### **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

### **IBO\_22** Educational Conference: The IBO Way to Excellence



# OUTLINE

# Introduction

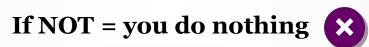
- □ Why a scientist should write a paper?
- Academic publishing of scientific papers:
  - Where and how students can find scientific publications in biology
- **Structure of a typical scientific paper in biological sciences**
- **Guidelines for an effective and critical reading of a scientific paper**
- □ 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 🗸





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:



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

Abdulsamie Hanano <sup>6,4</sup>, Mouhnad Shaban <sup>4</sup>, Ibrahem Almousally<sup>4</sup>, Mahmoud Al-Ktaifani <sup>b</sup> <sup>1</sup>Approved of Micraile Reize and Rockshall, Anno Roz (Constance of Spire (2003) RF 80: 801, Danase S, Spire <sup>1</sup>Approved of Micraile Reize and Rockshall, Anno Roz (Constance of Spire (2003) RF 80: 801, Danase S, Spire

#### HIGHLIGHTS

Article hassay: Excessed 4 September 2014 Excessed in moving finan 3 Newcosher Accepted 5 Newcosher 2014 Available online 27 Newcosher 2014

when I day I when I want

•A RCD-grown S cettrifiae represents an induced microsonal potenti of 593Da.
•A new gree of CMPDA study is closed and therefore refered as CMP2A58 CMP23A58 proteinably catalyses the terminal bytexplation of a beachcane.
•RCD-green year bar a modified competition is not wall manepreteins.
•RCD-green year bar a modified competition is not wall manepreteins.
•RCD-green year bar a modified competition is not in ICD-green SMPX.

#### ARTICLE INFO

A BATERACT International hydrocarbon controllectures has a writer based in hance both Masses. But any supported hydrocarbon, can be reasoned by writers optical of prace. We producedly identified international and the second second second second second second second second and on 100 km second second second second second second second second second and the second dense index (identified) second seco

Petroleum crude-oil (PCD) hydrocarhons are known to have a worying risk on organisms. Their contaminators in soil and aguiter are therefore of great environmental noners. The toxicity of alkanes, a major component of petroleum hydrocarhons, to microorganisms, plant, animals and humans is well established.

http://dx.dxi.org/10.1016/j.chemaphen.2014.11.07 0045-45 Mile 2014 Exercise 244 All rights married

Tel.: + 983 11 21 325 80; fax: +943 11 61 122 81

d) public or intercontension in 2000 fatal yrad Hangama, 2010 yr aid yrad, a balano shonshan yman yrad Hangama, 2010 el increasing attention tut two docados. Heren, few yrait grene namely, Candida, Yarrwak, and Achia were sidawid fonn pett ieun contamaated wrivonamets aid their potential abliens degrade a alkanes were studio. Coverstoally, be capability several ywart species to ure a illiann ii melaind bripeofil eny mate system, the instit acab heir go dhynd y orgenaa mate system, the instit acab heir go dhynd y orgenaa en ywart alkane were studio heir go dhynd y orgenaa mate system, the instit acab heir go dhynd y orgenaa en ywart alkane were studio heir go dhynd y orgenaa en ywart alkane ywart alkane were studio y orgenaa en ywart alkane ywart alkane were studio y orgenaa en ywart alkane ywart alkane were studio en ywart alkane ywart alkane ywart alkane ywart alkane en ywart alkane ywart alkane ywart alkane ywart alkane ywart alkane en ywart alkane ywart alkane ywart alkane were studio en ywart alkane ywart alka **Scientific knowledge** 

# 🖌 Paper design





□ 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?



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



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



### www.elsevier.com



Blocatalysis and Agricultural Blotechnology

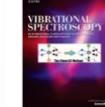


journal of

food engineering

Microbiological

Research



Vibrational

Spectroscopy

of Human

Animal

GENE

Evolutio

Journal of Human Evolution



Comparative

Molecular & Integrative Physiology

POLICY

Current Opinion In Environmental Blochemistry and Sustainability Physiology - Part A:

