



The 34th International Biology Olympiad_UAE_23

1

Educational Conference: The IBO Way to Excellence

“Open Biology Towards Sustainable Thinking”

Dr. Prof. Abdulsamie Hanano
President of Scientific Committee for Syrian Olympiad Biology
Head of Toxicology & Biochemistry Division
AECS, Damascus, Syria



OUTLINE

- **Introduction**
 - What is open Biology?
 - Is open biology really needed for young biologists?
- **Academic publishers of Open Biology:**
 - Where and how students can find scientific publications in open biology
- **Structure of a typical scientific paper in biological sciences**
- **Guidelines for an effective and critical reading of a scientific paper**
- **A brief highlight on the writing of a scientific paper: order of process**



□ Introduction: What is Open Biology?

Google says **Open Biology** is an online journal that welcomes original, high impact research in cell and developmental biology, molecular and structural biology, ...

This is good, however, this is not exactly what we are looking for!

Open access is a broad international movement that seeks to grant free and open online access to academic information, such as books, publications and data.

Making biological knowledge freely accessible for everyone



□ **Introduction: What is Open Biology?**

Making biological knowledge freely accessible for every surfer

Is this mission is achieved?

We can say YES, but not completely

**Over 30% of all scholarly articles are actually published as
paid-for open access**

82% of the peer-reviewed articles from 2021 of the 14 Dutch universities are Open Accessed available with a strong increase of 9% compared to 2020.

This is good, however, is open biology accessible for students?



□ Introduction: What is Open Biology?

Is open biology accessible for IBOs students?

Under-graduated students, our future IBOs, **have not**

- Scholar platform to access to open biology
- Enough information on how and where find open biology
- Very limited sources in the less developing countries



□ **Introduction:** Is open biology really needed for young biologists?

Yes, definitely

Because open biology offers:

- A large and varied collection of academic sources in biology: books, scientific publications, protocols and even more experimental videos,..
- New trends in biology: emergent pollutants, global warming, new tools in bioinformatics,..
- New techniques in biology: analytical and imaging



□ Academic publishers of Open Biology

- Where and how students can find scientific publications in biology?

This is the first question we try to answer and discuss in the welcome lecture with our new National Biology Olympians



□ Academic publishers of Open Biology

- Where and how students can find scientific publications in biology?

There are lots of scientific publications in different branches of biology



Where we find open-accessed scientific publications?



□ Academic publishers of Open Biology

- Where and how students can find scientific publications in biology?

Top 10 academic publishers 2023:

**Science group,
Springer Nature group,
Elsevier,
Cell Press,
Oxford Academic,
Wiley-Blackwell,
Taylor & Francis
BMC group
Frontiers group
PLOS group**

Published more than a half of peer-reviewed academic papers



□ Academic publishers of Open Biology

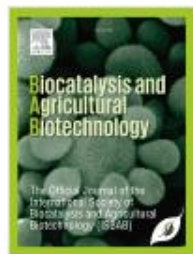
□ Where and how students can find scientific publications in biology?



ELSEVIER

www.elsevier.com

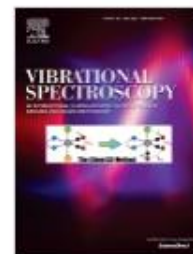
Life Sciences > Biological Sciences journals



Biocatalysis and
Agricultural
Biotechnology



Microbiological
Research



Vibrational
Spectroscopy



Journal of Human
Evolution



Comparative
Biochemistry and
Physiology - Part A:
Molecular &
Integrative Physiology



Current Opinion In
Environmental
Sustainability



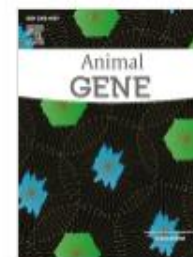
Journal of Nutrition
Education and
Behavior



Journal of Food
Engineering



Morphologie



Animal Gene



Resources Policy



International Journal
for Parasitology

In the most journals, there is no free access to full paper:
We should pay to have PDF!

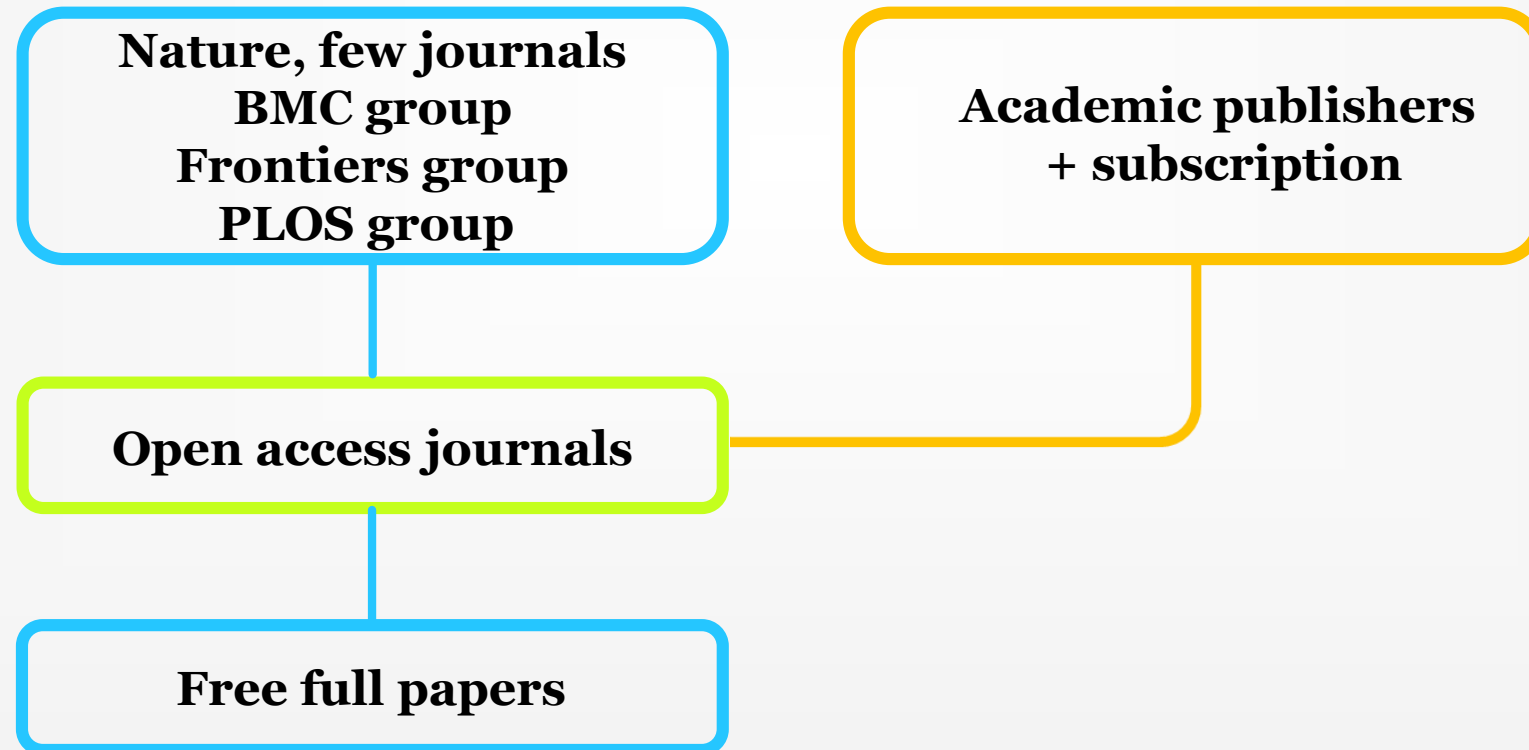
Fortunately the Open Access Journals are present



