Revealing unidentified heterogeneity in different epithelial cancers using heterocellular subtype classification

Pawan Poudel[1,&], Gift Nyamundanda[1], Chanthirika Ragulan[1], Rita T. Lawlor[2,3], Kakoli Das[4], Patrick Tan[4,5,6], Aldo Scarpa[2,3], Anguraj Sadanandam[1,&,*]

[1]Division of Molecular Pathology, Institute of Cancer Research (ICR), London, United Kingdom
[2]ARC-Net Centre for Applied Research on Cancer, University and Hospital Trust of Verona, Verona, Italy
[3]Department of Pathology and Diagnostics, University and Hospital Trust of Verona, Verona, Italy
[4]Cancer and Stem Cell Biology Program, Duke-NUS Medical School, Singapore
[5]Genome Institute of Singapore, Biopolis, Singapore
[6]Cancer Science Institute of Singapore, National University of Singapore, Singapore

[&] Equal contribution to the research work
* Corresponding author

Supplementary Information.

Systematic comparison of heterocellular subtypes to known intrinsic subtypes in multiple cancer types

Ovarian cancer

We compared six intrinsic OV subtypes[1] with CRC heterocellular subtypes (**Supplementary Figure 3E, Supplementary Table 1L-N**). There were significant associations (FDR<0.05) between intrinsic ovarian cancer subtypes and four CRC heterocellular subtypes except for the enterocyte subtype, reflecting low enrichment due to sample size. The stroma-rich C1 subtype was associated with its counterpart stem-like heterocellular subtype; the immune-rich C2 subtype with the inflammatory subtype; the low malignant potential (LMP) secretory cell-type enriched C3 and high-grade C4 with the goblet-like subtype; and Wnt signature-high C5 and C6 subtypes with the TA subtype (high Wnt signalling). As expected, both the C1 subtype and stem-like subtypes have poor prognosis, the C2 and inflammatory subtypes have intermediate prognosis, and the C5 and C6 subtypes and the TA subtype have good prognosis.

HNSC

The four intrinsic subtypes described for HNSC[2] were significantly associated with heterocellular subtypes (**Supplementary Figure 3F and Supplementary Table 1L-N**). The mesenchymal HNSC subtype was associated with the stem-like heterocellular subtype, while the atypical (human papilloma virus-positive) HNSC subtype was associated with goblet-like and enterocyte heterocellular subtypes, reflecting their well-differentiated status. The classical HNSC subtype enriched for the TA CRC heterocellular subtype and, of note, xenobiotic metabolic genes similar to the exocrine-PC subtype[3] and known to be associated with the TA CRC heterocellular subtype. Finally, the basal HNSC subtype was significantly enriched for the inflammatory heterocellular subtype.

Other cancer types

Among the other cancer types, the “mitotic” UCEC subtype[4] was significantly enriched for the inflammatory, and less significantly with the stem-like, subtypes (**Supplementary Figure GD and Supplementary Table 1L-N**). Similarly, the “hormonal” UCEC subtype with increased progesterone receptor was significantly enriched for the TA followed by the goblet-like subtypes, reflecting “well-differentiated” genes. Finally, the immunoreactive UCEC subtype was significantly enriched for stem-like, and less significantly with the goblet-like, subtypes, which might reflect different sets of immune cells enriched in these subtypes compared to inflammatory CRC heterocellular subtypes.

In BLCA, the “cluster I” subtype[5] with papillary histology and *FGFR3* aberrations was significantly enriched for the *FGFR3*-high TA subtype and less significantly with differentiated goblet-like and enterocyte subtypes (**Supplementary Figure 3H and Supplementary Table 1L-N**). The “cluster III” subtype with increased basal genes *KRT5* and *KRT14* was enriched for the inflammatory subtype, and the *ERBB2/HER2* aberration containing “cluster IV” was enriched, albeit non-significantly, for the inflammatory subtype. The urothelial differentiated “cluster II” subtype was significantly enriched for the stem-like subtype, representing potential heterogeneity similar to that of the PP/classical PC subtype.

