7900–7907 Nucleic Acids Research, 2011, Vol. 39, No. 18 Published online 30 June 2011
doi:10.1093/nar/gkr533

A global insight into a cancer transcriptional space using pancreatic data: importance, findings and flaws

Emanuela Gadaleta[1], Rosalind J. Cutts[1], Gavin P. Kelly[2], Tatjana Crnogorac-Jurcevic[1], Hemant M. Kocher[3], Nicholas R. Lemoine[1] and Claude Chelala[1,*]

[1]Centre for Molecular Oncology, Barts Cancer Institute, Queen Mary University of London, London EC1M 6BQ, [2]Bioinformatics and Statistics, Cancer Research UK London Research Institute, London WC2A 3PX and [3]Centre for Tumour Biology, Barts Cancer Institute, Queen Mary University of London, London EC1M 6BQ, UK

Received March 14, 2011; Revised May 17, 2011; Accepted June 11, 2011

ABSTRACT

Despite the increasing wealth of available data, the structure of cancer transcriptional space remains largely unknown. Analysis of this space would provide novel insights into the complexity of cancer, assess relative implications in complex biological processes and responses, evaluate the effectiveness of cancer models and help uncover vital facets of cancer biology not apparent from current small-scale studies. We conducted a comprehensive analysis of pancreatic cancer-expression space by integrating data from otherwise disparate studies. We found (i) a clear separation of profiles based on experimental type, with patient tissue samples, cell lines and xenograft models forming distinct groups; (ii) three subgroups within the normal samples adjacent to cancer showing disruptions to biofunctions previously linked to cancer; and (iii) that ectopic subcutaneous xenografts and cell line models do not effectively represent changes occurring in pancreatic cancer. All findings are available from our online resource for independent interrogation. Currently, the most comprehensive analysis of pancreatic cancer to date, our study primarily serves to highlight limitations inherent with a lack of raw data availability, insufficient clinical/histopathological information and ambiguous data processing. It stresses the importance of a global-systems approach to assess and maximise findings from expression profiling of malignant and non-malignant diseases.

INTRODUCTION

Every cancer reflects the highly heterogeneous make-up of the patient's genes, the stochastic mutational processes occurring within the tumour; the balance between these processes ultimately determining each tumour's unique profile (1,2). These inter-individual differences are evident in the variability of patient outcome. Under intensive investigation for the past decade, a wealth of information is now available on transcriptional regulation differences in tumourigenesis across multiple histological cancer types, cultured cells and xenograft models. Although critically needed to maximise cancer research, the interconnections between the molecular events that govern this transcriptional space in cancer are still largely unknown since most studies focus on molecular profiling a single sample type or condition. The same is true for non-malignant diseases where the amount of transcriptomic data obtained by molecular profiling a wide range of tissues and cells affected by the disease has grown exponentially.

A global integrated analysis of gene expression data generated by many different laboratories will reveal the overall structure of gene-expression space while enhancing the sensitivity of the analysis, by yielding improved statistical power and novel biological insight. This helps confidently assess sources of variability in the data and remove the poor-quality arrays that could compromise the statistical and biological significance of the original study (3).

There are two general approaches to comparative profiling analysis. The first method requires normalising and re-analysing the original data from each individual study. A limited number of meta-analyses tend to rely on this rigorous and reliable method because of problems associated with cross-platform analyses and most importantly

*To whom correspondence should be addressed. Tel: +44 (0)20 7882 3570; Fax: +44 (0)20 7882 3884; Email: c.chelala@qmul.ac.uk

the availability of both raw data and clinical information. The second approach is based on the assumption that essential genes will be consistently altered and relies on the identification of intersections between studies. While independent from the availability of raw data, this type of meta-analysis depends heavily on the pre-processing and analysis methods, the significance threshold and the annotation builds used in the original publication, not all of which are always accurate and reproducible (4).

A recent study has used the first method to produce a global map of human gene expression derived from the extensive analysis of 14 500 human genes across 5372 samples representing the structure of the expression space of 369 different cell and tissue types, disease states and cell lines (5). The authors showed that the major patterns are attributable to the tissue of origin independent of the disease state. Additional comparative analysis also determined that global patterns of tissue-specific expression of orthologous genes are conserved between human and mouse (6). The impact of such analyses on cancer research is hampered not only by the uniqueness and complexity of each cancer type but also by the quality and magnitude of raw and clinical annotations of cancer samples interrogated. In an attempt to assess and expand this global map to cancer research, we have analysed pancreatic cancer-specific expression data.