□ Academic publishers of Open Biology

- Where and how students can find scientific publications in biology?

Top academic publishers with Open Access policy:





□ Academic publishers of Open Biology

□ Where and how students can find scientific publications in biology?

Life Sciences > Biological Sciences journals > Open Access Journals

PLOS Publish with PLOS ▾ Research Communities Open Science ▾ Resources ▾ About PLOS

**Breaking boundaries.
Empowering researchers.
Opening Science.**

PLOS is a nonprofit, Open Access publisher empowering researchers to accelerate progress in science and medicine by leading a transformation in research communication.

About PLOS



□ Academic publishers of Open Biology

□ Where and how students can find scientific publications in biology?

Life Sciences > Biological Sciences journals > Open Access Journals

BROWSE PUBLISH ABOUT SEARCH advanced search

PLOS BIOLOGY

June 20, 2023

A new human embryonic cell type

The inner cell mass, from which the human fetus is derived, has proven hard to define. Laurence Hurst and colleagues perform single-cell analysis and embryo visualization and identify a common novel class of non-committed cells that undergo apoptosis and may reflect a quality control screening process.

Image credit: pbio.3002162



□ Structure of a typical scientific paper in biological sciences

Guidelines on how students can be familiar with scientific paper

A scientific paper provides information on:



✓ Scientific knowledge

✓ Paper design

✓ Scientific language



□ Structure of a typical scientific paper in biological sciences

In biological sciences, a scientific paper is typically composed of:

- Title
- Abstract
- Introduction
- Materials & methods
- Results
- Discussion
- Conclusions
- Acknowledgment
- References

nature
computational
science

ARTICLES

https://doi.org/10.1038/s43588-022-00263-8

OPEN

Minimal gene set discovery in single-cell mRNA-seq datasets with ActiveSVM

Xiaoqiao Chen¹, Sisi Chen^{2,3} and Matt Thomson^{1,2,3}

Sequencing costs currently prohibit the application of single-cell mRNA-seq to many biological and clinical analyses. Targeted single-cell mRNA-sequencing reduces sequencing costs by profiling reduced gene sets that capture biological information with a minimal number of genes. Here we introduce an active learning method that identifies minimal but highly informative gene sets that enable the identification of cell types, physiological states and genetic perturbations in single-cell data using a small number of genes. Our active feature selection procedure generates minimal gene sets from single-cell data by employing an active support vector machine (ActiveSVM) classifier. We demonstrate that ActiveSVM feature selection identifies gene sets that enable ~90% cell-type classification accuracy across, for example, cell atlas and disease-characterization datasets. The discovery of small but highly informative gene sets should enable reductions in the number of measurements necessary for application of single-cell mRNA-seq to clinical tests, therapeutic discovery and genetic screens.

Single-cell mRNA-seq methods have scaled to allow routine transcriptome-scale profiling of thousands of cells per experimental run. Although single-cell mRNA-seq approaches provide insights into many different biological and biomedical problems, high sequencing costs prohibit the broad application of single-cell mRNA-seq in many exploratory assays such as small-molecule and genetic screens, and in cost-sensitive clinical assays. The sequencing bottleneck has led to the development of targeted mRNA-seq strategies that reduce sequencing costs by up to 90% by focusing sequencing resources on highly informative genes for a given biological question or an analysis^{1–5}. Commercial gene-targeting kits, for example, reduce sequencing costs through selective amplification of specific transcripts using ~1,000 gene-targeting primers.

Cells modulate gene expression through the regulation of transcriptional programs or modules that contain multiple genes regulated by common sets of transcription factors⁶. Genes within transcriptional modules exhibit correlated gene expression due to co-regulation. Correlations in gene expression can enable the transcriptional state of a cell to be reconstructed through the targeted mRNA profiling of a small number of highly informative genes^{3,4}. However, such targeted sequencing approaches require computational methods to identify highly informative genes for specific biological questions, systems or conditions. A range of computational approaches, including differential gene expression analysis and principal components analysis (PCA), can be applied to identify highly informative genes⁷. Yet, current methods for defining minimal gene sets are computationally expensive to apply to large single-cell mRNA-seq datasets and often require heuristic user-defined thresholds for gene selection^{8,9}. As an example, computational approaches based on matrix factorization (PCA, non-negative matrix factorization) are typically applied to complete datasets and therefore are computationally intensive when datasets scale into the millions of cells¹⁰. Furthermore, gene set selection after matrix factorization requires heuristic strategies for thresholding coefficients in gene vectors extracted by PCA or non-negative matrix factorization, and then querying whether the selected genes retain core biological information.