In KIRC, the ”m4” subtype[6] with increased base-excision repair gene expression was enriched for the inflammatory CRC heterocellular subtype, possibly representing a hypermutated subtype (**Supplementary Figure 3I and Supplementary Table 1L-N**). Similarly, the “m2” subtype was enriched for the inflammatory subtype, the “m1” subtype was enriched for the TA subtype representing a potential chromatin-remodelling process. Finally, the “m3” subtype with CDKN2A deletion was enriched for the stem-like subtype, similar to some of the QM-PCs enriched for a stem-like subtype associated with cell lines derived from the *CDKN2A/ARF*-deleted PC mouse model.

Among the lung cancer histological subtypes[7], the well-differentiated adenocarcinoma subtype was significantly associated with the goblet-like subtype whereas the squamous subtype was primarily and significantly associated with a poorly differentiated and poor prognostic stem-like subtype (**Supplementary Figure 3J and Supplementary Table 1L-N**). However, there was a non-significant association between the squamous subtype and the TA CRC heterocellular, possibly representing a differentiation pathway in certain squamous lung tumours. To understand this further, we compared the TCGA LUAD intrinsic transcriptomic subtypes with CRC heterocellular subtypes (**Supplementary Figure 3K**), which revealed that a proximal proliferative (PRP) subtype was enriched for goblet-like and TA subtypes. These PRP subtypes were also enriched for *KRAS* mutations, similar to the overrepresentation of *KRAS* mutations in CMS3/goblet-like subtypes. Interestingly, we also observed that good prognosis terminal respiratory unit

(TRU) was associated with good prognosis enterocyte and goblet-like CRC heterocellular subtypes.

Supplementary Methods

Gene expression data processing. The raw data (CEL files) containing patient tumour gene expression profiles generated using Affymetrix GeneChip® Human Genome U133 Plus 2.0 arrays were downloaded from Gene Expression Omnibus[8] (GEO) (**Supplementary Table 6A)**. The CEL files were pre-processed and normalized using robust multi-array normalization (RMA) from bioconductor[9] packages - *affy*[10]. The following criteria were applied to select only those good quality microarrays with reduced repetitive samples. The Normalized Unscaled Standard Error (NUSE[11]; from *affyPLM*[12] package) median score (1±0.05) was used to select only high quality arrays. Repeated samples were removed based on the information from GEO, original publications or those two samples with Pearson correlation coefficient equal to or greater than 0.99. The cell lines gene expression profile data from lung, colorectal (large intestine) and pancreatic cancers were obtained from Cancer Cell Line Encyclopedia (CCLE)[13] as RMA processed probe level (Affymetrix GeneChip® Human Genome U133 Plus 2.0 array) data (**Supplementary Table 6D**). Next, the genes corresponding to each probe for Affymetrix GeneChip® Human Genome U133 Plus 2.0 array were annotated with Human Genome Organisation (HUGO) Gene Nomenclature Committee (HGNC)[14] identifiers using the R package - *hug133plus2db*[15]. Finally, a highly variable probe was selected for those genes associated with multiple Affymetrix GeneChip® probes, as described previously[16-18]. We used the following primary tumour datasets for different cancer types - GSE42568[19] (BRCA; n=104), GSE14333[20] (CRC; n=288) GSE37745[7] (LUAD/LUSC; n=168), GSE9891[1] (OV; n=177), GSE15471[21] (PC; n=36), GSE35809[22] (GC-1; n=68), GSE34942[22] (GC-2; n=56) and GSE15459[22] (GC-4; n=182).

