



ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Αθηνών
— ΙΔΡΥΘΕΝ ΤΟ 1837 —

ΙΑΤΡΙΚΗ ΣΧΟΛΗ ΑΘΗΝΩΝ

ΕΡΓΑΣΤΗΡΙΟ ΦΥΣΙΟΛΟΓΙΑΣ

Διευθύντρια: Καθηγήτρια Μαυραγάνη Κλειώ

Μια υπολογιστική προσέγγιση στην κατεύθυνση μιας βελτιωμένης διαδικασίας
χαρακτηρισμού και περιγραφής του μεταγραφώματος ενός κυτταρικού
πληθυσμού, βασισμένη σε τεχνολογίες αλληλούχισης RNA σε επίπεδο
μοναδιαίου κυττάρου

A computational approach towards a more accurate characterization and
annotation of a cell population's transcriptome based on single-cell RNA
sequencing technologies

Τζαφέρης Χρήστος

Βιοπληροφορικός

Διδακτορική Διατριβή

Αθήνα 2024



ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Αθηνών
— IΔΡΥΘΕΝ ΤΟ 1837 —

ΙΑΤΡΙΚΗ ΣΧΟΛΗ ΑΘΗΝΩΝ

ΕΡΓΑΣΤΗΡΙΟ ΦΥΣΙΟΛΟΓΙΑΣ

Διευθύντρια: Καθηγήτρια Μαυραγάνη Κλειώ

**Μια υπολογιστική προσέγγιση στην κατεύθυνση μιας βελτιωμένης διαδικασίας
χαρακτηρισμού και περιγραφής του μεταγραφώματος ενός κυτταρικού
πληθυσμού, βασισμένη σε τεχνολογίες αλληλούχισης RNA σε επίπεδο
μοναδιαίου κυττάρου**

**A computational approach towards a more accurate characterization and
annotation of a cell population's transcriptome based on single-cell RNA
sequencing technologies**

Τζαφέρης Χρήστος

Βιοπληροφορικός

Διδακτορική Διατριβή

Αθήνα 2024

Στοιχεία ταυτότητας διδακτορικής διατριβής

α) Ημερομηνία αιτήσεως του υποψηφίου: 04/09/2019

β) Ημερομηνία ορισμού 3μελούς Συμβουλευτικής Επιτροπής: 08/10/2019

γ) Τα μέλη της 3μελούς Συμβουλευτικής Επιτροπής:

- Κόλλιας Γεώργιος, Καθηγητής, Ιατρική Σχολή, ΕΚΠΑ
- Χατζηγεωργίου Αντώνιος, Αναπληρωτής Καθηγητής, Ιατρική Σχολή, ΕΚΠΑ
- Παυλόπουλος Γεώργιος, Ερευνητής Α', Ε.ΚΕ.Β.Ε «Αλέξανδρος Φλέμιγκ»

δ) Ημερομηνία ορισμού του Θέματος: 03/01/2020

ε) Ημερομηνία καταθέσεως της διδακτορικής διατριβής: 19/06/2024

ζ) Ημερομηνία δημόσιας υποστήριξης της διδακτορικής διατριβής: 15/07/2024

Επταμελής εξεταστική επιτροπή

1. Σφηκάκης Πέτρος, Καθηγητής, Ιατρική Σχολή, ΕΚΠΑ
2. Κόλλιας Γεώργιος, Καθηγητής, Ιατρική Σχολή, ΕΚΠΑ
3. Χατζηγεωργίου Αντώνιος, Αναπληρωτής Καθηγητής, Ιατρική Σχολή, ΕΚΠΑ
4. Παληκαράς Κωνσταντίνος, Επίκουρος Καθηγητής, Ιατρική Σχολή, ΕΚΠΑ
5. Αθανασιάδης Εμμανουήλ, Επίκουρος Καθηγητής, Πανεπιστήμιο Δυτικής Αττικής
6. Παυλόπουλος Γεώργιος, Ερευνητής Α', Ε.ΚΕ.Β.Ε «Αλέξανδρος Φλέμιγκ»
7. Νικολάου Χριστόφορος, Ερευνητής Β', Ε.ΚΕ.Β.Ε «Αλέξανδρος Φλέμιγκ»

Επιβλέπων

Κόλλιας Γεώργιος, Καθηγητής, Ιατρική Σχολή, ΕΚΠΑ

Πρόεδρος Ιατρικής Σχολής

Αρκαδόπουλος Νικόλαος, Καθηγητής Ιατρικής Σχολής Αθηνών

ΟΡΚΟΣ ΤΟΥ ΙΠΠΟΚΡΑΤΗ

Ὅμνυμι Ἀπόλλωνα ἰητρὸν, καὶ Ἀσκληπιὸν, καὶ Ὑγίαν, καὶ Πανάκειαν, καὶ θεοὺς πάντας τε καὶ πάσας, ἴστορας ποιούμενος, ἐπιτελέα ποιήσῃν κατὰ δύναμιν καὶ κρίσιν ἐμήν ὄρκον τόνδε καὶ συγγραφὴν τήνδε. Ἠγήσασθαι μὲν τὸν διδάξαντά με τὴν τέχνην ταύτην ἴσα γενέτησιν ἐμοῖσι, καὶ βίου κοινώσασθαι, καὶ χρεῶν χρηρίζοντι μετάδοσιν ποιήσασθαι, καὶ γένος τὸ ἐξ αὐτέου ἀδελφοῖς ἴσον ἐπικρινέειν ἄρρεσι, καὶ διδάξειν τὴν τέχνην ταύτην, ἣν χρηρίζωσι μανθάνειν, ἄνευ μισθοῦ καὶ συγγραφῆς, παραγγελίης τε καὶ ἀκροήσιος καὶ τῆς λοιπῆς ἀπάσης μαθήσιος μετάδοσιν ποιήσασθαι υἱοῖσί τε ἐμοῖσι, καὶ τοῖσι τοῦ ἐμὲ διδάξαντος, καὶ μαθηταῖσι συγγεγραμμένοισί τε καὶ ὠρκισμένοις νόμῳ ἰητρικῷ, ἄλλῳ δὲ οὐδενί. Διαιτήμασί τε χρήσομαι ἐπ' ὠφελείῃ καμνόντων κατὰ δύναμιν καὶ κρίσιν ἐμήν, ἐπὶ δηλήσει δὲ καὶ ἀδικίῃ εἴρξειν. Οὐ δώσω δὲ οὐδὲ φάρμακον οὐδενὶ αἰτηθεὶς θανάσιμον, οὐδὲ ὑφηγήσομαι ξυμβουλίην τοιήνδε. Ὅμοίως δὲ οὐδὲ γυναικὶ πεσσὸν φθόριον δώσω. Ἀγνῶς δὲ καὶ ὁσίως διατηρήσω βίον τὸν ἐμὸν καὶ τέχνην τὴν ἐμήν. Οὐ τεμέω δὲ οὐδὲ μὴν λιθιῶντας, ἐκχωρήσω δὲ ἐργάτησιν ἀνδράσι πρήξιος τῆσδε. Ἐς οἰκίας δὲ ὀκόσας ἂν ἐσίω, ἐσελεύσομαι ἐπ' ὠφελείῃ καμνόντων, ἐκτὸς ἐὼν πάσης ἀδικίης ἐκουσίης καὶ φθορίας, τῆς τε ἄλλης καὶ ἀφροδισίων ἔργων ἐπὶ τε γυναικείων σωμαίων καὶ ἀνδρώων, ἐλευθέρων τε καὶ δούλων. Ἄ δ' ἂν ἐν θεραπείῃ ἢ ἴδω, ἢ ἀκούσω, ἢ καὶ ἄνευ θεραπιῆς κατὰ βίον ἀνθρώπων, ἃ μὴ χρή ποτε ἐκλαλέεσθαι ἕξω, σιγήσομαι, ἄρρητα ἠγεύμενος εἶναι τὰ τοιαῦτα. Ὅρκον μὲν οὖν μοι τόνδε ἐπιτελέα ποιέοντι, καὶ μὴ συγγέοντι, εἴη ἐπαύρασθαι καὶ βίου καὶ τέχνης δοξαζομένῳ παρὰ πᾶσιν ἀνθρώποις ἐς τὸν αἰεὶ χρόνον. παραβαίνοντι δὲ καὶ ἐπιορκοῦντι, τάναντία τουτέων.

CURRICULUM VITAE

Personal information

First name/ Surname: Christos Tzaferis
Date of birth: 09/07/1990
Nationality: Greek
Address: Kyprou 11, Petroupoli 13231, Athens
Telephone: +306973936727
E-mail: christzaferis@gmail.com
LinkedIn name: Christos Tzaferis

Professional experience

08/2022 – present

Bioinformatician at Single Cell Analysis Unit of Biomedical Sciences Research Center "Alexander Fleming" Institute - lab: Institute for Bio-innovation - Kollias lab
Responsibilities: Bioinformatics analysis of -Omics data, development of pipelines and workflows for automated analysis and visualization of bulkRNA-seq, scRNA-seq, scATAC-seq and spatial transcriptomics data.
Supervisor: George Kollias, Research Group Leader

11/2020 – 07/2022

Bioinformatician at pMedGR infrastructure, Medical School of Athens.
Sector: Personalized medicine
Responsibilities: Bioinformatics analysis of Big data from NGS platforms, development of pipelines for data analysis and visualization purposes.
Supervisor: Petros Sfikakis, Scientific Manager

02/2018 – 10/2020

Research assistant in bioinformatics at Biomedical Sciences Research Center "Alexander Fleming" Institute - lab: Institute for Bio-innovation - Kollias lab
Responsibilities: Bioinformatics analysis of next generation sequencing (NGS) data, analysis and visualization of genomics, transcriptomics and pharmacogenomics data.
Supervisor: George Kollias, Research Group Leader

09/2013 – 02/2014

Internship at Janssen Pharmaceutical Companies of Johnson & Johnson.
Sector: Digital Communications, Business Intelligence
Responsibilities: Supportive tasks in the information system of the company, participation in the implementation of a new web portal for health care professionals and web testing of the new corporate website.
Supervisor: Tsigris Ksenofon, Digital Communications Manager

Education

10/2019 – present

PhD candidate, Department of Physiology, Medical School of Athens

Research topic: “A computational approach towards a more accurate characterization and annotation of a cell population’s transcriptome based on single-cell RNA sequencing technologies.”

PhD Supervisor: Prof. George Kollias

10/2014 – 02/2017

Master of Bioinformatics, Biology department, University of Athens

Master thesis: “Research on cellular communication and immunity”

09/2008 – 02/2014

Bachelor of Computer Science, Department of Informatics, Athens University of Economic and Business

Specialization: Computer Systems and Networks, Information Systems and Security

Publications and awards

- Papadopoulou D, Mavrikaki V, Charalampous F, Tzaferis C, Samiotaki M, Papavasileiou KD, Afantitis A, Karagianni N, Denis MC, Sanchez J, Lane JR, Faidon Brotzakis Z, Skretas G, Georgiadis D, Matralis AN, Kollias G. *Discovery of the First-in-Class Inhibitors of Hypoxia Up-Regulated Protein 1 (HYOU1) Suppressing Pathogenic Fibroblast Activation*. **Angew Chem Int Ed Engl.** (2024) 63(14):e202319157. doi: <https://doi.org/10.1002/anie.202319157>
- Tzaferis C, Karatzas E, Baltoumas FA, Pavlopoulos GA, Kollias G, Konstantopoulos D. *SCALA: A complete solution for multimodal analysis of single-cell Next Generation Sequencing data*. **Computational and Structural Biotechnology Journal** (2023) 21:5382-5393. <https://doi.org/10.1016/j.csbj.2023.10.032>
- Papadopoulou D, Roumelioti F, Tzaferis C, Chouvardas P, Pedersen AK, Charalampous F, Christodoulou-Vafeiadou E, Ntari L, Karagianni N, Denis MC, Olsen JV, Matralis AN, Kollias G. *Repurposing the antipsychotic drug amisulpride for targeting synovial fibroblast activation in arthritis*. **JCI Insight.** (2023) 8(9):e165024. doi: <https://doi.org/10.1172/jci.insight.165024>
- Armaka M, Konstantopoulos D, Tzaferis C, Lavigne MD, Sakkou M, Liakos A, Sfrikakis PP, Dimopoulos MA, Fousteri M, Kollias G. *Single-cell multimodal analysis identifies common regulatory programs in synovial fibroblasts of rheumatoid arthritis patients and modeled TNF-driven arthritis*. **Genome Medicine** (2022) 14(1):78. <https://doi.org/10.1186/s13073-022-01081-3>
- Kerdidani D, Aerakis E, Verrou KM, Angelidis I, Douka K, Maniou MA, Stamoulis P, Goudevenou K, Prados A, Tzaferis C, Ntafis V, Vamvakaris I, Kaniaris E, Vachlas K, Sepsas E, Koutsopoulos A, Potaris K, Tsoumakidou M. *Lung tumor MHCII immunity depends on in situ antigen presentation by fibroblasts*. **Journal of Experimental Medicine** (2022) 219(2):e20210815. <https://doi.org/10.1084/jem.20210815>
- Melissari MT, Henriques A, Tzaferis C, Prados A, Sarris ME, Chalkidi N, Mavroeidi D, Chouvardas P, Grammenoudi S, Kollias G, Koliaraki V. *Col6a1+/CD201+ mesenchymal cells regulate intestinal*

morphogenesis and homeostasis. Cellular and Molecular Life Sciences (2021) D79(1):1.
<https://doi.org/10.1007/s00018-021-04071-7>

- Koliaraki V, Chalkidi N, Henriques A, Tzaferis C, Polykratis A, Waisman A, Muller W, Hackam DJ, Pasparakis M, Kollias G. *Innate Sensing through Mesenchymal TLR4/MyD88 Signals Promotes Spontaneous Intestinal Tumorigenesis. Cell Reports (2019) 26(3):536-545.e4.*
<https://doi.org/10.1016/j.celrep.2018.12.072>

Best poster presentation award certificate in the 14th Conference of HSCBB19 for the study “A computational approach towards a more accurate characterization and annotation of a cell population’s transcriptome based on single-cell RNA sequencing technologies.”

Conferences

- 16th Conference of the Hellenic Society for Computational Biology and Bioinformatics (HSCBB22, 2022) [Oral presentation]
- 1st International Conference on Mesenchymal Cells in Health & Disease (2022) [Poster presentation]
- 14th Conference of the Hellenic Society for Computational Biology and Bioinformatics (HSCBB19, 2019) [Poster presentation]
- Mesenchymal cells in inflammation, immunity and cancer (EMBO Workshop, 2019)
- 17th European Conference on Computational Biology (ECCB2018, 2018)
- Genome Informatics, Precision Medicine & Clinical Omics in a world of Data Economies (H.bioinfo, 2018)
- 11th Conference of the Hellenic Society for Computational Biology and Bioinformatics (HSCBB16, 2016)
- 10th Conference of the Hellenic Society for Computational Biology and Bioinformatics (HSCBB15, 2015)

Core skills

Bioinformatics: Analysis of NGS data, Biological databases, Modeling of complex biological systems, Simulations of biological functions, Bioinformatics algorithms, Biological networks, Visualization of biological data

Programming languages: Java, R, R/Shiny, Python, BASH, C#, C++, C, Perl, JDBC, JSP, Java ME, VHDL, MIPS Assembly Language

Web technologies: XHTML, CSS, JavaScript

Databases: Microsoft SQL, MySQL, Transact SQL

Operating Systems programming (PCs and “smartphones”): Microsoft Windows, UNIX/LINUX, Android, Windows Phone, BlackBerry, Windows Mobile

Modeling tools and languages: Unified Modeling Language (UML), Entity–Relationship model (ER), Agent Based Modeling (ABM)

Languages

Greek: Native

English: Full professional proficiency

French: Elementary level

Acknowledgments

In this section, I would like to take the opportunity to express my gratitude to all those who helped me successfully complete my PhD thesis.

First and foremost, I would like to thank Professor George Kollias for giving me the opportunity to work in his lab, for trusting me with this fascinating project, and for his continuous support and guidance throughout my PhD. I also extend my sincere thanks to Associate Professors Georgios Pavlopoulos and Antonis Hatzigeorgiou for their invaluable contributions during the supervision of my thesis. Additionally, I am grateful to P. Sfikakis, K. Palikaras, E. Athanasiadis and C. Nikolaou who honored me with their presence on the 7-member committee for my PhD thesis.

I am deeply appreciative of all my colleagues who closely collaborated with me on the two projects related to my PhD thesis, including M. Fousteri, M. Armaka, M. Sakkou, M. Lavigne, T. Liakos, D. Konstantopoulos, V. Karatzas, and F. Baltoumas.

Special thanks go to Ana, Alex, Dimitra, Dora, Elena, Erifyli, Filippos, Fani, Katerina, Kostas, Lida, Lydia, Maria, Niki, and Vaso. Over the years, we shared the same office and created many great memories both inside and outside of working hours. They all created a wonderful working environment for me.

I would also like to thank Dr. V. Koliaraki, P. Moulos, P. Chouvardas, M. Rezcko, A. Dimopoulos, C. Nikolaou, M. Denis, N. Karagianni, E. Christodoulou, M. Tsoumakidou, A. Matralis, and G. Sofianatos for the stimulating scientific conversations and our collaboration on various projects.

I am grateful to my colleagues from other labs, including Dimitris, Dimitra, Manos, Arsenios, Ilias, Iliana, Apostolis, Elie, Kleio, and Athanasia, for our scientific discussions and interactions at the Fleming premises.

Finally, I would like to thank my family—my parents Loukas and Popi, and my sister Giota—for their unwavering love and support over the years, standing by my side through both the joyful and challenging times.

Σας ευχαριστώ για όλες τις αναμνήσεις που δημιουργήσαμε!

Χρήστος

Table of Contents

Table of Contents

1	Introduction	12
1.1	Chronic Inflammatory Diseases – Rheumatoid Arthritis	12
1.2	The role of cytokines in Rheumatoid Arthritis	14
1.3	The role of Tumor Necrosis Factor in Chronic Inflammatory Diseases	17
1.4	The use of animal models in Rheumatoid Arthritis.....	19
1.5	The function of fibroblast cells in homeostasis and disease	25
1.6	Single cell sequencing technology	29
1.7	Single cell application in biological systems	39
1.8	Computational methodologies for single cell data analysis	42
2	Material and methods.....	47
2.1	Implementation of a web-based application for SC data analysis.....	47
2.2	Data input.....	47
2.3	Workflow description	48
2.4	Quality control	48
2.5	Normalization and scaling of the data	50
2.6	Detection of highly variable genes.....	51
2.7	Principal Component Analysis	52
2.8	Latent Semantic Indexing.....	53
2.9	Clustering.....	53
2.10	Non-linear dimensionality reduction methods	54
2.11	Identification of marker genes	55
2.12	Inspection of features	56
2.13	Doublet detection.....	56
2.14	Cell cycle phase analysis	57
2.15	Functional/Motif enrichment analysis.....	58
2.16	Automated annotation of clusters	58
2.17	Multimodal integration analysis	59
2.18	Trajectory analysis	60
2.19	Cell-cell communication analysis	61
2.20	Gene regulatory network reconstruction.....	62
2.21	Visualization of epigenome signal tracks	63

2.22	Utility functions and code history	63
3	Results.....	67
3.1	Analysis of synovial fibroblasts in <i>hTNFtg</i> arthritis mouse model.....	67
3.2	Analysis using SCALA's scRNA-seq pipeline	67
3.3	Analysis using SCALA's scATAC-seq pipeline	71
3.4	Subclustering of Lining fibroblasts	73
3.5	Comparison of <i>hTNFtg</i> and STIA single cell RNA-seq data	73
3.6	Cross species integration of mouse and human patients single cell data	76
3.7	Analysis with alternative workflows	79
3.8	Benchmarking with other similar tools.....	82
3.9	Performance and scalability of the application.....	83
4	Discussion	85
5	Conclusions.....	87
6	Summary.....	89
7	Περίληψη.....	91
8	Acronyms	93
9	Παράρτημα	97
9.1	Προϋποθέσεις απόκτησης διδακτορικού	97
9.2	Δημοσιεύσεις	98
10	References	107

1 Introduction

1.1 Chronic Inflammatory Diseases – Rheumatoid Arthritis

Study of immune system, inflammatory processes as well as mechanisms leading to chronic inflammation, has contributed to the recognition of Chronic Inflammatory Diseases as one of the most significant causes of death today (Furman et al. 2019). Rheumatoid arthritis (RA), Crohn's disease (CD), Inflammatory Bowel Disease (IBD), psoriasis, and Systemic Lupus Erythematosus (SLE) are examples of such diseases. Although inflammation is a normal process that can protect the human organism from different types of hazards e.g., pathogens or toxins, it can also result in serious disorders when the resolving phase cannot be reached. Chronic inflammation is linked to aberrant cytokine and chemokine production, which in turn induce infiltration of immune cells and activation of fibroblasts, leading to temporary or permanent tissue damage (Hayden and Ghosh 2014).

RA is mainly characterized by inflammation of synovial membrane and pannus formation (the formation of an invasive synovial tissue), which can lead to cartilage destruction and bone erosion (Fig. 1) (Sudoł-Szopińska et al. 2012).

RA is estimated to affect on average 0.5 to 1.0 % of adult population in Western world, however geographic variability in RA incidence has been reported (Tobón, Youinou, and Saraux 2010; García-Alonso, Pérez-Naranjo, and Fernández-Caballero 2014). More precisely, women are more susceptible than men in developing this disease. That difference could be explained partially by the effect of estrogens in the regulation of immune system, although the overall influence of hormones in disease onset and progression remains a controversial topic (Ngo, Steyn, and McCombe 2014). Patients with RA are more likely to suffer also from cardiovascular diseases, diabetes mellitus and hyperlipidemia. Factors like abnormal immunity and unresolved chronic inflammation, could explain the increased risk of RA patients to develop heart disease compared to the general population (Crowson et al. 2013).

Due to the complex nature of RA, and its heterogeneity among individuals, it is difficult to conclude on a single cause of the disease. However, over the past years several risk factors have been associated with disease initiation and progression. More specifically, genetic susceptibility, epigenetic modifications, as well as smoking, diet, and microbiota are all factors that can trigger the emergence of disease (Smolen et al. 2018).

