SCIENTIFIC REPORTS
nature research

OPEN

Machine Learning and Network Analyses Reveal Disease Subtypes of Pancreatic Cancer and their Molecular Characteristics

Musalula Sinkala*, Nicola Mulder & Darren Martin

Given that the biological processes governing the oncogenesis of pancreatic cancers could present useful therapeutic targets, there is a pressing need to molecularly distinguish between different clinically relevant pancreatic cancer subtypes. To address this challenge, we used targeted proteomics and other molecular data compiled by The Cancer Genome Atlas to reveal that pancreatic tumours can be broadly segregated into two distinct subtypes. Besides being associated with substantially different clinical outcomes, tumours belonging to each of these subtypes also display notable differences in diverse signalling pathways and biological processes. At the proteome level, we show that tumours belonging to the less severe subtype are characterised by aberrant mTOR signalling, whereas those belonging to the more severe subtype are characterised by disruptions in SMAD and cell cycle-related processes. We use machine learning algorithms to define sets of proteins, mRNAs, miRNAs and DNA methylation patterns that could serve as biomarkers to accurately differentiate between the two pancreatic cancer subtypes. Lastly, we confirm the biological relevance of the identified biomarkers by showing that these can be used together with pattern-recognition algorithms to accurately infer the drug sensitivity of pancreatic cancer cell lines. Our study shows that integrative profiling of multiple data types enables a biological and clinical representation of pancreatic cancer that is comprehensive enough to provide a foundation for future therapeutic strategies.

Pancreatic cancer is a heterogeneous disease that is characterised by poor clinical outcomes and few effective treatment options. Attempts to define a standard classification for tumours of the pancreas have been ongoing for decades[1–3]. In general, the approaches that are currently used for making both outcome predictions and treatment decisions are based on histological subtyping and clinical parameters such as the disease stage, metastasis, and the resectability of tumours[4,5]. Recently, however, the advent of molecular profiling has laid the foundation for quantitatively profiling tumours based on their genome-wide gene transcription profiles, protein expression profiles and/or mutational landscapes[6–9]. These profiling methods promise a more accurate and precise definition of tumour subtypes and better predictions of how particular tumour types will respond to different treatments.

Further, molecular data that is used to construct the molecular profiles of particular cancers have been used to identify the perturbances in the cellular regulatory networks that characterize these cancers: often revealing numerous potential drug targets within various signalling pathways. This molecular data together with the known molecular profiles of numerous well characterized cancer cell lines can even be leveraged using machine learning methods to predict the responses of particular patient tumour subtypes to different anticancer drugs[10,11].

A crucial resource for the discovery of useful diagnostic biomarkers and potential anticancer drug targets are large-scale datasets comprising, among other data types, extensive genomic, transcriptomic and proteomic profiles of matched healthy and tumorous tissues. These datasets, which are compiled and maintained by The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) are helping us uncover the molecular characteristics and signalling pathway perturbations that define specific cancer subtypes[12,13].

Among the cancers that are well represented in these data collections is pancreatic cancer. Molecular profiling analyses of the pancreatic tumour datasets have identified both distinct pancreatic cancer subtypes, and mutations of the genes, KRAS, TP53, SMAD4 and CDKN2A as potential drivers of pancreatic cancer[14–18]. Although

University of Cape Town, School of Health Sciences, Department of Integrative Biomedical Sciences, Computational Biology Division, Anzio Rd, Observatory, 7925, Cape Town, South Africa. *email: smsinks@icloud.com

the biomarkers that differentiate between different pancreatic cancer subtypes could eventually inform treatment decisions, there are as yet no available subtype-specific treatment options for this type of cancer. There is, therefore, a pressing need to, firstly, find a set of biomarkers that can be used to accurately and sensitively diagnose pancreatic cancer subtypes and, secondly, to identify suitable targets for drug development among these biomarkers.

Definitions of disease subtypes is a perpetual process, with classifiers and cut-offs that differentiate between the subtypes, essentially needing to be continually re-defined and refined as more molecular data and better molecular profiling tools become available. As classification schemes for pancreatic cancers improve, it is expected that additional specific molecular correlates of patient survival, responses to anticancer drugs, and tumour aggressiveness will be uncovered. Armed with such knowledge, we could develop better prognostic and diagnostic methods, and select the best drugs to treat specific pancreatic cancer subtypes. Further, more subtype-specific molecular features could potentially enhance the accuracy with which machine learning methods could predict the drug response profiles of specific pancreatic tumours, thus leading to improved disease outcomes.

However, it remains technically difficult to effectively leverage the diverse and ever-increasing data relating to pancreatic tumours[19–21]. These difficulties include, but are not limited to, inconsistent classifications of patient tumours when the tumours are subtyped using different types of molecular data, and the efficient integration and analysis of different data types to yield consistent identifications of the causal disruptors of the molecular processes that underlie the observed differences between pancreatic cancer subtypes[19]. Ultimately, these difficulties undermine efforts to predict the responses of tumours to drugs: an endeavour involving comparisons between the relevant molecular features of a novel tumour with those of well-characterized tumour subtypes or tumour cell lines.

With these issues in mind, we attempted to identify clinically relevant subtypes of pancreatic cancer accounting for the full spectrum of molecular and clinical data available for pancreatic cancer tumours in the TCGA dataset. We address the problem of inconsistent tumour classifications that are obtained using different types of molecular data, by applying an integrative classification approach that considered all the available molecular data types. As expected, our analyses identified discrepancies between various classification schemes but ultimately supported the existence of two major pancreatic cancer subtypes. Besides uncovering the likely molecular causes of altered biological processes within the tumours of these two subtypes, we identified biomarker sets that can be used to accurately and sensitively classify novel pancreatic tumours. Further, in the face of multiple high-dimensional data types, we show that statistical models that capture the complexity of disease can aid in the identification of relevant drugs and drug targets that might offer substantial benefits for patients afflicted with tumours belonging to either of the pancreatic cancer subtypes.

Results

Subtypes of pancreatic cancer and their clinical characteristics. We applied K-means clustering to the reverse phase protein array (RPPA) determined proteomics data of the 45 high-purity pancreatic cancer samples that are available in the TCGA database to identify two coherent clusters of patient tumours (Fig. S1A)[22]. Then, we compared this clustering of pancreatic cancer samples to other subtypes that are reported in the literature for various other molecular data types (DNA methylation status, protein expression levels and mRNA/miRNA transcription levels) and established that the samples clustered differently depending on the specific molecular data type used (Fig. 1A).

To mitigate this problem, we applied a multi-platform integrative clustering method called similarity network fusion (SNF). SNF solves the disparate clustering problem by constructing similarity networks of samples for each available molecular data type and then efficiently fuses these into one network that represents clustering based on all the underlying data types (Fig. 1B)[19]. Using DNA methylation status, protein expression, mRNA transcription and miRNA data of the 45 high purity cancer tumour samples available in TCGA, we applied the SNF clustering method to identify two-cluster and three-cluster clustering solutions (Fig. 1C).

The pancreatic cancer subtypes in the two-cluster solution comprised 25 and 20 tumours, which we provisionally named as subtype-1 and subtype-2, respectively. Interestingly, the SNF clustering solutions were highly concordant with each of the clustering solutions obtained using individual molecular data types but were most similar to that obtained using the proteomics data (refer to Fig. 1C).

Next, we sought to understand whether the identified pancreatic cancer subtypes were associated with different clinical outcomes. Indeed, we found that the two groups of patients differed with respect to the overall percentages of individuals with progressive disease and the percentages of individuals who eventually died. Here we found that the patients with subtype-1 tumours were more likely to survive than those with subtype-2 tumours (75% vs 35% survival, respectively; Fig. S1C). We further observed a nearly 50% lower median disease-free survival (DFS) period for patients with subtype-2 tumours (DFS = 12.42 months) than for patients with subtype-2 tumours (DFS = 25.07 months; Fig. S1D). Likewise, the overall survival (OS) periods for the patients with subtype-2 tumours (OS = 16.05 months) were shorter than those with subtype-1 tumours (OS = 23.06 months; Fig. S1E). However, our analysis of OS and DFS periods using the Kaplan-Meier methods revealed no statistically significant difference between the pancreatic cancer subtypes; possibly due to the small sample size (Fig. S1D,E)[23].

Proteomics-based signalling pathway analyses distinguish disease subtypes. For each disease subtype, we compared the enrichment of KEGG pathways and Gene Ontology (GO) biological process classifications of proteins that were upregulated within tumour belonging to each of the subtypes using Enrichr[24]. We found that whereas certain pathways were differentially altered between tumours belonging to different subtypes, other pathways were consistently altered (albeit to different extents in some cases) in the tumours of both subtypes (Fig. 2A, also see Supplementary File 1).

The mTOR signalling pathway was altered in subtype-1 tumours but not in subtype-2 tumours (combined score = 85, hypergeometric test; $p = 2.1 \times 10^{-19}$). Within the mTOR pathway of subtype-1 tumours, we found

