Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Plant Biology

Advancing plant single-cell genomics with foundation models

Tran N. Chau[1,2], Xuan Wang[3], John M. McDowell[2] and Song Li[1,2,3]

Single-cell genomics, combined with advanced AI models, hold transformative potential for understanding complex biological processes in plants. This article reviews deep-learning approaches in single-cell genomics, focusing on foundation models, a type of large-scale, pretrained, multi-purpose generative AI models. We explore how these models, such as Generative Pre-trained Transformers (GPT), Bidirectional Encoder Representations from Transformers (BERT), and other Transformer-based architectures, are applied to extract meaningful biological insights from diverse single-cell datasets. These models address challenges in plant single-cell genomics, including improved cell-type annotation, gene network modeling, and multi-omics integration. Moreover, we assess the use of Generative Adversarial Networks (GANs) and diffusion models, focusing on their capacity to generate high-fidelity synthetic single-cell data, mitigate dropout events, and handle data sparsity and imbalance. Together, these AI-driven approaches hold immense potential to enhance research in plant genomics, facilitating discoveries in crop resilience, productivity, and stress adaptation.

Addresses
[1] Genetics, Bioinformatics, and Computational Biology, Virginia Tech, USA
[2] School of Plant and Environmental Sciences, Virginia Tech, USA
[3] Department of Computer Science, Virginia Tech, USA

Corresponding author: Li, Song (songli@vt.edu)

Current Opinion in Plant Biology 2024, **82:**102666

This review comes from a themed issue on **Genome studies and molecular genetics 2024**

Edited by **Leena Tripathi** and **Sushma Naithani**

For a complete overview see the Issue and the Editorial

Available online xxx

https://doi.org/10.1016/j.pbi.2024.102666

Advancements and challenges of single-cell data in plants

Single-cell RNA sequencing (scRNA-seq) capabilities have grown exponentially over the past decade, significantly increasing the number of cells that can be processed within a single experiment [1]. This technology advances agriculture by uncovering key genes and cellular processes across plants. It reveals genes for nitrogen fixation in soybean nodules [2], identifies suberin regulation for drought tolerance in tomato and *Arabidopsis* [3], informs leaf development patterns in *Brassica rapa* [4], enhances yield strategies by elucidating maize ear development [5], and uncovers xylem diversity in woody plants [6]. These discoveries have the potential to drive improvements in crop resilience and productivity.

However, plant-specific challenges persist, such as rigid cell walls that complicate cell isolation and protoplast generation. The diversity in cell sizes and types in plant tissues further hinders efficient capture and comprehensive representation in experiments. Additionally, plant single-cell genomics faces unique hurdles including (1) limited datasets (as compared to human/mouse), (2) high data sparsity with many zero measurements, (3) under-represented cell types in complex tissues, and (4) complexity in determining orthologous genes across species.

