


RESEARCH

Open Access



Transcriptional signals of transformation in human cancer

Gerda Kildisiute^{1†}, Maria Kalyva^{2†}, Rasa Elmentaite¹, Stijn van Dongen¹, Christine Thevanesan³, Alice Piapi³, Kirsty Ambridge¹, Elena Prigmore¹, Muzlifah Haniffa^{1,4}, Sarah A. Teichmann^{1,5}, Karin Straathof³, Isidro Cortés-Ciriano^{2*}, Sam Behjati^{1,6,7*} and Matthew D. Young^{1*} 

Abstract

Background As normal cells transform into cancers, their cell state changes, which may drive cancer cells into a stem-like or more primordial, foetal, or embryonic cell state. The transcriptomic profile of this final state may encode information about cancer's origin and how cancers relate to their normal cell counterparts.

Methods Here, we used single-cell atlases to study cancer transformation in transcriptional terms. We utilised bulk transcriptomes across a wide spectrum of adult and childhood cancers, using a previously established method to interrogate their relationship to normal cell states. We extend and validate these findings using single-cell cancer transcriptomes and organ-specific atlases of colorectal and liver cancer.

Results Our bulk transcriptomic data reveals that adult cancers rarely return to an embryonic state, but that a foetal state is a near-universal feature of childhood cancers. This finding was confirmed with single-cell cancer transcriptomes.

Conclusions Our findings provide a nuanced picture of transformation in human cancer, indicating cancer-specific rather than universal patterns of transformation pervade adult epithelial cancers.

[†]Gerda Kildisiute and Maria Kalyva contributed equally.

*Correspondence:

Isidro Cortés-Ciriano

icortes@ebi.ac.uk

Sam Behjati

sb31@sanger.ac.uk

Matthew D. Young

my4@sanger.ac.uk

¹ Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK

² EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridge, UK

³ University College London Cancer Institute and Great Ormond Street Biomedical Research Centre, London, UK

⁴ Biosciences Institute and Newcastle NIHR-BRC Dermatology, Newcastle University, Newcastle Upon Tyne, UK

⁵ Cavendish Laboratory, University of Cambridge, JJ Thomson Ave, Cambridge, UK

⁶ Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK

⁷ Department of Paediatrics, University of Cambridge, Cambridge, UK

Background

Malignant transformation is underpinned by changes in the cell state towards a less differentiated or stem-like cell state. Consequently, cancers may broadly retain the cell state of origin or transform to resemble a more primitive cell type, of tissue-specific foetal or primordial embryonic cells. Which cell state—primordial or otherwise—cancer assumes is a fundamental question of cancer biology as it provides a net readout of the consequences of cancer formation.

The cancer cell state can be studied using transcriptional readouts that represent cellular phenotypes and differentiation states. The key challenge in such an analysis is defining cellular states, such as stemness, embryonicness, foetalness, etc., in quantitative molecular terms. One approach is to use appropriate mRNA signals, identified using unsupervised



clustering or pattern recognition methods applied to single cell atlases, that embody the state the cancer may transform into. A number of studies have used this approach to investigate “stemness” across the entire spectrum of human cancer [1–3], for example, by measuring stemness signals derived from in vitro differentiating human embryonic stem cells [4]. These studies indicate that human cancer transcriptomes resemble transcriptional modules of in vitro differentiating embryonic stem cells. Whether the correlation of such modules represents a global transcriptional transformation towards an antenatal state remains unknown. Furthermore, it is conceivable that other developmental states, such as gastrulation, foetal tissues, or post-natal stem cells, may also contribute to the state of cancer transcriptomes.

Recent efforts, referred to as the Human Cell Atlas project [5], have begun to provide transcriptional definitions of human cells across all stages of development for multiple tissues using single-cell mRNA sequencing. These reference transcriptomes provide the opportunity to study cancer transformation using a more complete and nuanced set of reference states than has been possible in the past.

We approach the quantification of cancer cell state by probing cancer bulk transcriptomes for evidence of single-cell-defined human reference signals using a previously established method [6]. This method, which is conceptually similar to deconvolution, quantifies the extent to which reference signals explain the observed expression profile of a bulk transcriptome. Crucially, this approach also estimates the fraction of the bulk transcriptomic profile unexplained by the provided reference, which, together with goodness of fit metrics, assesses the extent to which reference signals account for bulk transcriptomes.

Methods

The human reference cell types we included in each analysis encompassed the entire spectrum of human tissue development: human pre-gastrulation epiblast and hypoblast cells [7]; cells representing the three germ layers, endoderm, mesoderm, and ectoderm in their earliest stages [8]; and tissue-specific foetal [9] and adult cells [10] (Additional file 1: Table S1). We applied this combined reference to a wide spectrum of childhood and adult solid tissue cancers [11–16] (Additional file 1: Table S2). We then extended these results using high-resolution atlases of specific tissues together with single-cell cancer transcriptomes for two types of adult epithelial cancers.

