IBO_22 Educational Conference: The IBO Way to Excellence

"A good reader today is a good writer tomorrow"

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

OUTLINE

- **Introduction**
  - Why a scientist should write a paper?
- **Academic publishing of scientific papers:**
  - Where and how students can find scientific publications in 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: Why a scientist should write a paper?

Science grows by communication

If you communicate your results = you do something for science ✓

If NOT = you do nothing ✗

If you choose ✓

A primary task that a researcher will front is the communication of his results to the broader scientific community

Introduction: Why a scientist should write a paper?

Science grows by communication

Communication starts by "Writing" and ends by "Publishing"

Writing a research manuscript is an intimidating process for many beginner writers in sciences

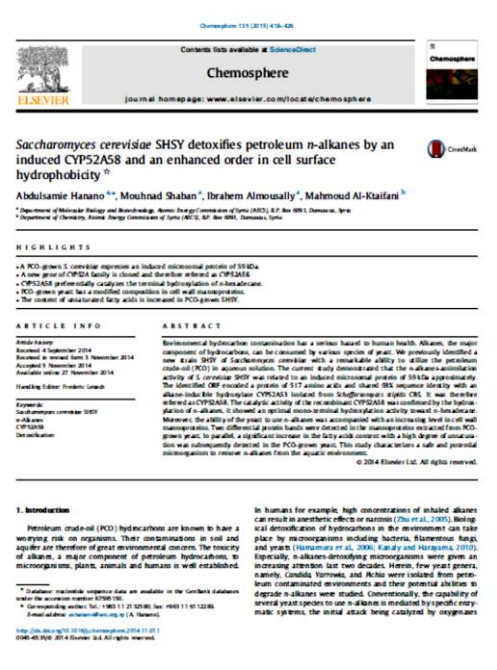
One of the stumbling blocks is the beginning of the process and creating the first draft

Introduction: Why a scientist should write a paper?

A good writing starts with a good "reading"

Do not lose a time! **Read as much as you can** before the writing day comes!!

A scientific paper provides information on:



Chemosphere 135 (2015) 416–426

Contents lists available at ScienceDirect

Chemosphere

journal homepage: www.elsevier.com/locate/chemosphere

Saccharomyces cerevisiae SHSY detoxifies petroleum *n*-alkanes by an induced CYP52A58 and an enhanced order in cell surface hydrophobicity ☆

Abdulsamie Hanano *, Mouhnad Shaban *, Ibrahem Almousally *, Mahmoud Al-Ktaifani b

HIGHLIGHTS

- A PCO-grown *S. cerevisiae* expresses an induced microsomal protein of 59 kDa.
- A new gene of CYP52A family is cloned and therefore refered as CYP52A58.
- CYP52A58 preferentially catalyzes the terminal hydroxylation of *n*-hexadecane.
- PCO-grown yeast has a modified composition in cell wall mannoproteins.
- The content of unsaturated fatty acids is increased in PCO-grown SHSY.

ARTICLE INFO

ABSTRACT

1. Introduction

- ✓ **Scientific knowledge**
- ✓ **Paper design**
- ✓ **Scientific language**

Academic publishing of scientific papers

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

There are lots of scientific publications in different branches of biology







Where we find scientific publications?

Academic publishing of scientific papers

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

Top 10 academic publishers 2022:

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 publishing of scientific papers

