From the archive

Early efforts to understand hay fever, and remembering an astronomer who made a discovery about Venus.

100 years ago

Hay fever is a catarrhal affection mainly of the upper respiratory passages ... Our knowledge of the disease may be divided into three periods .... clinical, experimental, and therapeutic. The clinical period was inaugurated by John Bostock, the physiologist, in 1819. He gave an admirable account of the symptoms as they occurred in himself every summer over a period of thirty-eight years ... The second period is typified by the masterly scientific research of Dr. Charles H. Blackley, of Manchester ... To study the distribution of pollen in the atmosphere Blackley ... undertook a long series of experiments with ingenious apparatus which he devised ... From May to the end of July he traced from day to day the pollen incidence in the air, and showed how it was subject to great fluctuations depending on temperature and moisture ... Blackley was so far ahead of his time, that despite the excellence of his work the causation of hay fever was still regarded as a *terra incognita*.

From *Nature* 16 June 1923

150 years ago

If national glory can ever be connected with a natural phenomenon, the transit of Venus over the sun's disc may be said to bring peculiar distinction to England ... [T]he phenomenon ... was for the first time in human history, accurately observed in a corner of England, by an English youth, self-taught, and provided with few of the appliances of scientific research. Now that the spectacle, so striking in itself, so sublime in the infrequent regularity of its recurrence, so important as the key to numerous astronomical problems, is again attracting ... attention ..., now that the expanse of ocean from Honolulu to Kerguelen's Land is about to be dotted with watchers from the other side of the earth, the occasion appears favourable for recalling the memory of the original observer, Jeremiah Horrox, curate of Hoole, near Preston.

From *Nature* 12 June 1873

Cancer

Patterns of tumour transcriptional variability

Raymond W. S. Ng & Sydney M. Shaffer

The compilation and analysis of a compendium of single-cell RNA-sequencing studies across various cancers reveals recurring gene-expression programs that underpin tumour heterogeneity. **See p.598**

Cancer is a highly variable disease, and understanding its intricacies might help to improve personalized (precision-medicine) therapeutic approaches. On page 598, Gavish *et al.*[1] shed light on the patterns of gene-expression variability found in tumours.

The sequencing of bulk RNA samples enables researchers to characterize tumours on the basis of gene expression. However, this method pools all the cells in a tumour for analysis, which masks the heterogeneity present at the single-cell level. The development of single-cell RNA sequencing (scRNA-seq) technology provides gene-expression data at an unprecedented level of detail through the sequencing of RNA from individual cells.

This scRNA-seq technology has revealed that, in addition to the variability between different tumours (intertumour heterogeneity), there is a tremendous amount of variation in gene expression between cells in the same tumour (intratumour heterogeneity). Individual cells in a tumour can be classified on the basis of their gene expression, and coordinated patterns of expression can be identified. Some such patterns of expression might correspond to those in a subset of tumour cells that have different functions or behaviours from other tumour cells.

Several studies[2–4] using scRNA-seq have characterized the variable expression programs in tumours for a single type of cancer. However, it was unclear whether these programs were cancer-type specific or were present in multiple cancers (possibly reflecting fundamental aspects of tumour biology). Addressing this requires scRNA-seq data from a large number of cells across multiple types of cancer. A 2022 study[5] using scRNA-seq to examine 62 tumour samples across 15 cancer types revealed 16 expression programs that were a general feature of all the cancer types. These programs were classified into named groups, such as 'stress', 'interferon response' and 'epithelial–mesenchymal transition (EMT)', with these data indicating that recurrent tumour gene-expression programs can exist across different types of cancer. Gavish and colleagues build on this work by providing a much larger scRNA-seq data set and by further characterizing gene-expression programs.

The authors have curated and analysed one of the largest compendiums of cancer-specific scRNA-seq data available so far, encompassing 1,163 examined samples across 24 cancer types and more than 2.5 million cells. The authors assigned cell types on the basis of gene expression and distinguished malignant cells from neighbouring non-cancer cells by identifying as cancer cells those that had alterations in the number of copies of genes (copy-number alterations). To aid future discoveries, the authors built a website to allow other researchers to use these data and to easily download them for their own analyses (see go.nature.com/3wvzbp8).

In their effort to capture gene-expression variability in tumours, Gavish and colleagues

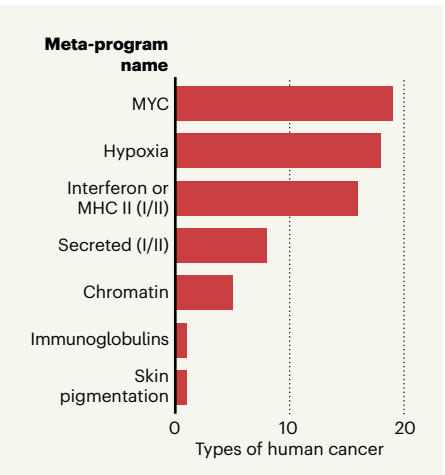

Figure 1 | Gene-expression programs identified in human cancer. Gavish *et al.*[1] examined data from RNA sequencing of single cells in 24 types of cancer. The authors found that these gene-expression patterns could be grouped into 41 distinct clusters, termed meta-programs, some named examples of which are shown here (including some combined meta-programs). Some meta-programs were expressed broadly across different types of tumour, whereas others were specific to particular tumour types.

© 2023 Springer Nature Limited. All rights reserved.

diverged twice from the typical analysis pipelines that are used to identify intertumour heterogeneity. First, the authors opted not to perform a procedure called batch correction, which is intended to minimize variability that can arise from the sequencing technique itself, because this correction method can reduce the capture of variability that arises naturally in a biological sample. Second, instead of defining gene-expression patterns in a combined data set involving all tumours, the authors initially defined variable gene-expression programs in each tumour sample and then compared the programs found across the samples.