Single-cell hepatoblastoma processing

Surplus tumour tissue obtained at diagnostic biopsy or tumour resection was processed immediately after receipt in the histopathology laboratory (<1 h after interventional radiology/surgical procedure). Tissue was minced using a scalpel and then incubated in RPMI 1640, supplemented with 10% foetal calf serum, 1% l-glutamine, and 1% penicillin/streptomycin, with collagenase IV (1.6 mg/ml; catalog no. 11410982; MP Biomedicals), for 30 min at 37 °C, inverting the tube every 10 min. The digested tissue was passed through a 70-µm filter and incubated in 1×RBC lysis buffer (catalog no. 420301; BioLegend) for 10 min at room temperature.

The obtained single-cell suspension was used for downstream processing. Part of the single-cell suspension was depleted of CD45+ cells to enrich for tumour cells using a CD45 MicroBeads kit (catalog no. 130–045-801; Miltenyi Biotec), following the manufacturer’s protocol. Both CD45 nondepleted and CD45-depleted single-cell suspensions were depleted of dead cells using a Dead Cell Removal kit (catalog no. 130–090-101; Miltenyi Biotec), following the manufacturer’s protocol. Obtained viable single-cell suspensions were processed on the 10× Chromium platform.

The concentration of single-cell suspensions was manually counted using a haemocytometer and adjusted to 1000 cells/µl or counted by flow cytometry. Cells were loaded according to the standard protocol of the Chromium Single Cell 3’ Kit (v2 and v3 chemistry). All the following steps were performed according to the standard manufacturer’s protocol. One lane of Illumina HiSeq 4000 per 10× chip position was used.

Single-cell RNA-seq data were mapped, and counts of molecules per barcode were quantified using the 10× software package cellranger (versions 2.0.2 and 3.0.2) to map sequencing data to version 2.1.0 of the build of the GRCh38 reference genome supplied by 10x.

Single-cell data processing

To perform cell signal analysis and/or logistic regression, the following were required for each single-cell dataset: (i) a count table that has undergone quality control and (ii) cell annotations. Where both could be obtained, no further action was taken. These were taken directly from the publication where available, or reproduced following the methods of the relevant publication. All datasets were 10x, except where specified below. All cell types with < 10 cells were removed. When merging single-cell data matrices from different sources, only common rows (ENSEMBL IDs wherever possible) were kept. Additional processing was performed for the following references:

Gastrulation data

QC'd count tables and annotations were obtained from the authors. As this was a Smart-Seq2 dataset, the raw count table was transformed to transcripts per kilobase million (TPM) to account for gene length bias.

Foetal liver

QC'd count tables and annotations were obtained from the authors. Sub-types of erythroid cells, B-cells, and dendritic cells were merged into one cell type to simplify the annotation (e.g. early, mid, and late erythroid cells were re-labelled to erythroid cells).

Adult liver

QC'd count tables and annotations were obtained from the authors. Clusters of hepatocytes, T-cells, and liver sinusoidal endothelial cells into one cell type to simplify the annotation.

Epithelial liver cell reference by Segal et al.

A QC'd raw count table was obtained from the authors. As this was a Smart-Seq2 dataset, raw count tables were transformed to transcripts per kilobase million to account for gene length bias. Adult and foetal HHyP and hepatocyte, and adult BECs, were annotated based on marker gene expression.

Hepatoblastoma

Single-cell RNA-seq data were mapped, and counts of molecules per barcode were quantified using the 10x software package cellranger (version 3.0.0) to map sequencing data to version 2.1.0 of the build of the GRCh38 reference genome supplied by 10x. Cells with >20% mitochondrial expression, fewer than 200 detected genes, or 500 UMIs were removed as low quality. Data were log normalised and clustered using a community detection method [17]. Clusters were annotated with the following markers: *CD45* (leukocyte marker), *HBA* and *HBB* (erythrocyte markers), *EPCAM* and *AFP* (tumour cell markers), *PECAMI1* (endothelial marker), and *ACTA2* (hepatic stellate marker). Only those cells that could be definitively annotated using these markers were used for the analysis.

Hepatocellular carcinoma

Single-cell RNA-seq data were mapped, and counts of molecules per barcode were quantified using the 10x software package cellranger (version 3.0.0) to map sequencing data to version 2.1.0 of the build of the GRCh38 reference genome supplied by 10x. Cells with >10% mitochondrial expression, fewer than 300

detected genes, or 1000 UMIs were removed as low quality. Cells were then annotated using marker gene expression [13].

Bulk data processing

To perform cell signal analysis, the following were required for each bulk sample: (i) gene counts and (ii) gene lengths. These were generated in the following manner:

TCGA data

Gene counts and lengths were taken from the recount2 mapping of the TCGA [18].

St Jude's

Raw sequencing reads were mapped against the GRCh38 reference v1.2.0 provided by 10X, using a pseudo-aligner [19], which produced both gene counts and lengths. Cancers which were not unique to childhood (osteosarcoma and melanoma) were removed.

Hepatoblastoma (tumour and normal)

Gene counts were provided by authors, and the same gene lengths as used for the TCGA were used.

Colorectal bulk (tumour, normal, adenomas)

Gene counts and lengths were provided by the authors.

Foetal liver bulk

Gene counts were provided by authors, and the same gene lengths as used for the TCGA were used.

Foetal gut bulk

Gene counts were provided by authors, and the same gene lengths as used for the TCGA were used.

Blastoid bulk

Gene counts were downloaded from GEO, and at accession GSE179040, the same gene lengths as used by the TCGA were used.

