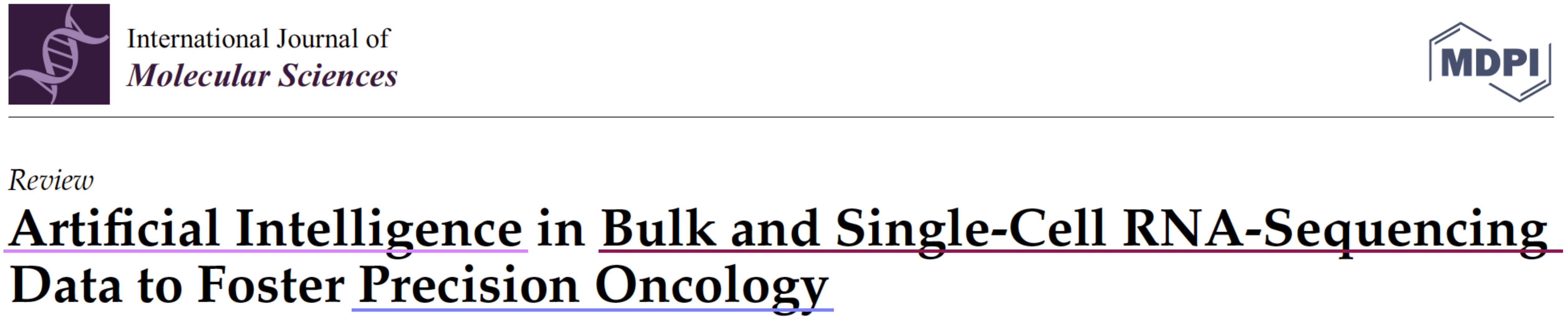
Understanding a review (while learning about machine learning;)

2024-04-18

Explore the title… what we will read about?



International Journal of *Molecular Sciences*

MDPI

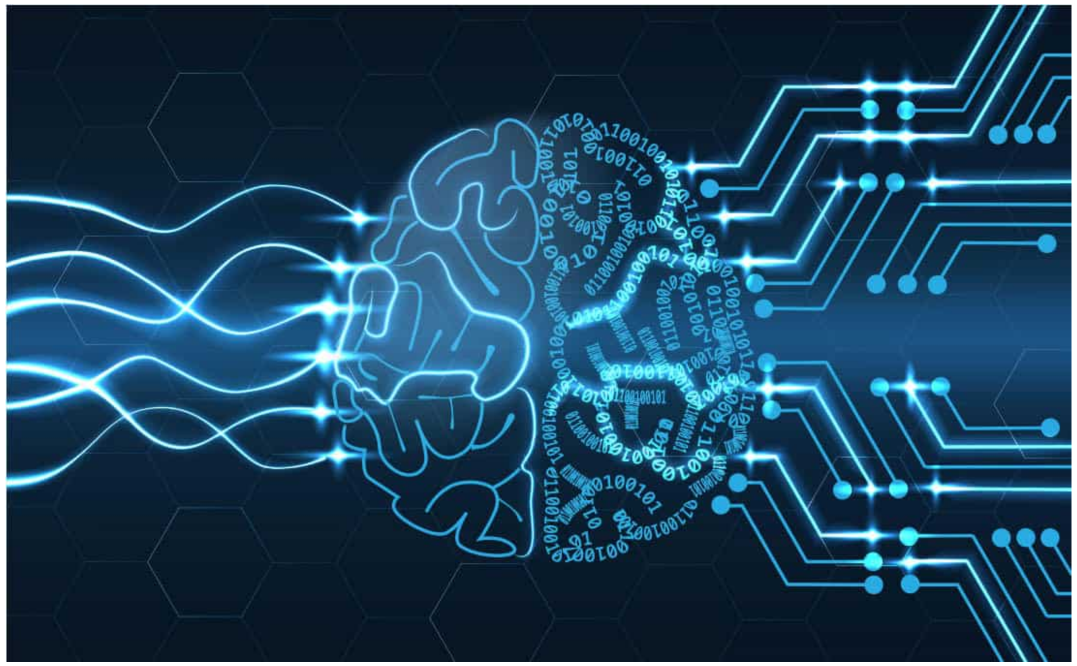
Review

Artificial Intelligence in Bulk and Single-Cell RNA-Sequencing Data to Foster Precision Oncology

Explore the title… what we will read about?

Artificial Intelligence

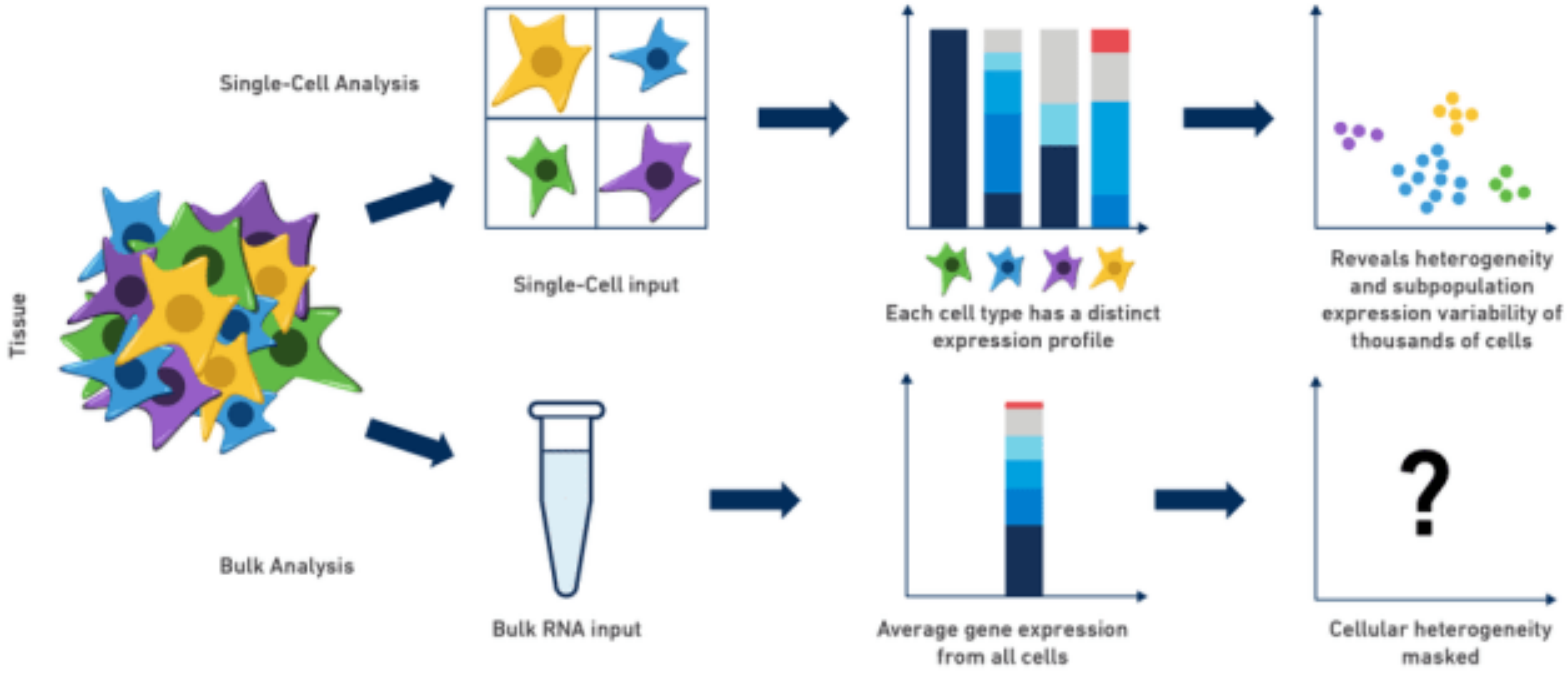
Artificial intelligence (AI) is the capability of a computer system to mimic human cognitive functions such as learning and problem-solving. Through AI, a computer system uses math and logic to simulate the reasoning that people use to learn from new information and make decisions.



Explore the title… what we will read about?

Bulk and Single-Cell RNA-Sequencing

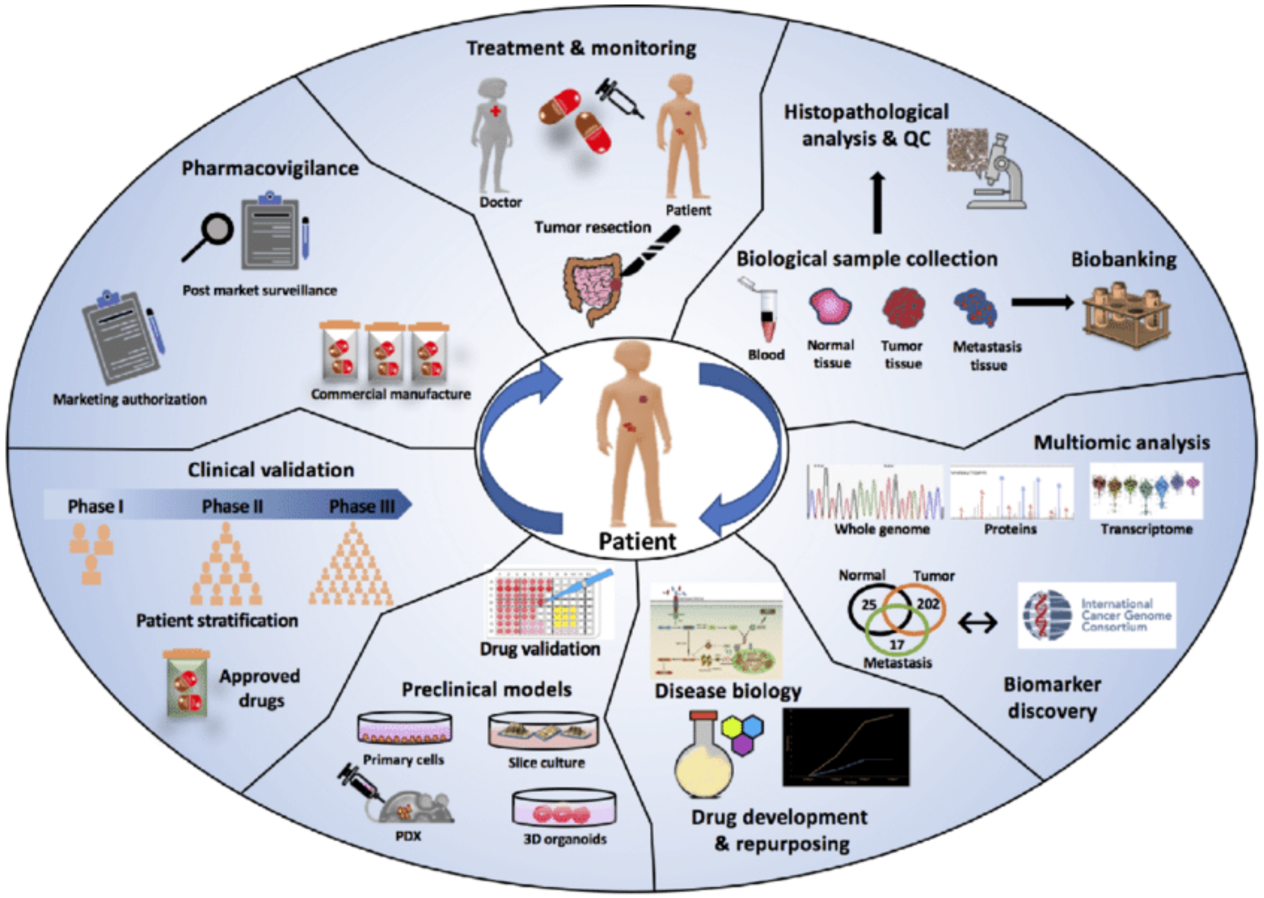
Bulk RNA-seq captures the average of expression profiles of thousands of cells, while **single-cell RNA-seq** allows the capture of individual measurements.


Tissue
Single-Cell Analysis
Single-Cell input
Each cell type has a distinct expression profile
Reveals heterogeneity and subpopulation expression variability of thousands of cells
Bulk Analysis
Bulk RNA input
Average gene expression from all cells
?
Cellular heterogeneity masked

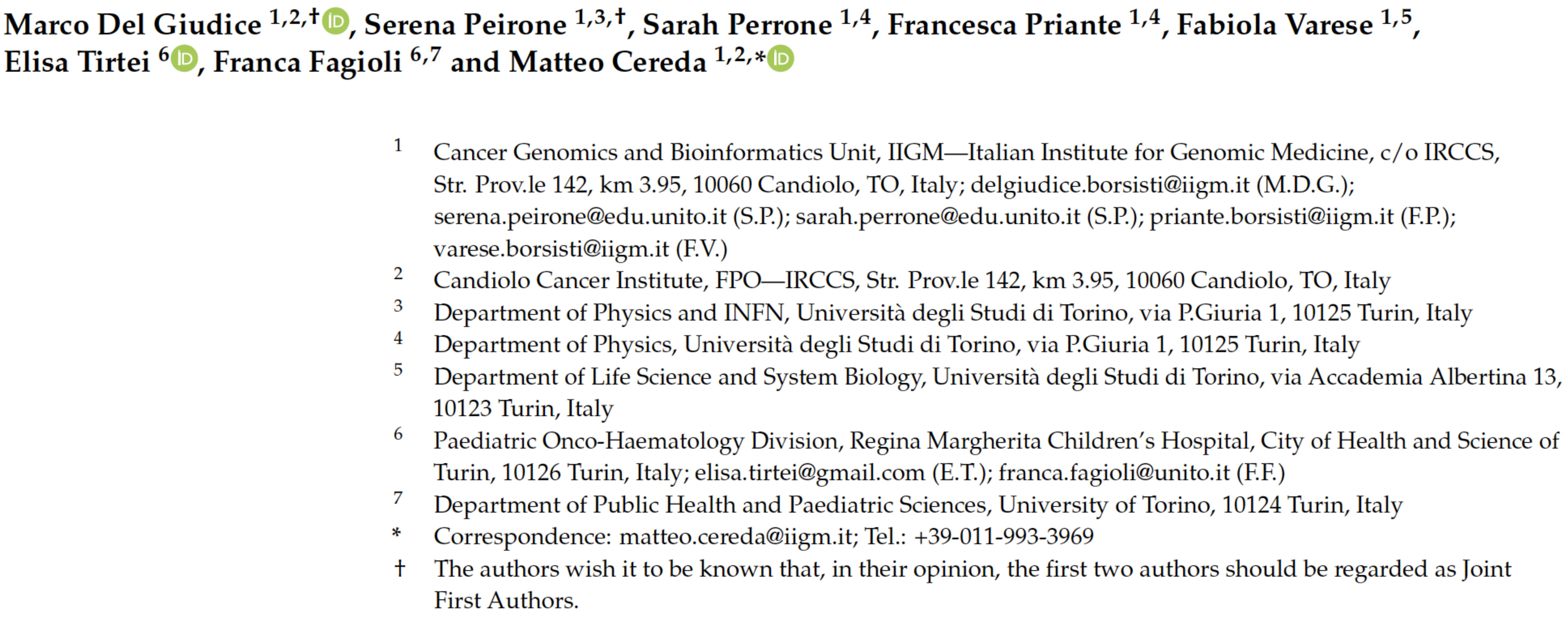
Explore the title… what we will read about?

Precision Oncology

Precision oncology is the molecular profiling of tumours to identify targetable alterations. The goal of precision medicine is to **deliver** the **personalised treatment** to **each patient**.



Explore the authors… do you know someone?



Marco Del Giudice [1,2,†], Serena Peirone [1,3,†], Sarah Perrone [1,4], Francesca Priante [1,4], Fabiola Varese [1,5], Elisa Tirtei [6], Franca Fagioli [6,7] and Matteo Cereda [1,2,*]

[1] Cancer Genomics and Bioinformatics Unit, IIGM—Italian Institute for Genomic Medicine, c/o IRCCS, Str. Prov.le 142, km 3.95, 10060 Candiolo, TO, Italy; delgiudice.borsisti@iigm.it (M.D.G.); serena.peirone@edu.unito.it (S.P.); sarah.perrone@edu.unito.it (S.P.); priante.borsisti@iigm.it (F.P.); varese.borsisti@iigm.it (F.V.)

[2] Candiolo Cancer Institute, FPO—IRCCS, Str. Prov.le 142, km 3.95, 10060 Candiolo, TO, Italy

[3] Department of Physics and INFN, Università degli Studi di Torino, via P.Giuria 1, 10125 Turin, Italy

[4] Department of Physics, Università degli Studi di Torino, via P.Giuria 1, 10125 Turin, Italy

[5] Department of Life Science and System Biology, Università degli Studi di Torino, via Accademia Albertina 13, 10123 Turin, Italy

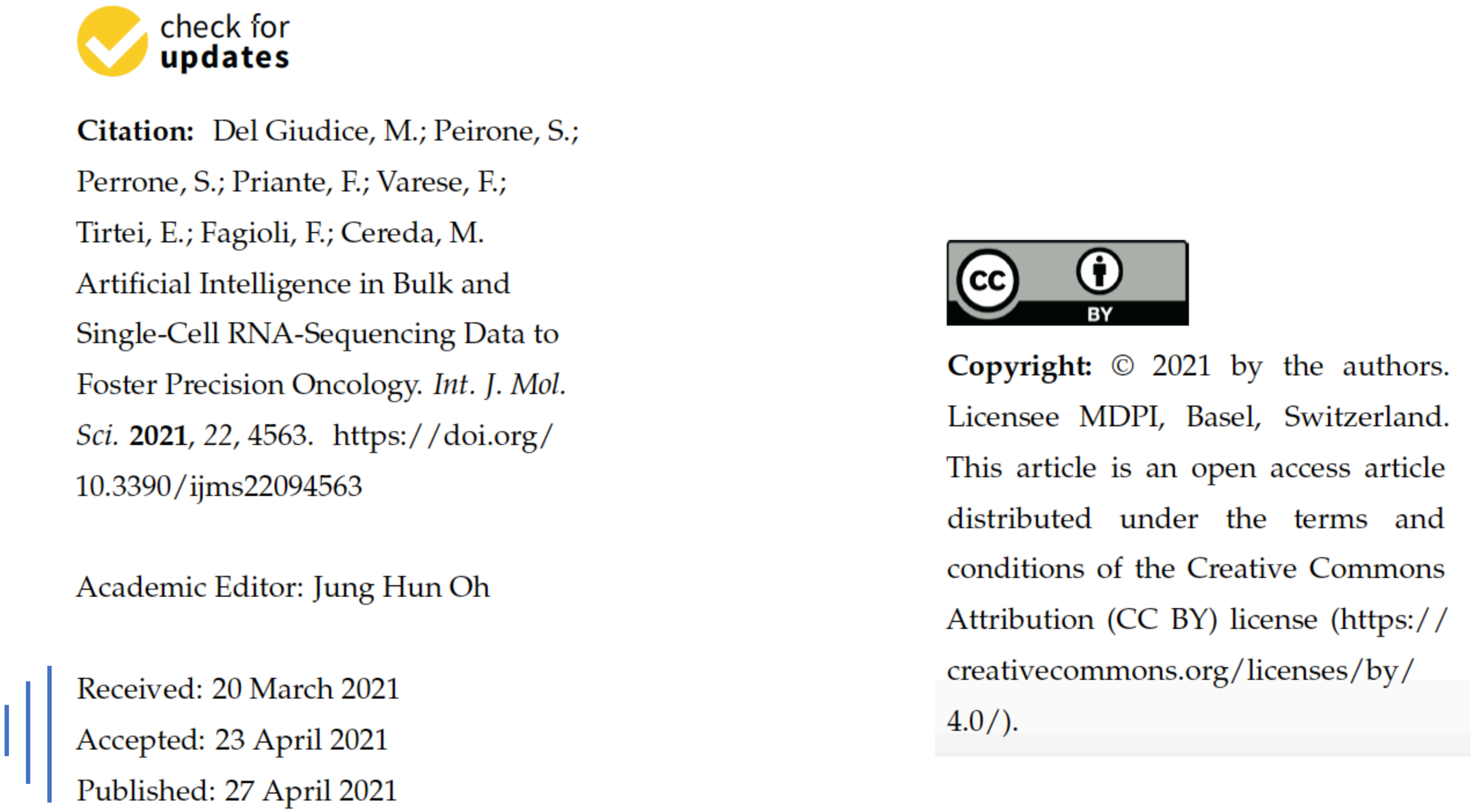
[6] Paediatric Onco-Haematology Division, Regina Margherita Children's Hospital, City of Health and Science of Turin, 10126 Turin, Italy; elisa.tirtei@gmail.com (E.T.); franca.fagioli@unito.it (F.F.)

[7] Department of Public Health and Paediatric Sciences, University of Torino, 10124 Turin, Italy

* Correspondence: matteo.cereda@iigm.it; Tel.: +39-011-993-3969

† The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Search the date… is it recent?



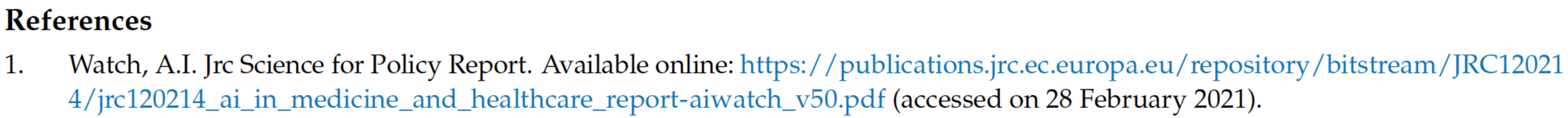
check for updates

Citation: Del Giudice, M.; Peirone, S.; Perrone, S.; Priante, F.; Varese, F.; Tirtei, E.; Fagioli, F.; Cereda, M. Artificial Intelligence in Bulk and Single-Cell RNA-Sequencing Data to Foster Precision Oncology. *Int. J. Mol. Sci.* **2021**, *22*, 4563. https://doi.org/10.3390/ijms22094563

Academic Editor: Jung Hun Oh

Received: 20 March 2021
Accepted: 23 April 2021
Published: 27 April 2021

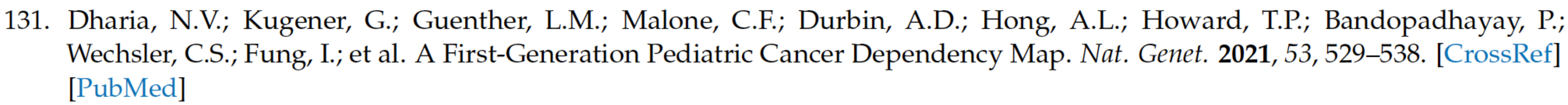
Notice the number of references



References

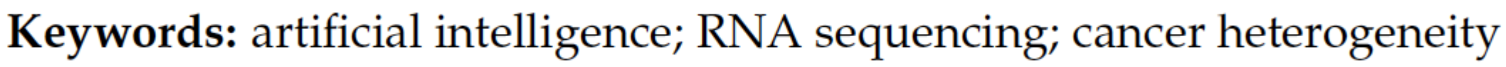
1. Watch, A.I. Jrc Science for Policy Report. Available online: https://publications.jrc.ec.europa.eu/repository/bitstream/JRC120214/jrc120214_ai_in_medicine_and_healthcare_report-aiwatch_v50.pdf (accessed on 28 February 2021).

…



131. Dharia, N.V.; Kugener, G.; Guenther, L.M.; Malone, C.F.; Durbin, A.D.; Hong, A.L.; Howard, T.P.; Bandopadhayay, P.; Wechsler, C.S.; Fung, I.; et al. A First-Generation Pediatric Cancer Dependency Map. *Nat. Genet.* **2021**, *53*, 529–538. [CrossRef] [PubMed]

Are there any keywords?



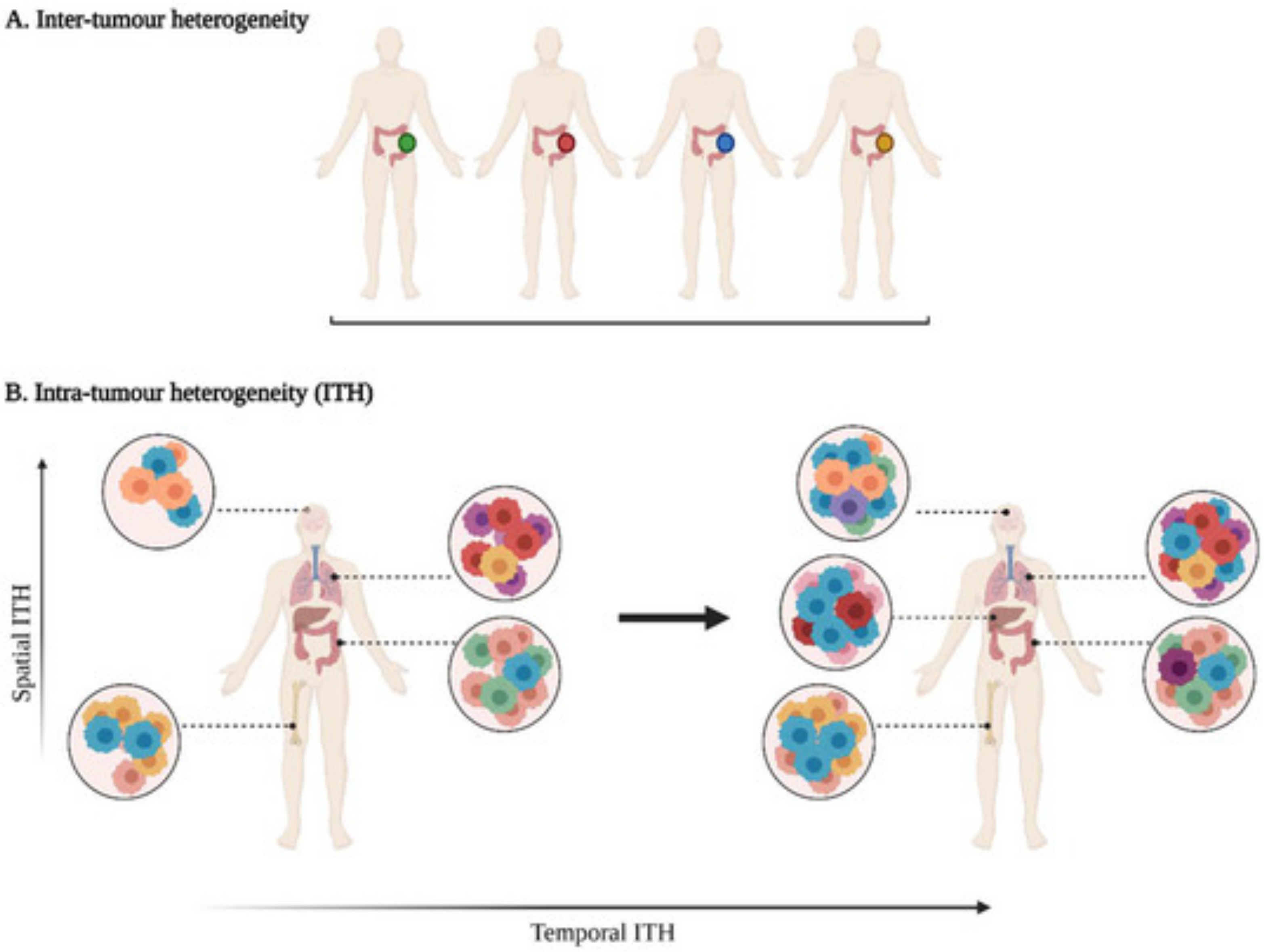
Keywords: artificial intelligence; RNA sequencing; cancer heterogeneity

Are there any keywords?

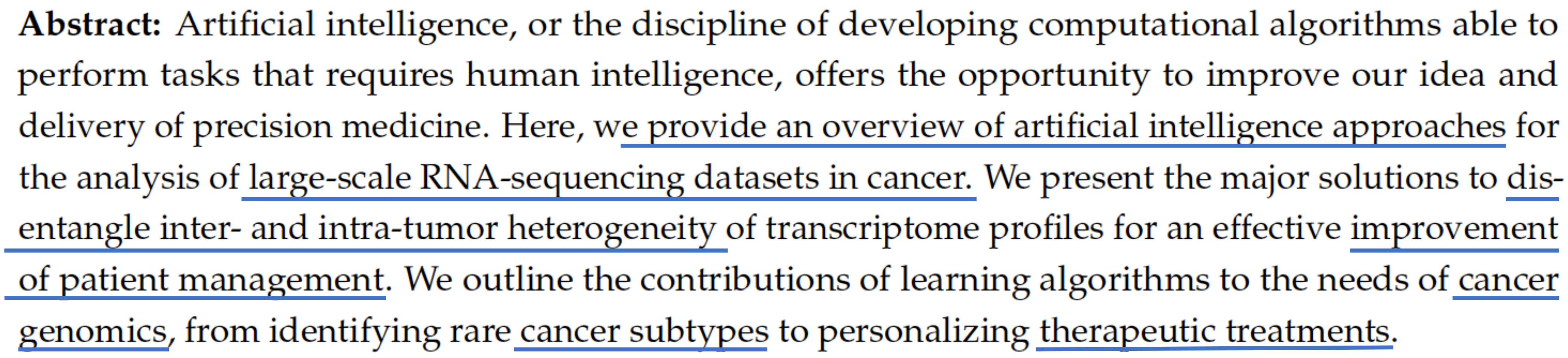
What’s heterogeneity in cancer? Heterogeneity between: **Tumor types**

Patients sharing the same tumor type

Different clones in the same tumor


A. Inter-tumour heterogeneity
B. Intra-tumour heterogeneity (ITH)
Spatial ITH
Temporal ITH

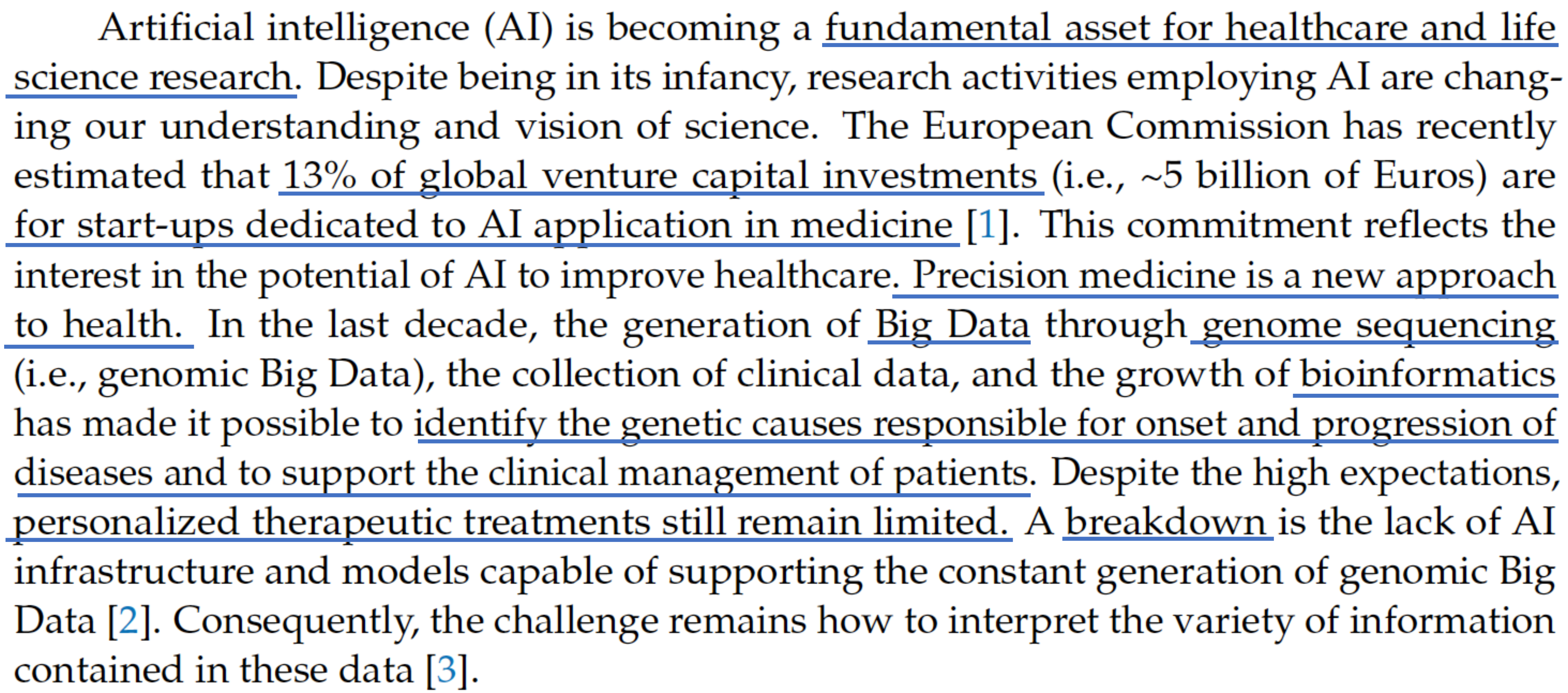
Understand the abstract



Abstract: Artificial intelligence, or the discipline of developing computational algorithms able to perform tasks that requires human intelligence, offers the opportunity to improve our idea and delivery of precision medicine. Here, we provide an overview of artificial intelligence approaches for the analysis of large-scale RNA-sequencing datasets in cancer. We present the major solutions to disentangle inter- and intra-tumor heterogeneity of transcriptome profiles for an effective improvement of patient management. We outline the contributions of learning algorithms to the needs of cancer genomics, from identifying rare cancer subtypes to personalizing therapeutic treatments.