Environmental Sustainability

3.99 6.34



International Journal for Parasitology

### 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

Nutrition Education and Behavior



**Journal of Food** Morphologle Engineering

MORPHOLOGIE

Animal Gene

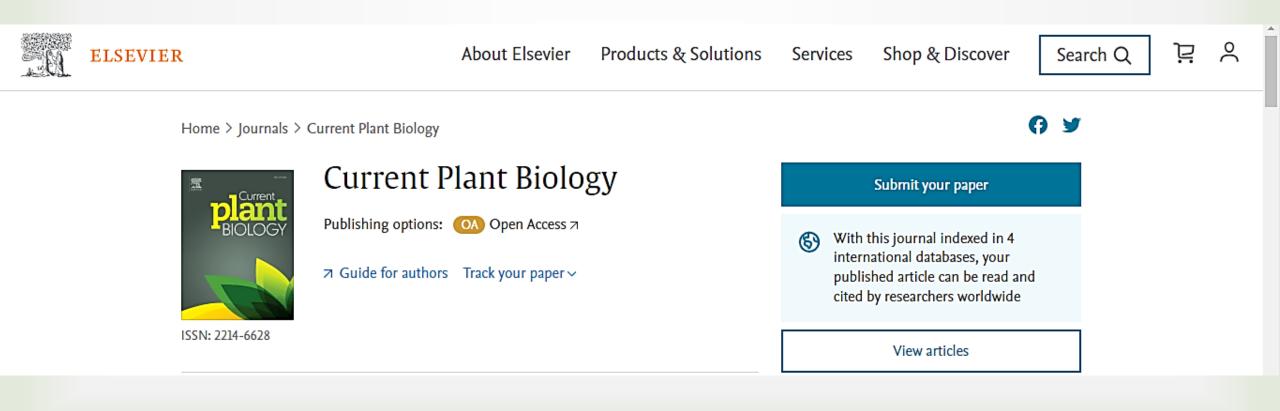






# Where and how students can find scientific publications in biology?

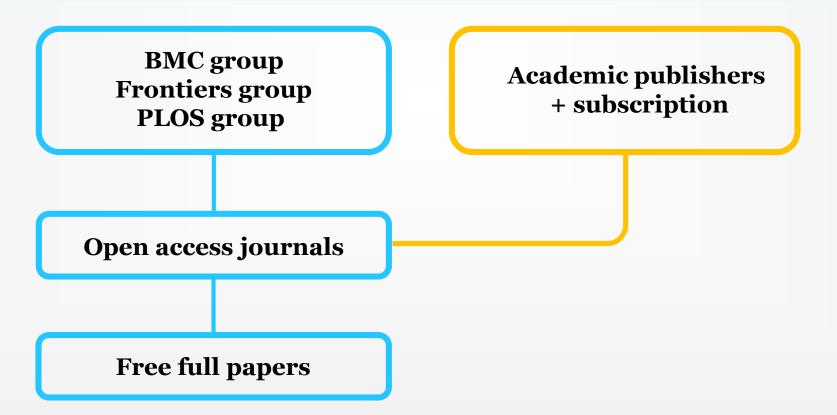
### Life Sciences > Biological Sciences journals > Open Access Journals





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

nature

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

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

ARTICLES computational science https://doi.org/10.1038/543588-022-00263-8 R Check for updates

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

Xiaogiao Chen<sup>1</sup>, Sisi Chen<sup>2,3</sup> and Matt Thomson<sup>01,2,3</sup>

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 gane sets that capture biological information with a minimal number of ganes. Here we introduce an active learning method that identifies minimal but highly informative gane a minimal numeer or genes. Here we increase a surve saming, mennes and sensitive 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 mean application of single-cell mRNA-seq to clinical tests, therapeutic discovery and genetic screens.

Single-cell mRNA-seq methods have scaled to allow routine factorization, and then querying whether the selected genes retain core biological information. I mental 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 mUNA-seq in many exploratory aways such as small-support vector machine classification task (ActiveSVM)<sup>44,1</sup>. The by focustng sequencing resources on highly informative genes for a given biological question or an analysis". Commercial gene-

primen transcriptional programs or modules that contain multiple genes dure actively queries the output of an SVM classifier for cells that regulated by common sets of transcription factors'. Genes within classify poorly, and then performs a detailed analysis of the mistranscriptional modules exhibit correlated gene expression due to classified cells to select maximally informative genes. By selecting co-regulation. Correlations in gene expression can enable the tran- minimal gene sets through a well-defined dassification task, we scriptional state of a cell to be reconstructed through the targeted ensure that the gene sets discovered by ActiveSVM retain biological mRNA profiling of a small number of highly informative genes". information. However, such targeted sequencing approaches require computational methods to identify highly informative genes for specific bio-scale to large single-cell datasets with more than one million cells as logical questions, systems or conditions. A range of computational the procedure focuses computational resources on poorly classified approaches, including differential gene expression analysis and principal components analysis (PCA), can be applied to identify that classify poorly with the current gene set, the method can be highly informative genes". Yet, current methods for defining minimal gene sets are computationally expensive to apply to large singlecell mRNA-seq datasets and often require heuristic user-defined thresholds for gene selectionar. As an example, computational brain dataset with 1.3 million cells in only hours of computational approaches based on matrix factorization (PCA, non-negative time. In addition to scaling, the ActiveSVM classification paradigm matrix factorization) are typically applied to complete datasets and generalizes to a range of single-cell data analysis tasks, including therefore are computationally intensive when datasets scale into the identification of disease markers, genes that respond to Cas9 the millions of cells". Furthermore, gene set selection after matrix perturbation and region-specific genes in spatial transcriptomics. factorization requires heuristic strategies for thresholding coef-