GTEX

TPM values provided by the GTEx consortium were used, with gene lengths set to 1.

Cell signal analysis: comparing bulk transcriptomes to a single cell reference

Cell signal analysis was performed as previously described [20]. For the pan-cancer analysis, we excluded mitochondrial and ribosomal genes and downweighted the likelihood of housekeeping genes by 50%. From the foetal reference, we excluded reference cell types from

the foetal heart as there were no corresponding cell types in the adult reference.

Cell similarity: comparing single-cell transcriptomes to a single-cell reference

To measure the similarity of a target single-cell transcriptome to a reference single-cell dataset, logistic regression was used to train a model on the reference single-cell dataset [21]. This model was then applied to predict the probability of similarity (normalised to 1 across all categories), between each cell type in the reference dataset and each cell in the target dataset.

Stemness score calculation

Code to calculate the stemness score as previously described was provided by the authors [4]. To provide a comparison between all samples, both adult and paediatric, the stemness score was calculated across all bulk samples from both the TCGA and St Judes simultaneously.

Cell signal summary score calculation

Cell signal analysis produces for each bulk transcriptome, the relative contribution of each single cell defined cell type provided in the reference. In this paper, we make use of various summary scores (embryonic, foetalness, etc.) which we define as the sum of all relevant cell signal analysis scores for a bulk transcriptome. For example, we defined the foetalness score as the sum of all the cell signal contributions across all cell types derived from foetal tissues, when cell signal analysis is performed using a reference consisting of both a foetal and mature tissue reference.

Results

The starting point of our analysis was to test whether our approach captured similar information to existing measures of “stemness” in human adult cancer. We compared previously published stemness scores that build on

gene sets to a broader signal of stemness, the transcriptomes of human pre-gastrulation embryo cells (hypoblast and epiblast). Despite these differences in underlying methodology, we observed that stemness score and our measure, the fraction of each cancer bulk transcriptome explained by human pre-gastrulation embryo cells, correlated strongly (Pearson correlation 0.45, Additional file 2: Fig. S1). However, we also found that only a small fraction of each cancer transcriptome could be explained by early embryonic signals (mean goodness of fit = 0.25; mean fraction of unexplained signal = 0.52), which was particularly apparent when compared to positive control transcriptomes (blastoid transcriptomes [22]). Examining which genes drove the stemness score, we found that genes associated with S and G2M phases of the cell cycle were significant drivers of stemness (Additional file 2: Fig. S1). Consistent with this observation, we found stemness scores were higher in dividing than non-dividing tissues (Additional file 2: Fig. S1). Overall, these results suggest that proliferation is a key driver of previously reported stemness scores. This then raises the question of whether cancers exhibit tissue-specific signals of reversion to an antenatal state beyond proliferative signals.

Accordingly, we re-examined bulk transcriptomes by progressively expanding the reference we used for comparison, which enabled us to assess at which point cancer transcriptomes were most completely accounted for. In successive iterations of the analysis, we expanded the reference of pre-gastrulation cells with the following human cell atlases: gastrulation embryo, foetal tissues, and adult tissues. For the control population (blastoids), the vast majority of the bulk transcriptomic signal was explained by the early embryo and gastrulation reference, even when 658,368 foetal and post-natal cells were provided (Fig. 1A, B, Additional file 2: Fig. S2-4). By contrast, very little of the early embryonic signals were retained by solid cancers once tissue-specific references were available (Fig. 1A, B, Additional file 2: Fig. S2-4). We also

(See figure on next page.)

Fig. 1 **A** Pan-cancer analysis of transformation state from pan-tissue single-cell reference atlases. Fit quality and embryonic signal of bulk transcriptomes with increasingly complete reference atlases: Fractional contribution of embryonic reference (*y*-axis, early embryo + gastrulation for full reference, early embryo otherwise) in explaining bulk transcriptomes (dots) as a function of goodness of fit (*x*-axis, pseudo *R*-squared) when fit using single cell reference consisting of cells from the early embryo (yellow), early embryo and gastrulation (dark red), or early embryo, gastrulation, foetal, and mature pan-tissue reference (green). Bulk cancer transcriptomes are circled in black and genuinely embryonic controls (blastoids) are circled in grey. **B** Relative contribution of references to explaining the bulk transcriptomes of a range of adult and childhood cancers: Average relative contribution of early embryo (yellow), gastrulation (dark red), foetal (purple), and adult (green) single cell reference populations in explaining bulk transcriptomes (*y*-axis) for different combinations of these references (*x*-axis, labels at top). Bulk transcriptomes are organised by source (labels on the right). **C** Childhood cancers have a stronger foetal contribution than adult cancers or control populations: Relative contribution of foetal reference (*y*-axis) in explaining bulk cancer transcriptomes (dots), when provided a complete set of the early embryo, gastrulation, foetal, and adult single-cell references. Bulk transcriptomes are split into childhood (purple), adult (green), and blastoid control (orange) and then by cancer type (*x*-axis). Distributions are summarised by median (horizontal lines), 1st and 3rd quartiles (horizontal lines for cancer types, shaded coloured areas for childhood/adult/control), and 1.5 times inter-quartile range (light-shaded areas)

