, since our methodology aimed to subtype cancer cells, we removed any tumors with <150 cancer cells. Finally, we did not include cells from the two metaplastic samples (CID4513 and CID4523) in the training data because this is a histological subtype not used in the original PAM50 training set. Using this approach, we identified 10 tumor samples in the training dataset: HER2E (CID3921, CID44991, CID45171), Basal-like (CID4495, CID44971, CID4515), LumA (CID4290, CID4530) and LumB (CID3948, CID4535). Only tumor cells with greater than 500 UMIs were used for training and test datasets in scSubtype (total of 24,889 cells).

Within each training set subtype, we utilized the cancer cells from each tumor sample and performed pairwise single cell integrations and differential gene expression calculations. The integration was carried out in a "within group" pairwise fashion using the *FindIntegrationAnchors* and *IntegrateData functions* in the Seurat v3.0.0 package[3]. Briefly, the first step identifies anchors between pairs of cells from each dataset using mutual nearest neighbors. The second step integrates the datasets together based on a distance based weights matrix constructed from the anchor pairs. Differentially expressed genes were calculated between each pair using a Wilcoxon Rank Sum test by the *FindAllMarkers function within Seurat v3*. As the number of cancer cells per tumor sample were highly variable, this strategy prevented a bias of identifying genes for a training group from a sample with the highest number of cells. The following pairs were analyzed: HER2E (CID3921-CID44991, CID44991-CID45171, CID45171-CID3921), Basal-like (CID4495-CID44971, CID44971-CID4515, CID4515-CID4495), LumA (CID4290-CID4530) and LumB (CID3948-CID4535). In this way we identified unique upregulated genes per sample, but also genes broadly highlighting cells within each respective training group or subtype. We removed any duplicate genes occurring between the 4 training groups, which yielded 4 sets of genes composed of 89 genes defining Basal_SC, 102 genes defining HER2E_SC, 46 genes defining LumA_SC and 65 genes defining LumB_SC, which we define as "scSubtype" gene signatures (Supplementary Table 4).

To assign a subtype call to a cell we calculated the average (i.e. mean) read counts for each of the 4 signatures for each cell. The SC subtype with the highest signature score was then assigned to each cell. We utilized this method to subtype all 24,489 neoplastic cells, from both our training samples (n=10) and the remaining test (n=10) set samples.

Gene module analysis of neoplastic intra-tumor heterogeneity

For each individual tumor, with more than 50 neoplastic cells, the neoplastic cells were clustered using Seurat v3.0.0[3] at five resolutions (0.4, 0.8, 1.2, 1.6, 2.0). MAST[11] (v1.12.0) was then used to identify the top-200 differentially regulated genes in each cluster. Only gene-signatures containing greater than 5 genes and originating from clusters of more than 5 cells were kept. In addition, redundancy was reduced by comparing all pairs of signatures within each sample and removing the pair with fewest genes from those pairs with a Jaccard index greater than 0.75. Across all tumors, a total of 574 gene-signatures of intra-tumor heterogeneity were identified.

Consensus clustering (using spherical k-means, skmeans, implemented in the cola R package (v1.2.0): https://www.bioconductor.org/packages/release/bioc/html/cola.html) of the Jaccard similarities between these gene-signatures was used to identify 7 robust groups, or gene-modules. For each of these, a gene module was defined by taking the 200 genes that had the highest frequency of occurrence across clusters and individual tumors. These are defined as gene-modules GM1 to GM7. A gene-module signature was calculated for each cell using AUCell[12] and each neoplastic cell was assigned to a module, using the maximum of the scaled AUCell gene-module signature scores. This resulted in 4,368, 3,288, 2,951, 4,326, 3,931, 2,500, 3,125 cells assigned to GM1 to GM7, respectively. These are defined as gene-module based neoplastic cell states. Selected breast cancer related gene-signatures[13-16] were used for pathway enrichment in Extended Data Figure 2b.