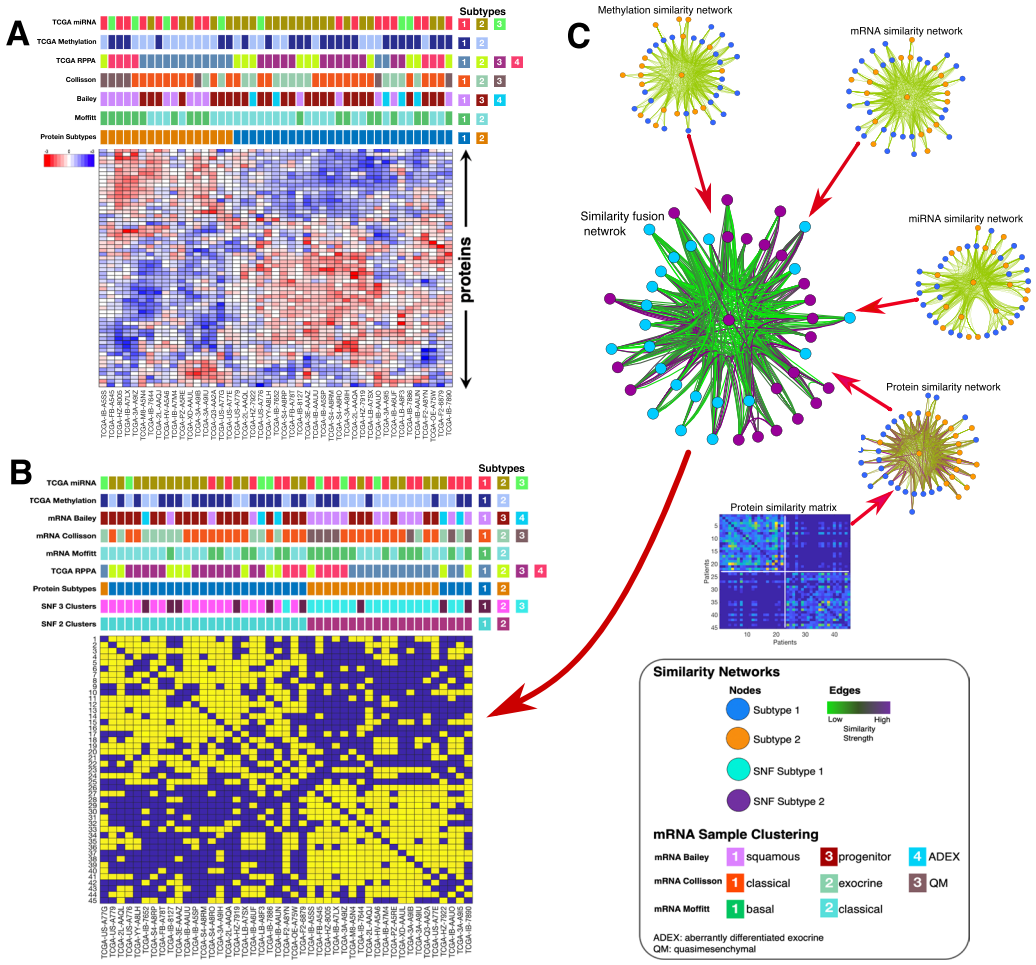

Figure 1. Classification of pancreatic cancer: (**A**) Comparison between the proteomics-based subtyping of pancreatic cancers using unsupervised hierarchical clustering, to other classification schemes from top to bottom: TCGA's (Raphael *et al.*, 2017) miRNA, RPPA, and DNA methylation; mRNA-based classification schemes using the gene biomarkers established by Collosson *et al.*; Bailey *et al.*; and Moffitt *et al.* (**B**) Illustrative example of SNF steps: similarity matrices are used to create patient networks from protein, mRNA, miRNA and DNA methylation data showing patient-to-patient similarities for the 45 pancreatic cancer patients. The network nodes represent patients. The colours of edges joining nodes indicate the degree of similarity between pairs of patients. The nodes of the fused network are coloured according to the subtypes to which the patient tumours were assigned using spectral clustering of the combined patient network. (**C**) Comparison between the SNF subtyping using spectral clustering to other classification schemes from top to bottom: TCGA's[14] miRNA, and DNA methylation classifications; mRNA-based classification schemes[8,48]; TCGA's RPPA classification, our K-means clustering classification; our 3-cluster SNF classification; and our 2-clusters SNF classification.

increased expression of well-documented oncoproteins including MTOR and BRAF: both of which have previously been linked to pancreatic carcinogenesis (Fig. 2B)[25–27].

Further, we found that proteins that are involved in the KEGG Cancer Pathways were dysregulated in both the subtype-1 and subtype-2 tumours; these pathways encompass several known oncoproteins (such as RAD51, BRAC1, and ERBB2) and tumour suppressor proteins (such as PTEN and CDK2A1)[28–30] (Fig. 2C). Despite the upregulation of the KEGG Cancer Pathways in tumours belonging to both subtypes, we found that the clustering of patients using only proteins within these cancer pathways was concordant with our subtype classification (Fig. 2D). Such a clustering pattern indicates that even when the same pathways are altered in both subtype-1 and subtype-2 tumours, the exact nature of the alterations within these pathways still differs between the two tumour subtypes. For example, whereas subtype-1 tumours exhibit hyperactivation of mTOR-associated signalling, subtype-2 tumours display increased activation of SMAD4-associated signalling. Also, we found that other proteins involved in mTOR signalling were both more strongly correlated and more highly expressed in subtype-1 tumours than they were in subtype-2 tumours, indicating the hyperactivation of this pathway requires the increased expression of most of the mTOR signalling proteins (Fig. 3A). Likewise, SMAD4 signalling pathway protein expression levels also differed significantly ($p = 2 \times 10^{-4}$) between these subtypes (Fig. 3B).

We further attempted to identify the kinases that likely phosphorylate substrates within the various signalling pathways of pancreatic tumour cells. Using kinase enrichment analysis (KEA), we found a subset of kinases that

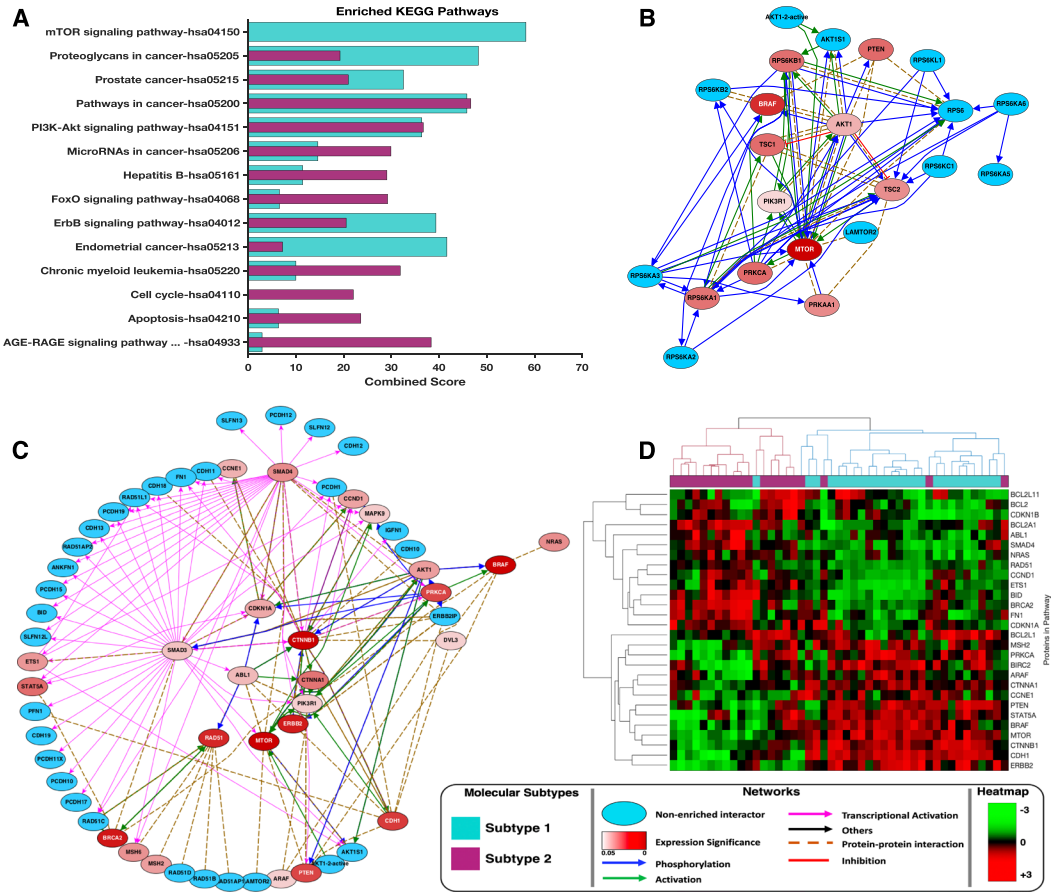

Figure 2. Pathway enrichment analyses: (**A**) KEGG pathways enrichment results of the most significantly altered pathways in tumours belonging to each of the inferred pancreatic cancer subtypes. Refer to Supplementary File 1 for the complete list of KEGG pathways enriched based on the proteomics data. (**B**) mTOR signalling pathways found to be uniquely altered in subtype-1 tumours. Blue nodes indicate proteins with expression levels that were either not significantly altered between the subtypes or were not measured by the TCGA. Red coloured nodes represent proteins with significantly altered expression levels with the degree of statistical significance being expressed as the negative logarithm of Benjamin-Hochberg adjusted p-values. The connectivity of network components was extracted from the KEA, ChEA, and UCSC super pathway databases. (**C**) KEGG cancer pathways found to be consistently altered in tumours belonging to both pancreatic cancer subtypes. (**D**) Clustergram of tumours using only the proteins that are members of the KEGG cancer pathways ontology.

might drive pancreatic carcinogenesis, including, among others (Supplementary File 1), AKT1 ($p = 8.2 \times 10^{-03}$), MTOR ($p = 0.011$), and RPS6KA1 ($p = 0.0499$) (Fig. 3C)[31]. We observed a moderate positive correlation between proteins involved in mTOR signalling and their phosphorylated forms (Fig. 3D). Further, our results show that the protein phosphorylation pattern among the two pancreatic cancer subtypes is distinctive. Here, we found that in subtype-1 tumours various phosphoproteins that participant in mTOR signalling – such as MTOR-pS2448, GSKB-pS21-S9, PDK-pS241 and growth factor receptors EGFR-pY1068 and ERBB-pY1248 – all exhibited increased phosphorylation (Fig. 3E)[32,33]. These phosphoproteomics analyses support our initial findings (using dephosphorylated proteins) that subtype-1 tumours display increased mTOR signalling. Conversely, for subtype-2 tumours, we found elevated phosphorylation levels of proteins such as CDK1-pY15, p27-pT158 and p27-pT198 (Fig. 3E) which are involved in cell-cycle-associated processes[34].

Overall, our findings suggest that for tumours of the two major pancreatic cancer subtypes, oncogenesis may be primarily driven by perturbation in either SMAD4 or mTOR signalling.