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

ELSEVIER

www.elsevier.com

Life Sciences > Biological Sciences journals

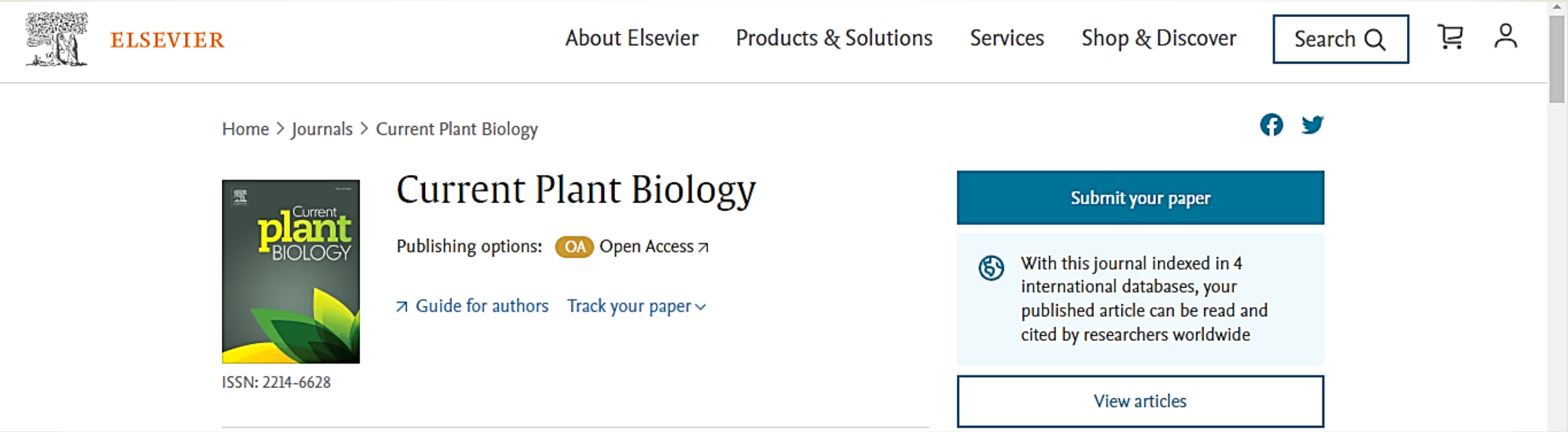
In the most journals, there is no free access to full paper:
We should pay to have publication!

Fortunately the Open Access Journals are present

Academic publishing of scientific papers

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

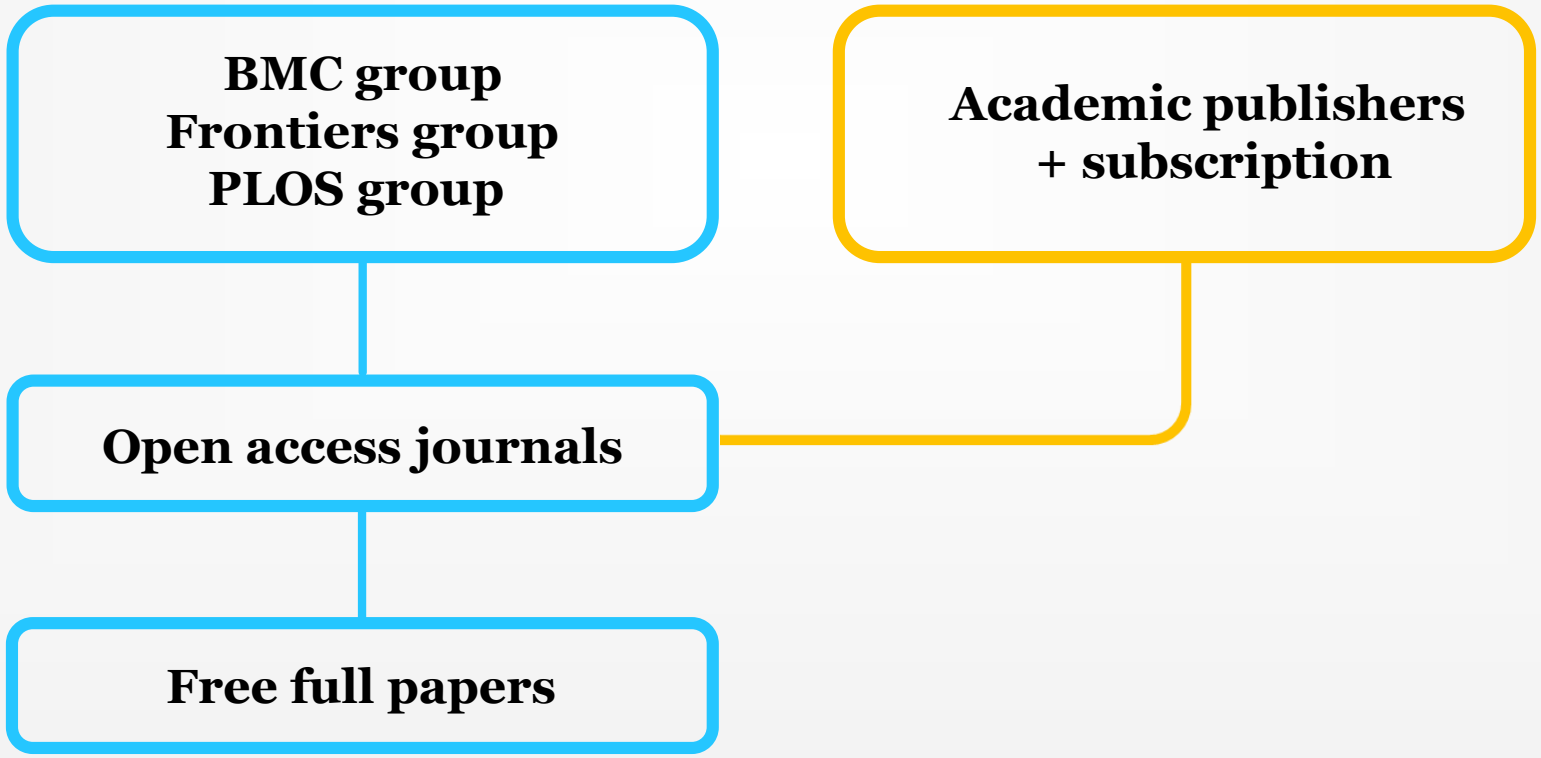
Life Sciences > Biological Sciences journals > Open Access Journals