Differential gene expression, module scoring and gene ontology enrichment

Differential gene expression was performed using the MAST method[11] (v1.8.2) in Seurat (*FindAllMarkers* step) using default cutoff parameters. All DEGs from each cluster (Supplementary Table 9 and 10) were used as input into the ClusterProfiler package[17] (v3.14.0) for gene ontology functional enrichment. All ontologies within the enrichGO databases were used with the human org.Hs.eg.db database. Results were clustered, scaled and visualised using the pheatmap package (v1.0.12) in *R*. Cytotoxic, TAM and Dysfunctional T-cell gene expression signatures were assigned using the *AddModuleScore* function in Seurat v3.0.0[3]. The list of genes used for dysfunctional T-cells were adopted from Li *et al.*[18]. The TAM gene list was adopted from Cassetta et al.[19]. The cytotoxic gene list consists of 12 genes which translate to effector cytotoxic proteins (*GZMA*, *GZMB*, *GZMH*, *GZMK*, *GZMM*, *GNLY*, *PRF1* and *FASLG*) and well described cytotoxic T-cell activation markers (*IFNG*, *TNF*, *IL2R* and *IL2*).

CITE-Seq antibody staining

Samples were stained with 10X Chromium 3' mRNA capture compatible TotalSeq-A antibodies (Biolegend, USA). Staining was performed as previously described by Stoeckius et. al[20] with a few modifications listed below. A total of four cases from our scRNA-Seq cohort were analyzed, including one luminal (CID4040), one HER2 (CID383) and two TNBC (CID4515 and CID3956). A panel of 157 barcoded antibodies (Supplementary Table 11) were used, which recognised a range of cell surface lineage and activation markers, in addition to a large collection of co-stimulatory and co-inhibitory receptors and ligands[21]. Briefly, a maximum of 1 million cells per sample was resuspended in 120 ul of cell staining buffer (Biolegend, USA) with 5 ul of Fc receptor Block (TrueStain FcX, Bioelegend, USA) for 15 min. This was followed by a 30 min staining of the antibodies at 4°C. A concentration of 1 ug / 100 ul was used for all antibody markers used in this study. The cells were then washed 3 times with PBS containing 10% FCS media followed by centrifugation (300 x g for 5min at 4°C) and expungement of supernatant. The sample was then resuspended in PBS with 10% FCS for 10X Chromium capture.

Visium spatial transcriptomics data processing

Reads were demultiplexed and mapped to the reference genome GRCh38 using the Space Ranger Software v1.0.0 (10X Genomics). Count matrices were loaded into the Seurat v3.2.0 (https://github.com/satijalab/seurat/tree/spatial) and STutility v0.1.0 (https://github.com/jbergenstrahle/STUtility) *R* packages for all subsequent data filtering, normalisation, filtering, dimensional reduction and visualization. All spatial spots determined to be over tissue regions by Space Ranger were kept for subsequent analysis. Poor quality tissue locations were then filtered out based on a cutoff of 500 unique genes. Genes detected in more than 10 locations were also kept for analysis. Data normalisation was performed on independent tissue sections using the variance stabilizing transformation method implemented in the *SCTransform* function in Seurat. We applied non-negative matrix factorization (NMF) to the normalised expression matrix using the STutility package (nfactors = 20). NMF reduction was then used for clustering using Seurat with all 20 factors as input (*RunUMAP*, *FindNeighbors* and *FindClusters* functions).