Pancreatic cancer is a major health problem and the fourth most common cause of cancer-related death world wide, with survival statistics relatively unchanged over the past 30 years (7). A multitude of studies have been dedicated to elucidating the pathogenesis of this disease, resulting in the generation and publication of an increasing volume of transcriptomics data (8). Hence, there is no shortage of pancreatic cancer raw data but an urgent need for robust and rigorous data analysis for establishing whether the results are valid and accurate. Unfortunately, relative to other malignancies, this is still somewhat under-investigated in the field of pancreatic cancer.

By enabling essential data integration and analysis, this is the first study that will allow the international community to assess and exploit the high volume of raw pancreatic-cancer data to maximal advantage. As the amount of cancer data continues to grow, this study is needed to evaluate the quality of the data generated and address the impact of molecular targets on cancer development, progression and resistance to treatment. Importantly, public repositories do not assess the quality of the deposited datasets or provide information on the data processing used. Therefore, one still has to carefully review the experimental and clinical annotations associated with the deposited raw data files.

This study overcomes these challenges to permit in-depth mining of different types of pancreatic cancer data. Our group has a lead in this development by having successfully established the Pancreatic Expression Database (PED) to address these challenges in published literature (9,10). PED is unique since it is the only tool currently available for mining of curated pancreatic cancer-literature data. This project is complementary but distinct from PED. It aims to perform an upstream analysis starting from the raw data in order to evaluate its quality and perform an in-depth data analysis of the structural space of pancreatic cancer expression data. All our findings will be linked to the PED portal making it a one-stop solution for the integration, analysis and visualisation of raw and literature-derived pancreatic cancer data.

MATERIALS AND METHODS

Data collection

A total of 309 raw pancreatic cancer expression data files, generated on the highly comprehensive Affymetrix GeneChip® Human Genome U133 Plus 2.0 array (~47 000 transcripts) were obtained from the Gene Expression Omnibus (GEO) (11), Array Express (12) and the expO project (https://expo.intgen.org/geo/home.do). When the data were available from GEO and ArrayExpress, the concordance was ascertained. Additional normal pancreas sample data were obtained from GEO and Affymetrix (www.affymetrix.com).

Data pre-processing

All processing was performed using Bioconductor packages (www.bioconductor.org) within the *R* statistical environment (www.r-project.org). The arrayMvout package was utilised to implement stringent quality control criteria to the datasets. Nine quantitative features of array quality, computed on each array, were evaluated: ABG (average background); SF (scale factor); PP (percent of present calls); AR (actin 3′/5′ratio); GR (GAPDH 3′/5′ ratio); median normalised unscaled standard error (NUSE); median relative log expression (RLE); RLE-IQR (interquartile range of IQR per array, to measure variability in RLE) and RNAS (slope of RNA degradation measure). These components were subsequently analysed using principal components 1, 2 and 3 analysis, followed by parametric multivariate outlier detection with calibration for multiple testing. A false positive rate of 0.01 for outlier detection, adjusting for multiple comparisons according to Caroni and Prescott's adaptation of Rosner was used (13,14). All data files passing the quality control checks were subsequently normalised jointly to create a global space of gene expression in pancreatic cancer.

Data normalisation and dimensionality reduction

The primary data from all studies were normalised using the GC-RMA hybrid algorithm with the GCRMA package. This normalising method was selected because, while being based on the Robust Multi-array Average (RMA) algorithm, it also accounts for GC-content background correction. Furthermore, GC-RMA has recently been recommended for the detection of differentially expressed genes (15,16). The probe sets were filtered using standard deviation calls to isolate a list of the top 10 000 most variable probe sets across all experiments.

Differential expression analysis

Genes differentially regulated between biological groups were identified using limma (17). We applied an

unbiased linear model and estimated the variance across the whole panel of samples so all comparisons have the same power. We used the duplicate correlation method from limma to adjust for replicates. The Benjamini and Hochberg (BH) false discovery rate (FDR) was used for multiple testing corrections. A double cut-off of FDR ≤ 0.01 and a fold change of ≥ 2 were set for the discovery analysis. Further, more stringent, thresholds of FDR ≤ 0.005 and fold changes of ≥ 2, 3 and 4 were also used to narrow down the list of deregulated genes and highlight the most affected pathways. We validated the differential pattern of expression for the biological comparisons taking into account within-group correlation, by treating study group as a random effect in a mixed effects model. Subsequently, Venn diagrams with four groups were generated by using an extension on the code from the limma package, available from http://bioinfo-mite.crb.wsu.edu/Rcode/Venn.R.

Principle component analysis and hierarchical clustering

PCA, using the R stats package, was applied, thereby reducing the dimensionality of the data to identify key components of variability. The R package pcaMethods was used to evaluate the cross-validation of our PCA model with the explained variance coefficients Q^2 and R^2, respectively. Moreover, as an internal validation, we have randomly removed 10% of the data to ensure the cluster tree was not affected. The cluster Bioconductor package was used for extended analysis, providing an increased insight of interactions via subgrouping of normal-adjacent expression data. Using the A2R package, unsupervised clustering, based on the Euclidean distance matrix and the average linkage algorithm, was performed.