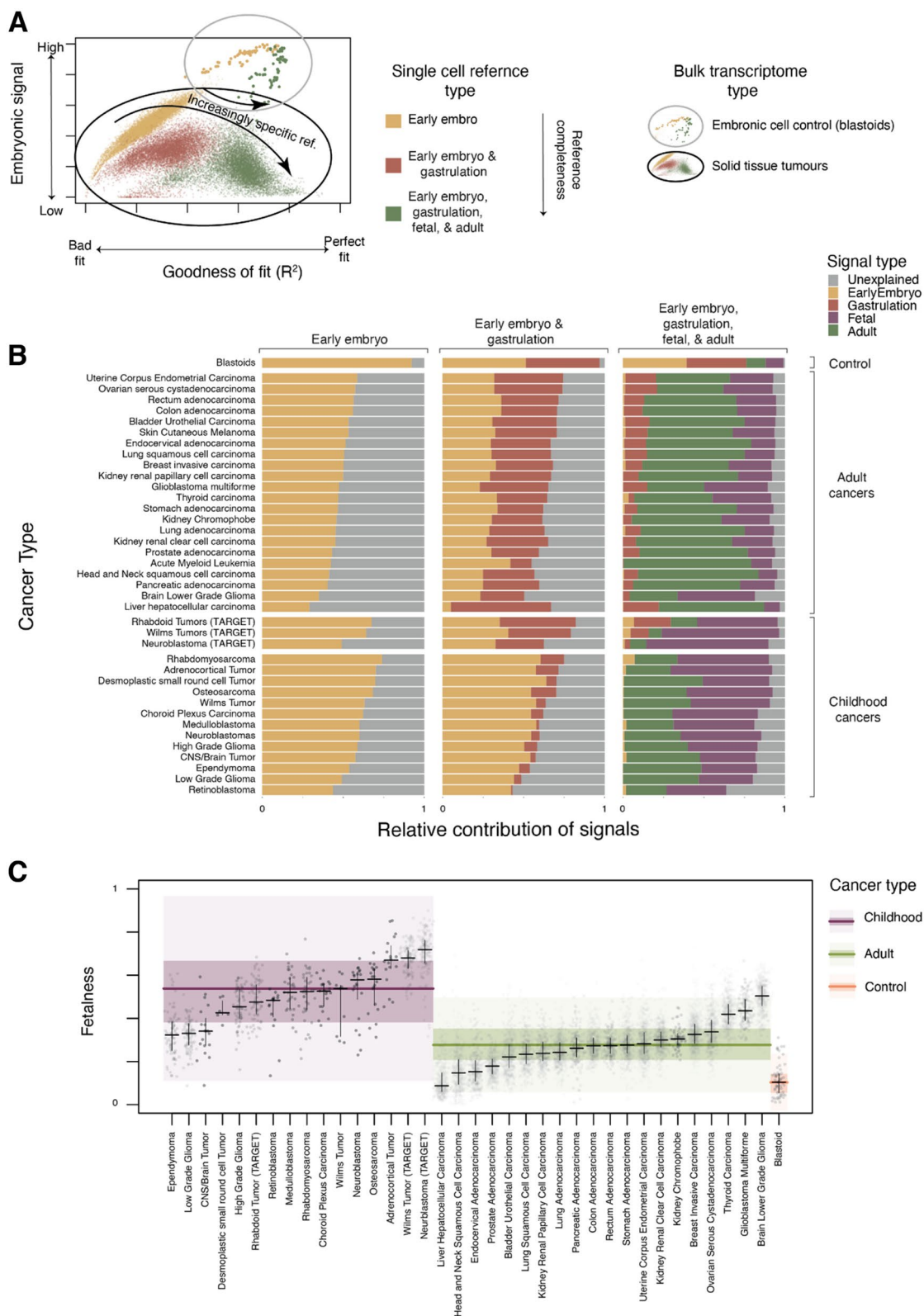


Fig. 1 (See legend on previous page.)

investigated whether aggregated signals from the least differentiated cell types once tissue-specific references are available are predictive of survival outcome. While in three adult tumour types and one paediatric type, a higher signal from these least differentiated cell types correlated with poorer survival outcome (Additional file 2: Fig. S7), this was not universal. Taken together, these results indicate that while cancer cells may share functional features with embryonic stem cells, the cancer transcriptome does not return to an embryonic state.

While embryonic signals seemed to have no discernable biological meaning once more mature references were available, we found a strong and consistent difference between the amount of retained foetal signal, or “foetalness”, in childhood versus adult tumours (Fig. 1C). This finding extends previous work in renal tumours that described a foetal-like transcriptome in childhood, but not adult kidney tumours [20]. One limitation of our approach is that pan-tissue human atlases, which we used in our analyses, inevitably lack the granularity of annotation and resolution of dedicated, tissue-specific cell atlases. Although our pan-tissue analyses are sufficient to make general statements about stemness, foetalness, etc., they are insufficient to make detailed comparisons to specific cell types. Therefore, in the second part of our analyses, we sought to study particular tumours utilising detailed tissue-specific cell atlases and single cancer cell transcriptomes

for validation of our findings. For this analysis, we focused on liver and colorectal cancer, as both developmental and adult cell atlases of normal liver and colorectal tissues are available.