Spatial deconvolution using Stereoscope

The Visium platform has not yet reached single cell resolution, but rather tends to host multiple cells at each capture location, potentially of different cell type identities. Thus, we performed deconvolution of spatial tissue locations using the method presented as Stereoscope[22] (v0.2.0), a probabilistic model for estimating cell type proportions using annotated scRNA-Seq data as input. Stereoscope models the observed expression vectors (associated to each spatial location) as a mixture of transcripts originating from one or more cells of equally many or less types. By assuming that both single cell and spatial expression data is negative binomial distributed, parameters characterizing the different cell types can be learnt from the former and transferred to the latter. Inferring proportion estimates for the spatial data, somewhat simplified, is thus equivalent to finding the combination of cell type parameters (estimated from the single cell data) that best explain the observed expression values. Implementation-wise, stochastic optimization using gradient descent to find the MAP estimates of the parameters is used.

Upon deconvolving the spatial data, we matched spatial and single cell data with respect to cancer subtype. Meaning that for any spatial sample of a given subtype, only cells originating from tissue of the same subtype were provided as input into Stereoscope and used when inferring type parameters. We deconvolved cell types across three tiers of classification including the major, minor and subset lineages, still maintaining the separation between different cancer subtypes. In all of our analyses, we used 50000 epochs during both steps (parameter inference and proportion estimation) of the analysis. Furthermore, the scRNA-Seq in each analysis was subsampled excluding those types with less than 25 members and using an upper bound of 500 cells per type. Finally, only the top 5000 highest expressed genes (in the single cell data) were used throughout the procedure. Both these approaches (of subsampling and top gene selection) were in line with the official documentation of Stereoscope. The batch size was set to 2,046 in both steps of the analysis. For the remaining set of parameters we used the default values. The results obtained when deconvolving the spatial data are proportion estimates of each type at every spatial location, represented using a [n_spots] x [n_types] matrix for every sample. The rows of this matrix always sum to one, due to the values representing proportions.

Tumor ecotype analysis using deconvolution of bulk sequencing patient cohorts

CIBERSORTx[23] and DWLS[24] were used to deconvolute predicted cell-fractions from a number of bulk transcript profiling datasets. To prevent confounding of cycling cell-types we first assigned all neoplastic epithelial cells with a proliferation score > 0 as cycling and then combined these with “cycling” cell states from all other cell-types to generate a single “Cycling” cell-state. To generate cell-type signature matrices for each of the tiers of cell-type annotation described in this study, we randomly subsampled 15% of cells from each level of annotation type.

CIBERSORTx

We then ran CIBERSORTx “cibersortx/fractions” to generate cell-type signature matrices using the following parameters: *--single_cell TRUE --G.min 300 --G.max 500 --q.value 0.01 --filter FALSE --k.max 999 --replicates 5 --sampling 0.5 --fraction 0.75*.

For cell-type deconvolution of bulk tumours we ran CIBERSORTx “cibersortx/fractions” to calculate the relative cell-type abundances in each tumour. S-mode batch correction was used for the METABRIC tumours.

DWLS

For deconvolution analysis using DWLS we used the functions in the “Deconvolution_functions.R” script obtained from https://github.com/dtsoucas/DWLS. Cell-type signature matrices were generated using the buildSignatureMatrixMAST() function and then filtered to only contain genes that are present in both the bulk and single-cell derived signature matrices, using the trimData() function. Cell-type abundances were then calculated using the solveDampenedWLS() function.

Bulk expression datasets

Pseudo-bulk expression matrices were generated from the scRNA-Seq datasets in this study by summing the UMIs for each gene across all cells for each tumor. Normalised METABRIC expression matrices, clinical information and PAM50 subtype classifications were obtained from https://www.cbioportal.org/study/summary?id=brca_metabric.

Tumour Ecotypes

Tumor ecotypes in the METABRIC cohort were identified using spherical k-means (skmeans) based consensus clustering (as implemented in the cola R package v1.2.0: https://www.bioconductor.org/packages/release/bioc/html/cola.html) of the predicted cell-fraction from either CIBERSORTx or DWLS, in each bulk METABRIC patient tumor. When comparing ecotypes between methods (i.e., consensus clustering results from using cell-abundances of all cell-types or just the 32 significantly significantly correlated cell-types from CIBERSORTx deconvolution and consensus clustering results from CIBERSORTx or DWLS cell-abundances) the number of tumour ecotypes was fixed as 9 and the tumour overlaps between all ecotype pairs was calculated (Supplementary Table 7 and 8). Common ecotypes were then identified by identifying the ecotype pairs with the largest average METABRIC tumour overlap.