Pathway analysis

Ingenuity Pathway Analysis (www.ingenuity.com), version 8.5, was used as a means to interrogate the data and identify various pathways and biofunctions deregulated within the different samples.

Accessibility of the data

An independent online resource has been constructed and implemented for the exploration of our findings. This user-friendly interface will enable researchers to query for a given gene(s) or probe(s); biological function(s) and canonical pathway(s); and then visualise the expression and differential expression levels across each of the sample types.

RESULTS

Sample selection

We collected pancreatic-related expression data files from public repositories. These included 309 samples (18–28) comprising of: four normal pancreas (commercial source); 55 normal-adjacent pancreas (pancreatic cancer patients); 96 pancreatic cancer [pancreatic cancer patients, representing 91 pancreatic ductal adenocarcinoma (PDAC) and five other pancreatic cancer types]; 65 pancreatic cancer cell lines (29 distinct cell lines); 24 saliva specimens (12 healthy subjects and 12 pancreatic cancer patients); 59 ectopic subcutaneous xenografts (pancreatic cancer patients) and six orthotopic xenografts (pancreatic cancer cell lines) (Supplementary Table S1). A total of 51 arrays, identified as outliers in quality control checks, were excluded from additional analyses. These included all the profiles representing saliva, three pancreatic cancer cell lines (Capan1, HPAC, SW1990), orthotopic xenografts from L3.6pl cell line and non-functional islet cell tumour (Supplementary Table S1). In addition, one of the raw expression files was corrupted (23) and hence also removed. Normalisation and statistical analyses were performed on the remaining data files.

Initial clustering analysis

PCA generated to view the underlying structure of the data showed two axes; microenvironment and malignancy, with patient tissue samples, xenograft models and cultured cell lines located in separate areas of the first empirical principal component and samples dispersed on the second empirical principal component based on their tumour content (Figure 1, Supplementary Figures S1 and S2). We found that the first two principal components explain ~42% of variability and have biological interpretations. Cross-validation showed that we do not appear to be overfitting the data by using two components. Unsupervised HCL yielded similar results, highlighting two predominant groups (Supplementary Figure S3). The first group comprised normal and cancer samples; and the second group consisted of xenograft models and cell lines. In accordance with the PCA, the clustering profiles from normal pancreatic tissue resected adjacent to cancer were found interspersed between normal donors (ND) and pancreatic cancer samples. Prior to conducting further analysis, PCA of sample origin was conducted to ascertain the premise that the biological effects were stronger than any laboratory influences (Supplementary Figure S4). Since the study groups are effectively nested within the microenvironment factor, estimating the relative effect sizes is not achievable. However, the clustering of the samples, resulting in study groups targeting the same microenvironments tending to cluster together, and then within samples from the same study group clustering, indicates that the biological factors are the primary effect. Further inspection of average similarities between samples from different studies but within the same biological group and samples from the same study but representing different biological groups showed that the biological effects were significantly stronger than the study effects ($P < 2.2^{-16}$) (Supplementary Figure S5). For biological groups to which only one study contributed, we cannot distinguish directly between the study and biological effects.

Biological interpretations from in-depth statistical analyses

Closer inspection of profiles from histologically 'normal-appearing' tissue adjacent to pancreatic cancer indicated the presence of three distinct subgroups (NAD1, NAD2 and NAD3) (Figure 2 and Supplementary Figure S6).