Dive into the introduction

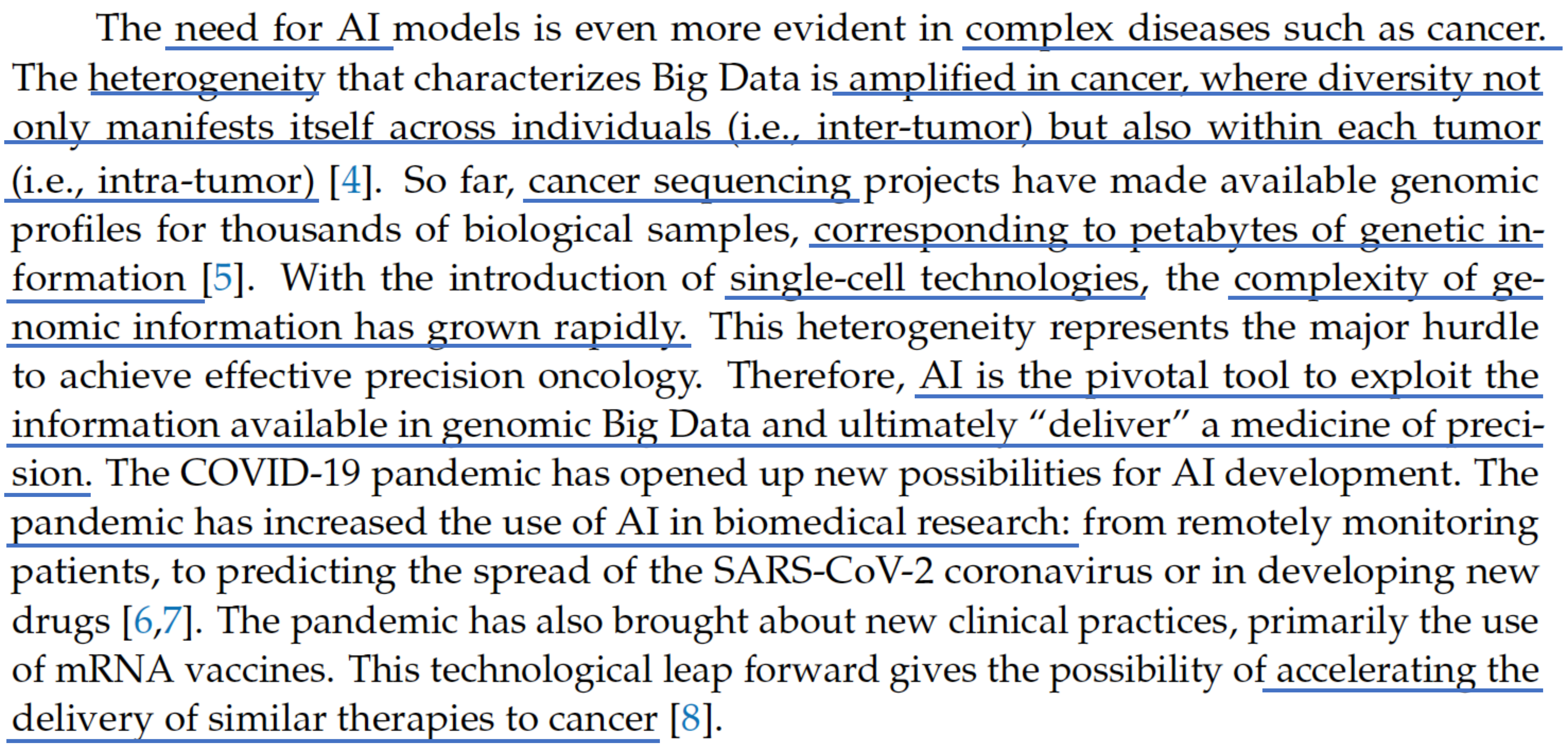
1st paragraph **AI progression, importance and limitations in precision medicine**



Artificial intelligence (AI) is becoming a fundamental asset for healthcare and life science research. Despite being in its infancy, research activities employing AI are changing our understanding and vision of science. The European Commission has recently estimated that 13% of global venture capital investments (i.e., ~5 billion of Euros) are for start-ups dedicated to AI application in medicine [1]. This commitment reflects the interest in the potential of AI to improve healthcare. Precision medicine is a new approach to health. In the last decade, the generation of Big Data through genome sequencing (i.e., genomic Big Data), the collection of clinical data, and the growth of bioinformatics has made it possible to identify the genetic causes responsible for onset and progression of diseases and to support the clinical management of patients. Despite the high expectations, personalized therapeutic treatments still remain limited. A breakdown is the lack of AI infrastructure and models capable of supporting the constant generation of genomic Big Data [2]. Consequently, the challenge remains how to interpret the variety of information contained in these data [3].

Dive into the introduction

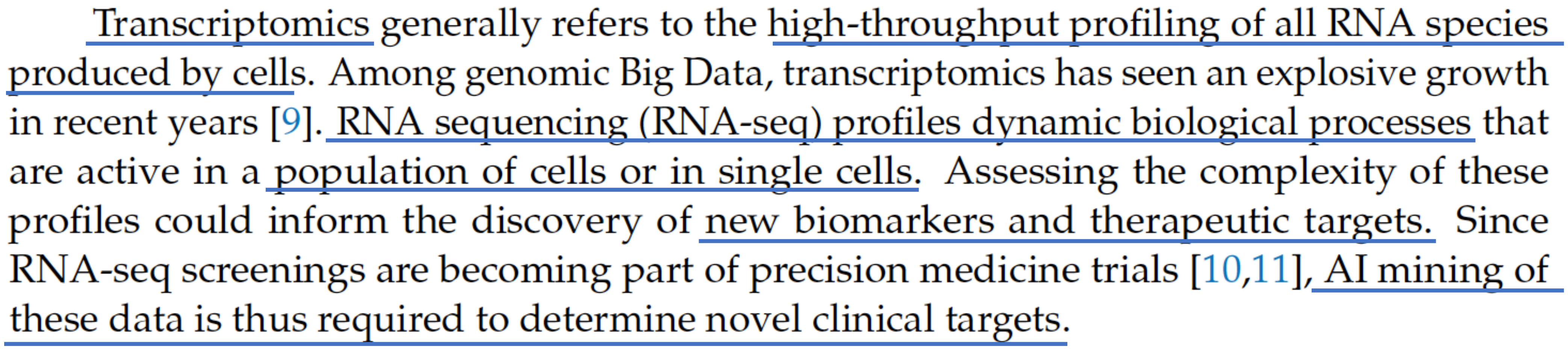
2nd paragraph **AI to resolve cancer heterogeneity**



The need for AI models is even more evident in complex diseases such as cancer. The heterogeneity that characterizes Big Data is amplified in cancer, where diversity not only manifests itself across individuals (i.e., inter-tumor) but also within each tumor (i.e., intra-tumor) [4]. So far, cancer sequencing projects have made available genomic profiles for thousands of biological samples, corresponding to petabytes of genetic information [5]. With the introduction of single-cell technologies, the complexity of genomic information has grown rapidly. This heterogeneity represents the major hurdle to achieve effective precision oncology. Therefore, AI is the pivotal tool to exploit the information available in genomic Big Data and ultimately “deliver” a medicine of precision. The COVID-19 pandemic has opened up new possibilities for AI development. The pandemic has increased the use of AI in biomedical research: from remotely monitoring patients, to predicting the spread of the SARS-CoV-2 coronavirus or in developing new drugs [6,7]. The pandemic has also brought about new clinical practices, primarily the use of mRNA vaccines. This technological leap forward gives the possibility of accelerating the delivery of similar therapies to cancer [8].

Dive into the introduction

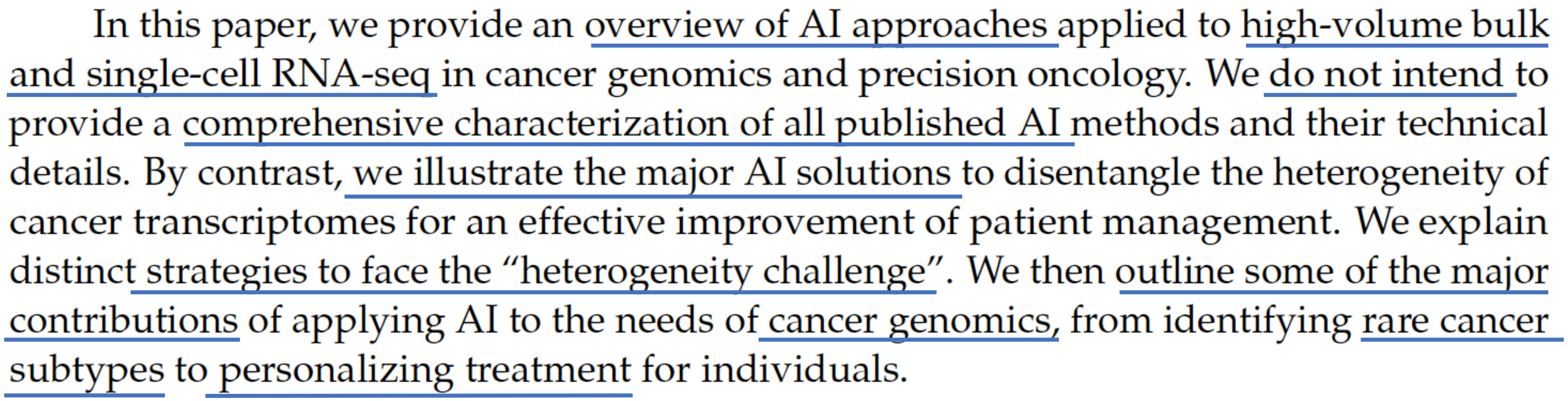
3rd paragraph **Transcriptomics & AI to find biomarkers**



Transcriptomics generally refers to the high-throughput profiling of all RNA species produced by cells. Among genomic Big Data, transcriptomics has seen an explosive growth in recent years [9]. RNA sequencing (RNA-seq) profiles dynamic biological processes that are active in a population of cells or in single cells. Assessing the complexity of these profiles could inform the discovery of new biomarkers and therapeutic targets. Since RNA-seq screenings are becoming part of precision medicine trials [10,11], AI mining of these data is thus required to determine novel clinical targets.

Dive into the introduction

4th paragraph **What we will and will not find while reading the paper**



In this paper, we provide an overview of AI approaches applied to high-volume bulk and single-cell RNA-seq in cancer genomics and precision oncology. We do not intend to provide a comprehensive characterization of all published AI methods and their technical details. By contrast, we illustrate the major AI solutions to disentangle the heterogeneity of cancer transcriptomes for an effective improvement of patient management. We explain distinct strategies to face the "heterogeneity challenge". We then outline some of the major contributions of applying AI to the needs of cancer genomics, from identifying rare cancer subtypes to personalizing treatment for individuals.

Figures and tables

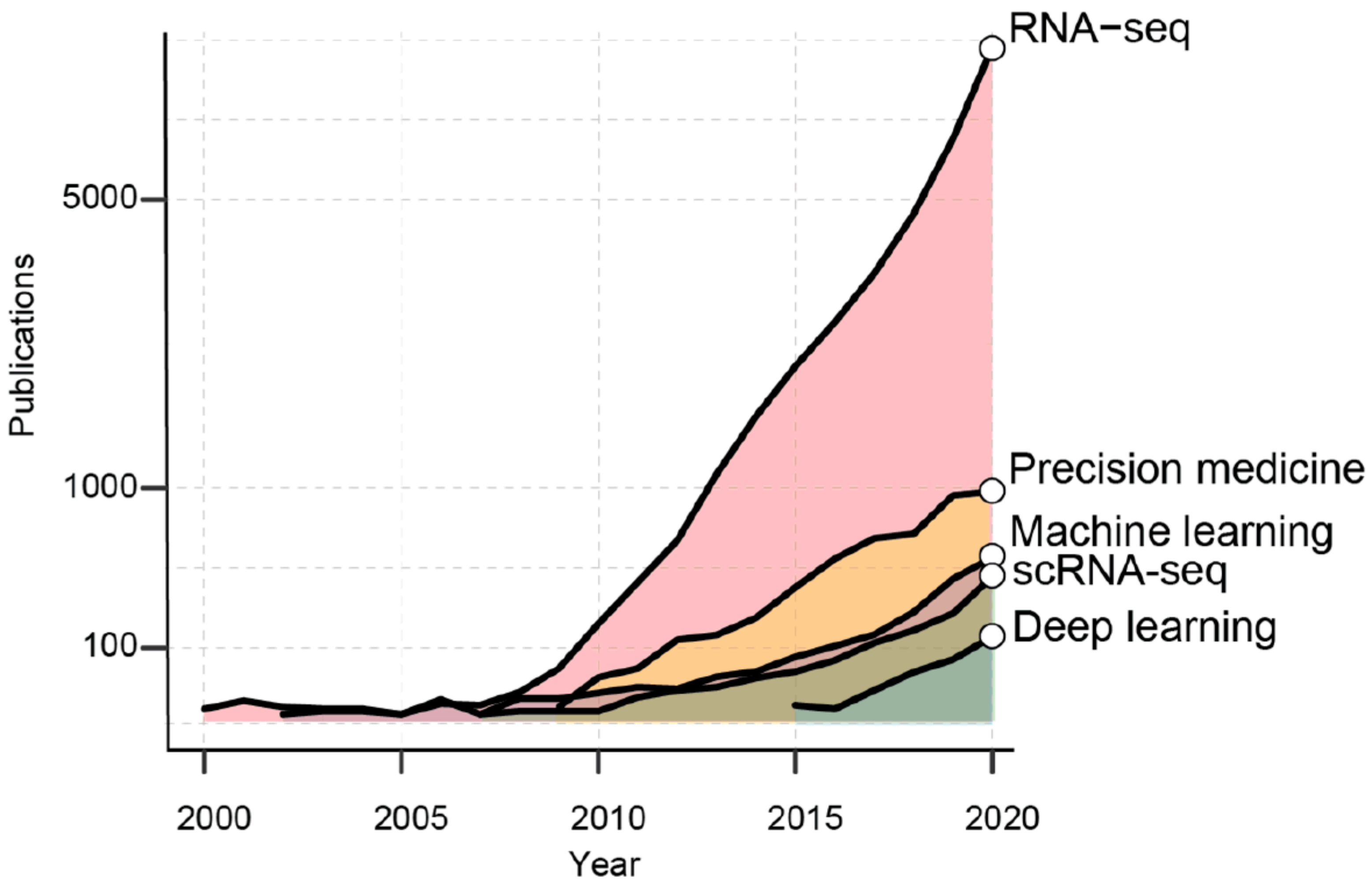


Figure 1. The graph shows the number of PubMed publications per years containing the reported keywords.

Figures and tables

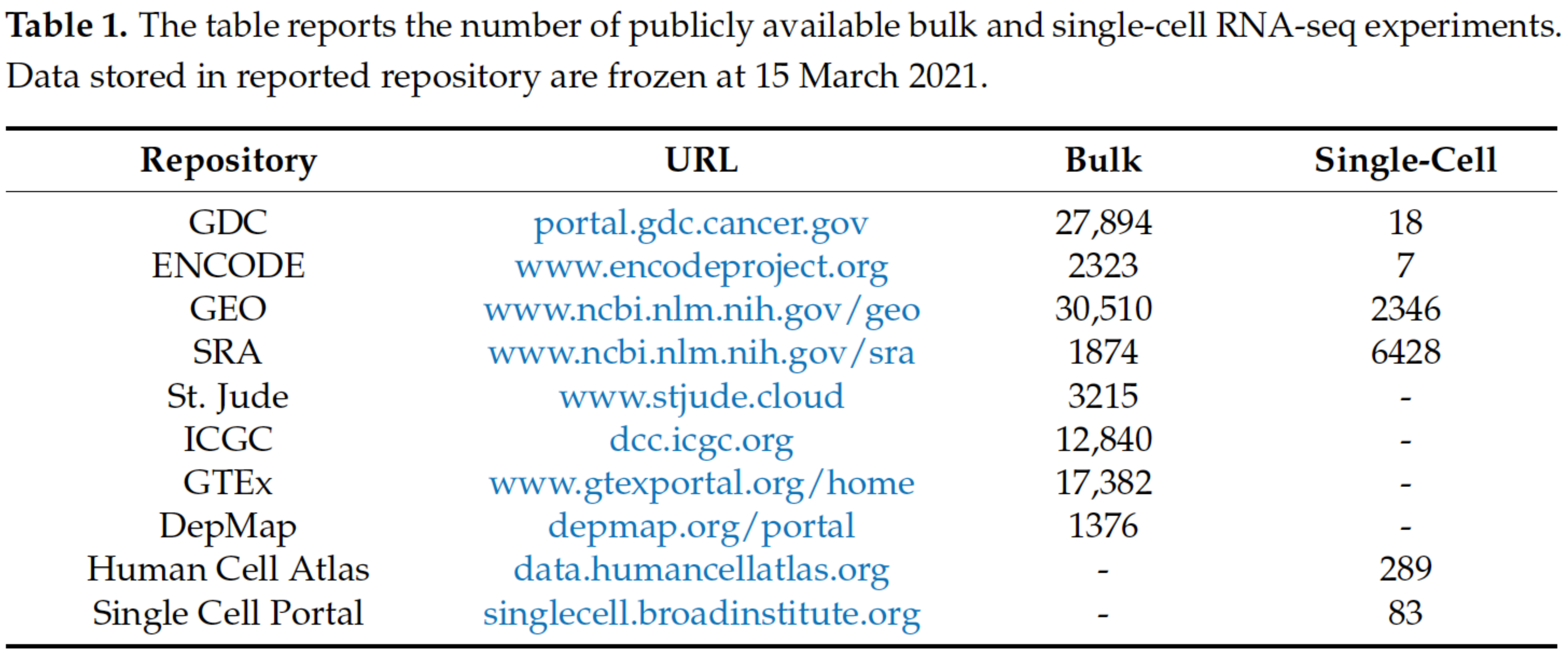


Table 1. The table reports the number of publicly available bulk and single-cell RNA-seq experiments. Data stored in reported repository are frozen at 15 March 2021.

Repository	URL	Bulk	Single-Cell
GDC	portal.gdc.cancer.gov	27,894	18
ENCODE	www.encodeproject.org	2323	7
GEO	www.ncbi.nlm.nih.gov/geo	30,510	2346
SRA	www.ncbi.nlm.nih.gov/sra	1874	6428
St. Jude	www.stjude.cloud	3215	-
ICGC	dcc.icgc.org	12,840	-
GTEx	www.gtexportal.org/home	17,382	-
DepMap	depmap.org/portal	1376	-
Human Cell Atlas	data.humancellatlas.org	-	289
Single Cell Portal	singlecell.broadinstitute.org	-	83

Figures and tables

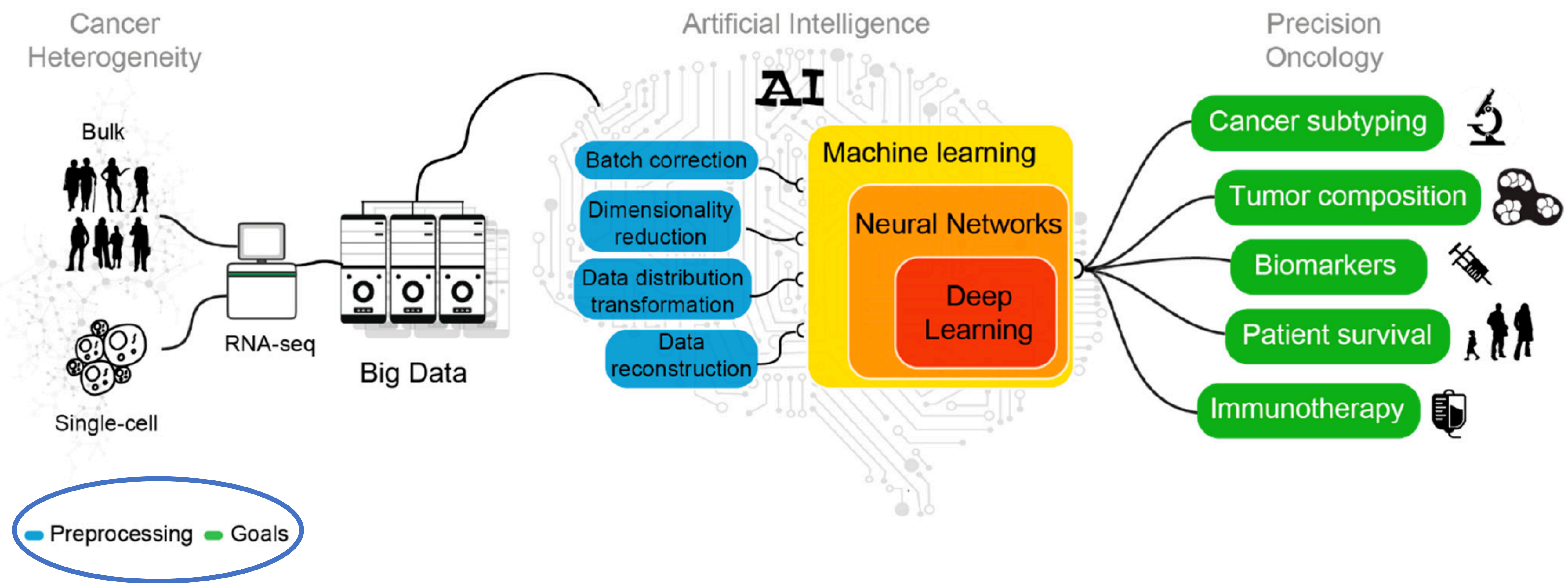


Figure 2. Sketch representing the analyses needed to decipher cancer heterogeneity and achieve an effective precision oncology.

Figures and tables

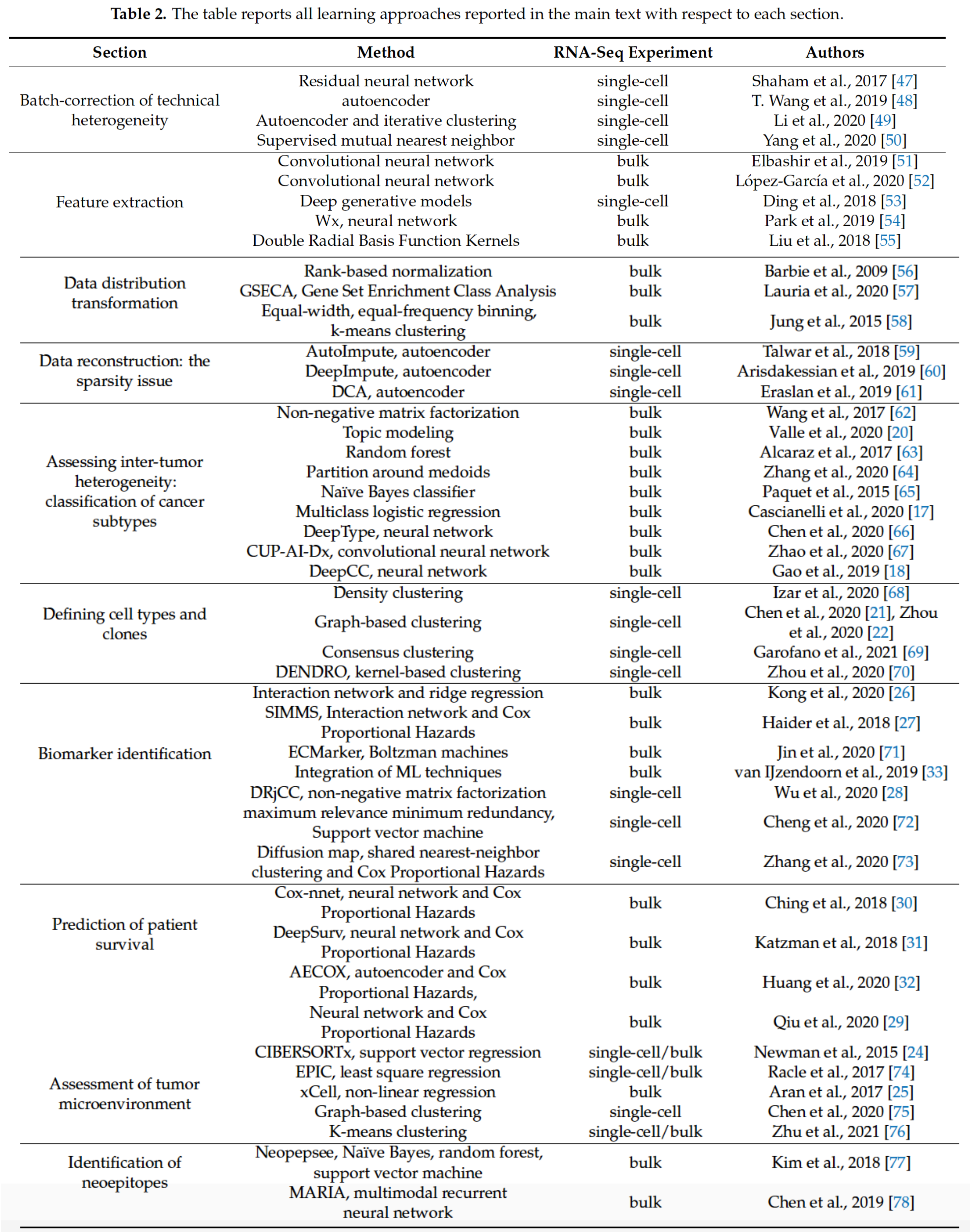


Table 2. The table reports all learning approaches reported in the main text with respect to each section.

Section	Method	RNA-Seq Experiment	Authors
Batch-correction of technical heterogeneity	Residual neural network	single-cell	Shaham et al., 2017 [47]
	autoencoder	single-cell	T. Wang et al., 2019 [48]
	Autoencoder and iterative clustering	single-cell	Li et al., 2020 [49]
	Supervised mutual nearest neighbor	single-cell	Yang et al., 2020 [50]
Feature extraction	Convolutional neural network	bulk	Elbashir et al., 2019 [51]
	Convolutional neural network	bulk	López-García et al., 2020 [52]
	Deep generative models	single-cell	Ding et al., 2018 [53]
	Wx, neural network	bulk	Park et al., 2019 [54]
	Double Radial Basis Function Kernels	bulk	Liu et al., 2018 [55]
Data distribution transformation	Rank-based normalization	bulk	Barbie et al., 2009 [56]
	GSECA, Gene Set Enrichment Class Analysis	bulk	Lauria et al., 2020 [57]
	Equal-width, equal-frequency binning, k-means clustering	bulk	Jung et al., 2015 [58]
Data reconstruction: the sparsity issue	AutoImpute, autoencoder	single-cell	Talwar et al., 2018 [59]
	DeepImpute, autoencoder	single-cell	Arisdakessian et al., 2019 [60]
	DCA, autoencoder	single-cell	Eraslan et al., 2019 [61]
Assessing inter-tumor heterogeneity: classification of cancer subtypes	Non-negative matrix factorization	bulk	Wang et al., 2017 [62]
	Topic modeling	bulk	Valle et al., 2020 [20]
	Random forest	bulk	Alcaraz et al., 2017 [63]
	Partition around medoids	bulk	Zhang et al., 2020 [64]
	Naïve Bayes classifier	bulk	Paquet et al., 2015 [65]
	Multiclass logistic regression	bulk	Cascianelli et al., 2020 [17]
	DeepType, neural network	bulk	Chen et al., 2020 [66]
	CUP-AI-Dx, convolutional neural network	bulk	Zhao et al., 2020 [67]
	DeepCC, neural network	bulk	Gao et al., 2019 [18]
Defining cell types and clones	Density clustering	single-cell	Izar et al., 2020 [68]
	Graph-based clustering	single-cell	Chen et al., 2020 [21], Zhou et al., 2020 [22]
	Consensus clustering	single-cell	Garofano et al., 2021 [69]
	DENDRO, kernel-based clustering	single-cell	Zhou et al., 2020 [70]
Biomarker identification	Interaction network and ridge regression	bulk	Kong et al., 2020 [26]
	SIMMS, Interaction network and Cox Proportional Hazards	bulk	Haider et al., 2018 [27]
	ECMarker, Boltzman machines	bulk	Jin et al., 2020 [71]
	Integration of ML techniques	bulk	van IJzendoorn et al., 2019 [33]
	DRjCC, non-negative matrix factorization	single-cell	Wu et al., 2020 [28]
	maximum relevance minimum redundancy, Support vector machine	single-cell	Cheng et al., 2020 [72]
	Diffusion map, shared nearest-neighbor clustering and Cox Proportional Hazards	single-cell	Zhang et al., 2020 [73]
Prediction of patient survival	Cox-nnet, neural network and Cox Proportional Hazards	bulk	Ching et al., 2018 [30]
	DeepSurv, neural network and Cox Proportional Hazards	bulk	Katzman et al., 2018 [31]
	AECOX, autoencoder and Cox Proportional Hazards,	bulk	Huang et al., 2020 [32]
	Neural network and Cox Proportional Hazards	bulk	Qiu et al., 2020 [29]
Assessment of tumor microenvironment	CIBERSORTx, support vector regression	single-cell/bulk	Newman et al., 2015 [24]
	EPIC, least square regression	single-cell/bulk	Racle et al., 2017 [74]
	xCell, non-linear regression	bulk	Aran et al., 2017 [25]
	Graph-based clustering	single-cell	Chen et al., 2020 [75]
	K-means clustering	single-cell/bulk	Zhu et al., 2021 [76]
Identification of neoepitopes	Neopepsee, Naïve Bayes, random forest, support vector machine	bulk	Kim et al., 2018 [77]
	MARIA, multimodal recurrent neural network	bulk	Chen et al., 2019 [78]

Figures and tables

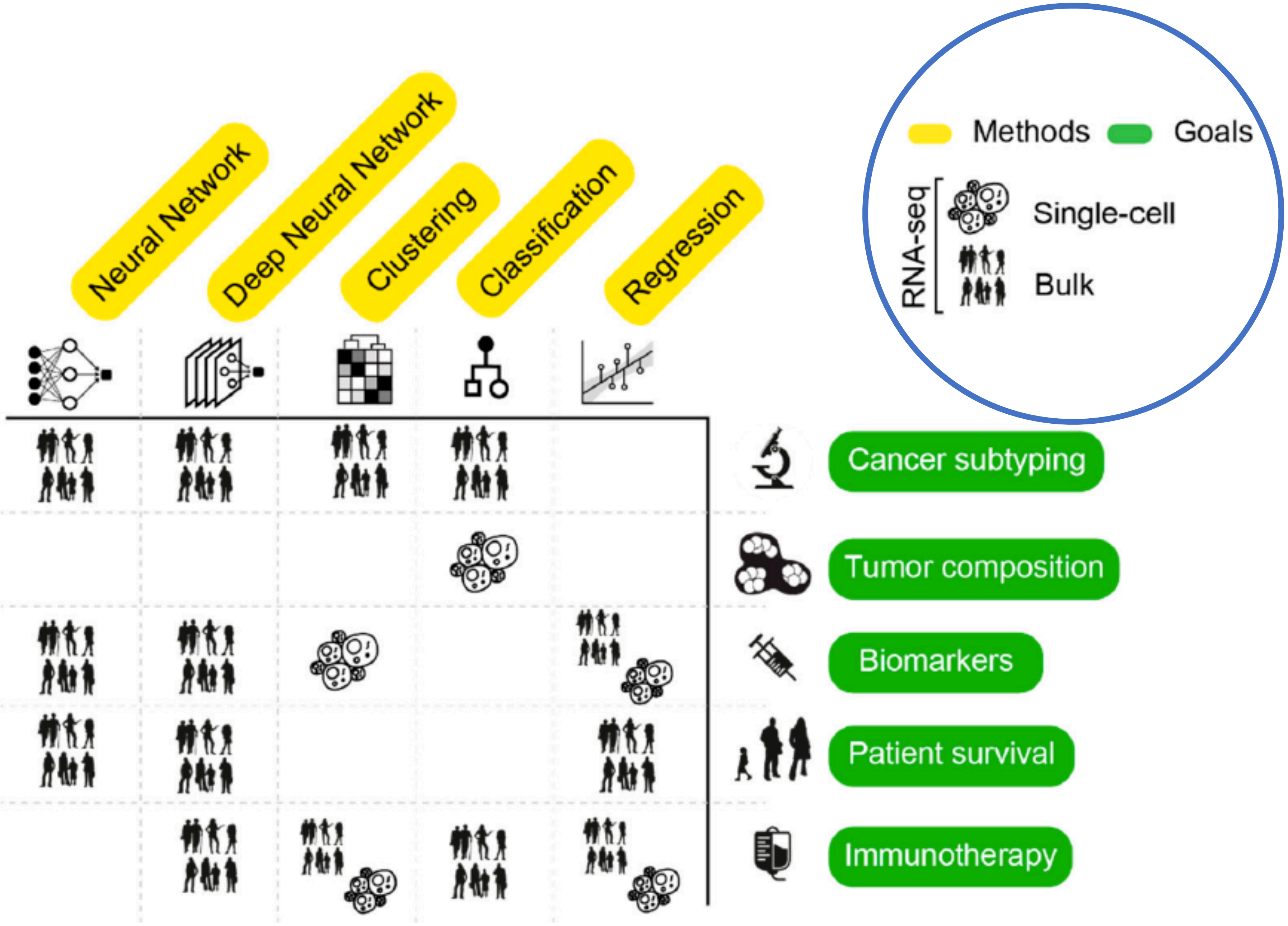
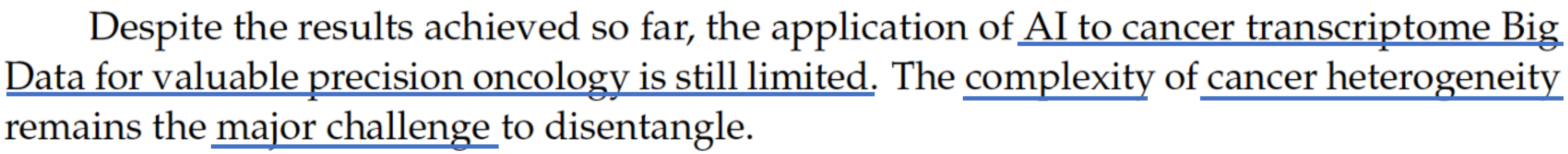


Figure 3. Graphical summary of AI approaches (columns) applied to solve tasks (rows) presented in this review. Cells show the RNA-seq data type used for the analysis. The "immunotherapy" task includes assessment of tumor microenvironment and identification of neoepitopes.

