The 34th International Biology Olympiad\_UAE\_23



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

### "Open Biology Towards Sustainable Thinking"

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

#### **IBO\_UAE\_23\_Educational Conference**

### OUTLINE

#### Introduction

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



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

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

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

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

Making biological knowledge freely accessible for everyone



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

#### Making biological knowledge freely accessible for every surfer

Is this mission is achieved?

#### We can say YES, but not completely Over 30% of all scholarly articles are actually published as paid-for open access

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

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

5

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

#### Is open biology accessible for IBOs students?

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

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

IBO\_UAE\_23\_Educational Conference



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

#### Yes, definitely

#### **Because open biology offers:**

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



Where and how students can find scientific publications in biology?

#### This is the first question we try to answer and discus in the welcome lecture with our news National Biology Olympians



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

There are lots of scientific publications in different branches of biology



Where we find open-accessed scientific publications?



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

Top 10 academic publishers 2023:

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

Published more than a half of peer-reviewed academic papers



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



#### www.elsevier.com



Blocatalysis and Agricultural Blotechnology



Research

**Journal of Food** 

Engineering

Microbiological

Vibrational Spectroscopy

Journal of Human Evolution

of Human

Evolutio



Comparative

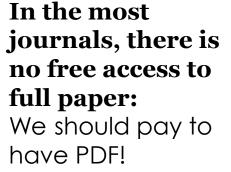
Molecular & Integrative Physiology

Blochemistry and



Environmental Sustainability

Current Opinion In Environmental Sustainability Physiology - Part A:







Journal of Nutrition Education and Behavlor



Morphologle





Resources Policy





RESOURCES

POLICY



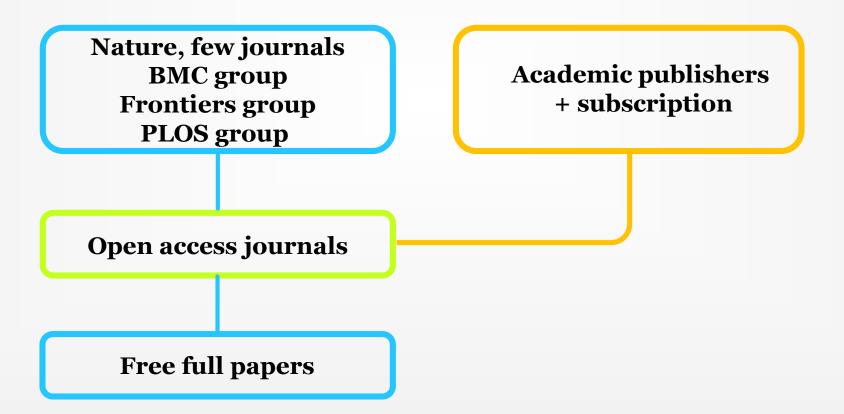


IJP



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

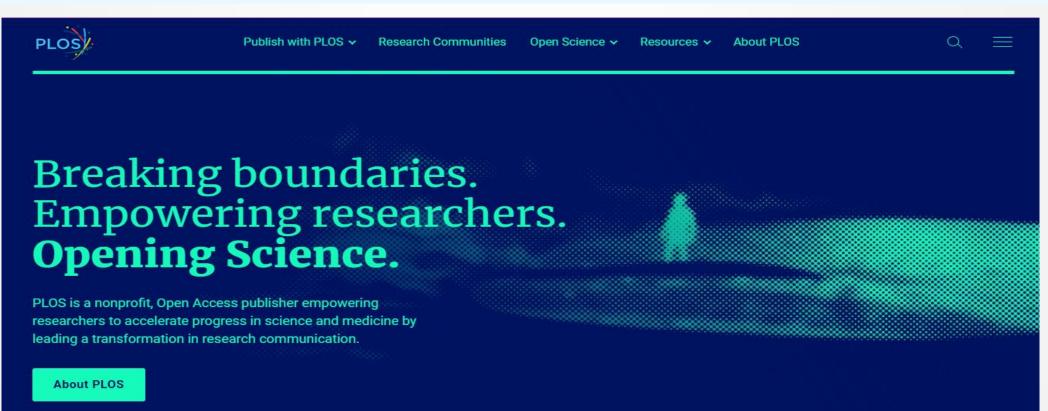
**Top academic publishers with Open Access policy:** 