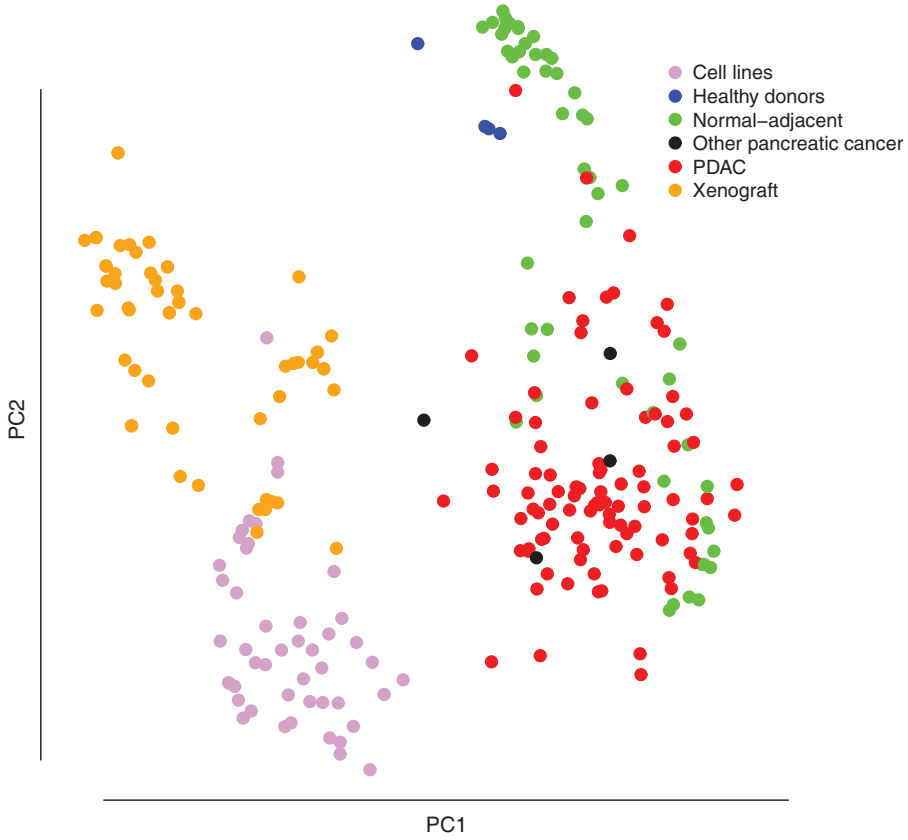

Figure 1. PCA plot of the global structure of pancreatic cancer-transcriptional space. The two main principle components can be visualised in this PCA. Each point represents the orientation of a single sample in a multi-dimensional gene expression space projected on the PCA, with different colours illustrating the biological group to which each sample belongs. The samples can be divided into three areas: xenograft models (yellow); cell-line models (purple) and human-derived profiles (green, red, black and blue).

The first subgroup comprises of ND and the normal-adjacent samples whose profiles most closely resemble those of ND. Sample profiles in the second group (NAD2) are sufficiently different to NAD1 to form a distinct subgroup. The third subgroup (NAD3) reflects the expression profiles of samples most closely resembling those of PDAC. Additional analyses were implemented in an attempt to characterise this heterogeneity.

To this end, the PDAC group and each normal-adjacent subgroup were initially compared against normal tissues from healthy donors. By making this direct comparison against normal pancreatic tissues, we took advantage of a 'subtractive' effect, which emphasised genes that consistently distinguished normal, normal-adjacent and cancer tissues. Subsequently, the common and unique genes identified as deregulated in PDAC and each normal-adjacent subgroup were assessed (Supplementary Figure S7). This circumvents issues with a lack of statistical power in other conventional disparate-study analyses with lower sample numbers, to identify a set of genes that define a molecular signature capable of distinguishing benign tissue resected adjacent to cancer from malignant tissue.

The top genes differentially expressed in each of the normal-adjacent subgroups and PDAC were selected and pathway analyses conducted. Three predominant groups of biofunctions were detected profoundly affected between all the normal subgroups and PDAC (Supplementary Figure S8). These pathways include: cancer-specific bioprocesses, such as deregulation in the apoptotic pathways, aberrant cellular growth, proliferation and development and disruption to cell cycle regulation; moieties involved in the inflammatory response pathways and other immunological responses; and an array of metabolic pathways, predominantly those involved in lipid metabolism, metabolic disease and vitamin and mineral metabolism. Each of these biofunctions have previously been linked to tumourigenesis (29–32)

Further comparisons of the signalling pathways and processes found altered in the 'normal-appearing' samples to those previously reported to be altered in pancreatic cancer generated an unexpected amount of overlap (Supplementary Table S2) (31). This suggests that varying similarities exist in the pathways deregulated between each of the normal-adjacent subgroups and PDAC.

To ensure a thorough examination of all the pancreatic-cancer data available, attempts were made to investigate whether genetic alterations associated with pancreatic cancer were preserved and reflected in a similar manner by the xenotransplanted samples and the cell lines. Since the aim of this study was to use available raw data to