Conclusions

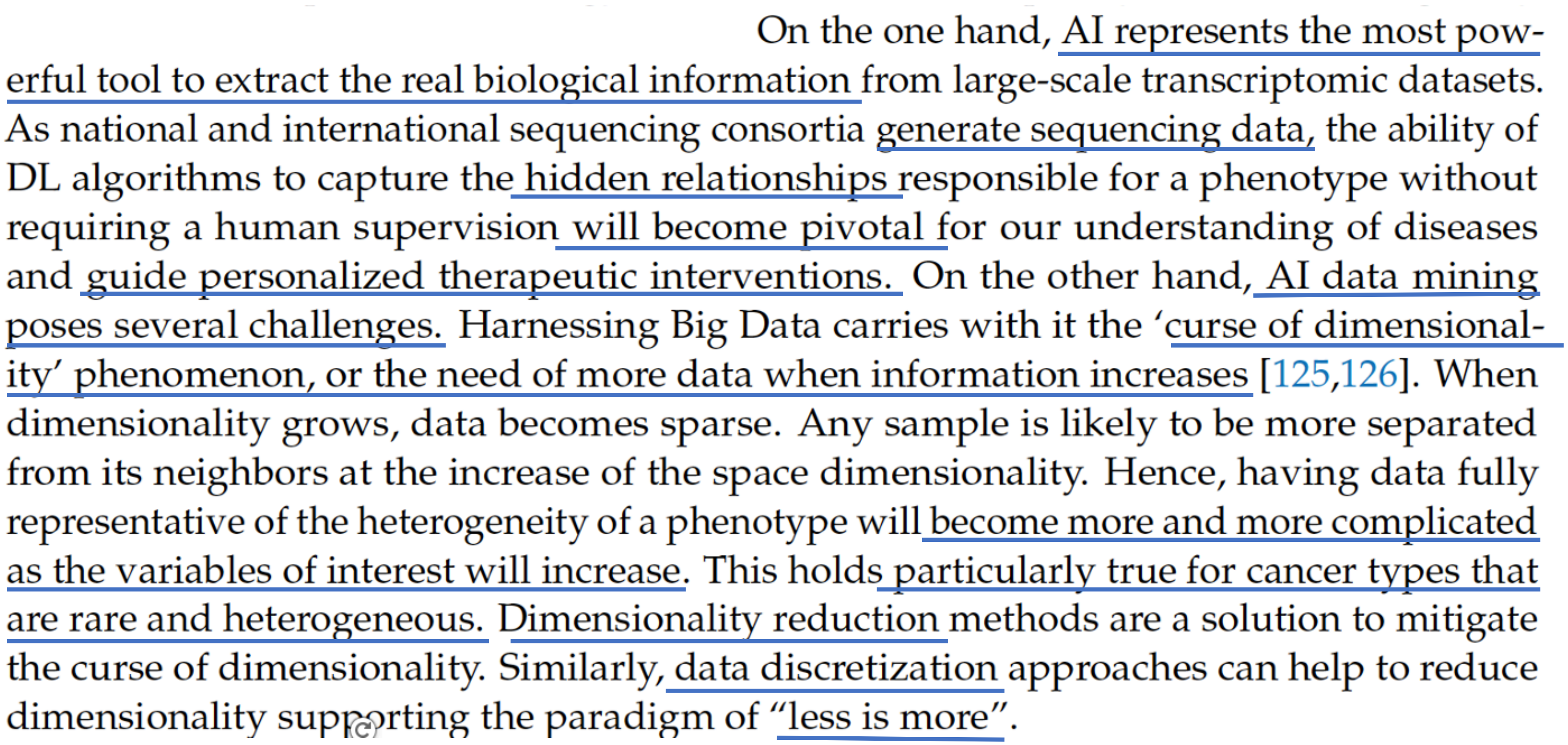
1st extract



Despite the results achieved so far, the application of AI to cancer transcriptome Big Data for valuable precision oncology is still limited. The complexity of cancer heterogeneity remains the major challenge to disentangle.

Conclusions

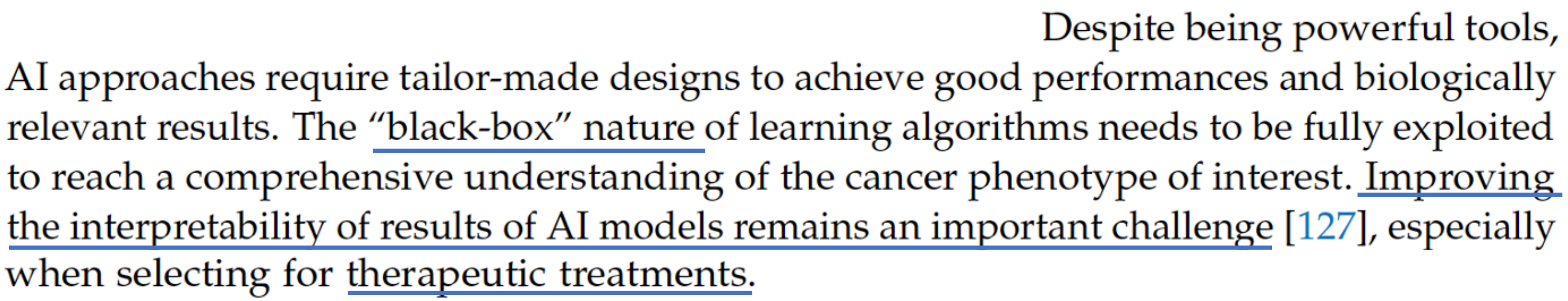
2nd extract



On the one hand, AI represents the most powerful tool to extract the real biological information from large-scale transcriptomic datasets. As national and international sequencing consortia generate sequencing data, the ability of DL algorithms to capture the hidden relationships responsible for a phenotype without requiring a human supervision will become pivotal for our understanding of diseases and guide personalized therapeutic interventions. On the other hand, AI data mining poses several challenges. Harnessing Big Data carries with it the 'curse of dimensionality' phenomenon, or the need of more data when information increases [125,126]. When dimensionality grows, data becomes sparse. Any sample is likely to be more separated from its neighbors at the increase of the space dimensionality. Hence, having data fully representative of the heterogeneity of a phenotype will become more and more complicated as the variables of interest will increase. This holds particularly true for cancer types that are rare and heterogeneous. Dimensionality reduction methods are a solution to mitigate the curse of dimensionality. Similarly, data discretization approaches can help to reduce dimensionality supporting the paradigm of "less is more".

Conclusions

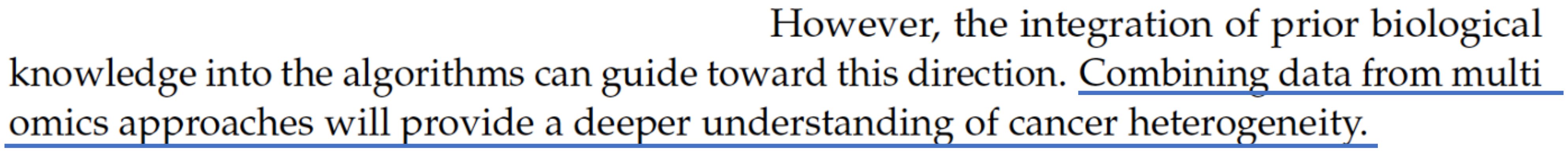
3rd extract



Despite being powerful tools, AI approaches require tailor-made designs to achieve good performances and biologically relevant results. The “black-box” nature of learning algorithms needs to be fully exploited to reach a comprehensive understanding of the cancer phenotype of interest. Improving the interpretability of results of AI models remains an important challenge [127], especially when selecting for therapeutic treatments.

Conclusions

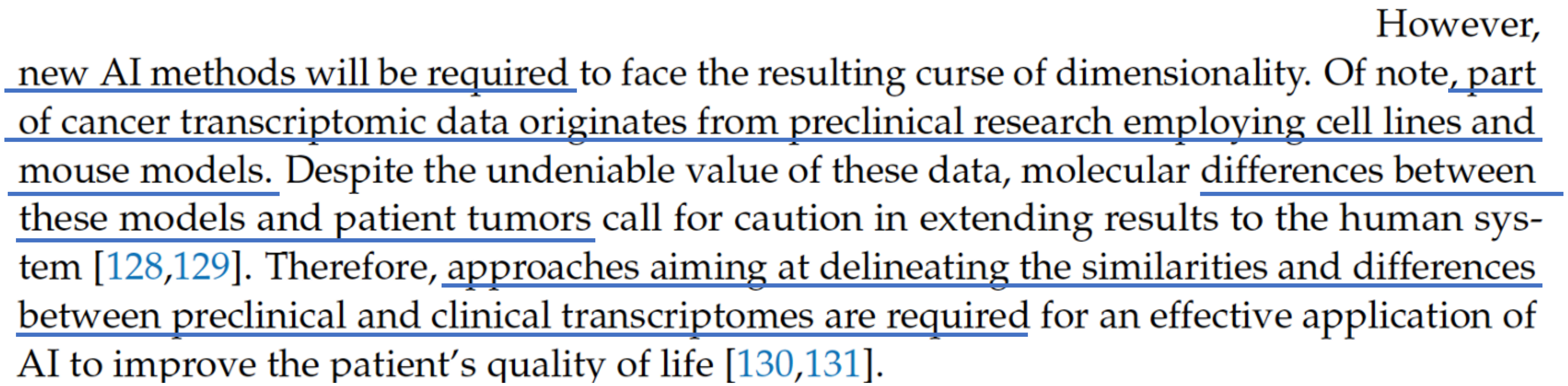
4th extract



However, the integration of prior biological knowledge into the algorithms can guide toward this direction. Combining data from multi omics approaches will provide a deeper understanding of cancer heterogeneity.

Conclusions

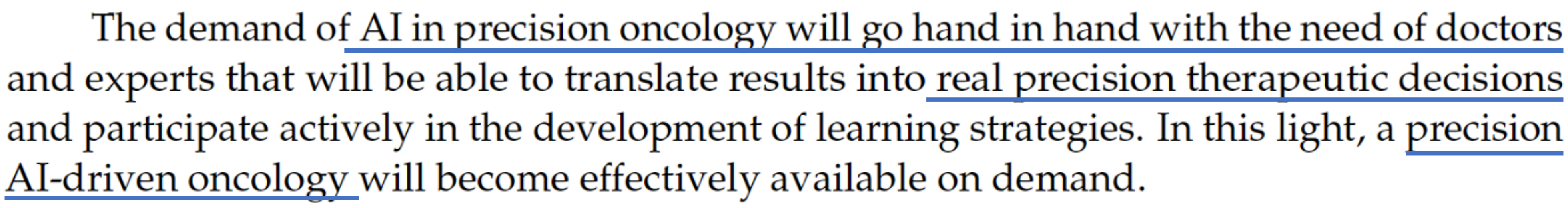
5th extract



However, new AI methods will be required to face the resulting curse of dimensionality. Of note, part of cancer transcriptomic data originates from preclinical research employing cell lines and mouse models. Despite the undeniable value of these data, molecular differences between these models and patient tumors call for caution in extending results to the human system [128,129]. Therefore, approaches aiming at delineating the similarities and differences between preclinical and clinical transcriptomes are required for an effective application of AI to improve the patient's quality of life [130,131].

Conclusions

6th extract



The demand of AI in precision oncology will go hand in hand with the need of doctors and experts that will be able to translate results into real precision therapeutic decisions and participate actively in the development of learning strategies. In this light, a precision AI-driven oncology will become effectively available on demand.

Paper structure

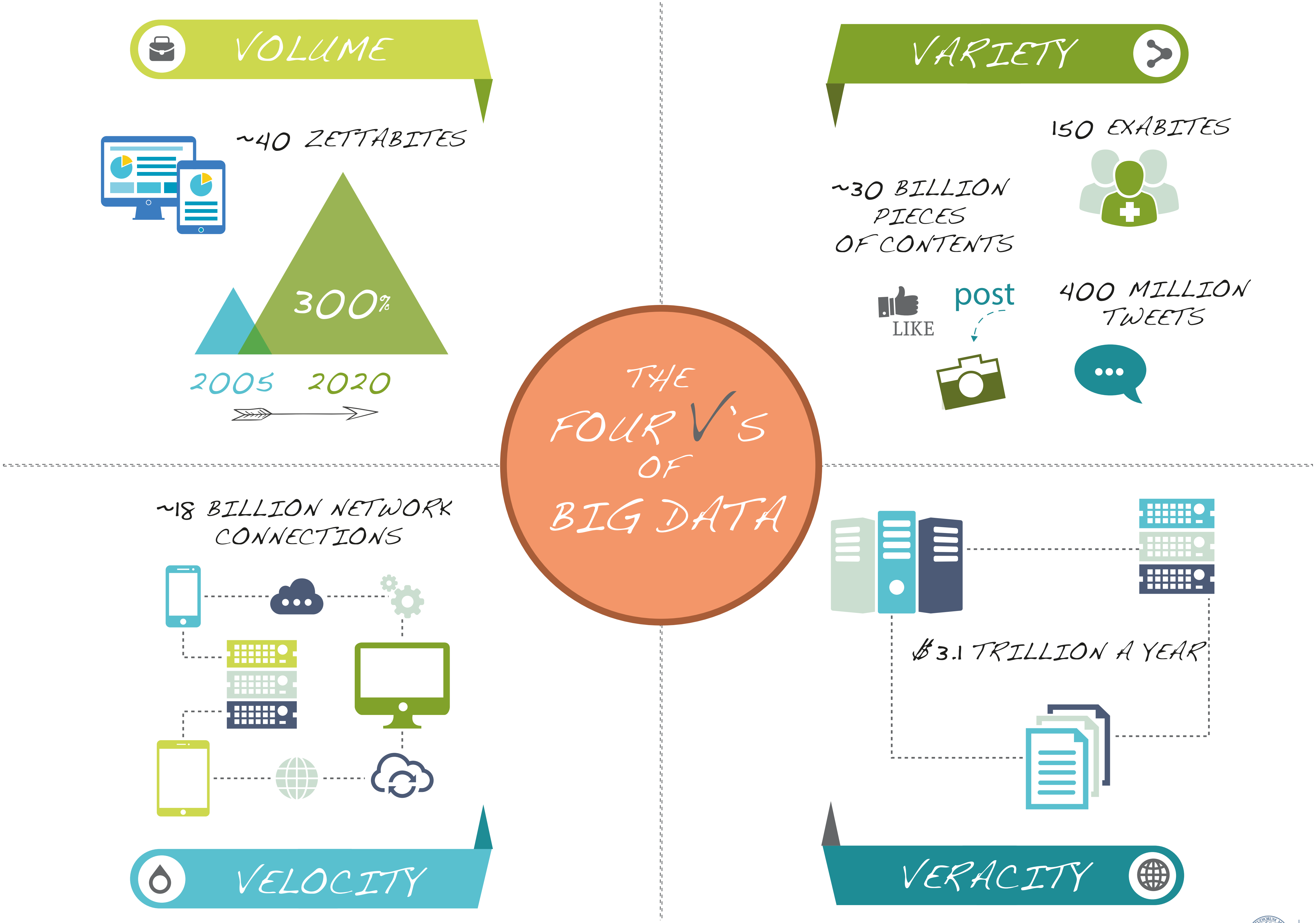
1. Introduction
2. AI in the Era of Transcriptomic Big Data
3. Managing the Heterogeneity of Cancer Transcriptomes
    - 3.1 Batch-Correction of Technical Heterogeneity
    - 3.2 Dimensionality Reduction Approaches
        - 3.2.1 Feature Extraction
        - 3.2.2 Feature Selection
    - 3.3 Data Distribution Transformation
    - 3.4 Data Reconstruction: The Sparsity Issue
4. AI Mining of Cancer Transcriptomes
    - 4.1 Assessing Inter-Tumor Heterogeneity: Classification of Cancer Subtypes
    - 4.2 Deciphering Intra-Tumor Heterogeneity
        - 4.2.1 Defining Cell Types and Clones
        - 4.2.2 Assessment of TME (*i.e.* Tumor MicroEnvironment)
    - 4.3 Biomarker Identification
    - 4.4 Prediction of Patient Survival
    - 4.5 Identification of Neoepitopes
5. Conclusions

2. AI in the era of transcriptomic big data

This is a general chapter, linked to figure 1 and 2, in which the relation between AI and Big Data in transcriptomics is explained.

What does we mean with “Big Data”?

AI is fundamental for the processing of Big Data



4Vs

High **Volume** → large file sizes with lots of observations

Wide **Variety** → lots of different types

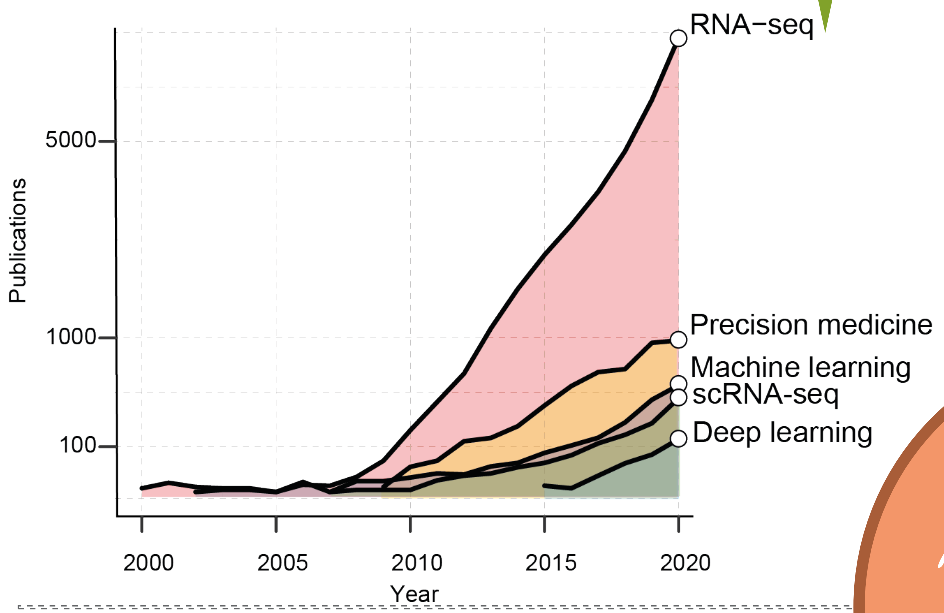
High **Velocity** → accumulating at a high rate

Compromised **Veracity** → variable quality that must be dealt otherwise downstream analyses will be compromised.

Big Data in biology

THE FOUR V'S OF BIG DATA

VOLUME



VARIETY

DNA-seq CHIP-seq

RNA-seq CLIP-seq

ATAC-seq others…

VELOCITY

NIH NATIONAL CANCER INSTITUTE GDC Data Portal

GEO Gene Expression Omnibus

GTEx

ICGC

VERACITY

GitHub

GitLab

ATLASSIAN Bitbucket

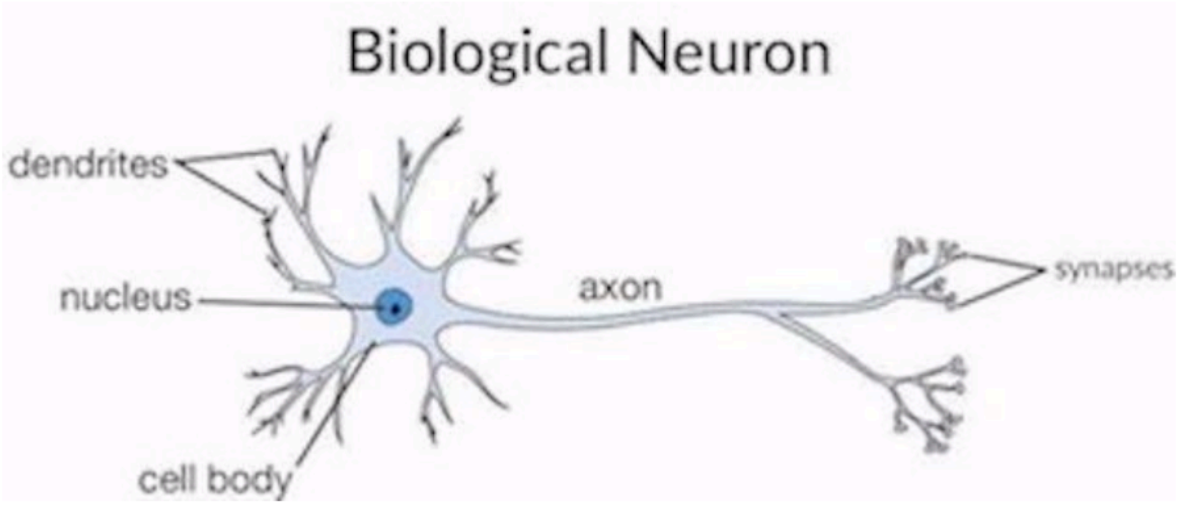
docker

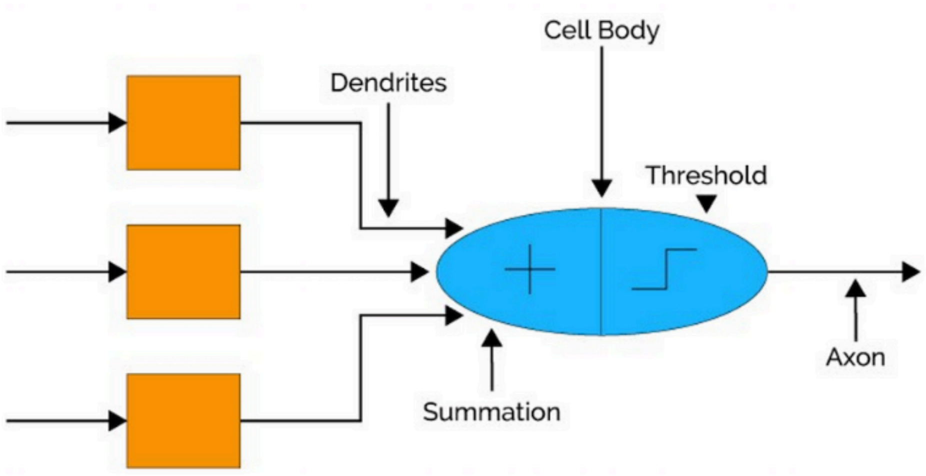
As a review is nothing without its citations, in each chapter from here on we will explore one of them (instead of reading the review's text itself).

Before diving into individual papers focusing on RNA-sequencing and AI, let's make a general introduction.

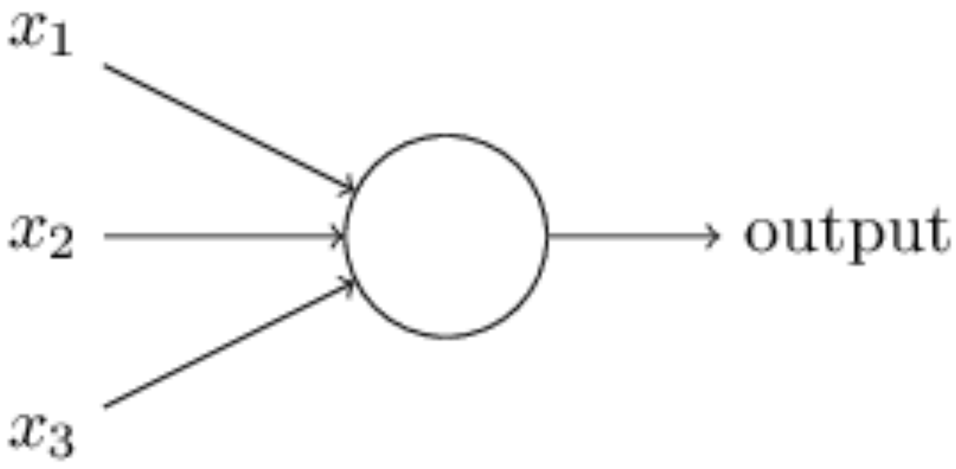
The idea of ML: how it was born…

The founding principle of ML is to emulate the functioning of biological neuron





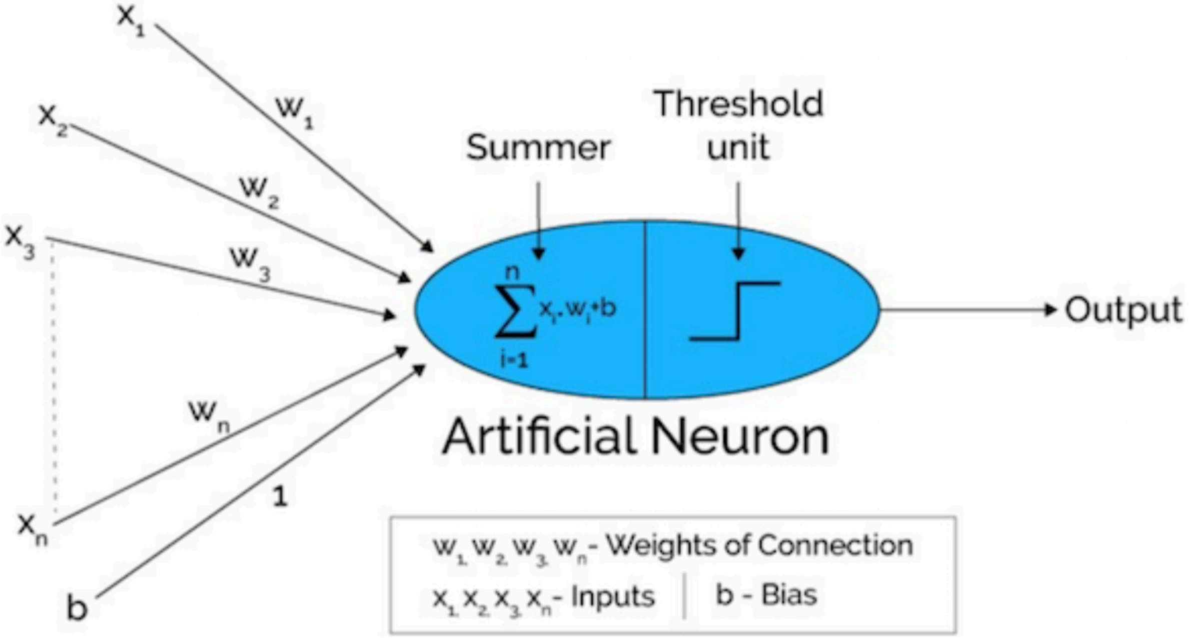
First ML approach… the perceptron

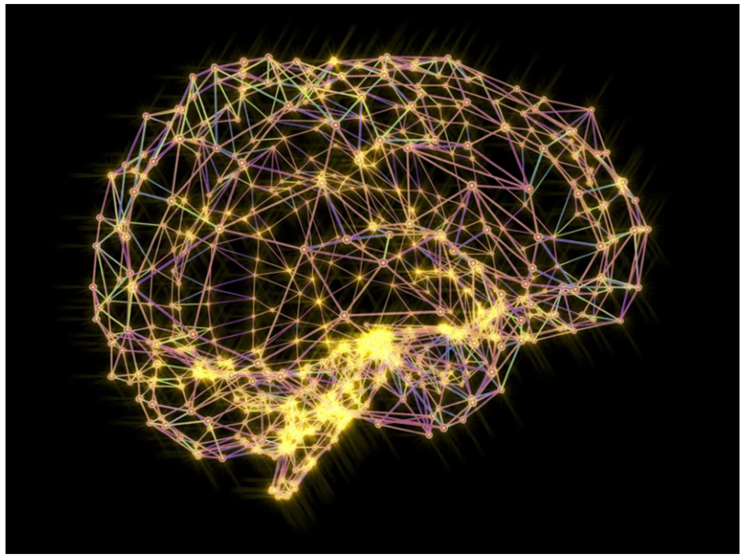


$$\text{output} = \begin{cases} 0 & \text{if } \sum_j w_j x_j \leq \text{ threshold} \\ 1 & \text{if } \sum_j w_j x_j > \text{ threshold} \end{cases}$$

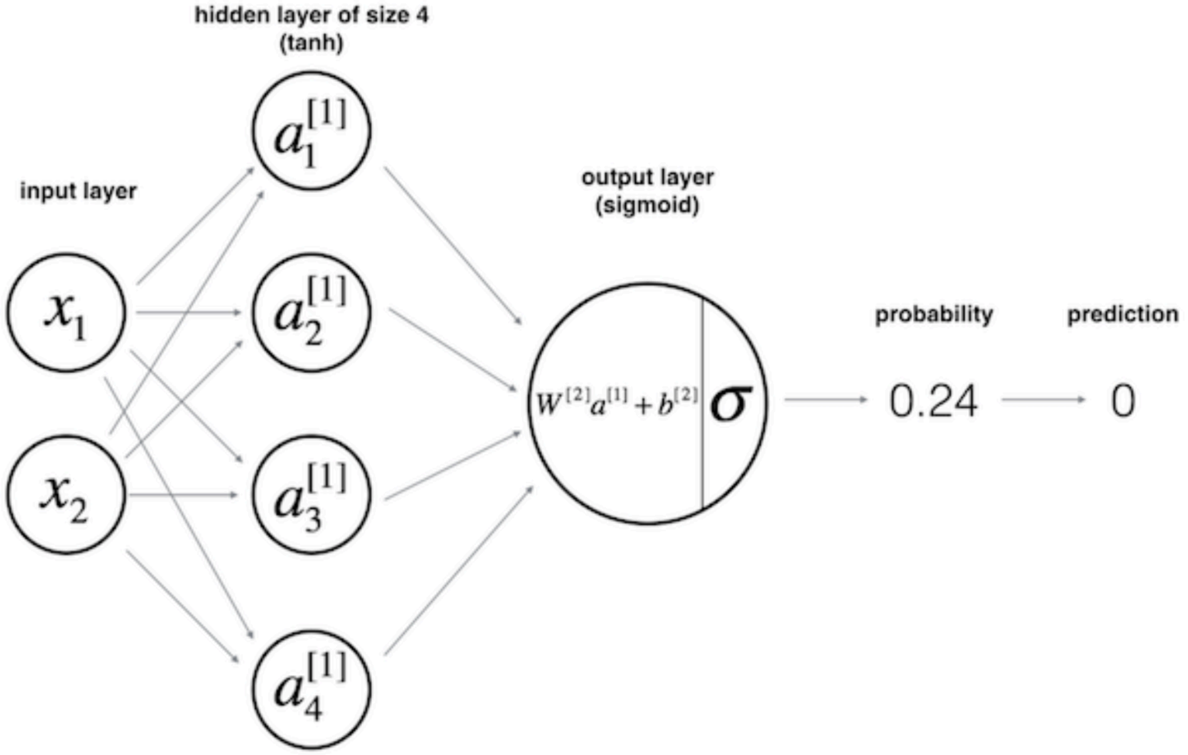
… and developed

Then perceptron evolved…

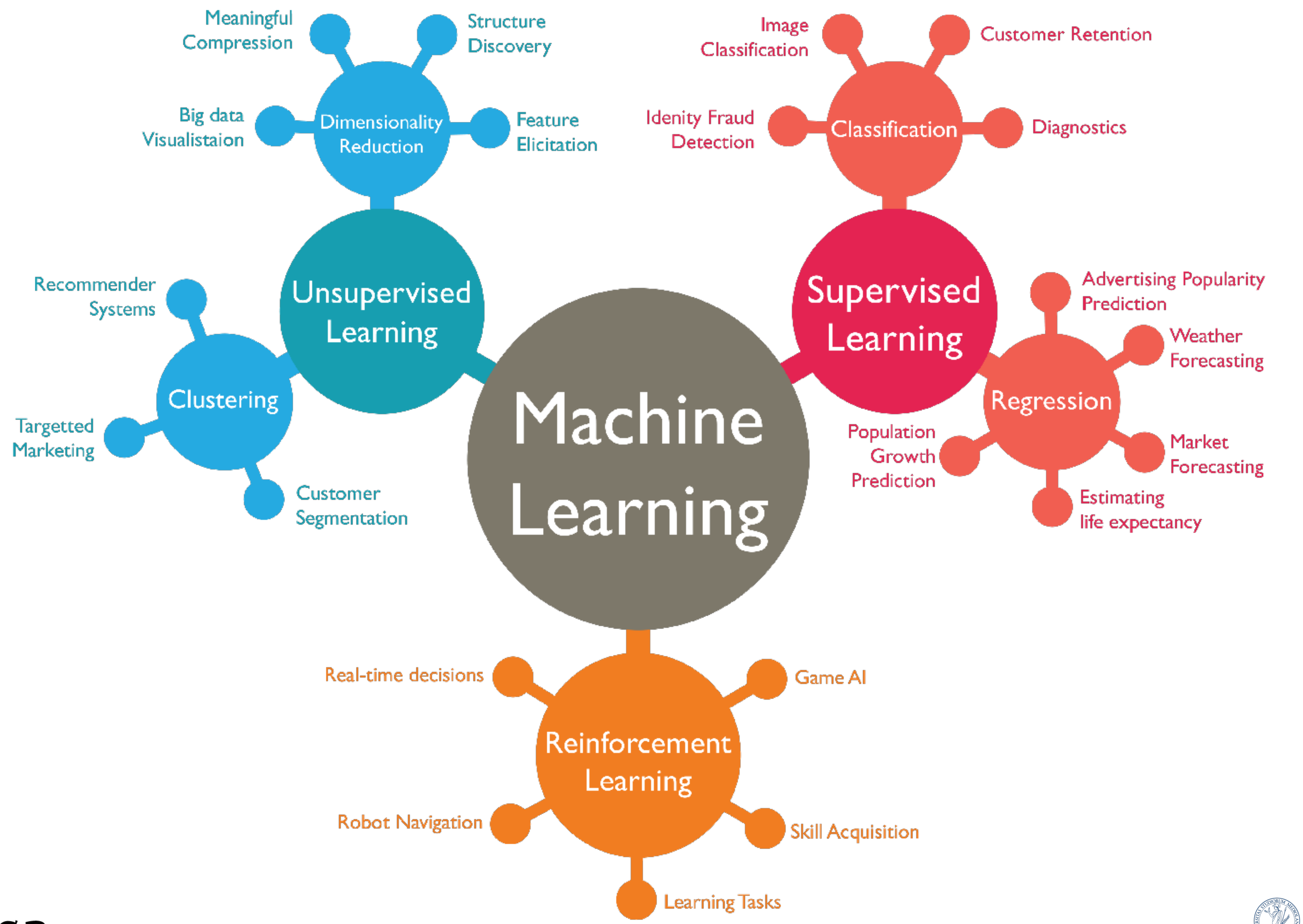

x_1
x_2
x_3
x_n
b
w_1
w_2
w_3
w_n
1
Summer
Threshold unit
$\sum_{i=1}^{n} x_i \cdot w_i + b$
Output
Artificial Neuron
w_1, w_2, w_3, w_n - Weights of Connection
x_1, x_2, x_3, x_n - Inputs
b - Bias