The processed and normalized TCGA Pan-Cancer gene expression profile data (log_2 transformed upper quartile normalized RSEM data) for 12 cancer types from Hoadley. *et al*[23] were obtained from Sage Bionetworks using the synapse id (syn1715755; **Supplementary Table 6B**). Those genes with missing values (a value of zero from log transformed RSEM data) in greater than 30% of the samples were removed, as described[23]. The repetitive samples from certain Pan-Cancer tumour types were removed based on the TCGA sample identifiers information. Again, those processed and normalized ICGC pancreatic cancer[24] and TCGA gastric cancer (STAD)[25] data were obtained from the original publications. Ensembl transcript identifiers from ICGC data were mapped to HGNC gene identifiers from the Ensembl[26] using web-based BioMart[27] (**Supplementary Table 6C**). The TCGA STAD data was log_2 transformed (transformed was added to the expression values before the logarithmic transformation) before further analysis.

Single sample classification: For the single sample classification, the "similarity-to-centroids" based approach was used, which is similar to the CRCAssigner and CMSclassifier[28]. The published[17] CRCassigner PAM centroids, for each subtype, were used to correlate the gene expression profiles of each sample to the centroids of each subtype. Using the Pearson correlation as a similarity measure, the samples were then assigned to one of the five CRCassigner subtypes based on highest correlation coefficient**.** Due to the existence of tumour heterogeneity only the high-confidence classification results were selected for downstream analysis. This approach of selecting high-confidence samples had been used previously for the CMS classification[28]. For the high-confidence classification, each sample must meet two criteria (i) the highest Pearson correlation coefficient should be greater than 0.15 and (ii) the difference between the highest and the second highest correlation coefficient must be greater than 0.06. Any sample that did not follow these criteria was classified as "mixed samples".

Before predicting CRCassigner subtypes using single sample classification approach, the following steps were performed: (i) the probes/identifiers without gene annotation were filtered out, (ii) genes having zero standard deviation across samples were removed, (iii) the genes were median centred across samples for each datasets separately, and (iv) only the genes in the CRCassinger classifier were selected. Due to gene annotations and platform differences, the selection of CRCassigner genes resulted in 750 (for microarrays), 643 (for Pan-Cancer), 721 (ICGC pancreatic cancer) and 757 (for TCGA gastric cancer) genes in different datasets.

Distance to the CRCassigner PAM centroids: A correlation analysis was performed to understand which heterocellular subtypes (from each cancer types) were closely associated with the CRCassigner subtypes. First, common CRCassigner genes were selected from the gene expression datasets from each cancer type. Next, the "mixed samples" (described above) were removed, and the genes were median centred across samples for each cancer/dataset. Subsequently, the median gene expression values (for all the CRCassigner genes) for five subtypes in each organ/datasets were obtained. Later, these median values were correlated (Pearson) with CRCassigner PAM centroids. The correlation coefficient score was used as an indicator to assess the similarity of heterocellular subtypes (from multiple cancers) to colon.

Tissue-specific gene analysis: The PAM centroid scores (for 786 CRCassinger genes and five subtypes) were used to assign a gene to one of the five subtypes. A gene was assigned to a subtype if it had the highest PAM score whilst the remaining subtypes have 0 or negative PAM scores. This process yielded 564 subtypes specific genes and the remaining 222 genes that were shared between subtypes. Next, the tissue specific gene expression[29] (TiGER; considering the genes in the expressed sequence tags) database was used to identify the colon

specific genes in each subtypes. Subsequently, the colon specific genes in CRCassinger classifier were estimated (for each subtype).

Visualisation of gene expression data: For heatmap visualisation, the following steps were performed: (i) CRCassigner genes (performed separately for each cancer type) were selected from gene expression data, (ii) "mixed samples" were removed, (iii) genes were median centred across samples, and (iv) samples were ordered by subtypes. Next, the genes were clustered (hierarchical clustering) using the java-based application cluster 3.0[30] with default settings. Finally, the clustered results were visualised using another java-based application called GENEE (https://software.broadinstitute.org/GENE-E/index.html), from the Broad Institute. The gene expression values were scaled to +/-3 for visualisation in GENEE.