Academic publishing of scientific papers

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

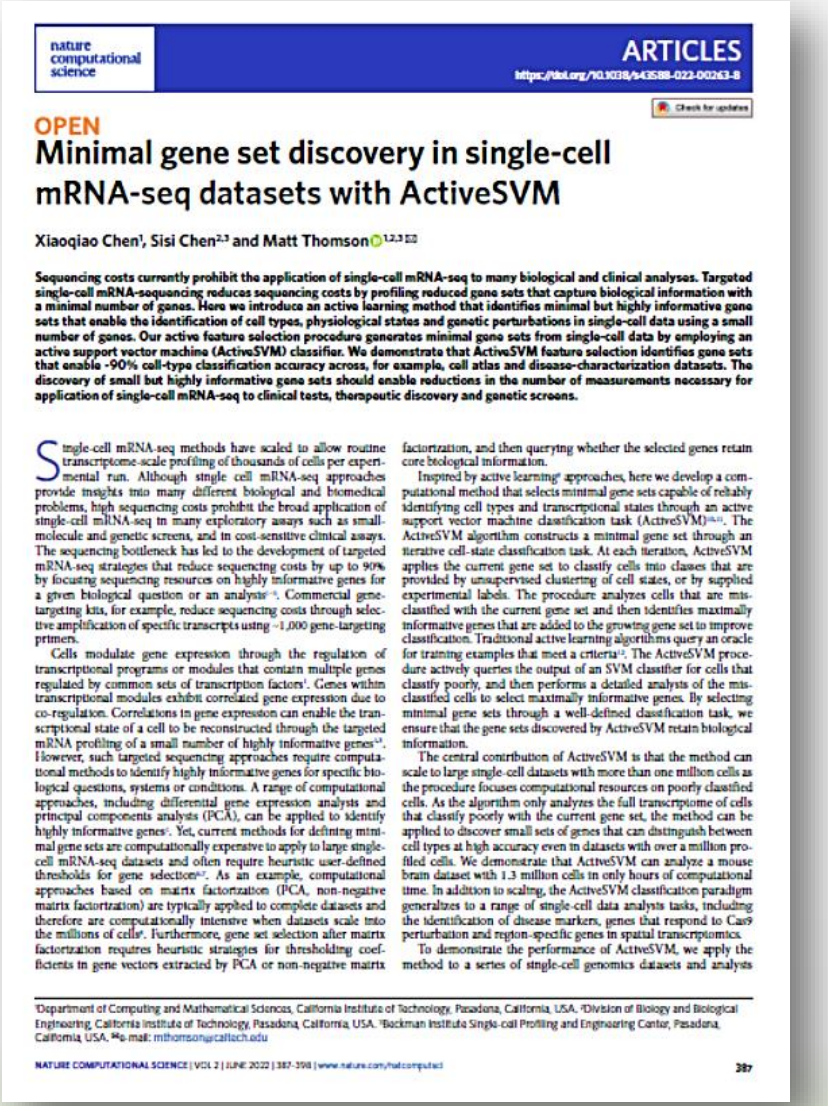
Top academic publishers with Open Access policy:



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 Chen[1], 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 factors[1]. 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[3,5]. 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[6]. 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[6,7]. 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 cells[8]. 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[9] 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)[10,11]. 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 criteria[12]. 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

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

Structure of a typical scientific paper in biological sciences

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

Check for updates

OPEN

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

Xiaoqiao Chen[1], 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 factors[1]. 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[1,3]. 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[6]. 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[6,7]. 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 cells[8]. 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[9] 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)[10,11]. 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 criteria[12]. 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

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

Guidelines for reading of a scientific paper

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 Chen[1], 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 factors[6]. 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[7,8]. 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[7]. 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[9–12]. 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 cells[9]. 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[13] 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)[14,15]. 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 criteria[13]. 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

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

From each result segment

Hypothesis/Question

Methods/How do that?

Result/Statement

Conclusion

Make a presentation!

Guidelines for reading of a scientific paper

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 Chen[1], 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[1–3]. 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 factors[4]. 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[5,6]. 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[7]. 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 cells[8]. 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[10] 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)[11–14]. 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 criteria[15]. 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

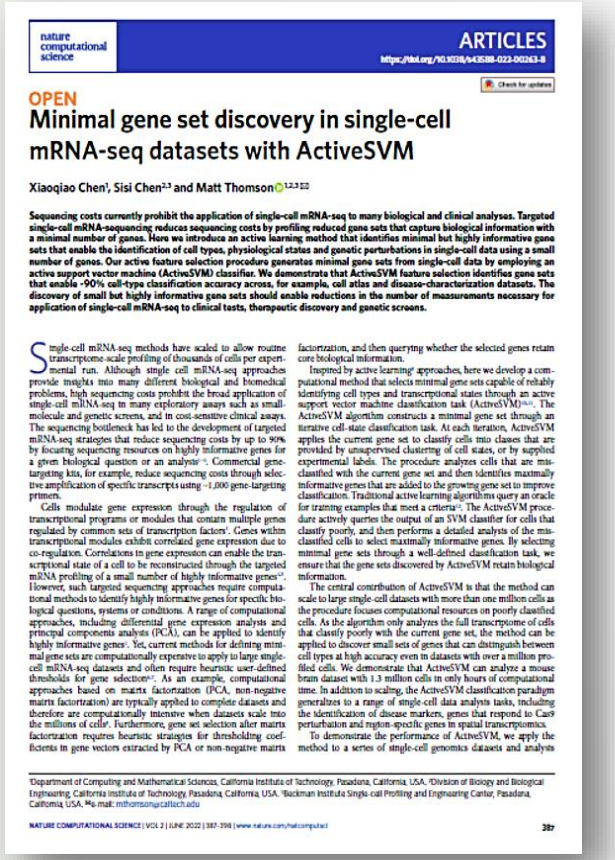
[1]Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, California, USA. [2]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, USA. [3]Beckman Institute Single-cell Profiling and Engineering Center, Pasadena,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

Skills acquired:

- Gain information on the construct of a scientific paper
- Familiarize with an advanced scientific language
- Gain more an elaborated scientific background
- Familiarize with different forms of illustrations
- Prepare a presentation
- And more importantly, to talk and discuss with friends

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

OPEN

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

Xiaoqiao Chen[1], 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 factors[6]. 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[7,8]. 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[9]. 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[10,11]. 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 cells[9]. 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[12] 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)[13,14]. 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 criteria[12]. 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

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

Parts of a paper are usually ordered by:

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

Which part is written first?

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 Chen[1], 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 factors[6]. 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[7,8]. 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[9]. 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[10,11]. 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 cells[9]. 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[12] 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)[13,14]. 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 criteria[12]. 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

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

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

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 Chen[1], 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 factors[1]. 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[6,7]. 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 genes[1]. 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[6,7]. 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 cells[8]. 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[9] 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)[10,11]. 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 criteria[12]. 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

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

Skills acquired:

- Gain information on the writing process of scientific paper
- Writing a brief reports
- Writing and design IBO international project

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

Critical reading and reviewing process

Young minds write → (A usual direction) → **Old minds review**

Old minds write → (Amazingly, an unusual direction) → **Young minds review**

frontiers FOR YOUNG MINDS

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

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

Critical reading and reviewing process

Thank you..