And evolved again…



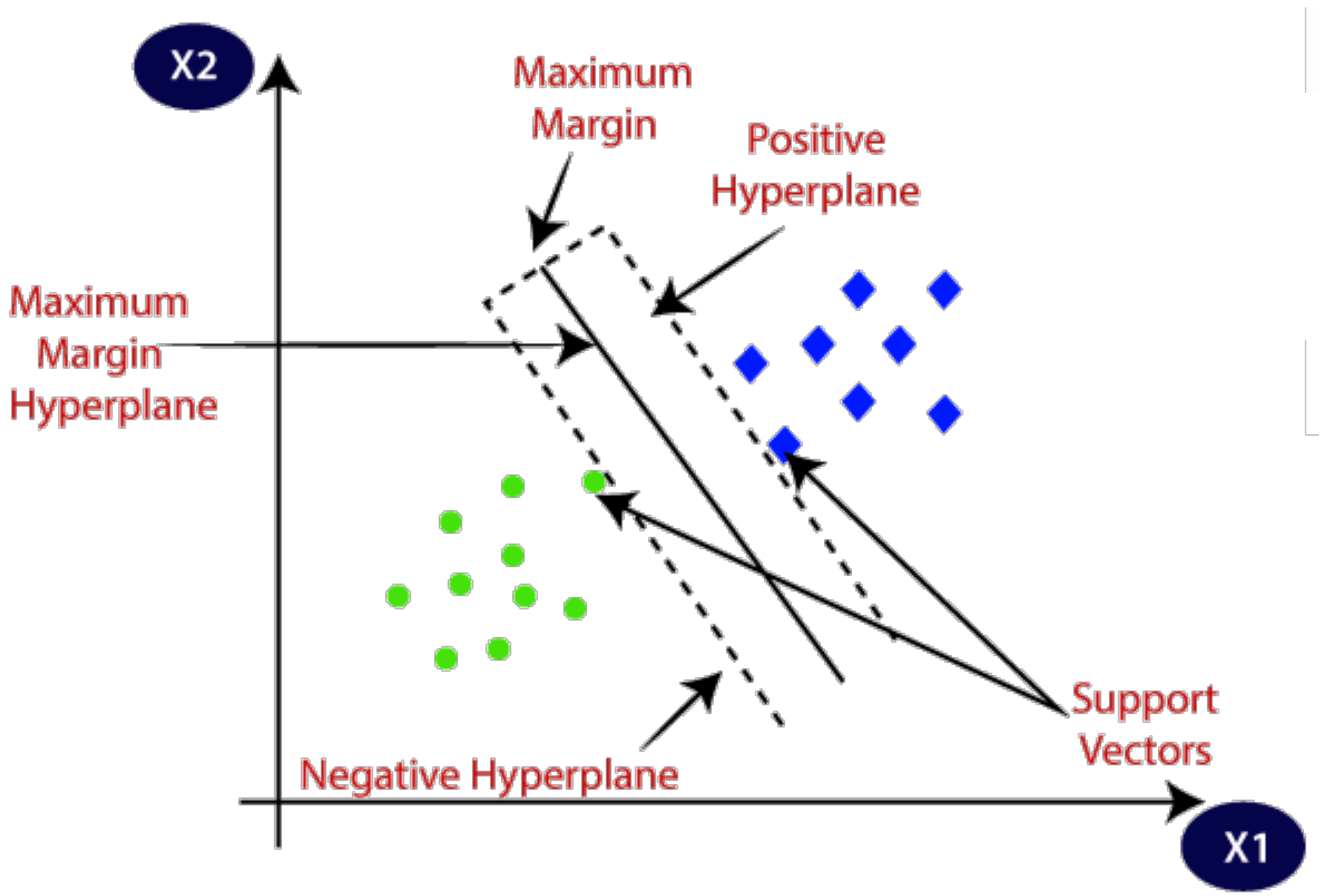
Machine Learning methods



Supervised Learning

In supervised learning data are labelled. Each point must present features (or covariates) and a label. The goal is to learn a function that maps covariates into the label.

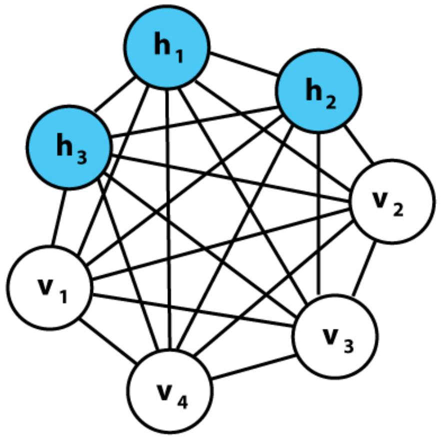
Examples of supervised learning methods are Support Vectors Machines (SVMs).

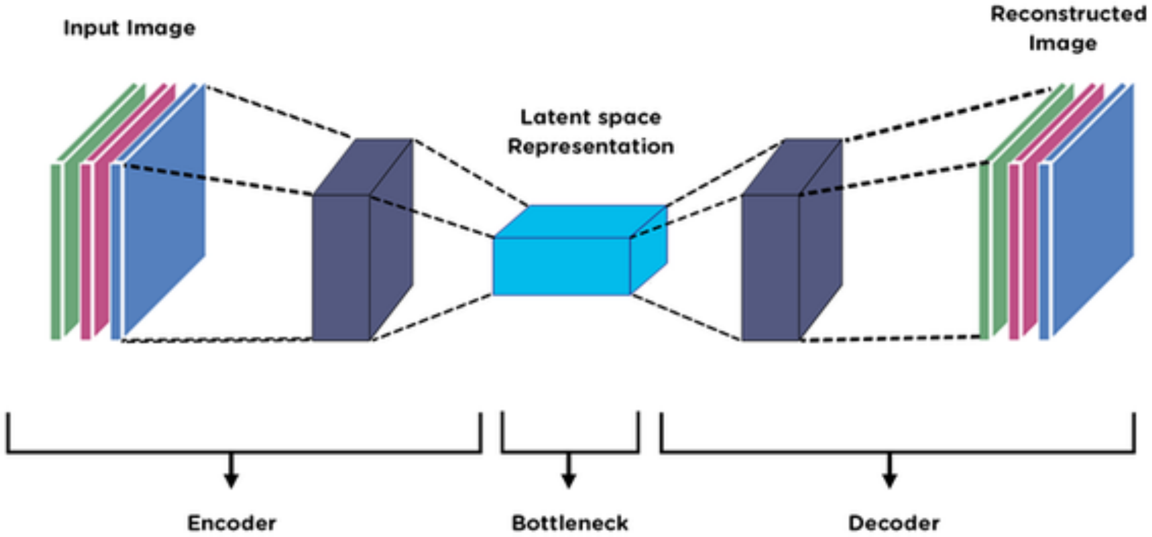


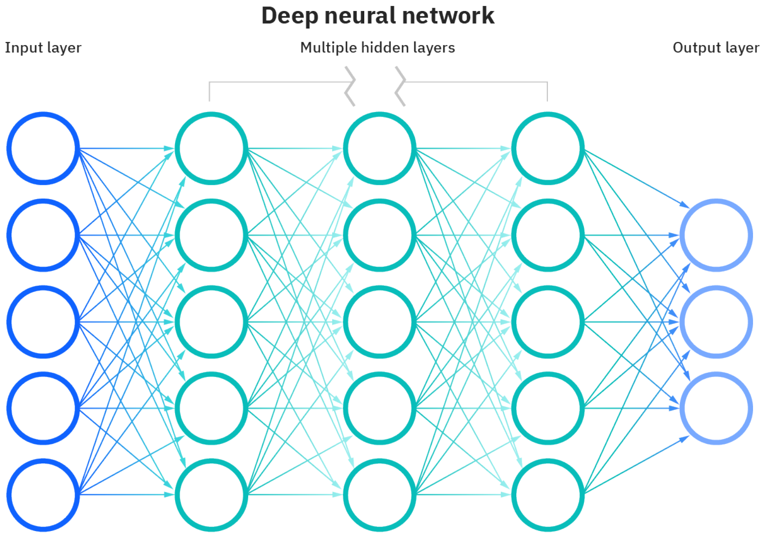
Unsupervised Learning

Unsupervised learning is an algorithm that tries learning patterns from non-labelled data. The algorithm is, therefore, forced to build a simplified representation of data, retrieving information from them.

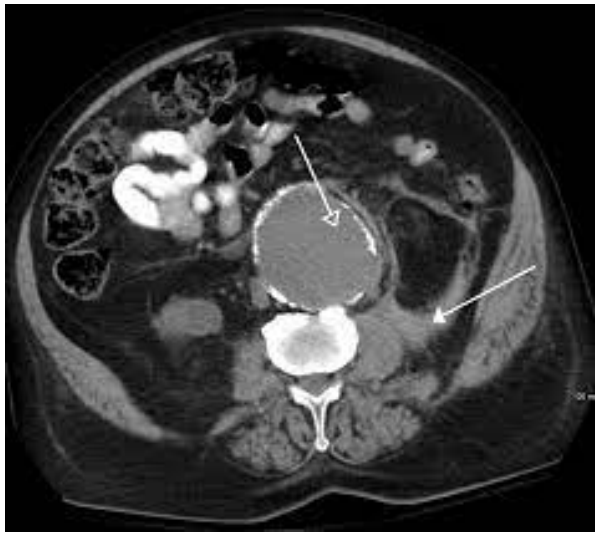
Examples of unsupervised learning methods are Convolutional Neural Networks, Autoencoders and Deep Neural Networks.

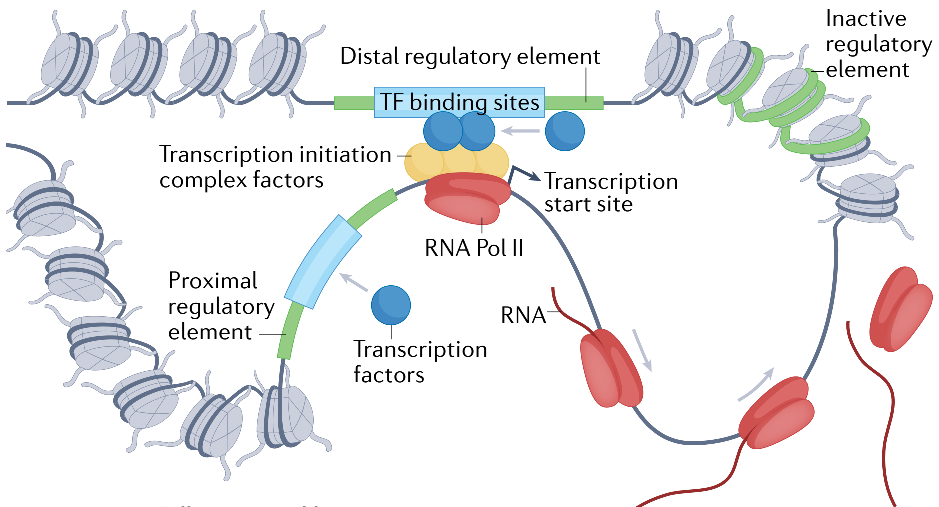


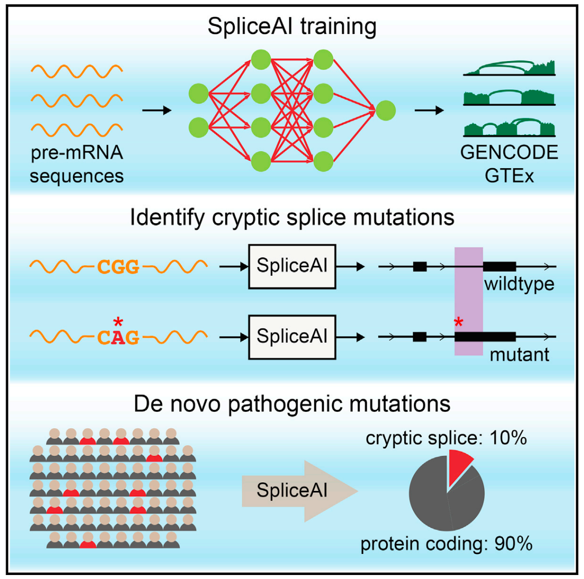


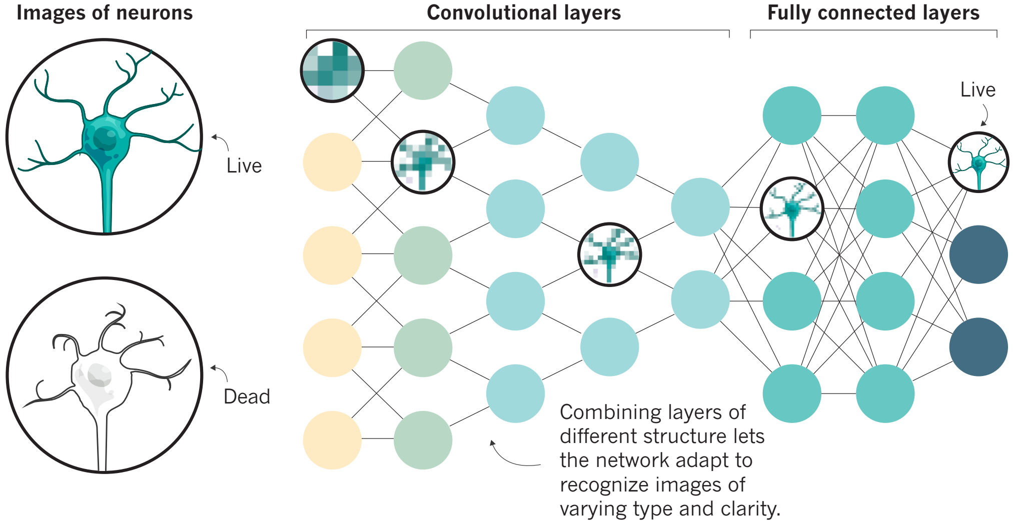


What are we able to do using machine learning in bioinformatics?

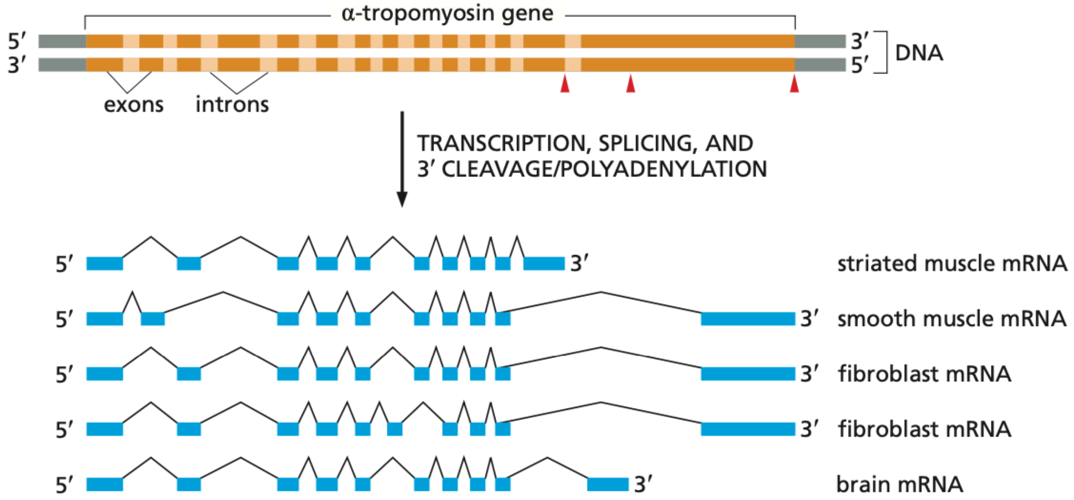


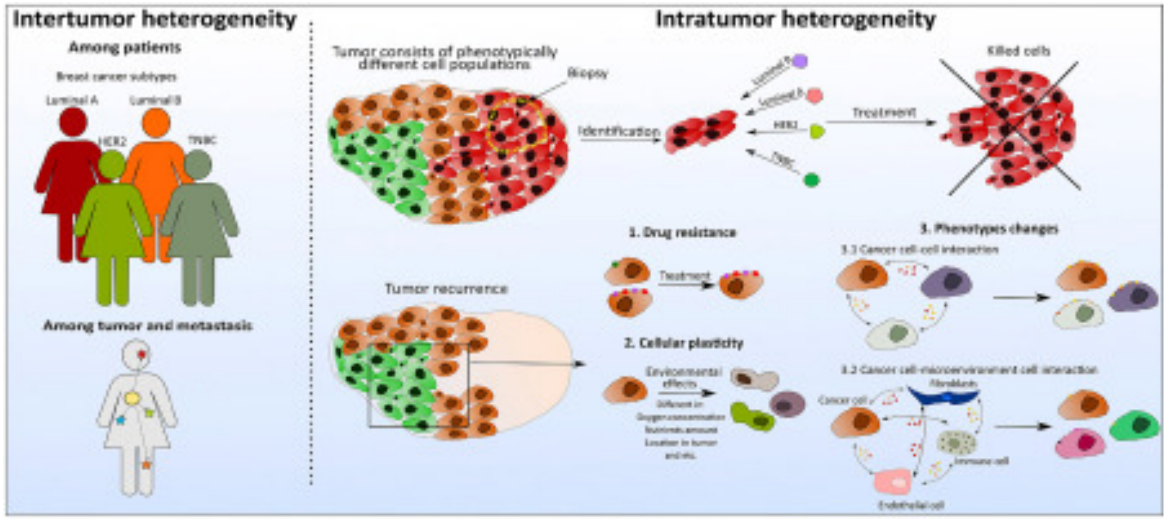

Distal regulatory element
TF binding sites
Inactive regulatory element
Transcription initiation complex factors
Transcription start site
RNA Pol II
Proximal regulatory element
Transcription factors
RNA

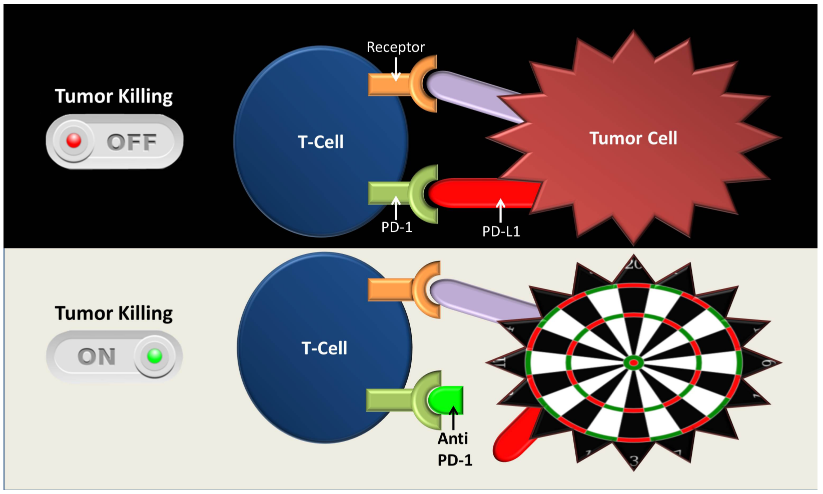




Combining layers of different structure lets the network adapt to recognize images of varying type and clarity.





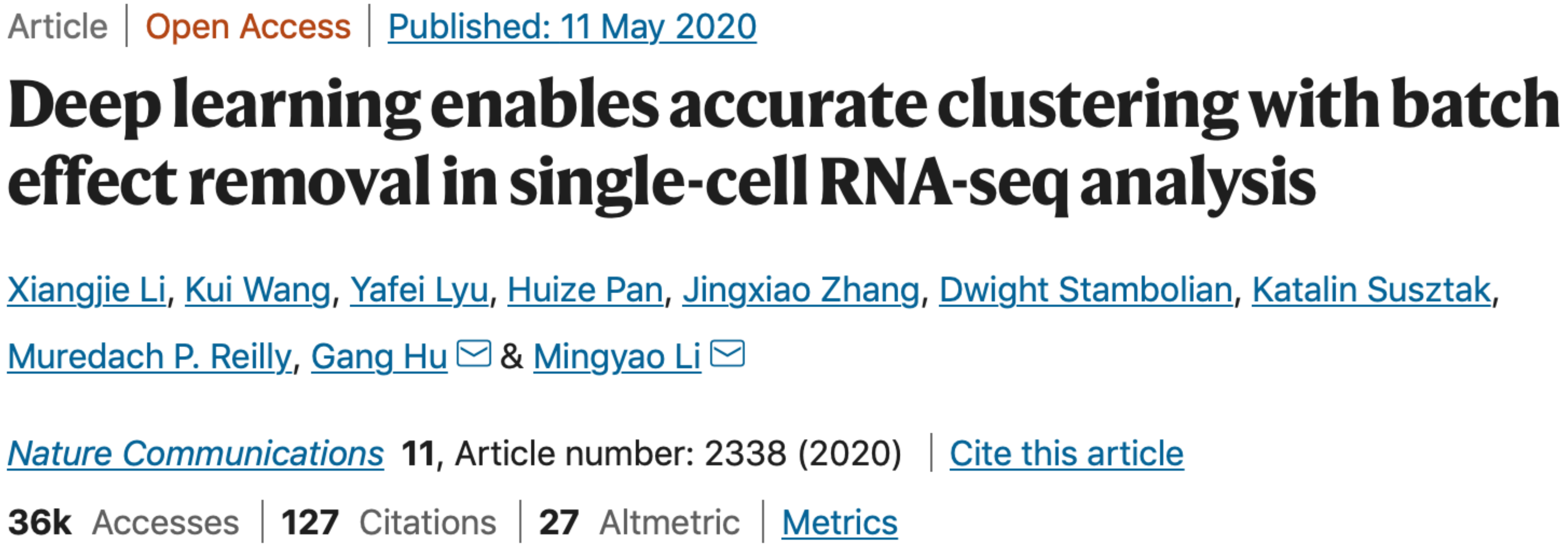


Let’s go back on our way… AI applied in RNA-seq analyses

3. Managing the Heterogeneity of Cancer Transcriptomes

3.1 Batch-Correction of Technical Heterogeneity

Cit. 49



Article | Open Access | Published: 11 May 2020

Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis

Xiangjie Li, Kui Wang, Yafei Lyu, Huize Pan, Jingxiao Zhang, Dwight Stambolian, Katalin Susztak, Muredach P. Reilly, Gang Hu ✉ & Mingyao Li ✉

Nature Communications **11**, Article number: 2338 (2020) | Cite this article

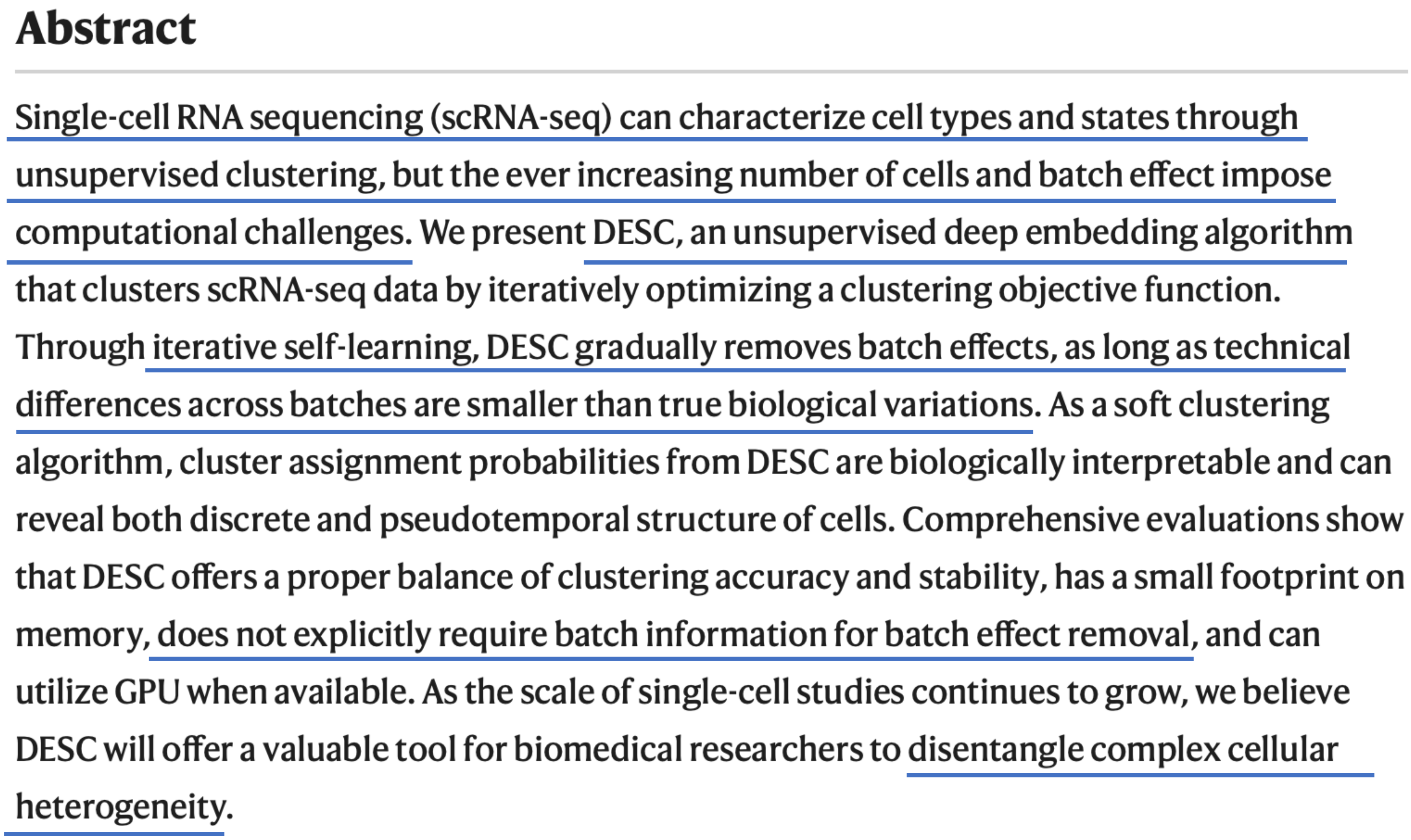
36k Accesses | **127** Citations | **27** Altmetric | Metrics

https://www.nature.com/articles/s41467-020-15851-3

3. Managing the Heterogeneity of Cancer Transcriptomes

3.1 Batch-Correction of Technical Heterogeneity

Cit. 49



Abstract

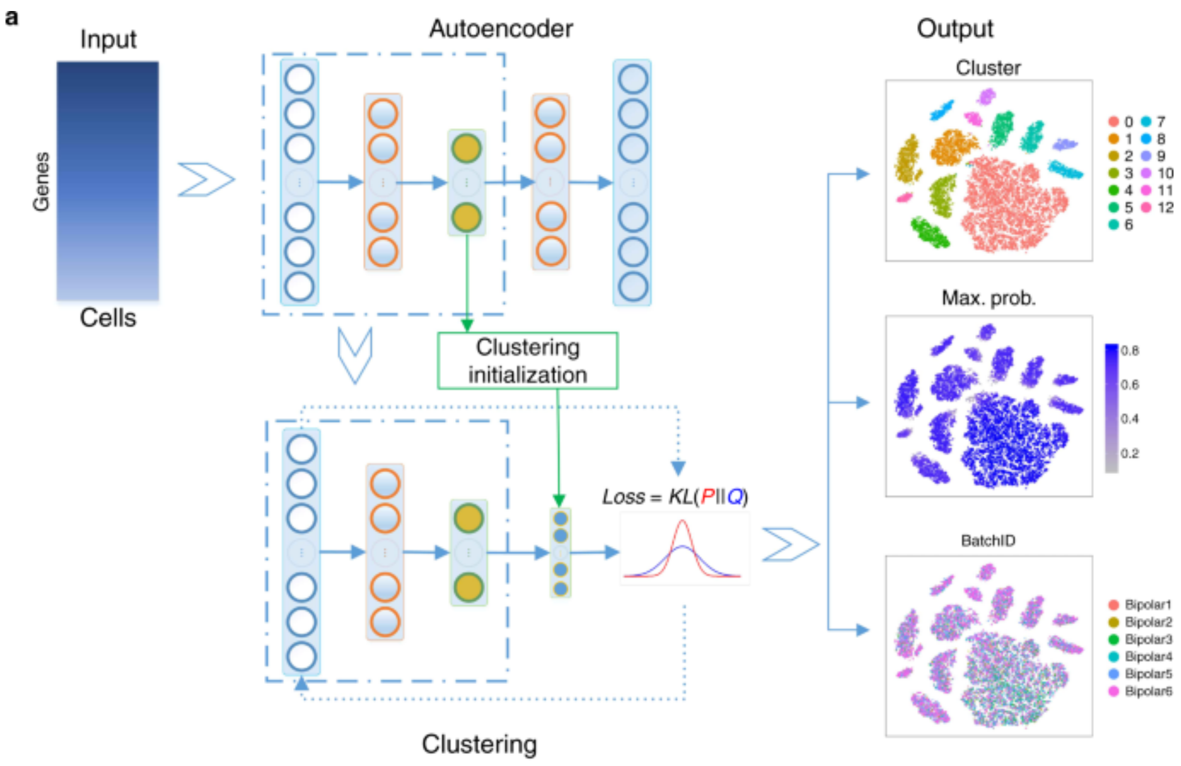
Single-cell RNA sequencing (scRNA-seq) can characterize cell types and states through unsupervised clustering, but the ever increasing number of cells and batch effect impose computational challenges. We present DESC, an unsupervised deep embedding algorithm that clusters scRNA-seq data by iteratively optimizing a clustering objective function. Through iterative self-learning, DESC gradually removes batch effects, as long as technical differences across batches are smaller than true biological variations. As a soft clustering algorithm, cluster assignment probabilities from DESC are biologically interpretable and can reveal both discrete and pseudotemporal structure of cells. Comprehensive evaluations show that DESC offers a proper balance of clustering accuracy and stability, has a small footprint on memory, does not explicitly require batch information for batch effect removal, and can utilize GPU when available. As the scale of single-cell studies continues to grow, we believe DESC will offer a valuable tool for biomedical researchers to disentangle complex cellular heterogeneity.