Reconciliation of subtypes. The intrinsic subtype classifications for each sample were downloaded from the respective published studies (**Supplementary Table 6E**). Next, the association between the intrinsic and the heterocellular subtypes were performed via sample enrichment using the hypergeometric test[31]. The p-value from the hypergeometric was corrected using false discovery rate (FDR) approach; the FDR was used to access the significance of association. Due to the unavailability of the intrinsic classification for the breast cancer data (GSE42568), the samples were classified to the intrinsic breast cancer subtypes using the R package - genefu[32].

Enrichment of stromal genes in PC: Following steps were performed on the pre-processed ICGC data (considering all the genes): (i) mixed samples were removed; (ii) gene-wise median centring was performed; (iii) highly variable genes were selected using the SD cut-off of 1. Later, these selected genes and samples were used to perform the GSEA between the stem-like versus all other subtypes using the Reactome[33] genesets. Additionally, a box plot was created to demonstrate the expression of some of the commonly sought after genes in immunotherapy[34] using the published ICGC data.

Heterogeneity of luminal A subtypes: The heterogeneity in stem-like luminal A subtypes and other heterocellular subtypes in luminal A was explored using gene set enrichment analysis (GSEA). The GSEA was performed between the two classes - stem-like (luminal A) and the other 4 heterocellular (luminal A) subtypes. Before GSEA, the following steps were performed: (i) all genes (n=14712) present in Pan-Cancer BRCA datasets were considered, (ii) only the luminal A samples from the Pan-Cancer BRCA data were selected, (iii) highly variable genes using the standard cut-off (SD) of 1.5 were selected, which yielded 876 genes and 131 samples.

Subsequently, the genes were median centred and the hierarchical clustering of genes was performed. The data was visualised via heatmaps using GENEE software. Additionally, published[35] classification

for the luminal A copy number subtypes (for samples in Pan-Cancer BRCA datasets) were associated with the heterocellular classification.

Comparing the microsatellite status across cancers: Microsatellite status for three cancers – CRC, STAD and UCEC - from the three published[4,25,36] studies were used to classify samples as microsatellite instable (MSI) and stable (MSS) (**Supplementary Table 6E**). Subsequently, the following steps were performed: (i) all the genes (n=14712) in each dataset were considered; (ii) microsatellite instability high (MSI-H) and low (MSI-L) were combined as MSI; (iii) mixed samples were removed; (iv) MSI samples in inflammatory and goblet-like were selected; (v) genes were median centred; and (vi) highly variable genes using the SD cut-off of 1 were selected for GSEA. Next, GSEA[37] was performed between the MSI inflammatory and the MSI goblet-like using published immune markers[38] as a geneset. For the additional validation, the Level 4 RPPA data from The Cancer Proteome Atlas (TCPA)[39] were used to compare the PDL-1 expression between MSI inflammatory and MSI goblet-like.

Prediction of *KRAS* dependency status and pathway analysis: The *KRAS* (**Supplementary Table 6F**) mutant samples (removing the "mixed" samples) from CRC, PC, and LUAD (analysis was performed separately for each datasets) were selected. Next, the *KRAS* dependency status for the *KRAS* mutant samples was predicted using the published[40] signature and the Nearest Template Prediction[41] (NTP) algorithm using default settings. Only those predictions having FDR less than 0.2 (as described in our previous publication[17]) were selected from NTP classification. Sample enrichment analysis between the heterocellular subtypes and the predicted *KRAS* dependency status (from NTP) were performed using the hypergeometric test. Based on the hypergeometric results, *KRAS*-dependent goblet-like (KD-GL) and the *KRAS*-independent (KID-SL) samples were selected to identify the commonly enriched pathways across the 3 cancer types. Before GSEA, the following steps were performed: (i) all the genes (n=14712) in each datasets were considered; (ii) only the samples in KD-GL and KID-SL were selected from each datasets; (iii) the data were median centred; (iv) highly variable genes using the SD cut-off of 1 were selected; and (v) GSEA was performed using the Hallmarks[42] and oncogenic[37] genes sets.