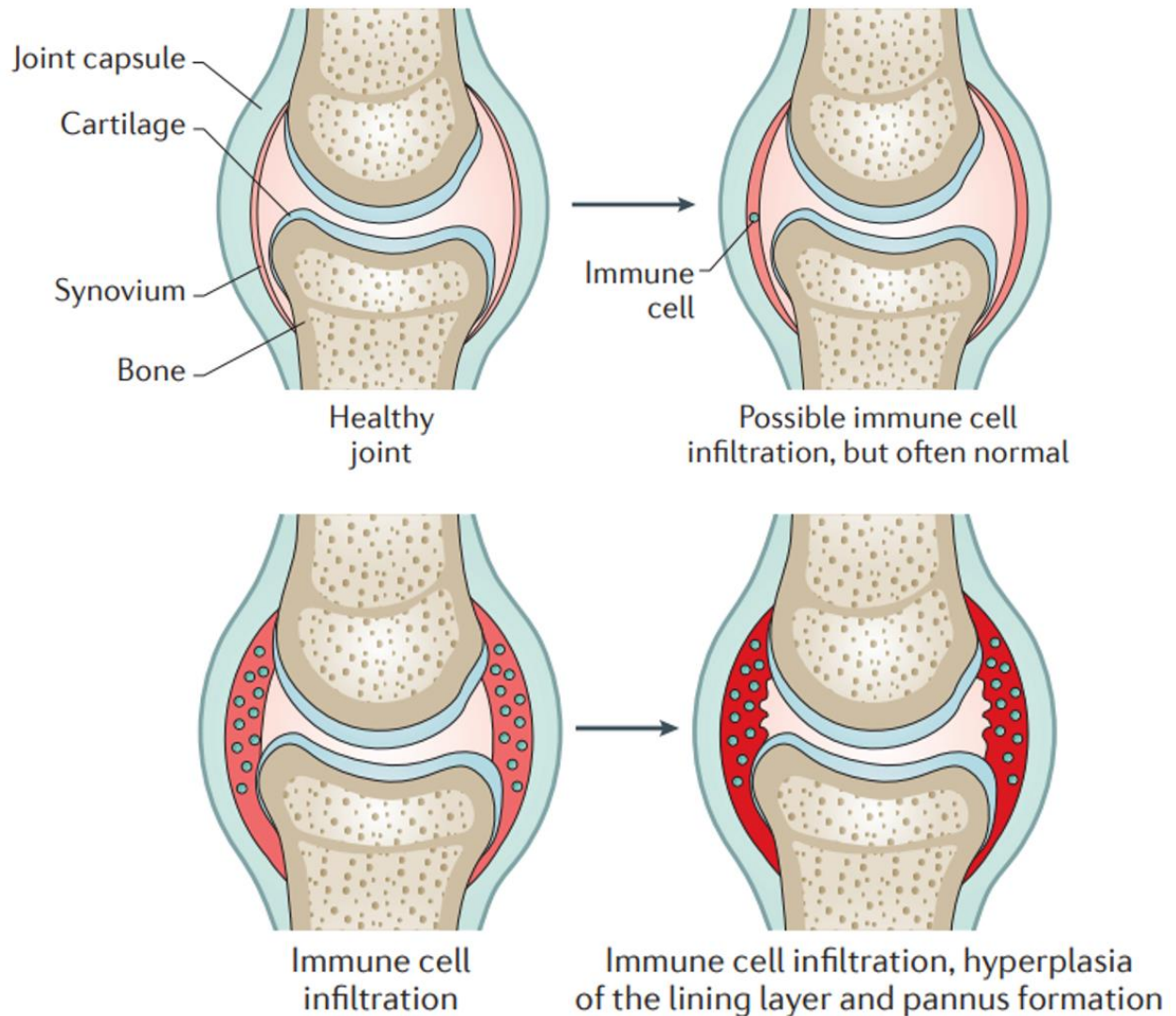


Figure 1. In this figure a knee joint is depicted before and after RA onset. In the top left panel healthy condition is shown. In the top right panel immune infiltration has been initiated in the lining layer. In the bottom panels synovial hyperplasia and pannus formation are evident, leading eventually to cartilage and bone damaging. (Adapted from Smolen et al., 2018)

As mentioned before, RA is a disease with variable clinical demonstrations and pathogenic characteristics among affected individuals. Hence, several different pathotypes of RA patients can be defined. One distinction is based upon the detection or not of autoantibodies in the blood of patients, dividing them in “seropositive” and “seronegative” respectively. In the first category autoantibodies against immunoglobulin G, also known as rheumatoid factor (RF), and citrullinated proteins, also referred as anti-citrullinated protein antibodies (ACPAs) can be detected. Interestingly though, this is not the case in the second category of patients, who are negative for those autoantibodies (Smolen et al. 2018). Another interesting categorization of patients has emerged based on linking gene expression signatures from synovium and

peripheral blood with clinical and imaging phenotypes, leading to the definition of three distinct RA pathotypes: “fibroblastic pauci-immune pathotype”, “macrophage-rich diffuse-myeloid pathotype”, and a “lympho-myeloid pathotype” (Fig.2). Those three pathotypes are characterized by different levels of immune cells infiltration. In the first category there is a lack of infiltrating cells, in the second one an enrichment of macrophages or monocytes is prevalent, while in the third B cells and T cells infiltration is observed (Lewis et al. 2019).

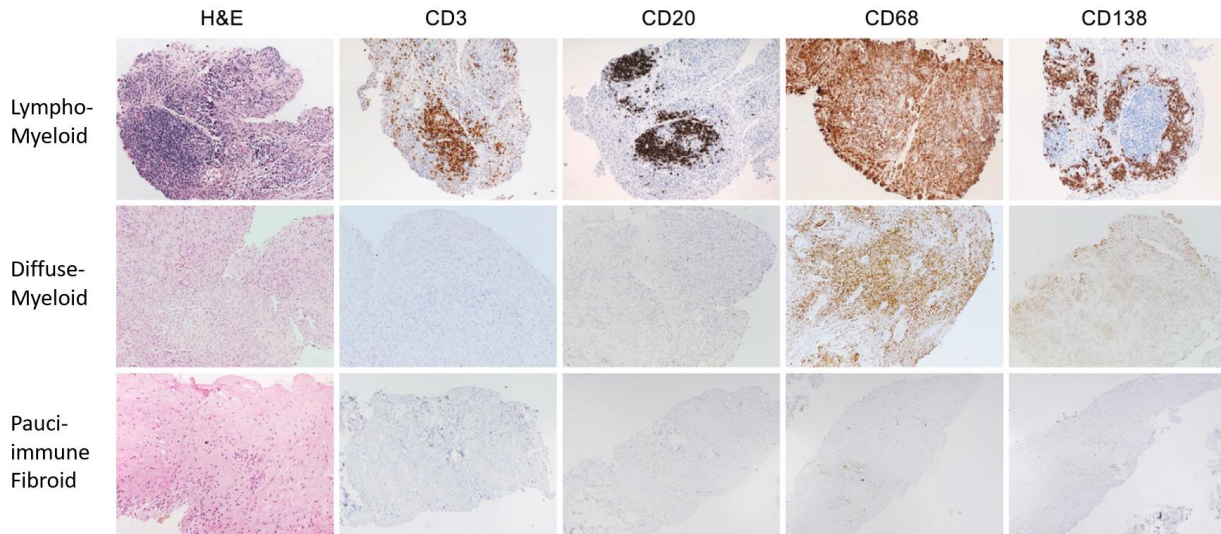


Figure 2. Immunohistochemistry of synovial biopsies from untreated patients with early RA for CD20+ B cells, CD3+ T cells and CD68+ macrophages and CD138+ plasma cells in lining and sublining compartments. They are categorized in three groups: lympho-myeloid (B cell aggregates are present), diffuse-myeloid (characterized by macrophage infiltration), or pauci-immune fibroid (lack of or low infiltration of immune cells). (Adopted from Lewis et Al., 2019)

1.2 The role of cytokines in Rheumatoid Arthritis

Over the past decades numerous studies have highlighted that cytokines have a crucial role in RA pathogenesis. In more detail, different cytokines can affect both innate and adaptive immune system responses as well as the stroma responses during disease (Fig. 3). Additionally, cytokines can contribute to the transition from a systemic to localized disease. Moreover, they influence the way patients respond to different therapeutical interventions and they can also affect the duration of remission period, as well as the probability for a recurring disease flare in the future (McInnes, Buckley, and Isaacs 2016).

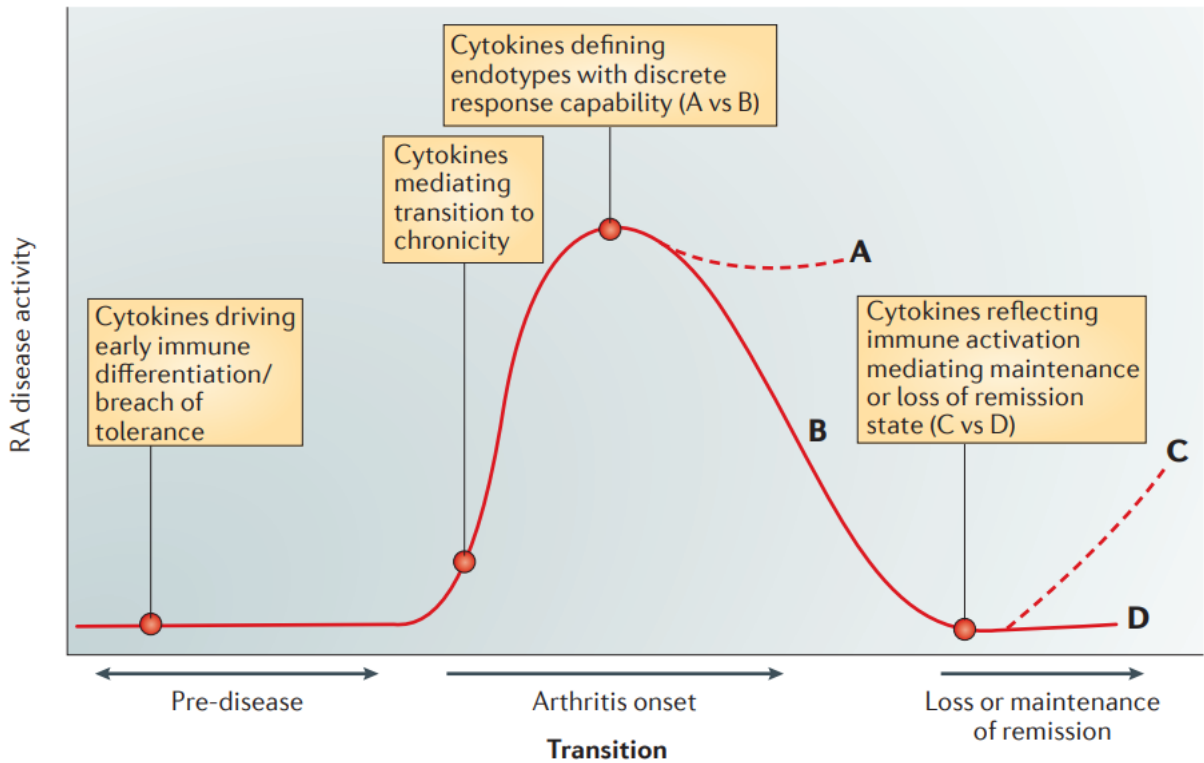


Figure 3. A schematic representation summarizing the different roles of cytokines during disease initiation, response to treatment and remission or relapse of the disease. (Adopted by McInnes et Al., 2015)

A wide range of cytokines can be present in the affected joints of RA patients. The most important of them include Tumor Necrosis Factor (TNF), Interleukins such as IL1, IL6, IL18, IL17, IL21, IL23, IL27, type I interferons (IFNs), as well as the Granulocyte- Macrophage Colony- Stimulating Factor (GM-CSF).

Before focusing on the pivotal role of TNF, it is important to discuss the effects of the rest of the cytokines mentioned in the previous section. Regarding IL6, it is worth stating that it is a cytokine participating in innate and adaptive immune response over the course of the disease. IL6 can activate other cell types both through cis (IL-6 binds to membrane IL-6 receptor) and trans-signaling (IL-6 & IL-6 soluble receptor binding and homodimerization with the subunits of glycoprotein 130). Notably, IL6 can aggravate synovitis and damage in the cartilage and bone of the affected joint, by promoting migration of neutrophils, aiding the maturation of osteoclasts, and supporting pannus proliferation through increased levels of vascular endothelial growth factor (VEGF) expression. Furthermore, it can influence the differentiation of B cells to plasma cells that produce antibodies. Moreover, IL6 in combination with transforming growth factor beta (TGF- β) in mouse models or IL-1 β and IL-23 in human can lead to the

differentiation of naïve T cells into T-helper 17 cells (Th17), which in turn secrete IL-17 (Srirangan and Choy 2010).

The IL-1 family contains members that can exhibit either pro-inflammatory or anti-inflammatory properties and can be detected in the joints of RA patients. The equilibrium between the two aforementioned categories can often affect the severity of disease symptoms (Dinarello 2019). Even though IL1 is connected to innate immune response, during the acute or chronic phase of the disease, its inhibition has not been proved as efficacious as expected against RA till this day. This could be attributed to the fact that pathways related to RA are mediated by IL1 and TNF in synergy, thus suggesting that the role of IL1 in the cytokine cascade of RA is not dominant (McInnes, Buckley, and Isaacs 2016).

Regarding IL-17A, it has been shown that it contributes to the secretion of other proinflammatory cytokines (such as TNF, IL-6, IL-1, GM-CSF), chemokines (like CXCL8, CCR2, CCR3) and matrix metalloproteinases (MMPs). Additionally, it influences processes like angiogenesis and activation of osteoclasts. Moreover, IL-17A in combination with other growth factors causes an anti-apoptotic effect in fibroblast like synoviocytes (FLS), T cells and B cells. Thus, its presence is associated both with inflammation and bone damage during the course of disease (McInnes, Buckley, and Isaacs 2016). Interestingly, a loop between IL-17 and IL-6 production is established since IL-17 promotes IL-6 production through FLS, while IL-6 contributes to IL-17 secretion through the stimulation of naïve T cells and its differentiation in Th17 cells.

Other cytokines that participate in RA disease are interleukins IL-12, IL-23, IL-27 and IL-35, all members of the IL12 family. Despite the fact that those cytokines share structural similarities (IL-12: subunits p40, p35, IL23: p40, p19, IL27: EBI3, p28, IL35: EBI3, p35), they can have different roles. IL-12 and IL-23 exhibit mainly proinflammatory attributes, while IL-35 poses a more immunoregulatory role. Interestingly, IL-27 can have both proinflammatory and immunoregulatory characteristics, depending on the maturation state of T cells.

Type I Interferons (IFNs) are critical components of the host defense mechanism against viral infections. However, they have also an active role in RA, as they can be detected in patients' synovial fluid and tissues (Conigliaro et al. 2010). Their biological activity in disease has also been supported from transcriptomics studies in synovium and leukocytes in peripheral blood. Although, they cannot be successfully targeted directly for RA therapy thus far, the IFN-stimulated response elements (ISREs) could have prognostic value

in predicting how patients will respond to treatment with certain biological drugs (McInnes, Buckley, and Isaacs 2016).

GM-CSF is a pleiotropic cytokine that can contribute to the activation of several cell types such as macrophages, neutrophils and dendritic cells (Fig. 4). This activation leads these cells to exhibit an inflammatory phenotype, followed by increased cytokine production and synthesis of prostanoids (McInnes, Buckley, and Isaacs 2016). GM-CSF has been detected in RA patients both in synovial fluid and in blood.

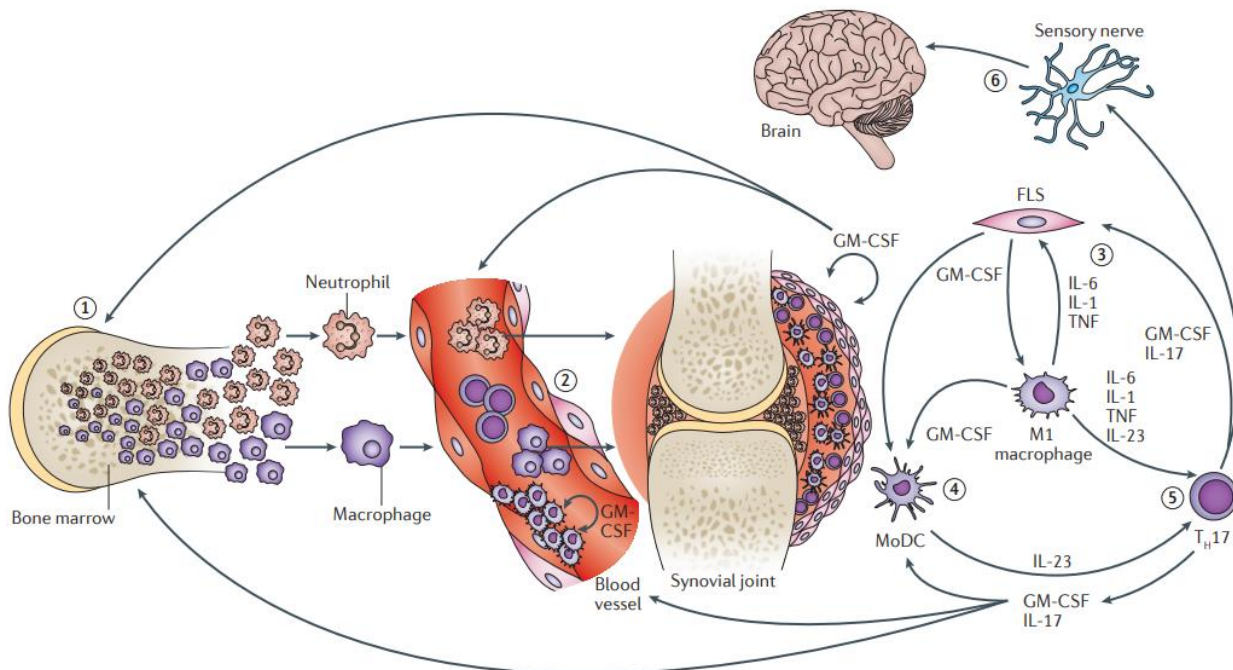


Figure 4. Schematic figure summarizing the contribution of GM-CSF in RA progression and pain elicitation. (Adopted by Wicks et Al., 2016)

Regarding the therapeutic potential of blocking this cytokine, early clinical trials with mavrilimumab have demonstrated positive outcomes, particularly in reducing disease activity and alleviating pain. (Wicks and Roberts 2016).

1.3 The role of Tumor Necrosis Factor in Chronic Inflammatory Diseases

Tumor Necrosis Factor (TNF) is one of the most important cytokines for both RA initiation and development (McInnes, Buckley, and Isaacs 2016).

TNF has a soluble (sTNF) and a transmembrane form (mTNF). After trimerization of the protein, it can bind to two different receptors TNFR1/p55 and TNFR2/p75. It's worth mentioning that soluble TNF exhibits a selective binding for TNFR1 receptor, while transmembrane TNF can bind to both receptors (TNFR1, TNFR2). Interestingly, the TNF- α converting enzyme, also known as TACE, can lead to the conversion of mTNF to its soluble form through enzymatic processing.

Binding of TNF to its receptors can induce various inflammatory signaling pathways. The TNFR1 receptor can be found universally expressed in almost every cell type. On the contrary TNFR2 is expressed mainly in neurons, oligodendrocytes, regulatory T cells (Tregs) and monocytes (Atrekhany et al. 2020; Madsen et al. 2020; 2016; Veroni et al. 2020; X. Chen et al. 2007; Polz et al. 2014). Through a complex procedure, which includes the recruitment of various molecules, followed by creation of different complexes (shown in a schematic representation in figure5), TNF-TNFR1 binding leads to the activation of NF- κ B signaling, which in turn induce transcription of pro-inflammatory genes. Another similar mechanism, associated with up-regulation of pro-inflammatory genes, includes the activation of the transcription factor (TF) AP1 through MAP kinases p38 and JNK. Additionally, the interaction between TNF and TNFR1 receptor is responsible for two biological processes related to cell death. In the first one, cell apoptosis is achieved in a caspase8 dependent manner, while in the second one necroptosis is caused by rupture in the cell membrane followed by intrusion of ions with a positive charge including Ca^{2+} , Na^+ and K^+ . Although TNF-TNFR2 interaction exhibits differences with the TNF-TNFR1, especially as regards the molecules required to be recruited for the signal transduction, they share similar downstream effects. In more detail, TNF-TNFR2 can induce both canonical and non-canonical NF- κ B signaling. In addition, it can reinforce TNFR1-driven apoptosis through controlling the TRAF2 expression levels in the cytoplasm. Of note, both TNF receptors can acquire soluble forms through proteolytic cleavage. These soluble receptors, known as sTNFR1 and sTNFR2, act as natural TNF antagonists.

TNF blockade has been proved a successful form of therapy for many RA patients. TNF can be mainly targeted with the use of biologic drugs – monoclonal antibodies such as infliximab, adalimumab, certolizumab pegol, golimumab or fusion protein Etanercept (Enbrel), as well as small molecule inhibitors that affect trimerization of TNF. Although anti-TNF treatment is widely recognized as one of the most effective therapeutic solutions for chronic inflammatory diseases, some patients do not respond to the therapy, experience a loss of response over time, or become more susceptible to infections due to the immune system suppression caused by the treatment. (McInnes, Buckley, and Isaacs 2016; Willrich, Murray, and Snyder 2015; Mazumdar and Greenwald 2009; Goel and Stephens 2010).

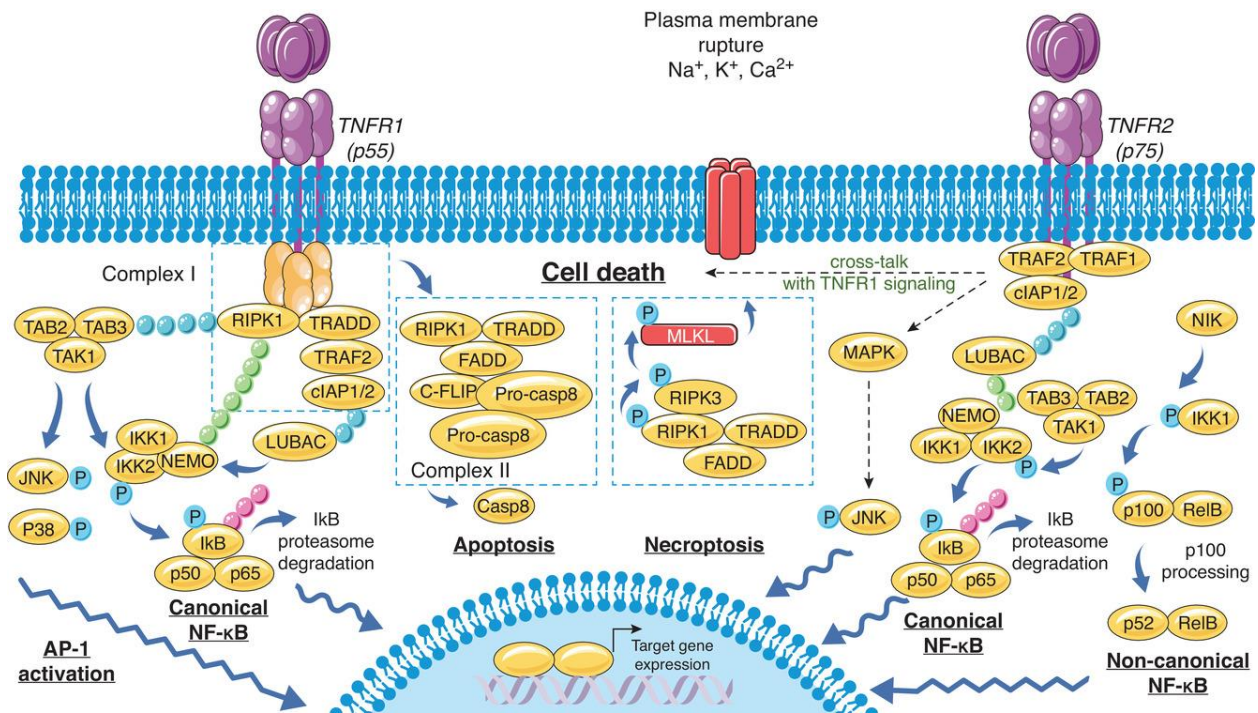


Figure 5. An overview of TNF binding to TNFR1 & TNFR2, summarizing the different downstream effects that are induced in a cell by those two interactions. (Adopted by Atrtekhyani et Al., 2020)

1.4 The use of animal models in Rheumatoid Arthritis

The use of animal models is very common in the biomedical field, as they are extremely useful for studying the underlying causes of disease pathology and testing novel therapeutic approaches before proceeding with clinical studies in human patients. In the case of RA, a wide variety of animal models have been generated to cover different aspects of the disease, including initiation, chronicity, relapse, and outcome (Kollias et al. 2011). In the following section a brief description of the most widely used animal models can be found.

➤ Adjuvant-induced arthritis model (AIA)

In this model of arthritis, the disease is induced by intradermally injecting a rat with mycobacterial cell walls diluted in mineral oil. However, this model fails to accurately describe the disease since it showcases characteristics of systemic inflammation. Another similar model to this arthritis model is the pristane-induced arthritis (PIA) model. This model uses a component of mineral oil called pristane and is T cell-

dependent. It can be induced in both mice and rats, with its main characteristics being joint swelling and infiltration of inflammatory cells, leading to chronic and recurring disease.