Inspired by active learning' approaches, here we develop a computational method that selects minimal gene sets capable of reliably molecule and genetic screens, and in cost-sensitive clinical assays. ActiveSVM algorithm constructs a minimal gene set through an The sequencing bottleneck has led to the development of targeted tterative cell-state classification task. At each tteration, ActiveSVM mRNA-seq strategies that reduce sequencing costs by up to 90% 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 mistargeting kits, for example, reduce sequencing costs through selec- classified with the current gene set and then identifies maximally tive amplification of specific transcripts using ~1,000 gene-targeting informative genes that are added to the growing gene set to improve classification. Traditional active learning algorithms query an oracle Cells modulate gene expression through the regulation of for training examples that meet a criteria". The ActiveSVM proce-

The central contribution of ActiveSVM is that the method can 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 To demonstrate the performance of ActiveSVM, we apply the ficients in gene vectors extracted by PCA or non-negative matrix method to a series of single-cell genomics datasets and analysis

Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadana, California, USA, "Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, USA, 'Beckman Institute Single-cali Profiling and Engineering Center, Pasadena, California USA, Re-mail: mithomsona/callech.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 П
- **Discussion:** Critical opinion
- Conclusions
- Acknowledgment: Say thank you
- **References**

#### ARTICLES https://doi.org/10.1038/s43588-022-00263-8 R Check for updates

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

Xiaogiao Chen<sup>1</sup>, Sisi Chen<sup>2,3</sup> and Matt Thomson<sup>()</sup><sup>12,3</sup>

nature

computational

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 gans sets that capture biological information with a minimal number of ganse. Here we introduce an active learning method that identifies minimal but highly informative gans a minima number of genes. Pere we introduce an active learning meetide that touchines minima to the give more than to the set that enables 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 softs 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 measu ments necessary for application of single-cell mRNA-seq to clinical tests, therapeutic discovery and genetic screens.

mental run. Although stngle cell mRNA-seq approaches provide insights into many different biological and biomedical by focusting sequencing resources on highly informative genes for a given biological question or an analysis". Commercial genetive amplification of specific transcripts using ~1,000 gene-targeting primers

regulated by common sets of transcription factors'. Genes within classify poorly, and then performs a detailed analysis of the mistranscriptional modules exhibit correlated gene expression due to classified cells to select maximally informative genes. By selecting 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". information. However, such targeted sequencing approaches require computational methods to identify highly informative genes for specific biological questions, systems or conditions. A range of computational the procedure focuses computational resources on poorly classified approaches, including differential gene expression analysis and cells. As the algorithm only analyzes the full transcriptiome of cells highly informative genes. Yet, current methods for defining minimal gene sets are computationally expensive to apply to large singlecell mRNA-seq datasets and often require heuristic user-defined filed cells. We demonstrate that ActiveSVM can analyze a mouse thresholds for gene selections. As an example, computational brain dataset with 1.3 million cells in only hours of computational approaches based on matrix factorization (PCA, non-negative time. In addition to scaling, the ActiveSVM classification paradigm matrix factorization) are typically applied to complete datasets and generalizes to a range of single-cell data analysis tasks, including therefore are computationally intensive when datasets scale into the identification of disease markers, genes that respond to Cars the millions of cells'. Furthermore, gene set selection after matrix perturbation and region-specific genes in spatial transcriptomics. factorization requires heuristic strategies for thresholding coefficients in gene vectors extracted by PCA or non-negative matrix method to a series of single-cell genomics datasets and analysis

Stingle-cell mRNA-seq methods have scaled to allow routine factorization, and then querying whether the selected genes retain core biological information.

