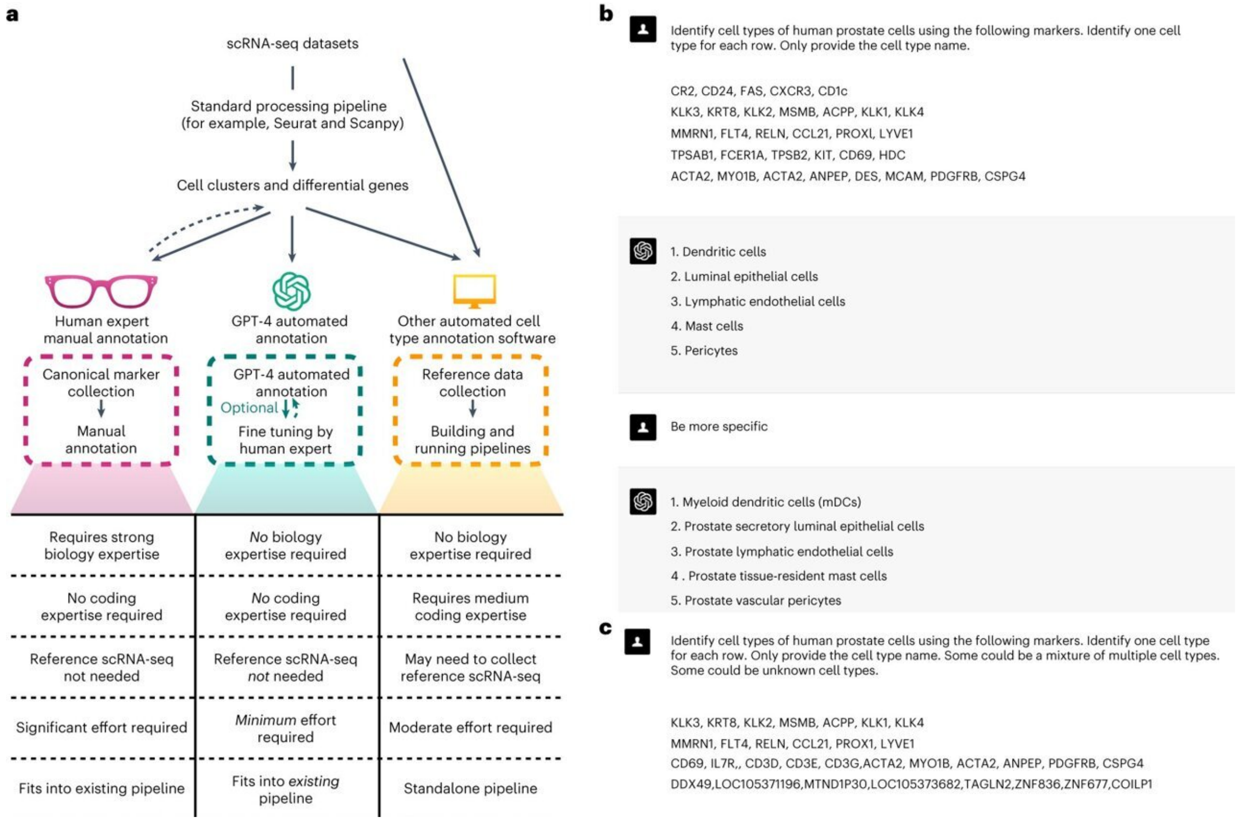
PHYS.ORG

GPT-4 for identifying cell types in single cells matches and sometimes outperforms expert methods

March 25 2024



Requires strong biology expertise	*No* biology expertise required	No biology expertise required
No coding expertise required	*No* coding expertise required	Requires medium coding expertise
Reference scRNA-seq not needed	Reference scRNA-seq *not* needed	May need to collect reference scRNA-seq
Significant effort required	*Minimum* effort required	Moderate effort required
Fits into existing pipeline	Fits into *existing* pipeline	Standalone pipeline

Identify cell types of human prostate cells using the following markers. Identify one cell type for each row. Only provide the cell type name.

CR2, CD24, FAS, CXCR3, CD1c
KLK3, KRT8, KLK2, MSMB, ACPP, KLK1, KLK4
MMRN1, FLT4, RELN, CCL21, PROXl, LYVE1
TPSAB1, FCER1A, TPSB2, KIT, CD69, HDC
ACTA2, MY01B, ACTA2, ANPEP, DES, MCAM, PDGFRB, CSPG4

1. Dendritic cells
2. Luminal epithelial cells
3. Lymphatic endothelial cells
4. Mast cells
5. Pericytes

Be more specific

1. Myeloid dendritic cells (mDCs)
2. Prostate secretory luminal epithelial cells
3. Prostate lymphatic endothelial cells
4 . Prostate tissue-resident mast cells
5. Prostate vascular pericytes

Identify cell types of human prostate cells using the following markers. Identify one cell type for each row. Only provide the cell type name. Some could be a mixture of multiple cell types. Some could be unknown cell types.