➤ Zymosan-Induced Arthritis (ZIA)

Zymosan induced arthritis (ZIA) model can be initiated by utilizing zymocan, an ingredient which can be found in *Saccharomyces cerevisiae*, and injecting it in mice or rats. Studies with TLR2 and C3 deficient mice suggested that both innate and acquired immune pathways are involved in ZIA. More specifically, TLR2 exhibited an important role in adaptive immune response, while C3 appeared to be less influential in this arthritis model (Frasnelli et al. 2005). The symptoms of the disease begin at the 3rd day post immunization and include immune cell infiltration, formation of pannus and cartilage destruction.

➤ Type II collagen induced arthritis (CIA)

In the collagen-induced arthritis (CIA) model, which is an induced model of rheumatoid arthritis, immunization with type II collagen (diluted in complete Freund's adjuvant) leads to Th17-mediated responses and the production of antibodies against joint collagens. This process ultimately triggers inflammation and pain in the affected area. CIA has been successfully tested in mice, rats, rabbits, and non-human primates. Disease onset typically occurs around day 12 post-immunization, with the peak of the disease being reached by day 30 (Trentham, Townes, and Kang 1977). While the CIA model shares many similarities with human rheumatoid arthritis, such as the development of rheumatoid factor (RF) and the presence of ACPAs, it also exhibits significant variability, often associated with the quality of the collagen II used during injection or group-related stress (R. Holmdahl et al. 1992).

➤ Collagen antibody-induced arthritis model (CAIA)

CAIA is an inducible mouse model of arthritis. The immunization can be achieved either by utilizing directly monoclonal antibodies, that target epitopes of Collagen type II (Rikard Holmdahl et al. 1986), or by serum transfer from other immunized mice or RA patients, given that their serum contains the relevant monoclonal antibodies (K. S. Nandakumar, Svensson, and Holmdahl 2003). The developed arthritis is characterized by the implication of macrophages and fibroblasts, while it is considered independent of B and T cells. The onset of disease is defined at 48 hours post immunization and the peak of disease is reached at 7 days after the injection. However, despite eliciting both innate and adaptive responses, the immune compartment is not significantly involved. As a result, this mouse model fails to capture important elements found in human disease.

➤ Serum-transfer Induced arthritis (STIA)

Although STIA is considered an inducible model of RA, is also highly dependent on a spontaneous model known as K/BxN. Disease is initiated after injection of anti-G6PI antibodies intraarticularly (K. S. elv. Nandakumar and Holmdahl 2006). Symptoms of RA can be manifested as early as 20 minutes after the injection, while the peak of the disease is reached approximately at the 14th day post induction. This procedure can be applied to a variety of mouse strains; however, variability in the phenotype of disease is observed. Recently, single-cell (SC) studies in STIA mouse model have highlighted the existence of distinct fibroblasts subsets that perform different functions during disease progression (Fig. 6). More particularly, FAP α + THY1+ fibroblasts are found in the synovial sub-lining compartment, while FAP α + THY1- fibroblasts are restricted to the lining layer of the synovial membrane. Interestingly, when these cell populations are adoptively transferred into the inflamed ankle joints of mice with STIA, FAP α + THY1- fibroblasts selectively mediate bone and cartilage damage with no significant contribution to inflammation. On the contrary, the transfer of FAP α + THY1+ fibroblasts leads in the development of a more severe and persistent inflammatory arthritis, with minimal effect on bone and cartilage integrity (Croft et al. 2019).

➤ Human chimeric transfer model

For the generation of the human chimeric transfer model, mice with severe combined immunodeficiency disease (SCID) are submitted to a surgical procedure, which enables the implantation of small fragments of tissue from human synovium (Geiler et al. 1994). These humanized mice exhibit pannus formation and cartilage destruction, recapitulating aspects of disease found in RA patients. A disadvantage that could impede the experimental usage of the model is the time required for progression of the disease. In more detail, signs of bone erosion can be detected histologically after the 35th day following induction, while pannus formation begins at the 105th day.

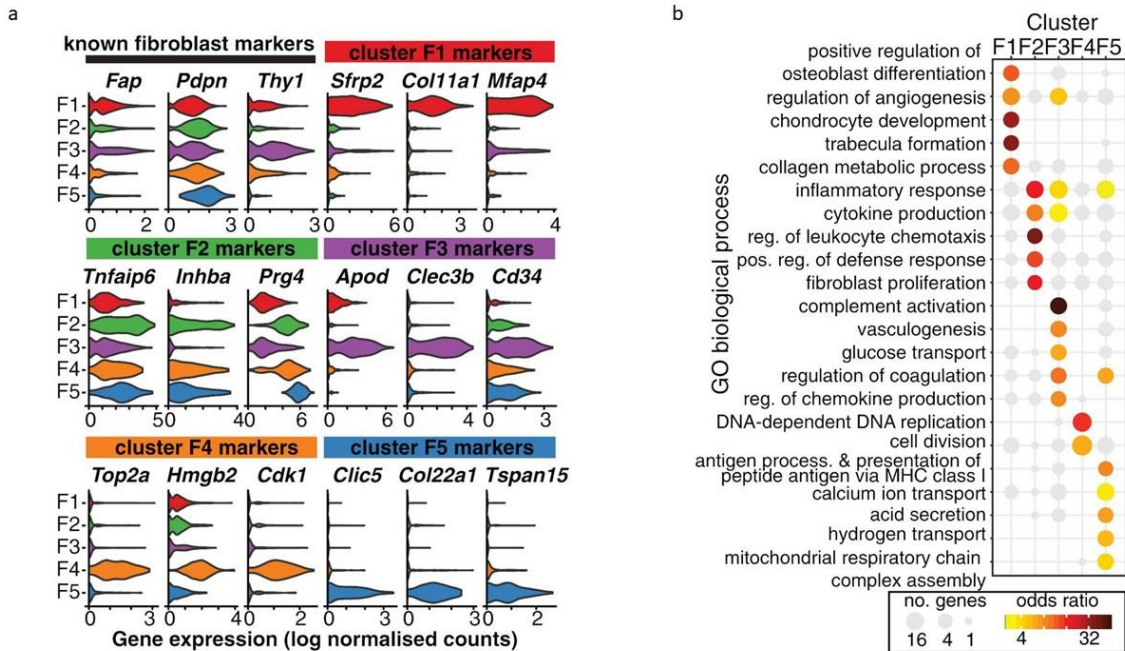


Figure 6. (a) Violin plots showcasing marker genes of the fibroblast subsets identified during the analysis of STIA single-cell RNA-sequencing data. (b) Dotplot depicting unique and shared enriched GO terms for the different fibroblast clusters. (Adopted from Croft et al, 2019)

➤ K/BxN model

K/BxN model was also mentioned briefly in the description of STIA model above. For the generation of this mice, a cross between non-obese diabetic mice (NOD) and mice with a KRN T-cell receptor transgene (K/B), is required. The TCR receptor can recognize a specific peptide from the glucose-6-phosphate isomerase (GPI) protein, which is presented through the major histocompatibility complex class II (MHC II) (Monach, Mathis, and Benoist 2008). K/BxN exhibits many similarities with RA pathology, including development of synovitis, high rates of fibroblast proliferation, adaptive immune response and destruction of bone and cartilage. Arthritis symptoms begin 15 days after the birth of the mice and the peak of the disease is reached at 3 months of age. Is worth repeating that the serum of these mice can induce arthritis in the STIA mice.

➤ IL-1ra-deficient mice

IL-1ra deficient mice is a T-cell dependent RA model. More precisely, a complete knockout of Interleukin 1 (IL1) receptor causes an increase in the expression of IL1 systematically, leading ultimately to the

development of chronic RA (Horai et al. 2000). In this mouse model, several aspects of human disease are represented. More particularly, a variety of cytokines, such as IL1, IL6, IL17, and TNF are up-regulated. Additionally, RF and autoantibodies (targeting collagen and double-stranded DNA) can be detected during the progression of disease. The initiation of RA is estimated at the 5th week of age, while the peak of the disease is reached at the 16th week.

➤ SKG mice

The SKG is a spontaneous model of RA, linked with a mutation in the ZAP-70 gene (zeta chain associated protein kinase 70kDa), which affects T-cell receptor signaling and leads to T-cell driven arthritis (Sakaguchi et al. 2003). Classical manifestations of human disease are also present in this model, including hyperplasia of the synovium and synovitis, accompanied by immune cells' infiltration, detection of RF and autoantibodies, as well as, formation of pannus and destruction of cartilage and bone. At two months postnatally the disease begins and reaches its peak at approximately 8 months of age. Since the point mutation mentioned above creates a genetic predisposition to the mice, it is worth mentioning that the disease onset can be accelerated when there is exposure to serum (containing antibodies against glycoprotein 39) from K/BxN mice.

➤ TNFΔARE/+ mice

The TNFΔARE mouse model is generated by the deletion of the AU-rich elements (ARE) from the 3' untranslated region of the TNF gene. These regulatory sequences have an important role in both the stability and translation of the mRNA molecule to the corresponding protein (Kontoyiannis et al. 1999). Ablation of ARE elements leads to a chronic over-expression of endogenous mouse TNF. This mouse model displays a phenotype that is characterized by comorbidities including chronic polyarthritis and inflammatory bowel disease. Arthritis manifestations can be detected as early as 3 weeks of age, while the peak of disease is reached approximately at 16 weeks post birth. Furthermore, the clinical and histological manifestations of arthritis in this model are exclusively dependent on the overexpression of TNF by synovial fibroblasts (SFs), rendering the involvement of B and T cells unnecessary for both the initiation and progression of the disease (Kontoyiannis et al. 1999; Maria Armaka et al. 2008).

➤ *hTNFtg* mice

The Tg197 mouse model is a spontaneous model that exhibits chronic inflammatory polyarthritis, sharing many characteristics with RA in humans. It provided the first in vivo evidence demonstrating the pathogenic role of the TNF molecule in RA. This mouse model contains five copies of the human TNF

transgene (hTNF), characterized by chronic overexpression of human TNF (Fig. 7). This overexpression is achieved by replacing the 3' untranslated region of the hTNF gene with the corresponding region from the human β -globin gene, resulting in the continuous expression of human TNF mRNA (Keffer et al. 1991). The symptoms experienced by the animals range from pain and swelling of back joints to inability of proper movement in the front limbs. Disease onset starts at 3 weeks of age, while in most of the cases mice die approximately at the 12th week, after having developed cachexia.

Histological study of the affected areas shows hyperplasia in the synovium, immune cell infiltration, cartilage destruction and bone erosion, however RF is absent at all stages of disease. Additionally, an increase in the expression of different metalloproteases, such as MMP (-3, -9, -13) is observed in both Tg197 mice and RA patients.

SFs possess a pivotal role in the initiation and development of the disease in Tg197 mouse model, as they are the main cell type responding to TNF signals. Moreover, it has been shown that they are capable of driving disease advancement in immunodeficient Rag^{-/-} mice (Aidinis et al. 2003). Although, TNFR1 receptor on SFs is crucial for the pathology progression, lack of TNFR1 signaling does not affect disease onset (Douni et al. 1995; Marietta Armaka et al. 2018). Another important aspect of SFs is that they exhibit an activated phenotype, they express cytokines and genes related to cell proliferation and migration, sharing overall common properties with the RA FLS, which are their human counterparts (Vasilopoulos et al. 2007; Ntougkos et al. 2017).

It is worth noting that Tg197 has also been used in drug testing. More specifically, anti-TNF, anti-IL1 and anti-IL6 therapies have been accessed utilizing this model.

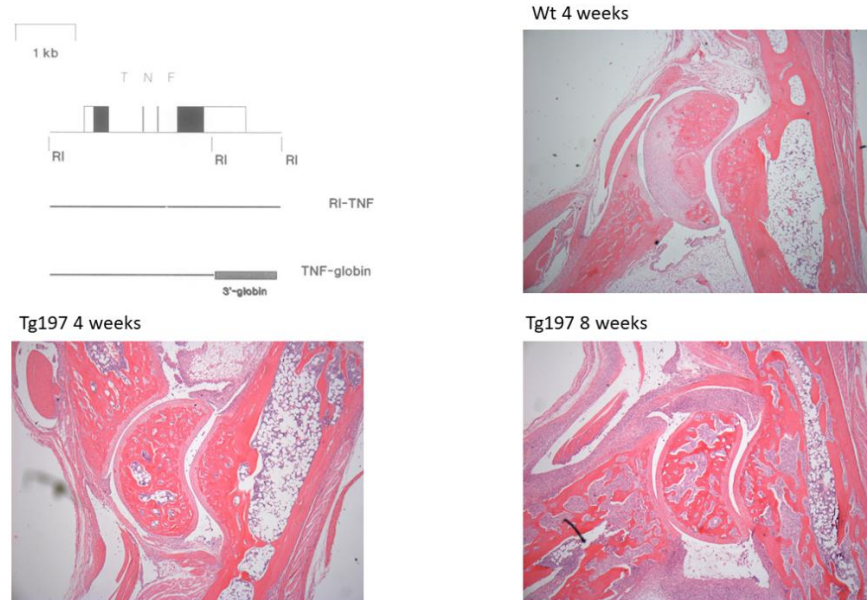


Figure 7. In the top left panel, a schematic representation of hTNF transgene is depicted. In the next three panels, hematoxylin and eosin (H&E) staining of the ankle joint at the talus level is exhibited for wild type mice and hTNFtg mice at the age of 4 and 8 weeks. (Adopted from Keffer et Al., 1991 and Armaka et Al., 2022)

1.5 The function of fibroblast cells in homeostasis and disease

Fibroblast cells originate from a non-hematopoietic lineage, they exhibit a spindle-shaped morphology, they possess the role of resident cells in many different tissues and are known for producing extracellular matrix proteins. Numerous studies over the years have highlighted their multifunctional role both in homeostatic and disease states. Fibroblasts are known to be implicated in various cancer types, inflammatory diseases, and fibrosis (Koliaraki et al. 2020). They can sense both molecular cues related to pathogens and mechanistic stress. They respond by secreting cytokines and chemokines, which play a role in the recruitment of leukocytes.

Studies at single cell level seek to identify marker genes uniquely expressed in specific fibroblast subsets, in order to facilitate their isolation. To date, fibroblasts are mainly selected based on the absence of protein markers such as CD45, CD31 and EPCAM, which are found in myeloid, endothelial and epithelial cells, respectively. Although positive markers like PDGFRA and PDPN are associated with fibroblasts, they are not exclusively expressed by them.

Recent studies have highlighted the existence of various fibroblast subsets that perform distinct functions across different tissues, under diverse conditions (Davidson et al. 2021; Buechler et al. 2021). Fibroblast heterogeneity is evident not only during disease, but also during homeostasis. For instance, fibroblasts located in the villus of the intestine express the genes WNTa and WNTb, however the fibroblasts in the lamina propria express only WNTb (Smillie et al. 2019). In the case of lung disease, a fibroblast subpopulation characterized as AXIN1⁺/ PDGFRA⁻ is connected to lung fibrosis, however a different population AXIN⁺/PDGFRA⁺ is essential for the maintenance of the alveolar epithelium during homeostasis.

There was an effort by (Buechler et al. 2021) to delineate different fibroblast subsets by building mouse and human single cell atlases of fibroblasts in steady and perturbed states in different tissues. Bioinformatics analysis of the single-cell transcriptomics profiles highlighted the existence of 10 distinct fibroblasts groups, characterized by the expression of different markers such as, Pi16, Col15a1, Ccl19, Coch, Comp, Cxcl12, Fbln1, Bmp4, Npnt and Hhip. From those clusters the Pi16⁺ fibroblast subset was predicted as an initial state in the differentiation process. Interestingly, the comparison between fibroblast clusters that were identified in steady and perturbed states revealed two groups that were present in both conditions (Pi16⁺ and Col15a1⁺) and 3 groups that emerged in the disease state (characterized by markers such as Lrrc15, Cxcl5 and Adamdec1). Besides differences in the gene expression profiles of the aforementioned clusters, it is suggested that they are also associated with separate biological processes. More particularly, the two universal populations are mainly involved in fibroblasts' development and ECM secretion, while the three disease emerging clusters are implicated in PI3K, TNF, NFκB and TGFβ signaling.

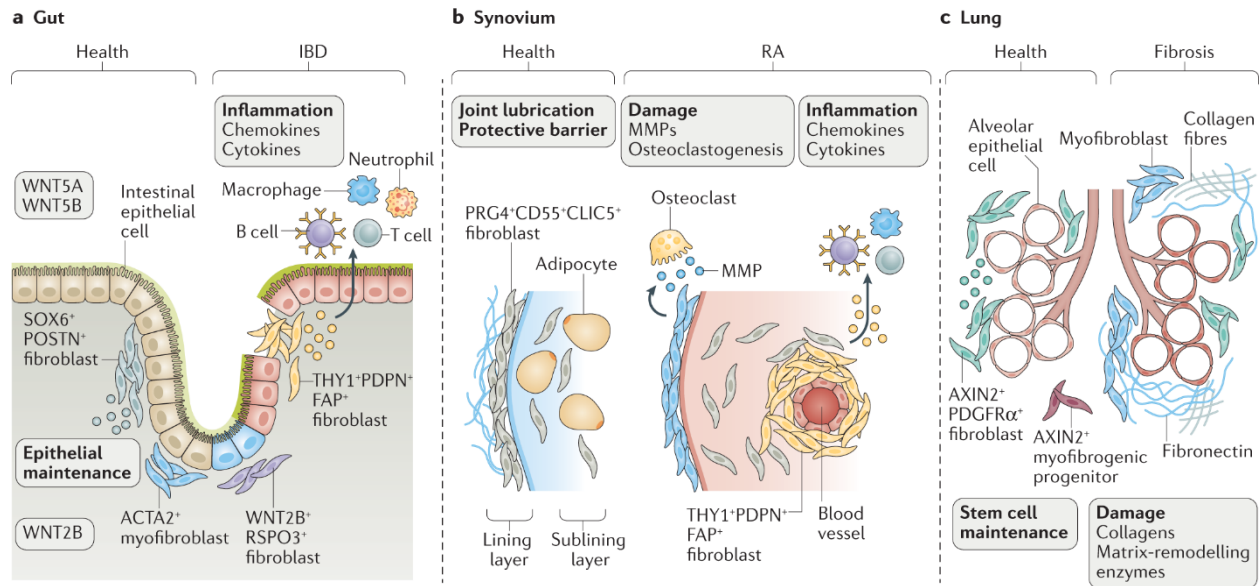


Figure 8. A schematic representation of distinct fibroblast subsets, showcasing the differences in their functional role during homeostasis and disease state in intestinal, synovial and lung tissues. (Adopted from Davidson et Al., 2021)

It is worth noting also that another category of fibroblasts, known as cancer associated fibroblasts (CAFs), can be found in various types of cancers and can demonstrate either tumor-promoting or tumor-restraining behavior, depending on the specific context or conditions within the tumor microenvironment (Fig. 9).

Focusing on the roles of SFs during RA disease, it is worth noting that activated SFs secrete cytokines, chemokines and metalloproteinases, which facilitate the inflammation of the joint, infiltration of immune cells, cartilage destruction and bone erosion. Notably, TNF signaling is both sufficient and necessary for the development of chronic polyarthritis in mice (Maria Armaka et al. 2008; Marietta Armaka et al. 2018). More specifically, it has been shown that SFs with pathogenic characteristics are able to initiate disease in mice with RAG knockout (Müller-Ladner et al. 1996; Lefèvre et al. 2009; Aidinis et al. 2003).

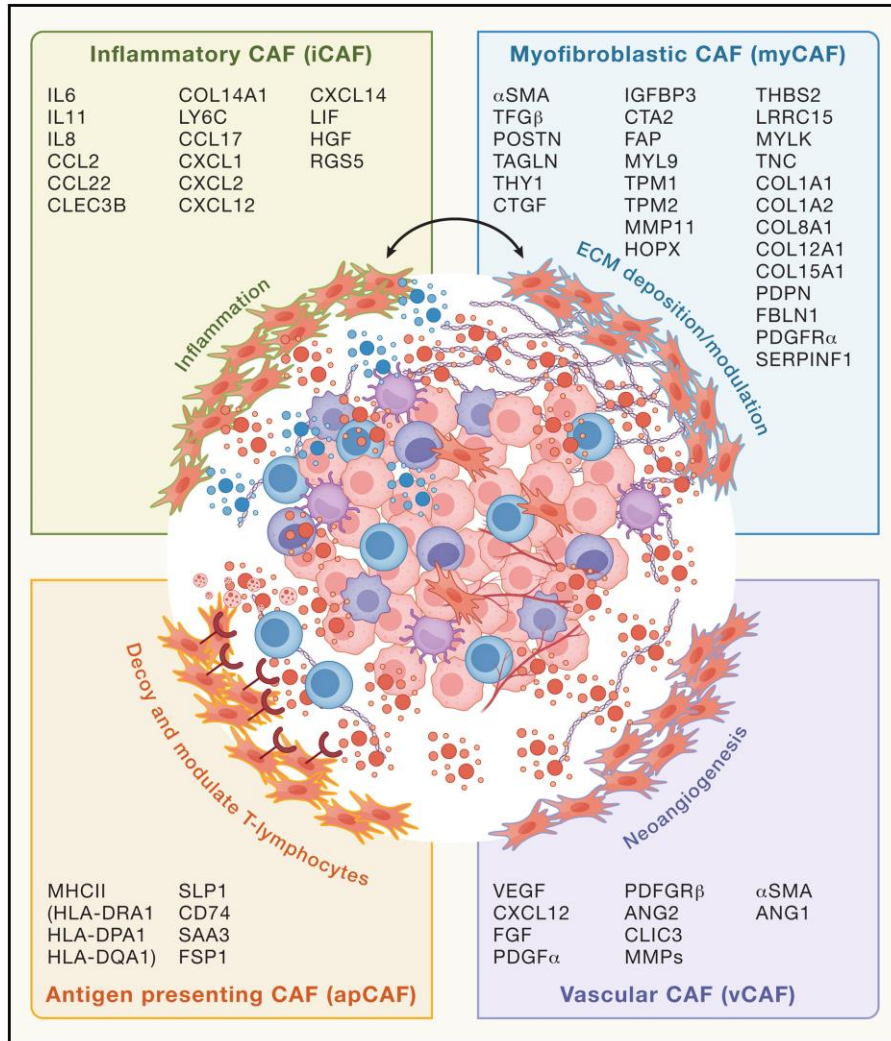


Figure 9. A simplified representation of the different types of CAFs. (Adopted from Chhabra et al., 2023)

As regards the characterization of fibroblast populations in RA at single cell resolution, several attempts have been made over the last five years. Single-cell RNA-sequencing (scRNA-seq) in RA and OA patients revealed the existence of four different SFs populations CD34⁺ (SC-F1), HLA-DRA^{hi} (SC-F2), DKK3⁺ (SC-F3) and CD55⁺ (SC-F4). The first three Thy1⁺ clusters belong to the sublining compartment, while the fourth Prg4^{hi} cluster belongs to the lining compartment (F. Zhang et al. 2019). Another recent single-cell study in RA patients highlighted the important role of Notch3 signaling in the differentiation processes of Thy1⁺ fibroblasts to Prg4⁺ through intermediate transcriptional states (Wei et al. 2020).