Inspired by active learning¹¹ approaches, here we develop a computational method that selects minimal gene sets capable of reliably identifying cell types and transcriptional states through an active support vector machine classification task (ActiveSVM)^{12–15}. The ActiveSVM algorithm constructs a minimal gene set through an iterative cell-state classification task. At each iteration, ActiveSVM applies the current gene set to classify cells into classes that are provided by unsupervised clustering of cell states, or by supplied experimental labels. The procedure analyzes cells that are misclassified with the current gene set and then identifies maximally informative genes that are added to the growing gene set to improve classification. Traditional active learning algorithms query an oracle for training examples that meet a criteria¹⁶. The ActiveSVM procedure actively queries the output of an SVM classifier for cells that classify poorly, and then performs a detailed analysis of the misclassified cells to select maximally informative genes. By selecting minimal gene sets through a well-defined classification task, we ensure that the gene sets discovered by ActiveSVM retain biological information.

The central contribution of ActiveSVM is that the method can scale to large single-cell datasets with more than one million cells as the procedure focuses computational resources on poorly classified cells. As the algorithm only analyzes the full transcriptome of cells that classify poorly with the current gene set, the method can be applied to discover small sets of genes that can distinguish between cell types at high accuracy even in datasets with over a million profiled cells. We demonstrate that ActiveSVM can analyze a mouse brain dataset with 1.3 million cells in only hours of computational time. In addition to scaling, the ActiveSVM classification paradigm generalizes to a range of single-cell data analysis tasks, including the identification of disease markers, genes that respond to Cas9 perturbation and region-specific genes in spatial transcriptomics.

To demonstrate the performance of ActiveSVM, we apply the method to a series of single-cell genomics datasets and analysts

Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, California, USA; ²Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, USA; ³Buckman Institute Single-cell Profiling and Engineering Center, Pasadena, California, USA. ✉: matt.thomson@caltech.edu

NATURE COMPUTATIONAL SCIENCE | VOL 2 | JUNE 2022 | 387–398 | www.nature.com/natcomp



□ Guidelines for reading of a scientific paper

A scientific paper is typically composed of:

- Title
- Abstract
- **Introduction:** Deep scientific background in a specific topic
- **Materials & methods:** Old & new technics
- **Results:** Figures > statements
- **Discussion:** Critical opinion
- Conclusions
- **Acknowledgment:** Say thank you
- References

nature
computational
science

ARTICLES

<https://doi.org/10.1038/s43588-022-00263-8>

OPEN

Minimal gene set discovery in single-cell mRNA-seq datasets with ActiveSVM

Xiaoqiao Chen¹, Sisi Chen^{2,3} and Matt Thomson^{1,2,3}✉

Sequencing costs currently prohibit the application of single-cell mRNA-seq to many biological and clinical analyses. Targeted single-cell mRNA-sequencing reduces sequencing costs by profiling reduced gene sets that capture biological information with a minimal number of genes. Here we introduce an active learning method that identifies minimal but highly informative gene sets that enable the identification of cell types, physiological states and genetic perturbations in single-cell data using a small number of genes. Our active feature selection procedure generates minimal gene sets from single-cell data by employing an active support vector machine (ActiveSVM) classifier. We demonstrate that ActiveSVM feature selection identifies gene sets that enable ~90% cell-type classification accuracy across, for example, cell atlas and disease-characterization datasets. The discovery of small but highly informative gene sets should enable reductions in the number of measurements necessary for application of single-cell mRNA-seq to clinical tests, therapeutic discovery and genetic screens.

Single-cell mRNA-seq methods have scaled to allow routine transcriptome-scale profiling of thousands of cells per experimental run. Although single-cell mRNA-seq approaches provide insights into many different biological and biomedical problems, high sequencing costs prohibit the broad application of single-cell mRNA-seq in many exploratory assays such as small-molecule and genetic screens, and in cost-sensitive clinical assays. The sequencing bottleneck has led to the development of targeted mRNA-seq strategies that reduce sequencing costs by up to 90% by focusing sequencing resources on highly informative genes for a given biological question or an analysis^{1–5}. Commercial gene-targeting kits, for example, reduce sequencing costs through selective amplification of specific transcripts using ~1,000 gene-targeting primers.

Cells modulate gene expression through the regulation of transcriptional programs or modules that contain multiple genes regulated by common sets of transcription factors⁶. Genes within transcriptional modules exhibit correlated gene expression due to co-regulation. Correlations in gene expression can enable the transcriptional state of a cell to be reconstructed through the targeted mRNA profiling of a small number of highly informative genes^{4,7}. However, such targeted sequencing approaches require computational methods to identify highly informative genes for specific biological questions, systems or conditions. A range of computational approaches, including differential gene expression analysis and principal components analysis (PCA), can be applied to identify highly informative genes⁸. Yet, current methods for defining minimal gene sets are computationally expensive to apply to large single-cell mRNA-seq datasets and often require heuristic user-defined thresholds for gene selection⁹. As an example, computational approaches based on matrix factorization (PCA, non-negative matrix factorization) are typically applied to complete datasets and therefore are computationally intensive when datasets scale into the millions of cells¹⁰. Furthermore, gene set selection after matrix factorization requires heuristic strategies for thresholding coefficients in gene vectors extracted by PCA or non-negative matrix factorization, and then querying whether the selected genes retain core biological information.

Inspired by active learning¹¹ approaches, here we develop a computational method that selects minimal gene sets capable of reliably identifying cell types and transcriptional states through an active support vector machine classification task (ActiveSVM)^{12,13}. The ActiveSVM algorithm constructs a minimal gene set through an iterative cell-state classification task. At each iteration, ActiveSVM applies the current gene set to classify cells into classes that are provided by unsupervised clustering of cell states, or by supplied experimental labels. The procedure analyzes cells that are misclassified with the current gene set and then identifies maximally informative genes that are added to the growing gene set to improve classification. Traditional active learning algorithms query an oracle for training examples that meet a criteria¹⁴. The ActiveSVM procedure actively queries the output of an SVM classifier for cells that classify poorly, and then performs a detailed analysis of the misclassified cells to select maximally informative genes. By selecting minimal gene sets through a well-defined classification task, we ensure that the gene sets discovered by ActiveSVM retain biological information.