Hepatocellular carcinoma (HCC) is the most common adult liver cancer. It is an epithelial cancer that arises from hepatocytes, often in the context of chronic liver disease. The precise state of hepatocellular carcinoma has not been established, compounded by an ongoing debate about the nature of hepatobiliary stem cells. It is noteworthy that some hepatocellular carcinomas resume expression of foetal albumin (alpha-fetoprotein, AFP), which may represent the reversion of some cancers towards a foetal hepatic state [23]. We assessed 360 hepatocellular carcinoma bulk transcriptomes from 3 cohorts, as well as 64 bulk transcriptomes of the childhood liver cancer, hepatoblastoma (HB) [11, 12, 15]. Using a highly detailed combined reference map of adult and foetal liver [24–26], together with the pre-gastrulation and early embryo references [7, 8], we asked which reference best explained the bulk transcriptomes state. As with our pan-tissue analysis (Fig. 1), the reference foetal and adult liver cells accounted for the majority of the transcriptome in hepatocellular carcinoma and hepatoblastoma (Fig. 2A, Additional file 1: Fig. S5). The state of hepatocellular carcinoma bulk transcriptomes was therefore most completely represented by liver cells, but not by more primordial human cell populations.

(See figure on next page.)

Fig. 2 Detailed analysis of bulk and single cell liver and gut cancers. **A** Contribution of different signal types to bulk transcriptomes of liver: Relative contribution (y-axis) of different reference single-cell populations (horizontal facets) in explaining bulk transcriptomes (dots) grouped by transcriptome type (x-axis within facets) as indicated by x-axis symbols and labels. The fit was performed with all single-cell reference atlases provided, related groups of reference cell populations indicated by hierarchical labels (top), and the distribution of groups of bulk transcriptomes summarised by their median (horizontal lines) and 1st/3rd quartiles (vertical lines). **B** Foetal hepatocyte contribution correlated with AFP expression: Foetal hepatocyte contribution to bulk transcriptomes of hepatocellular carcinoma (y-axis) plotted against alpha-fetoprotein (AFP) expression (log₁₀ of TPM, x-axis), for each bulk transcriptome (dots). A best fit linear trend line is shown along with its equation and associated *R* squared value ($p < 2.2 \times 10^{-16}$, *t*-test). The fit of the foetal hepatocyte signature to the bulk transcriptome was performed excluding the *AFP* gene. **C** UMAP of liver cancer transcriptomes: Dimensionality reduction analysis (UMAP) showing single cell transcriptomes (dots) derived from 1 hepatoblastoma (left) and 8 individual hepatocellular carcinomas (right), grouped by cell type (contours, labels, and colours). Different donors of cancer cells are indicated by different shades of blue. **D** Similarity of single cell HCC/HB transcriptomes with embryonic and liver reference atlases: Similarity score (logistic regression, colour value) of single-cell transcriptomes of liver cancers (C) grouped by cell type (y-axis) and compared to transcriptomes of reference cell types (x-axis) using a reference consisting of embryonic cells, developmental liver, and post-natal liver. Each rectangle represents a group of cells (indicated by y-axis label) and shows the distribution of similarity scores for those single cells compared to the reference cell population (indicated by x-axis label). **E** Contribution of different signal types to bulk transcriptomes of the intestines: Relative contribution (y-axis) of different reference single cell populations (horizontal facets) in explaining bulk transcriptomes (dots) grouped by transcriptome type (x-axis within facets) as indicated by x-axis symbols and labels. The fit was performed with all single cell references provided, related groups of reference cell populations indicated by hierarchical labels (top), and the distribution of groups of bulk transcriptomes summarised by their median (horizontal lines) and 1st/3rd quartiles (vertical lines). **F** UMAP of colorectal carcinoma cell transcriptomes from 25 individuals: Dimensionality reduction analysis (UMAP) showing single-cell transcriptomes (dots) derived from 25 individual HCCs, grouped by cell type (contours, labels, and colours). Different donors of cancer cells are indicated by different shades of blue. **G** Similarity of single-cell CRC transcriptomes with embryonic and intestine reference atlases: Similarity score (logistic regression, colour value) of single-cell transcriptomes of colorectal cancers (F) grouped by cell type (y-axis) and compared to transcriptomes of reference cell types (x-axis) using a reference atlas consisting of embryonic cells, developmental intestine, and post-natal intestine. Each rectangle represents a group of cells (indicated by the label on the y-axis) and shows the distribution of similarity scores for those single cells compared to the reference cell population (indicated by x-axis label)