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

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



PUBLISH

ABOUT

SEARCH

advanced search



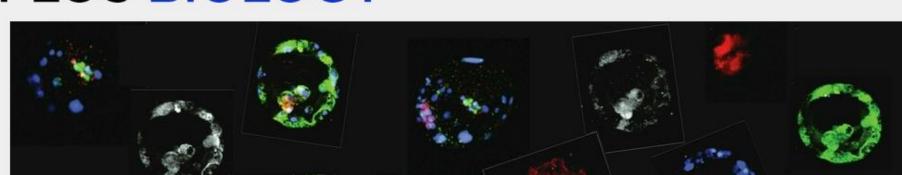
## Academic publishers of Open Biology

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

Life Sciences > Biological Sciences journals > Open Access Journals

BROWSE





### A new human embryonic cell type

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



### **U** Structure of a typical scientific paper in biological sciences

#### Guidelines on how students can be familiar with scientific paper

#### A scientific paper provides information on:



mRNA-seq datasets with ActiveSVM

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

promote that applications of rainings-call minor L-radge to many absolution and produces assequencing costs by profiling reducting gains set of the capture be carbino in coll types, physiological status and ganadic parturbations in sign in a charter selection procedure gasareture ninning agene sets from ingle-nies (ActiveSVM) classifies. We demonstrate that ActiveSVM feature selection calification accuracy accurs, for example, cal altas and disasse-chara by informative gana sets should anable reductions in the number of man Varianti gains acts should anable reductions in the number of man Varianti gains acts should anable reductions in the number of man

gle cell mRNA seq methods have scaled to allow routine. Excitoration, and then querying whother the selected genes retain maneprone, ack profiling of Documeh of cells per apper. Constraints and the selection of the selection

ActiveSVM algentia contracts a mitman gene set through a question to contracts a mitman gene set through a question to contract the question of the question to the question of question to the question of question to the question of question to question of question to question of question transcription (Gaussian et al., question), and question of question transcription (Gaussian et al., question), and question of question transcription (Gaussian et al., question), and question of question transcription (Gaussian et al., question), and question of question transcription (Gaussian et al., question), and question of question transcription (Gaussian), and question of question transcription (Gaussian et al., question), and question of question transcription (Gaussian), and question of question transcription (Gaussian), and question (Gaussian), and question (Gaussian), question (Ga proaches based on maritx factorization (PCA, non-negative time. In addition to scaling, the ActiveSVM classification paradigr ritation) are typically appled to complete dataset and generalizes to a range of engle-cell data analysis tasks, including computationally interime when datasets scale into of celler, Furthermore, gene at selection after materia requires barruits, trainages in thresholding cod-to dataset and the performance of AttueEVM, we apply the 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 Sichnology, Pasalara, California, USA. "Oxidan of Biology and Biologics Implements, California Institute of Hochnology, Pasadora, California, USA. "Bioclima institute Single- cali Profiling and Engineering Carlier, Pasadora, California, USA, #main: Imbiosciegnical Localu

AT HE COMPANY AND A STREET WAS A LONG TO LONG THE AND LONG TO LONG THE ADDREET AND ADDREET AND ADDREET ADDREET

✓ Scientific knowledge

**V** Paper design

**Scientific language** 



### Structure of a typical scientific paper in biological sciences

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

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

nature ARTICLES computational 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>012,3</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 number of genes. Here we introduce an active angument generating and the 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) classificar. We demonstrate that ActiveSVM feature selection identifies gene sets that enable -90% cell-type classification accuracy across, for example, cell attas and disease-characterization datasets. The discovery of small but highly informative gene sets should enable reductions in the number of meas application of single-cell mRNA-seq to clinical tests, therapeutic discovery and genetic screens.