Single cell studies in mouse models of RA have contributed significantly to the appreciation of the heterogeneity of SFs and the understanding of their functional roles during disease progression. As mentioned before (in the paragraph dedicated to STIA mouse model) (Croft et al. 2019), five distinct SFs

populations were identified in STIA mouse model. The ones belonging to the sublining layer (F1, F2, F3, F4) were mainly implicated in inflammatory responses, while the fifth population (F5) belonging to the lining layer was mainly associated with a destructive transcriptional profile. Additionally, sequencing techniques at single cell resolution were employed to study the ankle joints of Tg197 mice at early and established disease timepoints. Single cell data analysis pinpointed the existence of 9 distinct fibroblasts subsets that can be categorized in three main groups, namely sublining, intermediate and lining fibroblasts. Comparing the relative abundances of the different populations between the wild-type (wt) and disease states (Fig. 10) revealed the expansion of intermediate and inflammatory lining synovial fibroblasts (SFs) (Marietta Armaka et al. 2022). Since the aforementioned dataset served as a use-case scenario for the computational platform developed within the context of this PhD dissertation, accompanied by custom analysis tasks, a more detailed presentation of the biological findings will follow in the next sections. In both cases, bioinformatics methods, including correlation analysis between cluster marker genes and integration of mouse-human datasets, revealed correspondences between the fibroblast populations identified in human patients and the mouse model.

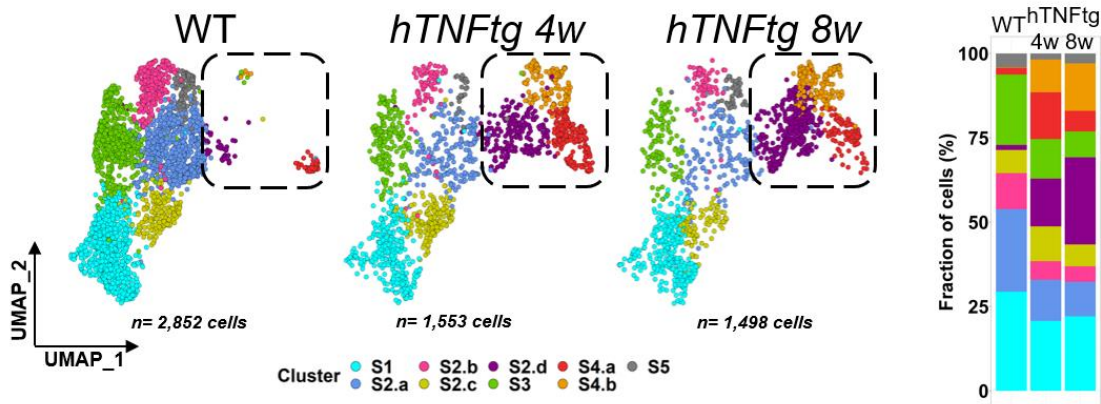


Figure 10. Comparison of the relative abundances of SFs populations across wild type and hTNFg arthritic mice. (Adopted from Armaka et Al., 2022)

1.6 Single cell sequencing technology

Single-cell sequencing technologies have transformed genomics, transcriptomics, and proteomics, enabling the examination of intricate biological systems at the level of individual cells. Yet, before focusing

on the contemporary techniques, it is valuable to briefly refer to some of their predecessors, which laid the foundation for the new era of omics and enabled the generation of large scale of biological data.

➤ Microarrays

Microarrays involve the binding of thousands of nucleic acids to a surface, allowing for the assessment of the relative concentration of nucleic acid sequences within a mixture through hybridization and subsequent detection of the hybridization events. Their most common application is in the measurement of gene expression (Bumgarner 2013). To achieve this RNA is extracted from target cells and is directly labeled or converted into labeled complementary DNA or RNA (cDNA or cRNA), further amplified through the Eberwine process (Van Gelder et al. 1990). Various labeling techniques are available, among these the most common methods involve incorporating fluorescently labeled nucleotides during cRNA or cDNA synthesis, or biotin-labeled nucleotides during cRNA synthesis (e.g., Affymetrix platform). The labeled samples are then hybridized onto the microarray, washed, and fluorescence is detected, typically using a scanning confocal microscope. The intensity of signals at each spot reflects the expression level of the corresponding gene.

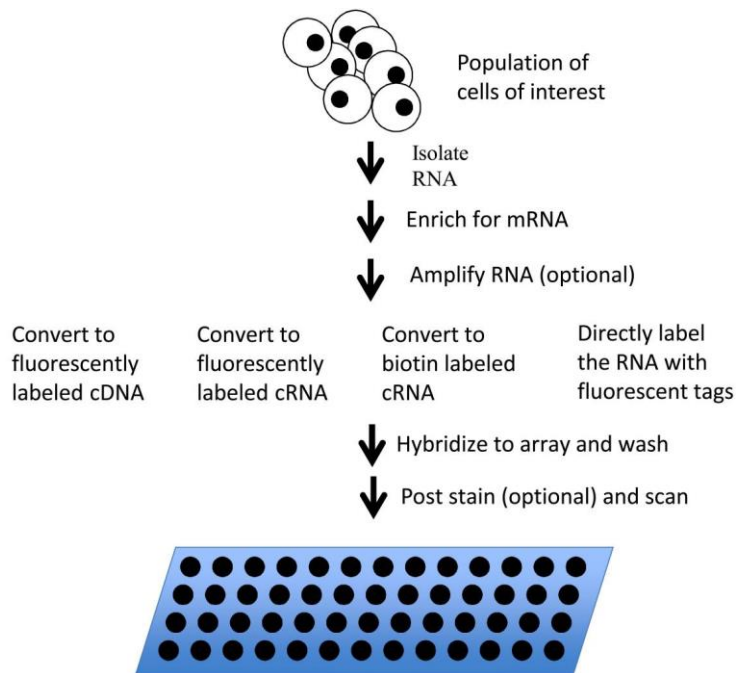


Figure 11. Schematic representation of a typical microarray experiment (Adopted from Bumgarner et Al., 2013).

Except for gene expression profiling, microarrays have been successfully used in the tasks of transcription factor binding analysis (Horak and Snyder 2002) and SNP genotyping. It's important to note that microarrays have been invaluable in comparing gene expression profiles between different conditions or

states, allowing for the identification of disparities between diseased and healthy samples, as well as the evaluation of the effects of various perturbagens such as smoke, radiation, or treatments. However, despite their significant contributions to biomedical research, microarrays do come with limitations. Specifically, the fluorescent signal detected on a microarray may not directly correlate with the concentration of the target species in solution, due to the dynamics of hybridization. Moreover, when studying genes or gene families with multiple spliced variants, cross-hybridization can occur, leading to an increase in false positives as similar sequences might bind to the same probes. Furthermore, the design of probes is constrained to genes cataloged in the reference genome of an organism, thereby excluding other potential targets such as microRNAs or long non-coding RNAs that have not yet been annotated.

➤ Bulk RNA sequencing

Bulk RNA sequencing (RNA-Seq) is a potent successor of microarrays, harnessing high-throughput sequencing methods to explore the transcriptome of a cell. It offers significantly enhanced coverage and resolution of transcriptome dynamics. Beyond simply measuring gene expression levels, bulk RNA-Seq data facilitates the discovery of novel transcripts, identification of alternatively spliced genes, and detection of allele-specific expression. Moreover, RNA-Seq is not limited to polyadenylated messenger RNA (mRNA) transcripts; it can also probe various RNA molecules, including total RNA, pre-mRNA, and noncoding RNA such as microRNA and long noncoding RNA (ncRNA) (Kukurba and Montgomery 2015). However, it is worth noting that the effectiveness of an RNA-seq experiment hinges on meticulous experimental design, tailored to address specific biological questions. This involves management of various technical considerations, including the choice of RNA-extraction protocols like polyA-selection or ribosomal depletion, and deciding between sequencing options such as single-end or paired-end reads. Moreover, determining the optimal sequencing depth and incorporating adequate technical and biological replicates are crucial steps to capture the inherent variability of biological systems. Properly managing these technical aspects lays the groundwork for robust downstream statistical analyses, gaining meaningful insights into gene expression dynamics and regulatory mechanisms (Conesa et al. 2016). The burgeoning utilization of RNA-seq analysis across diverse biomedical domains has spurred significant advancements within the bioinformatics community as well. This led to the development of numerous specialized software packages tailored to various stages of analysis. These tools encompass a wide array of functions, ranging from aligning sequences to a reference genome and summarizing read counts, to conducting differential expression analyses, reconstructing regulatory networks, and performing functional enrichment analyses. Such software contributions play a pivotal role in enhancing the efficiency

and accuracy of RNA-seq data interpretation by researchers. One of the major drawbacks of bulk RNA-seq, regarding its resolution, is that it provides only an averaged gene expression measurement across entire populations of cells. In more detail, the RNA is extracted from a large number of cells, pooled together, and sequenced collectively. This means that any differences in gene expression between individual cells within the population are lost, and the results represent an average expression profile of the entire population (Li and Wang 2021). To overcome this limitation and achieve higher resolution, scRNA-seq techniques have been developed.

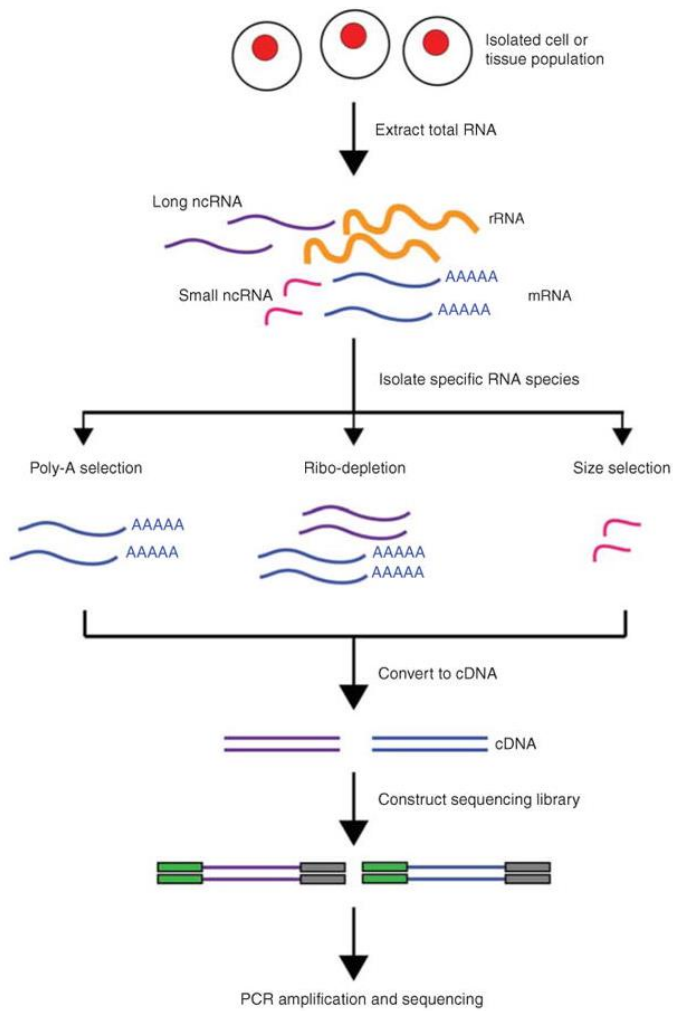


Figure 12. Brief overview of the main steps involved in a typical RNA-seq experiment (Adopted from Kukurba et Al., 2015).

➤ Single cell RNA sequencing

Over the past decade, single-cell RNA sequencing has revolutionized the landscape of transcriptomics. It was initially recognized as the Method of the Year for 2013 by Nature (“Method of the Year 2013”) and its

rapid advancement has empowered biomedical researchers to study gene expression profiles with unprecedented precision and resolution in many different complex biological systems. ScRNA-seq enables the interrogation of gene expression dynamics in individual cells, unraveling intricate molecular signatures and providing invaluable insights into many biomedical phenomena such as cellular heterogeneity in healthy tissues or tumor sites, developmental relationships among cell-types, and underlying disease mechanisms. Those new capabilities surpass the limitations posed by bulk RNA-seq methods, which can only capture the average expression of cell populations within a sample (Fig. 13).

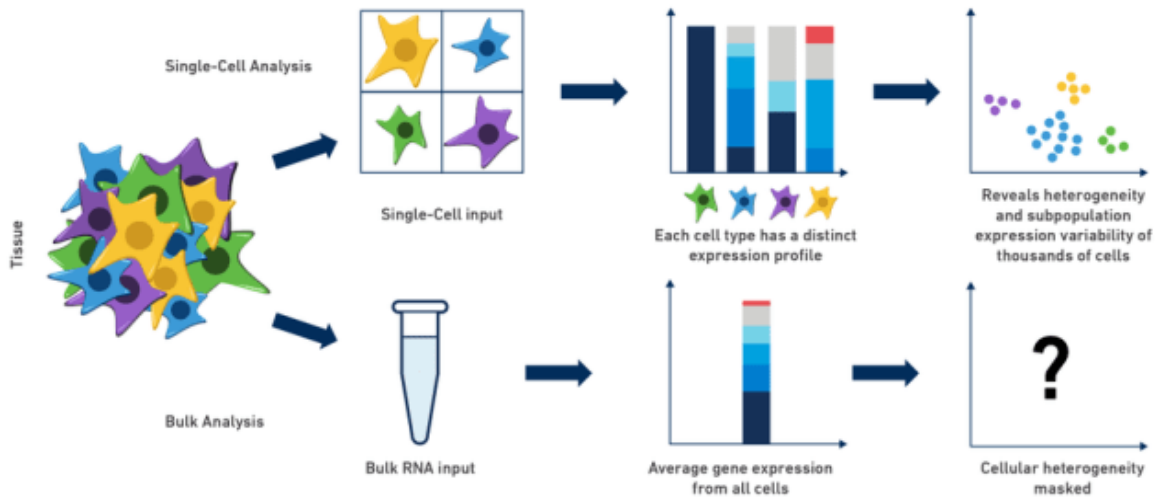


Figure 13. Schematic representation summarizing the differences between bulk RNA-seq and scRNA-seq (Adopted from 10x Genomics website).

The first single-cell transcriptome analysis employing a next-generation sequencing platform was performed in 2009 (F. Tang et al. 2009), allowing the characterization of cells during early developmental stages. Remarkably, substantial advancements in equipment, alongside significant improvements in the scalability of the software utilized for data analysis, have facilitated the profiling of hundreds of thousands, or even millions of cells within a single experimental procedure (Hwang, Lee, and Bang 2018).

Over the years, numerous protocols (Fig. 14) have been devised for single-cell RNA sequencing, each distinguished by factors such as the applied methodology for cell isolation, amplification, and sequencing (Papalexi and Satija 2018). Therefore, researchers often face the task of selecting the most appropriate protocol depending on the biological questions at hand. For instance, when the primary objective is to explore tissue heterogeneity and identify various cell populations, a protocol that allows profiling of a large number of cells, albeit with reduced sequencing depth, is typically preferred. On the contrary, in scenarios where researchers aim to dissect specific cell subsets in greater detail, protocols that focus on profiling

fewer cells but achieving deeper sequencing are often employed, allowing for the detection of a higher number of genes per individual cell.

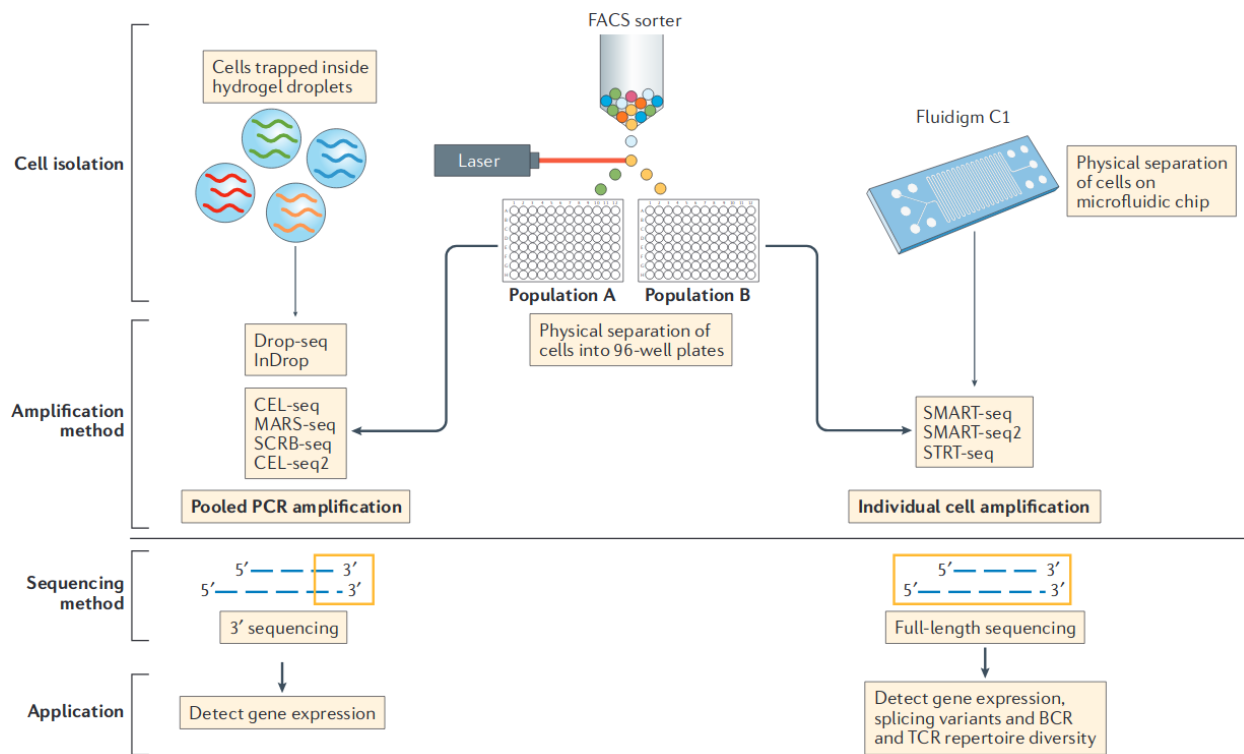


Figure 14. Different protocols of scRNA-seq experiments are showcased (Adopted from Papalexi et Al., 2017).

A significant effort in cataloguing the different scRNA-seq protocols and their major advantages and disadvantages was made by (G. Chen, Ning, and Shi 2019). These protocols can be further classified based on the isolation method employed. Established cell isolation techniques include limiting dilution, micromanipulation, flow-activated cell sorting (FACS), and laser capture microdissection.

Limiting dilution involves individual cell isolation via pipetting, while micromanipulation employs microscope-guided capillary pipettes to retrieve cells, limiting the throughput due to their low capacity.

In recent years, flow-activated cell sorting (FACS) has emerged as the leading method for isolating highly purified single cells. Initially, cells are tagged with fluorescent monoclonal antibodies, facilitating the recognition of specific surface markers for positive or negative cell selection.

Laser capture microdissection utilizes a laser system in conjunction with a computer to isolate cells from solid samples.

Notably, microfluidic technology for cell isolation has gained traction due to its advantages of minimal sample consumption, cost-effectiveness, and precise fluid control, with nanoliter-sized volumes reducing contamination risks. Platforms like Fluidigm C1 offer automated single-cell lysis, RNA extraction, and cDNA synthesis for hundreds of cells in parallel. Additionally, microdroplet-based microfluidics enables manipulation and profiling of thousands to millions of cells at a low cost, exemplified by the Chromium system from 10× Genomics, Drop-seq, and InDrop.

Another crucial distinction among scRNA-seq protocols lies in their ability to sequence the full-length transcript, demonstrated by Smart-seq2, SUPeR-seq, and MATQ-seq, or only capturing and sequencing the 3'-end or 5'-end of transcripts, as seen in Drop-seq, Seq-Well, DroNC-seq, SPLiT-seq, and STRT-seq. Full-length scRNA-seq methods excel in isoform analysis, allelic expression detection, and RNA editing identification due to their comprehensive transcript coverage and may outperform 3' sequencing methods in detecting lowly expressed genes. Moreover, certain scRNA-seq technologies, like SUPeR-seq and MATQ-seq, can capture both polyA⁺ and polyA⁻ RNAs, facilitating the sequencing of long noncoding RNAs (lncRNAs) and circular RNAs (circRNAs). This capability opens avenues for exploring the expression dynamics of both coding and noncoding RNAs at the single-cell level.

Despite their differences in technical details, the major steps of scRNA-seq remain the same, including isolation of the cells, cell lysis and RNA extraction, cDNA synthesis, PCR amplification, library construction, sequencing and quantification of gene expression measurements.

Finally, the single cell era propelled also the development of new bioinformatics methods and novel algorithms utilized in the analysis, visualization and interpretation of the data. Two reviews summarizing the different steps of analysis and the best practices in the field can be found online (Luecken and Theis 2019; Heumos et al. 2023), however a more thorough description of the available bioinformatics steps will follow in the section 1.8.

Methods	Transcript coverage	UMI possibility	Strand specific	References
Tang method	Nearly full-length	No	No	Tang et al., 2009
Quartz-Seq	Full-length	No	No	Sasagawa et al., 2013
SUPeR-seq	Full-length	No	No	Fan X. et al., 2015
Smart-seq	Full-length	No	No	Ramskold et al., 2012
Smart-seq2	Full-length	No	No	Picelli et al., 2013
MATQ-seq	Full-length	Yes	Yes	Sheng et al., 2017
STRT-seq and STRT/C1	5'-only	Yes	Yes	Islam et al., 2011, 2012
CEL-seq	3'-only	Yes	Yes	Hashimshony et al., 2012
CEL-seq2	3'-only	Yes	Yes	Hashimshony et al., 2016
MARS-seq	3'-only	Yes	Yes	Jaitin et al., 2014
CytoSeq	3'-only	Yes	Yes	Fan H.C. et al., 2015
Drop-seq	3'-only	Yes	Yes	Macosko et al., 2015
InDrop	3'-only	Yes	Yes	Klein et al., 2015
Chromium	3'-only	Yes	Yes	Zheng et al., 2017
SPLIT-seq	3'-only	Yes	Yes	Rosenberg et al., 2018
sci-RNA-seq	3'-only	Yes	Yes	Cao et al., 2017
Seq-Well	3'-only	Yes	Yes	Gierahn et al., 2017
DroNC-seq	3'-only	Yes	Yes	Habib et al., 2017
Quartz-Seq2	3'-only	Yes	Yes	Sasagawa et al., 2018

Figure 15. A table showing various scRNA-seq methods and their technical differences (Adopted from Chen et Al., 2019)

➤ Single cell ATAC sequencing

Assay for Transposase-Accessible Chromatin using sequencing at single-cell (scATAC-seq) has enabled the study of chromatin accessibility dynamics at an unprecedented resolution. More precisely, individual cells are isolated and subjected to the ATAC-seq protocol. Depending on the selected protocol, cells are sorted into individual wells of a microfluidic device or plates, where the steps of transposition, fragmentation,

and sequencing adapter ligation occur within each cell. More specifically, a transposase enzyme is utilized to efficiently cleave open chromatin DNA, while simultaneously specific sequences known as adapters are attached. The adapter-ligated DNA fragments are subsequently isolated, amplified via PCR, and prepared for next-generation sequencing. Analysis of sequencing data allows for the identification of regions with increased accessibility, indicating open chromatin. Additionally, it facilitates the mapping of transcription factor binding sites and the positioning of nucleosomes across the genome (Yan et al. 2020). Among the most common applications of scRNA-seq data analysis are the study of tumor heterogeneity across different cancer types, the reconstruction of gene regulatory relationships, lineage tracing and discovery of novel biomarkers.

➤ Spatially resolved transcriptomics

Spatially resolved transcriptomics techniques have become increasingly accessible in the past five years. These methods, whether used alone or in conjunction with other modalities like single-cell RNA sequencing (scRNA-seq), offer valuable insights into tissue architecture and cellular organization under both normal and diseased conditions. They allow for the characterization of transcriptional patterns and regulatory mechanisms in tissues, uncovering not only broad gene expression patterns but also subtle differences in tissue neighborhoods that may contribute to disease initiation or progression. A broad categorization (Williams et al. 2022) of spatial methods (Fig. 16) can be achieved by dividing them in:

1. Imaging-based methods, including in situ hybridization (ISH) and in situ sequencing (ISS), which offer visualization of mRNA molecules within the tissue of interest.
2. Sequencing-based methods, which extract mRNA and preserve at the same time spatial information for the upcoming next-generation sequencing (NGS).