Pancreatic cancer subtypes exhibit functional differences in mRNA levels and DNA methylation patterns. We attempted to determine whether any GO molecular functions were enriched for among the overexpressed genes that differentiated the two pancreatic cancer subtypes. Specifically, we queried Enrichr using mRNA transcripts that were significantly upregulated across all of the tumours belonging to a particular subtype (see Supplementary File 2)[24]. We found that the over-transcribed genes in subtype-2 tumours were enriched for, among others, molecular functions associated with transmembrane transporter and G-protein coupled receptor activities (Fig. 4A, also see Supplementary File 1). Alternatively, the over-transcribed genes in subtype-1 tumours

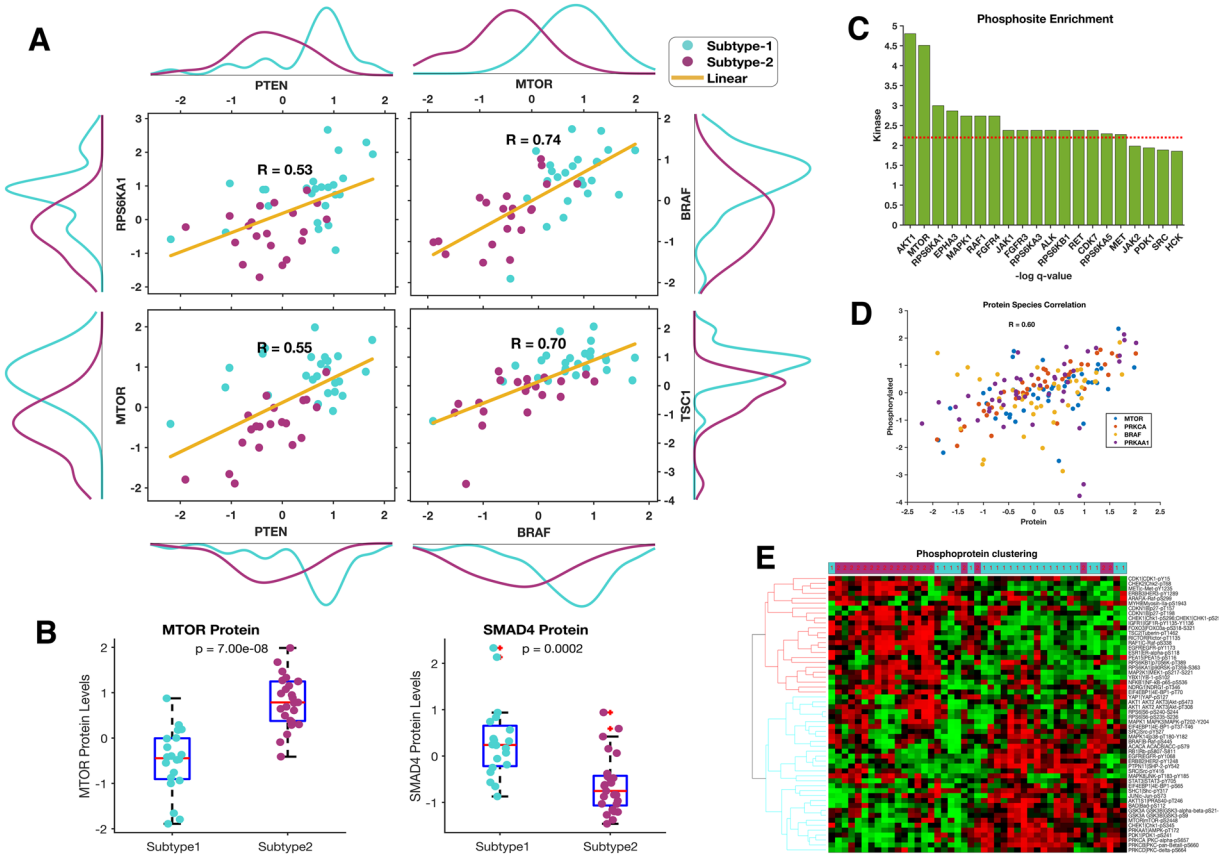

Figure 3. (**A**) Pearson's correlation values of some proteins involved in mTOR signalling. The plot shows relatively higher expression levels of these proteins in subtype-1 tumours compared to subtype-2 tumours. (**B**) Boxplots show mTOR and SMAD4 protein expression biomarker of the SNF subtypes. (**C**) Enriched phosphosites identified by kinase enrichment analysis: the negative logarithm values of the Benjamin-Hochberg adjusted p-value are plotted on the y-axis while kinases are plotted along the x-axis. The red line represents the cut-off values at the 10% false discovery rate. (**D**) Correlation between the phosphorylated and de-phosphorylated proteins species for proteins involved in the mTOR signalling pathway. (**E**) Unsupervised hierarchical clustergram of tumours phosphoproteins showing high concordance with the clustering obtained from all the proteins (de-phosphorylated and non-phosphorylated protein) profiled by the TCGA. The clustergram was produced using the Spearman correlation distance metric and the complete linkage.

were enriched for, among others, molecular functions that are associated with phosphoinositide 3-kinase signalling, peptidase enzyme activity and growth factor receptors (Fig. 4A, also see Supplementary File 1).

We explored the enriched KEGG pathways that were differentially expressed between the two pancreatic cancer subtypes using lists of genes with methylation profiles and mRNA transcription levels that differed between the subtypes (see Supplementary File 2). Interestingly, we found that only subtype-1 tumours displayed enrichment for pancreatic secretions (Fig. S2A). These results corroborate both our previously noted enrichment in subtype-1 tumours of mRNAs involved in transmembrane transport, and published observations that the secretion of compounds from the pancreas and other organs is associated with increased transmembrane transporter activity[35].

Similarly, for both enrichment analyses using differentially expressed mRNA and proteins, we found enrichment for components of the AGE-RAGE signalling pathway in subtype-2 tumours (Figs. 2A and S3A). The AGE-RAGE system promotes the development of various types of cancers, including those of the pancreas and prostate, through diminished apoptosis and increased cell viability[36,37]. Therefore, targeted inhibition of RAGE may serve as an effective treatment strategy against subtype-2 tumours.

In addition to these findings, the DNA methylation data revealed that while the methylation landscapes of subtype-1 and subtype-2 tumours were generally similar, the subtype-1 tumours had some additional genes displaying significantly increased DNA methylation (Supplementary File 2). We noted that these hypermethylated genes participate in various cellular pathways including focal adhesion, RAP1-signalling, and actin cytoskeleton regulation (Fig. S2B). Since these DNA methylation alterations are unique to subtype-1 tumours, they could be associated with reduced pancreatic tumour aggressiveness.

Unexpectedly, we observed no significant differences in mutation distributions and gene copy number alterations for the genes with transcription and translation profiles that differed between the two subtypes (Fig. 4B).

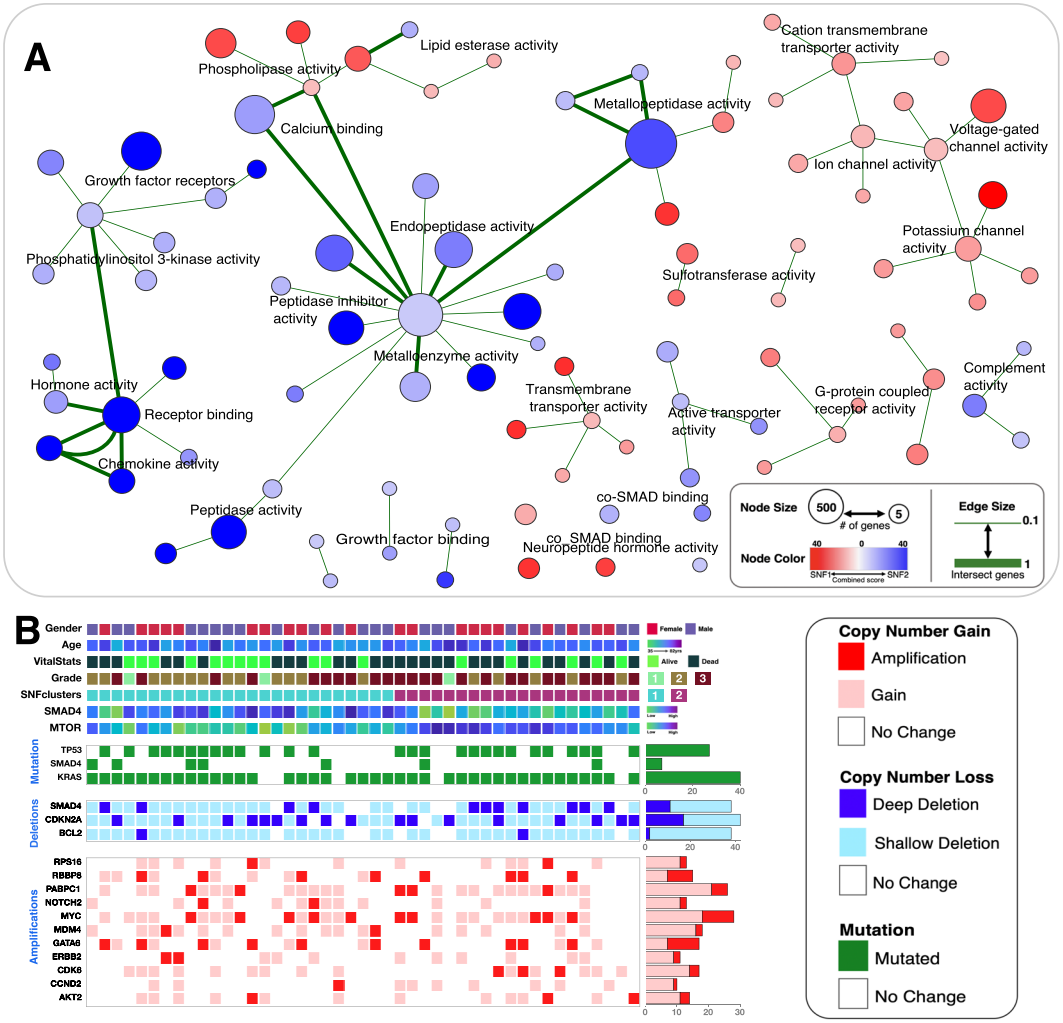

Figure 4. **(A)** Network of Gene Ontology (GO) molecular functions found enriched between the two pancreatic cancer subtypes. Enrichr was used to obtain enriched GO-terms that were visualised in Cytoscape (refer to the methods section). Each node represents a GO-term with similar nodes clustered together and connected by edges with the number of shared genes between the nodes being represented by the thickness of the edges. The size of each node denotes the gene set size of the represented GO-term. The colour of each node represents the magnitude of the combined enrichment score: red represent enrichment in subtype-1 tumours and blue represents enrichment in subtype-2 tumours. **(B)** The integrated plot of clinical and molecular features of 45 tumour samples ordered by their SNF clustering positions. From top to bottom panels indicate: patient gender; Age at which a condition or disease was first diagnosed; neoplasm histological grade; SNF subtype of tumour; SMAD4 protein expression level; mTOR protein expression level; significantly mutated genes: TP53, SMAD4 and KRAS gene mutations; SMAD4, CDKN2A and BCL2 gene deep deletion (dark blue) and shallow deletion (pale blue); gene amplification (red) and copy number gain (pink) of multiple genes.