C ingle-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 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 tierative cell-state classification task. At each tieration, ActiveSVM by focustng sequencing resources on highly informative genes for a given biological question or an analysis". Commercial genetargeting 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 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 classification 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 biological 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 selection\*7. 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 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 misclassification. 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 proce-

The central contribution of ActiveSVM is that the method can scale to large single-cell datasets with more than one million cells as 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

381

Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, California, USA, 4Division 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



### Guidelines for reading of a scientific paper

#### A scientific paper is typically composed of:

#### **D** 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 computational 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>012,3</sup>

nature

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 measur ments 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 factorization, and then querying whether the selected genes retain mental run. Although single cell mRNA-seq approaches. Inspired by active karming approaches, here we develop a com-

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

scriptional 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 selection#7. 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

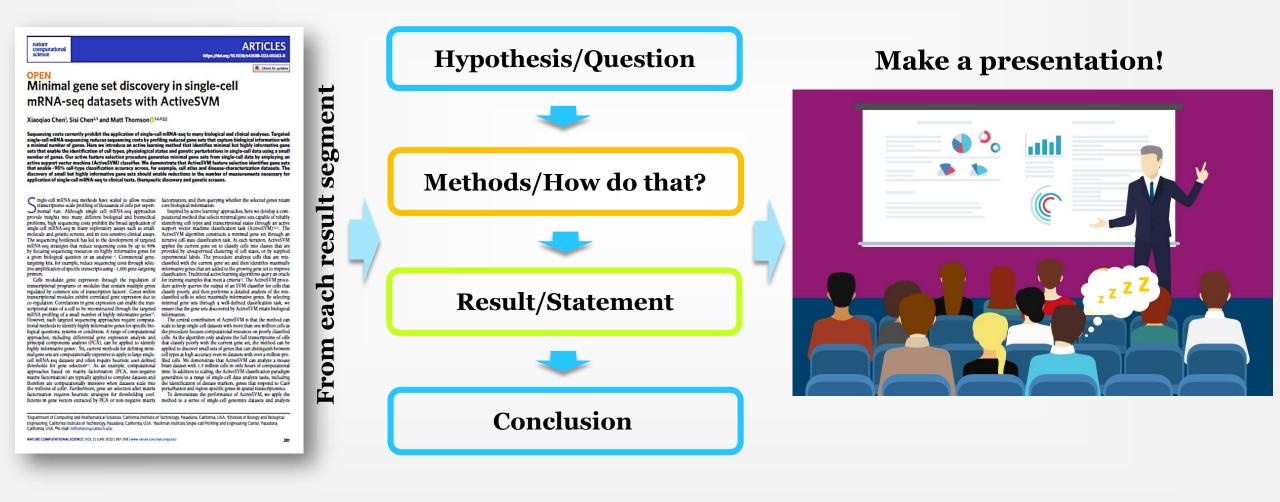
putational method that selects minimal gene sets capable of reliably problems, high sequencing costs prohibit the broad application of sidentifying cell types and transcriptional states through an active single-cell mIUNA-seq in many exploratory aways such as small-support vector machine classification task (ActiveSVM)<sup>34,1</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 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 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, Pasadana, California, USA, "Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, USA, 'Backman Institute Single-cell Profiling and Engineering Canter, Pasadena, California, USA, Re-mail: mthomsonaicallech.edu



### Guidelines for reading of a scientific paper





### Guidelines for reading of a scientific paper



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 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 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 iscovery of small but highly informative gene sets should enable reductions in the number of ma of single-cell mRNA-seq to cli inical 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 care 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 artilogies 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 holds on the constant of the classified cells to select matrixing informative genes. By selecting co-regulation. Correlations in gene expression can enable the transmitting irre 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 ngle-cell mRNA-seq in many exploratory aways such as small-support vector machine classification task (ActiveSVM)<sup>40,10</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