Survival Analysis

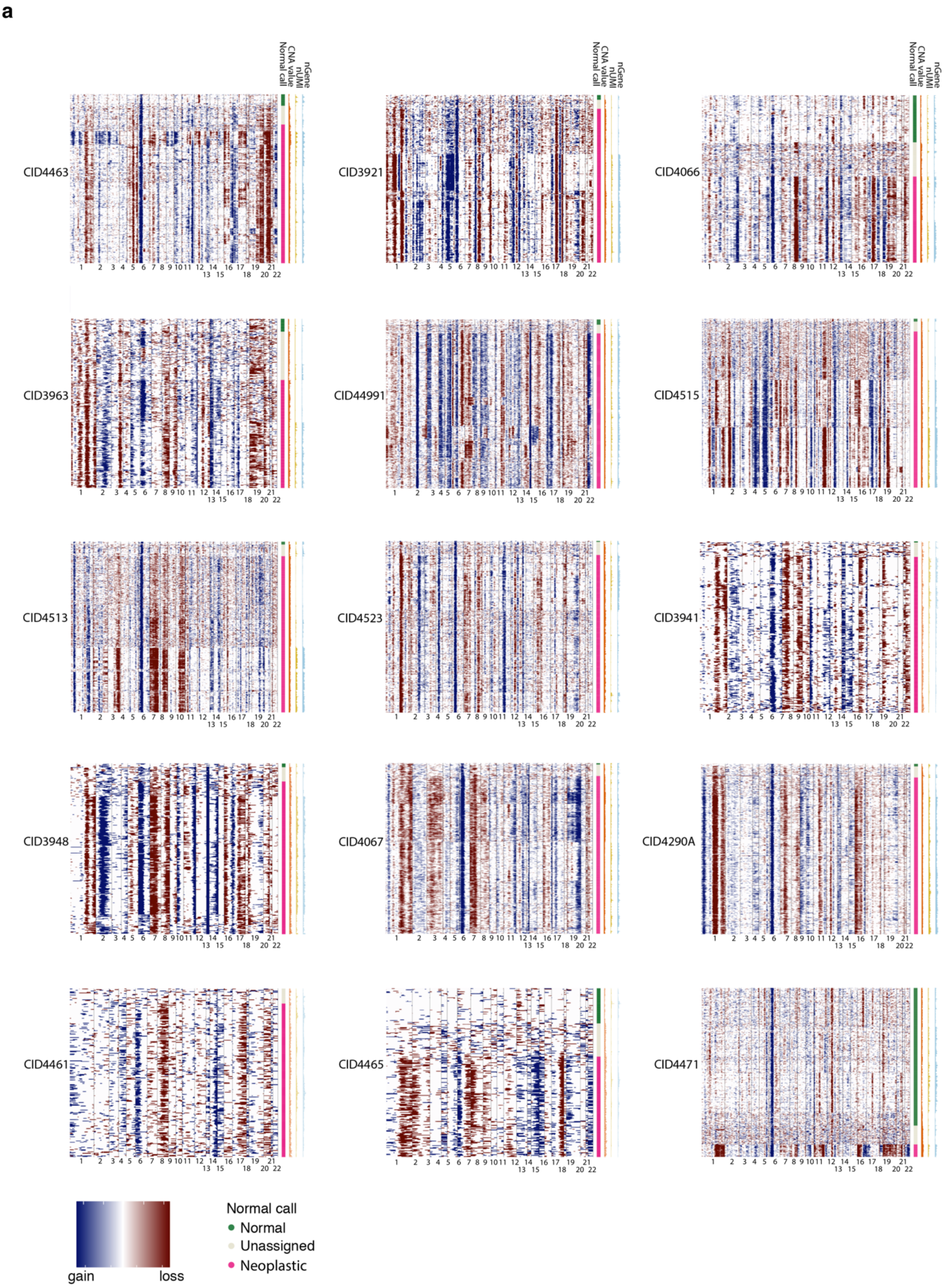
Differences in survival between ecotypes were assessed using Kaplan-Meier analysis and log-rank test statistics, using the *survival* (v2.44-1.1) and *survminer* (v0.4.7) R packages.