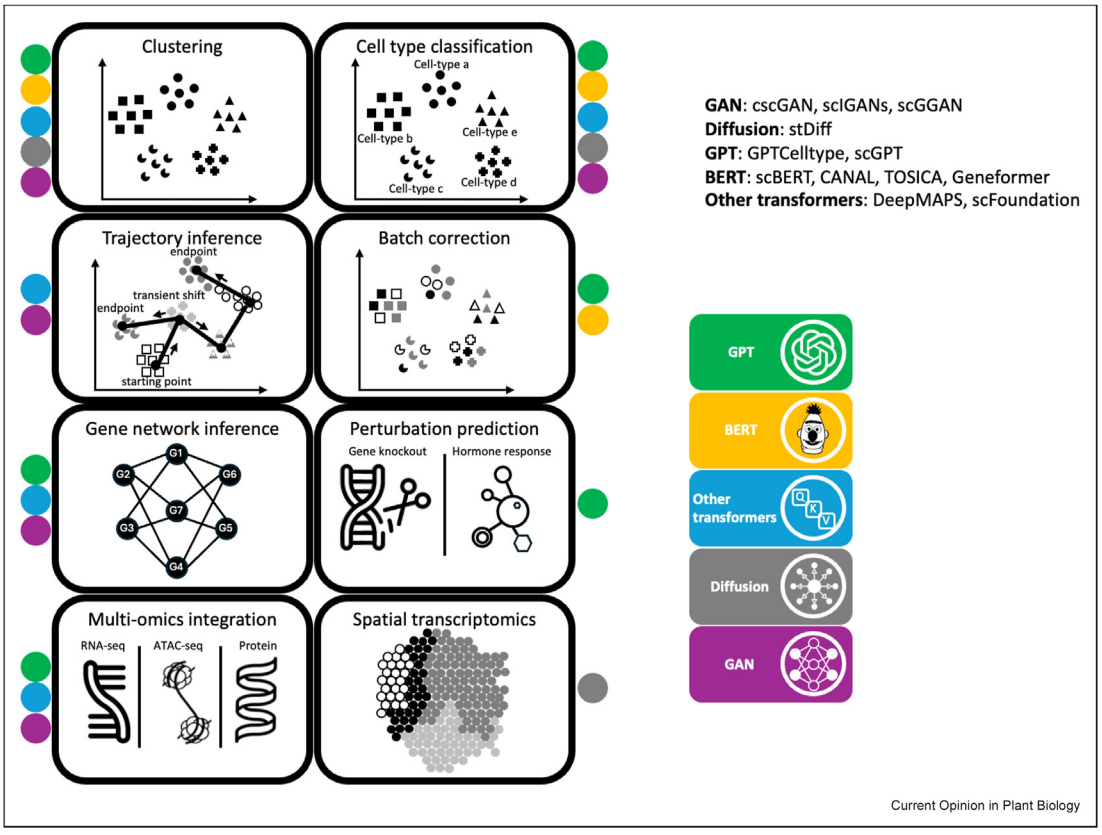
Addressing the hurdles through foundation models

To overcome these challenges, researchers are turning to foundation models - advanced deep learning frameworks trained on vast datasets that can be fine-tuned for a wide range of tasks [7]. These models act as adaptable platforms for AI applications, offering scalability and flexibility across multiple domains, including plant biology [7]. Although training foundation models requires large datasets, significant computational resources, and specialized expertise, their versatility makes them transformative once established [8]. By adapting successful AI models from biomedical research (Figure 1), pre-trained on extensive datasets, the transformative potential of foundation models can be harnessed to address the current gaps in plant genomics, offering more precise and comprehensive solutions to complex plant data analysis challenges.

Leveraging various single-cell analysis tasks with GPT

GPT (Generative Pre-trained Transformer) models, which are foundation models and Large Language