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 matrixes, ginen tata reprodu lo Carr the million of cells<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, Pacadena, California, USA, 'Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, USA, 'Beckman Institute Single-cali Profiling and Engineering Cantor, Pasadena, California USA, Re-mail: mthorrsonarcallech.edu

NALSOFICE LVOL 2 LENE 2022 LTRA-358 LVM

#### **Skills acquired:**

- To gain information on the construct of a scientific paper
- To be familiar with an advanced scientific language
- To gain more an elaborated scientific background
- To be familiar with different forms of illustrations

**To prepare a presentation** 

And more importantly, to talk and discuss with others



#### How a scientific paper is written?



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

ng costs currently prohibit the application of single-cell mRNA-seq to many biological and clinical analyses. Targeted Sequencing cost currently prohibit the application of single-call mRNA-seq to many biological and clinical analyses. Targeted single-call mRNA-sequencing endoces sequencing cost to by profiling endoced goes sets that capation biological information with start that anable the identification of call types, physiological extrans and genetic perturbations in ingle-call data using a small member of genese. On active startner solection proceeding segments and impact and genese that can start that anable the identification of call types, physiological extrans and genetic perturbations in ingle-call data using a small active support vector machine (ActiveSVM) Classifier. We demonstrate that ActiveSVM feature selection identifier gene sets that enable for call-type called Starts (or call starts, for example, call acts and genese-tharterized features). The discovery of small bot highly informative genes sets thom and genetic features and papification of single-call mRNA-seq to clinical texts, themport discovery and genetic to comes.

Single-cell mRNA-seq methods have scaled to allow routine factorization, and then querying whether the selected genes retain mential run. Mhodps intigle cell mRNA-seq approaches. Inspeed by active learning approaches, here we develop a com-provide mights into many different biological and biomedical practional mission tastics minimal gene sets capable of relative problems, they specimency costs problem for the biotograd and biomedical practice and the sets capable of the set of problems, they specimency costs problem is they specimency costs problem is the rout application of 'identifying an active and the set of the set of the set of the sets of the set single-cell mRNA-seq in many exploratory aways such as small-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 tarseted tterative cell-state classification task. At each tteration, ActiveSVM The sequencing bottleneck has led to the development of targeted a matter cell-state classification task. At each instance, Activa's M mRNA-seq strategies that reduce sequencing costs by por 90%, applies here current genes et al. (a dask of tar at by focuring sequencing neurons on highly informative genes for given biological question or an analysis<sup>10</sup>. Commercial genes. ereptermeter allelables. The proceeding analysis<sup>10</sup> and that are mis-trageting kits, for example, reduce sequencing costs through selec-classified with the current gene set and then identifies maximally

transcriptonal programs or modules that contain multiple genes dure actively queries the output of an SVM classifier for cells that regulated by common sets of transcripton factors'. Genes within classify poorly, and then perform a detailed analysis of the mi-transcriptional modules eithin correlated gene expression due to classific periods to set 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 ensure that the gene sets discovered by ActiveSVM retain biological mEXA preding of a small anaber of highly informative generation in the state of the principal composition analysis (CCC), and the applicable datamay highly microardine geness. Nex current herebooks of defining mini-aryptical observer stall sets of geness that can be applied by the appendix. The applicable data and the applicable between mail gene sets are computationally expensive to apply to large might-cell mRXN-set quarks and only mergen bearries used with the applicable data and the approaches based thresholds for gene selectory. As an example, computational based data is a set of the appendix of the appendix of the appendix on south ghe devices YM data data for open-ticable of the appendix matching of the appendix of the

tre amplification of specific transcripts using -1,000 gene-targeting informative genes that are added to the growing gene estio improve classification. Traditional active learning algorithms query an oracle Cells includate gene expression through the regulation of for transmise sharmest a citrary. The ActiveSVM proce-

approaces raises on markin taxonatore (irCA), non-negaries ume in autocomo is cautto, tai excess sis cautto, parategin markin faciotzatora in etypically application complete datacates acid generalizes to a range of mighe cill data analysis takis, including berefore are compariationally intensive when datacets scale into the identification of disease markers, gene that respond to Casi berefore are compariationally intensive when datacets scale into the identification of disease markers, gene that respond to Casi berefore are compariationally intensive scale scalescolar and markin-specific genesis in gatail transcriptionics factorization requires hearings for thresholding coef-ficients in gene vescine entitied by PAC or non-negative markin: moltool to a series of intiple-cill protonics datacets and analysis