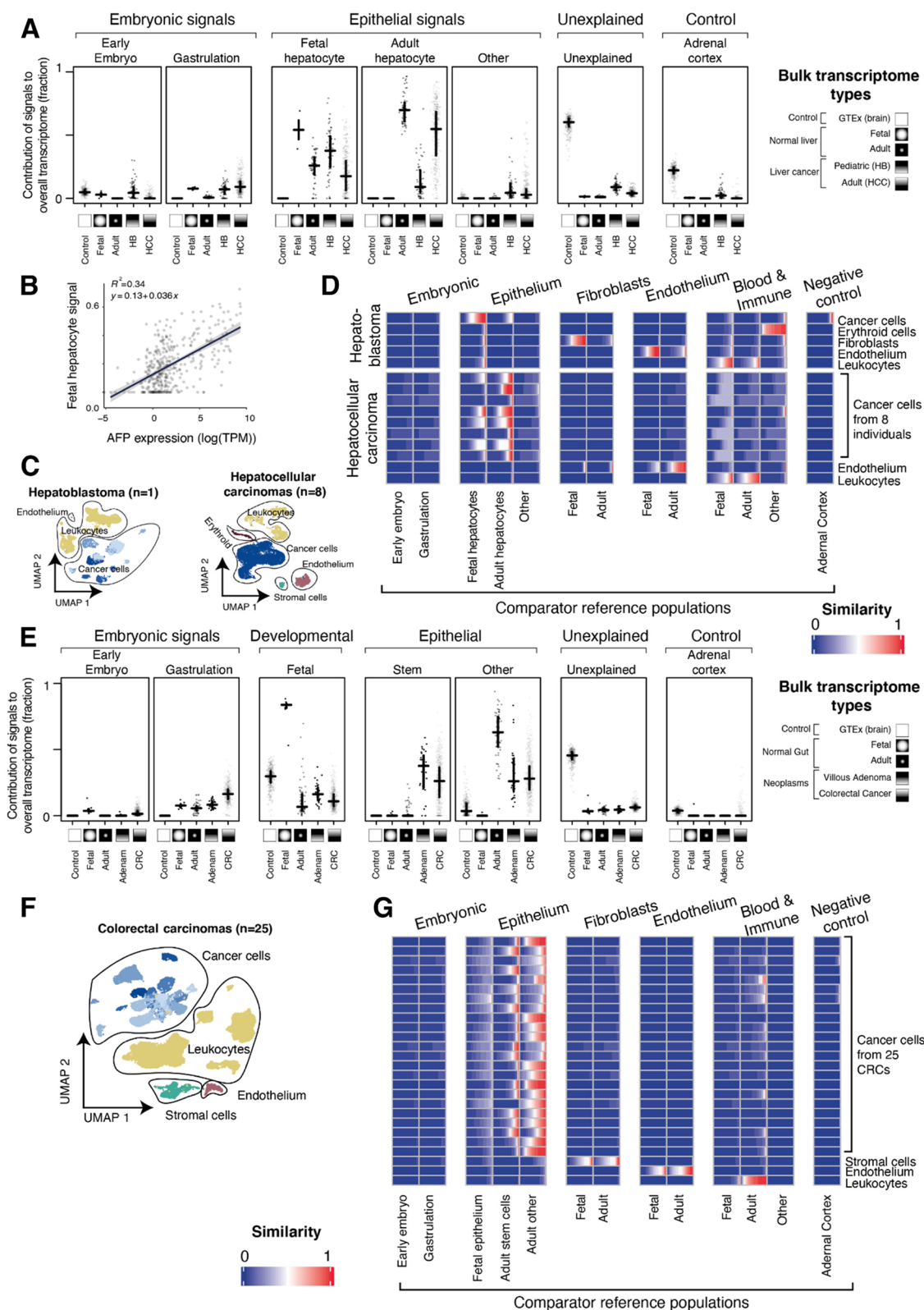


Fig. 2 (See legend on previous page.)

The next step in the analysis was to determine which precise liver cell type best explains the transcriptome in hepatoblastoma and hepatocellular carcinoma and to exclude that the signal derives from non-parenchymal cells. We found that adult and foetal hepatocytes, alone or in combination, were the main cell signal in adult hepatocellular carcinoma, with hepatoblastoma favouring the foetal hepatocytes as expected (Fig. 2A, Additional file 1: Fig. S5). In some HCCs, the foetal hepatoblast signal predominated, which correlated with serum AFP levels and AFP mRNA counts. The predominance of this signal persisted when we removed AFP from the reference transcriptome, indicating that it was not driven by this transcript alone (Fig. 2B). The foetal state of some hepatocellular carcinoma may be of prognostic significance, as AFP serum levels have previously been associated with poor outcomes in hepatocellular carcinoma [27]. Occasional adult tumours exhibited signals of other rare liver cells, namely foetal hepatobiliary hybrid progenitors (HHyP) or so-called adult BEC (hepatocytes and biliary epithelial cells, Additional file 2: Fig. S5). A clinical or pathological significance of these unusual cellular signatures was not apparent. Consistent with our pan-tissue analysis (Fig. 1C), we found that the foetal hepatocyte signal dominated in the childhood cancer hepatoblastoma (Fig. 2A).

To validate these findings, we integrated (published) single-cell transcriptomes from hepatocellular carcinomas [13] ($n=25,605$ cells from 8 tumours) and generated 13,180 single-cell transcriptomes from a hepatoblastoma using the Chromium 10X platform (Fig. 2C). For each single-cell transcriptome, we calculated a similarity score (using logistic regression [21]) against the same reference used to analyse the bulk transcriptomes (Fig. 2D). The single-cell mRNA analyses verified our bulk transcriptomic findings, namely that adult hepatocellular carcinoma may be viewed as aberrant adult hepatocytes, some of which transform towards a foetal hepatoblast state (Fig. 2D). By contrast, most childhood hepatoblastoma cells matched foetal hepatocytes (Fig. 2D).