Figure 1



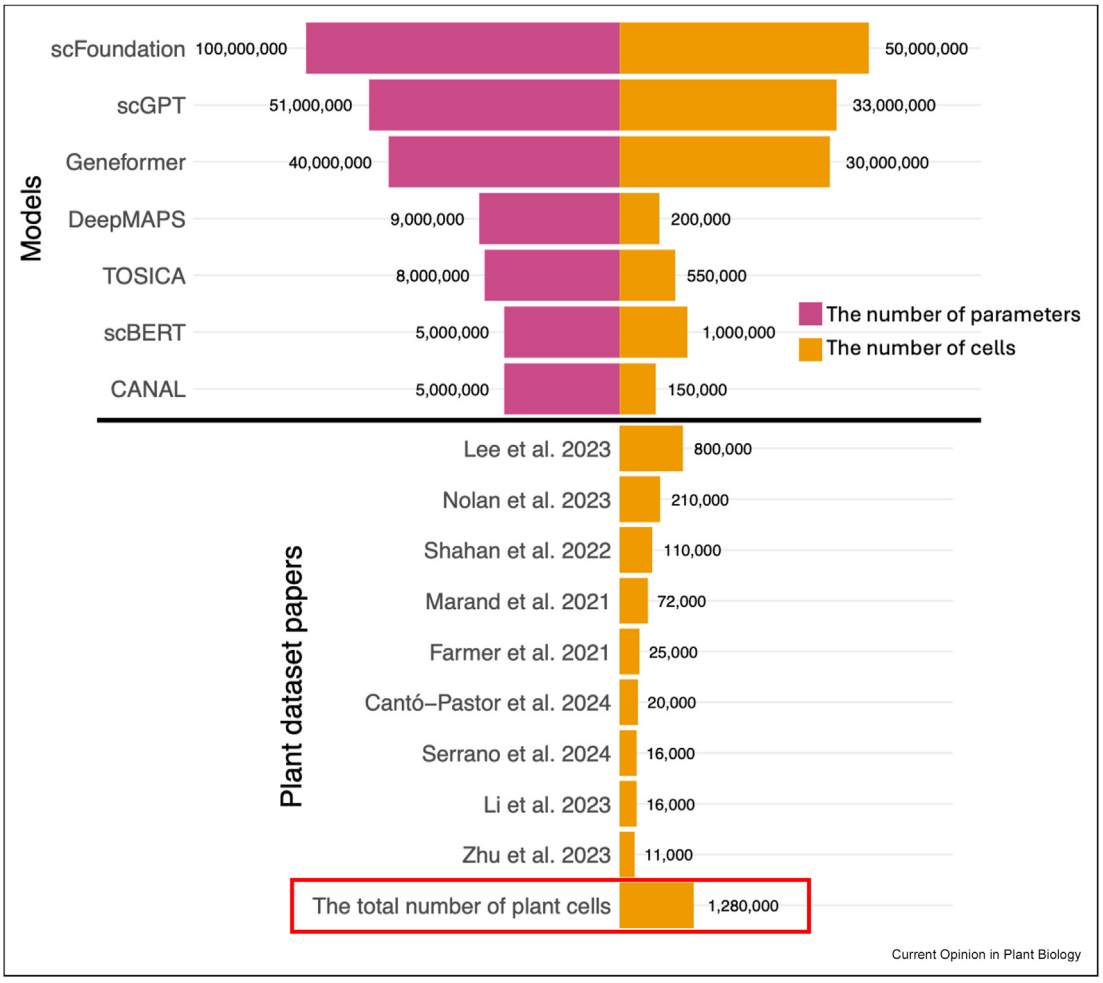
This figure illustrates the applications of GPT (GPTCelltype, scGPT), BERT (scBERT, CANAL, TOSICA, Geneformer), other Transformer models (DeepMAPS, scFoundation), GAN (cscGAN, scIGAN, scGGAN), and diffusion models (stDiff) in various single-cell analysis tasks. These tasks include clustering (grouping cells based on similar gene expression), cell type classification (identifying cell types based on gene markers), trajectory inference (tracking the progression of cells through different states), batch correction (removing technical variation between datasets), gene network inference (revealing interactions between genes), perturbation prediction (predicting how gene expression changes under different conditions), multi-omics integration (combining different types of molecular data for a comprehensive view of cellular processes), and spatial transcriptomics (mapping gene expression to specific locations within tissues, providing spatial context). Each model - represented by a specific color: green for GPT, yellow for BERT, blue for other Transformers, purple for GANs, and gray for diffusion models - is marked with a dot next to each task box to indicate its applicability. The figure summarizes model applications as described in their original papers.

Models (LLMs), use a decoder-only transformer architecture. This architecture generates data autoregressively, meaning it predicts each element based on previous ones, which enhances interpretability. In single-cell analysis, GPT models are used for various tasks including integration, synthetic data generation, and annotation (Figure 1). These models have been adapted for single-cell applications through implementations like GPTCelltype and scGPT.

GPTCelltype [9] leverages GPT-4's capacity of understanding natural language to automatically annotate cell types based on marker gene names and biomedical text. It integrates into single-cell analysis workflows like Seurat via an R package, enabling interactive annotation refinement for cell clusters. In contrast, scGPT [10] is designed to handle numerical gene expression data from scRNA-seq datasets. Pre-trained on 33 million human cells (Figure 2), it addresses various tasks such as cell-type annotation, multi-batch and multi-omics integration, genetic perturbation prediction, and gene network inference. With its decoder-only transformer architecture, scGPT generates synthetic gene expression profiles by making structured and biologically meaningful predictions based on previous elements in the dataset, rather than arbitrary outputs, reducing the risk of hallucinations seen in free-form text models like ChatGPT.