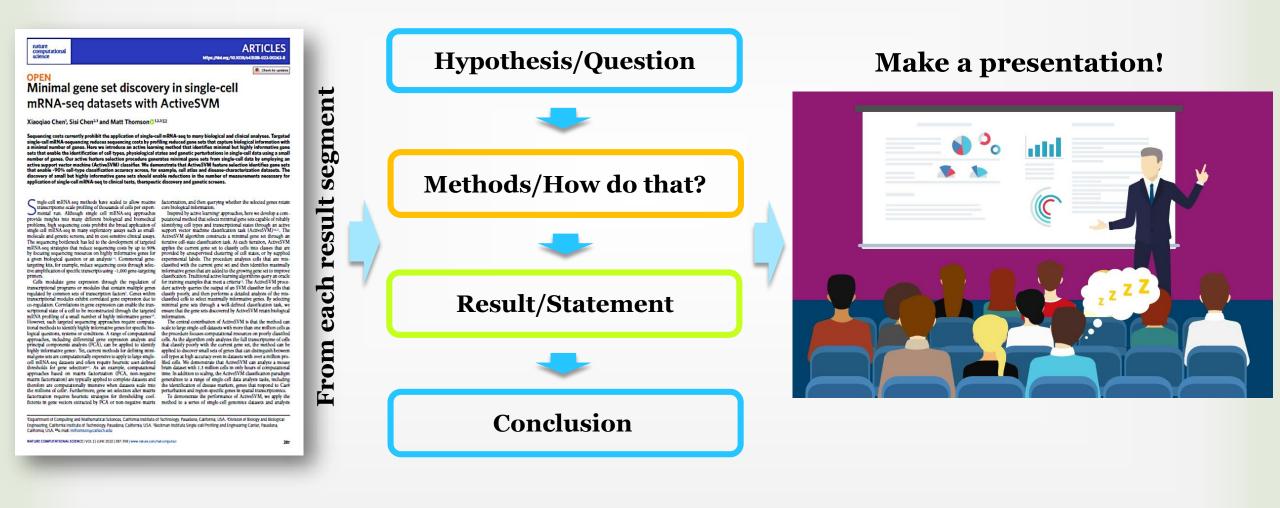
Inspired by active learning' approaches, here we develop a computational method that selects minimal gene sets capable of reliably problems, high sequencing costs prohibit the broad application of identifying cell types and transcriptional states through an active single-cell mIUNA-seq in many exploratory assays such as small-support vector machine dassification task (ActiveSVM)<sup>340</sup>. The molecule and genetic screens, and in cost-sensitive clinical assays. ActiveSVM algorithm constructs a minimal gene set through an The sequencing bottleneck has led to the development of targeted tterative cell-state classification task. At each tteration, ActiveSVM mRNA-seq strategies that reduce sequencing costs by up to 90% 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 mistargeting kits, for example, reduce sequencing costs through selec- classified 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 Cells modulate gene expression through the regulation of for training examples that meet a criteria<sup>10</sup>. The Active5VM procetranscriptional programs or modules that contain multiple genes dure actively queries the output of an SVM classifier for cells that ensure that the gene sets discovered by ActiveSVM retain biological

The central contribution of ActiveSVM is that the method can scale to large single-cell datasets with more than one million cells as principal components analysis (PCA), can be applied to identify 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 pro-To demonstrate the performance of ActiveSVM, we apply the

Department of Computing and Mathematical Sciences. California Institute of Technology. Peradena. California, USA. "Division of Biology and Biological Engineering, California Institute of Technology, Pasadana, California, USA, 'Beckman Institute Single-cali Profiling and Engineering Center, Pasadana, California, USA, Re-mail: mthomsonaicallech.edu



# Guidelines for reading of a scientific paper





# Guidelines for reading of a scientific paper

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

Xiaoqiao Chen<sup>1</sup>, Sisi Chen<sup>2,3</sup> and Matt Thomson<sup>012,3</sup>

nature computationa science

ncing costs currently prohibit the application of single-cell mRNA-seq to many biological and clinical analyses. Targeted c-all mRNA-sequencing reduces sequencing costs by profiling reduced gene sets that capture biological information with mail number of genes. Here we introduce an active learning method that identifies minimal but highly informative single-cell mRNAsets 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 semploying an active support vector machine (ActiveSYM) classifier. We demonstrate that ActiveSVM feature selection identifies gene sets able -90% cell-type classification accuracy across, for example, cell atlas and disease-characte discovery of small but highly informative gene sets should enable reductions in the number of me of single-cell mRNA-seq to clinical tests, therapeutic discovery and genetic scr

C ingle-cell mRNA-seq methods have scaled to allow routine factorization, and then querying whether the selected genes retain cratectorization, and then querying whether the selected genes retain core biological information.

molecule and genetic screens, and in cost-sensitive clinical assays. ActiveSVM algorithm constructs a minimal gene set through an The sequencing bottleneck has led to the development of targeted tterative cell-state classification task. At each iteration, ActiveSVM mRNA-seq artileges that reduce sequencing costs by up to 90%, applies the current gene set to classify cells imit draws that are by focuting sequencing resources on highly informative genes for provided by unsupervised clustering of cell states, or by supplied a given biological question or an analysis<sup>10</sup>. Commercial gene- experimential labels. The procedure analyses cells that are mistargeting kits, for example, reduce sequencing costs through selec-tive amplification of specific transcripts using -1,000 gene-targeting primen.