3. Managing the Heterogeneity of Cancer Transcriptomes

3.1 Batch-Correction of Technical Heterogeneity

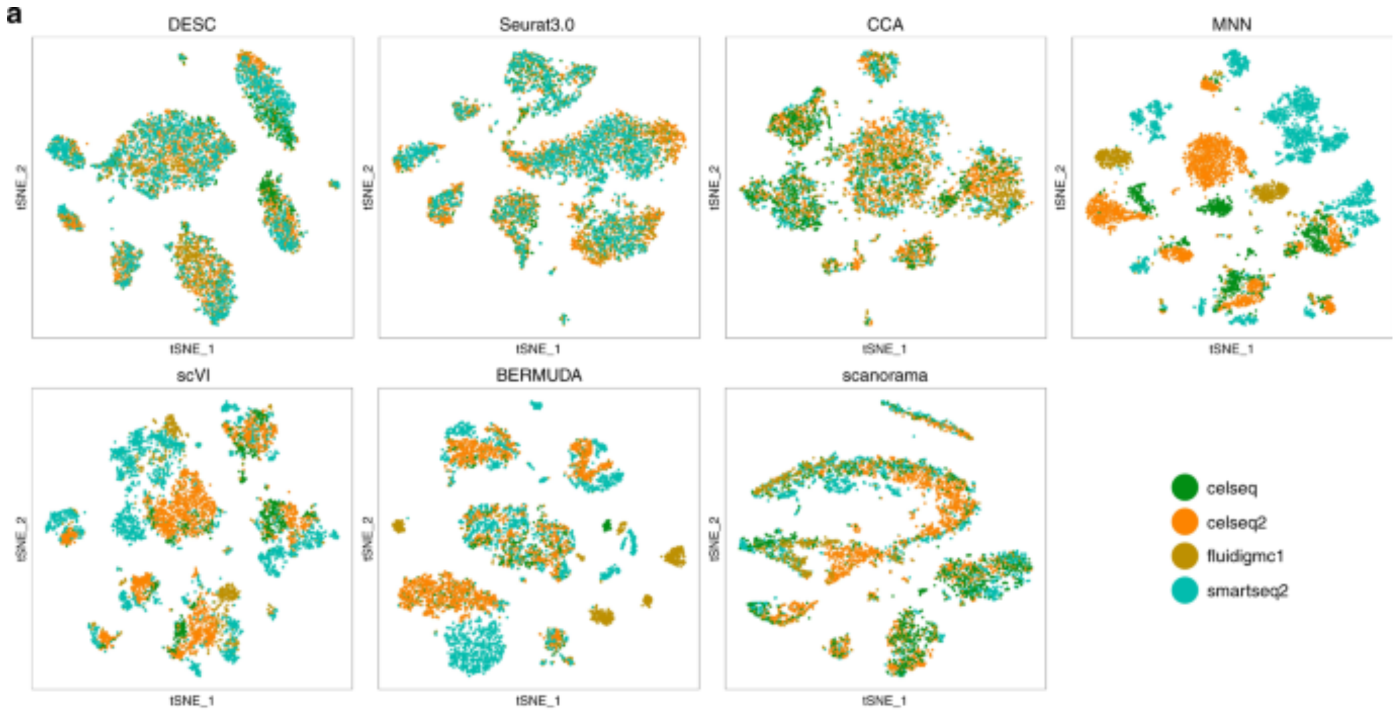
Cit. 49

Fig. 1



a Overview of the DESC framework. DESC starts with parameter initialization in which a stacked autoencoder is used for pretraining and learning a low-dimensional representation of the input gene expression matrix. The resulting encoder is then added to the iterative clustering neural network to cluster cells iteratively. The final output of DESC includes cluster assignment, the corresponding probabilities for cluster assignment for each cell, and the low-dimensional representation of the data;

Fig. 4

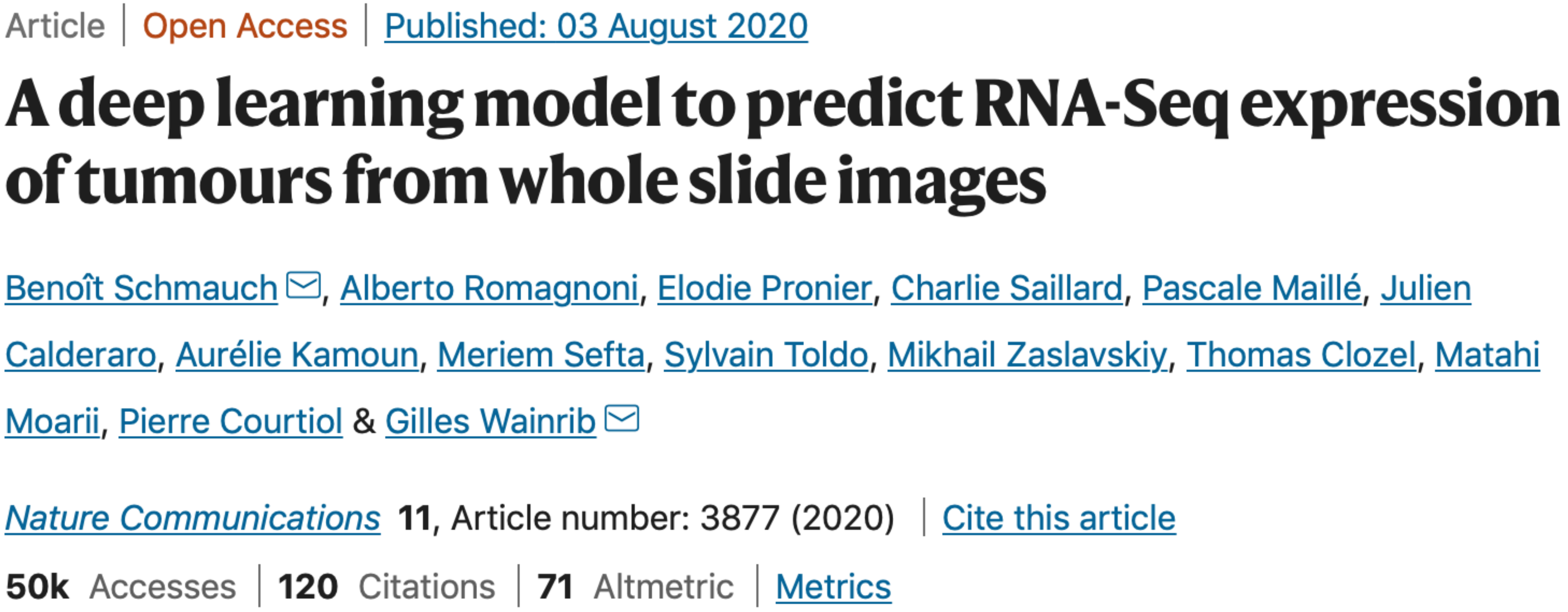


a The t-SNE plots in which cells were colored by batch.

3. Managing the Heterogeneity of Cancer Transcriptomes

3.2 Dimensionality reduction approaches
3.2.1 Feature Extraction

Cit. 93



Article | Open Access | Published: 03 August 2020

A deep learning model to predict RNA-Seq expression of tumours from whole slide images

Benoît Schmauch, Alberto Romagnoni, Elodie Pronier, Charlie Saillard, Pascale Maillé, Julien Calderaro, Aurélie Kamoun, Meriem Sefta, Sylvain Toldo, Mikhail Zaslavskiy, Thomas Clozel, Matahi Moarii, Pierre Courtiol & Gilles Wainrib

Nature Communications **11**, Article number: 3877 (2020) | Cite this article

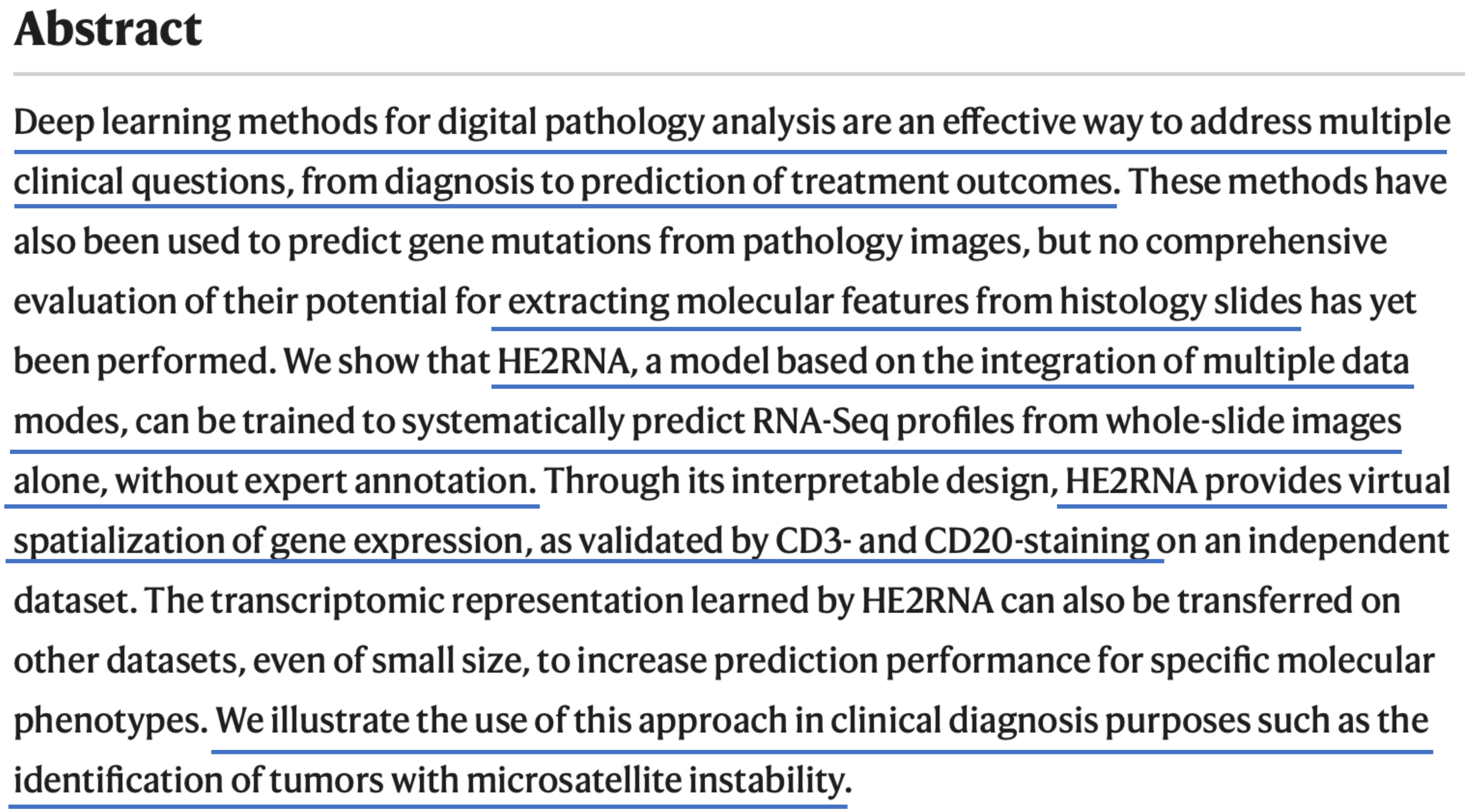
50k Accesses | **120** Citations | **71** Altmetric | Metrics

https://www.nature.com/articles/s41467-020-17678-4

3. Managing the Heterogeneity of Cancer Transcriptomes

3.2 Dimensionality reduction approaches
3.2.1 Feature Extraction

Cit. 93



Abstract

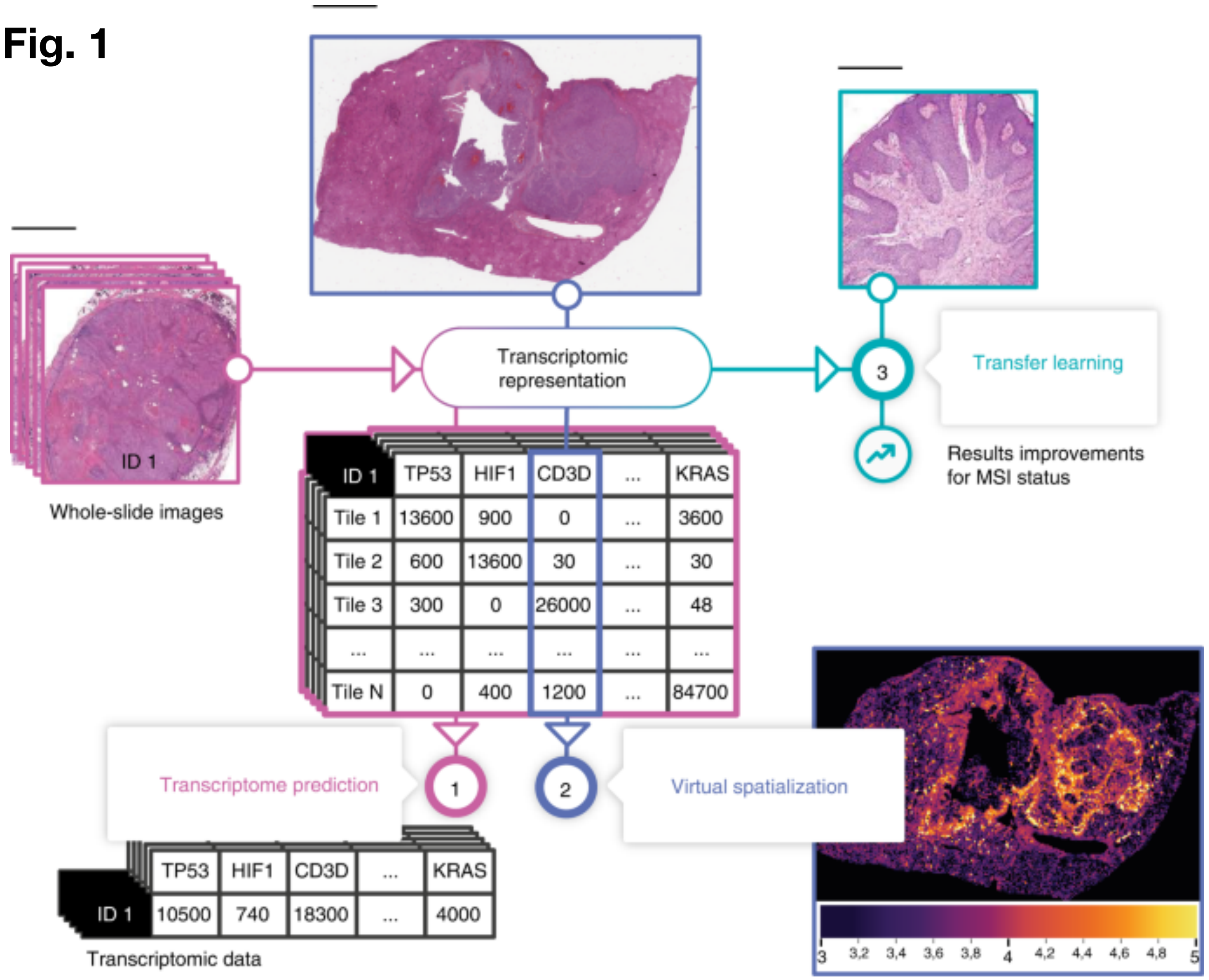
Deep learning methods for digital pathology analysis are an effective way to address multiple clinical questions, from diagnosis to prediction of treatment outcomes. These methods have also been used to predict gene mutations from pathology images, but no comprehensive evaluation of their potential for extracting molecular features from histology slides has yet been performed. We show that HE2RNA, a model based on the integration of multiple data modes, can be trained to systematically predict RNA-Seq profiles from whole-slide images alone, without expert annotation. Through its interpretable design, HE2RNA provides virtual spatialization of gene expression, as validated by CD3- and CD20-staining on an independent dataset. The transcriptomic representation learned by HE2RNA can also be transferred on other datasets, even of small size, to increase prediction performance for specific molecular phenotypes. We illustrate the use of this approach in clinical diagnosis purposes such as the identification of tumors with microsatellite instability.

3. Managing the Heterogeneity of Cancer Transcriptomes

3.2 Dimensionality reduction approaches
3.2.1 Feature Extraction

Cit. 93

Fig. 1



ID 1	TP53	HIF1	CD3D	...	KRAS
Tile 1	13600	900	0	...	3600
Tile 2	600	13600	30	...	30
Tile 3	300	0	26000	...	48
...	...	...	...	...	...
Tile N	0	400	1200	...	84700

	TP53	HIF1	CD3D	...	KRAS
ID 1	10500	740	18300	...	4000

Hematoxylin & eosin (H&E)-stained histology slides and RNA-Seq data (FPKM-UQ values) for 28 different cancer types and 8725 patients were collected from The Cancer Genome Atlas (TCGA) and used to train the neural network HE2RNA to predict transcriptomic profile from the corresponding high-definition whole-slide images (WSI). During this task, the neural network learned an internal representation encoding both information from tiled images and gene expression levels. This transcriptomic representation can be used for: (1) transcriptome prediction from images without associated RNA sequencing. (2) The virtual spatialization of transcriptomic data. For each predicted coding or noncoding gene, a score is calculated for each tile on the corresponding WSI, which can be interpreted as the predicted gene expression for this tile (even though the real value is available only for the slide). These predictive scores can be used to generate heatmaps for each gene for which expression is significantly predicted. (3) Improving predictive performances for different tasks, in a transfer learning framework, as shown here for a realistic setup, for microsatellite instability (MSI) status prediction from non-annotated WSIs. Scale bar: 5 mm.

3. Managing the Heterogeneity of Cancer Transcriptomes

3.2 Dimensionality reduction approaches
3.2.1 Feature Selection

Cit. 55

Research article | Open Access | Published: 29 October 2018

Feature selection of gene expression data for Cancer classification using double RBF-kernels

Shenghui Liu, Chunrui Xu, Yusen Zhang ✉, Jiaguo Liu, Bin Yu, Xiaoping Liu ✉ & Matthias Dehmer

BMC Bioinformatics **19**, Article number: 396 (2018) | Cite this article

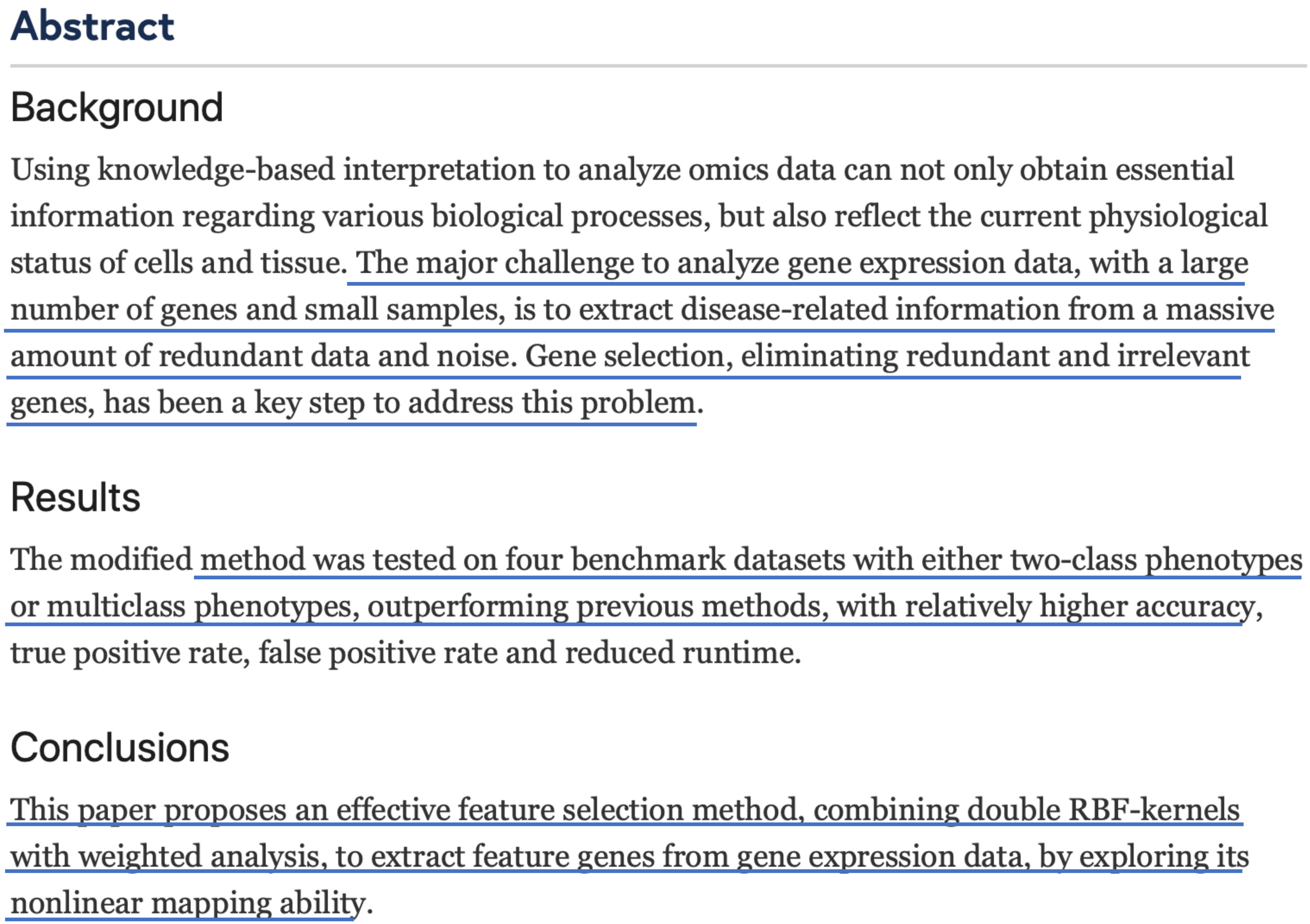
10k Accesses | **39** Citations | **1** Altmetric | Metrics

https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2400-2

3. Managing the Heterogeneity of Cancer Transcriptomes

3.2 Dimensionality reduction approaches
3.2.1 Feature Selection

Cit. 55



Abstract

Background

Using knowledge-based interpretation to analyze omics data can not only obtain essential information regarding various biological processes, but also reflect the current physiological status of cells and tissue. The major challenge to analyze gene expression data, with a large number of genes and small samples, is to extract disease-related information from a massive amount of redundant data and noise. Gene selection, eliminating redundant and irrelevant genes, has been a key step to address this problem.

Results

The modified method was tested on four benchmark datasets with either two-class phenotypes or multiclass phenotypes, outperforming previous methods, with relatively higher accuracy, true positive rate, false positive rate and reduced runtime.

Conclusions

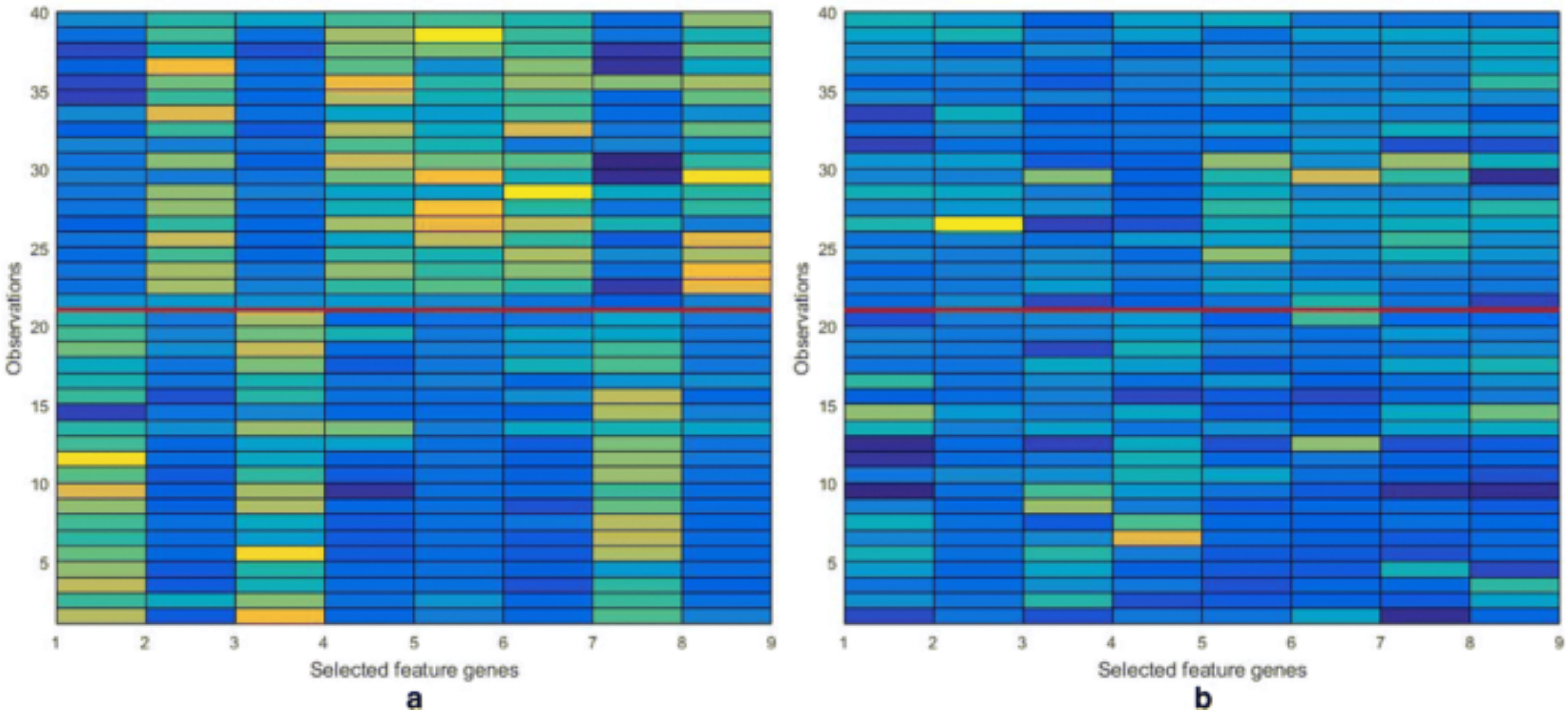
This paper proposes an effective feature selection method, combining double RBF-kernels with weighted analysis, to extract feature genes from gene expression data, by exploring its nonlinear mapping ability.

3. Managing the Heterogeneity of Cancer Transcriptomes

3.2 Dimensionality reduction approaches
3.2.1 Feature Selection

Cit. 55

Fig. 6

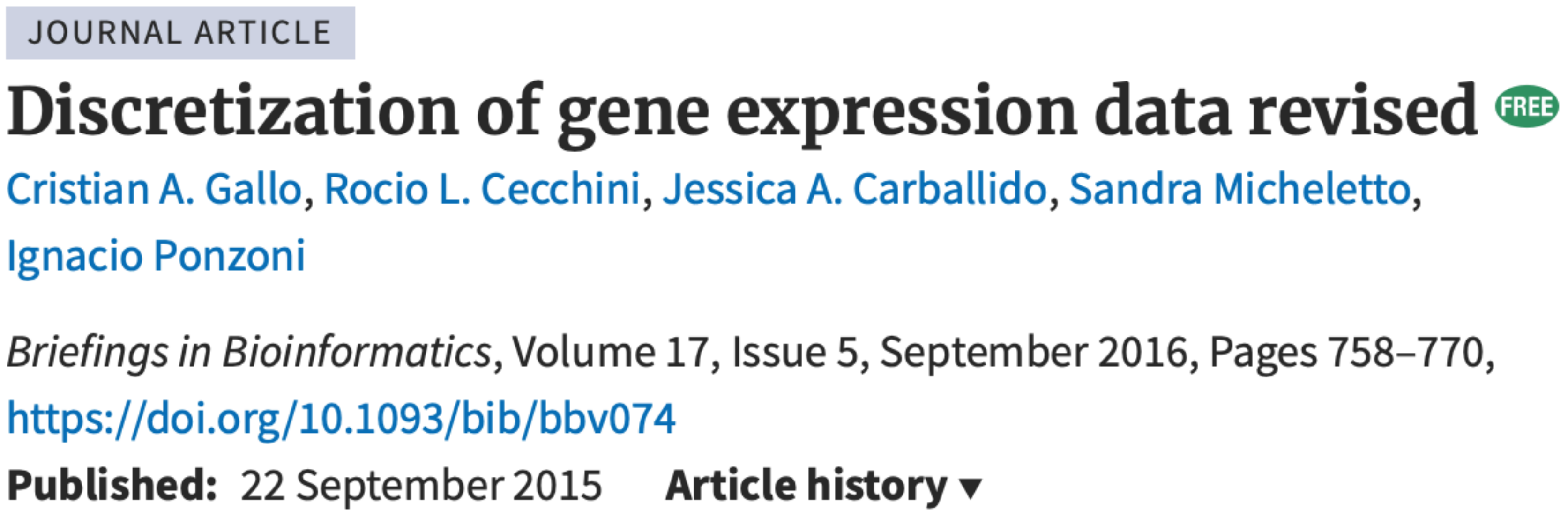


The colormap of the expression profiles for nine most significant genes selected by DKBCGS (**a**) and for 9 randomly chosen genes (**b**). The red line distinguishes between cancer samples and normal samples

3. Managing the Heterogeneity of Cancer Transcriptomes

3.3 Data distribution transformation

Cit. 104



JOURNAL ARTICLE

Discretization of gene expression data revised

FREE

Cristian A. Gallo, Rocio L. Cecchini, Jessica A. Carballido, Sandra Micheletto, Ignacio Ponzoni

Briefings in Bioinformatics, Volume 17, Issue 5, September 2016, Pages 758–770, https://doi.org/10.1093/bib/bbv074

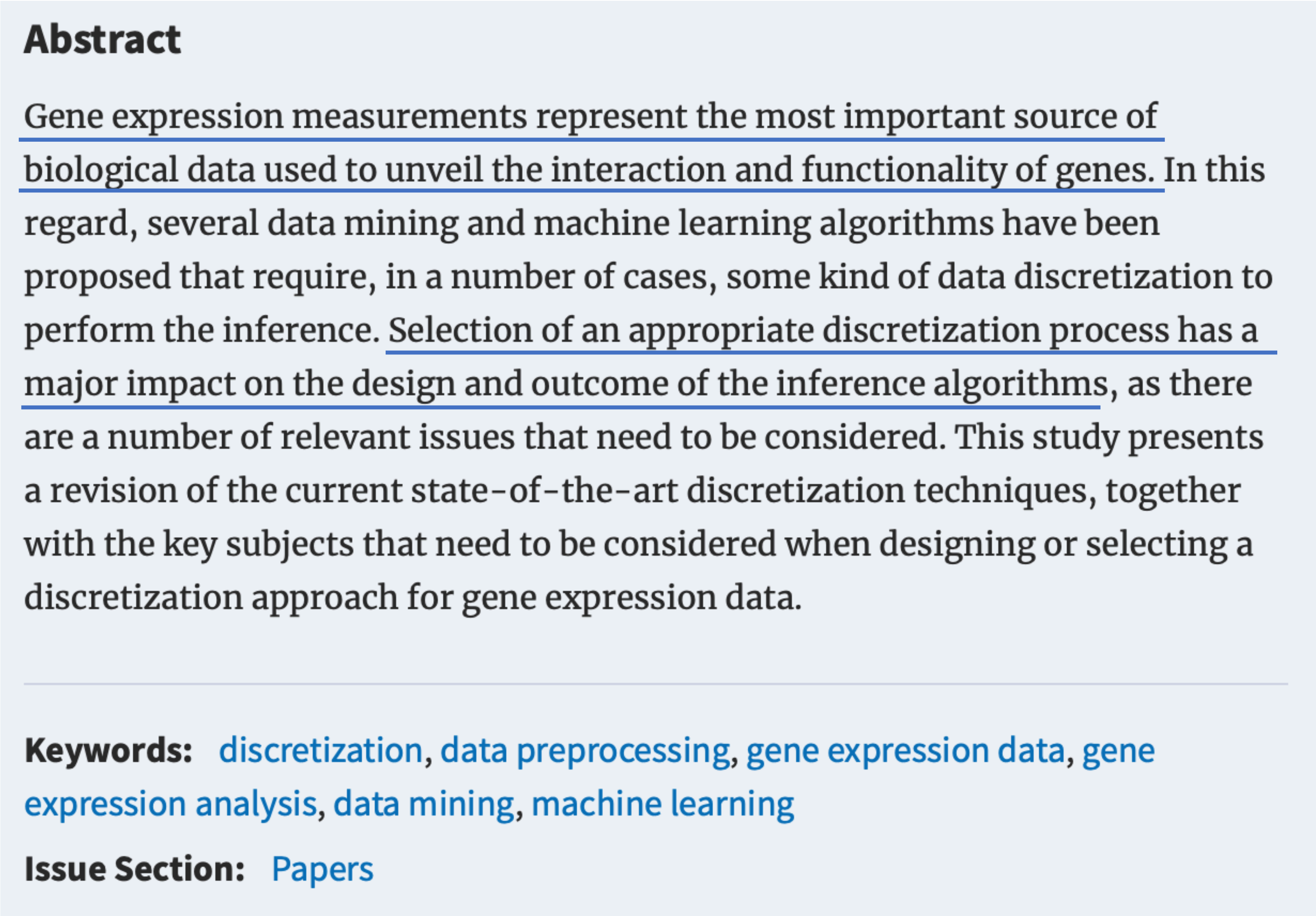
Published: 22 September 2015 **Article history** ▾

https://academic.oup.com/bib/article/17/5/758/2261412

3. Managing the Heterogeneity of Cancer Transcriptomes

3.3 Data distribution transformation

Cit. 104



Abstract

Gene expression measurements represent the most important source of biological data used to unveil the interaction and functionality of genes. In this regard, several data mining and machine learning algorithms have been proposed that require, in a number of cases, some kind of data discretization to perform the inference. Selection of an appropriate discretization process has a major impact on the design and outcome of the inference algorithms, as there are a number of relevant issues that need to be considered. This study presents a revision of the current state-of-the-art discretization techniques, together with the key subjects that need to be considered when designing or selecting a discretization approach for gene expression data.