Biomarker genes, proteins and miRNA sets that define the pancreatic cancer subtypes. Given that different types of molecular data yield different patterns of tumour clustering, we attempted to identify biomarker genes, proteins or miRNA sets that best differentiated between the two pancreatic cancer subtypes. It was anticipated that these sets of biomarker genes might allow for consistent classification of pancreatic cancer patients using machine learning methods applied to only one category of molecular data.

To extract relevant features for each category of molecular data, we applied the diagonal adaptation of neighbourhood component analysis (NCA) for classification with regularisation[38]. NCA learns feature weights for minimisation of an objective function that measures the average leave-one-out classification loss over the training data (Fig. S3A,B)[38].

Using NCA, we identified biomarker sets comprising 50 mRNAs, 49 methylated genes, 14 proteins, and 20 miRNAs. For these biomarker sets, we separately applied hierarchical clustering to each of the different molecular data categories to consistently and accurately reproduce the pancreatic cancer subtype classifications (Fig. 5A–D). Also, we individually applied supervised machine learning methods to the 50 mRNA, and the 49 methylated gene sets to classify tumours into subtype-1 and subtype-2 categories. For this, we used the K-nearest neighbour (KNN) algorithm for the mRNA expression data and the support vector machines (SVM) classifier (see

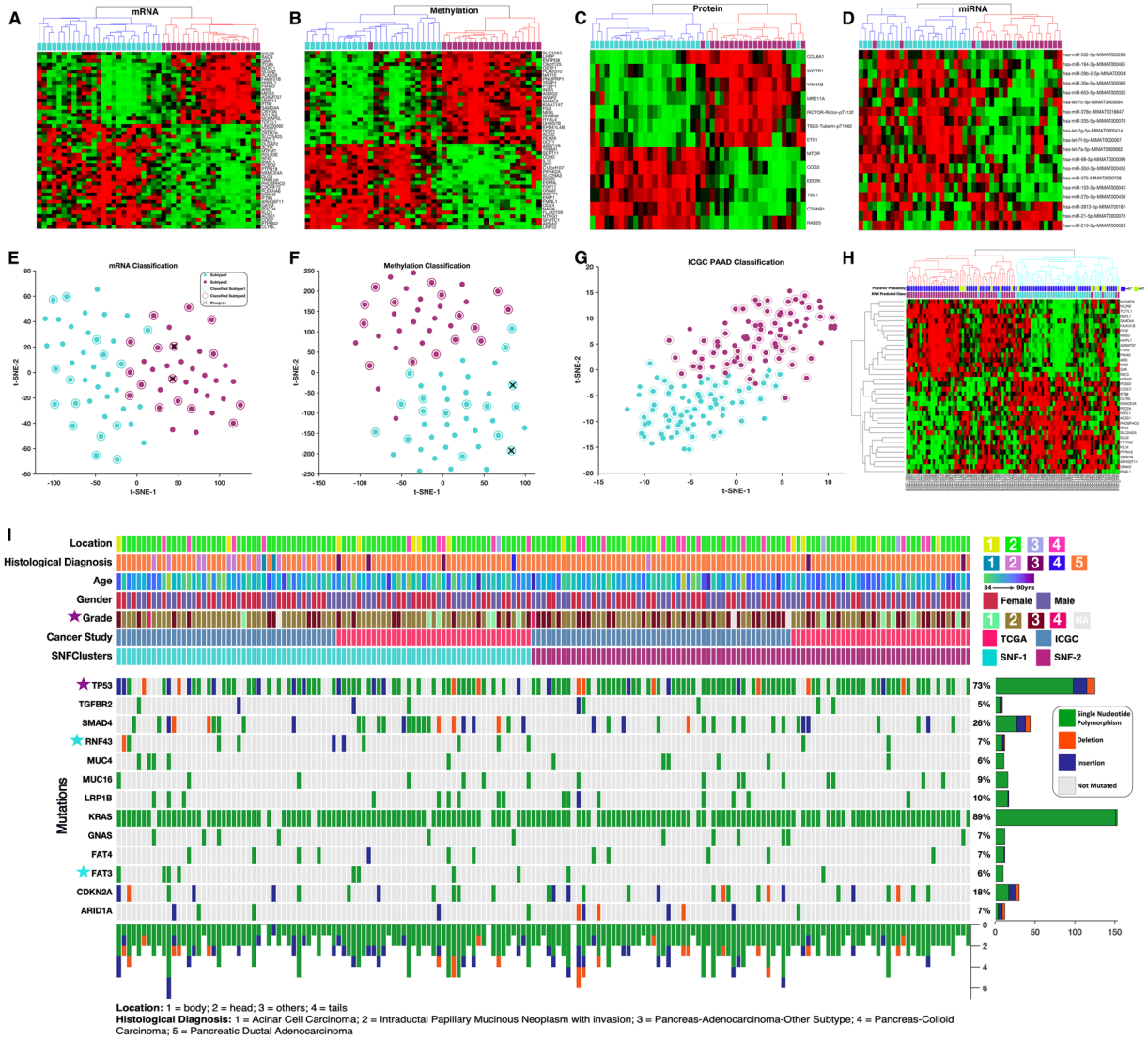

Figure 5. Clustered heatmap of tumours using the **(A)** mRNA biomarker gene set, **(B)** DNA methylation biomarker gene set, **(C)** protein biomarker set, and **(D)** miRNA biomarker set. All the heatmaps (In **A**–**D**) were produced using unsupervised hierarchical clustering with the cosine distance metric and complete linkage. The coloured bars on each clustergram shows the original subtype classification of each patient's tumour found by applying SNF and spectral clustering to all molecular data sets. **(E)** Supervised classification of cancer patients using the mRNA biomarker set trained on the KNN-machine learning model. **(F)** Supervised classification of cancer patients using the DNA methylation biomarker set trained on an SVM-machine learning model. For both plots (**E**,**F**), t-SNE was used to visualise the tumour classes using the exact algorithm and squared Euclidean distance metric. Circled points represent newly classified TCGA pancreatic cancer patients, whereas un-circled points represent the original 45 tumour samples that were used to train the models. Crossed points represent disagreement between the mRNA-based model and the DNA methylation-based model. **(G)** Supervised classification of ICGC cancer patients using the mRNA-based KNN model trained on TCGA data. Circled points represent newly classified ICGC pancreatic cancer patients, whereas un-circled points represent the original 45 TCGA tumour samples that were used to train the model. **(H)** Unsupervised hierarchical clustering of the ICGC patients using the mRNA biomarker gene. The coloured bar on the clustergram shows the KNN model predicted class. (**I**) The integrated plot of clinical and molecular features for the TCGA and ICGC patient's data, ordered by their integrative (SNF) clustering. From top to bottom panels indicate primary tumour location; neoplasm histological type; patient gender; age at diagnosis; neoplasm histological grade; cancer study; integrative tumour subtypes; non-silent gene mutations. The key to the number coding of tumour location and histological diagnosis is at the bottom.

methods section) for the DNA methylation data to achieve very accurate subtype classifications of the tumours (Fig. 5E,F)[39,40]. Specifically, we observed five-fold cross-validation classification accuracies of 99% for the mRNA-based KNN classifier and 98% for the DNA methylation-based SVM classifier, with an agreement of 97% (see methods sections).

Decreasing the number of biomarker genes needed to accurately classify tumours from new pancreatic cancer patients would improve the utility of these sets in a clinical diagnostic setting. To identify smaller biomarker gene sets, we used supervised machine learning methods (see methods section) to define a biomarker set of fewer than ten genes, miRNA or proteins that would minimise incorrect classifications (Fig. S3E). Also, we used these biomarker sets to consistently re-classify TCGA pancreatic cancer patients using hierarchical clustering (Fig. S3F–I). These results imply that smaller gene sets could potentially be useful in a clinical diagnostic setting.

To validate the performance of our 50-mRNA biomarker set, we downloaded pancreatic cancer data from the ICGC data portal[12]. Using the mRNA-based KNN classifier that was trained on TCGA data, we tested the reproducibility of the two-subtype classification scheme by classifying 96 ICGC pancreatic cancer patients into subtypes-1 and subtypes-2 (Fig. 5G). We also applied unsupervised hierarchical clustering to the mRNA biomarker set extracted from the ICGC RNAseq data to reproduce a two-subtype classification analogous to that obtained fusing the TCGA datasets (Fig. 5H). The grouping of ICGC patients yielded by the supervised "TCGA classifier" and the unsupervised "ICGC classifier" agreed on the classifications of 94% of the patients. We observed that 5% of patients with posterior subtype membership probabilities that were less than 0.7, were more likely to be among the discordant cases, accounting for five out of the seven discordant patients (Fig. 5H)[41].

We examined mutational data for the genes that are frequently altered in pancreatic cancer together with the clinical features of subtype-1 and subtype-2 tumours from all of the patients represented in the TCGA and ICGC datasets (Fig. 5I). Here, we found no significant differences in the gene mutations between the tumour subtypes (see Supplement Table 2). Also, we observed that no genes were consistently altered in all of the tumours belonging to either of the subtypes. Similar to other studies, we discovered that some tumours lack mutations in any of the frequently mutated genes[42,43]. This diversity in the mutational landscape of pancreatic cancer tumours is likely to complicate the discovery of broadly applicable treatment regimens that target driver mutations[43].

Concerning histological features of tumours that might be useful for differentiating between the subtypes, we observed that only subtype-1 tumours displayed evidence of intraductal papillary mucinous neoplasm, whereas only subtype-2 tumours were categorised by histological inspection as being pancreatic adenocarcinomas (Fig. 5I). Further, we found that subtype-1 tumours tended to be assigned a lower grade than subtype-2 tumours ($\chi^2 = 10.3$, $p < 0.01$).

Subtyping pancreatic cancer cell lines and predicting drug responses. We obtained mRNA expression and drug response data for 45 pancreatic cancer cell lines from the Cancer Cell Line Encyclopaedia (CCLE)[10]. We attempted to subtype these cell lines using the biomarker gene set identified using the KNN classifier that we trained on the TCGA mRNA data (Fig. 6A). It is known that cell lines with similar transcription profiles are likely to exhibit similar responses to drug perturbations[10,44,45]. It follows, therefore, that the drug response profiles of cell lines should be predictable based on their gene expression profiles[10,45,46].