References

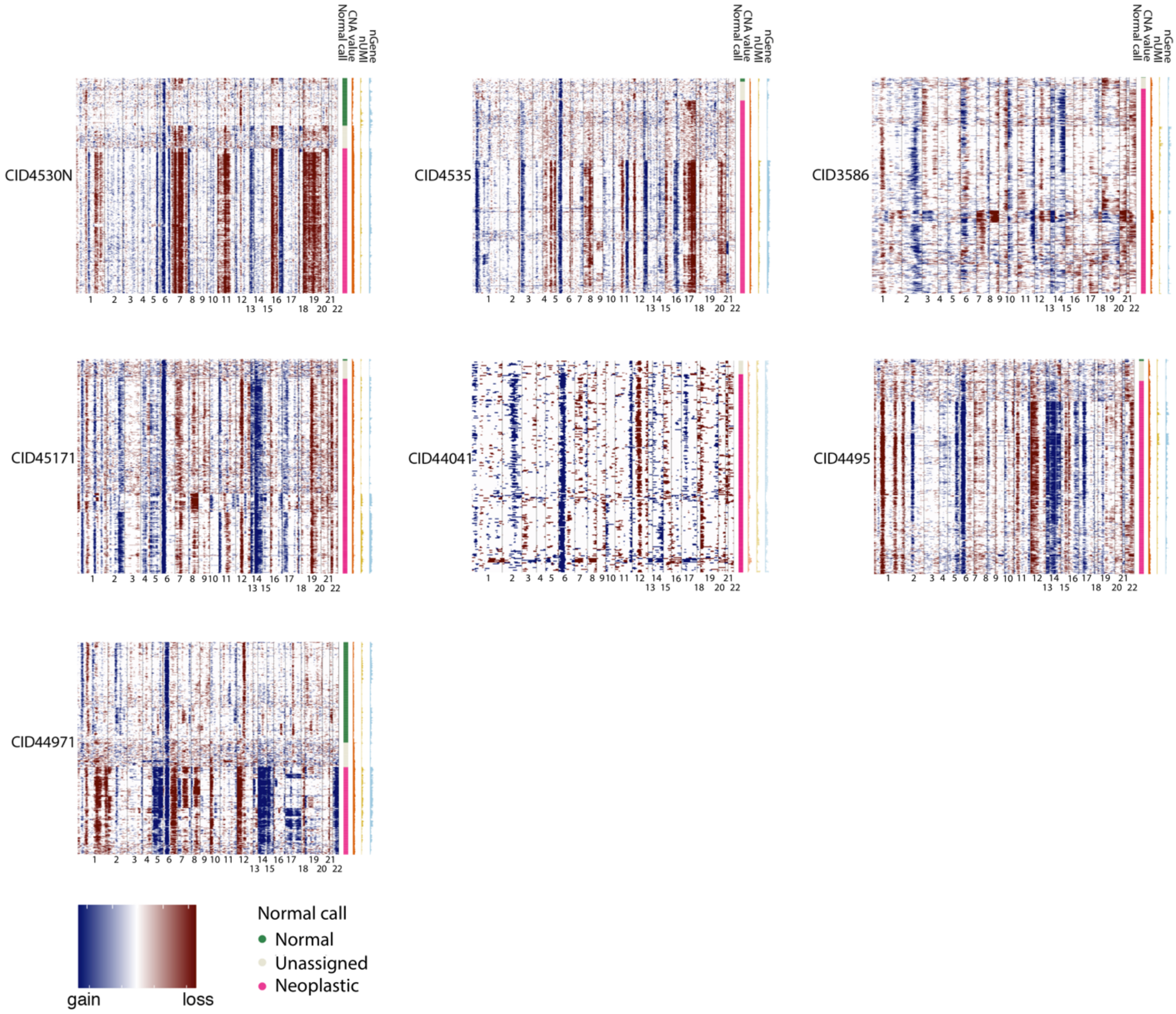
1. Wu, S.Z. *et al.* Cryopreservation of human cancers conserves tumour heterogeneity for single-cell multi-omics analysis. *Genome Medicine* **13**, 81 (2021).
2. Lun, A.T.L. *et al.* EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol* **20**, 63 (2019).
3. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e21 (2019).
4. Pliner, H.A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *Nature Methods* **16**, 983-986 (2019).
5. Lim, E. *et al.* Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat Med* **15**, 907-13 (2009).
6. Aran, D., Hu, Z. & Butte, A.J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol* **18**, 220 (2017).
7. Neftel, C. *et al.* An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. *Cell* **178**, 835-849 e21 (2019).
8. Parker, J.S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* **27**, 1160-7 (2009).
9. Zhao, X., Rodland, E.A., Tibshirani, R. & Plevritis, S. Molecular subtyping for clinically defined breast cancer subgroups. *Breast Cancer Res* **17**, 29 (2015).
10. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14**, 417-419 (2017).
11. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* **16**, 278 (2015).
12. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* **14**, 1083-1086 (2017).
13. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).
14. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417-425 (2015).
15. Gatza, M.L., Silva, G.O., Parker, J.S., Fan, C. & Perou, C.M. An integrated genomics approach identifies drivers of proliferation in luminal-subtype human breast cancer. *Nat Genet* **46**, 1051-9 (2014).
16. Prat, A. *et al.* Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res* **12**, R68 (2010).
17. Yu, G., Wang, L.G., Han, Y. & He, Q.Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284-7 (2012).