In the one hand imaging-based methods rely on fluorescently labeled probes or direct sequencing of amplified mRNAs. On the other hand, sequencing-based methods preserve spatial information through either microdissection or the existence of spatially barcoded probes. Spatial methodologies are employed in a wide range of applications in many different research areas such as cancer, neuroscience, developmental biology, auto-immune diseases, and others. It is worth clarifying that some of them achieve single cell resolution e.g. CosMx platform, while others approach single cell resolution like 10x Visium, where 1-10 cells are captured in a single spot. Finally, most of the bioinformatics steps required for data analysis are almost identical to the ones implemented for scRNA-seq.

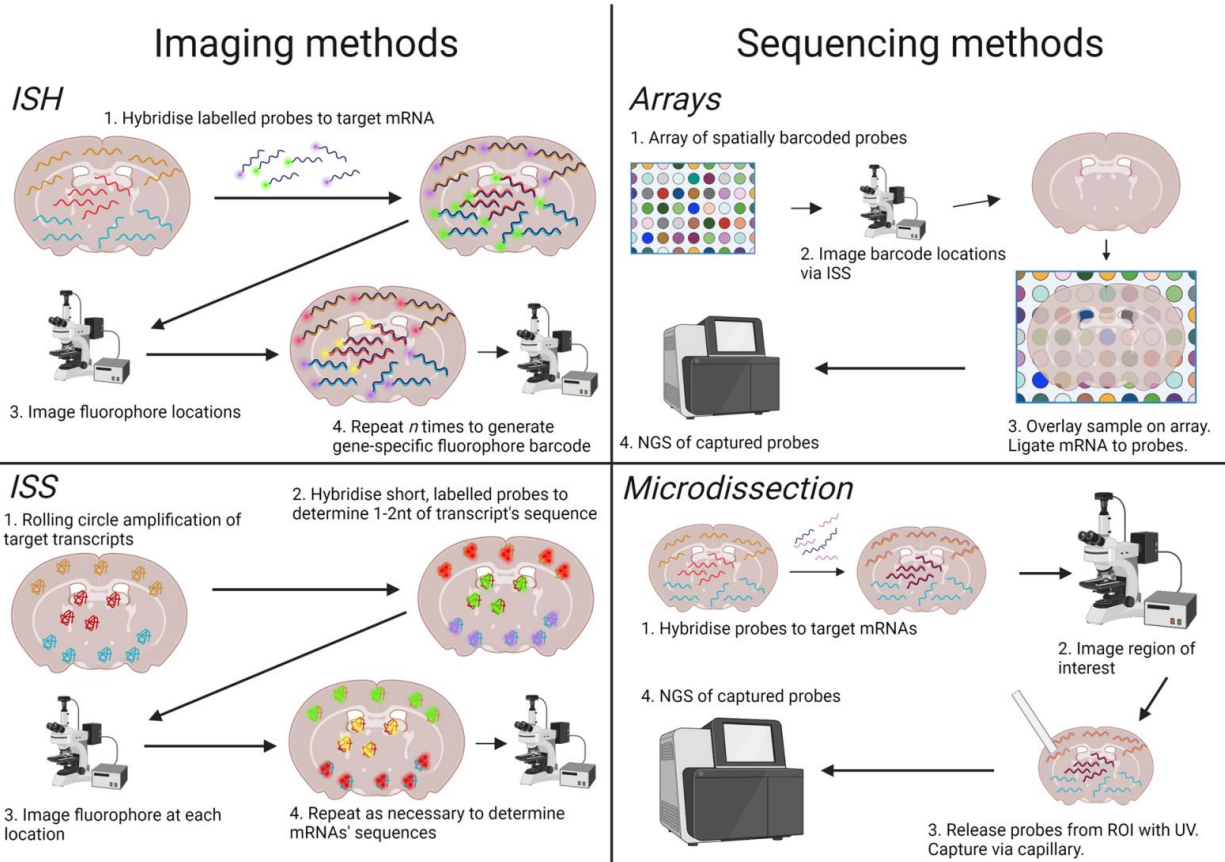


Figure 16. The two main categories of spatial transcriptomics methods are showcased (Adopted from Williams et Al., 2022)

➤ Multimodal single cell assays

In the previous paragraphs, we provided a concise overview of various single-cell assays, highlighting their advancements over their predecessors and outlining their main applications. Multimodal single-cell assays, by leveraging the principles underlying these methodologies, can enable the simultaneous measurement of multiple modalities within the same cell (Fig. 17). More specifically, parallel profiling of a cell's genome and transcriptome was achieved by G&T-seq and DR-seq followed by methods such as TARGET-seq and SIDR. Most of these methods enable the simultaneous studying of gene expression dynamics and possible mutations. Regarding the combination of RNA and open chromatin information more than twenty methods have been reported in the literature. Among them is the 10x multiome that enables epigenomics (chromatin accessibility), and transcriptomics (RNA) measurements, from the same single cell. Additionally, TEA-seq, a method for trimodal single-cell measurements permitting study of transcripts, epitopes and chromatin accessibility at the same time. Moreover, the pairing of transcriptomics and proteomics approaches led to the development of techniques such as CITE-seq or REAP-seq. In the first, high-throughput detection of protein markers is integrated with unbiased

transcriptome profiling for thousands of individual cells simultaneously. In the second, a unified workflow permits the quantification of surface proteins utilizing 82 antibodies and the concurrent genome wide mRNA analysis (Baysoy et al. 2023).

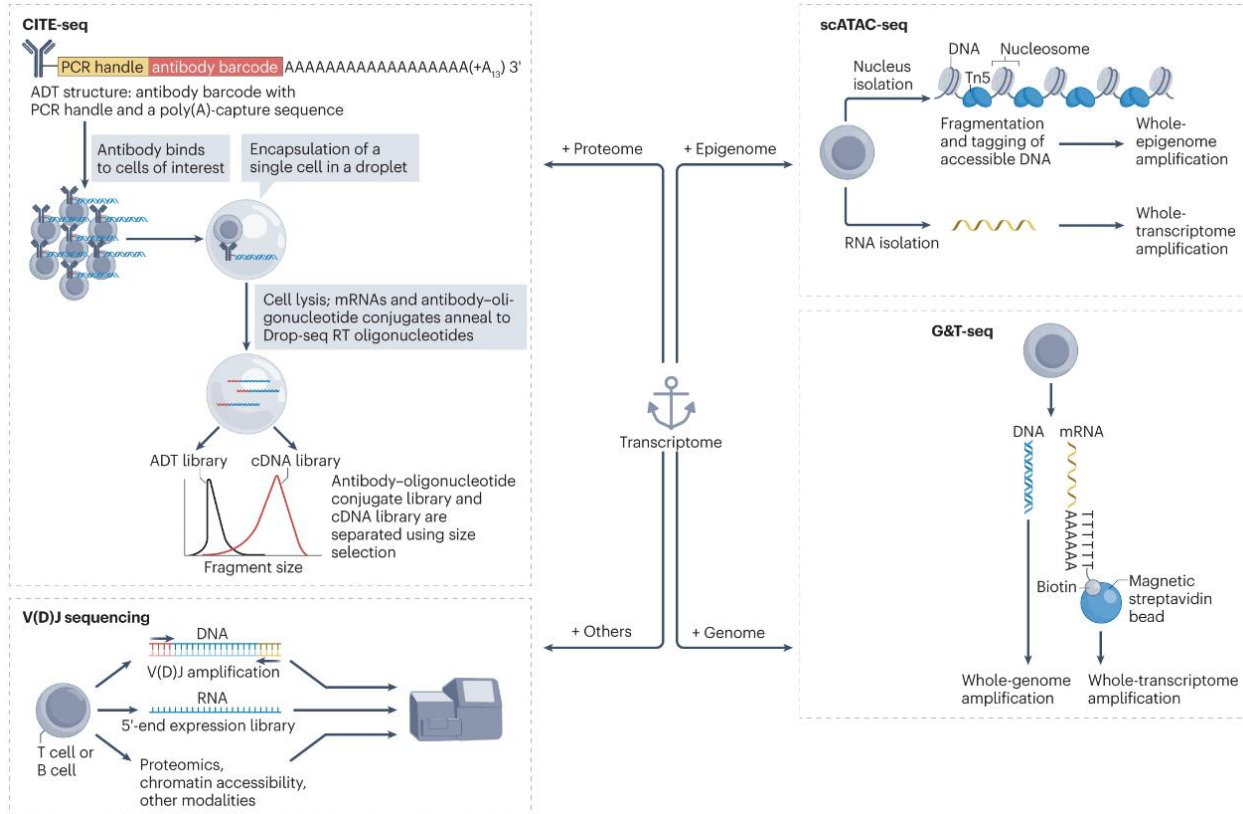


Figure 17. Different combinations of multiomics assays are presented (Adopted from Baysoy et Al., 2023).

1.7 Single cell application in biological systems

ScRNA-seq and scATAC-seq assays have been used in the study of many different biological systems and experimental settings in *Homo sapiens* and other organisms like *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans* and *Danio rerio* and many others. The capabilities of single cell technology have enabled shedding new light on different aspects of biology such as organogenesis, cell development, cancer, wound healing/tissue repair mechanisms and disease conditions (e.g. auto-immune diseases), to mention just a few.

In the past decade, numerous groundbreaking scientific publications have leveraged SC technology to provide new insights into various biological processes. These discoveries have enhanced our

understanding of various diseases and have also facilitated the application of novel therapeutic strategies. We will briefly mention some relevant publications in the following lines, however it's important to note that this is not an exhaustive list.

In 2019 (Packer et al. 2019) have compiled a map of embryonic cells from *C. elegans* consisting of 86,024 single cell transcriptomes. Additionally, (Sommarin et al. 2023) have characterized the transcriptional profiles of individual hematopoietic cells, unravelling the molecular cues involved in the emergence and maturation of hematopoietic stem cells during human fetal development. As regards lung cancer, in the publication of (D. He et al. 2021) the usage of SC transcriptomics enabled the examination of the cellular composition of early-stage lung adenocarcinomas harboring EGFR mutations (Fig. 18). The tumors studied included populations of both tumor cells and immune cells exhibiting heterogeneous properties. The analysis showcased diverse cellular subtypes of tumor cells with distinct gene expression profiles, highlighting the existence of intra-tumoral heterogeneity. Furthermore, different immune cell populations were identified within the tumor microenvironment. Overall, the findings emphasized the underlying complexity of early-stage lung adenocarcinomas. Another interesting publication in 2022 (Sinha et al. 2022) stratified SC transcriptomics in order to decipher the role of fibroblasts subsets in reindeer upon injury (Fig. 19). More specifically velvet fibroblasts, found in the antlers of the deer, enable regeneration of the injury site by adopting an immunosuppressive phenotype that accelerates resolution. On the contrary, skin fibroblasts found in the back of the animals (resembling fibroblast subtypes in human and mice) express inflammatory molecules and promote leukocyte infiltration, leading to difficulties in completion of the repair process. Although, the transplantation of velvet fibroblasts to scar-forming back skin initially enables regeneration in the injury site, eventually leads to fibrosis, resembling the fetal-to-scar-forming transition that is also observed in humans. Conclusively, the study proposes reindeer as a valuable model for studying wound healing and suggests that the of targeting fibroblast-immune interactions could be proved beneficial to mitigate scarring in humans. Finally, significant progress has been achieved also in the sector of various disease conditions, where publications utilizing SC have enabled the study of novel aspects in human patients and disease models or organoids. A prominent example is the work of (Martin et al. 2019), in which a cell module consisting of IgG plasma cells, inflammatory mononuclear phagocytes, activated T cells, endothelial cells and fibroblasts (called GIMATS) proved to be predictive of anti-TNF response in IBD patients (Fig. 20). On the other hand, anti-TNF non-responders exhibited enriched interaction of IL1 (produced from inflammatory macrophages) and IL-1R (expressed by fibroblasts).

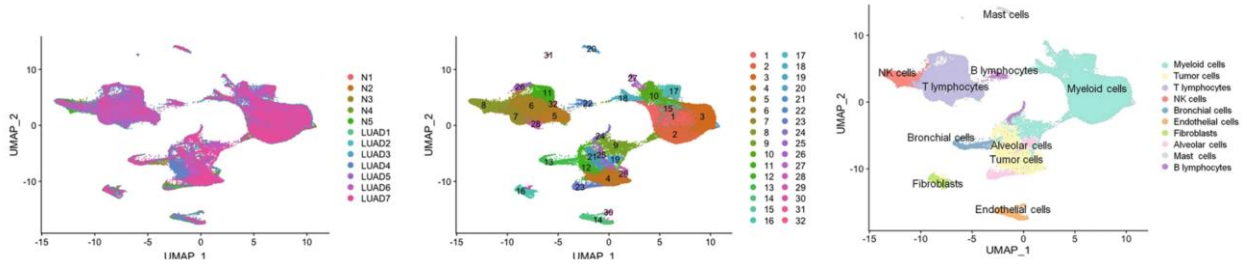


Figure 18. UMAP plot of 125,674 cells originated from Lung adenocarcinoma and healthy samples (Adopted from He et Al., 2021).

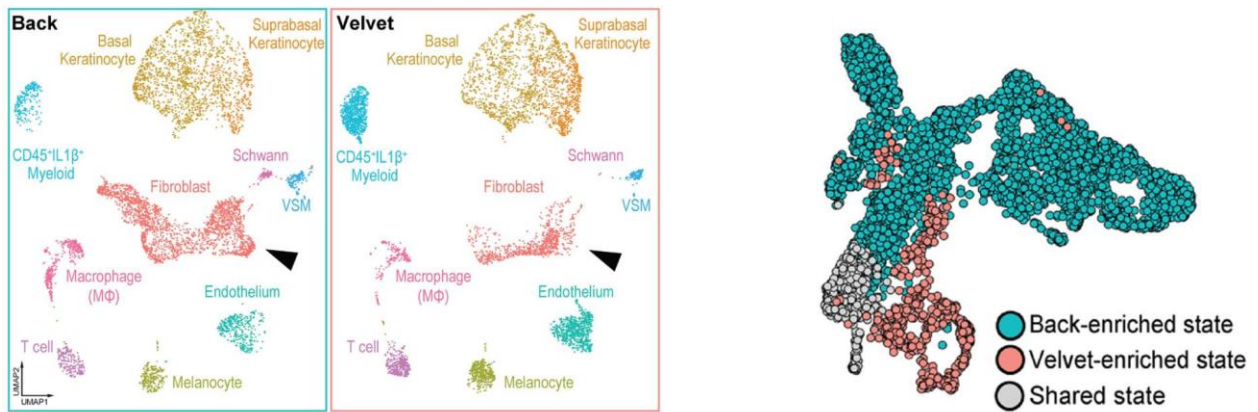


Figure 19. UMAP plot depicting the different cell types in wound healing skin of reindeer (Adopted from Sinha et Al., 2022).

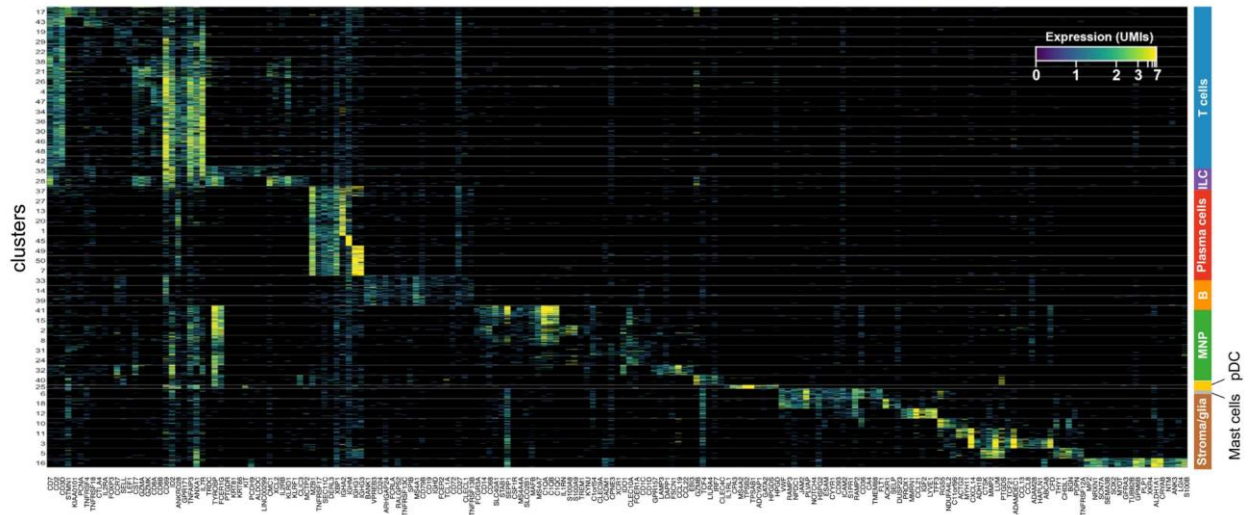


Figure 20. Heatmap depicting gene expression patterns across different clusters in patients with Crohn's disease (Adopted from Martin et Al., 2019).

The coordinated efforts of large consortia were crucial for the development of public databases containing vast numbers of SC datasets from different tissues. The Human Cell Atlas, the Tabula Muris and the Fly Cell Atlas belong among the most distinguished resources in single-cell community. More specifically, the Human Cell Atlas takes advantage of scRNA-seq, imaging technologies and computational approaches to reconstruct maps of human cells from all different tissues. The uploaded datasets are organized in 18 broad categories (Adipose, Gut, Lung, Pancreas, Breast, Heart, Musculoskeletal, Reproduction, Development, Immune, Nervous System, Skin, Eye, Kidney, Oral & Craniofacial, Genetic diversity, Liver, Organoid) and are consisted of ~ 59 million cells in total originated from ~ 8.6 thousands of donors. Regarding its mouse counterpart, known as Tabula Muris, it is a compendium of scRNA-seq datasets encompassing ~ 100,000 cells from 20 different organs or tissues. Interestingly, the Fly Cell Atlas project employs SC genomics, transcriptomics and epigenomics methodologies in order to construct a set of cellular atlases representing distinct developmental or disease states of drosophila.

In parallel, new repositories such as CellPortal (Tarhan et al. 2023) and CellxGene (Chan Zuckerberg Initiative, n.d.) have been created to facilitate collection and easy access in processed data that are publicly available through scientific publications. Both support online exploration of the uploaded SC datasets, as well as downloading of the data in well-established formats, which can allow further processing by software packages in R or python programming languages.

1.8 Computational methodologies for single cell data analysis

Since the emergence of single cell technology, thousands of software applications have been developed for data analysis and visualization purposes (Fig. 21). Databases such as scTools (Zappia and Theis 2021) provide extensive catalogues of those tools accompanied by additional information including links to the source code, manuscripts, vignettes, etc. In a previous section, several single cell techniques were described. However, In the context of the current dissertation, we will focus only on methodologies and software packages utilized for the analysis and representation of scRNA-seq and scATAC-seq data. Most of the steps implemented are common in both assays, however there are also analytical tasks specific for each modality. It is worth noting that the preliminary steps of analysis, including alignment to reference genome and quantification of gene expression, are usually performed by sequencing facilities or by utilizing software packages tailored to the sequencing platform and the experimental protocol that was selected. For these reasons those steps are not discussed in the following sections.

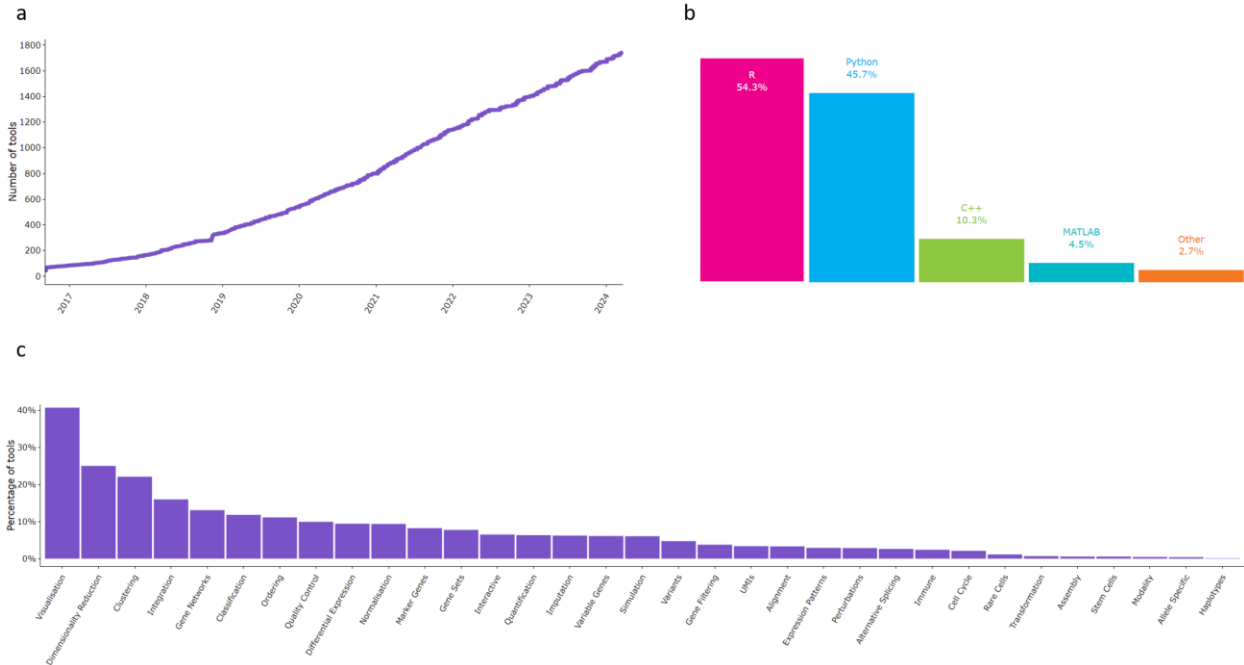


Figure 21. (a) Number of tools developed for the analysis of scRNA-seq data. (b) Programming languages utilized for the development of the tools. (c) Analytical tasks implemented and included in the different available tools (Adopted from <https://www.scrna-tools.org/>).

Regarding scRNA-seq data, the first step of the analysis is devoted to quality control ensuring that low quality cells are filtered out. After that, normalization of the data is performed to mitigate differences in the sequencing depth between the different cells. Next the most highly variable genes of the dataset are detected. Following this step principal component analysis (PCA), a well-established linear method of dimensionality reduction, is employed. The most informative principal components (PCs) derived from the PCA analysis are then used as input in non-linear dimensionality reduction algorithms like t-Distributed Stochastic Neighbor Embedding (tSNE) (Maaten and Hinton 2008) or uniform manifold approximation and projection (UMAP) (Becht et al. 2018) to produce new embeddings for the visualization of the cells in 2D or 3D space. Additionally, PCs are utilized to construct a shared nearest neighbor graph (SNN) of the cells. Since the primary objective of the main analysis is the detection of cells populations, graph-based clustering with louvain or leiden algorithm is applied. The clusters originating from the previous step are subjected to marker gene analysis, which enables the identification of genes that exhibit preferential expression in specific clusters.

In the case of the scATAC-seq data, the analysis begins with quality control as well. However, normalization, detection of highly variable genes and PCA are replaced by the latent semantic indexing (LSI) dimensionality reduction method. Non-linear dimensionality reduction methods e.g. tSNE and UMAP are

utilized for cell visualization purposes, followed by clustering and marker gene detection. Some additional steps available in the ATAC modality are the detection of marker peaks, that enables the search of peaks that show high accessibility in specific clusters and the motif enrichment analysis, which can recognize enriched TF-specific binding sites on the marker peaks.