We predicted the anti-cancer drug responses of the cell lines from the drug response profiles of the cell lines that are most similar (i.e., the nearest neighbours) to each "query" cell line as determined using an exhaustive KNN searcher model[47]. The Searcher model quantified and stored information concerning similarities between the transcription profiles of all the cell lines. Next, we retrieved the drug response profile of a "query" cell line and those of its nearest neighbours based on squared Euclidean distances from the Searcher model. To infer the drug response of the query cell line, we calculated the median drug response of the retrieved nearest neighbour cell lines to each of the 24 anticancer drugs that were profiled by the CCLE (Fig. 5B). For example, in Fig. 5B, the cell lines SU8686 and PANC1005 both have available drug response profiles in the CCLE database, and both are the nearest neighbours of the cell line, PANC0203. Therefore, we used the mean drug responses of SU8686 and PANC1005 to predict the drug responses of PANC0203 (see methods section) (Fig. 5C).

After predicting the drug responses of all the pancreatic cancer cell lines that also had observed drug response data, we compared the predictions to the observed drug responses. Our drug response predictions displayed substantial agreement with the actual drug responses in that they yielded an average Kappa statistic of 0.67 (Fig. 5D).

Discussion

We conducted a comprehensive analysis of clinically relevant patterns of mutation, methylation, transcription, protein expression, and miRNA synthesis within pancreatic cancer tumours. Several pancreatic cancer studies have previously highlighted the limitations of utilising a single molecular data type to accurately classify pancreatic cancers (Fig. 2A)[8,9,14,48]. Here, we attempted to resolve this issue by employing a multidimensional clustering method capable of simultaneously utilising protein expression, mRNA transcription, DNA methylation and miRNA synthesis data. We found that by integrating across all these molecular data types, pancreatic cancer tumours could be classified into two clinically distinct subtypes: which we have simply named subtype-1 and subtype-2.

We observed that subtype-1 tumours were characterised by alterations of the mTOR signalling pathway, and the expression levels of different mTOR pathway proteins were positively correlated to each other (Fig. 1B–D). This finding is consistent with previous studies based on analyses of mRNA transcription and mutation data which also observed alterations of the mTOR pathway in pancreatic cancers[49–51]. Further, it is well established that some pancreatic cancer subtypes respond well to drugs which inhibit the mTOR pathway[52,53]. Accordingly, we anticipate that subtype-1 tumours will likely be more responsive to such therapies than will subtype-2 tumours.

Interestingly, subtype-2 tumours display unique alterations to cell cycle pathways (Fig. 1F,G). This is consistent with the observation that subtype-2 tumours are clinically more aggressive than subtype-1 tumours in that an element of aggressiveness is the hyperactivation of the cell cycle processes that accelerate tumour growth[54–57].

We noted that, in addition to differences in patterns of protein expression, the two pancreatic cancer subtypes differ with respect to patterns of protein phosphorylation, implying that the kinases that are involved in oncogenic transformation differ between the subtypes. Specifically, whereas subtype-1 tumours show upregulation of

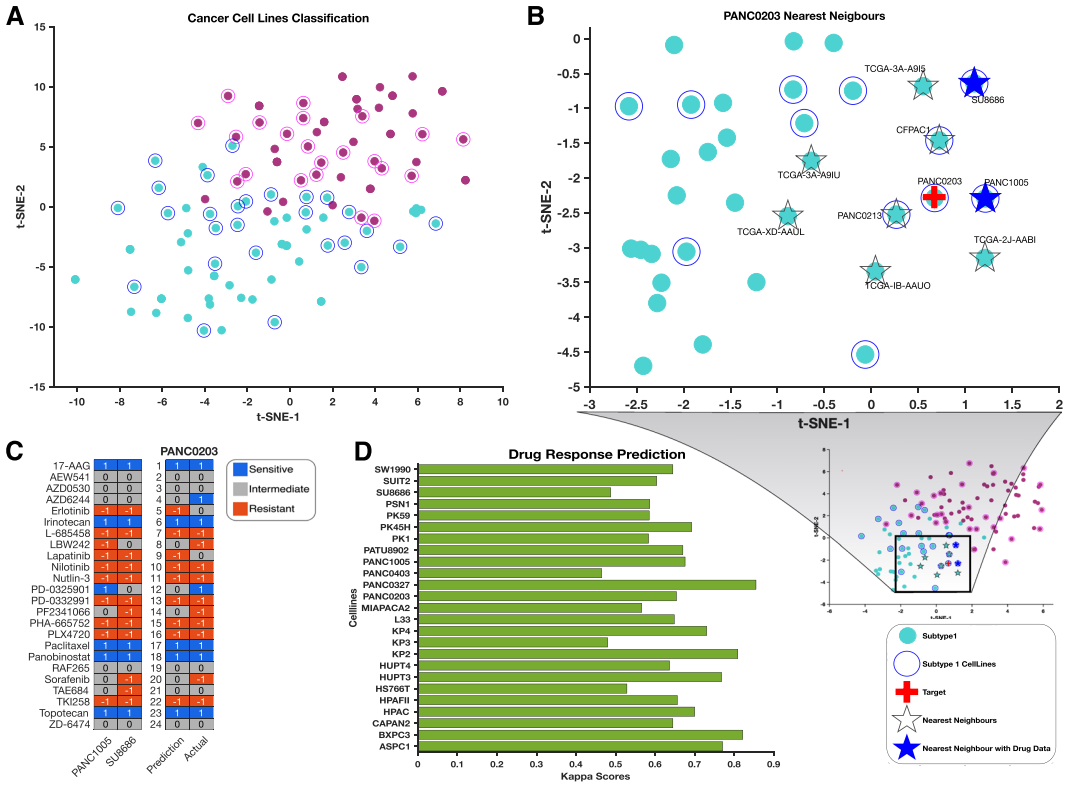

Figure 6. (**A**) Supervised classification of CCLE pancreatic cancer cell lines using the mRNA-based KNN-model trained on TCGA data. t-SNE was used to visualise the tumour classes using the exact algorithm and squared Euclidean distance metric. Circled points represent classified CCLE cell lines, whereas un-circled points represent the TCGA samples used to train the models. (**B**) The t-SNE plot represents the KNN search for the nearest neighbours of PANC0203 in the exhaustive searcher model. Refer to the legend at the right bottom of the figure for interpretation. (**C**) Drug response prediction: first two lanes represent the ranked drug responses to the 24 anticancer drugs of the PANC0203 nearest neighbours (PANC1005 and SU8686) for which such data is available. The last two lanes represent PANC0203's predicted drug responses and its actual drug responses. (**D**) Kappa scores of all CCLE pancreatic cancer cell lines with drug data. The kappa score was calculated using the quadratic method by comparing the actual and predicted drug responses of cell lines to the 24 CCLE anticancer.

mTOR signalling associated kinases (among others, MTOR-pS2448, GSKB-pS21-S9, and PDK-pS241), subtype-2 tumours display upregulation of cell cycle associated kinases (among others, CDK1-pY15, p27-pT158, and p27-pT198; Fig. 3E). Most of these kinases represent credible targets for small molecule inhibitors that might prove useful for subtype-specific anticancer therapies. Such small molecule kinase inhibitors are currently either being tested in clinical trials or are already in use as cancer therapies[57–60].

In addition to displaying alterations in the mTOR signalling pathway, subtype-1 tumours also display evidence of elevated ion channel (Fig. 4A,D) and secretion pathway activities: a phenotype that is likely associated with increased trans-membrane transport of cell products (Fig. S2A). Changes in the expression patterns of ion channel proteins are also found in breast and prostate cancers[61,62]. In pancreatic cancers, ion channel proteins likely play crucial roles in cellular processes that are integral to oncogeneses such as cellular proliferation, motility, tissue invasion, and the excretion of lactic acid produced as a consequence of anaerobic respiration[63,64]. It is plausible therefore that subtype-1 tumours may be responsive to anti-cancer treatments that target ion channels and membrane pump proteins[63].

Subtype-2 tumours on the other hand display elevated peptidase activities (Fig. 4A,C). Peptidases regulate various proteins that play essential roles in regulatory signalling networks. As is presently the case for tumours of the kidney, peptidases may be useful as diagnostic and/or prognostic biomarkers of subtype-2 pancreatic cancers[65,66].

We found no significant differences in the mutational landscape between the two pancreatic cancer subtypes, indicating that the accumulation of similar genetic mutations drives the formation of tumours belonging to both subtypes. Recently, the paradigm of oncogenesis has been expanded beyond the classical view that oncogenesis is entirely driven by the accumulation of genetic mutations[67,68]. This paradigm now includes the disruption of epigenetic regulatory mechanisms and variations in miRNA expression[68–73]. Unlike with mutations, we currently lack adequate conceptual knowledge and the analytical framework needed to identifying putative driver and passenger changes in epigenetic and miRNA based regulatory processes[73–76].

Nevertheless, we observed several differences between subtype-1 and subtype-2 tumours with respect to epigenetic (DNA methylation profile) and miRNA signatures. These suggest that epigenetic and/or miRNA

variations may be primarily drivers of the differences in the transcriptome and proteome profiles of subtype-1 and substype-2 tumours.

In line with other studies that have identified biomarkers to classify tumour subtypes, some of which have important treatment and prognostic implications, we identified biomarker mRNA, DNA methylation, protein or miRNA sets that could be used to accurately subtype pancreatic tumours[8,9,48,77]. We are optimistic that any of these four biomarker sets could be individually used to obtain accurate subtype classifications for new pancreatic tumours. Nevertheless, the utility of these four biomarkers sets for predicting clinical outcomes and guiding treatment strategies will need to be evaluated in future studies.

Encouragingly, we were able to demonstrate that, by focusing on just the transcription levels of the mRNA molecules that are represented in our mRNA biomarker set, we could accurately predict the drug responses of cancer cell lines based on the drug responses of other cancer cell lines with similar mRNA expression profiles. Although others have also been able to predict the drug responses of cell lines using similar machine learning approaches[10,11,45,46,78], our approach is novel in that it utilizes tumour subtyping based on all available molecular data to mine for biomarkers that differentiate disease subtypes: biomarkers which are then used to inform our KNN exhaustive search model with respect to quantifying the similarity of cell lines. What this means is that our approach is capable of utilizing matched molecular data and drug responses from either cancer patients or cell lines to predict, with reasonable accuracy, the drug responses of tumours for which we have only information on the concentrations of the mRNAs, proteins or miRNAs that are included within the biomarker sets which we have identified. As with other machine learning based inference schemes, the accuracy of the predictions that are made should improve given additional matched molecular and drug response data[79].

Altogether, our analyses have revealed the molecular underpinnings of, and potential treatment strategies for, two clinically distinct forms of pancreatic cancer. We are optimistic that an approach such as we have used, where multiple different molecular data types are leveraged to subtype and characterise particular tumour variants, could yield valuable insights into the management of other difficult to treat cancers such as those of the lungs and triple negative breast cancer.