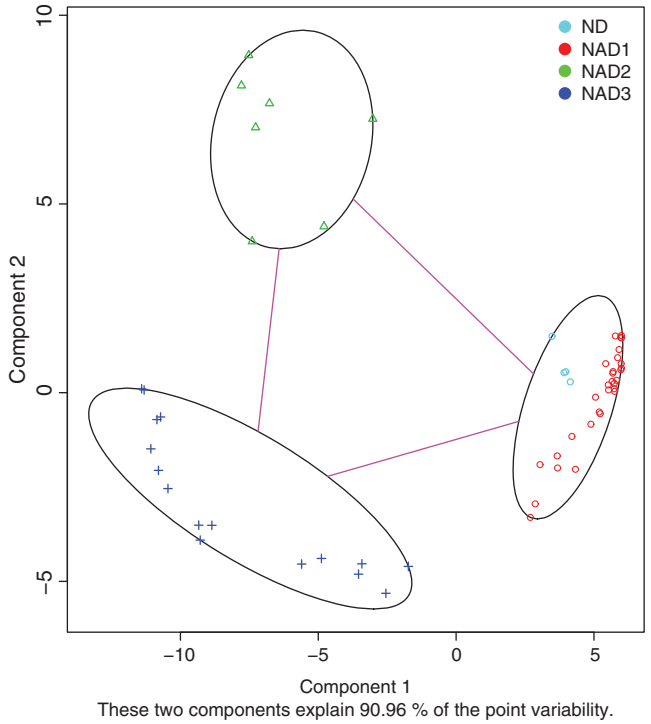

Figure 2. Bivariate clustering of the normal and normal-adjacent expression profiles. The emergence of three distinct subgroups, NAD1, NAD2 and NAD3, respectively, can be noted. The first subgroup (blue and red circles) comprises of ND and the normal-adjacent samples whose profiles most closely resemble those of ND (NAD1). Sample profiles in NAD2 (green triangles) are sufficiently different to NAD1 to form a distinct subgroup. The third subgroup (NAD3) reflects the expression profiles of samples most closely resembling those of PDAC (dark blue crosses).

evaluate the global structure of the transcriptional space of a cancer-specific environment across its different samples and models, confirming or disproving the original findings was not our intention and should not be taken as evidence that the published analyses are wrong.

The xenograft models were found to have a greater similarity to cell lines than to PDAC, suggesting that these models are unable to truly represent alterations occurring in pancreatic tumours (Figure 1). Differential expression relating to the effect of various drugs and treatments on ectopic xenografts and pancreatic cancer cell lines were also investigated. Although we were not able to fully replicate the results at the gene level, our overall findings were, to a certain extent, in concordance with the reported findings at the pathway level.

Attempts were made to identify key biological characteristics, which could account for any variability or similarities between the different cell lines: AsPC1, YPAC, HuPT4 and MiaPaCa2 (well-characterised pancreatic-cancer cell lines with at least three replicates available) (Supplementary Figure S9). Overall, the cell lines separated according to phenotypic characteristics, with some effect being exerted based on the origin of the cell line (Supplementary Figure S10). Pathway analysis on the four, aforementioned, cell lines appeared to support this, with invasion- and metastasis-associated pathways and molecules being deregulated to a greater extent in the more aggressive cell lines. An interesting observation was that the MiaPaCa2 cell line appears to be unique in its own right. The expression profiles of this cell line, obtained from multiple studies and exposed to different treatments, remained consistently distinct from all other pancreatic cancer cell lines. As before, the risk of bias from individual studies was assessed, with the biological effects being greater than any laboratory effects (Supplementary Figure S11).

Portal for data-mining

The web portal (http://www.pancreasexpression.org/PancreaticCancerLandscape.html) accompanying this article was developed as an accompanying tool to add confidence to gene-expression queries made using PED. In addition to allowing a rapid querying of pre-processed analyses of the data for all the well-represented defined subtypes, it also offers visualisation of the expression boxplots and heatmaps. At present, the four primary modes for querying these analyses are by: gene/probe, comparison, biological function and pathway.

Similarly to PED, scientists are able both to follow their gene of interest and mine for additional genes, thereby facilitating the design of validation experiments. In addition, they are also able to observe the pathways to which these genes are linked. This increased dimensionality to PED queries will facilitate novel findings and conclusions.

DISCUSSION

Our study encompasses the largest number of pancreatic-cancer profiles generated on a high-resolution expression array thereby ensuring that this is the most comprehensive analytical analysis of pancreatic cancer to date. By subjecting all the datasets from all the studies to the same rigorous quality control criteria and treating them jointly to a single, well-annotated, data processing pipeline, we have eliminated any differences in results attributable to analytical diversity. Furthermore, the increased statistical power and the capacity to generate comparisons not available from the disparate studies increase the ability to provide novel insights. This robust integrative analysis of multiple and diverse pancreatic cancer datasets has highlighted some important findings and issues.

Our main finding suggests that normal tissues, often used as a baseline against which cancer profiles are compared, may have already acquired a number of genetic alterations. This is of relevance not only in highlighting the possibility of 'field change' in cancer, and the genes that characterise this, but importantly, that the 'normal' matched samples in many studies may not be an appropriate baseline for comparison with cancers. This may partially contribute to the lack of reproducibility between studies. We also show distinctions between profiles generated by studies based on cell lines and cell line-derived models to those using tumour samples. The ectopic subcutaneous xenograft models do not appear capable of accurately representing tumour behaviour, instead sharing

greater similarities with pancreatic cancer cell lines (33). Furthermore, our observations, especially those relating the MiaPaCa2 cell lines, suggest that it is important for the research community to examine the origins and properties of cancer cell lines prior to use, thereby ensuring optimal recapitulation of the characteristics of the cancer being investigated.