Department of Computing and Mathematical Sciences, California Institute of Technology, Pacadena, California, USA, "Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, USA, "Backman Institute Single-cal Profiling and Engineering Canter, Pasadena, California, USA, He-mail: mthoreonigicallech.edu

NATURE COMPUTATIONAL SCIENCE I VOL 2 LENE 2022 | 382, 358 | www.tature.com/todown

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?

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

#### 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>012,3</sup>

nature computational science

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 gasentase minimal gase sets from single-call data by employing active support sector machine (ActiveSVI) classific. We downorize that ActiveSVI hatture selection identifies gase sets active support sector machine (ActiveSVI) classific. We downorize that ActiveSVI hatture selection identifies gase sets active support sector machine (ActiveSVI) classific. We downorize that ActiveSVI hatture selection identifies gase sets active support sector machine (ActiveSVI) classific. We downorize that ActiveSVI hatture selection identifies gase sets active support sector machine (ActiveSVI) classific. We downorize that ActiveSVI hatture selection identifies gase sets active support sector machine (ActiveSVI) classific. We downorize that ActiveSVI hatture selection identifies gase sets active support sector machine (ActiveSVI) classific. We active SVI hatture SVI hatture selection identifies gase sets active support sector machine (ActiveSVI) classific. We setter setter sector se 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 olication of single-cell mRNA-see 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 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 Cells modulate gene expression through the regulation of for training examples that meet a criteria<sup>10</sup>. The ActiveSVM proce-

transcriptional 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 schibt 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 biogical questions, systems or conditions. A range of computational nformative genes". Yet, current methods for defining mini-

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 classifies provaches, including differential gene expression analysis and cells. As the algorithm only analyzes the full transcriptione of cells intrictial components analysis (PCA), can be applied to identify that classify poorly with the current 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-florests in gene vectors extracted by PCA or non-negative matrix method to a series of imple-cell genemics datasets and analysis

Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadona, California, USA, "Division of Biology and Biological earing, California Institute of Technology, Pasadena, California, USA. "Beckman Institute Single-cali Profiling and Engineering Canter, Pasaden amia, USA. "Ae-mat: mthomsong-califoch-edu

NATURE COMPUTATIONAL SCIENCE | VOL 2 | ILINE 2022 | 387-398 | www.halum.com/halcompu

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

20



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

#### 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 in the factor synthese static in blanching genes 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 and the statices are static 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-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 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 factor

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 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

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 targeted 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 thio clases 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". 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-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<sup>12</sup>. The ActiveSVM procetranscriptional programs or modules that contain multiple genes dare actively queries the output of an SVM classifier for cells that regulated by common sets of transcription factors'. Genes within classify poorly, and then perform a detailed analysis of the mu-transcriptional modules schibt correlated gene expression due to classific dells to select maximally informative genes. By selecting

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 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

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 Cantor, Pasaden California, USA. "As mal: mthomsongicaltech.edu

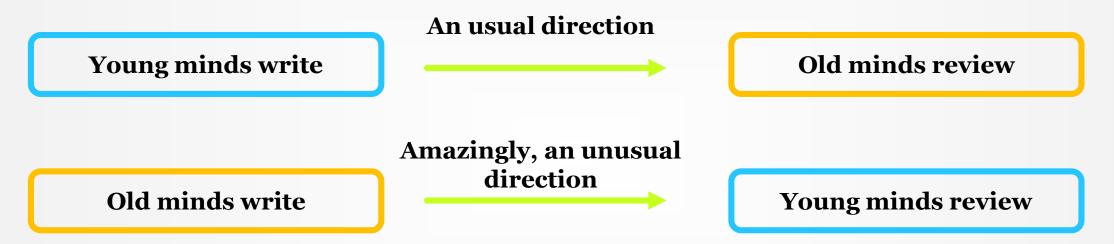
NATURE COMPUTATIONAL SCIENCE | VOL 2 | JUNE 2022 | 387-398 | www.hature.com/hatcomputer