The central contribution of ActiveSVM is that the method can scale to large single-cell datasets with more than one million cells as the procedure focuses computational resources on poorly classified cells. As the algorithm only analyzes the full transcriptome of cells that classify poorly with the current gene set, the method can be applied to discover small sets of genes that can distinguish between cell types at high accuracy even in datasets with over a million profiled cells. We demonstrate that ActiveSVM can analyze a mouse brain dataset with 1.3 million cells in only hours of computational time. In addition to scaling, the ActiveSVM classification paradigm generalizes to a range of single-cell data analysis tasks, including the identification of disease markers, genes that respond to Cas9 perturbation and region-specific genes in spatial transcriptomics.

To demonstrate the performance of ActiveSVM, we apply the method to a series of single-cell genomics datasets and analysis

Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, California, USA. ²Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, USA. ³Beckman Institute Single-cell Profiling and Engineering Center, Pasadena, California, USA. ✉e-mail: mthomson@caltech.edu

NATURE COMPUTATIONAL SCIENCE | VOL 2 | JUNE 2022 | 387–398 | www.nature.com/natcomp



Guidelines for reading of a scientific paper

nature
computational
science

ARTICLES

https://doi.org/10.1038/s43588-022-00243-8

OPEN
Minimal gene set discovery in single-cell mRNA-seq datasets with ActiveSVM

Xiaoqiao Chen¹, Sisi Chen^{2,3} and Matt Thomson^{1,2,3,4}✉

Sequencing costs currently prohibit the application of single-cell mRNA-seq to many biological and clinical analyses. Targeted single-cell mRNA-sequencing reduces sequencing costs by profiling reduced gene sets that capture biological information with a minimal number of genes. Here we introduce an active learning method that identifies minimal but highly informative gene sets that enable the identification of cell types, physiological states and genetic perturbations in single-cell data using a small number of genes. Our active feature selection procedure generates minimal gene sets from single-cell data by employing an active support vector machine (ActiveSVM) classifier. We demonstrate that ActiveSVM feature selection identifies gene sets that enable ~90% cell-type classification accuracy across, for example, cell atlas and disease-characterization datasets. The discovery of small but highly informative gene sets should enable reductions in the number of measurements necessary for application of single-cell mRNA-seq to clinical tests, therapeutic discovery and genetic screens.

Single-cell mRNA-seq methods have scaled to allow routine transcriptome-scale profiling of thousands of cells per experimental run. Although single cell mRNA-seq approaches provide insights into many different biological and biomedical problems, high sequencing costs prohibit the broad application of single-cell mRNA-seq in many exploratory assays such as small-molecule and genetic screens, and in cost-sensitive clinical assays. The sequencing bottleneck has led to the development of targeted mRNA-seq strategies that reduce sequencing costs by up to 90% by focusing sequencing resources on highly informative genes for a given biological question or an analysis¹⁻⁵. Commercial gene-targeting kits, for example, reduce sequencing costs through selective amplification of specific transcripts using ~1,000 gene-targeting primers.

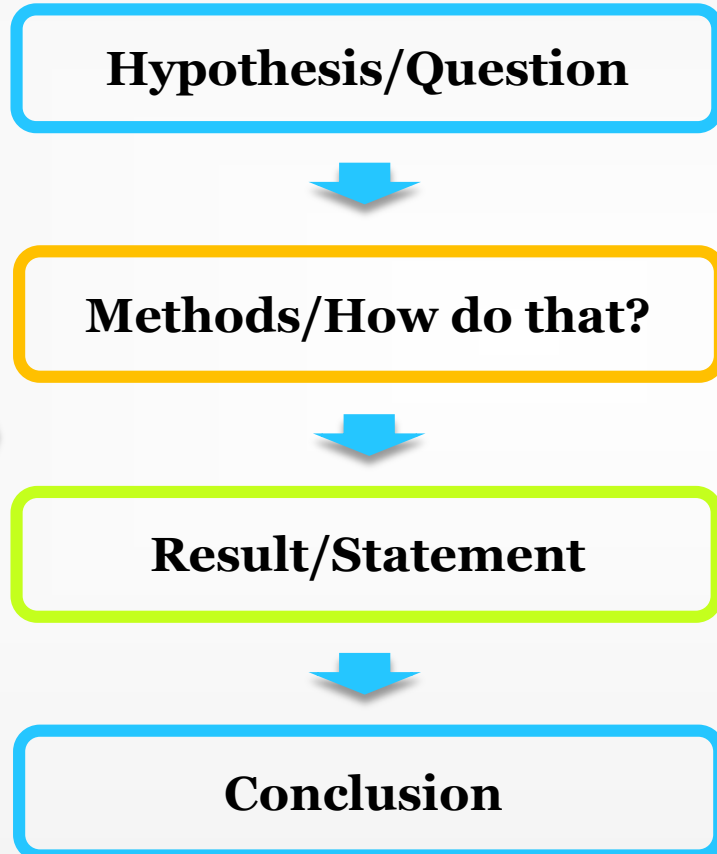
Cells modulate gene expression through the regulation of transcriptional programs or modules that contain multiple genes regulated by common sets of transcription factors⁶. Genes within transcriptional modules exhibit correlated gene expression due to co-regulation. Correlations in gene expression can enable the transcriptional state of a cell to be reconstructed through the targeted mRNA profiling of a small number of highly informative genes⁷. However, such targeted sequencing approaches require computational methods to identify highly informative genes for specific biological questions, systems or conditions. A range of computational approaches, including differential gene expression analysis and principal components analysis (PCA), can be applied to identify highly informative genes. Yet, current methods for identifying minimal gene sets are computationally expensive to apply to large single-cell mRNA-seq datasets and often require heuristic user-defined thresholds for gene selection^{8,9}. As an example, computational approaches based on matrix factorization (PCA, non-negative matrix factorization) are typically applied to complete datasets and therefore are computationally intensive when datasets scale into the millions of cells¹⁰. Furthermore, gene set selection after matrix factorization requires heuristic strategies for thresholding coefficients in gene vectors extracted by PCA or non-negative matrix factorization, and then querying whether the selected genes retain core biological information.