Keywords: discretization, data preprocessing, gene expression data, gene expression analysis, data mining, machine learning

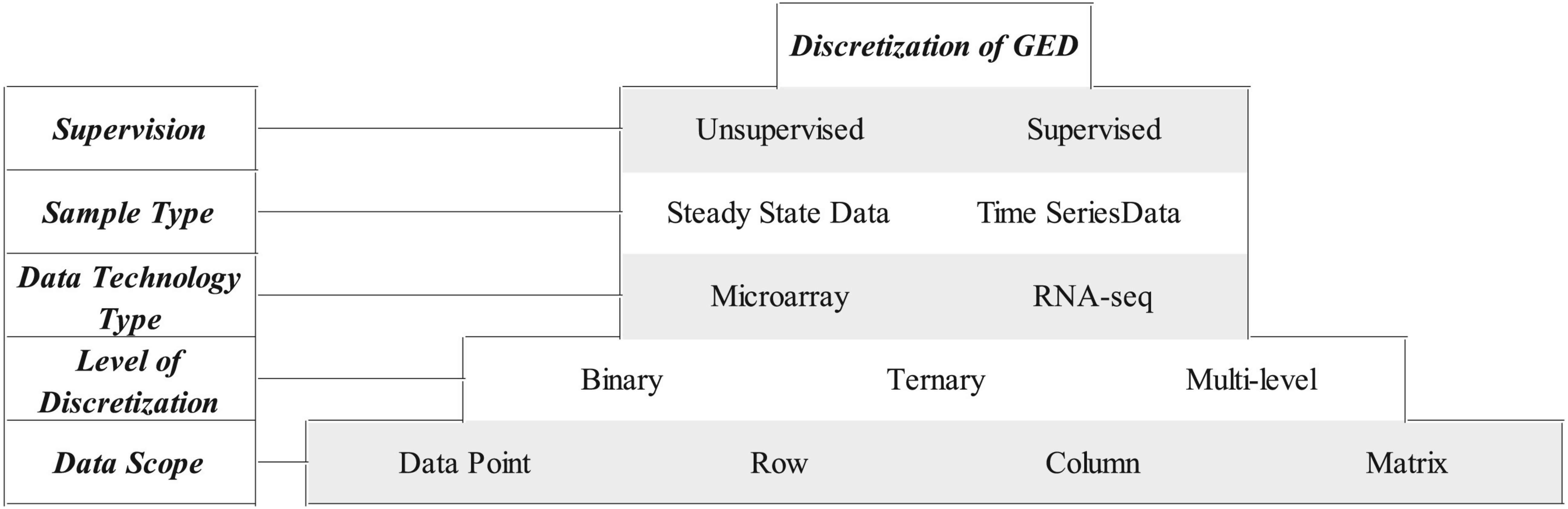
Issue Section: Papers

3. Managing the Heterogeneity of Cancer Transcriptomes

3.3 Data distribution transformation

Cit. 104

Fig. 2



Discretization of GED

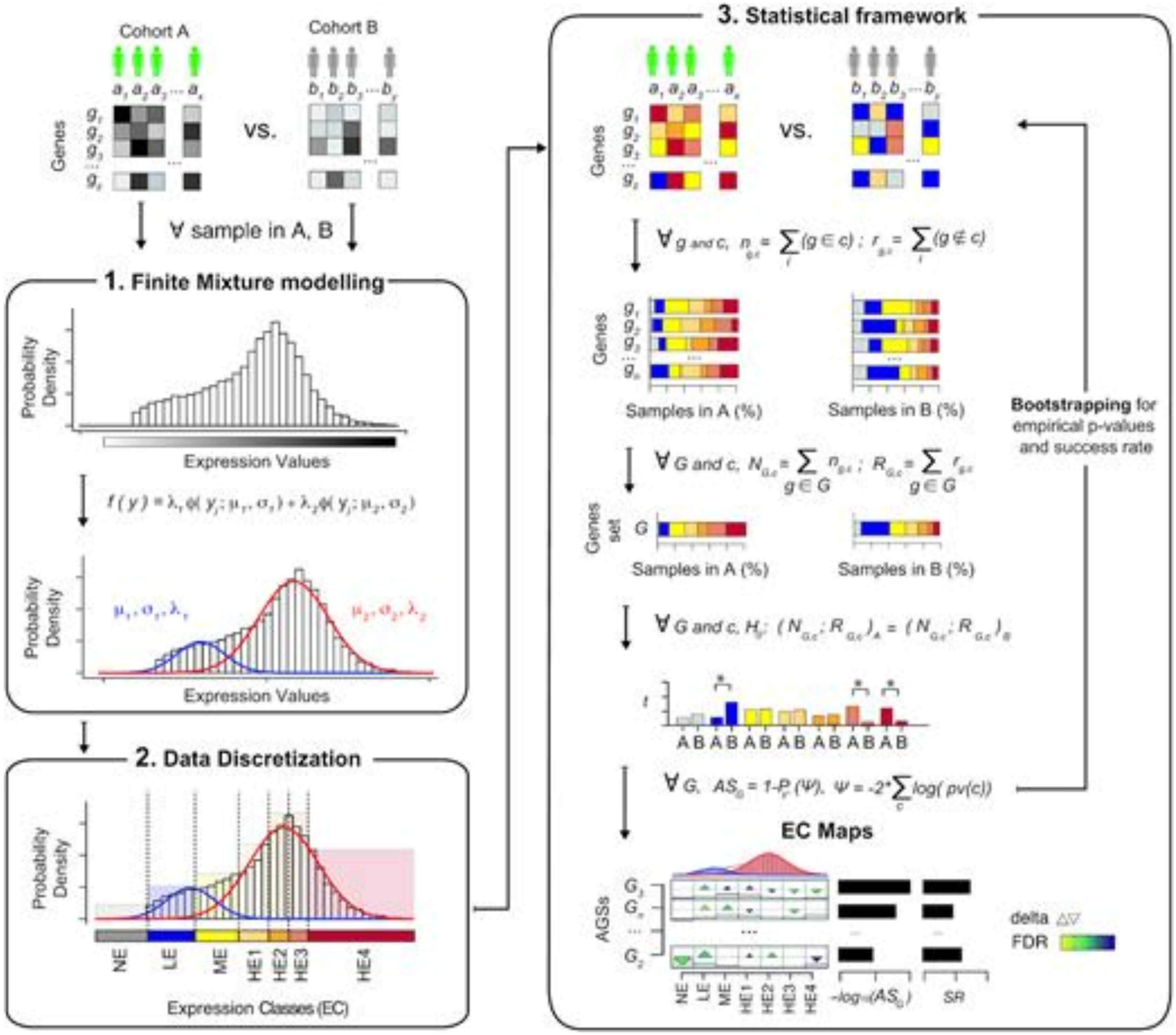
Supervision	Unsupervised	Supervised		
Sample Type	Steady State Data	Time SeriesData		
Data Technology Type	Microarray	RNA-seq		
Level of Discretization	Binary	Ternary	Multi-level	
Data Scope	Data Point	Row	Column	Matrix

Main features of gene expression discretization with their multiple variants.

3. Managing the Heterogeneity of Cancer Transcriptomes

3.3 Data distribution transformation

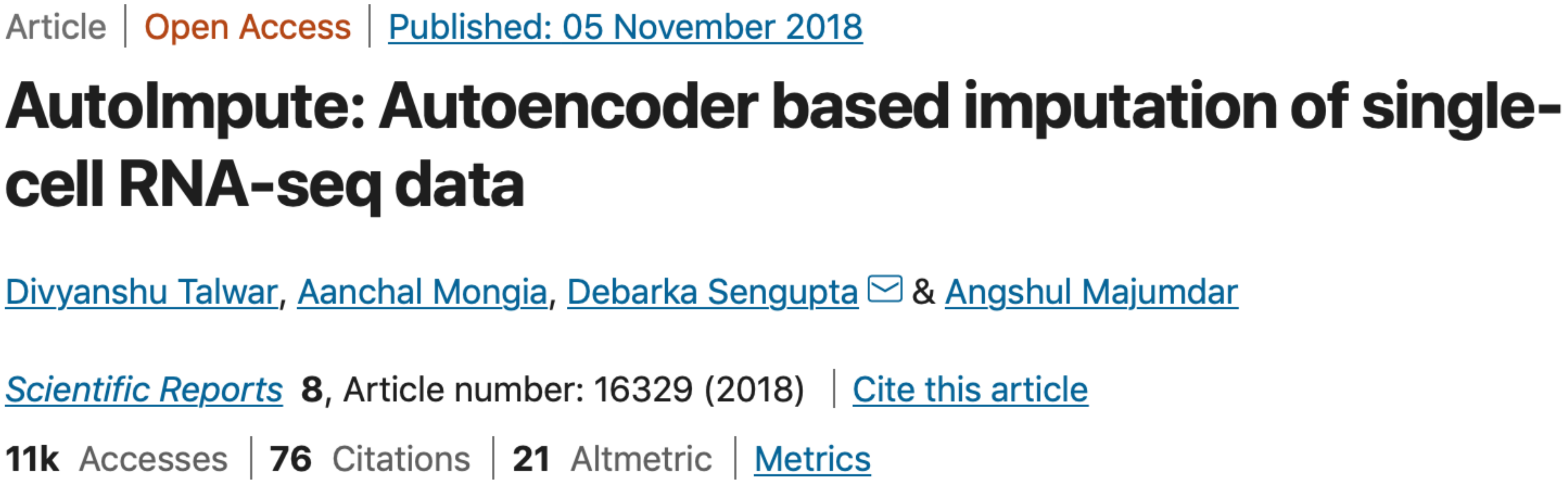
An example of expression distribution discretisation… The Gene Set Enrichment Class Analysis (GSECA)


Cohort A
Cohort B
$a_1\ a_2\ a_3 \cdots a_x$
$b_1\ b_2\ b_3 \cdots b_y$
Genes
g_1
g_2
g_3
g_z
vs.
∀ sample in A, B
1. Finite Mixture modelling
Probability Density
Expression Values
$f(y) = \lambda_1\phi(y_i;\mu_1,\sigma_1) + \lambda_2\phi(y_i;\mu_2,\sigma_2)$
$\mu_1, \sigma_1, \lambda_1$
$\mu_2, \sigma_2, \lambda_2$
2. Data Discretization
NE
LE
ME
HE1
HE2
HE3
HE4
Expression Classes (EC)
3. Statistical framework
$\forall$ g and c, $n_{g,c} = \sum_i (g \in c)$; $r_{g,c} = \sum_i (g \notin c)$
Samples in A (%)
Samples in B (%)
$\forall$ G and c, $N_{G,c} = \sum_{g \in G} n_{g,c}$; $R_{G,c} = \sum_{g \in G} r_{g,c}$
Genes set
G
$\forall$ G and c, H_0: $(N_{G,c} : R_{G,c})_A = (N_{G,c} : R_{G,c})_B$
t
A B A B A B A B A B A B A B
$\forall$ G, $AS_G = 1 - P_{\chi^2}(\Psi)$, $\Psi = -2 * \sum_c \log(pv(c))$
Bootstrapping for empirical p-values and success rate
EC Maps
AGSs
G_1
G_2
G_z
$-\log_{10}(AS_G)$
SR
delta △▽
FDR

3. Managing the Heterogeneity of Cancer Transcriptomes

3.4 Data reconstruction: the sparsity issue

Cit. 59



Article | Open Access | Published: 05 November 2018

AutoImpute: Autoencoder based imputation of single-cell RNA-seq data

Divyanshu Talwar, Aanchal Mongia, Debarka Sengupta & Angshul Majumdar

Scientific Reports **8**, Article number: 16329 (2018) | Cite this article

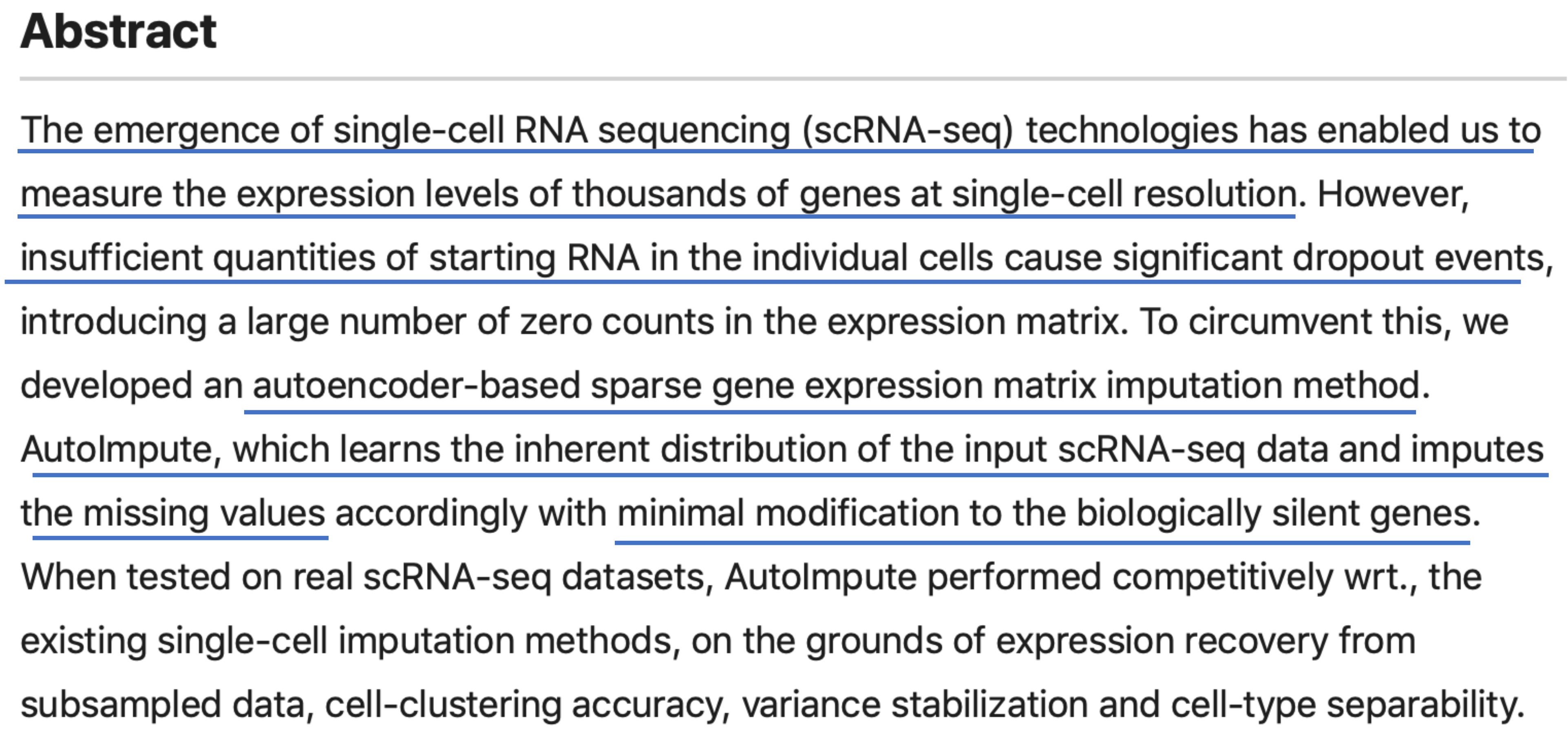
11k Accesses | **76** Citations | **21** Altmetric | Metrics

https://www.nature.com/articles/s41598-018-34688-x

3. Managing the Heterogeneity of Cancer Transcriptomes

3.4 Data reconstruction: the sparsity issue

Cit. 59



Abstract

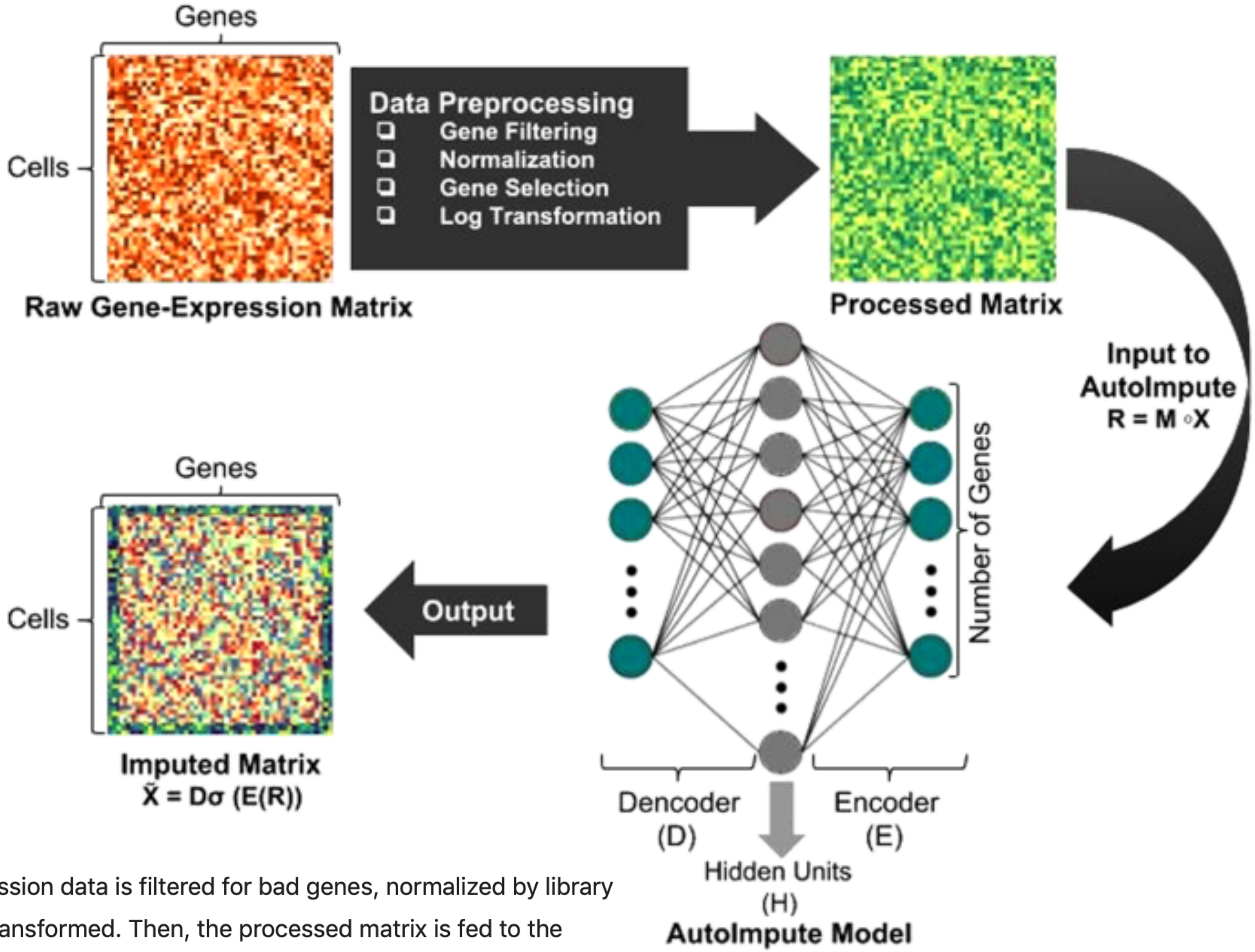
The emergence of single-cell RNA sequencing (scRNA-seq) technologies has enabled us to measure the expression levels of thousands of genes at single-cell resolution. However, insufficient quantities of starting RNA in the individual cells cause significant dropout events, introducing a large number of zero counts in the expression matrix. To circumvent this, we developed an autoencoder-based sparse gene expression matrix imputation method. AutoImpute, which learns the inherent distribution of the input scRNA-seq data and imputes the missing values accordingly with minimal modification to the biologically silent genes. When tested on real scRNA-seq datasets, AutoImpute performed competitively wrt., the existing single-cell imputation methods, on the grounds of expression recovery from subsampled data, cell-clustering accuracy, variance stabilization and cell-type separability.

3. Managing the Heterogeneity of Cancer Transcriptomes

3.4 Data reconstruction: the sparsity issue

Cit. 59

Fig. 1

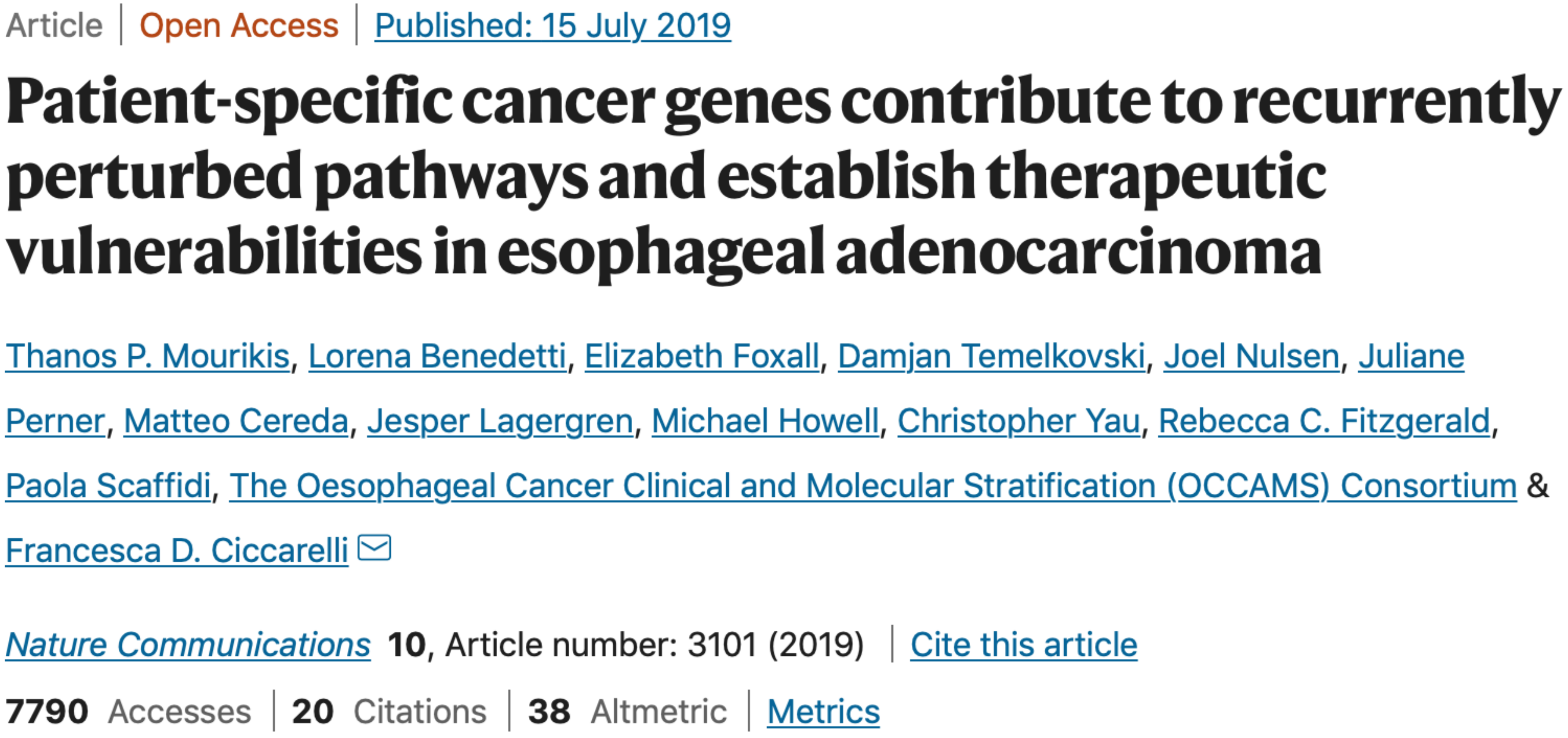


AutoImpute pipeline: The raw gene expression data is filtered for bad genes, normalized by library size, pruned by gene-selection and log transformed. Then, the processed matrix is fed to the AutoImpute model for learning expression data representation and finally reconstructing the imputed matrix.

4. AI mining of cancer transcriptomes

3.4 Assessing inter-tumor heterogeneity: classification of cancer subtypes

Cit. 111



Article | Open Access | Published: 15 July 2019

Patient-specific cancer genes contribute to recurrently perturbed pathways and establish therapeutic vulnerabilities in esophageal adenocarcinoma

Thanos P. Mourikis, Lorena Benedetti, Elizabeth Foxall, Damjan Temelkovski, Joel Nulsen, Juliane Perner, Matteo Cereda, Jesper Lagergren, Michael Howell, Christopher Yau, Rebecca C. Fitzgerald, Paola Scaffidi, The Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) Consortium & Francesca D. Ciccarelli ✉

Nature Communications **10**, Article number: 3101 (2019) | Cite this article

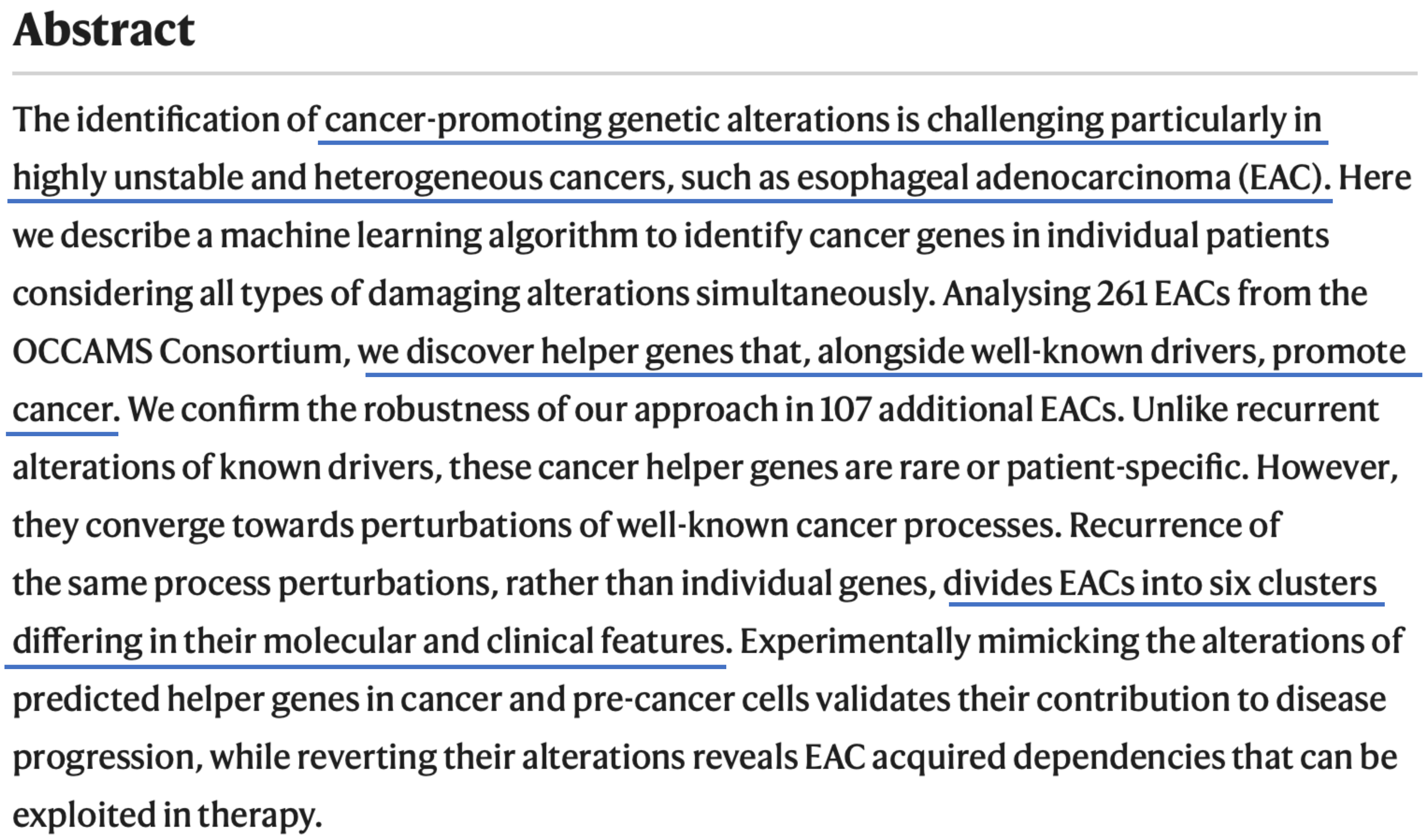
7790 Accesses | **20** Citations | **38** Altmetric | Metrics

https://www.nature.com/articles/s41467-019-10898-3

4. AI mining of cancer transcriptomes

3.4 Assessing inter-tumor heterogeneity: classification of cancer subtypes

Cit. 111



Abstract

The identification of cancer-promoting genetic alterations is challenging particularly in highly unstable and heterogeneous cancers, such as esophageal adenocarcinoma (EAC). Here we describe a machine learning algorithm to identify cancer genes in individual patients considering all types of damaging alterations simultaneously. Analysing 261 EACs from the OCCAMS Consortium, we discover helper genes that, alongside well-known drivers, promote cancer. We confirm the robustness of our approach in 107 additional EACs. Unlike recurrent alterations of known drivers, these cancer helper genes are rare or patient-specific. However, they converge towards perturbations of well-known cancer processes. Recurrence of the same process perturbations, rather than individual genes, divides EACs into six clusters differing in their molecular and clinical features. Experimentally mimicking the alterations of predicted helper genes in cancer and pre-cancer cells validates their contribution to disease progression, while reverting their alterations reveals EAC acquired dependencies that can be exploited in therapy.