#### **Skills acquired:**

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

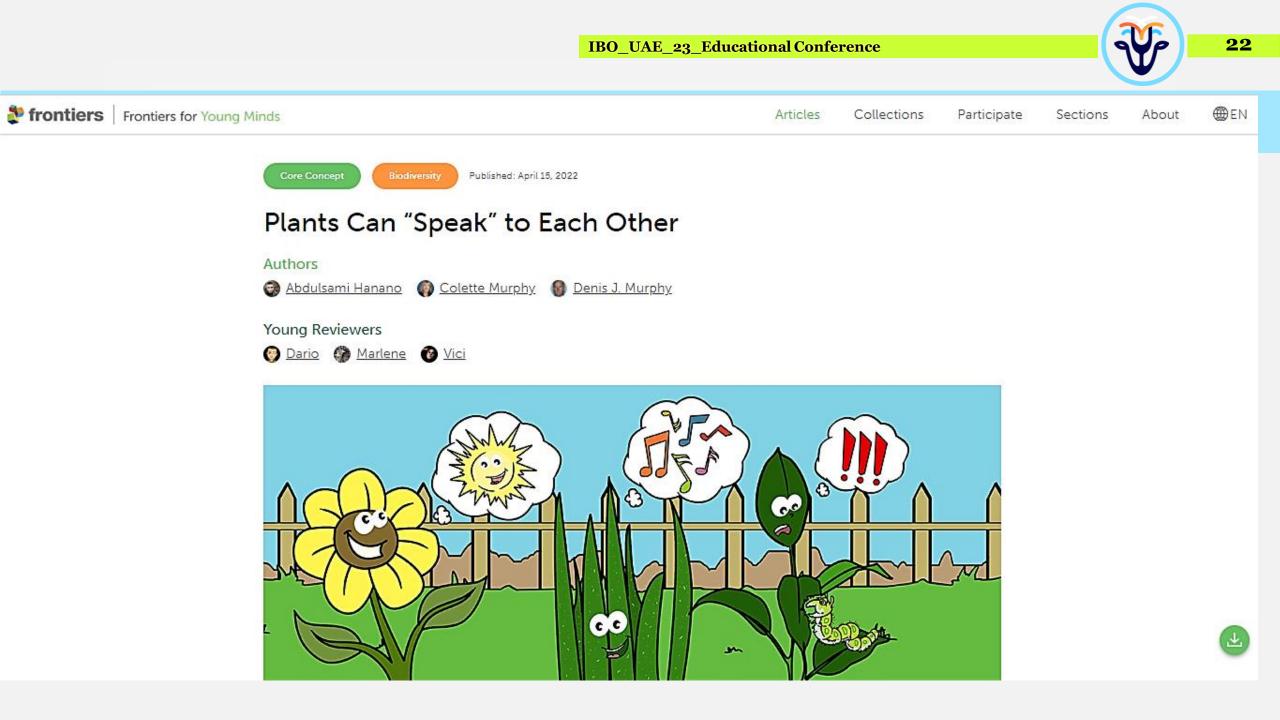
#### □ A brief highlight on the writing of a scientific paper: order of process

#### Critical reading and reviewing process





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



**IBO\_UAE\_23\_Educational Conference** 



### "Open Biology Towards Sustainable Thinking"

Dr. Prof. Abdulsamie Hanano President of Scientific Committee for Syrian Olympiad Biology Head of Toxicology & Biochemistry Division AECS, Damascus, Syria Email: <u>ashanano@aec.org.sy</u>

Thank you..