Inspired by active learning approaches, here we develop a computational method that selects minimal gene sets capable of reliably identifying cell types and transcriptional states through an active support vector machine classification task (ActiveSVM)^{11,12}. The ActiveSVM algorithm constructs a minimal gene set through an iterative cell-state classification task. At each iteration, ActiveSVM applies the current gene set to classify cells into classes that are provided by unsupervised clustering of cell states, or by supplied experimental labels. The procedure analyzes cells that are misclassified with the current gene set and then identifies maximally informative genes that are added to the growing gene set to improve classification. Traditional active learning algorithms query an oracle for training examples that meet a criteria¹³. The ActiveSVM procedure actively queries the output of an SVM classifier for cells that classify poorly, and then performs a detailed analysis of the misclassified cells to select maximally informative genes. By selecting minimal gene sets through a well-defined classification task, we ensure that the gene sets discovered by ActiveSVM retain biological information.

The central contribution of ActiveSVM is that the method can scale to large single-cell datasets with more than one million cells as the procedure focuses computational resources on poorly classified cells. As the algorithm only analyzes the full transcriptome of cells that classify poorly with the current gene set, the method can be applied to discover small sets of genes that can distinguish between cell types at high accuracy even in datasets with over a million profiled cells. We demonstrate that ActiveSVM can analyze a mouse brain dataset with 1.3 million cells in only hours of computational time. In addition to scaling, the ActiveSVM classification paradigm generalizes to a range of single-cell data analysis tasks, including the identification of disease markers, genes that respond to Cas9 perturbations and region-specific genes in spatial transcriptomics.

To demonstrate the performance of ActiveSVM, we apply the method to a series of single-cell genomics datasets and analyze

From each result segment

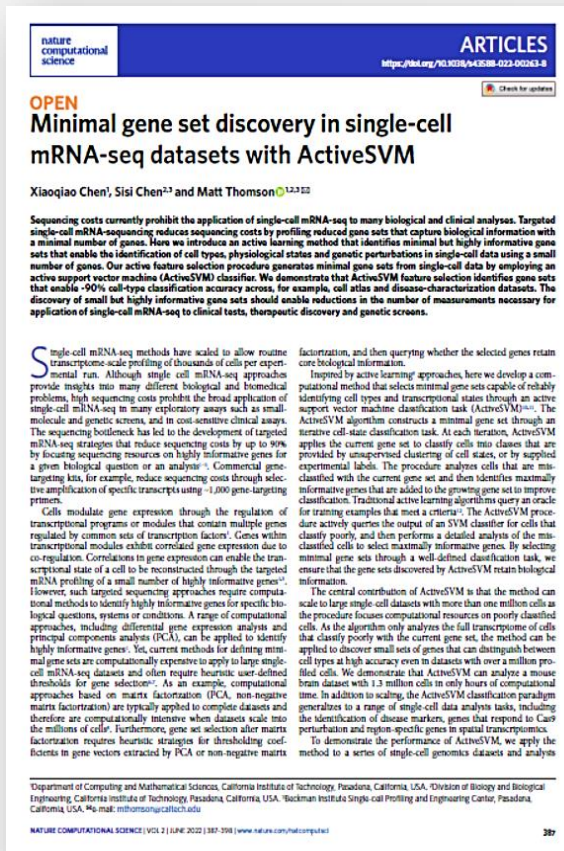


Make a presentation!

¹Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, California, USA. ²Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, USA. ³Buckman Institute Single-cell Profiling and Engineering Center, Pasadena, California, USA. ✉e-mail: mthomson@caltech.edu



Guidelines for reading of a scientific paper



Skills acquired:

- ❑ To gain information on the construct of a scientific paper
- ❑ To be familiar with an advanced scientific language
- ❑ To gain more an elaborated scientific background
- ❑ To be familiar with different forms of illustrations
- ❑ To prepare a presentation
- ❑ And more importantly, to talk and discuss with others



□ A brief highlight on the writing of a scientific paper: order of process

□ How a scientific paper is written?

Parts of a paper are usually ordered by:

- Title
- Abstract
- Introduction
- Materials & methods
- Results
- Discussion
- Conclusions
- Acknowledgment
- References

Which part is written first?



OPEN Minimal gene set discovery in single-cell mRNA-seq datasets with ActiveSVM

Xiaoqiao Chen¹, Sisi Chen^{2,3} and Matt Thomson^{1,2,3,4}

Sequencing costs currently prohibit the application of single-cell mRNA-seq to many biological and clinical analyses. Targeted single-cell mRNA-seq reduces sequencing costs by profiling reduced gene sets that capture biological information with a minimal number of genes. Here we introduce an active learning method that identifies minimal but highly informative gene sets that enable the identification of cell types, physiological states and genetic perturbations in single-cell data using a small number of genes. Our active feature selection procedure generates minimal gene sets from single-cell data by employing an active support vector machine (ActiveSVM) classifier. We demonstrate that ActiveSVM feature selection identifies gene sets that enable ~90% cell-type classification accuracy across, for example, cell atlas and disease-characterization datasets. The discovery of small but highly informative gene sets should enable reductions in the number of measurements necessary for application of single-cell mRNA-seq to clinical tests, therapeutic discovery and genetic screens.

Single-cell mRNA-seq methods have scaled to allow routine transcriptome-scale profiling of thousands of cells per experimental run. Although single-cell mRNA-seq approaches provide insights into many different biological and biomedical problems, high sequencing costs prohibit the broad application of single-cell mRNA-seq in many exploratory assays such as small-molecule and genetic screens, and in cost-sensitive clinical assays. The sequencing bottleneck has led to the development of targeted mRNA-seq strategies that reduce sequencing costs by up to 90% by focusing sequencing resources on highly informative genes for a given biological question or an analysis^{1–5}. Commercial gene-targeting kits, for example, reduce sequencing costs through selective amplification of specific transcripts using ~1,000 gene-targeting primers.

Cells modulate gene expression through the regulation of transcriptional programs or modules that contain multiple genes regulated by common sets of transcription factors⁶. Genes within transcriptional modules exhibit correlated gene expression due to co-regulation. Correlations in gene expression can enable the transcriptional state of a cell to be reconstructed through the targeted mRNA profiling of a small number of highly informative genes^{7–11}.