Supplementary Figure Legends and Tables.

Supplementary Figure 1: A. Proportions of mixed subtypes in CRC-1 (GSE14333[20]; n=288), GC-1 (GSE35809[22]; n=68), OV-1 (GSE9891[1]; n=177), PC-1 (GSE15471[21]; n=36) and BRCA-1 (GSE42568[19], n=104) and LUAD/LUSC (GSE37745[7]; n=168). **B**. Proportions of mixed subtypes in additional gastrointestinal cancers - gastric [(GC-2/GSE34942[22]; n=56) and GC-4/GSE15459[22]; n=182)], GC-3[25] (gastric data from TCGA; n=239) and pancreatic cancer[24] (including PC and other histological subtypes from ICGC; PC-2; n=96). **C.** Proportions of mixed subtypes in

TCGA Pan-cancer datasets[23] (syn1715755) – CRC-2 (n= 262); OV-2 (n=259); BLCA (n=122), LUAD (n=351); KIRC (n=480); HNSC (n=303), BRCA-2 (n=835), LUSC (n=257), UCEC (n=370). **D**. Heatmap showing correlation coefficient comparing CRCassigner PAM centroids and median values of the corresponding genes across samples within each subtype and cancer type (considering all the cancers used in this study). **E.** Venn diagram showing the number of colon-specific genes (from TiGER[29] database) present in CRCassigner gene signature. **F.** Proportion of CRC heterocellular subtypes in different cancer types from all the data sources used in this study. **G-I.** Heatmaps showing the variability in CRCassigner genes in three different cancer types – (G.) gastric cancer[22] (GC-1/GSE35809), (H.) PC[21] (PC-1/GSE15471) and (I.) breast cancer[19] (BRCA-1/GSE42568).

Supplementary Figure 2: A. Bar plot showing the proportions of Bailey's[24] subtypes in PDAassigner[43] subtypes using ICGC datasets. **B-C.** Heatmap showing hypergeometric test-based FDR comparing PDAassigner subtypes (x-axis) with the published (B) Bailey's subtypes[24] (y-axis) using ICGC dataset[24] and (C) CRC heterocellular subtypes (y-axis) using GSE15471[21] datasets. **D-E.** Proportions of different PC histotypes in (D) CRC heterocellular subtypes (ICGC[24] datasets), (E) goblet-like subtype. **F.** Proportions of CRC heterocellular subtypes (ICGC dataset) in different PC histotypes. **G.** Proportions of Bailey's subtypes[24] in CRC heterocellular subtypes (ICGC dataset). **H**. GSEA analysis showing enrichment of collagen formation (reactome[33,42] gene set) representing increased desmoplastic reaction in stem-like PC subtype. **I**. Heatmap showing the expression of highly variable (SD>1) marker genes associated with immunogenic, inflammatory, stem-like and goblet-like subtypes.

Supplementary Figure 3: A-C. Heatmap showing hypergeometric test-based FDR values comparing CRC heterocellular subtypes (y-axis) with intrinsic gene expression subtypes from Lei *et al.*[22] in two different gastric cancer datasets (x-axis) (A.) GSE35809[22] and (B) GSE34942[22] and (C) integrative subtype[25] from TCGA gastric cancer. **D.** Kaplan-Meier survival curve showing significant prognostic (overall survival) difference between surgery (n=72) and adjuvant 5-FU (n=22) treatment groups in combined data consisting of stem-like, inflammatory, TA and the mixed subtype. **E-L.** Heatmap showing hypergeometric test-based FDR values comparing CRC heterocellular subtypes (y-axis) with intrinsic gene expression subtypes (x-axis) from (E.) ovarian (GSE9891[1]), (F.) HNSC[2] (TCGA), (G.) UCEC[4] (TCGA), (H.) BLCA[5] (TCGA), (I.) KIRC[6] (TCGA), (J.) LAUD/LUSC (GSE37745[7]), (K) LUAD[44] (TCGA) and (L) LUSC[45] (TCGA). **M.** GSEA results showing the enrichment of macrophages in MSI inflammatory samples from CRC, GC and UCEC. **Note:** Immuno reactive (IR), chromatin remodelling (CR), base excision repair (BER), genomically stable (GS), Epstein-Barr virus (EBV), chromosomal instability (CIN).

