Title: Identification of cancer cell states associated with patient survival at high resolution via combinatorial gene expression dependencies

Abstract:

To design more effective cancer prevention strategies and improve prognostic accuracy, we need a deeper understanding of the mechanisms driving tumour initiation and the evolution of both cancer cells and their surrounding microenvironment. Previously, we have demonstrated that cancer cell states associated with poor prognosis can be identified at high resolution using single cell RNA-sequencing (scRNAseq) data from established tumours, via structure learning and quantification of higher-order gene expression dependencies. Recently, we conducted a time-resolved scRNAseqing analysis of preneoplastic cells (PNCs) and associated innate immune cells within 24 hours of oncogene activation in a zebrafish model. This led to the identification of an EMT/CSC-like PNC cluster, with its marker genes predicting poor prognosis in several carcinomas in the TCGA database. Furthermore, this cluster appears to drive tumour-promoting neutrophil development, suggesting its critical role in malignant progression. In the current project, using our zebrafish dataset as the initial state of tumour initiation, we aim to map cellular states onto single-cell RNAseq and spatial transcriptomic datasets from a wider range of cancer patients and mammalian models. We will employ our recently developed computational tool, Stator, and test publicly available data integration tools such as SATURN to investigate how the EMT/CSC-like state, identified at the onset of oncogene activation, evolves throughout cancer development and how host immune cells might co-evolve in this process. Our goal is to uncover the mechanisms that either promote or suppress the EMT/CSC state, thereby identifying novel targets for cancer prevention. We will seek to discover potential biomarkers to enhance early detection and improve prognosis. Additionally, we will develop a software and user-friendly visualisation tool for identification of cancer cell states associated with prognosis from scRNA-seq.

Introduction

The best way to treat cancer is through early detection and prevention. Therefore, a better understanding of the key mechanisms driving tumour initiation and progression is crucial for the development of strategies for early detection and cancer prevention. Inflammation is known to play various promotive roles during cancer development. Specifically, innate immune cells can adopt a tumour-promoting phenotype, supporting cancer cell proliferation and progression within the tumour microenvironment. However, the mechanisms mediating the recruitment of innate immune cells to tumours, their phenotypic switching during tumour development, and how these events affect prognosis, remain to be elucidated. Cells carrying oncogenic mutations acquire phenotypic plasticity, which is one of the key events driving tumour progression. However, when and how this occurs during tumour initiation and progression remain largely unknown.

Previously, we have demonstrated that cancer cell states associated with poor prognosis can be identified at high resolution using single cell RNA-sequencing data, via structure learning and quantification of higher-order gene expression dependencies. However, a more targeted identification of such cells states, including information on the mutations present

and their interaction with the immune microenvironment will provide additional mechanistic insights and thus offers informed potential therapeutic interventions. This requires an integrative pan-cancer analysis of cancer cells states in human tissues and animal models. This is because through the latter, cancer mutations and immune interactions leading to these states can be both induced and validated in a well-controlled environment.

Zebrafish has recently emerged as an important model organism in cancer biology. Its optical translucency and genetic tractability enable detailed live imaging of tumour cell and host innate immune cell interactions during the earliest stages of tumourigenesis in vivo. The Feng lab has developed a tamoxifen-inducible preneoplastic cell (PNC) development model in zebrafish larval skin. In this model, we observed a rapid neutrophil response to developing cancer cell precursors (preneoplastic cells) at their inception. Importantly, we demonstrated that neutrophils play a role in promoting PNC proliferation. This mirrors tumour promoting neutrophils identified in tumour microenvironment in human cancers.

Recently, we conducted a time-course single-cell RNA sequencing analysis of FACS-isolated PNCs and innate immune cells. In our analysis, we identified a unique PNC cluster undergoing mesenchymal transition (EMT) and exhibiting cancer stem cell (CSC) features. Crucially, we found that the EMT/CSC-like PNC signature predicts poor prognosis across multiple cancer types in the TCGA database, suggesting that the persistence of this cellular state is associated with worse disease progression. In addition, we observed that a moderate neutrophil presence correlates with poor prognosis in the TCGA dataset, indicating that heterogeneous neutrophil phenotypes may exist within the tumour microenvironment.

Furthermore, our dataset revealed that from inception, PNCs first undergo a rapid dedifferentiation and a bifurcate phenotypic transition, either towards a more differentiated state or the EMT/CSC state, suggesting a possible intervention point for blocking tumour progression. Notably, EMT/CSC-like cells express elevated levels of factors known to modulate neutrophil development and function, corresponding to the tumour-promoting neutrophil phenotype observed in this zebrafish tumour initiation model. Further support the possible role of EMT/CSC PNC cluster in driving tumour progression.