Methods

We analysed data from 185 of the pancreatic cancer patients who had contributed samples to the TCGA project[13]. Data on these patient samples within the TCGA included: reverse phase protein array-based proteomics data (RPPA; n = 45), whole exome sequencing data (n = 76), transcriptome data determined using RNAseq (n = 76); DNA copy number and mutation data (n = 76), miRNA data (n = 56), and comprehensive clinical data. For our analyses, we only considered the 76 "high purity" samples for which transcriptome and whole exome sequencing data was available. Out of these 76 samples only 45 have RPPA data. All data used in our analyses were obtained from cBioPortal (http://www.cbioportal.org)[80].

RPPA-based Classification of Pancreatic Cancer. K-means clustering of proteomic data was performed to identify subtypes of the 45 high purity TCGA pancreatic tumour datasets with available RPPA data[39]. To find the most informative number of clusters, K-means clustering was run over 500 iterations for cluster sizes (K values) of two, three, four, and five (i.e., K = 2 to 5). The average silhouette values for each value of K were compared, revealing that the two-cluster solution had the highest mean silhouette value and was therefore deemed to be the most coherent (Fig. S1B). To aid in visualizing the most informative features that differentiated between the two inferred tumour subtypes, the 112 proteins with the highest entropy values across samples were used to reproduce the two-cluster K-mean classification using semi-supervised hierarchical clustering (Fig. 1A)[81]. The clustering pattern thus obtained was visualised using a principal component analysis plot (Fig. S1A)[82]. The clustering of these 45 pancreatic cancer tumours based on protein, miRNA and DNA methylation data has been previously published by Raphael *et al.*[14], and the results of these clustering analyses were extracted from the supplementary file of that publication.

Integrative Subtyping of Pancreatic Cancer. Similarity Network Fusion (SNF) is a clustering method that considers information from multiple molecular profiles. It has previously been used to segregate tumours of various cancer types based on multiple different sources of molecular data[19]. Briefly, standard normalised protein, mRNA, miRNA, and DNA methylation data derived from the 45 high-purity samples were used to create patient similarity networks (Fig. 2B). Next, we ran SNF to fuse the similarity networks over 25 iterations, with hyperparameter settings of 24 and 0.7 for the number of neighbours and alpha value, respectively. Finally, spectral clustering with two specified as the best number of clusters (identified according to the eigengap) was applied to the unified similarity network to obtain the final tumour classification (Fig. 2C)[19].

Patient's clinical characteristics of the pancreatic cancer subtypes. The Kaplan-Meier method was used to compare overall survival and the duration of progression-free survival of patients with tumours belonging to the different pancreatic cancer subtypes[23].

Pathways and kinase enrichment analyses. The differentially expressed proteins between the pancreatic cancer subtypes were identified using the Student *t*-test with unequal variance and with the Benjamin-Hochberg correction applied to p-values[83,84]. Further, we queried Enrichr with two lists of 60 and 30 proteins found to be upregulated in subtype-1 and subtype-2 tumours, respectively, to return enriched KEGG pathways for each subtype (see Supplement File 1)[24,85]. The enriched KEGG pathways were compared to identify pathways that are unique to each of the disease subtypes[86]. The proteins that participate in pathways that are uniquely altered in sybtype-1 or subtype-2 tumours were used to construct protein-protein interaction networks using known interactions from each of the following databases: the University of California Santa Cruz Super pathway (101,525 protein-protein interactions), the Kinase Enrichment Analysis (428 kinases and their

10,792 targets), and Chromatin Immunoprecipitation Enrichment Analysis 2016 (667 transcription factors and their 464,967 targets)[31,87,88]. We visualised the resulting networks in yEd (Fig. 1B,C). Lastly, Kinase Enrichment Analysis was used to computationally identify the kinases that are responsible for the observed phosphorylation patterns in pancreatic cancer[31].

The moderated student *t*-test based on the negative binomial model was used to identify differentially expressed mRNAs and variations in DNA methylation patterns (see Supplementary File 2)[89,90]. Additionally, functional enrichment analyses were performed using lists of differentially expressed mRNA transcripts or altered DNA methylation patterns associated with each disease subtype. These were used to query Enrichr to return Gene Ontology (GO) molecular functions and KEGG pathways enriched for each disease subtype (Figs. 4A, S2A,B, Supplementary File 2). A custom MATLAB script was used to create an enrichment network based on the enriched GO-molecular function designations. This enrichment network was visualised in Cytoscape(Fig. 4A)[91].

Identification and evaluation of biomarker sets. We used various data mining and machine learning methods to identify biomarker sets of mRNAs, DNA methylation, miRNAs or proteins that individually and consistently best stratified the two pancreatic cancer subtypes. The diagonal adaption of neighbourhood component analysis (NCA) with regularisation method was used to select the most useful features for each molecular data type[38]. Briefly, NCA attached feature weights to each attribute where the feature weights are used to select the most important attributes for classification. For each molecular biomarker dataset identified using NCA, unsupervised hierarchical clustering was applied to the TCGA datasets to reproduce the two-subtype pancreatic cancer classification (Fig. 5A–D). To apply supervised machine learning methods that accurately predict the tumour subtypes while utilising only one molecular data type, 23 different machine learning classifiers were trained ranging from linear discriminate analysis, support vector machines, decision trees, logistic regression, ensemble trees, and K-nearest neighbour algorithms. Then, the best performing classifier for each molecular biomarker dataset was selected based on their 5-fold cross-validation accuracy and area under the receiver operating characteristic curve. The selected models were the cubic K-nearest neighbour for the mRNA biomarker set (98.7% accuracy), quadratic SVM for the DNA methylation biomarker set (97.8% accuracy), Ensemble bagged trees for the protein biomarker set (95.6%), and the course Gaussian SVM for the miRNA biomarker set (93.3% accuracy)[92].

To improve the accuracy of these models, the optimal hyperparameters that minimise the five-fold cross-validation loss were obtained using Bayesian hyperparameter optimisation (Fig. S3C,D)[93–95]. This improved the overall classification accuracy of the models on the cross-validation set to 100% for the mRNA-based KNN model and 99% for the DNA methylation-based SVM model. The trained models were then used to classify 31 other high-purity pancreatic tumours from the TCGA (Figs. 4F and 5E). Supervised learning models based on the proteomic or miRNA biomarkers datasets were not trained because there were too few other high purity samples profiled by TCGA for these data types. Further, for each molecular data biomarker set, between three and ten features were selected based on the lowest cross-validation loss of the best performing algorithm (Fig. S3E). These features were then used to classify TCGA pancreatic cancer samples using unsupervised hierarchical clustering (Figs. S3F–I).

Validating biomarker molecular datasets. To evaluate the performance of the biomarker mRNA on a different pancreatic cancer dataset, we downloaded pancreatic cancer data from the ICGC data portal[12]. From the initial 50 mRNA biomarker set identified using the TCGA dataset, only 45 had corresponding data in the ICGC mRNA dataset. Therefore, we extracted the 45 gene biomarker set from both the TCGA and ICGC data. The mRNA-based KNN model was then re-trained on the TCGA 45 mRNA biomarker set. Here, standard normalisation was applied as a pre-processing step both to avoid platform associated biases, and because it was previously performed on the data before SNF clustering. Thereafter, the TCGA mRNA-based KNN model was used to predict the subtype of tumours in the ICGC dataset using a standard normalised mRNA biomarker set that was extracted from the ICGC RNAseq data (Fig. 5G). Also, unsupervised hierarchical clustering was applied to the ICGC biomarker gene set (Fig. 5F). Finally, the mutational landscape and clinical characteristics of the two pancreatic cancer subtypes of both the ICGC and TCGA datasets were compared (Fig. 5I).

Subtype classification of cell lines. mRNA expression data from 45 pancreatic cancer cell lines together with their response profiles to 24 anticancer drugs were downloaded from the Cancer Cell Line Encyclopaedia[10]. The 50-mRNA biomarker set was extracted from the mRNA expression dataset and standard normalised. Then, the normalised CCLE mRNA biomarker genes were to subtype the cell lines by running the mRNA transcript levels for these genes through the mRNA-based KNN-model trained on TCGA data. The predicted subtypes of the CCLE cell lines were visualised using t-distributed stochastic neighbour embedding (t-SNE) (Fig. 6A).

Machine learning method to predict a cell line's drug response. An exhaustive nearest neighbour searcher model was created using standard normalised mRNA biomarker sets of both the CCLE cell lines and TCGA tumours[96]. The exhaustive searcher model takes as input the training data (in this case the mRNA biomarkers), distance metrics, and parameter values of the distance metrics for an exhaustive nearest neighbour search and can then be used to identify the nearest neighbours to a particular patient tumour or cell line within a specified radius of the distance matric. Here, the nearest neighbours to a particular cell line suggest similarity at the molecular level based on mRNA, DNA methylation, protein and miRNA data encoded in the SNF subtyping. The ten nearest neighbouring cell lines or tumours were determined using a nearest neighbour search algorithm based on a squared Euclidean distance metric (see Fig. 5B for intuition). After that, the drug response activity areas of the nearest neighbour cell lines were z-normalized and categorised as sensitive (for z-scored activity areas > 0.8), intermediate (for z-scored activity areas between 0.8 and -0.8), or resistant (for z-scored activity areas < -0.8). A simple prediction model was employed where the median responses to a particular drug of the nearest neighbouring cell lines was used to infer a target cell line's drug response (Fig. 6B,C). Following this the

quadratic Cohen's Kappa score was used to evaluate the goodness of fit between the predicted and the actual drug response profiles of the cell lines (Fig. 6D)[97].

Statistical analyses. All statistical analyses were performed in MATLAB 2018a except where stated otherwise. Fisher's exact tests were used to assess associations between categorical variables. Wilcoxon rank sum tests or independent sample Student *t*-tests were used for continuous variables where appropriate. Statistical tests were considered significant at $p < 0.05$ for single comparisons, and for Benjamini-Hochberg adjusted p-values < 0.05 for multiple comparisons.

Ethics approval. The University of Cape Town; Health Sciences Research Ethics Committee (HREC) IRB00001938 approved the protocol of this study.