regulated by common sets of transcription factors'. Genes within classify poorly, and then performs a detailed analysis of the mis-transcriptional modulary informative genes by selecting on classified with the selectin mattainly informative genes. By selecting co-regulation. Correlations in gene expression can enable the transmitting gene sets through a well-defined duaditation task, we scriptional state of a cell to be reconstructed through the targeted mRNA profiling of a small number of highly informative genesity Howvers, such targeted sequencing approaches require computa-tion the central contribution of ActiveSVM is that the method can

Omental run. Although single cell mRNA-seq approaches, provide insights into many different biological and biomedical provide insights into many different biological and biomedical roblems, high sequencing costs prohibit the broad application of ingle-cell mRNA-seq in many exploratory aways such as small- support vector machine classification task (ActiveSVM)<sup>400</sup>. The Cells modulate gene expression through the regulation of for training examples that meet a criteria<sup>10</sup>. The Active/VM proce-transcriptional programs or modules that contain multiple genes dure actively queries the output of an SVM classifier for cells that

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

R. Check for update

tional methods to identify highly informative genes for specific bio-logical questions, systems or conditions. A range of computational the procedure focuses computational resources on poorly classified mal gene sets are computationally expensive to apply to large single-cell types at high accuracy even in datasets with over a million pro-cell mRNA-seq datasets and often require heuristic user-defined thresholds for gene selection<sup>24</sup>. As an example, computational brain dataset with 15 million cells in only hours of computational umonous or gene sector<sup>1</sup>. At al campe, computatoral prim datade visit 3 minuto das in ony nauro o computatoral approades based on matrix factorization (PCA, non-tegritter tunne la dation isolatil), tede AdverSM datadisticon paraligne matrix factorization) are typically applied to complete datasets and generalizes to a range of single-cil data analysis tasks, including beneficar ere computatorally interiorize when datasets starting the the identification of dataset markers, ginen tata reprodu lo Carri the million of celle<sup>4</sup>. Furthermore, gene set selection aller matrix clictorization requestre barratis strategies for thresholding is for thresholding.

ficients in gene vectors extracted by PCA or non-negative matrix method to a series of single-cell genomics datasets and analysis Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadona, California, USA, 'Division of Biology and Biological Engineering, California Institute of Technology, Pasadona, California, USA, 'Beckman Institute Single-cali Profiling and Engineering Cantor, Pasadona, California USA, Re-mail: mthorrsonarcallech.edu

NALSOFICE LVOL 2 LENE 2022 LTRA-358 LVM

# **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



## How a scientific paper is written?

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

#### R. Check for update Minimal gene set discovery in single-cell mRNA-seq datasets with ActiveSVM

Xiaoqiao Chen<sup>1</sup>, Sisi Chen<sup>2,3</sup> and Matt Thomson<sup>012,3</sup>

nature computational science

ing costs currently prohibit the application of single-cell mRNA-seq to many biological and clinical analyses. Targetee Sequencing costs currently promint the application or angle-cal ment-seq to many security and clinical sharpest. Linguista minimal number of genes. Here we initrative an accessible homing method that identifies minimal but highly interactive gene sets that enable the identification of call types, physiological states and genetic perturbations in angle-call data using a samular number of genes. On active statem solutions physiological states and genetic perturbations in angle-call data using a samular state enable the identification of call types, physiological states and genetic perturbations in angle-call data using a samular active support vector machine (ActiveSVM) classifier. We demonstrate that ActiveSVM feature selection identifies gene sets that enable cold call type classifies accuracy across, for sample, call taits and feature. very of small but highly informative gene sets should enable reductions in the number of m ation of single-cell mRNA-seq to clinical tests, therapeutic discovery and genetic screens.

Single-cell mRNA-seq methods have scaled to allow routine factorization, and then querying whether the selected genes retain mental run. Although strate cell mRNA-seq upproaches. Large transport and the selected genes retain figure and the selected genes retain genes and the selected genes genes genes genes genes and the selected genes gen