Among the cancer cells, the authors found more than 5,500 variable gene-expression programs that were robustly observed across tumours. The authors grouped the programs into 41 clusters, which they defined as meta-programs (MPs). For each cluster, Gavish *et al.* identified 50 'consensus' genes, which provided a signature of the MP. Of the 41 MPs, 16 were similar to ones described previously[5]. The authors also applied this analysis to non-cancer cells in the tumour microenvironment and to cancer cell lines grown *in vitro*.

The authors found that MPs in cancer cells recur across a range of cancer types (Fig. 1). To identify the fundamental biological processes that these MPs represent, Gavish and colleagues categorized them into 11 hallmark categories, such as 'cell cycle', 'mesenchymal' and 'senescence'. These groupings indicate that tumours are composed of subpopulations of cancer cells that might drive different aspects of tumour progression, such as its growth, its spread to other sites (metastasis) and resistance to drug treatment. Characterizing these subpopulations and their vulnerabilities for each tumour might therefore be biologically informative and clinically useful.

Gavish *et al.* discovered that the identity of MPs in non-cancer cells helps to explain some of the MPs present in cancer cells. They observed that several MPs in cancer cells were also present in their non-cancer counterparts, suggesting that much of the heterogeneity seen in cancer cells is a combination of fundamental cellular heterogeneity and heterogeneity that arises during tumour formation. Interestingly, this variability was independent of the degree of genetic diversity of the cancer cells. Moreover, the authors identified co-occurrences of various MPs in different cell types, which might indicate that evolutionarily conserved factors drive MPs in both cancer cells and other cells in the tumour microenvironment.

The data also highlight the usefulness of cell lines for studying intratumour heterogeneity. Gavish and colleagues discovered that although cell lines do not recapitulate the full extent of heterogeneity in tumour samples from patients, data from cell lines nevertheless capture a subset of the MPs and certain hallmark-category processes. This result suggests that cell lines can be valuable tools for exploring certain aspects of tumour heterogeneity and for advancing precision medicine. Moreover, the research provides specific information about which variable features of tumours are captured by the cell lines.

The authors grouped together many gene-expression programs to find recurring MPs. However, a limitation of this approach is that it might miss some biologically meaningful complexity, such as gene-expression programs that involve only a few genes or that exist in rare cells. Furthermore, scRNA-seq data cannot capture sources of heterogeneity beyond the level of messenger RNA abundance, such as variability in the amount of protein produced. Nonetheless, the authors have greatly advanced our understanding of how gene-expression programs are patterned in tumours and across cancers.

Gavish and colleagues open up several exciting prospects for future areas of cancer research. As the size of this scRNA-seq compendium grows, it might be possible to define MPs in more cancer contexts or in rare types of cancer cell. The emergence of data sets that capture RNA-sequencing data together with information about the spatial context of cells in tissues will offer a way to further validate the cell–cell interactions that occur in the tumour microenvironment and that determine the extent to which the heterogeneity of hallmark processes is organized in specific locations. Future scRNA-seq experiments, paired with more data, might shed light on the determinants of transcriptional heterogeneity (for example, through the use of single-cell 'multi-omics' analysis), and examining clinical data could offer insights into the functional outcomes of this heterogeneity.

The evolutionarily conserved pathways described by Gavish and colleagues present opportunities to identify vulnerabilities across different cancers that could be exploited for therapeutic purposes. Future studies in this area could increase our understanding of how precision medicine should take intratumour heterogeneity into account.

Raymond W. S. Ng and **Sydney M. Shaffer** are in the Department of Pathology and Laboratory Medicine and the Department of Bioengineering, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.
e-mail: sydshaffer@gmail.com

1. Gavish, A. *et al. Nature* **618**, 598–606 (2023).
2. Couturier, C. P. *et al. Nature Commun.* **11**, 3406 (2020).
3. Raghavan, S. *et al. Cell* **184**, 6119–6137 (2021).
4. Jerby-Arnon, L. *et al. Cell* **175**, 984–997 (2018).
5. Barkley, D. *et al. Nature Genet.* **54**, 1192–1201 (2022).

The authors declare no competing interests.
This article was published online on 31 May 2023.

Artificial intelligence

Online tools help language models to solve problems

Aleksandra Piktus

The large language models popularized by chatbots are being taught to alternate reasoning with calls to external tools, such as Wikipedia, to boost their accuracy. The strategy could improve fact-finding outcomes, as well as online shopping.

As the success of ChatGPT shows, large language models (LLMs) are becoming ever easier to use. The power of these systems lies in their ability to respond to textual prompts by generating natural-sounding language. But what makes them even more effective is their ability to learn by example: by taking in just a few demonstrations, LLMs can markedly improve their performance on complex tasks[1]. Precisely how to formulate these examples to elicit accurate answers is an open problem, but researchers have some ideas. Reporting at the Eleventh International Conference on Learning Representations (go.nature.com/42qwbwg), Yao *et al.*[2] propose ReAct — a prompting strategy that improves on existing methods by breaking down a multifaceted reasoning problem into bite-sized tasks and outsourcing them to external tools.

The name ReAct refers to Yao and colleagues' integration of an 'acting' step into the process of solving a problem through reasoning. This step gives LLMs a way to interact with other tools, such as Wikipedia, by executing programming requests known as API (application programming interface) calls. The authors incorporated this step into an existing strategy called chain-of-thought prompting[3], in which tasks requiring complex reasoning are decomposed into more-granular steps.

ReAct prompts LLMs to work through a trajectory of reasoning tasks and 'actions' to solve a given problem (Fig. 1). A reasoning step generates text, which Yao and colleagues call

© 2023 Springer Nature Limited. All rights reserved.