Initially conducted to provide an insight to the molecular events governing the transcriptional space of pancreatic cancer, this study serves to highlight concerns with the availability and quality of publicly accessible gene expression data. The main issue encountered was the lack of sufficient clinical and histopathological information available in the public domain, all of which act as a barrier to the accurate interpretation of cancer expression data. Similarly, a lack of detailed experimental and analytical documentation also hampers conclusive data evaluation and reproducibility. Attempts to overcome some of these limitations by obtaining access to additional clinical data were not met with success, precluding further investigation in the majority of cases. Heterogeneity of information currently available to the pancreatic cancer community means that, while our findings are interesting, without additional documentation regarding these samples, accurate evaluations of the data are not possible.

Our study shows the limitations that lack of complete good quality data can have on interpretation of results. First, in order to fully determine whether the seemingly normal-adjacent tissues have undergone alteration when compared to tissues obtained from ND, this study would have needed to analyse tissue selected to ensure similar tissue composition and exclude areas of inflammation and fibrosis. Unfortunately, chronic pancreatitis data profiled on the same platform were not available in public repositories for a direct comparison of expression signatures. This would have enabled considerably more reliable evidence for or against a 'cancer field effect' hypothesis in the normal-adjacent subgroup. Furthermore, information on the distance of the margins of resection along with additional histopathological data was not provided in the majority of studies used in this analysis. Lack of standardisation of histopathological examinations have recently been shown to influence the reporting of resection margin status (34). Therefore, it is possible that some of our observations are attributable to a contamination artefact with differing degrees of cancer cell infiltration within samples utilised as normal.

Increased confidence to our second observation, that ectopic subcutaneous xenograft and cell-line models do not appear to represent tumour behaviour *in vivo*, could be attained by ensuring that profiles obtained using the same sampling procedures are used for comparison. In this instance, use of bulk tissues, containing not only cancer cells but also stromal cells, from primary tumour samples was compared to high-cellularity pancreatic-carcinoma profiles from the xenograft models. Thus, it is possible that some of the observations could be attributed to differences in contaminating stromal cells. In addition, some of the differences in gene expression could be attributed to the contribution of desmoplastic stroma, of which pancreatic tumours tend to have a large proportion, which is not reproduced in cell line and xenograft models (35). This hypothesis could be either supported or disproved by a comparison of primary tissue profiles with micro-dissected profiles, in which desired cells are isolated from the xenograft specimens from the same tumour. Unfortunately, the low quality of the tissues available for this meta-analysis did not allow for this comparison to be generated.

To further enhance transparency and avoid publication and/or author bias, all our findings have been made publicly available in an online tool. Designed to complement PED queries, this tool will enable the pancreatic cancer research community to interrogate and compare results from the two main comparative analysis approaches; literature analysis and re-analysis of raw data. Integration of the two types of results will increase the utility of both resources and will help address issues associated with publication bias.

We have shown that data founded on poor quality assurance and insufficient documentation provision and uniformity is problematic to re-analyse, restricting the evaluative capabilities of our aggregate analysis, these limitations have also proven dangerous in the clinical environment (36). Moreover, research that bases findings predominantly on the results of a small number of individual studies, for which detailed documentation is missing, will contain any errors made in the original studies.

So, while integrative global analyses are capable of providing a wealth of valuable data and insights unachievable from single studies, the poor information on sample histology (linked to the raw data available in public databases) makes it difficult to interpret, with confidence, any results towards new biological insight. Our study adds to the proof that incomplete information on tissue composition and/or the use of well-defined tissues severely hampers any meaningful data analysis and conclusive data interpretation. This problem seems to be prominent in a disease as complex as pancreatic cancer, known for its confounding characteristics such as poor cellularity; high-stromal reaction; accompanying pancreatitis; and various tumour-associated inflammatory cells. Due to the plasticity of pancreatic cells and reports of multiple cell types being potential precursor cells, it remains an open question whether the relevant normal cells to be used as a baseline for comparison to ductal pancreatic cancer are normal ductal cells or acinar cells. The data from pancreatic cancer and normal-adjacent tissues available for this study in public repositories rely on studies using whole tissues with no attached information on histological quality criteria. Thus, the criteria used to define normal-adjacent tissues are unclear from these studies.

For most journals, submitting raw data to public repositories is mandatory. However, while public repositories verify the file format, the quality of clinical and raw data is not assessed. It is vital that both original submitters and public repositories ensure that good quality raw data in conjunction with full histological and clinical data, detailed experimental design and unambiguous analysis pipelines are made publicly available. Without addressing these

obstacles, the increasing abundance of any disease data becoming available cannot be used to its maximal potential.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Cancer Research UK (programme grant C355/A6253); Breast Cancer Campaign (to R.J.C.). Funding for open access charge: Cancer Research UK (programme grant C355/A6253).