Cells modulate gene expression through the regulation of for training examples that meet a criteria? The ActiveSVM proce Cells modulae gene expression through the regulation of for training examples has meet a criteria<sup>1</sup>. The ActiveSVM proce-transcriptional programs or modules that contain multiple genes date actively queries the output of an SVM classifier for cell that regulated by common sets of transcription factors'. Genes within corregulation. Correlations target expression can enable the tran-scriptional state of cell to be reconstrained genes capression that the trans-terprivation state of cell to be reconstrained genes of pression that the genes data for correlation target and mRNA profiling of a small number of highly informative genes." The formation of a single supervised in regulation of the genes esite discored by ActiveSVM transcholmed to actual descriptions of the single supervised in state the single state actual state of the single state to actual description of the single supervised in state of the single state actual state of the single state of the single single

provide mights into many different biological and biomedical putational method that selects minimal gene sets capable of rehaby problems, high sequencing costs problem the broad application of salentifying cell systes and transcriptional states broady and single-cill milliNaseq in many exploratory aways such as small- an support vector manhine dashtication task (AttreSVM)<sup>20,20</sup>. The molecule and genetic screens, and in cost-sensitive clinical aways. ActiveSVM algorithm constructs a minimal gene set through an The sequencing bottleneck has led to the development of targeted iterative cell-state classification task. At each iteration, ActiveSVM The expending contents in the to the development of targets in the expending contents in the total development of targets in focusing supervisor of targets in the expension of the expension of

approaches, including differential gene expression analysis and principal components analysis (PCA), can be applied to identify that classify poorly with the current gene set, the method can be highly informative genesi. Yet, current methods for defining mini-mal gene sets are computationally expensive to apply to large single-cell mytes at high accuracy even in datasets with over a million pro-cell mXNA-seq datasets and olient require hearitist user-defined if lide colls. We demonstrate that ArthresVM can analyze a mouse thresholds for gene selection.<sup>27</sup>. As an example, computational approaches based on matrix factorization (PCA, non-negative time. In addition to scaling, the ActiveSVM classification paradigm matrix factorization) are typically applied to complete datasets and generalizes to a range of single-cell data analysis tasks, including herefore are computationally intensive when datasets scale into the milions of cells. Furthermore, gene set selection after marks factorization requires hearistics trainges for thresholding coelficients in gene vectors extracted by PCA or non-negative matrix 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. "Beckman Institute Single-cali Profiling and Engineering Canter, Pasadena, California USA Re.mail: mthomsonescattech.edu

NATURE COMPUTATIONAL SCIENCE | VOL 2 | JUNE 2022 | 387-398 | www.neture.com/halcomp

Parts of a paper are usually ordered by:

Title П

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

Which part is written first?



# How a scientific paper is written?



Xiaoqiao Chen<sup>1</sup>, Sisi Chen<sup>2,3</sup> and Matt Thomson<sup>012,3</sup>

sts currently prohibit the application of single-cell mRNA-seg to many biological and clinical analyses. Targete sequencing total currently points the approximation of supportant marks seq to their provinging a mic tanks an analysis. Enginese ingle-call mRN-sequencing reduces sequencing costs by porfiling reduced gas sets that capture biological information with a minimal number of gases. How we introduce an active learning method that identifies minimal but highly informative gase sets that enable in identification of call types, physiological states and gasets growthowing missional gale-call data taby employing number of gases. Our active feature selection procedure gasenates minimal gase sets from single-call data by employing active support sector machine (ActiveSVI) classific. We downorize that ActiveSVI hatture selection identifies gase sets to main (active support sector machine). enable -90% cell-type classification accuracy across, for example, cell atlas and disease-cha very of small but highly informative gene sets should enable reductions in the number of m application of single-cell mRNA-seq to clinical tests, therapeutic discovery and genetic screens

Single-cell mRNA-seq methods have scaled to allow routine factorization, and then querying whether the selected genes retain structurerytome scale profiling of thousands of cells per open-mental run. Although single cell mRNA-seq approaches provide moghts mino many different bological and bromedical patatonal method that selects immung gene sets capable of relakity The sequencing bottleneck has led to the development of tarseted terrative cell-state classification task. At each iteration, ActiveSVM mRNA-seq strategies that reduce sequencing costs by up to 90% applies the current gene set to classify cells into classes that are by focustry sequencing resources on highly informative genes for provided by unsupervised clustering of cell states, or by supplied a given biological question or an analysis". Commercial gene-targeting kits, for example, reduce sequencing costs through selec-classified with the current gene set and then identifies maximally