18. Li, H. *et al.* Dysfunctional CD8 T Cells Form a Proliferative, Dynamically Regulated Compartment within Human Melanoma. *Cell* **176**, 775-789 e18 (2019).
19. Cassetta, L. *et al.* Human Tumor-Associated Macrophage and Monocyte Transcriptional Landscapes Reveal Cancer-Specific Reprogramming, Biomarkers, and Therapeutic Targets. *Cancer Cell* **35**, 588-602 e10 (2019).
20. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* **14**, 865-868 (2017).
21. Marin-Acevedo, J.A. *et al.* Next generation of immune checkpoint therapy in cancer: new developments and challenges. *J Hematol Oncol* **11**, 39 (2018).
22. Andersson, A. *et al.* Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Communications Biology* **3**, 565 (2020).
23. Newman, A.M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology* **37**, 773-782 (2019).

24. Tsoucas, D. *et al.* Accurate estimation of cell-type composition from gene expression data. *Nat Commun* **10**, 2975 (2019).

Supplementary Figure 1



Supplementary Figure 1 (cont.)



Supplementary Figure 1. Identification of malignant epithelial cells using inferCNV

a, InferCNV heatmaps showing all epithelial cells and their associated inferCNV based classification for all tumors. For each cell, the normal cell call, copy number alteration (CNA) values, number of unique molecular identifiers (UMIs) and genes per cell are plotted on the right. Normal cell calls were classified as either Normal (green), Unassigned (grey) or Neoplastic (pink). These classification are derived from the a genomic instability score, which is estimated by the inferred changes at each genomic loci, as determined by inferCNV. High UMI and gene metrics in normal cells importantly show that they are not a product of coverage or low sequencing depth.