In plant research, GPTCelltype can be adapted by fine-tuning, which involves retraining the pre-trained model on plant-specific texts like botanical literature, plant gene databases, or plant gene markers. It can handle text-based biological knowledge, useful for annotating plant cell types based on gene markers and literature. On the other hand, scGPT requires fine-tuning on large plant scRNA-seq datasets to capture plant-specific gene

Figure 2



This figure displays the number of parameters in the models, represented by the pink bar chart on the left, and the number of cells in the models, shown by the orange bar chart on the right, both above the horizontal line. Below the horizontal line, the bar chart illustrates the number of plant cells in each plant dataset paper, with the red box highlighting the total number of plant cells used in all these studies.

expression patterns. Fine-tuning is more efficient than building a model from scratch, as it leverages the general patterns already captured by the pre-trained model. It adjusts the existing architecture to incorporate plant-specific features, enhancing performance on plant-related tasks. However, scGPT's large size (51M parameters) and complexity make it computationally demanding, requiring significant resources and extensive fine-tuning for specialized tasks.

Annotating single-cell cluster with BERT

BERT (Bidirectional Encoder Representations from Transformers) models use an encoder-only transformer architecture, focusing on processing and understanding existing data rather than generating new outputs. Unlike GPT, which uses autoregressive pretraining, BERT employs masked language model pretraining, learning bidirectionally to produce representations or classifications (Table 1).

In single-cell analysis, models like scBERT [11] enhance cell-type annotation in scRNA-seq data by leveraging this encoder-only architecture. It involves self-supervised pretraining to learn gene interaction patterns from large-scale data, followed by supervised fine-tuning for cell type prediction. scBERT can handle numerous gene inputs without relying on highly variable genes or dimensionality reduction, maintaining gene-level resolution and interpretability. However, Khan and colleagues [12] highlight scBERT's limitations with imbalanced cell-type distributions and suggest a subsampling technique to address these limitations.

Adapting scBERT model, CANAL [13] is designed to continually annotate cell types in scRNA-seq data. To

Table 1

Definitions.

Model Type	Architecture	Training Approach	Data Processing	Example Models	Link
GPT	Decoder-only architecture that generates data autoregressively	Self-supervised pretraining	Processes data sequentially, predicting each element based on previous ones	scGPT, GPTCelltype	https://github.com/bowang-lab/scGPT
BERT	Encoder-only architecture with self-attention mechanisms	Masked language modeling	Processes data bidirectionally, understanding context from all tokens	scBERT, CANAL, TOSICA, Geneformer	https://github.com/TencentAILabHealthcare/scBERT https://github.com/aster-ww/CANAL-torch https://github.com/JackieHanLab/TOSICA https://huggingface.co/ctheodoris/Geneformer
Encoder-decoder	Combines encoder and decoder for sequence-to-sequence tasks	Seq2seq pretraining	Encodes input into a latent representation and decodes output	scFoundation	https://github.com/biomap-research/scFoundation
Graph-based	Transformer adapted for graph-structured data	Graph-based attention	Handles structured, graph-based data with attention to relationships	DeepMAPS	https://github.com/OSU-BMBL/deepmaps
GAN	Generator-discriminator framework	Adversarial training (generator vs. discriminator)	Generates data in parallel from random noise	cscGAN, scIGAN, scGGAN	https://github.com/imsb-uke/scGAN https://github.com/xuyungang/scIGANs https://www.sdu-idea.cn/codes.php?name=scGGAN
Diffusion model	Iteratively denoises data through probabilistic modeling	Progressive denoising to approximate data distributions	Sequentially removes noise to generate data closer to the original distribution	stDiff	https://github.com/fdu-wangfeilab/stDiff

prevent knowledge loss when learning new data, CANAL saves examples of rare cell types to review later and compares older and newer outputs to maintain important information while still learning new things. This allows CANAL to continuously expand its annotation capabilities.

Another encoder-only transformer TOSICA [14] is applied for annotating cell types in scRNA-seq data. Unlike scBERT, which uses self-supervised pretraining, TOSICA is trained directly on labeled data without pretraining. While CANAL emphasizes continuous learning, TOSICA excels in one-time annotation. It effectively handles batch effects without needing explicit batch information, offering interpretable insights at both gene and pathway levels, making it useful for trajectory analysis and dataset integration.

Geneformer [15] (40M parameters, 30M cells) demonstrates superior performance in cell type annotation and gene function prediction over other single-cell transcriptomics LLMs [16], including scBERT (5M parameters, 1M cells), due to their larger model sizes and extensive pre-training data (Figure 2). Geneformer also surpasses scGPT due to ranking genes based on importance, which enhances its ability to prioritize distinguishing genes while minimizing the effects of housekeeping genes and technical artifacts.