Cells modulate gene expression through the regulation of for training examples that meet a criteria<sup>10</sup>. The ActiveSVM procetranscriptional programs or modules that contain multiple genes date activity queries the output of an SVM classifler for cells that regulated by common dusts of transcription factors. Genes within classify poorly, and then performs a detailed anytis of the min-transcriptional modules shifts correlated gene expression due to classific delts to select maximally informative genes. By selecting co-regulation. Correlations in gene expression can enable the tran-scriptional state of a cell to be reconstructed through the targeted ensure that the gene sets discovered by ActiveSVM retain biological mRNA profiling of a small number of highly informative genes". However, such targeted sequencing approaches require computational methods to identify highly informative genes for specific bioformative genes'. Yet, current methods for defining mini-

tre amplification of specific transcripts using ~1,000 gene-largeting primers. Informative genes that are added to the growing gene set to improve classification. Traditional active learning algorithms query an oracle

information. The central contribution of ActiveSVM is that the method can scale to large single-cell datasets with more than one million cells as ogical questions, systems or conditions. A range of computational the procedure focuses computational resources on poorly classified protaches, including differential gene expression analysis and cells. As the algorithm only analyses the full transcriptione of cells intropal comproaching analysis and cells. As the algorithm only analysis the full transcriptione of cells and the 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 promalgine sets are compationally expensive to apply to large imple-ced mRNA-req dutases and other negative borraris user-facilities of the dutases and the AttaceVM can analyze a mouse thresholds for gene electron<sup>24</sup>. As an example, computational brain dataset with 1.3 million colis in only hours of computational approaches based on matrix factorization (FOA, non-negative time in addition to scaling, the AttaceVM can attachen paradaging approaches based on matrix factorization (FOA, non-negative time in addition to scaling, the AttaceVM can attachen paradaging approximation is an applicable applicable of complete datasets and generalizes to a range of single cell data analysis tasks, including therefore are computationally intensive when datasets scale into the identification of disease markers, genes that respond to Cars the millions of cellsr. Furthermore, gene set selection alter matrix perturbation and region-specific genes in spatial transcriptomics. Tactorization requires heuristic strategies for thresholding coef-ficients in gene vectors extracted by PCA or non-negative matrix method to a series of impleced genomes datasets and analysis

Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadona, California, USA, "Division of Biology and Biological earing, California linstitute of Technology, Pasadena, California, USA. "Backman Institute Single-cali Profiling and Engineering Cantor, Pasaderu amia, USA. "Ae-mat: mthomsong-calitech.edu

NATURE COMPUTATIONAL SCIENCE | VOL 2 | JUNE 2022 | 387-398 | www.nature.com/tolcompute

Writing a paper is usually proceeded in the following order:

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



## How a scientific paper is written?



Sequencing costs currently prohibit the application of single-call mRAr-seq to many biological and clinical analyses. Targeted single-call mRAr-sequencing reduces sequencing costs by profiling reduced genes sort that capture biological information with a minimal number of genes. Here we infroduce an active learning method that identifies minimal bet highly informative genes sets that enable the identification of call types, physiological states and genetic perturbations in single-call data tasing a small number of genes. Our active fasture selection procedure generates minimal gene sets from single-call data tasing as active support vector maching ActiveSVO classific. We demonstrate that ActiveSVM harbure selection identificages are statice support vector maching ActiveSVO classification we demonstrate that ActiveSVM harbure selection identificages are statices support vector maching ActiveSVO classification we demonstrate that ActiveSVM harbure selection identificages are statices and the statice selection procedure generates minimal genes sets from single-call data tas genes statices support vector maching ActiveSVO classification we demonstrate that ActiveSVM harbure selection identificages statices and the statices and the statices and the statices are statices and the statices and the statices and the statices are statices and the statices and the statices are statices and the statices and the statices and the statices are statices and the statices and the statices are statices and the statices are statices are statices and the statices are statices are statices and the statices are associated statices are active support vector mathine occurs y in classifier, the demonstrate that active support that enable -90% cell-type classification accuracy across, for example, cell atlas and disease-char discovery of small but highly informative gene sets should enable reductions in the number of me application of single-cell mRNA-soq to clinical tests, therapeutic discovery and genetic screens.

Single cell mRNA seq methods have scaled to allow routine factorization, and then querying whether the selected genes retain factorization, and then querying whether the selected genes retain factorization, and then querying whether the selected genes retain factorization, and then querying whether the selected genes retain factorization, and then querying whether the selected genes retain factorization, and then querying whether the selected genes retain factorization, and then querying whether the selected genes retain factorization, and then querying whether the selected genes retain factorization, and then querying whether the selected genes retain factorization, and then querying whether the selected genes retain factorization, and then querying whether the selected genes retain factorization, and then querying whether the selected genes retain factorization, and then querying whether the selected genes retain factorization, and then querying whether the selected genes retain factorization, and then querying whether the selected genes retain factorization fac

problems, high sequencing costs probabilities broad application of sidentifying cell types and transcriptional status through an active single-cell mRNA-seq in many exploratory aways such as small — support vector machine datafication task (ActiveSVM)<sup>400</sup>. The molecule and genetic screens, and in cost-sensitive Critical aways. A ActiveSVM alignenthm constructs are initial agenes est through an The sequencing bottleneck has led to the development of tarseted tterative cell-state classification task. At each tteration, ActiveSVM