We next applied our analytical approach to bulk and single-cell transcriptomes of colorectal cancers ($n=65,362$ cells from 23 patients), as well as to adenomas (polyps) [14, 16]. Adenomas are low-grade neoplastic lesions of colorectal epithelium that have the potential to progress to carcinomas, via the sequential acquisition of cancer-causing somatic mutations (adenoma-carcinoma sequence). In the first instance, we assessed which reference most fully accounted for adenoma and cancer transcriptomes (Fig. 2E, Additional file 1: Fig. S6) and again found that the tissue-specific colorectal reference, but not more primordial references, provided the best fit and that foetal cells

contributed little signal to this adult cancer. We then analysed which specific cell type of the colorectal tissue reference mostly explained adenoma bulk transcriptomes (Fig. 2E). We found that adult, but not foetal, colorectal stem cell signals predominated across cancers, with a roughly comparable contribution from more mature epithelial cells. Interestingly, the same stem cell signal pervaded adenomas. Thus, transcriptional, genetic, and histological differences notwithstanding, from a cell state perspective there was no obvious change (e.g. further reversion) from adenomas to carcinomas. We next verified cell signals in colorectal cancers by comparing single cancer cells [14] with normal reference cells, which confirmed that colorectal cancer cells resemble a mixture of stem and more mature epithelial cells (Fig. 2F,G) do not dedifferentiate to more primordial states. Overall these findings indicate that the predominant cell state of premalignant and malignant colorectal tumours is the colorectal stem cell [28]. This may indicate that the cell of origin of these neoplasms is the colorectal stem cells or, that irrespective of where in the differentiation hierarchy of colorectal epithelium tumours originate, they ultimately converge at the stem cell state.

Discussion

The study of the dedifferentiation state of cancer in transcriptional terms has largely focused on similarity to stem cells, or “stemness”. This study demonstrates that when a wider set of reference states is considered, the transcriptional state obtained by cancer cells is more nuanced, with adult cancers best explained by post-natal cells. As such, future work aiming to understand the transcriptional behaviour of adult cancers should consider post-natal tissue-specific models rather than embryonic stem cells.

We also found a clear categorical difference between adult and childhood cancers, with childhood cancer transcriptomes more closely resembling foetal cells. This is consistent with the theory that childhood cancers arise during development. Either way, these findings demonstrate that childhood cancers should be considered differently, at least in transcriptional terms, than adult cancers of the same tissue. Future work may be able to exploit these differences for therapeutic or diagnostic purposes.

It should be recognised that the reference data used in various analyses may not be entirely inclusive and might be missing cell types necessary for interpreting specific samples. With the availability of more extensive single-cell datasets, we anticipate improved inferential precision, although, to our knowledge, such a resource is not yet at hand.

Conclusions

Our analyses reveal a nuanced picture of the differentiation state of adult cancers, wherein most adult tumours were best explained by post-natal cells, but not by a reversion to an antenatal state. By contrast, a “foetal-like” transcriptome was a near-universal feature of childhood tumours, likely as a consequence of their probable origins in development. Overall, these findings suggest that cancer transformation, at the level of the transcriptome, is a highly tissue and cancer-type-specific process, rather than a general hallmark of adult cancer.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13073-023-01279-z>.

Additional file 1: Table S1. Summary of all single cell transcriptomic datasets used in this study, along with the source from which they were obtained. **Table S2.** Summary of all bulk transcriptomic samples used in this study, along with the source from which they were obtained.

Table S3. Summary of results from clinical predictability of dedifferentiated signals both in TCGA & TARGET datasets.

Additional file 2: Fig. S1. Relationship between stemness and early embryo scores. **Fig. S2.** Cell signal analysis of all TCGA samples. **Fig. S3.** Cell signal analysis of all TARGET samples. **Fig. S4.** Cell signal analysis of all StJudes samples. **Fig. S5.** Cell signal analysis of all liver samples. **Fig. S6.** Cell signal analysis of all gut samples. **Fig. S7.** Survival prediction for TCGA and TARGET cancers

Acknowledgements

Not applicable.

Authors' contributions

M.D.Y., S.B., and I.C.-C conceived and supervised the project and wrote the manuscript. G.K., M.K., S.v.D., and M.D.Y. performed the analysis and data processing. C.T., A.P., K.A., E.P., and K.S. collected samples and generated data. S.T., M.H., and R.A. contributed to understanding the reference data. All authors read and approved the final manuscript.

Funding

Wellcome Trust (personal fellowship to S.B., institutional grant to the Wellcome Sanger Institute; references 220540/Z/20/A and 223135/Z/21/Z). M.K. and I.C.-C thank EMBL for funding.

Availability of data and materials

The data used in this study was obtained from the public sources indicated in Additional file 1: Table S1-2 and processed as described in the “Methods” section. Additionally, single-cell hepatoblastoma transcriptomes were generated and can be obtained from the EGA under accession number EGAS00001002325 [29] (<https://ega-archive.org/studies/EGAS00001002325>). The code used to process the data and generate these figures has been made available in a public repository [30] accessible here <https://github.com/constAntAmateur/dediffPaperCode>.