Furthermore, additional modes of analysis can be used to perform more advanced analysis tasks containing functional enrichment analysis of the clusters, cell-cycle phase analysis, trajectory inference, automatic cell type annotation, Gene Regulatory Network (GRN) reconstruction and cell-cell communication analysis. Most of the previously mentioned algorithms are mainly developed in R and Python programming languages. Various software applications and packages have been introduced to execute these tasks. Notable ones include Seurat (Stuart et al. 2019), Scanpy (Wolf, Angerer, and Theis 2018), Monocle (Trapnell et al. 2014; Qiu et al. 2017), scater (McCarthy et al. 2017), slingshot (Street et al. 2018), scvelo (Bergen et al. 2020), SCENIC (Aibar et al. 2017), decoupleR (Badia-I-Mompel et al. 2022), cellphoneDB (Efremova et al. 2020), nichenetR (Browaeys, Saelens, and Saeys 2020), cellchat (Jin et al. 2021), singleR (Aran et al. 2019), CIPR (Ekiz et al. 2020b), Cicero (Pliner et al. 2018), Signac (Stuart et al. 2021), EpiScanpy (Danese et al. 2021), cisTopic (Bravo González-Blas et al. 2019), and ArchR (Granja et al. 2021b), which are some of the most widely used R and Python libraries. Seurat and Scanpy are primarily employed for analyzing single-cell RNA sequencing (scRNA-seq) data, offering functionalities ranging from QC to population identification and integration of multiple datasets. Signac and EpiScanpy extend these functionalities to process single-cell ATAC sequencing (scATAC-seq) data. ArchR focuses on analyzing single-cell chromatin accessibility data, offering standard analysis steps and advanced features like Positive Regulator identification, (TF) footprinting, and trajectory inference. Monocle provides a widely used pseudo-temporal cell ordering framework for scRNA-seq analysis, while Cicero extends it for scATAC-seq analysis. Scater focuses mainly on the initial quality control (QC) of the data, while SCENIC and decoupleR are utilized in GRN analysis. Slingshot and scVelo are specialized in the task of trajectory inference. CellphoneDB, nichenetR and CellChat are well established tools for the analysis of cell-cell communication interactions. Finally, SingleR and CIPR are two packages that perform automatic cell type annotation on the identified clusters.

Moreover, there are software applications also offering a Graphical User Interface (GUI) such as Scope, CZ CELLxGENE (Chan Zuckerberg Initiative, n.d.), Azimuth (Hao et al. 2021), Cerebro (Hillje, Pelicci, and Luzi 2020), iCellR (K. H. Tang et al. 2022), ICARUS (Jiang et al. 2022) and SeuratWizard (Yousif et al. 2020). The existence of a GUI, which in many cases is also accompanied by a web service, facilitates the engagement

of scientists with no prior computational experience in the analysis of their datasets. In terms of GUI tools, Scope offers various visualization options, including comparative views at cluster and gene levels for datasets containing multiple samples or conditions, although it lacks further downstream data analysis support. CZ CELLxGENE facilitates exploration of single-cell datasets and gene expression visualization across tissues in published datasets but lacks complex analytical capabilities. Azimuth specializes in basic scRNA-seq analysis steps and characterizing identified populations using a 'reference-based mapping' approach but lacks customization options. SeuratWizard follows standard analysis steps, while Cerebro expands upon them, offering additional modes such as signature scoring, cell cycle phase analysis, and trajectory inference. iCellR covers basic analyses for both scRNA-seq and scATAC-seq but does not include ligand-receptor (L-R) and GRN reconstruction functionalities. ICARUS performs all the aforementioned modes of analysis and at the same time offers a lightweight implementation of GRN analysis with SCENIC as well as cell-cell communication analysis mode based upon CellChat.

Advances in the field of machine learning (ML) and artificial intelligence (AI) have led to the development of software packages leveraging those approaches to perform SC analytical tasks. Neural Networks and Variational Autoencoders are two ML methodologies that are widely used for clustering, dimensionality reduction, annotation of cells as well as integration of SC datasets or SC modalities. Examples of developed tools based on ML principles are scAce (X. He et al. 2023), Midas (Z. He et al. 2024) and SUPREME (Kesimoglu and Bozdog 2023). Following the increasing interest of researchers on the AI based chatbots such as ChatGPT (Meyer et al. 2023), web-based AI tools performing data interpretation or data analysis have emerged. In the first category we can find tools like BioChatter, while in the second tools like scGPT (Cui et al. 2024).

Another interesting category of utility packages contains tools such as dittoSeq (Bunis et al. 2021), Scillus ("GitHub - Xmc811/Scillus: R Package for Single-Cell Dataset Processing and Visualization" n.d.), scPubR (Blanco-Carmona 2022) and scCustomize ("Samuel-Marsh/ScCustomize: Version 2.1.2" n.d.). These tools utilize already analyzed single cell datasets to offer enhanced visualization of cells in 2D/3D space and gene expression patterns. In more detail, scatterplots of single cells, violin plots, heatmaps, dotplots and swarm plots are implemented building upon basic packages of R or python such as ggplot2 and plotly (Fig. 22).

Concluding this introductory part, it is noteworthy to highlight that several computational methods discussed earlier have led to the development of publicly accessible databases. These databases can be employed for various analytical purposes, either similar or distinct from those described already.

Regarding regulatory interactions, two notable online databases, Dorothea and CollecTRI (Müller-Dott et al. 2023), were compiled by integrating interactions between transcription factors (TFs) and their target genes. These databases are valuable resources for inferring transcription factor activity. Moreover, cellphoneDB and cellchatDB are repositories containing curated L-R pairs that document autocrine and paracrine interactions among different cell types in humans and mice.

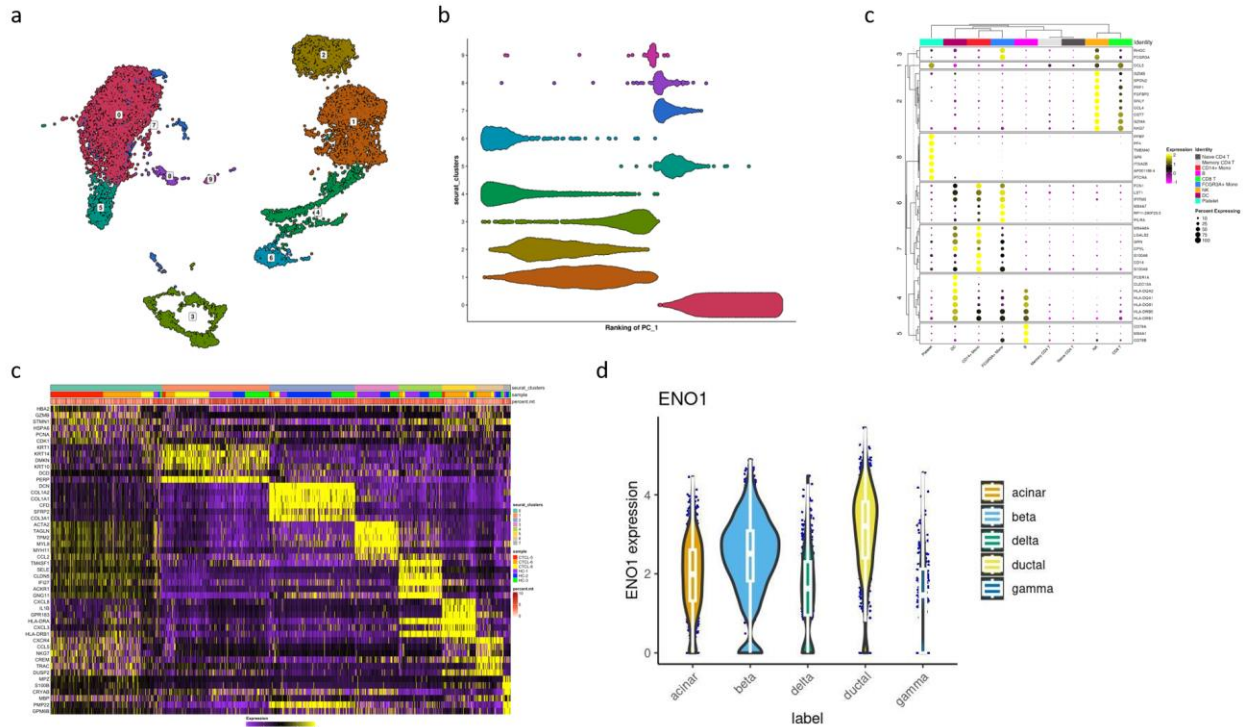


Figure 22. Various plots used in scRNA-seq data visualization including: (a) UMAP plot (b) Swarm plot (c) Dotplot (d) Heatmap and (e) Violinplot.

2 Material and methods

2.1 Implementation of a web-based application for SC data analysis

Building upon the intricate challenges previously discussed in SC data analysis, we embarked on the development of an interactive web application. This platform provides automated analysis, visualization, and exploration functionalities for both scRNA-seq and scATAC-seq datasets. Our approach involved integrating a range of software technologies including R/Shiny, HTML, JavaScript, and CSS, resulting in the establishment of a robust pipeline capable of executing diverse analytical tasks on single-cell data (Fig. 23).

Central to our design is a user-friendly graphical interface (GUI), tailored to accommodate researchers without prior programming experience. More specifically, the R/Shiny framework seamlessly merges the capabilities of R programming language with web technologies like HTML, CSS, and JavaScript. This integration empowers the creation of an interactive environment featuring dynamic plots, data tables, widgets, buttons, and comprehensive instructions to aid users in navigating their preferred analysis methodologies.

To provide a variety of analysis options for both assays, we carefully combined numerous state-of-the-art software packages and facilitated their seamless integration, resulting in the creation of a comprehensive pipeline, which was named SCALA (**S**ingle **C**ell **A**na**L**ysis for **A**ll). Moreover, to ensure reproducibility, we have made the code of the developed application accessible through a public GitHub repository. Finally, extensive segments from Material & Methods, Results and Discussion sections of the current dissertation, have also been published (in their current form or with slight modifications) in two research articles. One of them focuses on the detailed characterization of gene expression and chromatin accessibility profiles of *hTNFtg* mouse model at single cell resolution (Marietta Armaka et al. 2022) while the second is oriented in delineating the capabilities of our application and illustrating its functionality across different use case scenarios (Tzaferis et al. 2023).

2.2 Data input

SCALA supports various input data types. For scRNA-seq analysis, the primary input data is a unique molecular identifier (UMI) count matrix. Users can provide this matrix by either uploading a gene-by-cell

tab-delimited text file (where rows represent features and columns represent barcodes), including both row and column names, or by uploading the output of the 10X cellranger pipeline located at “filtered_bc_matrix” folder. In the latter case, the “cellranger count” output folder should include three files: “barcodes.tsv.gz” containing detected cellular barcodes in gzip CSV format, “features.tsv.gz” with features (genes) corresponding to row indices in gzip TSV format, and a feature-barcode count matrix in gzip Market Exchange Format (MEX). Additionally, users have the option to load a pre-analyzed Seurat object in RDS (R saved object) format. In the latter case, the condensed RDS format can allow uploads with hundred thousand cells, something that would be extremely difficult in the case of txt count matrix.

For scATAC-seq analysis, SCALA currently only accepts arrow files. This file format stores all associated data, including metadata, accessible fragments, and data matrices of a sample. Users can create arrow files using the provided “create_arrow_file.R” helper script from SCALA's GitHub repository or directly with the ArchR package. It's important to state that the analysis of human and mouse datasets is supported in both modalities.

2.3 Workflow description

Once the input files have been loaded, SCALA's primary workflow can be applied to both single-cell pipelines. The workflow includes the following steps: (i) Quality Control (QC), (ii) data normalization and scaling, (iii) detection of variable features, (iv) dimensionality reduction using Principal Component Analysis (PCA), (v) dimensionality reduction using Latent Semantic Indexing (LSI), (vi) clustering, (vii) additional dimensionality reduction methods, (viii) inspection of features, (ix) identification of markers, (x) analysis of cell cycle phases, (xi) functional/motif enrichment analysis, (xii) annotation of clusters, (xiii) trajectory analysis, (xiv) analysis of Ligand-Receptor (L-R) interactions, (xv) analysis of Gene Regulatory Networks (GRNs), and (xvi) visualization of epigenome signal tracks.

2.4 Quality control

In single-cell datasets, identifying and discarding “low quality” cells (such as empty, stressed, broken, or dead cells) and non-informative genes is crucial for downstream analysis. The developed application facilitates this process by allowing users to explore quality control (QC) plots and filter out cell barcodes based on user-defined thresholds. Common QC criteria for scRNA-seq include: (i) the number of unique

features detected per cell, (ii) the number of detected UMIs per cell, and (iii) the percentage of mitochondrial content per cell. Cells containing low numbers of unique features and UMIs are typically excluded as low-quality, while those with very high numbers may indicate capturing RNA material from multiple cells. Cells displaying a high percentage of mitochondrial UMIs are also flagged as low-quality or potentially dying cells (Fig. 24).

For scATAC-seq, typical QC metrics include: (i) transcription start site (TSS) enrichment and (ii) the number of unique nuclear fragments in logarithmic scale ($\log_{10}(nFrag)$). In most cell types, there is usually a notable enrichment of ATAC-seq signal in the transcription start site (TSS) regions of actively expressed genes, serving as a traditional indicator of the quality of the assay. For the calculation of the first metric a comparison between the enrichment of ATAC-seq signal at TSS regions to the enrichment observed in flanking regions, extending 2 kilobases (kb) away from the TSS, is performed. Additionally, regarding the second metric, cells with too few nuclear fragments should be discarded to prevent the inclusion of non-interpretable data.

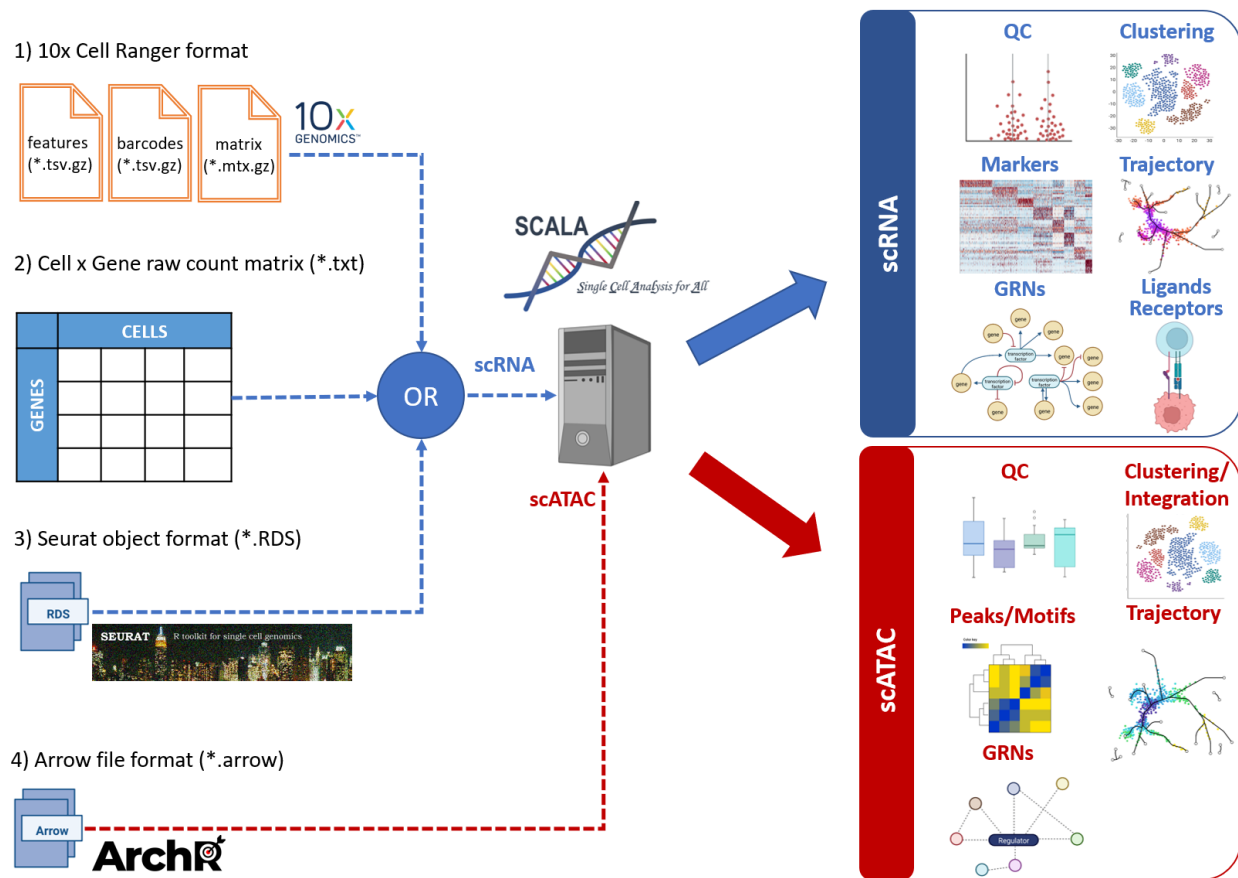


Figure 23. A schematic workflow of the developed application called SCALA (Adopted from Tzaferis et Al., 2023).

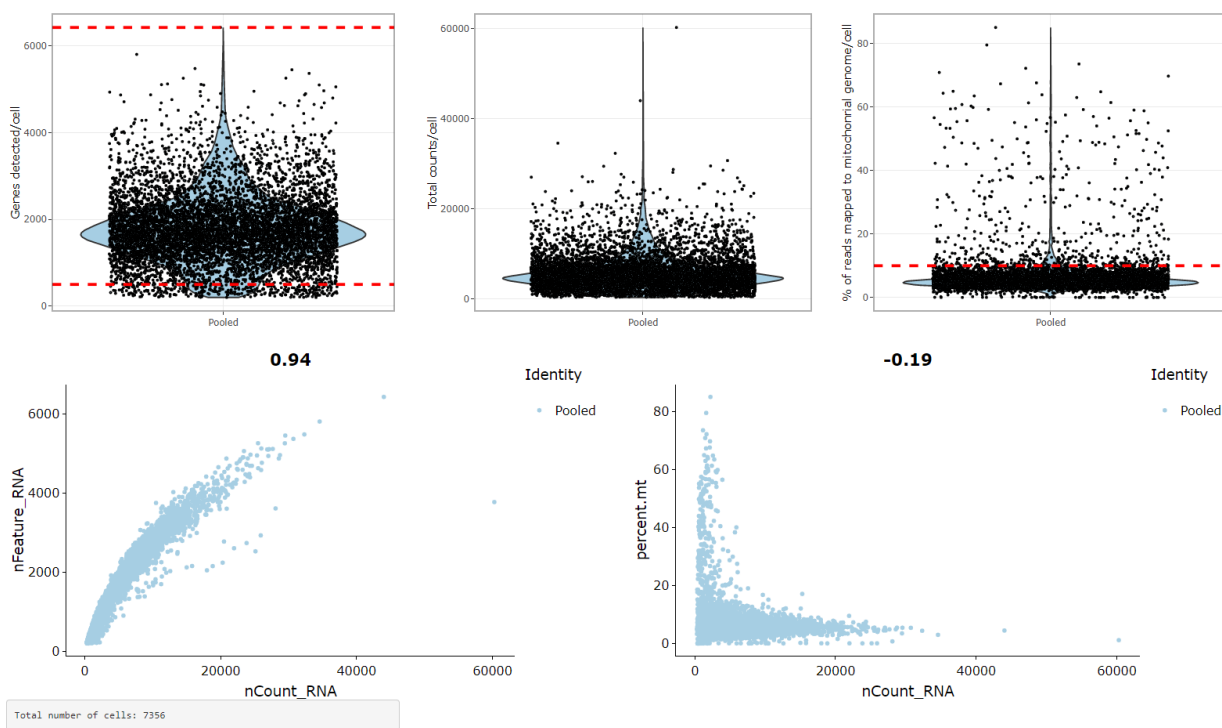


Figure 24. A set of scatter plots and violin plots utilized to guide quality control and cell filtering procedures.

2.5 Normalization and scaling of the data

Normalization and scaling of scRNA matrices are crucial steps that focus on mitigating biases originating mainly from differences in cell depth and ensuring proper transformation of the data before subsequent analyses such as variable feature detection and dimensionality reduction. In our implementation, data normalization follows a global-scaling approach (Hao et al. 2021), wherein the gene count for each barcode is normalized by the total barcode counts, multiplied by 10,000, and subjected to logarithmic transformation. These normalized values are stored within a matrix data structure, and are further standardized to z-scores, ensuring a column-wise mean expression of 0 and a variance of 1. Moreover, to address additional unwanted sources of variation, users have the option to specify metadata variables as covariates. In such cases, these variables are regressed against each feature, followed again by scaling and centering of the resulting residuals.

2.6 Detection of highly variable genes

During this step, the normalized RNA data matrix is utilized to identify genes that exhibit the highest variation among cells. This subset of features is crucial for uncovering the underlying biological patterns within single-cell datasets in a computationally efficient manner, reducing the initial dimensions of the matrix to less than 3,000 features. Three methods are supported for detecting the most variable features including “Variance Stabilizing Transformation” (VST), “Mean-Variance Plot” selection (MVP), and “Dispersion”.

VST (Fig. 25) involves fitting a line to the log-variance/log-mean relationship using local polynomial regression. Subsequently, feature values are standardized based on the observed mean and expected variance, with feature variance then calculated for standardized values. This procedure typically returns a fixed number of variable features (usually set at 2,000 by default).

MVP calculates average gene counts and gene dispersions using a designated function. Specifically, genes are divided into 20 bins based on their average read counts, and dispersion z-scores are computed for each gene group.

Finally, for the “Dispersion” method, genes with the highest dispersion values are retained. Both MVP and “Dispersion” methods return a variable number of features not determined by the user.

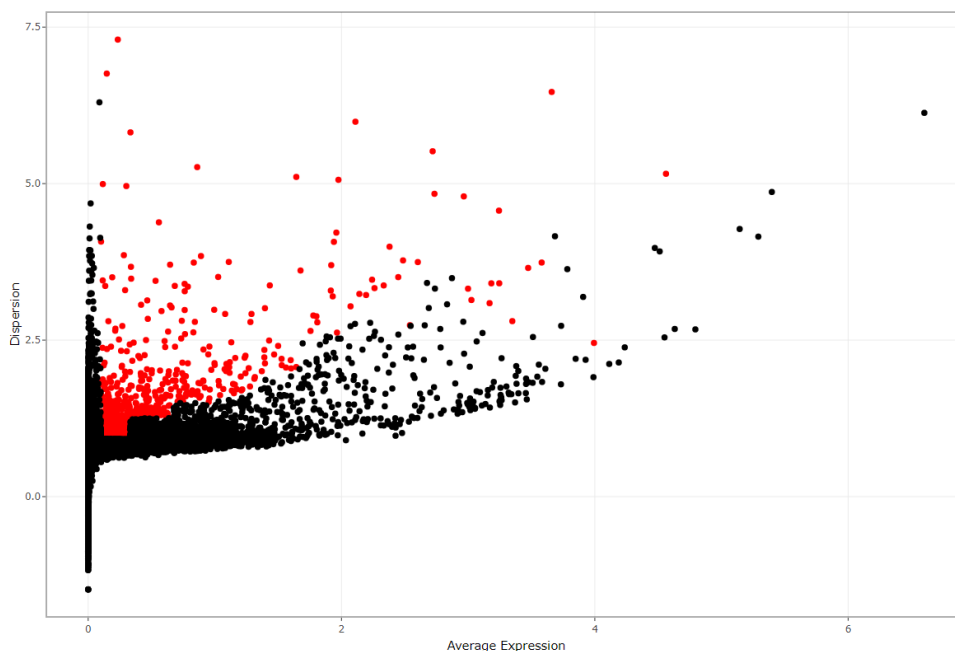


Figure 25. Scatter plot of genes in the single cell dataset. Most highly variable genes are depicted in red color.