Conflict of interest statement. None declared.

REFERENCES

1. Schneider,J.A., Pungliya,M.S., Choi,J.Y., Jiang,R., Sun,X.J., Salisbury,B.A. and Stephens,J.C. (2003) DNA variability of human genes. *Mech. Ageing Dev.*, **124**, 17–25.
2. Hanahan,D. and Weinberg,R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.
3. Dudley,J.T., Tibshirani,R., Deshpande,T. and Butte,A.J. (2009) Disease signatures are robust across tissues and experiments. *Mol. Syst. Biol.*, **5**, 307.
4. Ioannidis,J.P., Allison,D.B., Ball,C.A., Coulibaly,I., Cui,X., Culhane,A.C., Falchi,M., Furlanello,C., Game,L., Jurman,G. *et al.* (2009) Repeatability of published microarray gene expression analyses. *Nat. Genet.*, **41**, 149–155.
5. Lukk,M., Kapushesky,M., Nikkila,J., Parkinson,H., Goncalves,A., Huber,W., Ukkonen,E. and Brazma,A. (2010) A global map of human gene expression. *Nat. Biotechnol.*, **28**, 322–324.
6. Zheng-Bradley,X., Rung,J., Parkinson,H. and Brazma,A. Large scale comparison of global gene expression patterns in human and mouse. *Genome Biol.*, **11**, R124.
7. Hariharan,D., Saied,A. and Kocher,H.M. (2008) Analysis of mortality rates for pancreatic cancer across the world. *HPB*, **10**, 58–62.
8. Goonetilleke,K.S. and Siriwardena,A.K. (2008) Current status of gene expression profiling of pancreatic cancer. *Int. J. Surg.*, **6**, 81–83.
9. Chelala,C., Hahn,S.A., Whiteman,H.J., Barry,S., Hariharan,D., Radon,T.P., Lemoine,N.R. and Crnogorac-Jurcevic,T. (2007) Pancreatic Expression database: a generic model for the organization, integration and mining of complex cancer datasets. *BMC Genomics*, **8**, 439.
10. Cutts,R.J., Gadaleta,E., Hahn,S.A., Crnogorac-Jurcevic,T., Lemoine,N.R. and Chelala,C. (2011) The Pancreatic Expression database: 2011 update. *Nucleic Acids Res.*, **39**, D1023–D1028.
11. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M., Marshall,K.A. *et al.* (2009) NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res.*, **37**, D885–D890.
12. Parkinson,H., Kapushesky,M., Kolesnikov,N., Rustici,G., Shojatalab,M., Abeygunawardena,N., Berube,H., Dylag,M., Emam,I., Farne,A. *et al.* (2009) ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.*, **37**, D868–D872.
13. Rosner,B. (1983) Percentage points for a generalized ESD many-outlier procedure. *Technometrics*, **25**, 165–172.
14. Asare,A.L., Gao,Z., Carey,V.J., Wang,R. and Seyfert-Margolis,V. (2009) Power enhancement via multivariate outlier testing with gene expression arrays. *Bioinformatics*, **25**, 48–53.
15. Wu,Z. and Irizarry,R.A. (2004) Preprocessing of oligonucleotide array data. *Nat. Biotechnol.*, **22**, 656–658; author reply 658.
16. Mieczkowski,J., Tyburczy,M.E., Dabrowski,M. and Pokarowski,P. (2010) Probe set filtering increases correlation between Affymetrix GeneChip and qRT-PCR expression measurements. *BMC Bioinformatics*, **11**, 104.
17. Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, 1.
18. Badea,L., Herlea,V., Dima,S.O., Dumitrascu,T. and Popescu,I. (2008) Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia. *Hepatogastroenterology*, **55**, 2016–2027.
19. Maupin,K.A., Sinha,A., Eugster,E., Miller,J., Ross,J., Paulino,V., Keshamouni,V.G., Tran,N., Berens,M., Webb,C. *et al.* Glycogene expression alterations associated with pancreatic cancer epithelial-mesenchymal transition in complementary model systems. *PLoS ONE*, **5**, e13002.
20. Humbert,M., Casteran,N., Letard,S., Hanssens,K., Iovanna,J., Finetti,P., Bertucci,F., Bader,T., Mansfield,C.D., Moussy,A. *et al.* (2010) Masitinib combined with standard gemcitabine chemotherapy: in vitro and in vivo studies in human pancreatic tumour cell lines and ectopic mouse model. *PLoS ONE*, **5**, e9430.