Received: 28 May 2019; Accepted: 9 January 2020;
Published online: 27 January 2020

References

1. Isaji, S., Kawarada, Y. & Uemoto, S. Classification of pancreatic cancer: comparison of Japanese and UICC classifications. *Pancreas* **28**, 231–4 (2004).
2. Baylor, S. M. & Berg, J. W. Cross-classification and survival characteristics of 5,000 cases of cancer of the pancreas. *J. Surg. Oncol.* **5**, 335–58, https://doi.org/10.1002/jso.2930050410 (1973).
3. Cubilla, A. L. & Fitzgerald, P. J. Classification of pancreatic cancer (nonendocrine). *Mayo Clin. Proc.* **54**, 449–58 (1979).
4. Varadhachary, G. R. *et al.* Borderline Resectable Pancreatic Cancer: Definitions, Management, and Role of Preoperative Therapy. *Ann. Surg. Oncol.* **13**, 1035–46, https://doi.org/10.1245/ASO.2006.08.011 (2006).
5. Hidalgo, M. Pancreatic Cancer. *N. Engl. J. Med.* **362**, 1605–17, https://doi.org/10.1056/NEJMra0901557 (2010).
6. Biankin, A. V. *et al.* Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nat.* **491**, 399–405, https://doi.org/10.1038/nature11547 (2012).
7. Waddell, N. *et al.* Whole genomes redefine the mutational landscape of pancreatic cancer. *Nat.* **518**, 495–501, https://doi.org/10.1038/nature14169 (2015).
8. Bailey, P. *et al.* Genomic analyses identify molecular subtypes of pancreatic cancer. *Nat.* **531**, 47–52, https://doi.org/10.1038/nature16965 (2016).
9. Moffitt, R. A. *et al.* Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat. Genet.* **47**, 1168–78, https://doi.org/10.1038/ng.3398 (2015).
10. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nat.* **483**, 603–7, https://doi.org/10.1038/nature11003 (2012).
11. Menden, M. P. *et al.* Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. *PLoS One* **8**, e61318, https://doi.org/10.1371/journal.pone.0061318 (2013).
12. Zhang, J. *et al.* International Cancer Genome Consortium Data Portal–a one-stop shop for cancer genomics data. Database (Oxford) 2011, bar026. https://doi.org/10.1093/database/bar026 (2011).
13. Chang, K. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–20, https://doi.org/10.1038/ng.2764 (2013).
14. Cancer Genome Atlas Research Network. Electronic address: andrew_aguirre@dfci.harvard.edu TCGAR, Cancer Genome Atlas Research Network. Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. *Cancer Cell* **32**, 185–203.e13, https://doi.org/10.1016/j.ccell.2017.07.007 (2017).
15. Sinkala, M., Mulder, N. & Martin, D. P. Integrative landscape of dysregulated signaling pathways of clinically distinct pancreatic cancer subtypes. *Oncotarget* **9**, 29123–39, https://doi.org/10.18632/oncotarget.25632 (2018).
16. Dreyer, S. B., Chang, D. K., Bailey, P. & Biankin, A. V. Pancreatic Cancer Genomes: Implications for Clinical Management and Therapeutic Development. *Clin. Cancer Res.* **23**, 1638–46, https://doi.org/10.1158/1078-0432.CCR-16-2411 (2017).
17. Costello, E., Greenhalf, W. & Neoptolemos, J. P. New biomarkers and targets in pancreatic cancer and their application to treatment. *Nat. Rev. Gastroenterol. Hepatol.* **9**, 435–44, https://doi.org/10.1038/nrgastro.2012.119 (2012).
18. Bournet, B. *et al.* KRAS G12D Mutation Subtype Is A Prognostic Factor for Advanced Pancreatic Adenocarcinoma. *Clin. Transl. Gastroenterol.* **7**, e157, https://doi.org/10.1038/ctg.2016.18 (2016).
19. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **11**, 333–7, https://doi.org/10.1038/nmeth.2810 (2014).
20. Bauer, D. C. *et al.* Genomics and personalised whole-of-life healthcare. *Trends Mol. Med.* **20**, 479–86, https://doi.org/10.1016/J.MOLMED.2014.04.001 (2014).
21. Keogh, E. & Mueen, A. Curse of Dimensionality. Encycl. Mach. Learn. Data Min., Boston, MA: Springer US; p. 314–5, https://doi.org/10.1007/978-1-4899-7687-1_192 (2017).
22. ACM Special Interest Group for Algorithms and Computation Theory. D, SIAM Activity Group on Discrete Mathematics. S, Association for Computing Machinery., Society for Industrial and Applied Mathematics. Proceedings of the eighteenth annual ACM-SIAM Symposium on Discrete Algorithms. Association for Computing Machinery; (2007).
23. Goel, M. K., Khanna, P. & Kishore, J. Understanding survival analysis: Kaplan-Meier estimate. *Int. J. Ayurveda Res.* **1**, 274–8, https://doi.org/10.4103/0974-7788.76794 (2010).
24. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinforma.* **14**, 128, https://doi.org/10.1186/1471-2105-14-128 (2013).
25. Ishimura, N. *et al.* BRAF and K-ras gene mutations in human pancreatic cancers. *Cancer Lett.* **199**, 169–73, https://doi.org/10.1016/S0304-3835(03)00384-7 (2003).
26. Heidorn, S. J. *et al.* Kinase-Dead BRAF and Oncogenic RAS Cooperate to Drive Tumor Progression through CRAF. *Cell* **140**, 209–21, https://doi.org/10.1016/J.CELL.2009.12.040 (2010).
27. Testa, J. R. & Bellacosa, A. AKT plays a central role in tumorigenesis. *Proc. Natl Acad. Sci. USA* **98**, 10983–5, https://doi.org/10.1073/pnas.211430998 (2001).
28. Liu, Y., Sun, J. & Zhao, M. ONGene: A literature-based database for human oncogenes. *J. Genet. Genomics* **44**, 119–21, https://doi.org/10.1016/J.JGG.2016.12.004 (2017).
29. Eyhp, L. & Muller, W. J. Oncogenes and tumor suppressor genes. *Cold Spring Harb. Perspect. Biol.* **2**, a003236, https://doi.org/10.1101/cshperspect.a003236 (2010).
30. de Leon, M. P. Oncogenes and Tumor Suppressor Genes, Springer, Berlin, Heidelberg; p. 35–47, https://doi.org/10.1007/978-3-642-85076-9_4 (1994).
31. Lachmann, A. & Ma'ayan, A. KEA: kinase enrichment analysis. *Bioinforma.* **25**, 684–6, https://doi.org/10.1093/bioinformatics/btp026 (2009).
32. Schmid, K. *et al.* Dual inhibition of EGFR and mTOR pathways in small cell lung cancer. *Br. J. Cancer* **103**, 622–8, https://doi.org/10.1038/sj.bjc.6605761 (2010).