2.7 Principal Component Analysis

Principal Component Analysis (PCA) is a linear dimensionality reduction technique applied to the scaled values of the most variable features, resulting in the computation of 'meta-genes,' which are linear combinations of genes within the assay. The most informative Principal Components (PCs) are subsequently identified and employed in downstream steps such as cell clustering and cluster visualization (often using non-linear dimensionality reduction methods). Determining the optimal number of PCs exhibiting the highest variation in the scRNA matrix can be achieved either automatically through a 10-fold Singular Value Decomposition (SVD) cross-validation process or manually by inspecting the incremental variance ranking of each PC (via an elbow plot). For large datasets automatic calculation is not advised, as it could take hours to be completed.

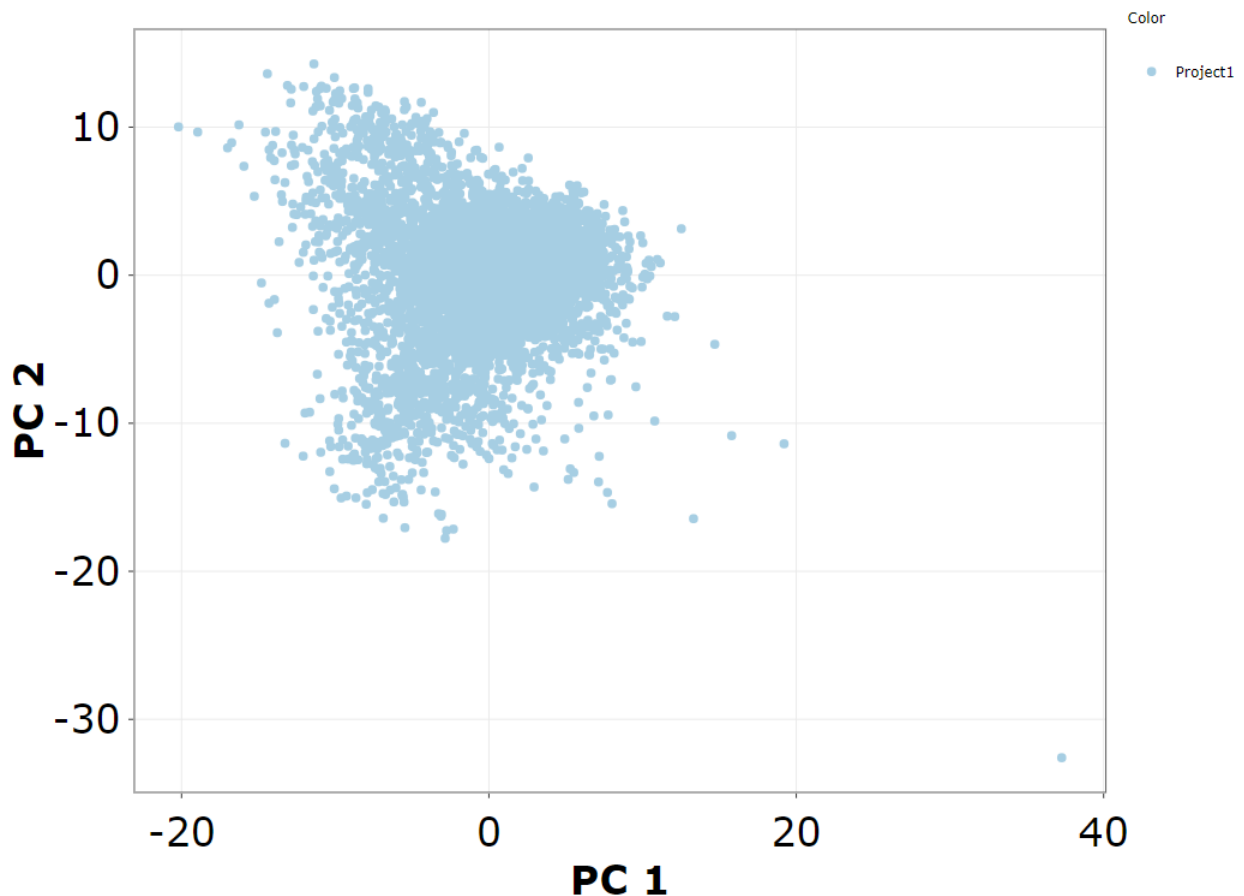


Figure 26. Cells depicted in PCA space after PCA analysis.

2.8 Latent Semantic Indexing

In scATAC-seq matrices, the methodology of Latent Semantic Indexing (LSI) is applied using genome-wide 500 base pairs (bp) tile counts (Granja et al. 2021). Initially, tile-counts undergo normalization to mitigate cell depth bias, utilizing a constant of 10,000, followed by inverse document frequency normalization and log-transformation. Throughout this process, the most variable features (in our case tiles) are discerned. LSI transformation is applied in an iterative manner using the most accessible features (tiles), thereby uncovering lower resolution clusters that are free from batch confounding factors. Subsequently, the average accessibility for each of these clusters is computed across all features. Finally, the most variable features across low-resolution clusters are identified and used as input for the next LSI iteration.

2.9 Clustering

Graph-based clustering is executed on scRNA-seq and scATAC-seq data to delineate cell types and/or cellular states. Initially, a Shared-Nearest Neighbor (SNN) graph structure of the cells is constructed using Euclidean distances in the PCA/LSI space. Cells sharing similar gene expression/chromatin accessibility profiles are connected by edges. Subsequently, the graph is partitioned into densely connected communities utilizing the Louvain algorithm (Blondel et al. 2008).

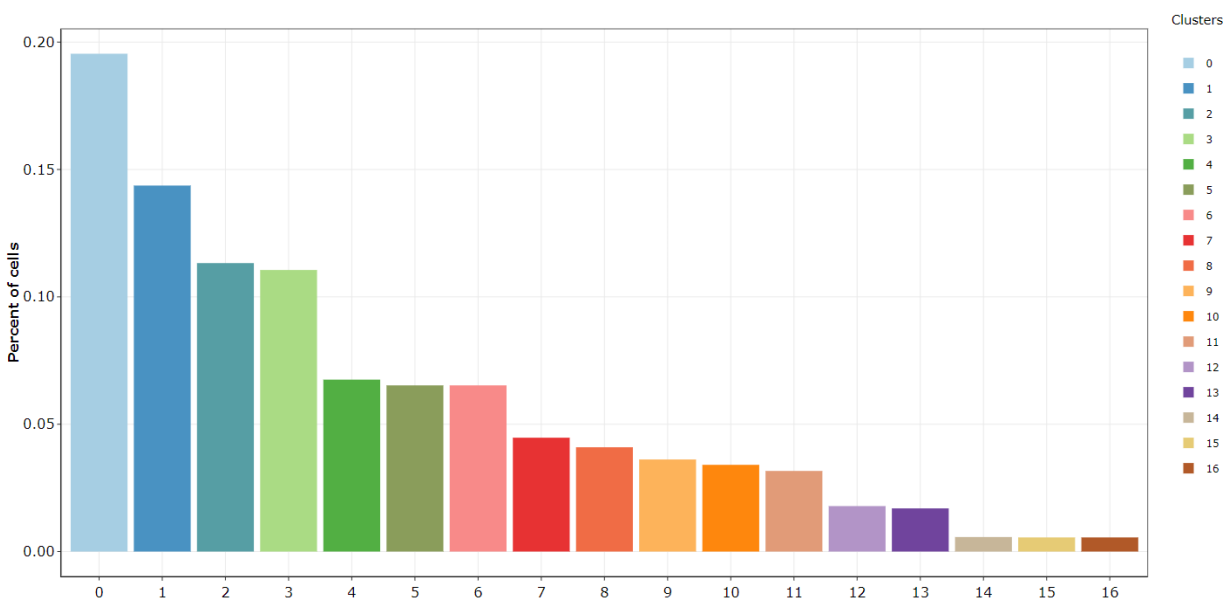


Figure 27. Barplot summarizing the results of Louvain clustering procedure.

2.10 Non-linear dimensionality reduction methods

To enhance the visualization of cells, cell clusters, and their relationships in both 2D and 3D space, a variety of nonlinear dimensionality reduction techniques are employed. Traditional linear methods like PCA or multi-dimensional scaling (MDS) may fail to capture complex patterns effectively, prompting the utilization of alternative methodologies such as UMAP, tSNE, diffusion maps (Haghverdi, Buettner, and Theis 2015), and Potential of Heat-diffusion for Affinity-based Trajectory Embedding (PHATE) (van Dijk et al. 2018). These methods play a crucial role in unraveling the underlying structure of the datasets while facilitating feature inspection, exploration of cluster structures, and trajectory inference. More specifically, UMAP and tSNE plots' inspection can serve as qualitative criterion for evaluating clustering success (Fig. 28). Additionally, they are very useful in exploring gene expression patterns of single genes or gene signatures. On the other hand, plots generated by PHATE and Destiny packages (e.g. diffusion maps) are valuable for exploring trajectory dynamics and lineage relationships between different cell populations.

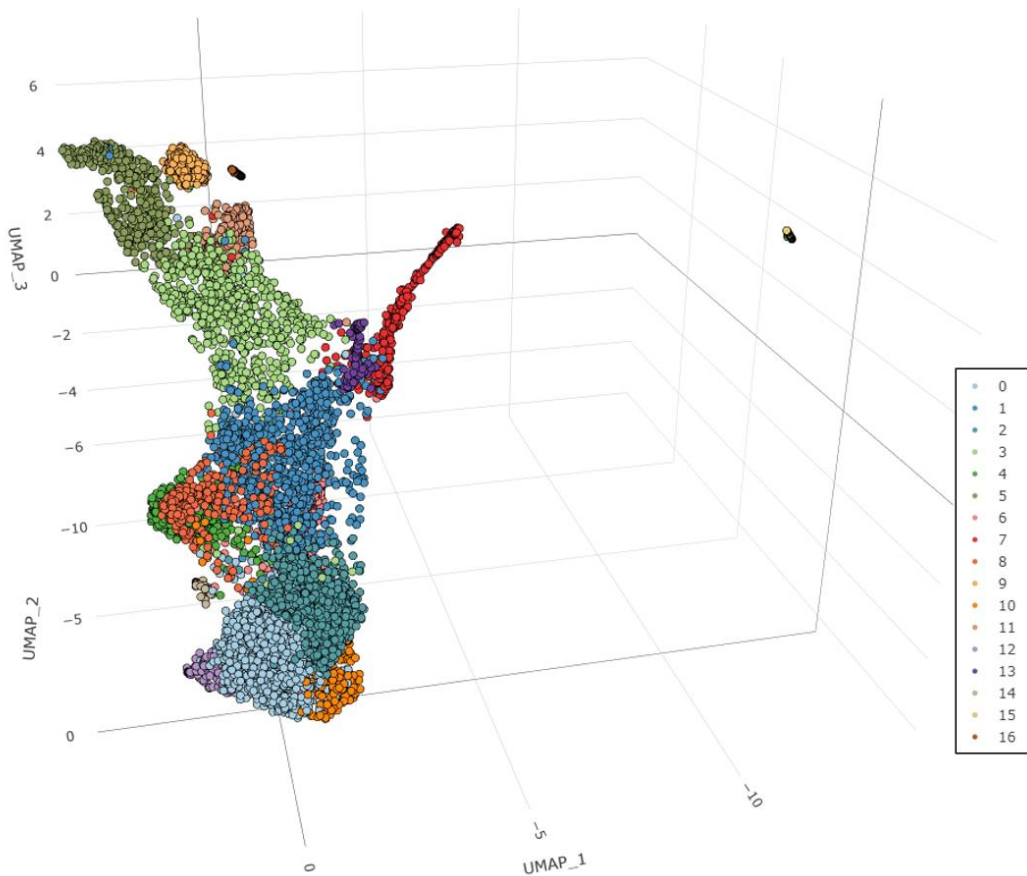


Figure 28. A 3D plot depicting cells in UMAP space.

2.11 Identification of marker genes

Differential expression analysis, as well as differential accessibility analysis, facilitate the identification of marker genes and marker peaks respectively, having a major contribution in the proper annotation and characterization of cell clusters based on already known cell type markers from the literature. This analytical approach aids in pinpointing crucial transcriptional and regulatory programs driving different biological processes such as development, progression of a disease, etc. The analysis is conducted in a cluster-specific manner. In more detail, cells within each cluster are compared against the rest of the cells in the dataset. Plenty of statistical tests are available for scRNA-seq analysis, including the Wilcoxon rank sum test, likelihood-ratio test for single-cell feature expression (McDavid et al. 2013), standard Area Under the Curve (AUC) classifier, Student's t-test, MAST (Finak et al. 2015), and DESeq2. Similarly, for scATAC-seq, available tests contain the Wilcoxon rank sum test, Student's t-test, and binomial test.

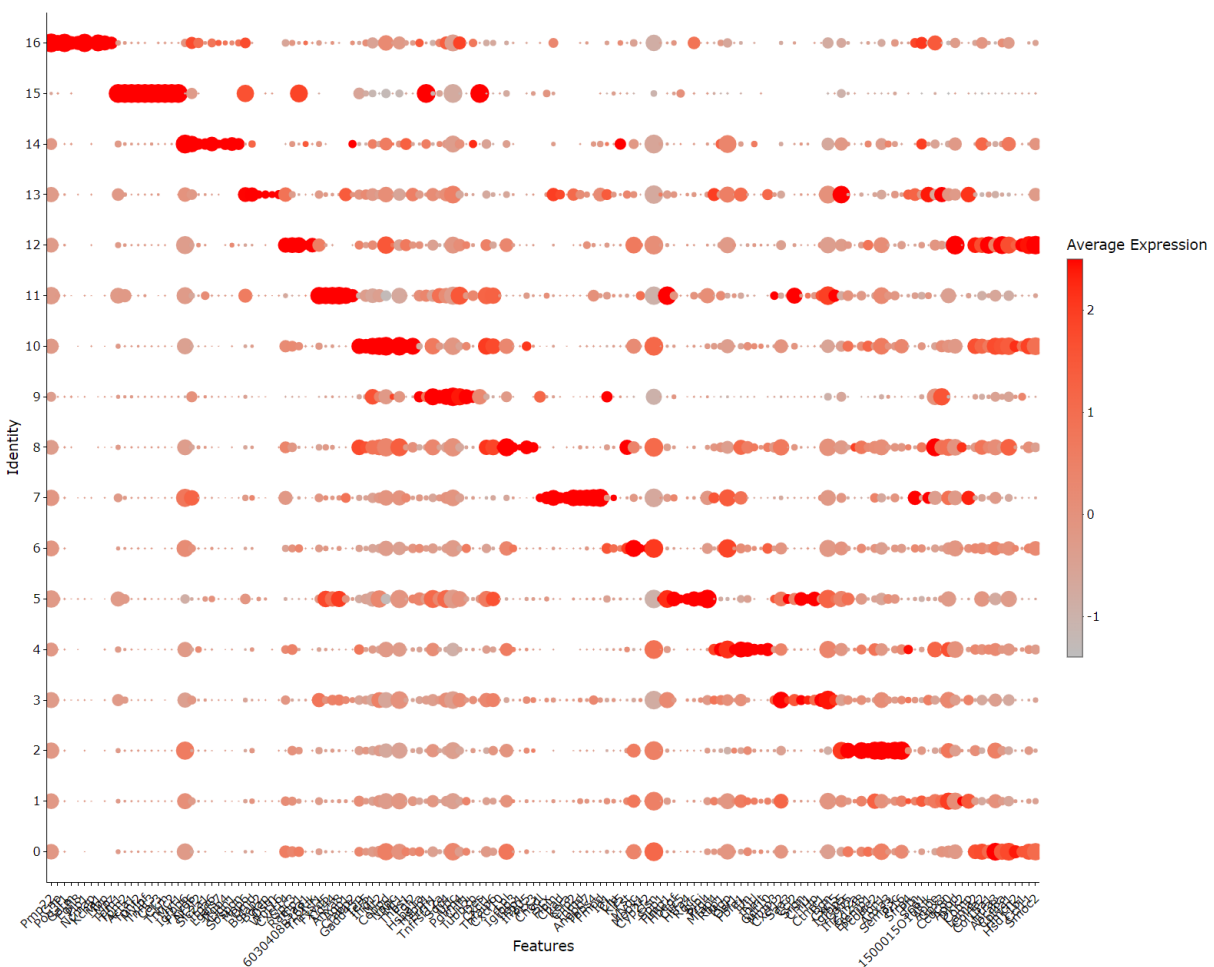


Figure 29. Dotplot showing normalized expression of top marker genes per cluster.

2.12 Inspection of features

Exploration of feature expression and chromatin activity can be conducted through inspection of cell scatter plots in reduced dimensional space (e.g., UMAP, tSNE, etc.), or via dotplots (Fig. 29), heatmaps, and violin plots (Fig. 30). In scRNA-seq datasets, gene signatures can additionally be computed using the UCell package and visualized as outlined above. Furthermore, quality control metrics such as the total number of reads per cell and genes detected per cell can be visualized using scatter plots and violin plots at a cluster level.

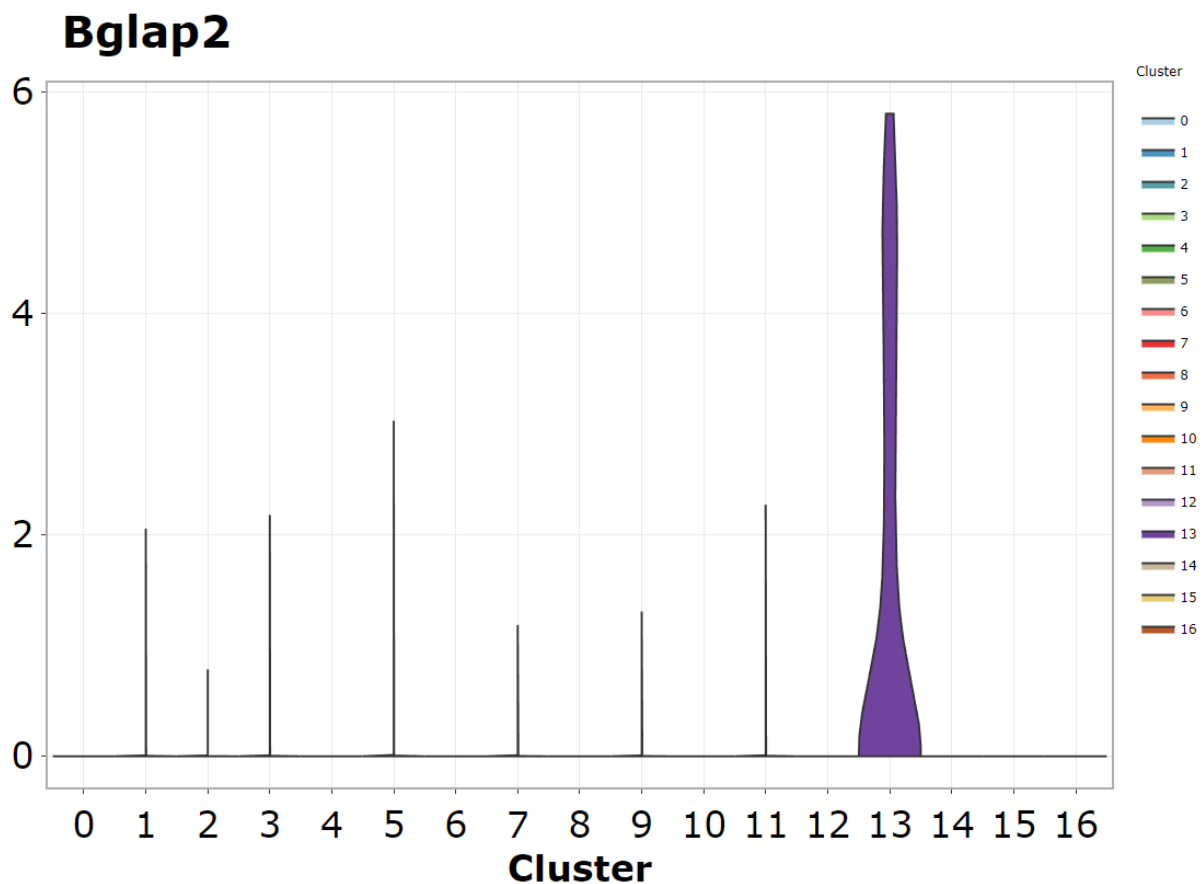


Figure 30. Violin plot used for inspection of genes across clusters.

2.13 Doublet detection

Doublet detection in scRNA-seq datasets is carried out by using the R package DoubletFinder (McGinnis, Murrow, and Gartner 2019). Initially, artificial doublets are simulated and merged with the original data.

Cells exhibiting a high number of artificial neighbors in the gene expression space are then characterized as potential doublets and can be excluded from subsequent analysis. This methodology demonstrates enhanced accuracy in detecting doublets arising from transcriptionally distinct cell types (increased performance in heterotypic doublets over homotypic). Similarly, for scATAC-seq datasets, a similar approach, implemented in ArchR package, is stratified to identify potential doublets. After computing doublet enrichment measurements, users can filter out doublets by specifying their preferred thresholds.

2.14 Cell cycle phase analysis

Cell cycle phase scores are computed for all cells based on canonical markers linked to S, G2/M, and G1 phase. If cluster-specific patterns of cell cycle biases are identified, users have the option to utilize the "regress out" feature during the scaling step to mitigate the cell-cycle effect. The results of this analysis can be visualized either in a scatter plot, where cells are projected into reduced spaces (PCA, UMAP, tSNE, diffusion map, PHATE), and colored according to the predicted phase of the cell cycle (Fig. 31), or as a bar plot summarizing the percentages of cells assigned to each cell cycle phase within each cluster.

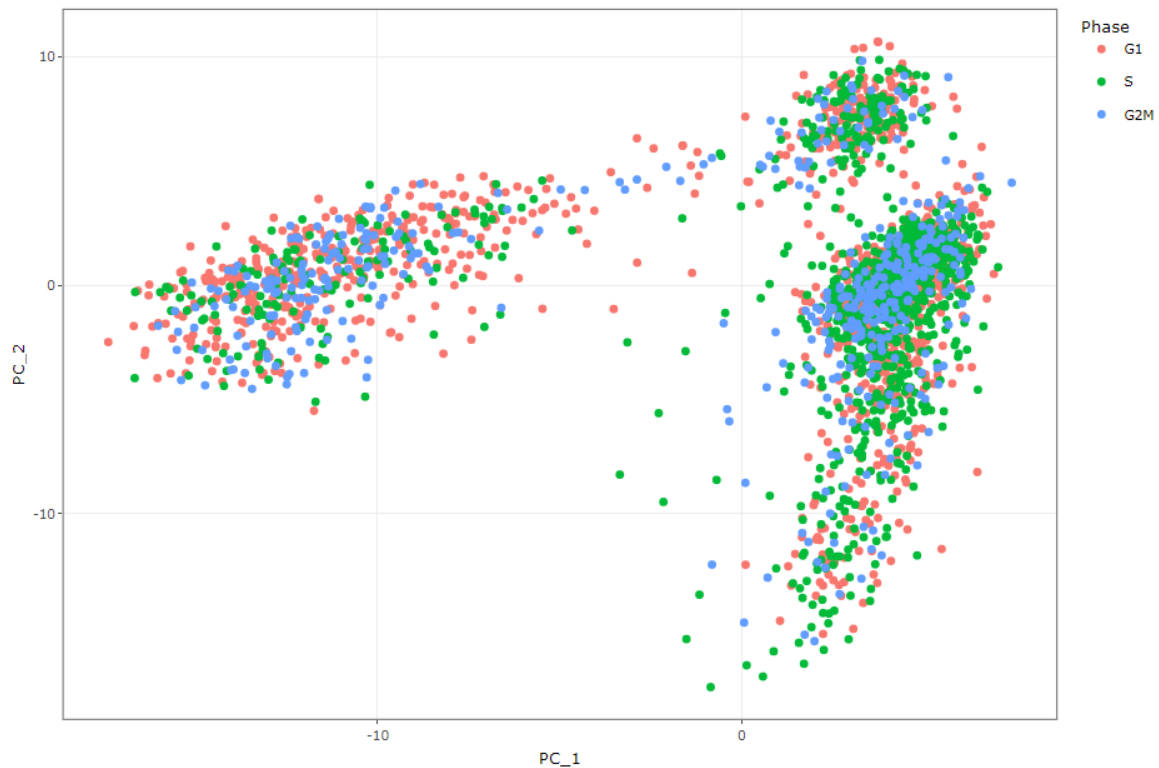


Figure 31. Scatter plot depicting cells colored by predicted cell cycle phase in PCA space.