These pre-trained models, initially trained on human cells, can be fine-tuned for plant studies to improve cell-type annotation in plant scRNA-seq datasets. As encoder-only models, they specialize in understanding and annotating existing data without the ability to generate new data like GAN or scGPT, making them particularly well-suited for clustering and classification tasks in single-cell RNA-seq data (Figure 1).

Enhancing single-cell and multi-omics analysis with other transformer [17] architectures

Building on transformer models' capabilities in analyzing single-cell data, DeepMAPS [18] uses a heterogeneous graph transformer (HGT) to infer cell-type-specific networks from scMulti-omics data. Unlike encoder-only or decoder-only transformer architectures, HGT integrates multiple data types, making it ideal for inferring complex biological networks from multi-modal data. HGT can be applied to multi-omics plant data, such as scRNA-seq and scATAC-seq, to uncover how different plant cell types interact and regulate gene expression in response to environmental factors like drought, nutrient fluctuations, or pathogen attacks.

Recently published, scFoundation [19] is a large pretrained model based on over 50 million scRNA-seq profiles. With an asymmetric encoder-decoder transformer-like architecture and 100 million parameters, it captures complex gene−gene relationships across diverse cell types and states, excelling in tasks such as gene expression enhancement, cell-type annotation, drug response prediction, perturbation prediction, and gene regulation networks (Figure 1). However, since scFoundation is pretrained on human data, capturing plant-specific gene interactions would require significant retraining on plant datasets, and this process demands substantial computational resources.

Solving single-cell data limitation and sparsity with GANs

Most traditional Generative Adversarial Network (GAN) models are not foundation models, but they offer solutions for data sparsity in single-cell analysis by generating synthetic data (Table 1). For example, cscGAN [20] generates specific cell types on demand, useful for augmenting sparse populations. scIGANs [21] employed a CNN-based GAN to impute data by generating synthetic cells, balancing performance between major and rare cell populations. scGGAN [22] integrates Graph Convolutional Networks with GANs, considering gene expression as controlled by related genes and using both single-cell and bulk RNA-seq data to construct a comprehensive gene relation network.

These GAN models are useful for imputing missing data and generating synthetic plant cell data in scRNA-seq datasets, especially when dealing with high dropout rates or rare cell types. They can augment small datasets to balance cell-type representation. However, GANs can be difficult to train and can suffer from mode collapse, where the generator produces limited diversity in output. They are not pre-trained and typically need to be trained from scratch for each new dataset. To evaluate the quality of synthetic data, real and synthetic datasets can be compared using UMAP or t-SNE for clustering, gene expression correlations, gene distribution patterns, or online tools such as Root Cell Atlas [23], Single Cell Expression Atlas [24], and Cella [25].

Enhancing spatial transcriptome data with diffusion models

The diffusion model stDiff [26] enhances spatial transcriptomics (ST) data by imputing missing gene expressions using relationships learned from scRNA-seq data. This model employs two Markov processes: one adds noise to gene expression data, while the other denoises it to recover the missing portions. By integrating scRNA-seq and spatial data, stDiff preserves spatial structures and accurately reconstructs gene expression patterns. This approach is particularly useful for addressing the limitations in plant spatial transcriptomics (Figure 1).

Opportunities and challenges for foundation models in plant genomics

Recent plant studies have generated many single-cell data to understand genetic perturbation, cell-type classification, and developmental trajectories. Shahan et al. [27] mapped over 110,000 Arabidopsis root cells, revealing changes in cell identity due to SHORTROOT and SCARECROW mutations. Another study [28] investigated 210,000 Arabidopsis root cells for Brassinosteroid responses, identifying HAT7 and GTL1 as key regulators of root cell elongation and confirming the cortex's role in hormone-mediated cell expansion. scBERT might be particularly useful for annotating cell types in these mid-sized datasets, while scGPT could be applied for multi-batch integration and gene network inference. In larger datasets, such as the study by Lee et al. [29] which profiled 800,000 single-nucleus data across various organs and developmental stages, models like scGPT or scFoundation are ideal for fine-tuning to predict cell types and analyze more complex gene relationships. Beyond Arabidopsis, a tomato roots study [30] profiled over 20,000 cells to investigate suberin's role in drought tolerance, identifying a novel cell type called exodermis which is absent in Arabidopsis. It would be interesting to explore how to apply foundation models trained in Arabidopsis to identify novel cell types in other plant species. However, the effectiveness of this cross-species application remains an open research question, with a main challenge being the need to create common gene names across species.