33. Zarogoulidis, P. *et al*. mTOR pathway: A current, up-to-date mini-review (Review). *Oncol. Lett.* **8**, 2367–70, https://doi.org/10.3892/ol.2014.2608 (2014).
34. Harashima, H., Dissmeyer, N. & Schnittger, A. Cell cycle control across the eukaryotic kingdom. *Trends Cell Biol.* **23**, 345–56, https://doi.org/10.1016/J.TCB.2013.03.002 (2013).
35. Frizzell, R. A. & Hanrahan, J. W. Physiology of epithelial chloride and fluid secretion. *Cold Spring Harb. Perspect. Med.* **2**, a009563, https://doi.org/10.1101/cshperspect.a009563 (2012).
36. Kang, R. *et al*. The receptor for advanced glycation end products (RAGE) sustains autophagy and limits apoptosis, promoting pancreatic tumor cell survival. *Cell Death Differ.* **17**, 666–76, https://doi.org/10.1038/cdd.2009.149 (2010).
37. Abe, R. & Yamagishi, S. AGE-RAGE System and Carcinogenesis. *Curr. Pharm. Des.* **14**, 940–5, https://doi.org/10.2174/138161208784139765 (2008).
38. Yang, W., Wang, K. & Zuo, W. Neighborhood Component Feature Selection for High-Dimensional. *Data*. https://doi.org/10.4304/jcp.7.1.161-168 (2012).
39. Wu, Y., Ianakiev, K. & Govindaraju, V. Improved k-nearest neighbor classification. *Pattern Recognit.* **35**, 2311–8, https://doi.org/10.1016/S0031-3203(01)00132-7 (2002).
40. Kecman, V., Huang, T.-M. & Vogt, M. Iterative Single Data Algorithm for Training Kernel Machines from Huge Data Sets: Theory and Performance, Springer, Berlin, Heidelberg; p. 255–74, https://doi.org/10.1007/10984697_12 (2005).
41. Platt, J. C. & Platt, J. C. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Adv LARGE MARGIN Classif: 61–74 (1999).
42. Marusyk, A., Almendro, V. & Polyak, K. Intra-tumour heterogeneity: a looking glass for cancer? *Nat. Rev. Cancer* **12**, 323–34, https://doi.org/10.1038/nrc3261 (2012).
43. Witkiewicz, A. K. *et al*. Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. *Nat. Commun.* **6**, 6744, https://doi.org/10.1038/ncomms7744 (2015).
44. Dhar, S. *et al*. Anti-cancer drug characterisation using a human cell line panel representing defined types of drug resistance. *Br. J. Cancer* **74**, 888–96, https://doi.org/10.1038/bjc.1996.453 (1996).
45. Bansal, M. *et al*. A community computational challenge to predict the activity of pairs of compounds. *Nat. Biotechnol.* **32**, 1213–22, https://doi.org/10.1038/nbt.3052 (2014).
46. Geeleher, P., Cox, N. J. & Huang, R. Clinical drug response can be predicted using baseline gene expression levels and *in vitro* drug sensitivity in cell lines. *Genome Biol.* **15**, R47, https://doi.org/10.1186/gb-2014-15-3-r47 (2014).
47. Friedman, J. H., Bentley, J. L. & Finkel, R. A. An algorithm for finding best matches in logarithmic expected time (1975).
48. Collisson, E. A. *et al*. Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nat. Med.* **17**, 500–3, https://doi.org/10.1038/nm.2344 (2011).
49. Mohammed, A. *et al*. Antidiabetic Drug Metformin Prevents Progression of Pancreatic Cancer by Targeting in Part Cancer Stem Cells and mTOR Signaling. *Transl. Oncol.* **6**, 649–IN7, https://doi.org/10.1593/TLO.13556 (2013).
50. Jiao, Y. *et al*. DAXX/ATRX, MEN1, and mTOR Pathway Genes Are Frequently Altered in Pancreatic Neuroendocrine Tumors. *Sci.* **331**, 1199–203, https://doi.org/10.1126/SCIENCE.1200609 (2011).
51. Morran, D. C. *et al*. Targeting mTOR dependency in pancreatic cancer. *Gut* **63**, 1481–9, https://doi.org/10.1136/gutjnl-2013-306202 (2014).
52. Soares, H. P. *et al*. Dual PI3K/mTOR Inhibitors Induce Rapid Overactivation of the MEK/ERK Pathway in Human Pancreatic Cancer Cells through Suppression of mTORC2. *Mol. Cancer Ther.* **14**, 1014–23, https://doi.org/10.1158/1535-7163.MCT-14-0669 (2015).
53. Ning, C. *et al*. Targeting ERK enhances the cytotoxic effect of the novel PI3K and mTOR dual inhibitor VS-5584 in preclinical models of pancreatic cancer. *Oncotarget* **8**, 44295–311, https://doi.org/10.18632/oncotarget.17869 (2017).
54. Loddo, M. *et al*. Cell-cycle-phase progression analysis identifies unique phenotypes of major prognostic and predictive significance in breast cancer. *Br. J. Cancer* **100**, 959–70, https://doi.org/10.1038/sj.bjc.6604924 (2009).
55. Teodoro, A. *et al*. Effect of lycopene on cell viability and cell cycle progression in human cancer cell lines. *Cancer Cell Int.* **12**, 36, https://doi.org/10.1186/1475-2867-12-36 (2012).
56. Williams, G. H. & Stoeber, K. The cell cycle and cancer. *J. Pathol.* **226**, 352–64, https://doi.org/10.1002/path.3022 (2012).
57. Diaz-Moralli, S., Tarrado-Castellarnau, M., Miranda, A. & Cascante, M. Targeting cell cycle regulation in cancer therapy. *Pharmacol. Ther.* **138**, 255–71, https://doi.org/10.1016/J.PHARMTHERA.2013.01.011 (2013).
58. Dickson, M. A. Molecular pathways: CDK4 inhibitors for cancer therapy. *Clin. Cancer Res.* **20**, 3379–83, https://doi.org/10.1158/1078-0432.CCR-13-1551 (2014).
59. McCubrey, J. A. *et al*. GSK-3 as potential target for therapeutic intervention in cancer. *Oncotarget* **5**, 2881–911, https://doi.org/10.18632/oncotarget.2037 (2014).
60. Madhok, B. M., Yeluri, S., Perry, S. L., Hughes, T. A. & Jayne, D. G. Dichloroacetate induces apoptosis and cell-cycle arrest in colorectal cancer cells. *Br. J. Cancer* **102**, 1746–52, https://doi.org/10.1038/sj.bjc.6605701 (2010).
61. Fraser, S. P. *et al*. Voltage-Gated Sodium Channel Expression and Potentiation of Human Breast Cancer Metastasis. *Clin. Cancer Res.* **11**, 5381–9, https://doi.org/10.1158/1078-0432.CCR-05-0327 (2005).
62. Furuya, Y., Lundmo, P., Short, A. D., Gill, D. L. & Isaacs, J. T. The role of calcium, pH, and cell proliferation in the programmed (apoptotic) death of androgen-independent prostatic cancer cells induced by thapsigargin. *Cancer Res.* **54**, 6167–75, https://doi.org/10.1158/0008-5472.can-04-2146 (1994).
63. Pedersen, S. F. & Stock, C. Ion Channels and Transporters in Cancer: Pathophysiology, Regulation, and Clinical Potential. *Cancer Res.* **73**, 1658–61, https://doi.org/10.1158/0008-5472.CAN-12-4188 (2013).
64. Monteith, G. R., Davis, F. M. & Roberts-Thomson, S. J. Calcium channels and pumps in cancer: changes and consequences. *J. Biol. Chem.* **287**, 31666–73, https://doi.org/10.1074/jbc.R112.343061 (2012).
65. Varona, A. *et al*. Altered levels of acid, basic, and neutral peptidase activity and expression in human clear cell renal cell carcinoma. *Am. J. Physiol. Physiol* **292**, F780–8, https://doi.org/10.1152/ajprenal.00148.2006 (2007).
66. Larrinaga, G. *et al*. The impact of peptidase activity on clear cell renal cell carcinoma survival. *Am. J. Physiol. Physiol* **303**, F1584–91, https://doi.org/10.1152/ajprenal.00477.2012 (2012).
67. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. vol. 144. Elsevier, https://doi.org/10.1016/j.cell.2011.02.013 (2011).
68. Duesberg, P. *et al*. Genetic instability of cancer cells is proportional to their degree of aneuploidy. *Proc. Natl Acad. Sci.* **95**, 13692–7, https://doi.org/10.1073/pnas.95.23.13692 (1998).
69. Coyle, K. M., Boudreau, J. E. & Marcato, P. Genetic Mutations and Epigenetic Modifications: Driving Cancer and Informing Precision Medicine. *Biomed. Res. Int.* **2017**, 9620870, https://doi.org/10.1155/2017/9620870 (2017).
70. Sharma, S., Kelly, T. K. & Jones, P. A. Epigenetics in cancer. *Carcinogenesis* **31**, 27–36, https://doi.org/10.1093/carcin/bgp220 (2010).
71. Reddy, K. B. MicroRNA (miRNA) in cancer. *Cancer Cell Int.* **15**, 38, https://doi.org/10.1186/s12935-015-0185-1 (2015).
72. Mishra, N. K. & Guda, C. Genome-wide DNA methylation analysis reveals molecular subtypes of pancreatic cancer. *Oncotarget* **8**, 28990–9012, https://doi.org/10.18632/oncotarget.15993 (2017).
73. Khatri, I. *et al*. Systems Biology Approach to Identify Novel Genomic Determinants for Pancreatic Cancer Pathogenesis. *Sci. Rep.* **9**, 123, https://doi.org/10.1038/s41598-018-36328-w (2019).
74. Kazanets, A., Shorstova, T., Hilmi, K., Marques, M. & Witcher, M. Epigenetic silencing of tumor suppressor genes: Paradigms, puzzles, and potential. *Biochim. Biophys. Acta - Rev. Cancer* **1865**, 275–88, https://doi.org/10.1016/J.BBCAN.2016.04.001 (2016).

75. Chatterjee, A., Rodger, E. J. & Eccles, M. R. Epigenetic drivers of tumourigenesis and cancer metastasis. *Semin. Cancer Biol.* **51**, 149–59, https://doi.org/10.1016/J.SEMCANCER.2017.08.004 (2018).
76. Shen, H. & Laird, P. W. Interplay between the Cancer Genome and Epigenome. *Cell*. **153**, 38–55, https://doi.org/10.1016/J.CELL.2013.03.008 (2013).
77. Prat, A., Parker, J. S., Fan, C. & Perou, C. M. PAM50 assay and the three-gene model for identifying the major and clinically relevant molecular subtypes of breast cancer. *Breast Cancer Res. Treat.* **135**, 301–6, https://doi.org/10.1007/s10549-012-2143-0 (2012).
78. Volm, M. & Efferth, T. Prediction of Cancer Drug Resistance and Implications for Personalized Medicine. *Front. Oncol.* **5**, 282, https://doi.org/10.3389/fonc.2015.00282 (2015).
79. Chu, C., Hsu, A.-L., Chou, K.-H., Bandettini, P. & Lin, C. Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *Neuroimage* **60**, 59–70, https://doi.org/10.1016/J.NEUROIMAGE.2011.11.066 (2012).
80. Cerami, E. *et al*. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–4, https://doi.org/10.1158/2159-8290.CD-12-0095 (2012).
81. Hastie, T & Tibshirani, R, Friedman, J. Unsupervised Learning, Springer, New York, NY; p. 485–585, https://doi.org/10.1007/978-0-387-84858-7_14 (2009).
82. Jolliffe, I. Principal Component Analysis. *Int. Encycl. Stat. Sci.*, Berlin, Heidelberg: Springer Berlin Heidelberg; p. 1094–6, https://doi.org/10.1007/978-3-642-04898-2_455 (2011).
83. Benjamini, Y. Discovering the false discovery rate. *J. R. Stat. Soc. Ser. B (Stat. Methodol).* **72**, 405–16, https://doi.org/10.1111/j.1467-9868.2010.00746.x (2010).
84. Lakens, D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Front. Psychol.* **4**, 863, https://doi.org/10.3389/fpsyg.2013.00863 (2013).
85. Gene Ontology Consortium: going forward. *Nucleic. Acids. Res.* **43**, D1049–56, https://doi.org/10.1093/nar/gku1179 (2015).
86. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–61, https://doi.org/10.1093/nar/gkw1092 (2017).
87. Lachmann, A. *et al*. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinforma.* **26**, 2438–44, https://doi.org/10.1093/bioinformatics/btq466 (2010).
88. Wong, C. K. *et al*. The UCSC Interaction Browser: multidimensional data views in pathway context. *Nucleic Acids Res.* **41**, W218–24, https://doi.org/10.1093/nar/gkt473 (2013).
89. Brooks, A. N. *et al*. Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Res.* **21**, 193–202, https://doi.org/10.1101/gr.108662.110 (2011).
90. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–8, https://doi.org/10.1038/nmeth.1226 (2008).
91. MathWorks, T. MATLAB (R2017b). MathWorks Inc 2017. https://doi.org/10.1007/s10766-008-0082-5.
92. Harris, E. K. & Boyd, J. C. On dividing reference data into subgroups to produce separate reference ranges. *Clin. Chem.* **36**, 265–70 (1990).
93. Research AB-J of ML, undefined. Convergence rates of efficient global optimization algorithms. JmlrOrg n.d. (2011).
94. Snoek, J., Larochelle, H. & Adams, R. P. Practical Bayesian Optimization of Machine Learning Algorithms: 2951–9 (2012).
95. Gelbart, M. A., Snoek, J. & Adams, R. P. Bayesian Optimization with Unknown Constraints (2014).
96. Friedman, J. H., Bentley, J. L. & Finkel, R. A. An algorithm for finding best matches in logarithmic expected Time. (1975).
97. Ben-David, A. Comparison of classification accuracy using Cohen's Weighted Kappa. *Expert. Syst. Appl.* **34**, 825–32, https://doi.org/10.1016/J.ESWA.2006.10.022 (2008).

Acknowledgements

Student bursary funding for this project was provided by H3ABioNet, supported by the National Institutes of Health Common Fund under grant number U24HG006941. The content of this publication is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

Conceptualisation, M.S., N.M. and D.P.M.; Methodology, M.S., N.M. and D.P.M.; Formal Analysis, M.S. and D.P.M.; Writing – Original Draft, M.S. and D.P.M.; Writing – Review & Editing; M.S., N.M. and D.P.M.; Visualisation, M.S.; Supervision, N.M. and D.P.M.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-58290-2.

Correspondence and requests for materials should be addressed to M.S.

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.