co-regulation. Correlations in gene expression can enable the tran-scriptional state of a cell to be reconstructed through the targeted ensure that the gene sets through a well-defined dassefication task, we mRNA profiling of a small number of highly informative genes". Information. The central contribution of ActiveSVM is that the method can The central contribution of ActiveSVM is that the method can be accessed as a second seco

mRNA-seq strategies that reduce sequencing costs by up to 90% applies the current gene set to classify cells thin classes that are by focusting sequencing resources on highly informative genes for provided by unsupervised clustering of cell states, or by supplied a given biological question or an analysis". Commercial gene-targeting kis, for example, reduce sequencing costs through selec-clastified with the current gene set and then identifies maximally tive amplification of specific transcripts using -1,000 gene-largeting informative genes that are added to the growing gene set to improve classification. Traditional active learning algorithms query an oracle Cells modulate gene expression through the regulation of for training examples that meet a criteria<sup>12</sup>. The ActiveSVM procetranscriptional programs or modules that contain multiple genes dare actively queries the output of an SVM classifler for cells that regulated by common sets of transcription factors'. Genes within classify poorly, and then perform a detailed anytis of the mi-transcriptional modules exhibit correlated gene expression due to dualitied cells to sets maximally informative genes. By selecting

tional methods to identify highly informative genes for specific bio-scale to large single-cell datasets with more than one million cells as logical questions, systems or conditions. A range of computational the procedure focuses computational resources on poorly classified approaches, including differential gene expression analysis and cells. As the algorithm only analyses the full transcriptione of cells principal components analysis (CA), can be applied to identify that classify poorly with the current gene set, the method can be principal composition analysis (rCA), can be approved to saming) that causing poorty with the current gives set, the method can be highly informative gives. Yet, current muchods for defining multi-applied to double current multi-set of general hash can during multi-between multi-general sets are compatibility appendive in apply to longe might-cold mRNAs equilates and often requires howing to set of the definition of the more a milline provide the more and multi-approaches based for gene selections<sup>2</sup>. As an example, compatibility and muse thresholds for gene selections<sup>2</sup>. As an example, compatibility the multi-asset with 1.3 million of his north have of compatibility approaches based on matrix factoration (ICC), non-negative time in haldbaset with 1.3 million of his north have of compatibility approaches based on matrix factoration (ICC), non-negative time in haldbaset with 1.3 million of his north have of compatibility approaches based on matrix factoration (ICC), non-negative time in haldbaset with 1.3 million of his north have of compatibility approaches based on matrix factoration (ICC), non-negative time in haldbaset with 1.3 million of his north have of compatibility approaches based on matrix factoration (ICC) non-negative time in haldbaset with 1.3 million of his north have of compatibility approaches based on matrix factoration (ICC) non-negative time in haldbaset with 1.3 million of his north have of compatibility of the matrix of the m important factor reaction are typically applied to complete datasets and generalizes to a range of single-cell data analysis tasks, including therefore are computationally intensive when datasets scale into the identification of disease markers, genes that respond to Carr the millions of cells. Furthermore, gene set selection aller matrix factorization requires hearinic strategies for thresholding coef-floretism nge wetches extracted by PCA or non-negative matrix method to a series of stuffe-Cell genesis datasets and analysis

Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadona, California, USA, "Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, USA. 'Backman Institute Single-cali Profiling and Engineering Canter, Pasadena California, USA. No. mat: mthomsong califoch edu

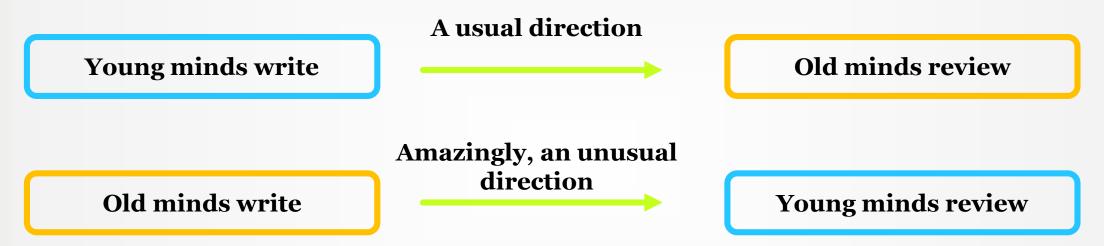
NATURE COMPUTATIONAL SCIENCE | VOL 2 | JUNE 2022 | 387-398 | www.nature.com/tofcomputed

# **Skills acquired:**

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



# 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



# **Critical reading and reviewing process**



Thank you..