Declarations

Ethics approval and consent to participate

Written informed consent to participate in the study was obtained from participants (or their carers). Studies underlying this paper have received appropriate approval by ethics review boards as per national legislation. UK tumour samples were collected under the following study: National Health Service (NHS) National Research Ethics Service reference 16/EE/0394. Our research conformed to the principles of the Helsinki Declaration.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 23 January 2023 Accepted: 18 December 2023

Published online: 09 January 2024

References

- Miranda A, et al. Cancer stemness, intratumoral heterogeneity, and immune response across cancers. *Proc Natl Acad Sci.* 2019;116:9020–9.
- Shi X, et al. Cancer stemness associated with prognosis and the efficacy of immunotherapy in adrenocortical carcinoma. *Front Oncol.* 2021; 11: 651622
- Xiao L, et al. Alternative splicing associated with cancer stemness in kidney renal clear cell carcinoma. *BMC Cancer.* 2021;21:703.
- Malta TM, et al. Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell.* 2018;173:338–354.e15.
- AR, et al. The Human Cell Atlas. *eLife.* 2017; 6. <https://pubmed.ncbi.nlm.nih.gov/29206104/>.
- Young MD, et al. Single cell derived mRNA signals across human kidney tumors. *Nat Commun.* 2021;12:3896.
- Molè MA, et al. A single cell characterisation of human embryogenesis identifies pluripotency transitions and putative anterior hypoblast centre. *Nat Commun.* 2021;12:3679.
- Tyser RCV, et al. Single-cell transcriptomic characterization of a gastrulating human embryo. *Nature.* 2021;600:285–9.
- XH, et al. Construction of a human cell landscape at single-cell level. *Nature.* 2020; 581. <https://pubmed.ncbi.nlm.nih.gov/32214235/>.
- The tabula sapiens consortium. The Tabula Sapiens: a multiple-organ, single-cell transcriptomic atlas of humans. *Science.* 2022; 376: eabl4896.
- Downing JR, et al. The Pediatric Cancer Genome Project. *Nat Genet.* 2012;44:619–22.
- AB, PW, JCZ. SnapShot: TCGA-Analyzed Tumors. *Cell.* 2018;173. <https://pubmed.ncbi.nlm.nih.gov/29625059/>.
- Ho DW-H, et al. Single-cell RNA sequencing shows the immunosuppressive landscape and tumor heterogeneity of HBV-associated hepatocellular carcinoma. *Nat Commun.* 2021;12:3684.
- Lee H-O, et al. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nat Genet.* 2020;52:594–603.
- Sekiguchi M, et al. Integrated multiomics analysis of hepatoblastoma unravels its heterogeneity and provides novel druggable targets. *Npj Precis Oncol.* 2020;4:1–12.
- Druliner BR, et al. Early genetic aberrations in patients with sporadic colorectal cancer. *Mol Carcinog.* 2018;57:114–24.
- Hao Y, et al. Integrated analysis of multimodal single-cell data. *Cell.* 2021;184:3573–3587.e29.
- Collado-Torres L, et al. Reproducible RNA-seq analysis using recount2. *Nat Biotechnol.* 2017;35:319–21.
- Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods.* 2017;14:417–9.
- Young MD, et al. Single cell derived mRNA signals across human kidney tumors. *Nat Commun.* 2021;12:1–19.
- Dominguez Conde C, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science.* 2022; 376: eabl5197.
- Kagawa H, et al. Human blastoids model blastocyst development and implantation. *Nature.* 2022;601:600–5.
- Bialecki ES, Di Bisceglie AM. Diagnosis of hepatocellular carcinoma. *HPB.* 2005;7:26–34.
- Popescu D-M, et al. Decoding human fetal liver haematopoiesis. *Nature.* 2019;574:365–71.
- MacParland SA, et al. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat Commun.* 2018;9:4383.

26. Segal JM, et al. Single cell analysis of human foetal liver captures the transcriptional profile of hepatobiliary hybrid progenitors. *Nat Commun.* 2019;10:3350.
27. Zhang J, et al. The threshold of alpha-fetoprotein (AFP) for the diagnosis of hepatocellular carcinoma: A systematic review and meta-analysis. *PLoS ONE.* 2020;15: e0228857.
28. Barker N, et al. Crypt stem cells as the cells-of-origin of intestinal cancer. *Nature.* 2009;457:608–11.
29. Kildisiute G, Kalyva M, Elmentaite R, van Dongen S, Thevanesan C, Piapi A, Ambridge K, Prigmore E, Haniffa M, Teichmann SA, Straathof K, Cortés-Ciriano I, Behjati S, Young MD. Transcriptional signals of transformation in human cancer. EGAS00001002325, European Genome-Phenome Archive. 2023. <https://ega-archive.org/studies/EGAS00001002325>.
30. Kildisiute G, Kalyva M, Elmentaite R, van Dongen S, Thevanesan C, Piapi A, Ambridge K, Prigmore E, Haniffa M, Teichmann SA, Straathof K, Cortés-Ciriano I, Behjati S, Young MD. Transcriptional signals of transformation in human cancer. github. 2023. <https://github.com/constantAmateur/dediffPaperCode>.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