Importantly, the cellular states observed during the earliest stages of tumourigenesis in our model are also present in established cancers in humans. This highlights the need for a deeper understanding of the co-evolution between cancer cells and their microenvironment, as well as how host tissue responses either promote or constrain tumour progression.

Research Challenge

We aim to identify cancer cell states associated with poor/good survival, across various cancer types in humans and animal models, with a focus on the interactions between cells carrying mutations and their immune microenvironment. Specifically, to trace the evolution of PNCs and associated innate immune cells, we plan to leverage publicly available datasets from cancer patients and mammalian models. By using our zebrafish dataset as an initial

state, we aim to map cellular states from our dataset onto single-cell RNAseq and spatial transcriptomic datasets from a broader range of cancer patients or mammalian models. This approach will allow us to determine how the EMT/CSC-like state we identified at the onset of oncogene activation evolves during cancer development. We aim to establish the mechanisms that either promote or suppress the EMT/CSC state, thereby providing novel targets for cancer prevention. Additionally, we will identify potential biomarkers to improve early detection and prognosis.

Data & Methodology

Data: We currently possess a time-course zebrafish scRNAseq dataset, derived from RAS-driven basal skin keratinocyte preneoplastic cells and associated macrophages and neutrophils, covering 24 hours of RAS activation. We also have access to scRNAseq datasets from human cutaneous squamous cell carcinoma (SCC), provided by Dr Gareth Inman, as well as publicly available datasets from the GEO database.

Methodology: We will use Stator, a ML methodology and software that quantifies cell types and states at high resolution, using structure learning and combinatorial gene expression dependence. Stator does not rely on clustering of cell in expression space, and instead allows cell to be multiply labelled by cell type, subtype and multiple biological states (e.g. cell cycle subphase, activation, …). We will develop Stator further for cross-species analyses as part of this project. Additionally, we will utilise various existing cross-species data integration tool, such as SATURN, to identify common cancer initiating cell states across different tissue and species. We will compare these results with Stator's cross-species state analysis.

Responsible AI/Ethical Considerations

Produced code and analyses will be made publicly available on GitHub. In compliance with the 3R principle, our zebrafish single-cell RNAseq data was generated using larvae less than 5 days old, which are not classified as protected animals. All sequencing data used in this project are either available in the GEO database or will be uploaded to the GEO database upon publication.

Expected Outcome & Impact

We aim to determine how cells with EMT/CSC features evolve throughout cancer development and to identify the key mechanisms by which these cells contribute to tumour progression and cancer prognosis. This research could lead to the identification of novel intervention targets or biomarkers for disease prognosis. We will disseminate our findings and the corresponding computational methods through publications and open-source software.

In addition, in collaboration with Genenet Technology (uk) Limited, we will create an industry-quality version of the software tool for performing integrative cancer cell state quantification and prognostic analysis, as well as a user-friendly visual interface. We will

further explore whether this software and visualisation tool can be commercialised as joint IP between UoE and Genenet, in conversation with Edinburgh Innovations.

References

Elliot A. M., Ribeiro Bravo I., Astorga Johansson J., Hutton E., Cunningham R., Myllymäki H., Chang K. Y., Cholewa-Waclaw J., Zhao Y., Beltran M., Dobie R., Lewis A., Elks P. M., Hansen C., Henderson N., **Feng Y*.** Oncogenic Ras activation in permissive somatic cells triggers rapid onset phenotypic plasticity and elicits a tumour-promoting neutrophil response. **BioRxiv** 2023.11.10.566547; doi: https://doi.org/10.1101/2023.11.10.566547 (updated version 2, 2024)

Rosen, Y., Brbić, M., Roohani, Y. *et al.* Toward universal cell embeddings: integrating single-cell RNA-seq datasets across species with SATURN. *Nat Methods* **21**, 1492–1500 (2024). https://doi.org/10.1038/s41592-024-02191-z

Abel Jansma, Yuelin Yao, Jareth Wolfe, Luigi Del Debbio, Sjoerd Beentjes, Chris P. Ponting, **Ava Khamseh*** High order expression dependencies finely resolve cryptic states and subtypes in single cell data, Accepted in EMBO Molecular System Biology, bioRxiv 2023.12.18.572232; doi: https://doi.org/10.1101/2023.12.18.572232