KLK3, KRT8, KLK2, MSMB, ACPP, KLK1, KLK4
MMRN1, FLT4, RELN, CCL21, PROX1, LYVE1
CD69, IL7R,, CD3D, CD3E, CD3G,ACTA2, MYO1B, ACTA2, ANPEP, PDGFRB, CSPG4
DDX49,LOC105371196,MTND1P30,LOC105373682,TAGLN2,ZNF836,ZNF677,COILP1

Examples of GPT-4's cell type annotation and comparisons with other methods. Credit: *Nature Methods* (2024). DOI: 10.1038/s41592-024-02235-4

GPT-4 can accurately interpret types of cells important for the analysis of single-cell RNA sequencing—a sequencing process fundamental to interpreting cell types—with high consistency to that of time-consuming manual annotation by human experts of gene information, according to a study at Columbia University Mailman School of Public Health. The findings are published in the journal *Nature Methods*.

GPT-4 is a large language model designed for speech understanding and generation. Upon assessment across numerous tissue and cell types, GPT-4 has demonstrated the ability to produce cell type annotations that closely align with manual annotations of human experts and surpass existing automatic algorithms.

This feature has the potential to significantly lessen the amount of effort and expertise needed for annotating cell types, a process that can take months. Moreover, the researchers have developed GPTCelltype, an R software package, to facilitate the automated annotation of cell types using GPT-4.

"The process of annotating cell types for single cells is often time-consuming, requiring human experts to compare genes across cell clusters," said Wenpin Hou, Ph.D., assistant professor of Biostatistics at Columbia Mailman School.

"Although automated cell type annotation methods have been developed, manual methods to interpret scientific data remain widely used, and such a process can take weeks to months. We hypothesized that GPT-4 can accurately annotate cell types, transitioning the process from manual to a semi- or even fully automated procedure and be cost-efficient and seamless."

The researchers assessed GPT-4's performance across ten datasets covering five species, hundreds of tissue and cell types, and including both normal and cancer samples. GPT-4 was queried using GPTCelltype, the software tool developed by the researchers. For competing purposes, they also evaluated other GPT versions and manual methods as a reference tool.

As a first step, the researchers first explored the various factors that may affect the annotation accuracy of GPT-4. They found that GPT-4 performs best when using the top 10 different genes and exhibits similar accuracy across various prompt strategies, including a basic prompt strategy, a chain-of-thought-inspired prompt strategy that includes reasoning steps, and a repeated prompt strategy. GPT-4 matched manual analyses in over 75% of cell types in most studies and tissues demonstrating its competency in generating expert-comparable cell type annotations.

In addition, the low agreement between GPT-4 and manual annotations in some cell types does not necessarily imply that GPT-4's annotation is incorrect. In an example of stromal or connective tissue cells, GPT-4 provides more accurate cell type annotations. GPT-4 was also notably faster.

Hou and her colleague also assessed GPT-4's robustness in complex real data scenarios and found that GPT-4 can distinguish between pure and mixed cell types with 93% accuracy, and differentiated between known and unknown cell types with 99% accuracy. They evaluated the performance of reproducing GPT-4's methods using prior simulation studies. GPT-4 generated identical notations for the same marker genes in 85% of cases.

"All of these results demonstrate GPT-4's robustness in various scenarios," observed Hou.

While GPT-4 surpasses existing methods, there are limitations to consider, according to Hou, including the challenges for verifying GPT-4's quality and reliability because it discloses little about its training proceedings.

"Since our study focuses on the standard version of GPT-4, fine-tuning GPT-4 could further improve cell type annotation performance," said Hou.

Zhicheng Ji of Duke University School of Medicine is a co-author.

More information: Wenpin Hou et al, Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis, *Nature Methods* (2024). DOI: 10.1038/s41592-024-02235-4

Provided by Columbia University's Mailman School of Public Health

Citation: GPT-4 for identifying cell types in single cells matches and sometimes outperforms expert methods (2024, March 25) retrieved 27 April 2024 from https://phys.org/news/2024-03-gpt-cell-cells-outperforms-expert.html