Supplementary Figure 4: A. Enrichment of hallmarks[42] genesets in *KRAS*-independent stem-like subtypes compared to the *KRAS*-dependent goblet-like in CRC[36], PC[24] and LUAD[44]. **B.** Proportion of *KRAS*-independent and dependent cell lines in goblet-like and stem-like subtypes in PC and LUAD. The KRAS dependency status for cell lines was consolidated from the published studies by Collission et al.[43] and Singh et al.[40], which used *KRAS* lentiviral assay to suppress *KRAS* expression followed by proliferation analysis in *KRAS* mutant cell lines across lung and PC cell lines. **C.** Proportion of *KRAS*-independent and -dependent CRC cell lines in stem-like and goblet-like subtypes. **D.** Status of *PIK3CA* in *KRAS* dependent and independent CRC cell lines present in Singh et al[40].

Supplementary Figure 5:
A. Heatmap showing hypergeometric test-based FDR values comparing CRC heterocellular subtypes (y-axis) with intrinsic gene expression BRCA from GSE42568[19] (x-axis). Due to the unavailability of the original subtypes in BRCA (GSE42568[19]) the samples were reclassified using the genefu[32] package. **B.** Kaplan-Meier survival curve showing recurrence free survival (RFS) difference between Luminal-A stem-like (n=34) versus the other subtypes (n=37). C. GSEA plot showing enrichment of stem cell genes in stem-like subtype of luminal A BRCA compared to the other CRC-SET subtypes of luminal A. D. Heatmap showing the expression of top highly variable genes (SD>1.5; n=876) between stem-like and other subtypes within luminal A BRCA subtype.

Legends for the Tables:

Supplementary Table 1: A-F. Results from the supervised classification of multiple cancers to CRC heterocellular subtypes in all the datasets used in this study. **G.** Comparison of the similarities of the heterocellular subtypes from multiple cancers to the CRCassigner subtypes**. H.** Assessing the subtype-specific genes in the CRCassigner-786 gene signature. **I.** Colon-specific genes present in the TiGER database. **J.** Assessment of the tissue-specific genes in the CRCassigner signatures. **K.** Chi-square test comparing the CRC heterocellular subtypes with the intrinsic subtypes from cancers percentage of tissue specific genes CRCassigner signatures. **L-N.** Table showing the counts, proportions and hypergeometric test results comparing the intrinsic classification with the CRC heterocellular classification.

Supplementary Table 2: A-B. Results showing the comparison of CRC heterocellular subtypes from PC with the histological subtypes. **C.** GSEA results showing the enrichment of Reactome genesets in stem-like heterocellular subtypes compared to the others. **D.** Expression of 5 common immune genes in immunogenic and inflammatory samples. **E.** Custom genesets created using the published immune markers. **F.** GSEA results showing the enrichment of these immune markers in inflammatory compared to the immunogenic subtypes.

Supplementary Table 3: A-I. Results showing the microsatellite status (counts of proportions) in CRC heterocellular subtypes from CRC, GC and UCEC. **J-L.** GSEA results showing the enrichment of immune pathways in inflammatory MSI compared to the goblet-like MSI samples in CRC, GC and UCEC. **M.** RPPA data showing the expression of PDL1 proteins in 3 cancers.