However, such targeted sequencing approaches require computational methods to identify highly informative genes for specific biological questions, systems or conditions. A range of computational approaches, including differential gene expression analysis and principal components analysis (PCA), can be applied to identify highly informative genes¹². Yet, current methods for defining minimal gene sets are computationally expensive to apply to large single-cell mRNA-seq datasets and often require heuristic user-defined thresholds for gene selection¹³. As an example, computational approaches based on matrix factorization (PCA, non-negative matrix factorization) are typically applied to complete datasets and therefore are computationally intensive when datasets scale into the millions of cells¹⁴. Furthermore, gene set selection after matrix factorization requires heuristic strategies for thresholding coefficients in gene vectors extracted by PCA or non-negative matrix

factorization, and then querying whether the selected genes retain core biological information.

Inspired by active learning approaches, here we develop a computational method that selects minimal gene sets capable of reliably identifying cell types and transcriptional states through an active support vector machine classification task (ActiveSVM)^{15,16}. The ActiveSVM algorithm constructs a minimal gene set through an iterative cell-state classification task. At each iteration, ActiveSVM applies the current gene set to classify cells into classes that are provided by unsupervised clustering of cell states, or by supplied experimental labels. The procedure analyzes cells that are misclassified with the current gene set and then identifies maximally informative genes that are added to the growing gene set to improve classification. Traditional active learning algorithms query an oracle for training examples that meet a criterion¹⁷. The ActiveSVM procedure actively queries the output of an SVM classifier for cells that classify poorly, and then performs a detailed analysis of the misclassified cells to select maximally informative genes. By selecting minimal gene sets through a well-defined classification task, we ensure that the gene sets discovered by ActiveSVM retain biological information.

The central contribution of ActiveSVM is that the method can scale to large single-cell datasets with more than one million cells as the procedure focuses computational resources on poorly classified cells. As the algorithm only analyzes the full transcriptome of cells that classify poorly with the current gene set, the method can be applied to discover small sets of genes that can distinguish between cell types at high accuracy even in datasets with over a million profiled cells. We demonstrate that ActiveSVM can analyze a mouse brain dataset with 1.3 million cells in only hours of computational time. In addition to scaling, the ActiveSVM classification paradigm generalizes to a range of single-cell data analysis tasks, including the identification of disease markers, genes that respond to Cas9 perturbation and region-specific genes in spatial transcriptomics.

To demonstrate the performance of ActiveSVM, we apply the method to a series of single-cell genomics datasets and analysis

Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, California, USA; Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, USA; Buckman Institute Single-Cell Profiling and Engineering Center, Pasadena, California, USA; *e-mail: mthomson@caltech.edu

NATURE COMPUTATIONAL SCIENCE | VOL 2 | JUNE 2022 | 387–398 | www.nature.com/natcomp

387



□ A brief highlight on the writing of a scientific paper: order of process

How a scientific paper is written?

Writing a paper is usually proceeded in the following order:

1. **Materials & methods:** accurate, reproducible
2. **Illustrations & tables:** clear and understandable without text
3. **Results:** direct, concise and comparative
4. **Introduction:** well constructed, concise, elusive
5. **Discussion:** comparison, interpretation and explication
6. **Conclusions:** significance and applications
7. **Abstract:** a very compact version of the paper
8. **Title:** one statement covering the most important result
9. **Acknowledgment**
10. **References:** RF management software: EndNote

nature
computational
science

ARTICLES
<https://doi.org/10.1038/s43588-022-00263-8>

Check for updates

OPEN
Minimal gene set discovery in single-cell mRNA-seq datasets with ActiveSVM

Xiaoqiao Chen¹, Sisi Chen^{2,3} and Matt Thomson^{1,2,3,4}

Sequencing costs currently prohibit the application of single-cell mRNA-seq to many biological and clinical analyses. Targeted single-cell mRNA-sequencing reduces sequencing costs by profiling reduced gene sets that capture biological information with a minimal number of genes. Here we introduce an active learning method that identifies minimal but highly informative gene sets that enable the identification of cell types, physiological states and genetic perturbations in single-cell data using a small number of genes. Our active feature selection procedure generates minimal gene sets from single-cell data by employing an active support vector machine (ActiveSVM) classifier. We demonstrate that ActiveSVM feature selection identifies gene sets that enable ~90% cell-type classification accuracy across, for example, cell atlas and disease-characterization datasets. The discovery of small but highly informative gene sets should enable reductions in the number of measurements necessary for application of single-cell mRNA-seq to clinical tests, therapeutic discovery and genetic screens.

Single-cell mRNA-seq methods have scaled to allow routine transcriptome-scale profiling of thousands of cells per experimental run. Although single cell mRNA-seq approaches provide insights into many different biological and biomedical problems, high sequencing costs prohibit the broad application of single-cell mRNA-seq in many exploratory assays such as small-molecule and genetic screens, and in cost-sensitive clinical assays. The sequencing bottleneck has led to the development of targeted mRNA-seq strategies that reduce sequencing costs by up to 90% by focusing sequencing resources on highly informative genes for a given biological question or an analysis¹. Commercial gene-targeting kits, for example, reduce sequencing costs through selective amplification of specific transcripts using ~1,000 gene-targeting primers.

Cells modulate gene expression through the regulation of transcriptional programs or modules that contain multiple genes regulated by common sets of transcription factors². Genes within transcriptional modules exhibit correlated gene expression due to co-regulation. Correlations in gene expression can enable the transcriptional state of a cell to be reconstructed through the targeted mRNA profiling of a small number of highly informative genes³. However, such targeted sequencing approaches require computational methods to identify highly informative genes for specific biological questions, systems or conditions. A range of computational approaches, including differential gene expression analysis and principal components analysis (PCA), can be applied to identify highly informative genes⁴. Yet, current methods for defining minimal gene sets are computationally expensive to apply to large single-cell mRNA-seq datasets and often require heuristic user-defined thresholds for gene selection⁵. As an example, computational approaches based on matrix factorization (PCA, non-negative matrix factorization) are typically applied to complete datasets and therefore are computationally intensive when datasets scale into the millions of cells⁶. Furthermore, gene set selection after matrix factorization requires heuristic strategies for thresholding coefficients in gene vectors extracted by PCA or non-negative matrix factorization, and then querying whether the selected genes retain core biological information.