GAN and diffusion models improve scRNA-seq data resolution, particularly for lowly expressed genes, benefiting plant biologists using single-nucleus sequencing instead of protoplast-based methods. These models improve gene detection in emerging spatial transcriptome technologies for plants. For example, Serrano and colleagues [31] used single-nucleus and spatial RNA sequencing to study the symbiosis between *Medicago truncatula* and the fungus *Rhizophagus irregularis*, revealing distinct transcriptomic profiles in various root cell types. Zhu et al. [32] employed scRNA-seq on 11,000 *Arabidopsis thaliana* leaf cells, along with confocal imaging, to study responses to *Pseudomonas syringae* infection, identifying a progression from immune to susceptible host cell states. Utilizing diffusion models like stDiff could expand the size of spatial single-cell data, enhancing our understanding of plant−pathogen interactions.

Li and colleagues [33] explored the MIA biosynthetic pathway in *Catharanthus roseus*, revealing gene clusters, duplications, and chromatin interactions using multi-omics approaches. With scRNA-seq, cell-type-specific gene expression and metabolomics lead to the discovery of a reductase for bis-indole alkaloid production. Some metabolic genes are undetectable by scRNA-seq, but GAN models could address this limitation. Farmer et al. [34] employed single-nucleus RNA and ATAC sequencing in Arabidopsis roots, discovering new cell subtypes and linking chromatin accessibility to gene activity. A maize study [35] used scATAC-seq and scRNA-seq on 72,090 nuclei, revealing human selection and retrotransposons' effects on the CRE landscape. These studies could benefit from DeepMAPs which enhance multi-omics analyses in plants or GAN and GPT models by increasing dataset sizes, improving gene expression detection, and facilitating cross-species gene function predictions.

While the natural language processing capabilities of models like ChatGPT can be fine-tuned for plant cell-type annotation, it is unclear how to implement this due to the unknown amount of plant science literature in its original training. Foundation models trained on human and mouse gene expression also face challenges when applied to plants due to differences in gene names. One solution is transfer learning with gene ortholog groups to translate gene names across species. This approach enables pre-trained models to adapt efficiently by retraining only specific layers with plant-specific data. This solution requires improving gene family functional annotation and establishing association of the individual family members with the specific pathways or the reactions in the model plants for important evolutionary clade. One such example is the curated S-domain subfamily of receptor-like kinases in rice [36]. Such analysis is crucial to be performed in other important plant clades. Another option is fine-tuning the entire model on plant-specific data, using ortholog groups to bridge shared functions across species. Lastly, a specialized approach is to train a new model from scratch using exclusively plant data, ensuring that the model learns plant biology directly without relying on knowledge from human or mouse data.

Conclusions and future directions

Many AI tools have been developed for biomedical research in the past few years. Applying foundation models could significantly enhance single-cell data availability and extend their applications to various plant species, improving our understanding of biological processes in plants and eventually leading to improved crop production. Beyond single-cell genomics, foundation models have been applied in genotyping and phenotyping analysis as well. For example, foundation models for sequence homology search [37], genome annotation [38], data mining [39,40], and image analysis [41] have surpassed the traditional machine learning models. Notably, the ability to perform "few-shot" and "zero-shot" learning, where models make accurate predictions with minimal or no task-specific data, is a unique advantage of new foundation models. In the coming

years, adopting foundation models from other domains offers major opportunities for plant science. Expert curation of gene families and gold-standard sets of gene functions, such as gene ontology and plant ontology [42], are urgently needed to help fine-tune these novel foundation models for plants. These capacities hold great promise to democratize advanced AI use in both plant biology research and agriculture production.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by the Plant Genome Research Program (PGRP, Grant No. 2344169) and the National Science Foundation EArly-concept Grants for Exploratory Research (NSF-EARGE, Grant No. 2218234). X.W. received support from the NSF NAIRR Pilot with PSC Neocortex and NCSA Delta, Commonwealth Cyber Initiative, Children's National Hospital, Franlin Biomedical Research Institute (Virginia Tech), Sanghani Center for AI and Data Analytics (Virginia Tech), Virginia Tech Innovation Campus, and a generous gift from the Amazon + Virginia Tech Center for Efficient and Robust Machine Learning.