Supplementary Table 4: A-F. Results showing the *KRAS* mutation status in CRC, LUAD and PC. **G-L**. *KRAS* dependency status predicted using the NTP. **M-O**. GSEA enrichment analysis showing the enrichment of hallmark genesets in stem-like *KRAS* independent versus the goblet-like KRAS dependent in 3 cancers. **P-R**. KRAS enrichment status detected using the lentiviral assay in cell lines from 3 cancers.

Supplementary Table 5: A. Results showing the GSEA enrichment analysis between the luminal A stem-like versus all other luminal A subtypes. **B-D.** Counts and proportions comparing the luminal A copy number subtypes with the CRC heterocellular subtypes.

References:

1 Tothill, R. W. *et al.* Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clinical cancer research : an official journal of the American Association for Cancer Research* **14**, 5198-5208, doi:10.1158/1078-0432.CCR-08-0196 (2008).

2 Cancer Genome Atlas, N. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576-582, doi:10.1038/nature14129 (2015).

3 Noll, E. M. *et al.* CYP3A5 mediates basal and acquired therapy resistance in different subtypes of pancreatic ductal adenocarcinoma. *Nature medicine* **22**, 278-287, doi:10.1038/nm.4038 (2016).

4 Cancer Genome Atlas Research, N. *et al.* Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67-73, doi:10.1038/nature12113 (2013).

5 Cancer Genome Atlas Research, N. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315-322, doi:10.1038/nature12965 (2014).

6 Cancer Genome Atlas Research, N. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499**, 43-49, doi:10.1038/nature12222 (2013).

7 Botling, J. *et al.* Biomarker discovery in non-small cell lung cancer: integrating gene expression profiling, meta-analysis, and tissue microarray validation. *Clinical cancer research : an official journal of the American Association for Cancer Research* **19**, 194-204, doi:10.1158/1078-0432.CCR-12-1139 (2013).

8 Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic acids research* **41**, D991-995, doi:10.1093/nar/gks1193 (2013).
9 Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome biology* **5**, R80, doi:10.1186/gb-2004-5-10-r80 (2004).
10 Kohl, M. & Deigner, H. P. Preprocessing of gene expression data by optimally robust estimators. *BMC bioinformatics* **11**, 583, doi:10.1186/1471-2105-11-583 (2010).
11 Brettschneider J, C. F., Bolstad BM, and Speed TP. Quality assessment for short oligonucleotide arrays. *Technometrics* (2007).
12 Bolstad BM, C. F., Brettschneider J, Simpson K, Cope L, Irizarry RA, and Speed TP. *Quality Assessment of Affymetrix GeneChip Data in Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. (2005).
13 Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607, doi:10.1038/nature11003 (2012).
14 Bruford, E. A. *et al.* The HGNC Database in 2008: a resource for the human genome. *Nucleic acids research* **36**, D445-448, doi:10.1093/nar/gkm881 (2008).
15 Carlson, M. hgu133plus2.db: Affymetrix Human Genome U133 Plus 2.0 Array annotation data (chip hgu133plus2).
16 Sadanandam, A. *et al.* A Cross-Species Analysis in Pancreatic Neuroendocrine Tumors Reveals Molecular Subtypes with Distinctive Clinical, Metastatic, Developmental, and Metabolic Characteristics. *Cancer Discov* **5**, 1296-1313, doi:10.1158/2159-8290.CD-15-0068 (2015).
17 Sadanandam, A. *et al.* A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nature medicine* **19**, 619-625, doi:10.1038/nm.3175 (2013).
18 Sadanandam, A. A cross-species analysis of pancreatic neuroendocrine tumours identifies novel molecular subtypes with distinct cellular origin, metabolism and metastatic potential. *Cancer Discovery* (2015).
19 Clarke, C. *et al.* Correlating transcriptional networks to breast cancer survival: a large-scale coexpression analysis. *Carcinogenesis* **34**, 2300-2308, doi:10.1093/carcin/bgt208 (2013).
20 Jorissen, R. N. *et al.* Metastasis-Associated Gene Expression Changes Predict Poor Outcomes in Patients with Dukes Stage B and C Colorectal Cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **15**, 7642-7651, doi:10.1158/1078-0432.CCR-09-1431 (2009).
21 Badea, L., Herlea, V., Dima, S. O., Dumitrascu, T. & Popescu, I. Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia. *Hepato-gastroenterology* **55**, 2016-2027 (2008).
22 Lei, Z. *et al.* Identification of molecular subtypes of gastric cancer with different responses to PI3-kinase inhibitors and 5-fluorouracil. *Gastroenterology* **145**, 554-565, doi:10.1053/j.gastro.2013.05.010 (2013).