21. Jimeno,A., Tan,A.C., Coffa,J., Rajeshkumar,N.V., Kulesza,P., Rubio-Viqueira,B., Wheelhouse,J., Diosdado,B., Messersmith,W.A., Iacobuzio-Donahue,C. *et al.* (2008) Coordinated epidermal growth factor receptor pathway gene overexpression predicts epidermal growth factor receptor inhibitor sensitivity in pancreatic cancer. *Cancer Res.*, **68**, 2841–2849.
22. Nakamura,T., Kuwai,T., Kitadai,Y., Sasaki,T., Fan,D., Coombes,K.R., Kim,S.J. and Fidler,I.J. (2007) Zonal heterogeneity for gene expression in human pancreatic carcinoma. *Cancer Res.*, **67**, 7597–7604.
23. Pei,H., Li,L., Fridley,B.L., Jenkins,G.D., Kalari,K.R., Lingle,W., Petersen,G., Lou,Z. and Wang,L. (2009) FKBP51 affects cancer cell response to chemotherapy by negatively regulating Akt. *Cancer Cell*, **16**, 259–266.
24. Selga,E., Oleaga,C., Ramirez,S., de Almagro,M.C., Noe,V. and Ciudad,C.J. (2009) Networking of differentially expressed genes in human cancer cells resistant to methotrexate. *Genome Med.*, **1**, 83.
25. Song,D., Chaerkady,R., Tan,A.C., Garcia-Garcia,E., Nalli,A., Suarez-Gauthier,A., Lopez-Rios,F., Zhang,X.F., Solomon,A., Tong,J. *et al.* (2008) Antitumor activity and molecular effects of the novel heat shock protein 90 inhibitor, IPI-504, in pancreatic cancer. *Mol. Cancer Ther.*, **7**, 3275–3284.
26. Wang,Y., Gangeswaran,R., Zhao,X., Wang,P., Tysome,J., Bhakta,V., Yuan,M., Chikkanna-Gowda,C.P., Jiang,G., Gao,D. *et al.* (2009) CEACAM6 attenuates adenovirus infection by antagonizing viral trafficking in cancer cells. *J. Clin. Invest.*, **119**, 1604–1615.
27. Yauch,R.L., Gould,S.E., Scales,S.J., Tang,T., Tian,H., Ahn,C.P., Marshall,D., Fu,L., Januario,T., Kallop,D. *et al.* (2008) A paracrine requirement for hedgehog signalling in cancer. *Nature*, **455**, 406–410.
28. Zhang,L., Farrell,J.J., Zhou,H., Elashoff,D., Akin,D., Park,N.H., Chia,D. and Wong,D.T. (2009) Salivary transcriptomic biomarkers for detection of resectable pancreatic cancer. *Gastroenterology*, **138**, 949–957, e941–947.
29. Alderton,G.K. (2010) Transcriptomics: common disease pathogenesis pathways. *Nat. Rev. Cancer*, **10**, 387.
30. Hirsch,H.A., Iliopoulos,D., Joshi,A., Zhang,Y., Jaeger,S.A., Bulyk,M., Tsichlis,P.N., Shirley Liu,X. and Struhl,K. (2010) A transcriptional signature and common gene networks link cancer with lipid metabolism and diverse human diseases. *Cancer Cell*, **17**, 348–361.
31. Jones,S., Zhang,X., Parsons,D.W., Lin,J.C., Leary,R.J., Angenendt,P., Mankoo,P., Carter,H., Kamiyama,H., Jimeno,A. *et al.* (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*, **321**, 1801–1806.

32. Balkwill,F. (2004) Cancer and the chemokine network. *Nat. Rev. Cancer*, **4**, 540–550.
33. van Staveren,W.C., Solis,D.Y., Hebrant,A., Detours,V., Dumont,J.E. and Maenhaut,C. (2009) Human cancer cell lines: experimental models for cancer cells in situ? For cancer stem cells? *Biochim. Biophys. Acta.*, **1795**, 92–103.
34. Verbeke,C.S., Leitch,D., Menon,K.V., McMahon,M.J., Guillou,P.J. and Anthoney,A. (2006) Redefining the R1 resection in pancreatic cancer. *Br. J. Surg.*, **93**, 1232–1237.
35. Froeling,F.E., Mirza,T.A., Feakins,R.M., Seedhar,A., Elia,G., Hart,I.R. and Kocher,H.M. (2009) Organotypic culture model of pancreatic cancer demonstrates that stromal cells modulate E-cadherin, beta-catenin, and Ezrin expression in tumor cells. *Am. J. Pathol.*, **175**, 636–648.
36. Baggerly,K.A. and Coombes,K.R. (2009) Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *Ann. Appl. Stat.*, **3**, 1309–1334.