Inspired by active learning⁷ approaches, here we develop a computational method that selects minimal gene sets capable of reliably identifying cell types and transcriptional states through an active support vector machine classification task (ActiveSVM⁸). The ActiveSVM algorithm constructs a minimal gene set through an iterative cell-state classification task. At each iteration, ActiveSVM applies the current gene set to classify cells into classes that are provided by unsupervised clustering of cell states, or by supplied experimental labels. The procedure analyzes cells that are misclassified with the current gene set and then identifies maximally informative genes that are added to the growing gene set to improve classification. Traditional active learning algorithms query an oracle for training examples that meet a criteria⁹. The ActiveSVM procedure actively queries the output of an SVM classifier for cells that classify poorly, and then performs a detailed analysis of the misclassified cells to select maximally informative genes. By selecting minimal gene sets through a well-defined classification task, we ensure that the gene sets discovered by ActiveSVM retain biological information.

The central contribution of ActiveSVM is that the method can scale to large single-cell datasets with more than one million cells as the procedure focuses computational resources on poorly classified cells. As the algorithm only analyzes the full transcriptome of cells that classify poorly with the current gene set, the method can be applied to discover small sets of genes that can distinguish between cell types at high accuracy even in datasets with over a million profiled cells. We demonstrate that ActiveSVM can analyze a mouse brain dataset with 1.3 million cells in only hours of computational time. In addition to scaling, the ActiveSVM classification paradigm generalizes to a range of single-cell data analysis tasks, including the identification of disease markers, genes that respond to Cas9 perturbation and region-specific genes in spatial transcriptomics.

To demonstrate the performance of ActiveSVM, we apply the method to a series of single-cell genomics datasets and analysis

¹Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, California, USA. ²Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, USA. ³Blackman Institute Single Cell Profiling and Engineering Center, Pasadena, California, USA. ⁴✉ matt.thomson@caltech.edu

NATURE COMPUTATIONAL SCIENCE | VOL 2 | JUNE 2022 | 387–398 | www.nature.com/naturecomputational 387



- A brief highlight on the writing of a scientific paper: order of process
- How a scientific paper is written?

nature
computational
science

ARTICLES
<https://doi.org/10.1038/s43588-022-00263-8>

Check for updates

OPEN
Minimal gene set discovery in single-cell mRNA-seq datasets with ActiveSVM

Xiaoqiao Chen¹, Sisi Chen^{2,3} and Matt Thomson^{1,2,3}✉

Sequencing costs currently prohibit the application of single-cell mRNA-seq to many biological and clinical analyses. Targeted single-cell mRNA-sequencing reduces sequencing costs by profiling reduced gene sets that capture biological information with a minimal number of genes. Here we introduce an active learning method that identifies minimal but highly informative gene sets that enable the identification of cell types, physiological states and genetic perturbations in single-cell data using a small number of genes. Our active feature selection procedure generates minimal gene sets from single-cell data by employing an active support vector machine (ActiveSVM) classifier. We demonstrate that ActiveSVM feature selection identifies gene sets that enable ~90% cell-type classification accuracy across, for example, cell atlas and disease-characterization datasets. The discovery of small but highly informative gene sets should enable reductions in the number of measurements necessary for application of single-cell mRNA-seq to clinical tests, therapeutic discovery and genetic screens.

Single-cell mRNA-seq methods have scaled to allow routine transcriptome-scale profiling of thousands of cells per experimental run. Although single cell mRNA-seq approaches provide insights into many different biological and biomedical problems, high sequencing costs prohibit the broad application of single-cell mRNA-seq in many exploratory assays such as small-molecule and genetic screens, and in cost-sensitive clinical assays. The sequencing bottleneck has led to the development of targeted mRNA-seq strategies that reduce sequencing costs by up to 90% by focusing sequencing resources on highly informative genes for a given biological question or an analysis¹. Commercial gene-targeting kits, for example, reduce sequencing costs through selective amplification of specific transcripts using ~1,000 gene-targeting primers.

Cells modulate gene expression through the regulation of transcriptional programs or modules that contain multiple genes regulated by common sets of transcription factors². Genes within transcriptional modules exhibit correlated gene expression due to co-regulation. Correlations in gene expression can enable the transcriptional state of a cell to be reconstructed through the targeted mRNA profiling of a small number of highly informative genes³. However, such targeted sequencing approaches require computational methods to identify highly informative genes for specific biological questions, systems or conditions. A range of computational approaches, including differential gene expression analysis and principal components analysis (PCA), can be applied to identify highly informative genes⁴. Yet, current methods for defining minimal gene sets are computationally expensive to apply to large single-cell mRNA-seq datasets and often require heuristic user-defined thresholds for gene selection. As an example, computational approaches based on matrix factorization (PCA, non-negative matrix factorization) are typically applied to complete datasets and therefore are computationally intensive when datasets scale into the millions of cells⁵. Furthermore, gene set selection after matrix factorization requires heuristic strategies for thresholding coefficients in gene vectors extracted by PCA or non-negative matrix factorization, and then querying whether the selected genes retain core biological information.

Inspired by active learning⁶ approaches, here we develop a computational method that selects minimal gene sets capable of reliably identifying cell types and transcriptional states through an active support vector machine classification task (ActiveSVM)⁷. The ActiveSVM algorithm constructs a minimal gene set through an iterative cell-state classification task. At each iteration, ActiveSVM applies the current gene set to classify cells into classes that are provided by unsupervised clustering of cell states, or by supplied experimental labels. The procedure analyzes cells that are misclassified with the current gene set and then identifies maximally informative genes that are added to the growing gene set to improve classification. Traditional active learning algorithms query an oracle for training examples that meet a criteria⁸. The ActiveSVM procedure actively queries the output of an SVM classifier for cells that classify poorly, and then performs a detailed analysis of the misclassified cells to select maximally informative genes. By selecting minimal gene sets through a well-defined classification task, we ensure that the gene sets discovered by ActiveSVM retain biological information.

The central contribution of ActiveSVM is that the method can scale to large single-cell datasets with more than one million cells as the procedure focuses computational resources on poorly classified cells. As the algorithm only analyzes the full transcriptome of cells that classify poorly with the current gene set, the method can be applied to discover small sets of genes that can distinguish between cell types at high accuracy even in datasets with over a million profiled cells. We demonstrate that ActiveSVM can analyze a mouse brain dataset with 1.3 million cells in only hours of computational time. In addition to scaling, the ActiveSVM classification paradigm generalizes to a range of single-cell data analysis tasks, including the identification of disease markers, genes that respond to Cas9 perturbation and region-specific genes in spatial transcriptomics.