4. AI mining of cancer transcriptomes

3.4 Assessing inter-tumor heterogeneity: classification of cancer subtypes

Cit. 111

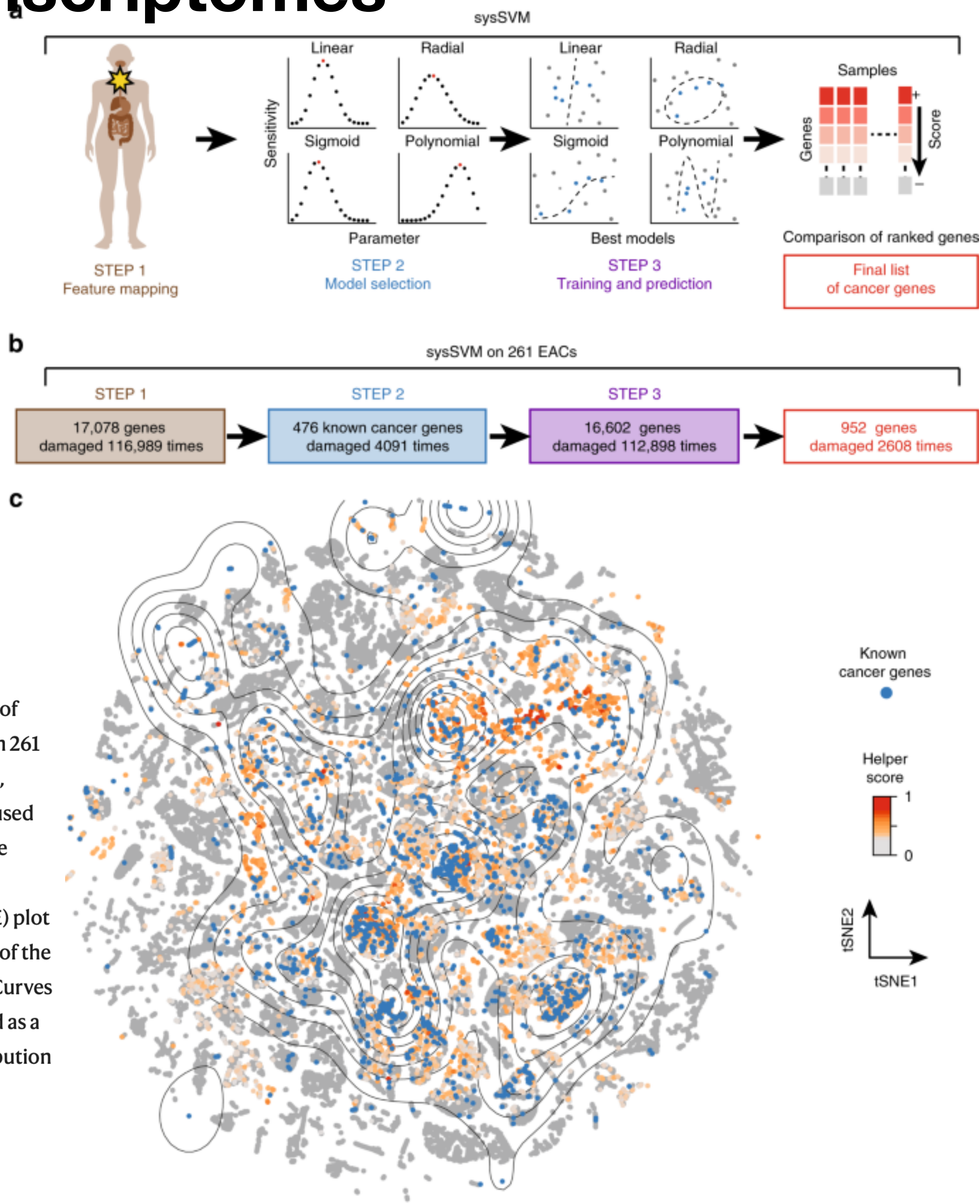


Fig. 1

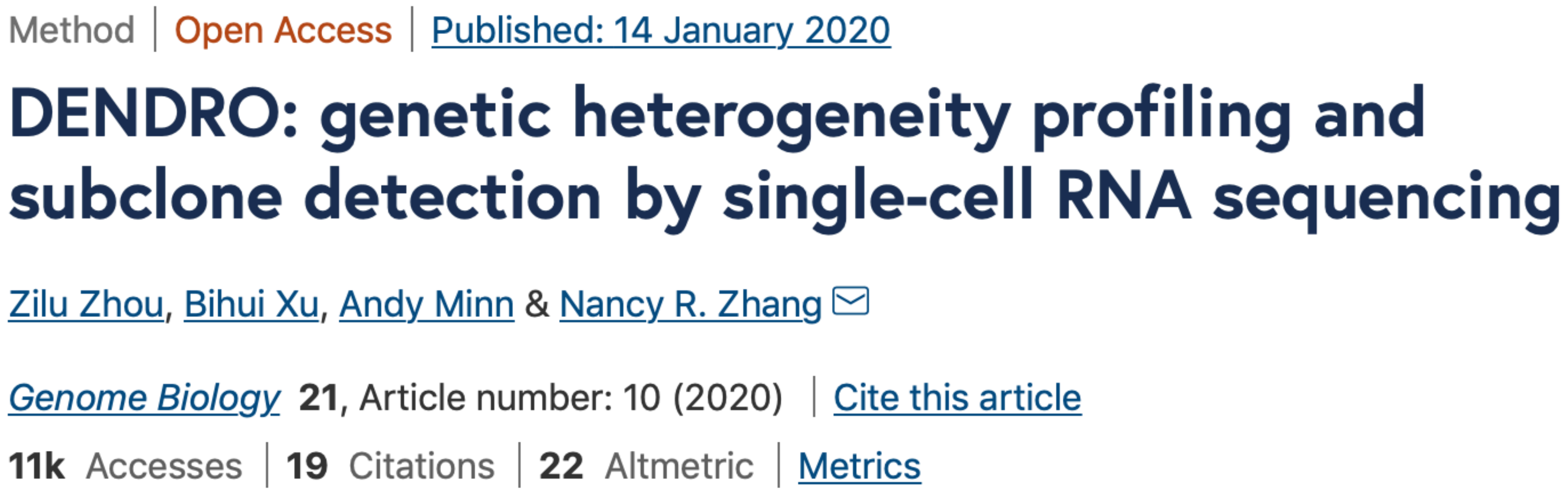
Cancer helper genes in 261 EACs. **a** Schematic workflow of the sysSVM algorithm. **b** Application of sysSVM to 261 EACs. Genes with somatic damaging alterations (*n* = 116,989) were extracted from 261 EACs and divided into training (known cancer genes, blue) and prediction (rest of altered genes, purple) sets. sysSVM was trained on the properties of known drivers and the best models were used for prediction. All altered genes were scored in each patient individually and the top 10 hits were considered as the cancer helper genes in that patient, for a total of 2608 helper alterations, corresponding to 952 unique hits (red). **c** t-distributed Stochastic Neighbour Embedding (t-SNE) plot of 116,989 altered genes in 261 EACs. Starting from the 34 properties used in sysSVM, a 2-D map of the high-dimensional data was built using Rtsne package (https://github.com/jkrijthe/Rtsne) in R. Curves are coloured according to the density of 476 known cancer genes altered 4091 times (blue) used as a training set and the rest of altered genes are coloured according to their sysSVM score. **d** Distribution

4. AI mining of cancer transcriptomes

4.2 Deciphering intra-tumor heterogeneity

4.2.1 Defining Cell Types and Clones

Cit. 70



Method | Open Access | Published: 14 January 2020

DENDRO: genetic heterogeneity profiling and subclone detection by single-cell RNA sequencing

Zilu Zhou, Bihui Xu, Andy Minn & Nancy R. Zhang

Genome Biology **21**, Article number: 10 (2020) | Cite this article

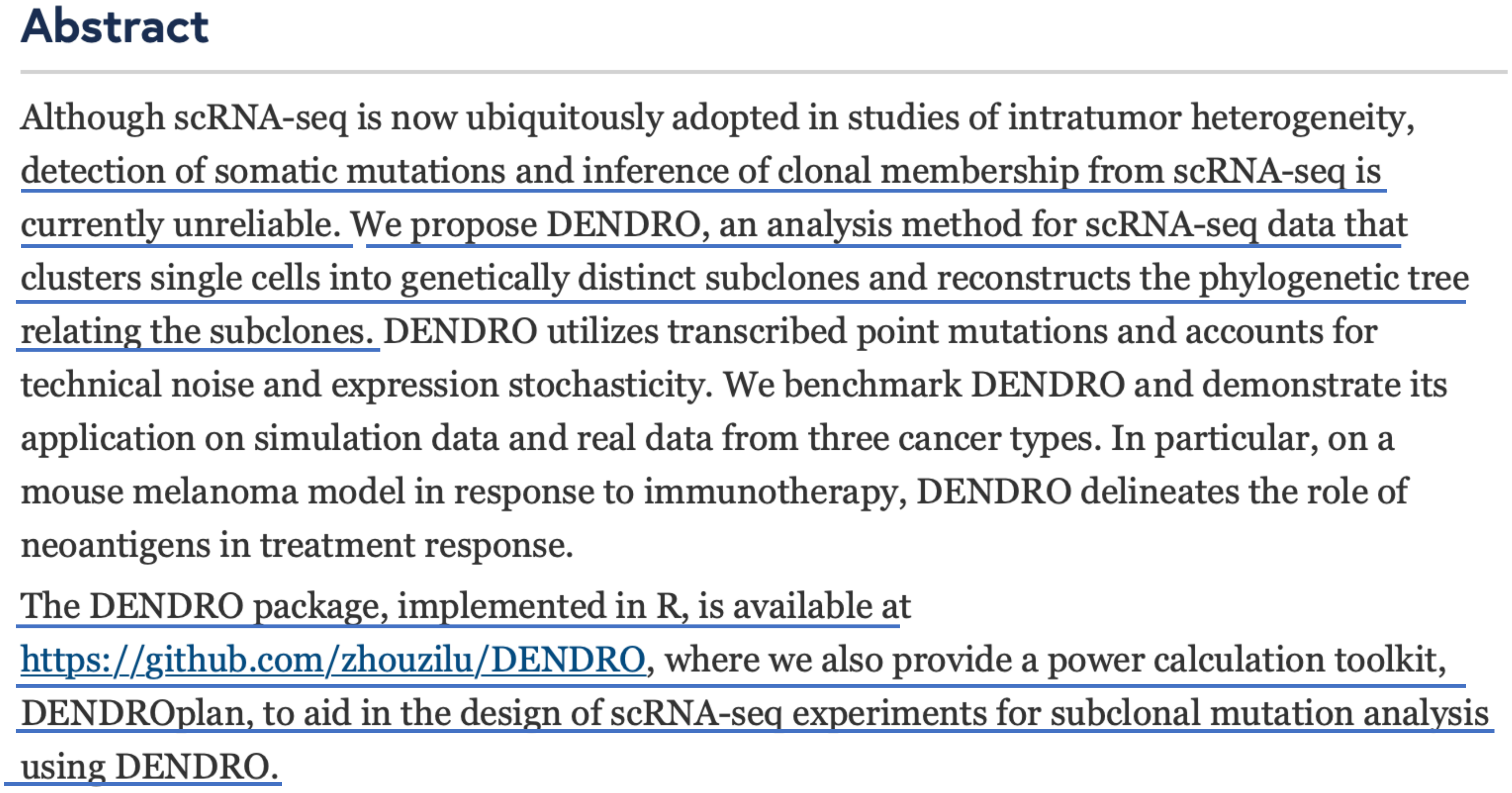
11k Accesses | **19** Citations | **22** Altmetric | Metrics

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1922-x

4. AI mining of cancer transcriptomes

4.2 Deciphering intra-tumor heterogeneity
4.2.1 Defining Cell Types and Clones

Cit. 70



Abstract

Although scRNA-seq is now ubiquitously adopted in studies of intratumor heterogeneity, detection of somatic mutations and inference of clonal membership from scRNA-seq is currently unreliable. We propose DENDRO, an analysis method for scRNA-seq data that clusters single cells into genetically distinct subclones and reconstructs the phylogenetic tree relating the subclones. DENDRO utilizes transcribed point mutations and accounts for technical noise and expression stochasticity. We benchmark DENDRO and demonstrate its application on simulation data and real data from three cancer types. In particular, on a mouse melanoma model in response to immunotherapy, DENDRO delineates the role of neoantigens in treatment response.

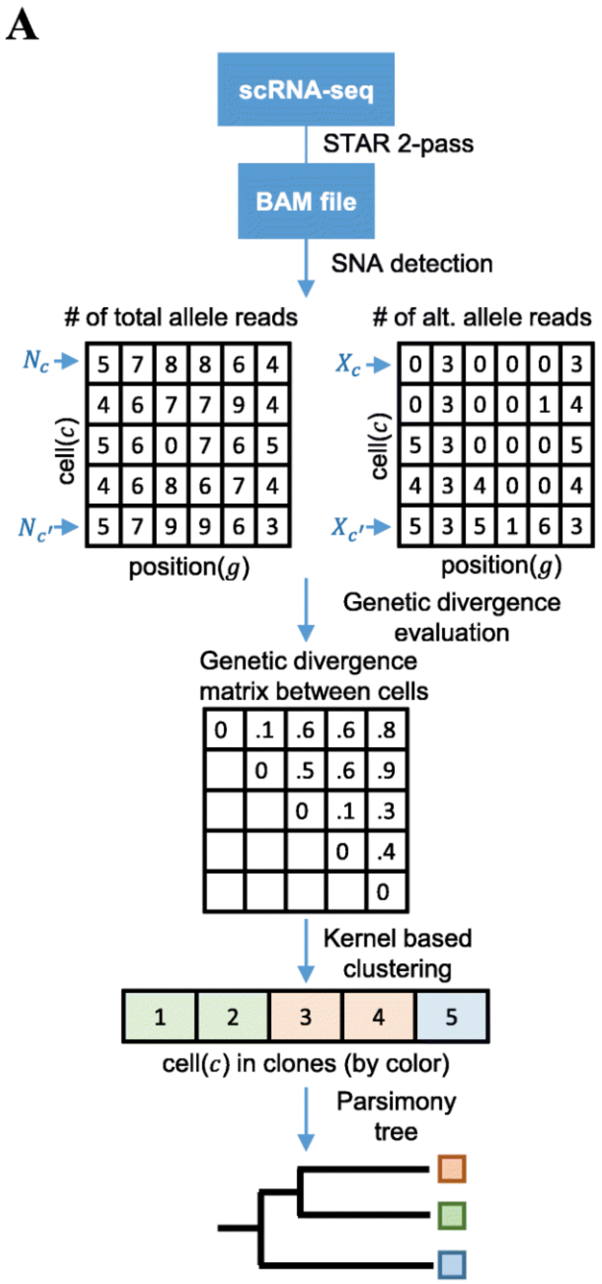
The DENDRO package, implemented in R, is available at https://github.com/zhouzilu/DENDRO, where we also provide a power calculation toolkit, DENDROplan, to aid in the design of scRNA-seq experiments for subclonal mutation analysis using DENDRO.

4. AI mining of cancer transcriptomes

4.2 Deciphering intra-tumor heterogeneity
4.2.1 Defining Cell Types and Clones

Cit. 70

Fig. 1



# of total allele reads

N_c	5	7	8	8	6	4
	4	6	7	7	9	4
	5	6	0	7	6	5
	4	6	8	6	7	4
$N_{c'}$	5	7	9	9	6	3

# of alt. allele reads

X_c	0	3	0	0	0	3
	0	3	0	0	1	4
	5	3	0	0	0	5
	4	3	4	0	0	4
$X_{c'}$	5	3	5	1	6	3

Genetic divergence matrix between cells

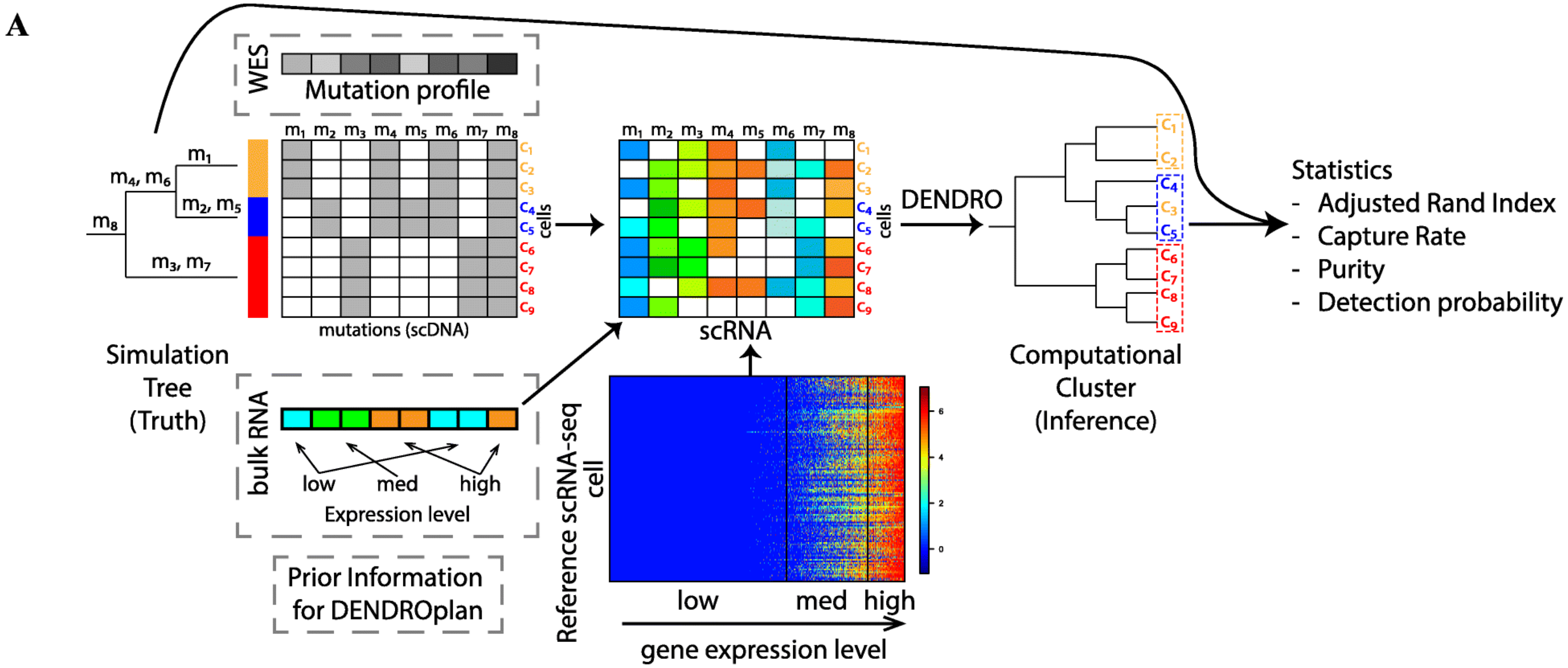
0	.1	.6	.6	.8
	0	.5	.6	.9
		0	.1	.3
			0	.4
				0

cell(c) in clones (by color)

1	2	3	4	5

DENDRO analysis pipeline and genetic divergence evaluation. **a** DENDRO analysis pipeline overview.

Fig. 2



DENDRO accuracy assessment. **a** The overall simulation analysis pipeline. Mutation matrix (cell-by-loci) is generated according to a simulated evolutionary tree, where the leaves are subclones and mutations can be placed on the branches. Matrices of alternative allele (X_{cg}) and total read counts (N_{cg}) are sampled from a scRNA-seq dataset with known transcriptomic allele-specific read counts. DENDRO cluster is further applied, and its performance is assessed by adjusted Rand index (global accuracy), capture rate (subclone-specific sensitivity), and purity (subclone-specific precision). See Additional file 2: Supplementary Materials for detailed definition. Gray dashed line indicates optional input for DENDROplan, where bulk DNA-seq and bulk RNA-seq can guide the tree simulation and read count sampling procedure. **b** Cluster accuracy via simulation studies. Various parameters show

4. AI mining of cancer transcriptomes

4.2 Deciphering intra-tumor heterogeneity
4.2.2 Assessment of TME

Cit. 22

Article | Open Access | Published: 10 December 2020

Single-cell RNA landscape of intratumoral heterogeneity and immunosuppressive microenvironment in advanced osteosarcoma

Yan Zhou, Dong Yang, Qingcheng Yang, Xiaobin Lv, Wentao Huang, Zhenhua Zhou, Yaling Wang, Zhichang Zhang, Ting Yuan, Xiaomin Ding, Lina Tang, Jianjun Zhang, Junyi Yin, Yujing Huang, Wenxi Yu, Yonggang Wang, Chenliang Zhou, Yang Su, Aina He, Yuanjue Sun, Zan Shen, Binzhi Qian, Wei Meng, Jia Fei, ... Haiyan Hu ✉ + Show authors

Nature Communications **11**, Article number: 6322 (2020) | Cite this article

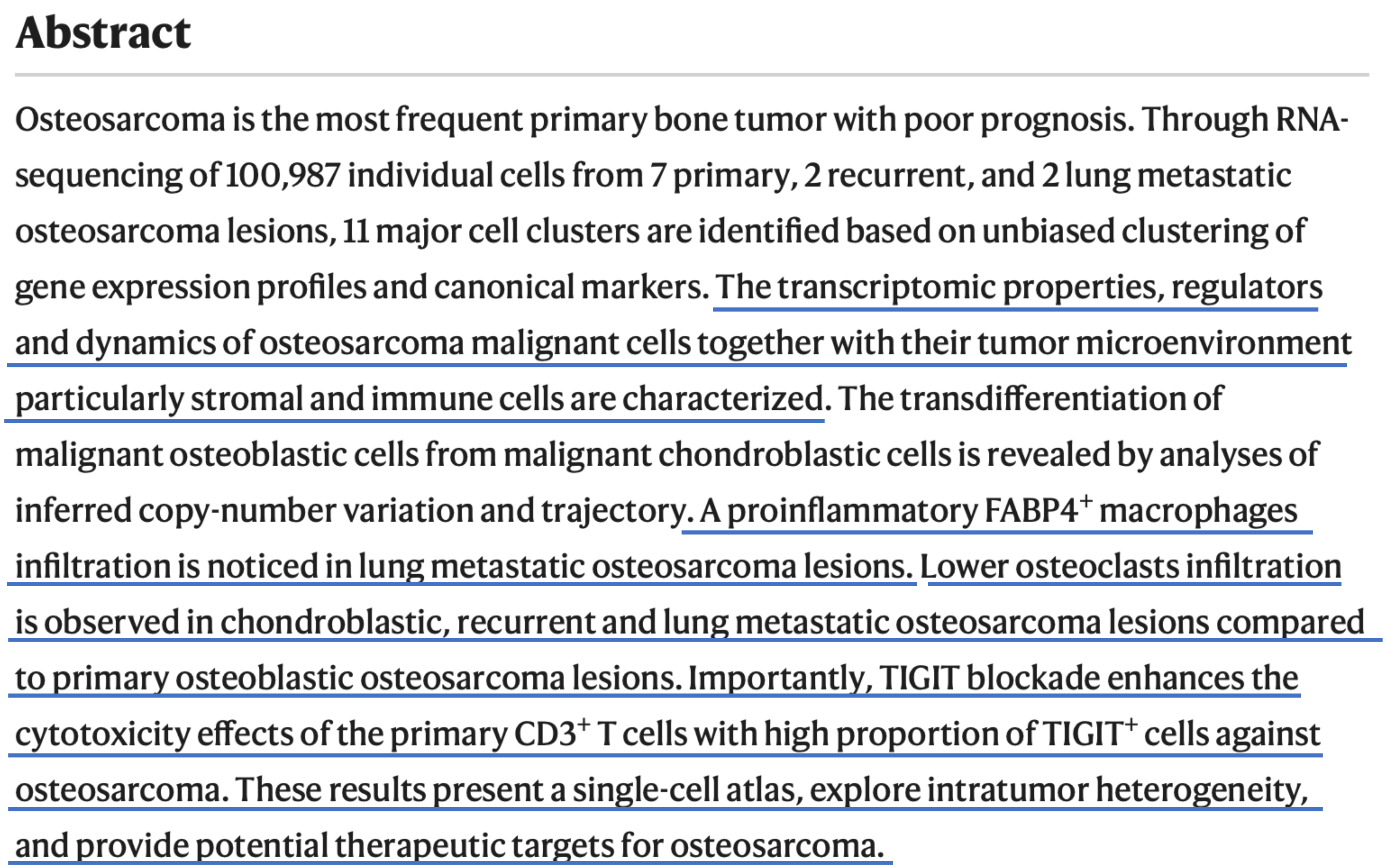
29k Accesses | **127** Citations | **11** Altmetric | Metrics

https://www.nature.com/articles/s41467-020-20059-6

4. AI mining of cancer transcriptomes

4.2 Deciphering intra-tumor heterogeneity
4.2.2 Assessment of TME

Cit. 22



Abstract

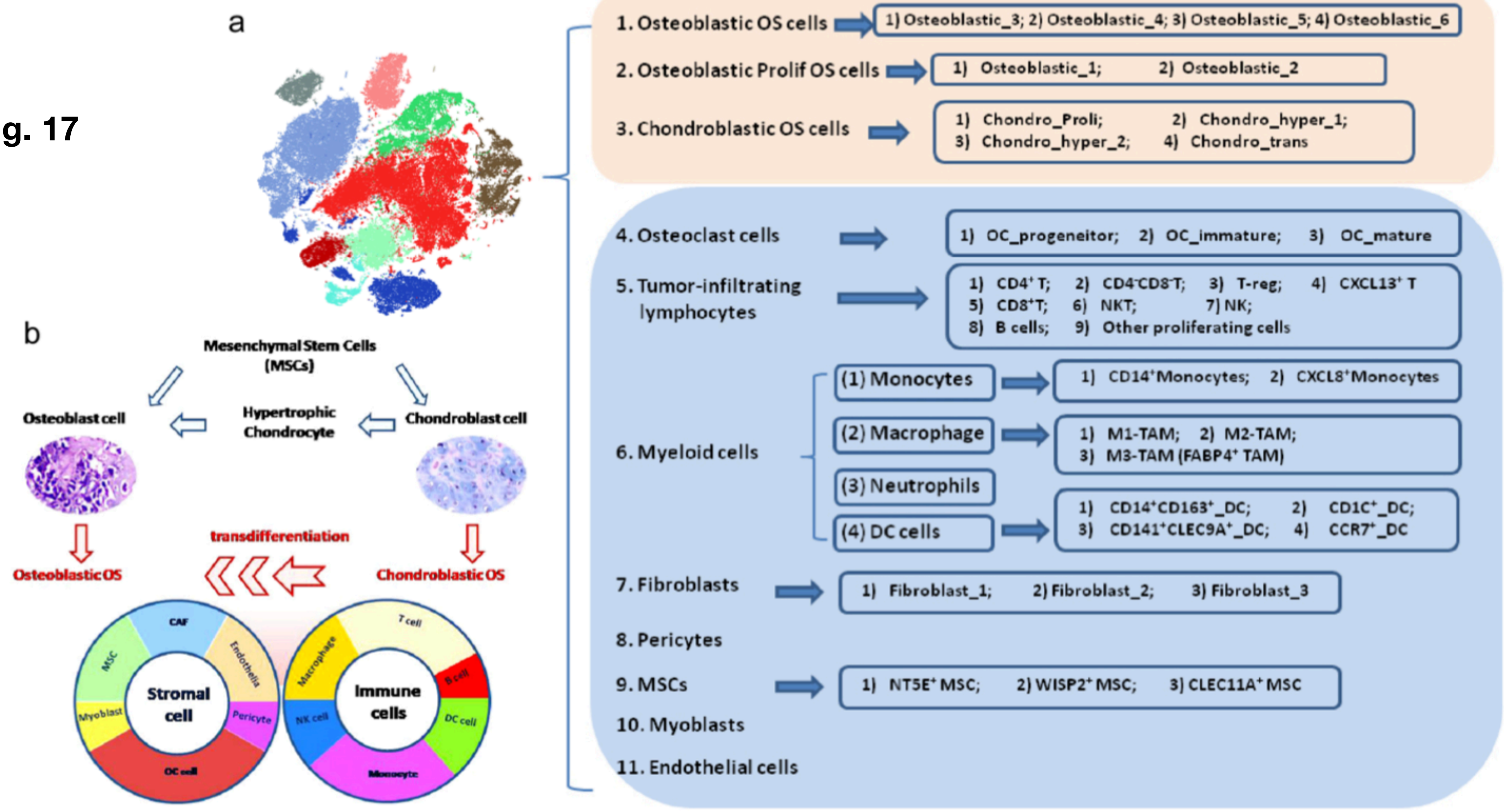
Osteosarcoma is the most frequent primary bone tumor with poor prognosis. Through RNA-sequencing of 100,987 individual cells from 7 primary, 2 recurrent, and 2 lung metastatic osteosarcoma lesions, 11 major cell clusters are identified based on unbiased clustering of gene expression profiles and canonical markers. The transcriptomic properties, regulators and dynamics of osteosarcoma malignant cells together with their tumor microenvironment particularly stromal and immune cells are characterized. The transdifferentiation of malignant osteoblastic cells from malignant chondroblastic cells is revealed by analyses of inferred copy-number variation and trajectory. A proinflammatory FABP4$^+$ macrophages infiltration is noticed in lung metastatic osteosarcoma lesions. Lower osteoclasts infiltration is observed in chondroblastic, recurrent and lung metastatic osteosarcoma lesions compared to primary osteoblastic osteosarcoma lesions. Importantly, TIGIT blockade enhances the cytotoxicity effects of the primary CD3$^+$ T cells with high proportion of TIGIT$^+$ cells against osteosarcoma. These results present a single-cell atlas, explore intratumor heterogeneity, and provide potential therapeutic targets for osteosarcoma.

4. AI mining of cancer transcriptomes

4.2 Deciphering intra-tumor heterogeneity
4.2.2 Assessment of TME

Cit. 22

Supplementary Fig. 17



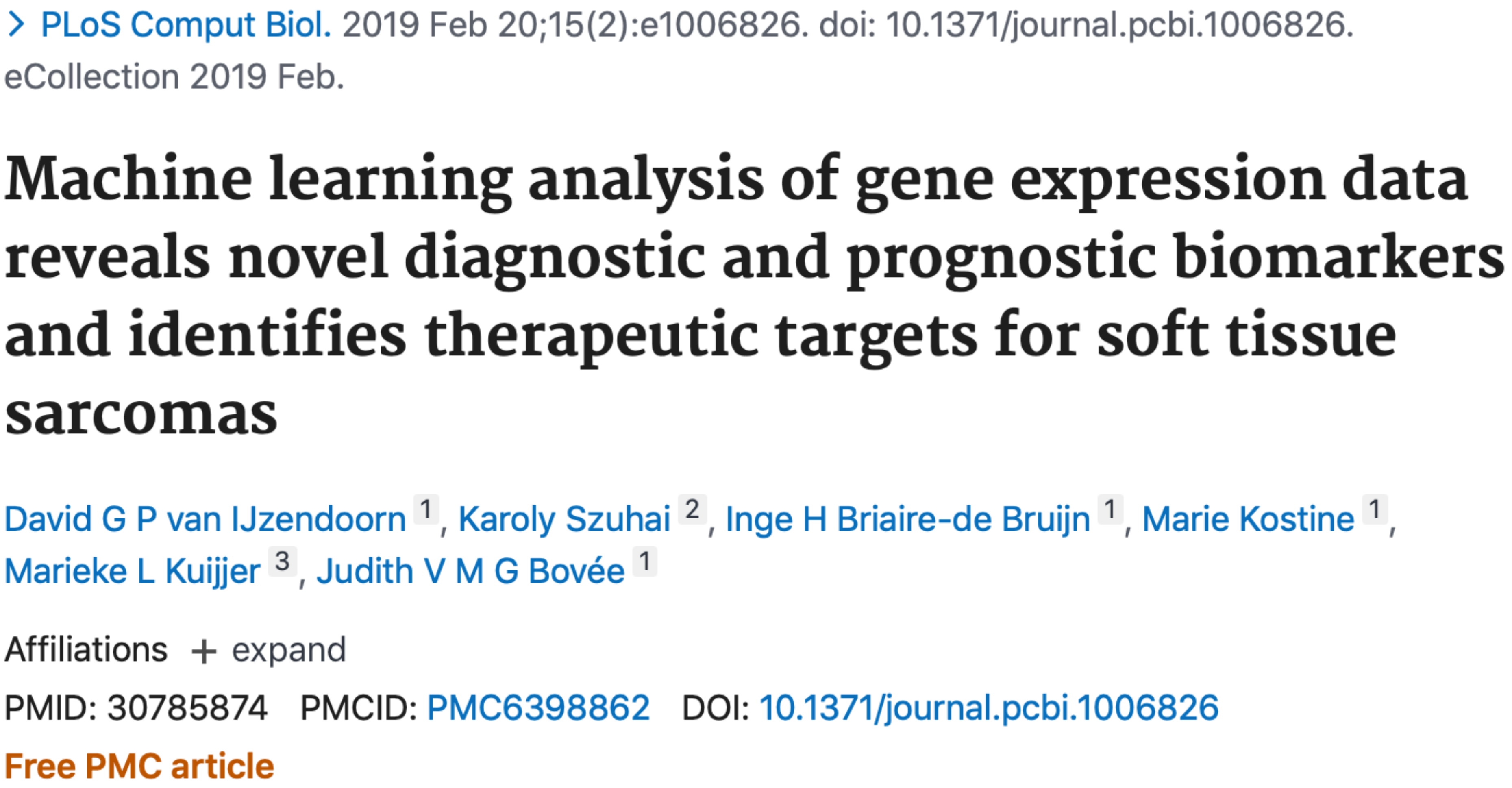
1. Osteoblastic OS cells → 1) Osteoblastic_3; 2) Osteoblastic_4; 3) Osteoblastic_5; 4) Osteoblastic_6
2. Osteoblastic Prolif OS cells → 1) Osteoblastic_1; 2) Osteoblastic_2
3. Chondroblastic OS cells → 1) Chondro_Proli; 2) Chondro_hyper_1; 3) Chondro_hyper_2; 4) Chondro_trans
4. Osteoclast cells → 1) OC_progeneitor; 2) OC_immature; 3) OC_mature
5. Tumor-infiltrating lymphocytes → 1) CD4$^+$ T; 2) CD4$^-$CD8$^-$T; 3) T-reg; 4) CXCL13$^+$ T 5) CD8$^+$T; 6) NKT; 7) NK; 8) B cells; 9) Other proliferating cells
6. Myeloid cells
   - (1) Monocytes → 1) CD14$^+$Monocytes; 2) CXCL8$^+$Monocytes
   - (2) Macrophage → 1) M1-TAM; 2) M2-TAM; 3) M3-TAM (FABP4$^+$ TAM)
   - (3) Neutrophils
   - (4) DC cells → 1) CD14$^+$CD163$^+$_DC; 2) CD1C$^+$_DC; 3) CD141$^+$CLEC9A$^+$_DC; 4) CCR7$^+$_DC
7. Fibroblasts → 1) Fibroblast_1; 2) Fibroblast_2; 3) Fibroblast_3
8. Pericytes
9. MSCs → 1) NT5E$^+$ MSC; 2) WISP2$^+$ MSC; 3) CLEC11A$^+$ MSC
10. Myoblasts
11. Endothelial cells

Supplementary Fig. 17. Overview of the identified cellular subclusters in scRNA-seq data of the OS lesions. a A summary of the cellular clusters and the subclusters of the 11 main cell types identified in OS lesions. **b** A schematic diagram displayed the malignant OS transdifferentiation cells and tumor microenvironment components.

4. AI mining of cancer transcriptomes

4.3 Biomarker Identification

Cit. 33



> PLoS Comput Biol. 2019 Feb 20;15(2):e1006826. doi: 10.1371/journal.pcbi.1006826. eCollection 2019 Feb.

Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas

David G P van IJzendoorn [1], Karoly Szuhai [2], Inge H Briaire-de Bruijn [1], Marie Kostine [1], Marieke L Kuijjer [3], Judith V M G Bovée [1]

Affiliations + expand

PMID: 30785874 PMCID: PMC6398862 DOI: 10.1371/journal.pcbi.1006826

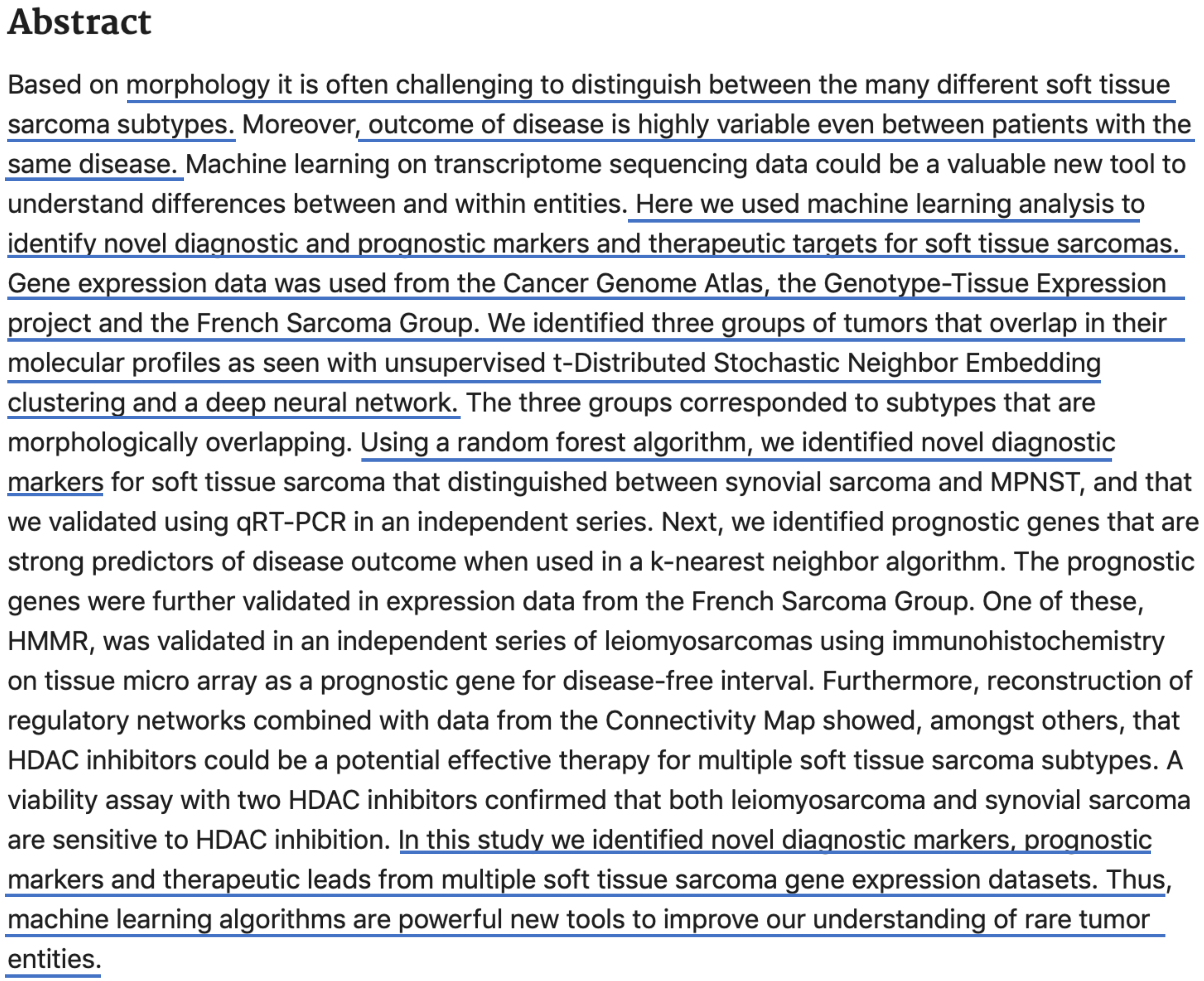
Free PMC article

https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006826

4. AI mining of cancer transcriptomes

4.3 Biomarker Identification

Cit. 33



Abstract

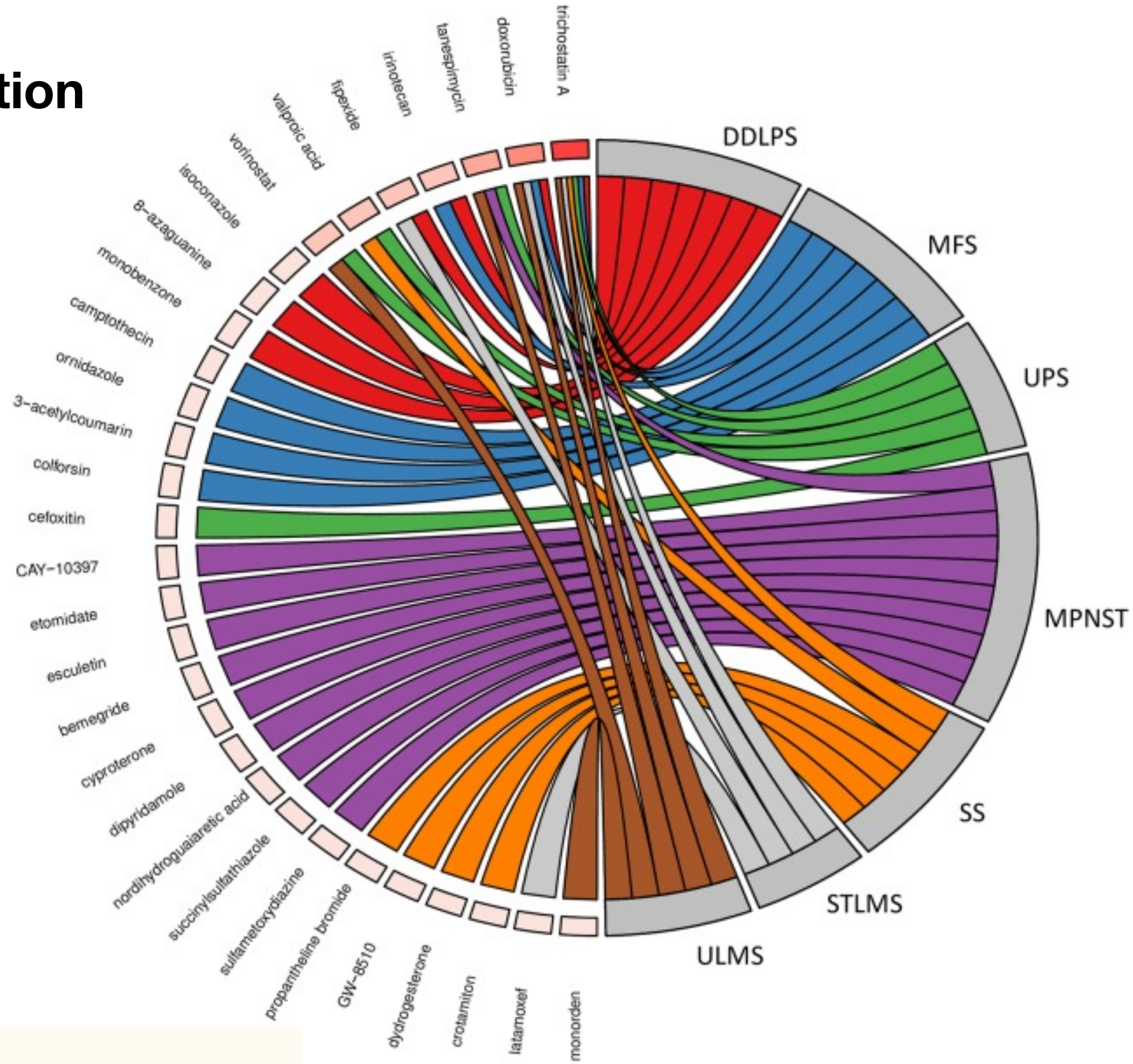
Based on morphology it is often challenging to distinguish between the many different soft tissue sarcoma subtypes. Moreover, outcome of disease is highly variable even between patients with the same disease. Machine learning on transcriptome sequencing data could be a valuable new tool to understand differences between and within entities. Here we used machine learning analysis to identify novel diagnostic and prognostic markers and therapeutic targets for soft tissue sarcomas. Gene expression data was used from the Cancer Genome Atlas, the Genotype-Tissue Expression project and the French Sarcoma Group. We identified three groups of tumors that overlap in their molecular profiles as seen with unsupervised t-Distributed Stochastic Neighbor Embedding clustering and a deep neural network. The three groups corresponded to subtypes that are morphologically overlapping. Using a random forest algorithm, we identified novel diagnostic markers for soft tissue sarcoma that distinguished between synovial sarcoma and MPNST, and that we validated using qRT-PCR in an independent series. Next, we identified prognostic genes that are strong predictors of disease outcome when used in a k-nearest neighbor algorithm. The prognostic genes were further validated in expression data from the French Sarcoma Group. One of these, HMMR, was validated in an independent series of leiomyosarcomas using immunohistochemistry on tissue micro array as a prognostic gene for disease-free interval. Furthermore, reconstruction of regulatory networks combined with data from the Connectivity Map showed, amongst others, that HDAC inhibitors could be a potential effective therapy for multiple soft tissue sarcoma subtypes. A viability assay with two HDAC inhibitors confirmed that both leiomyosarcoma and synovial sarcoma are sensitive to HDAC inhibition. In this study we identified novel diagnostic markers, prognostic markers and therapeutic leads from multiple soft tissue sarcoma gene expression datasets. Thus, machine learning algorithms are powerful new tools to improve our understanding of rare tumor entities.

4. AI mining of cancer transcriptomes

4.3 Biomarker Identification

Cit. 33

Fig. 04



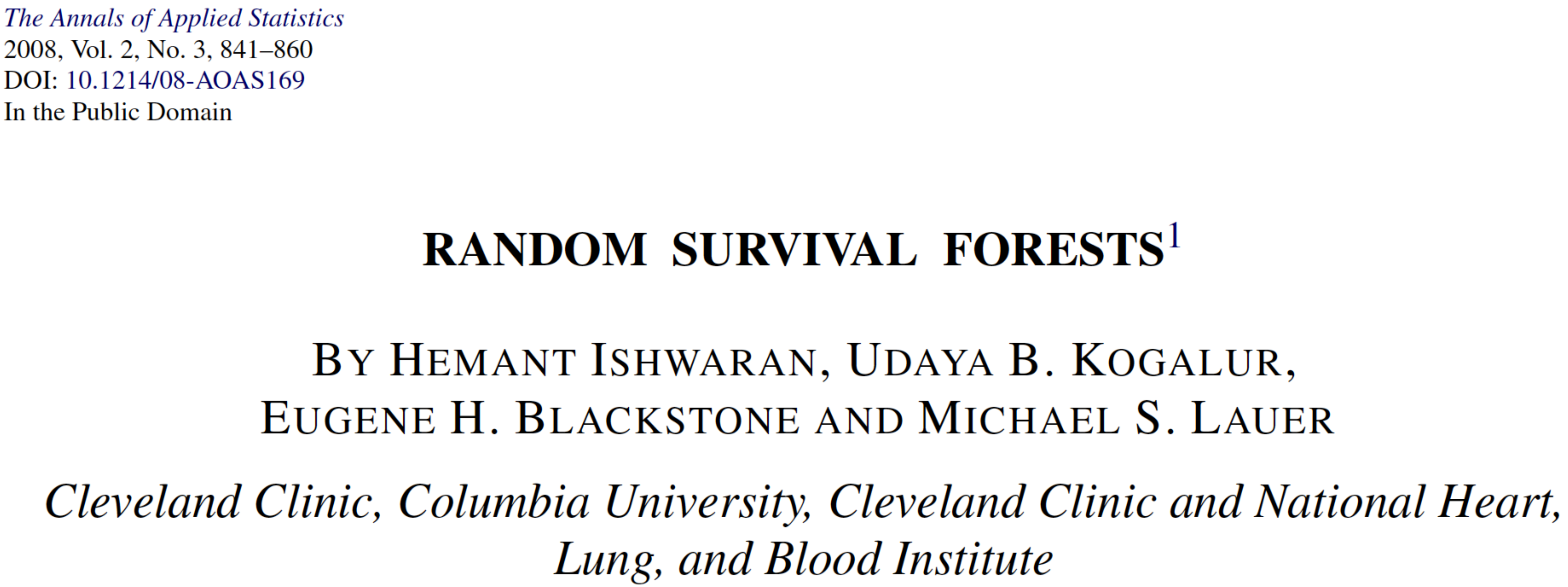
CMAP analysis to identify novel therapies.

(**a**) CMAP analysis identifies potential drugs based on the expression profile. The chord diagram shows links between the drugs and soft tissue sarcoma subtypes. Some compounds such as trichostatin A, doxorubicin and tanespimycin show connections with multiple soft tissue sarcoma subtypes, which is illustrated by the box color for each drug (darker red indicates more connections). (**b**) The dose response curves are shown for both trichostatin A (TSA) and quisinostat as tested

4. AI mining of cancer transcriptomes

4.4 Prediction of Patient Survival

Cit. 117



The Annals of Applied Statistics
2008, Vol. 2, No. 3, 841–860
DOI: 10.1214/08-AOAS169
In the Public Domain

RANDOM SURVIVAL FORESTS[1]

BY HEMANT ISHWARAN, UDAYA B. KOGALUR,
EUGENE H. BLACKSTONE AND MICHAEL S. LAUER

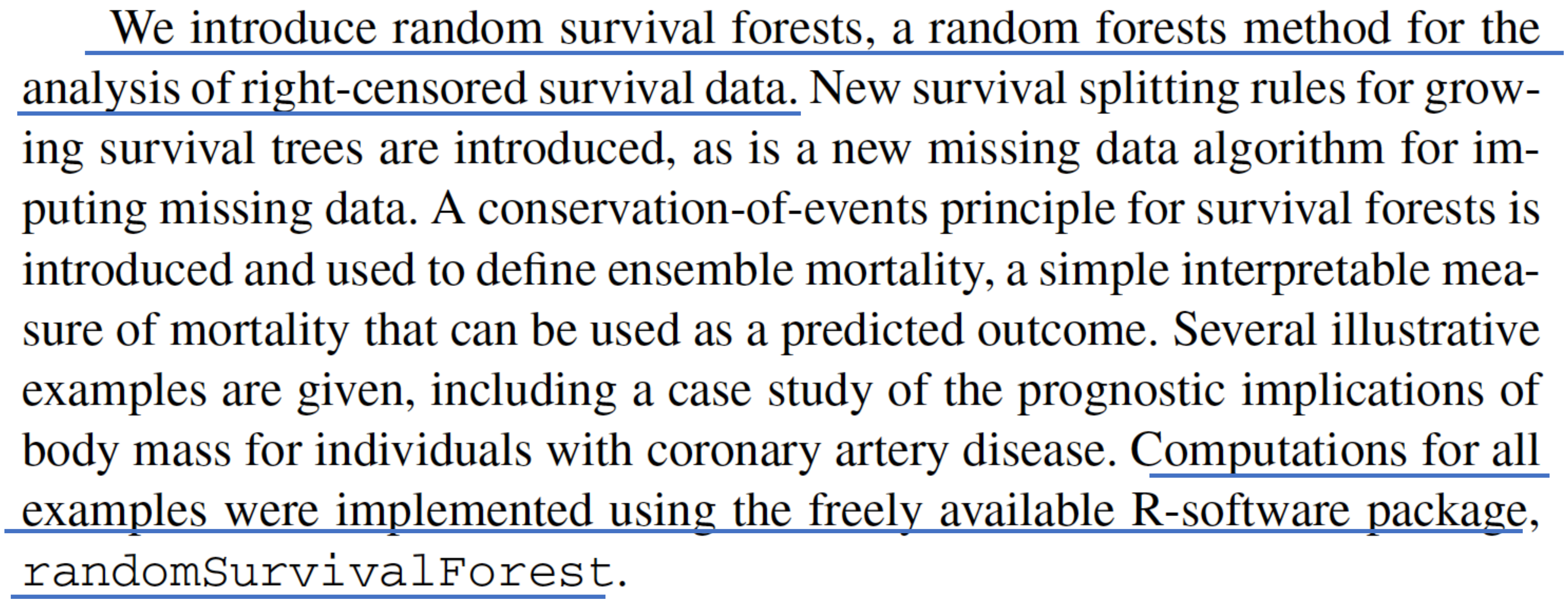
Cleveland Clinic, Columbia University, Cleveland Clinic and National Heart, Lung, and Blood Institute

https://projecteuclid.org/journals/annals-of-applied-statistics/volume-2/issue-3/Random-survival-forests/10.1214/08-AOAS169.full

4. AI mining of cancer transcriptomes

4.4 Prediction of Patient Survival

Cit. 117



We introduce random survival forests, a random forests method for the analysis of right-censored survival data. New survival splitting rules for growing survival trees are introduced, as is a new missing data algorithm for imputing missing data. A conservation-of-events principle for survival forests is introduced and used to define ensemble mortality, a simple interpretable measure of mortality that can be used as a predicted outcome. Several illustrative examples are given, including a case study of the prognostic implications of body mass for individuals with coronary artery disease. Computations for all examples were implemented using the freely available R-software package, `randomSurvivalForest`.

4. AI mining of cancer transcriptomes

4.4 Prediction of Patient Survival

Cit. 117

Fig. 01

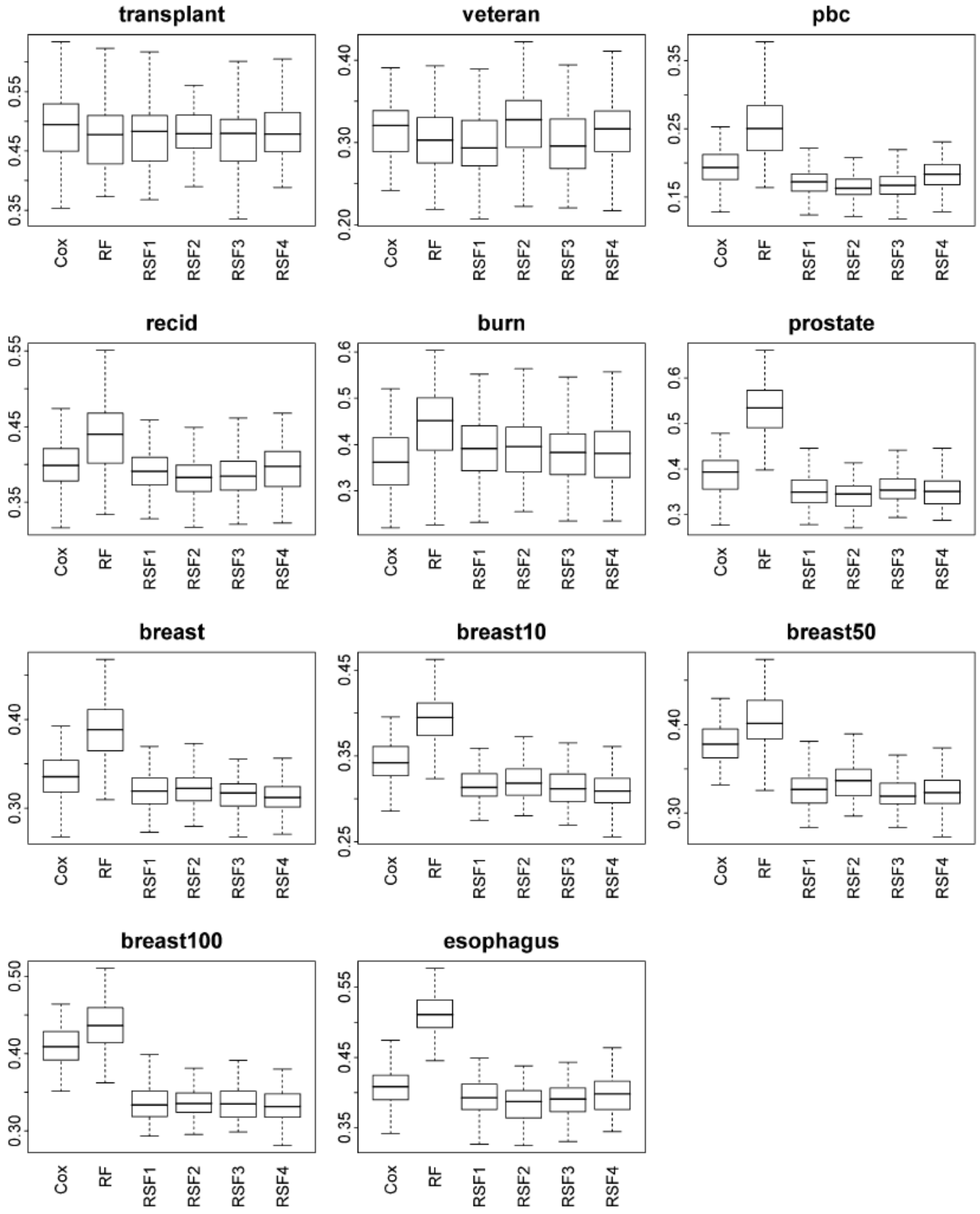
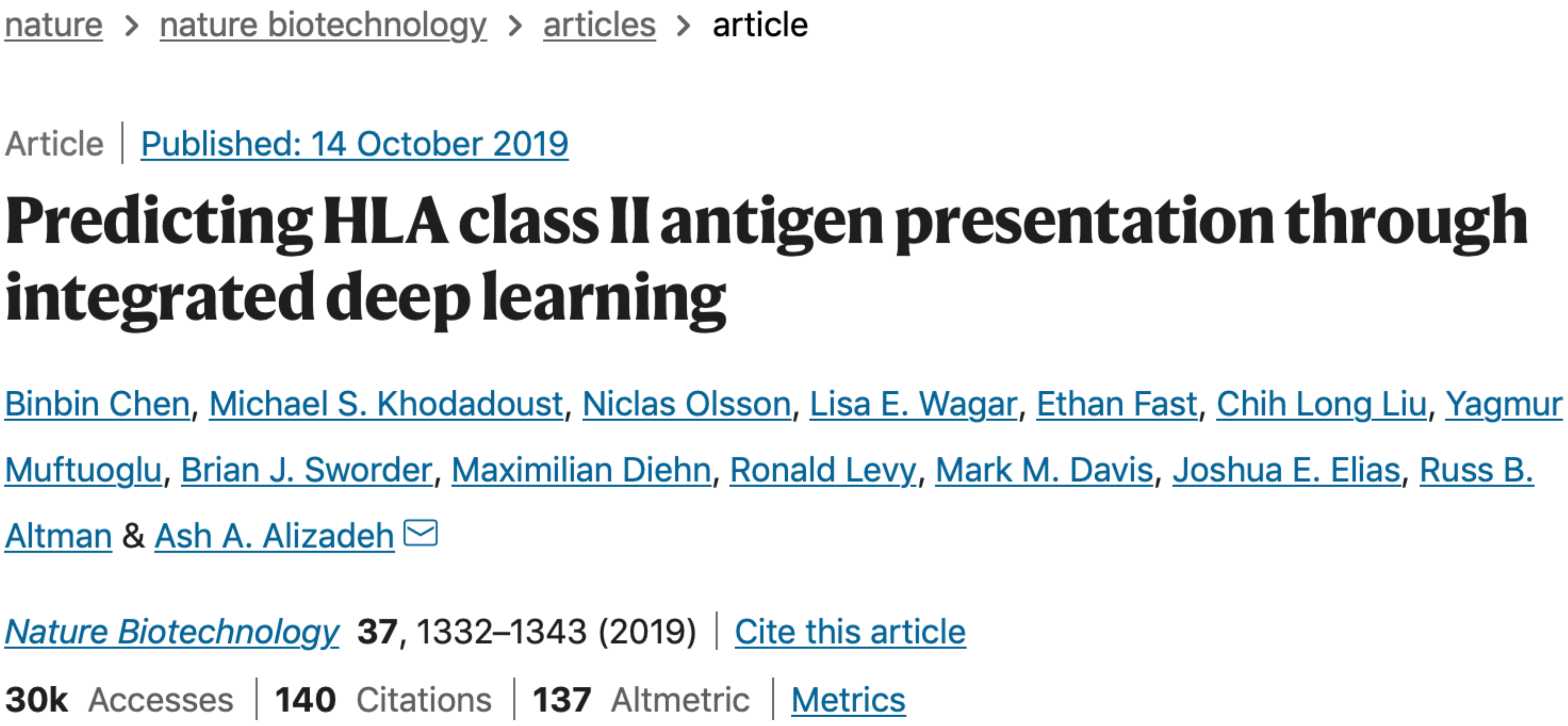


FIG. 1. *Boxplots of estimated prediction error (calculated using by C-index of Section* 5) *from* 100 *independent bootstrap replicates. Prediction error estimated on out-of-bag data. Dots in boxplots indicate mean values; horizontal lines are medians. Methods compared were as follows: Cox (Cox-regression); RF (RF for censored data [Hothorn et al.* (2006)]; *RSF*1 *through RSF*4 *(RSF using log-rank, conservation-of-events, log-rank score and random log-rank splitting). All forest analyses comprised* 1000 *trees. Datasets are indicated above each boxplot in bold.*

4. AI mining of cancer transcriptomes

4.5 Identification of Neoepitopes

Cit. 78



nature > nature biotechnology > articles > article

Article | Published: 14 October 2019

Predicting HLA class II antigen presentation through integrated deep learning

Binbin Chen, Michael S. Khodadoust, Niclas Olsson, Lisa E. Wagar, Ethan Fast, Chih Long Liu, Yagmur Muftuoglu, Brian J. Sworder, Maximilian Diehn, Ronald Levy, Mark M. Davis, Joshua E. Elias, Russ B. Altman & Ash A. Alizadeh ✉

Nature Biotechnology **37**, 1332–1343 (2019) | Cite this article

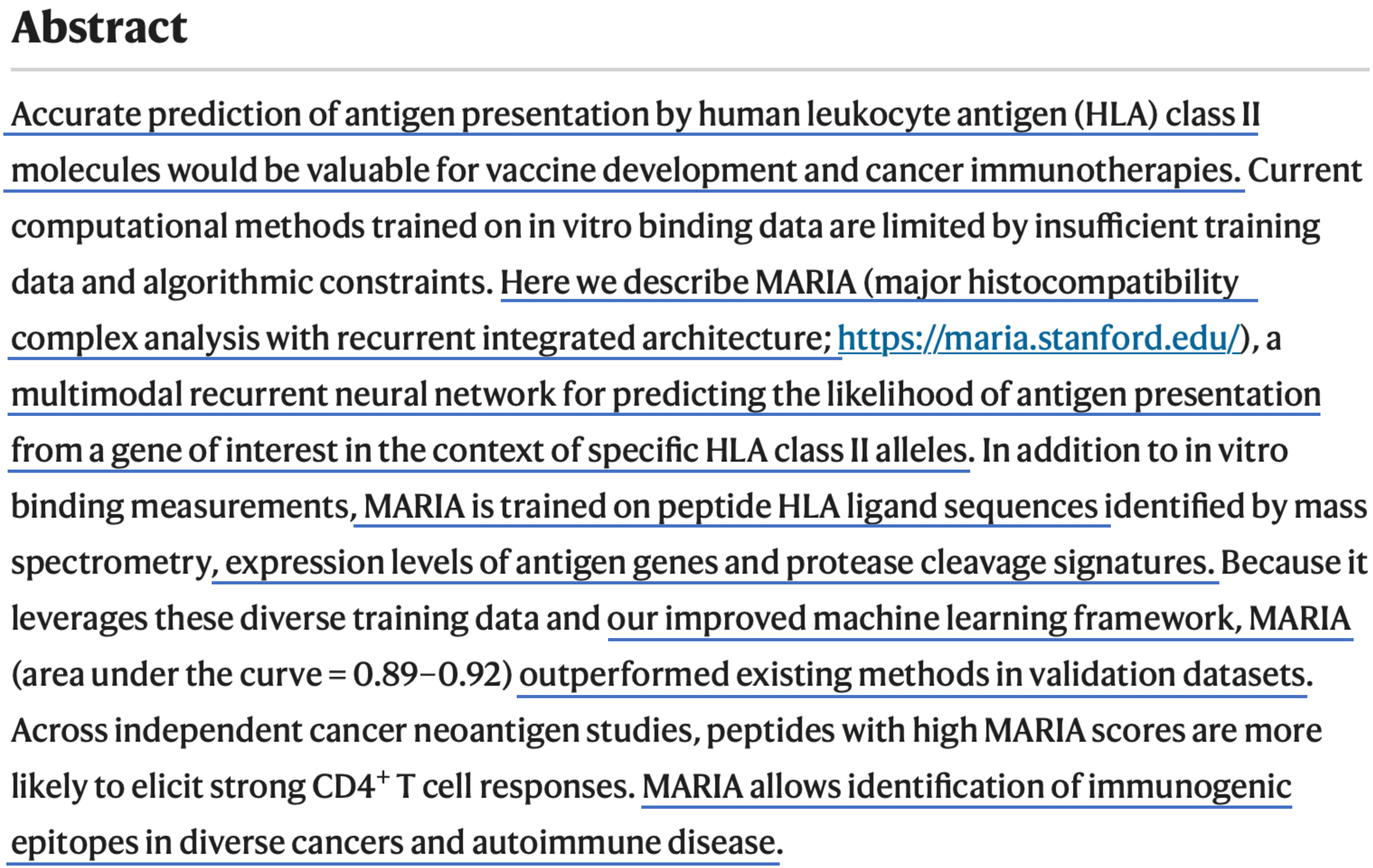
30k Accesses | **140** Citations | **137** Altmetric | Metrics

https://www.nature.com/articles/s41587-019-0280-2

4. AI mining of cancer transcriptomes

4.5 Identification of Neoepitopes

Cit. 78



Abstract

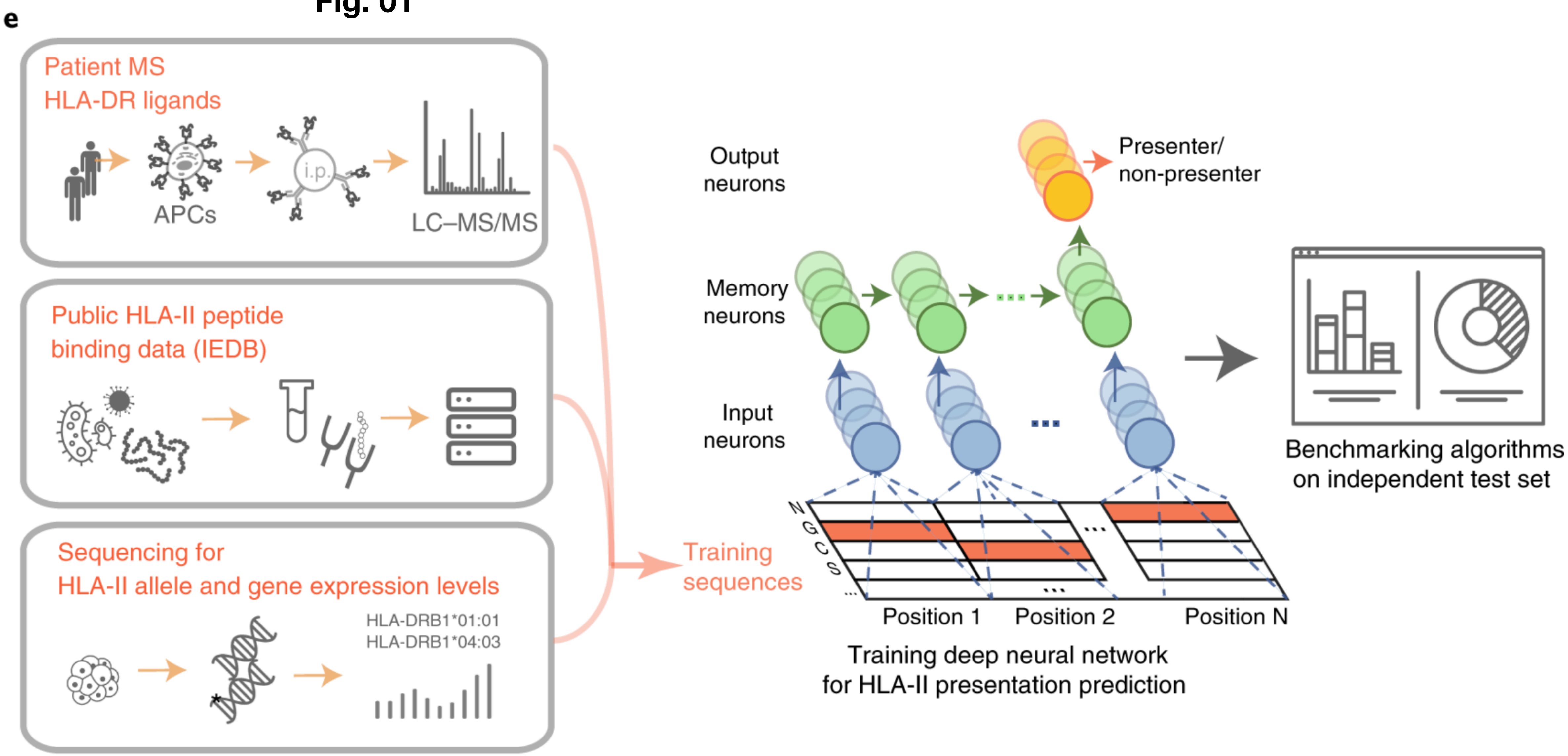
Accurate prediction of antigen presentation by human leukocyte antigen (HLA) class II molecules would be valuable for vaccine development and cancer immunotherapies. Current computational methods trained on in vitro binding data are limited by insufficient training data and algorithmic constraints. Here we describe MARIA (major histocompatibility complex analysis with recurrent integrated architecture; https://maria.stanford.edu/), a multimodal recurrent neural network for predicting the likelihood of antigen presentation from a gene of interest in the context of specific HLA class II alleles. In addition to in vitro binding measurements, MARIA is trained on peptide HLA ligand sequences identified by mass spectrometry, expression levels of antigen genes and protease cleavage signatures. Because it leverages these diverse training data and our improved machine learning framework, MARIA (area under the curve = 0.89–0.92) outperformed existing methods in validation datasets. Across independent cancer neoantigen studies, peptides with high MARIA scores are more likely to elicit strong CD4$^+$ T cell responses. MARIA allows identification of immunogenic epitopes in diverse cancers and autoimmune disease.

4. AI mining of cancer transcriptomes

4.5 Identification of Neoepitopes

Cit. 78

Fig. 01



proteins (*y* axis); remaining peptides are separately depicted in Supplementary Fig. 1. **e**, Training and evaluation scheme of MARIA, as a new machine learning framework for more accurate prediction of HLA-II ligands. Positive examples are HLA-II ligand peptide sequences directly identified by antigen presentation profiling of human cells and tissues by immunoprecipitation (i.p.) and MS, and negative examples are length-matched random human peptides (decoys). The model separately considers binding affinities estimated using in vitro binding data. Patient HLA-II allele or genotype and gene expression information are obtained from next-generation sequencing. A RNN integrates information and produces a predictor for HLA-II ligand presentation by minimizing training errors. Independent test sets determine the final performance of the model. See Supplementary Fig. 2 for detailed machine learning schemes.