Data availability

No data was used for the research described in the article.

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest

1. Svensson V, Vento-Tormo R, Teichmann SA: **Exponential scaling of single-cell RNA-seq in the past decade**. *Nat Protoc* 2018, **13**:599–604.

2. Cervantes-Pérez SA, *et al.*: **Single-cell transcriptome atlases of soybean root and mature nodule reveal new regulatory programs that control the nodulation process**. *Plant Commun* 2024, **5**, 100984.

3. Cantó-Pastor A, *et al.*: **A suberized exodermis is required for tomato drought tolerance**. *Nat Plants* 2024, **10**:118–130.

4. Guo X, *et al.*: **Single-cell transcriptome reveals differentiation between adaxial and abaxial mesophyll cells in *Brassica rapa***. *Plant Biotechnol J* 2022, **20**:2233–2235.

5. Xu X, *et al.*: **Single-cell RNA sequencing of developing maize ears facilitates functional analysis and trait candidate gene discovery**. *Dev Cell* 2021, **56**:557–568.e6.

6. Tung C-C, *et al.*: **Single-cell transcriptomics unveils xylem cell development and evolution**. *Genome Biol* 2023, **24**:3.

7. ** Bommasani R, *et al.*: *On the opportunities and risks of foundation models*. 2021.
This is a report by center for research on foundation models at Stanford University. The authors highlighted the technical capacities of foundation models, their different application domains, and the potential impact on society.

8. Samsi S, *et al.*: **From words to watts: benchmarking the energy costs of Large Language model inference**. In *2023 IEEE high performance extreme computing conference (HPEC)*. IEEE; 2023:1–9, https://doi.org/10.1109/HPEC58863.2023.10363447.

9. Hou W, Ji Z: **Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis**. *Nat Methods* 2024, https://doi.org/10.1038/s41592-024-02235-4.

10. Cui H, *et al.*: **scGPT: toward building a foundation model for single-cell multi-omics using generative AI**. *Nat Methods* 2024, https://doi.org/10.1038/s41592-024-02201-0.

11. Yang F, *et al.*: **scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data**. *Nat Mach Intell* 2022, **4**:852–866.

12. Khan SA, *et al.*: **Reusability report: learning the transcriptional grammar in single-cell RNA-sequencing data using transformers**. *Nat Mach Intell* 2023, **5**:1437–1446.

13. Wan H, Yuan M, Fu Y, Deng M: **Continually adapting pre-trained language model to universal annotation of single-cell RNA-seq data**. *Briefings Bioinf* 2024, **25**.

14. Chen J, *et al.*: **Transformer for one stop interpretable cell type annotation**. *Nat Commun* 2023, **14**:223.

15. Theodoris CV, *et al.*: **Transfer learning enables predictions in network biology**. *Nature* 2023, **618**:616–624.

16. Liu T, Li K, Wang Y, Li H, Zhao H: **Evaluating the utilities of foundation models in single-cell data analysis**. *bioRxiv* 2024, https://doi.org/10.1101/2023.09.08.555192.

17. Vaswani A, *et al.*: *Attention is all you need*. 2017.

18. Ma A, *et al.*: **Single-cell biological network inference using a heterogeneous graph transformer**. *Nat Commun* 2023, **14**:964.

19. ** Hao M, *et al.*: **Large-scale foundation model on single-cell transcriptomics**. *Nat Methods* 2024, https://doi.org/10.1038/s41592-024-02305-7.
this is the largest foundation model published to date on single cell data, using more than 50 million cells and 100 million parameters.

20. Marouf M, *et al.*: **Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks**. *Nat Commun* 2020, **11**:166.

21. Xu Y, *et al.*: **scIGANs: single-cell RNA-seq imputation using generative adversarial networks**. *Nucleic Acids Res* 2020, **48**:e85. e85.