To demonstrate the performance of ActiveSVM, we apply the method to a series of single-cell genomics datasets and analysis

¹Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, California, USA. ²Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, USA. ³Jackman Institute Single Cell Profiling and Engineering Center, Pasadena, California, USA. ✉matt.thomson@caltech.edu

NATURE COMPUTATIONAL SCIENCE | VOL 2 | JUNE 2022 | 387–398 | www.nature.com/naturecomputational

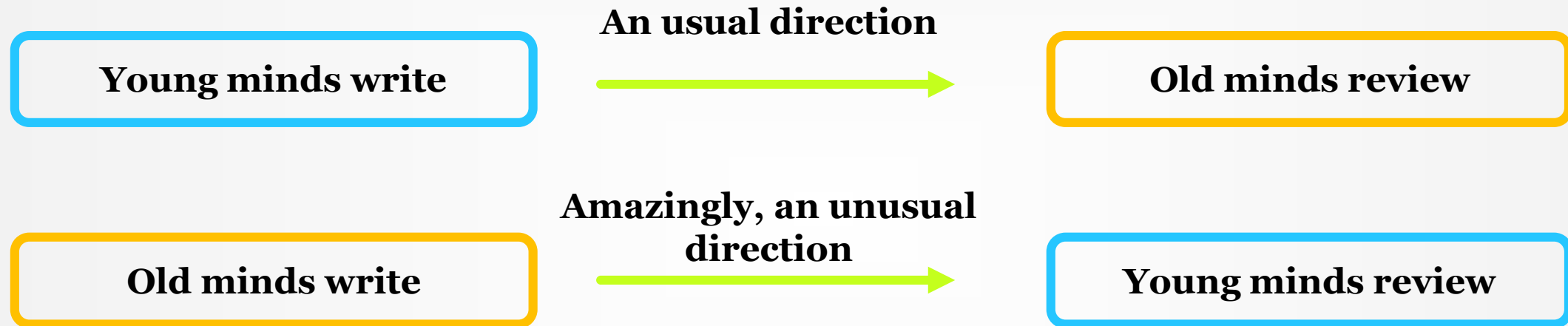
387

Skills acquired:

- To gain information on the writing process of scientific paper
- To write a brief report
- To write and design an IBO international project



- A brief highlight on the writing of a scientific paper: order of process
- Critical reading and reviewing process



Frontiers for Young Minds believes that the best way to make cutting-edge science discoveries available to younger audiences is to enable young people and scientists to work together to create articles that are both accurate and exciting



Core Concept

Biodiversity

Published: April 15, 2022

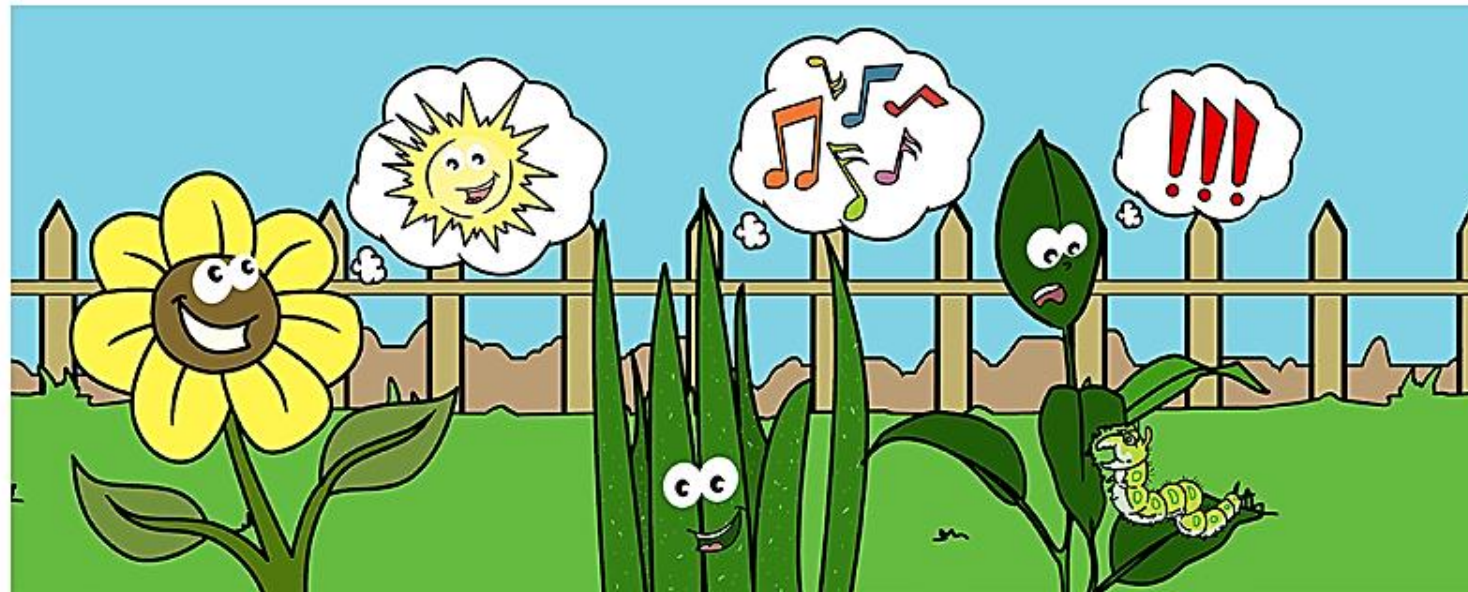
Plants Can “Speak” to Each Other

Authors

 [Abdulsami Hanano](#)  [Colette Murphy](#)  [Denis J. Murphy](#)

Young Reviewers

 [Dario](#)  [Marlene](#)  [Vici](#)





“Open Biology Towards Sustainable Thinking”

Dr. Prof. Abdulsamie Hanano

President of Scientific Committee for Syrian Olympiad Biology

Head of Toxicology & Biochemistry Division

AECS, Damascus, Syria

Email: ashanano@aec.org.sy

Thank you..