23 Hoadley, K. A. *et al.* Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929-944, doi:10.1016/j.cell.2014.06.049 (2014).

24 Bailey, P. *et al.* Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* **531**, 47-52, doi:10.1038/nature16965 (2016).

25 Cancer Genome Atlas Research, N. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**, 202-209, doi:10.1038/nature13480 (2014).

26 Flicek, P. *et al.* Ensembl 2014. *Nucleic acids research* **42**, D749-755, doi:10.1093/nar/gkt1196 (2014).

27 Zhang, J. *et al.* BioMart: a data federation framework for large collaborative projects. *Database : the journal of biological databases and curation* **2011**, bar038, doi:10.1093/database/bar038 (2011).

28 Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nature medicine*, doi:10.1038/nm.3967 (2015).

29 Liu, X., Yu, X., Zack, D. J., Zhu, H. & Qian, J. TiGER: a database for tissue-specific gene expression and regulation. *BMC bioinformatics* **9**, 271, doi:10.1186/1471-2105-9-271 (2008).

30 Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 14863-14868 (1998).

31 Sadanandam, A. *et al.* Reconciliation of classification systems defining molecular subtypes of colorectal cancer: interrelationships and clinical implications. *Cell cycle* **13**, 353-357, doi:10.4161/cc.27769 (2014).

32 Gendoo, D. M. *et al.* Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. *Bioinformatics* **32**, 1097-1099, doi:10.1093/bioinformatics/btv693 (2016).

33 Milacic, M. *et al.* Annotating cancer variants and anti-cancer therapeutics in reactome. *Cancers* **4**, 1180-1211, doi:10.3390/cancers4041180 (2012).

34 Le, D. T. *et al.* PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *The New England journal of medicine* **372**, 2509-2520, doi:10.1056/NEJMoa1500596 (2015).

35 Ciriello, G. *et al.* The molecular diversity of Luminal A breast tumors. *Breast cancer research and treatment* **141**, 409-420, doi:10.1007/s10549-013-2699-3 (2013).

36 Cancer Genome Atlas, N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-337, doi:10.1038/nature11252 (2012).

37 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).

38 Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G. & Hacohen, N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160**, 48-61, doi:10.1016/j.cell.2014.12.033 (2015).

39 Li, J. *et al.* TCPA: a resource for cancer functional proteomics data. *Nat Methods* **10**, 1046-1047, doi:10.1038/nmeth.2650 (2013).

40 Singh, A. *et al.* A gene expression signature associated with "K-Ras addiction" reveals regulators of EMT and tumor cell survival. *Cancer cell* **15**, 489-500, doi:10.1016/j.ccr.2009.03.022 (2009).
41 Hoshida, Y. Nearest template prediction: a single-sample-based flexible class prediction with confidence assessment. *PloS one* **5**, e15543, doi:10.1371/journal.pone.0015543 (2010).
42 Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell systems* **1**, 417-425, doi:10.1016/j.cels.2015.12.004 (2015).
43 Collisson, E. A. *et al.* Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nature medicine* **17**, 500-503, doi:10.1038/nm.2344 (2011).
44 Cancer Genome Atlas Research, N. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543-550, doi:10.1038/nature13385 (2014).
45 Cancer Genome Atlas Research, N. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519-525, doi:10.1038/nature11404 (2012).