22. Huang Z, Wang J, Lu X, Mohd Zain A, Yu G: **scGGAN: single-cell RNA-seq imputation by graph-based generative adversarial network**. *Briefings Bioinf* 2023, **24**.

23. ***Root cell atlas***. https://rootcellatlas.org/.

24. Moreno P, *et al.*: **User-friendly, scalable tools and workflows for single-cell RNA-seq analysis**. *Nat Methods* 2021, **18**:327–328.

25. Su C, *et al.*: **Cella: <scp>3D</scp> data visualization for plant single-cell transcriptomics in Blender**. *Physiol Plantarum* 2023, **175**.

26. Li K, Li J, Tao Y, Wang F: **stDiff: a diffusion model for imputing spatial transcriptomics through single-cell transcriptomics**. *Briefings Bioinf* 2024, **25**.

27. Shahan R, *et al.*: **A single-cell Arabidopsis root atlas reveals developmental trajectories in wild-type and cell identity mutants**. *Dev Cell* 2022, **57**:543–560.e9.

28. * Nolan TM, *et al.*: **Brassinosteroid gene regulatory networks at cellular resolution in the *Arabidopsis* root**. *Science* 2023:379.1979.
This paper and Shahan et al. provided over 300,000 cells, including all major cell types in the Arabidopsis roots. The number of cells is comparable to those used in smaller LLMs in human/mouse research such as CANAL and DeepMAPS.

29. * Lee, T. A. *et al.* A single-nucleus atlas of seed-to-seed development in Arabidopsis.

At the time of this review article being developed, this preprint (DOI: 10.1101/2023.03.23.533992) has not been published in any peer reviewed journals. The number of cells and tissue types analyzed in this article is the largest to date (July 2024) in any plant species.

30. * Cantó-Pastor A, *et al.*: **A suberized exodermis is required for tomato drought tolerance**. *Nat Plants* 2024, **10**:118–130.
This article provides one example where a cell type (exodermis) is not present in the model species Arabidopsis but exists in most other plant species. The authors used a combination of co-expression analysis, marker enrichment, comparison with bulk RNA-seq data and gene function analysis to determine the cell type for the exodermal cell cluster.

31. Serrano K, *et al.*: **Spatial co-transcriptomics reveals discrete stages of the arbuscular mycorrhizal symbiosis**. *Nat Plants* 2024, **10**:673–688.

32. Zhu J, *et al.*: **Single-cell profiling of Arabidopsis leaves to Pseudomonas syringae infection**. *Cell Rep* 2023, **42**, 112676.

33. Li C, *et al.*: **Single-cell multi-omics in the medicinal plant Catharanthus roseus**. *Nat Chem Biol* 2023, **19**:1031–1041.

34. Farmer A, Thibivilliers S, Ryu KH, Schiefelbein J, Libault M: **Single-nucleus RNA and ATAC sequencing reveals the impact of chromatin accessibility on gene expression in Arabidopsis roots at the single-cell level**. *Mol Plant* 2021, **14**: 372–383.

35. Marand AP, Chen Z, Gallavotti A, Schmitz RJ: **A cis-regulatory atlas in maize at single-cell resolution**. *Cell* 2021, **184**: 3041–3055.e21.

36. Naithani S, Dikeman D, Garg P, Al-Bader N, Jaiswal P: **Beyond gene ontology (GO): using biocuration approach to improve the gene nomenclature and functional annotation of rice S-domain kinase subfamily**. *PeerJ* 2021, **9**, e11052.

37. Liu W, *et al.*: **PLMSearch: protein language model powers accurate and fast sequence search for remote homology**. *Nat Commun* 2024, **15**:2775.

38. Zhou Z, *et al.*: *DNABERT-2: efficient foundation model and benchmark for multi-species genome*. 2023.

39. Wan M, *et al.*: *TnT-LLM: text mining at scale with Large Language Models*. 2024.

40. Wornow M, *et al.*: **The shaky foundations of large language models and foundation models for electronic health records**. *NPJ Digit Med* 2023, **6**:135.

41. Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M: *Hierarchical text-conditional image generation with CLIP latents.* 2022.

42. Walls RL, *et al.*: **The plant ontology facilitates comparisons of plant development stages across species**. *Front Plant Sci* 2019, **10**.