2.15 Functional/Motif enrichment analysis

Utilizing the previously identified marker genes and marker peaks, functional enrichment analysis, encompassing pathways and Gene Ontology (GOs) terms, as well as motif enrichment analysis, can be conducted for each cluster in scRNA-seq and scATAC-seq data respectively. Specifically, in scRNA-seq data, genes upregulated or downregulated (in the clusters identified during the previous steps) are assessed for enriched GO terms or KEGG pathways using the g:Profiler package (Raudvere et al. 2019). The enriched terms are presented in a tabular format alongside information relative to statistical significance and gene overlap between the input list and the term of interest. Furthermore, a bubble plot summarizing the enriched terms for the selected databases is also available to the users (Fig. 32). Regarding motif enrichment analysis, marker peaks identified in the step of marker peaks detection, are examined for enrichment of binding sites of specific transcription factors. Additionally, more comprehensive functional enrichment analysis with enhanced visualization options is provided by the external application Flame (Thanati et al. 2021). This analysis can be conducted either iteratively for each cluster, or by selecting as input multiple gene lists (up to 10 clusters) for simultaneous processing using interactive UpSet plots and many other additional features.

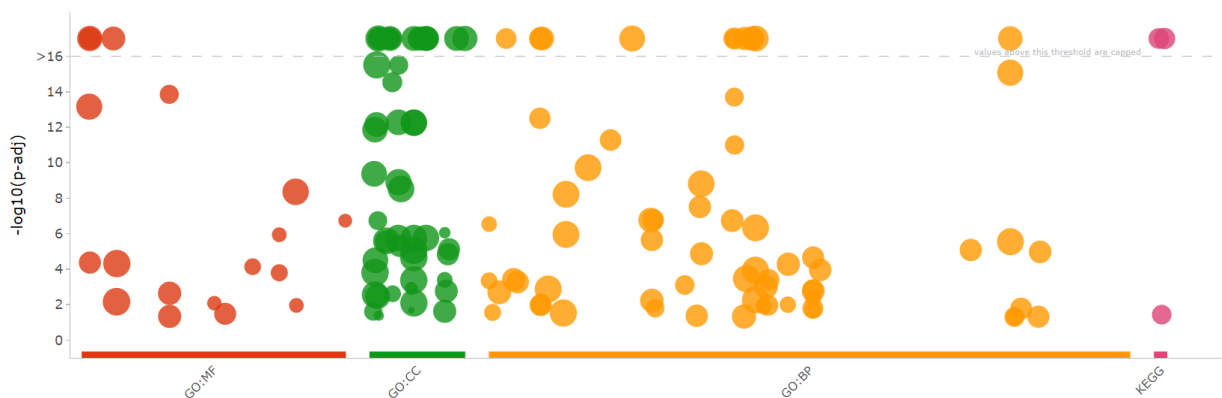


Figure 32. Bubble plot showing enriched functional terms of a cluster.

2.16 Automated annotation of clusters

This module is implemented only for scRNA-seq datasets. The CIPR package is employed (Ekiz et al. 2020), providing reference datasets, containing various cell types, for both human and mouse organisms. Users can choose a reference dataset and specify the type of analysis to be conducted for producing the final predictions per cluster, either by considering normalized expression measurements from all genes in the

dataset or fold change values only from the differentially expressed ones. Additionally, users have the flexibility to select the correlation metric (Pearson or Spearman) which will be used during the calculations. In terms of results' visualization, the output includes a table displaying all predicted cell type annotations per cluster, along with a dot plot that depicts the top 5 predictions for each cluster (Fig. 33).

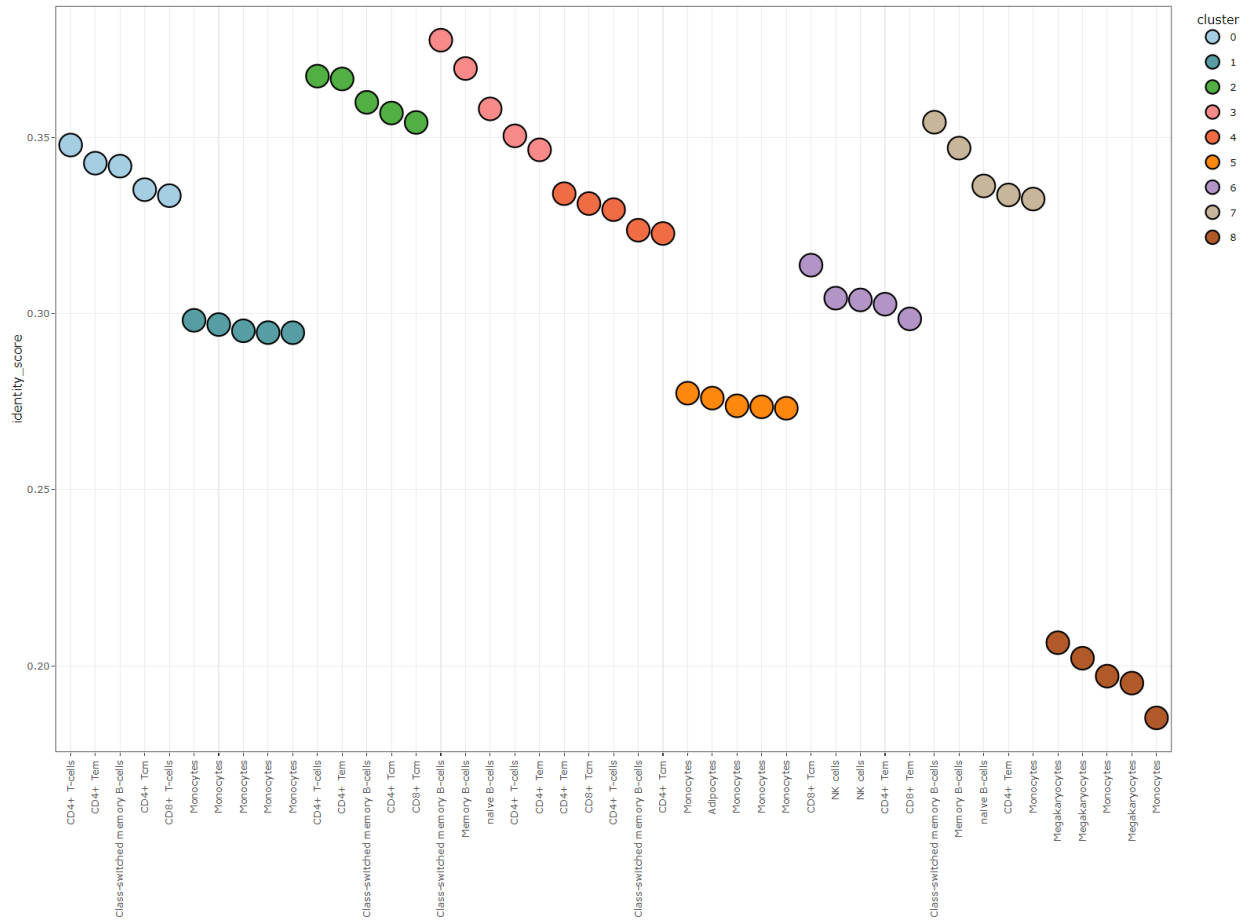


Figure 33. Dotplot showcasing cell type annotation predictions across clusters.

2.17 Multimodal integration analysis

This analysis mode is dedicated to scATAC-seq datasets. Specifically, users have the option to upload a pre-processed scRNA-seq dataset to perform integration analysis with the currently loaded scATAC-seq dataset. During this process, gene activity scores from the ATAC assay and gene expression values from the RNA assay are utilized to align cells between the datasets. The result of this integration analysis enables transferring labels from scRNA clusters to cells within the scATAC dataset. Subsequently, the newly

assigned clustering identities of the cells can be further utilized in various downstream tasks such as marker peak detection, trajectory analysis, and other analytical steps that require cluster information.

2.18 Trajectory analysis

The ordering of cells using pseudotime analysis can be very helpful in unravelling the underlying processes of differentiation and development, guiding cells through transitions between different cellular states. In our application, Slingshot package is employed for this purpose, utilizing both clustering information and dimensionality reduction coordinates for all cells within a dataset, Slingshot constructs a Minimal Spanning Tree (MST) at the cluster level. In this tree, nodes represent clusters, while edges signify relationships between them. Users have the option to select the dimensionality reduction method employed for Slingshot execution (PCA, UMAP, tSNE, diffusion map, or PHATE), along with defining the initial and final states of the trajectory. The MST is depicted in a UMAP plot, while pseudotime values are computed per lineage and can be further illustrated in UMAP space as a distinct scatter plot (Fig. 34). Cell pseudotime values close to zero indicate cells belonging to the root of the trajectory, while higher values denote cells associated with the final state.

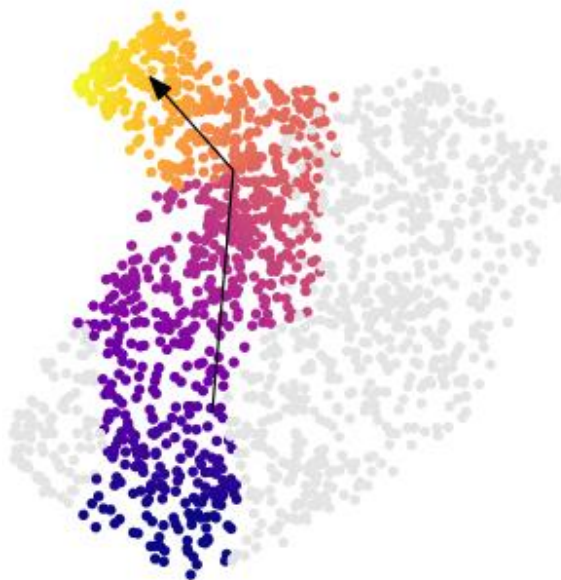


Figure 34. Scatterplot showcasing trajectory results. Arrow shows the direction of the lineage, while cells are colored according to pseudotime values.

2.19 Cell-cell communication analysis

The prediction of ligand-receptor interactions is a significant step for deciphering cell-to-cell communication patterns in different tissues. Inspection of communication networks between different cell types can contribute to the detection of key interactions, which can lead to gene expression alterations (downstream of signaling pathways) in healthy and disease contexts. SCALA incorporates the analysis framework of nichenetR. More precisely, after clustering the user needs to select a pair of clusters that will be examined for active L-R interactions among them. First, overexpressed genes are calculated in each cluster. Next, the reported interactions are ranked by considering a "prior interaction potential" score that is calculated in the initial steps, when the protein-protein interaction model is constructed. Regarding the visualization of results, a heatmap that summarizes all the interactions that have been detected between the two clusters of interest is provided (Fig. 35). L-R interactions along with their respective scores are available in a table format including the prior interaction potential score that signifies the strength of the predicted interaction.

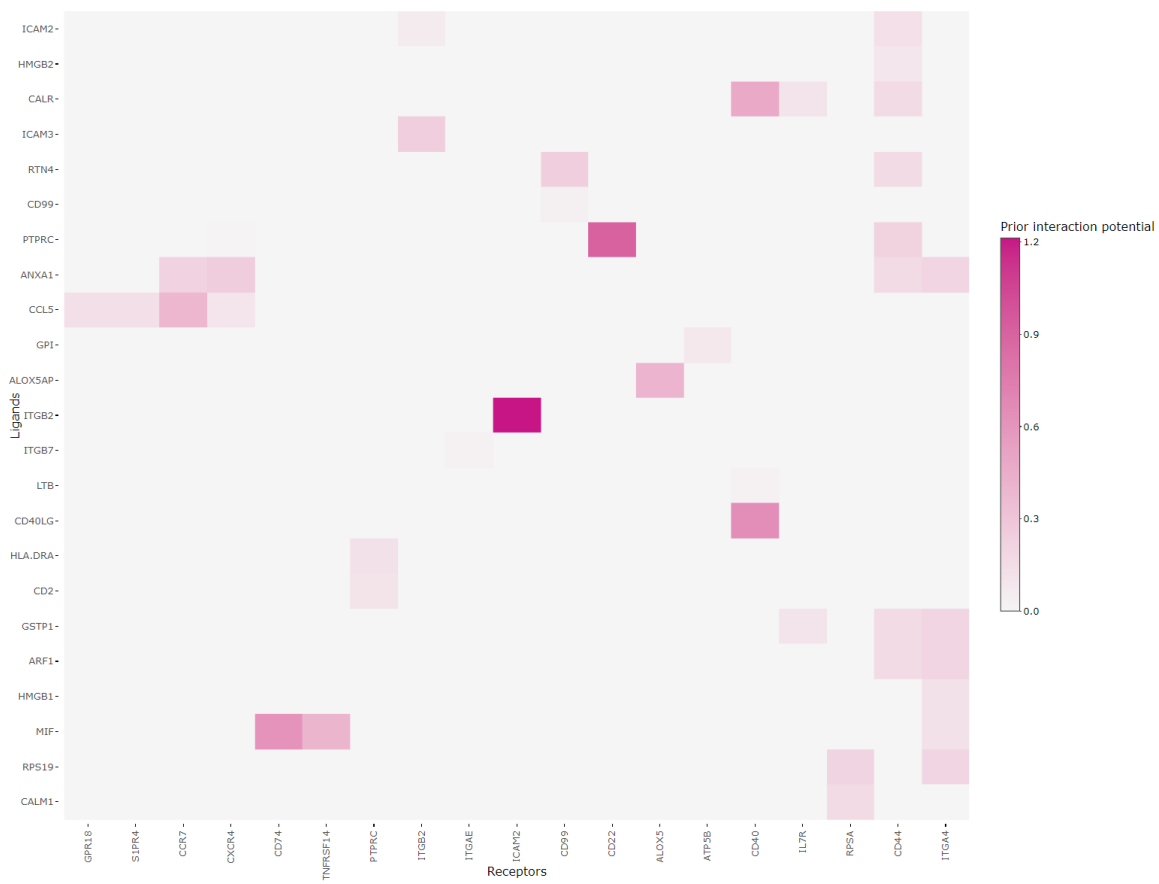


Figure 35. Heatmap summarizing L-R interactions between clusters.

2.20 Gene regulatory network reconstruction

In our application, users have the option to choose between two methodologies for this step of the analysis in scRNA-seq datasets.

The first method adopts the SCENIC workflow. Initially, co-expression modules of TFs and their target genes are detected through co-expression and TF motif analysis. Subsequently, the AUCell package calculates AUC scores per cell, representing the activity of a regulon—a group of genes containing a TF and its targets. These AUC values, along with Regulon Specificity Score (RSS) scores, indicating the activity and specificity of regulons respectively, are used for the visualization of active regulatory networks in heatmap format. Additionally, users can identify cluster-specific regulons within the dataset, by inspecting the average activity values across clusters. Due to runtime limitations in R environments, users are provided with instructions and custom scripts to externally execute certain parts of the analysis in Python. They can then import the result files back into our application for visualization and exploration.

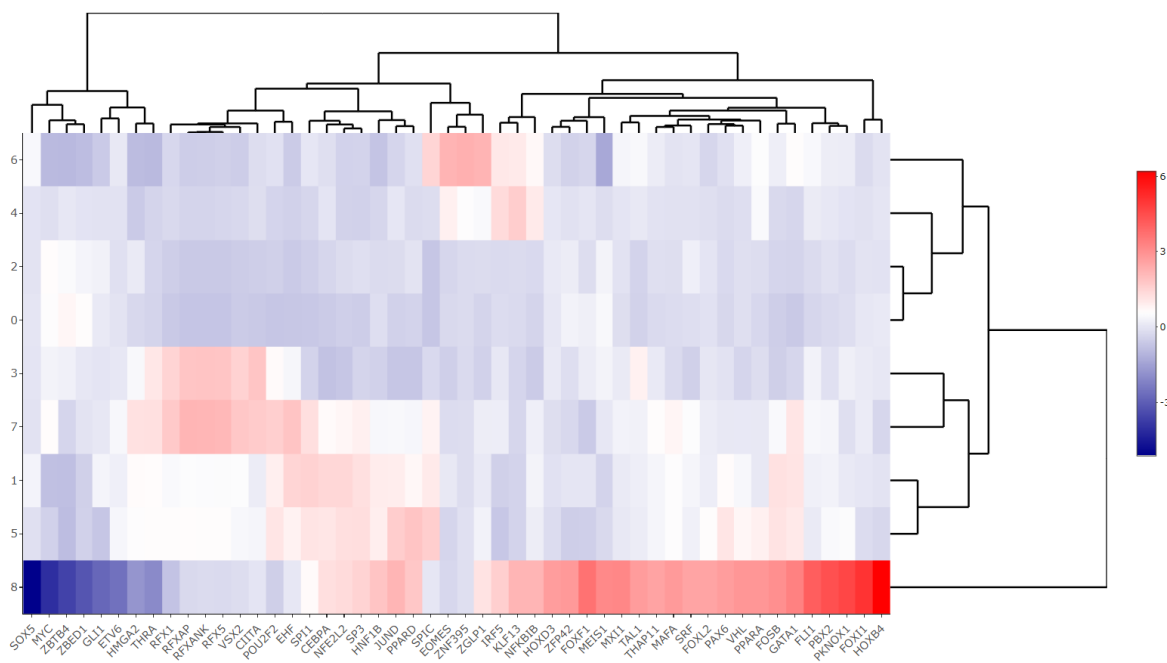


Figure 36. Heatmap showing scale TF activity scores for 50 TFs across clusters.

The second methodology offers an alternative option for users who prefer not to use SCENIC analysis. This approach follows the method proposed by decoupler to infer TF activity levels at the single-cell level. Specifically, it utilizes a curated resource of interactions between TFs and their target genes (CollectRI). For each cell in the dataset and each TF, a linear model is fitted to predict observed gene expression based

solely on the TF's TF-Gene interaction weights from the CollecTRI resource. The resulting t-value of the slope serves as the TF activity score in the cell, where positive values indicate TF activity and negative values indicate the opposite. These scores are scaled for visualization, and average activity values per cluster are provided. The TFs showcasing high variability among the clusters are depicted in a heatmap.

For scATAC-seq datasets, gene regulation analysis at the chromatin level aims to identify cluster-specific TFs whose expression correlates strongly with chromatin accessibility changes at genomic sites, including their DNA binding motifs, a process known as the identification of positive regulators.

2.21 Visualization of epigenome signal tracks

Chromatin accessibility tracks serve as an alternative to feature plots, which typically display gene activity scores in reduced space, such as UMAP or tSNE plots. In our application, users have the option to select a specific gene and define a genomic interval of interest by specifying the number of bases upstream and downstream. By examining the generated plot via a genome browser snapshot (Fig. 37), users can identify chromatin accessibility patterns within the gene body or within upstream/downstream gene regulatory elements, including promoters, enhancers, and silencers.

2.22 Utility functions and code history

Since the developed application is more useful to users with limited computational expertise, we tried to enhance the overall user experience and accessibility. This was achieved by integrating features like comprehensive instructions accompanied by explanatory screenshots, as well as a command history log detailing actions performed during basic analyses, as well as various utility operations.

Each analysis task within SCALA is accompanied by a series of explanatory pages and tabs. These contain brief descriptions of the ongoing operation, input guidelines, and explanations of the output. Certain modes also feature an instant help section, conveniently located at the top of the page in the form of a collapsible window. Furthermore, to alert users to potentially longer processing times, especially for large datasets, a banner displaying estimated processing times is also provided (Fig. 38).

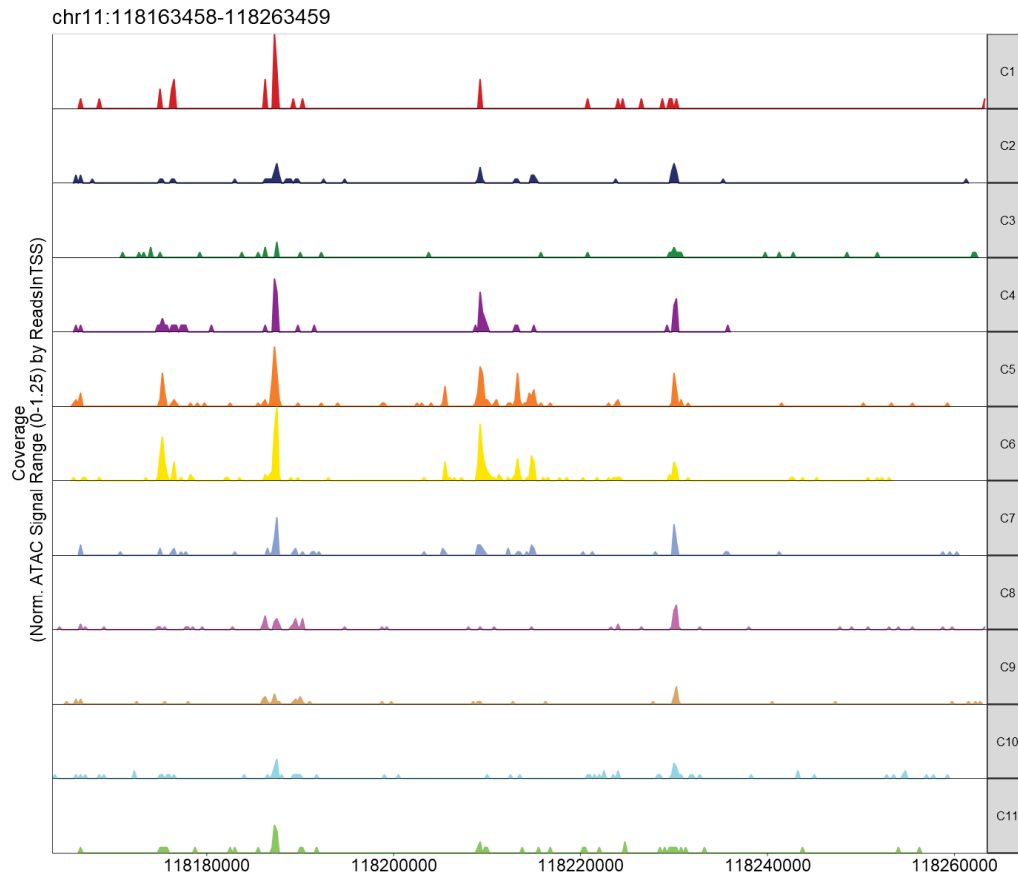


Figure 37. Example of ATAC tracks for an immune marker gene across different clusters.

Regarding the code history function, for scRNA-seq data, the basic analysis R commands and their parameters are displayed in a dedicated text area at the “utility” tab. For ATAC-seq data, the executed commands and the selected parameters are stored in a text file within the sub-folder of each analysis step. These sub-folders are automatically saved under a directory dedicated to storing all the analysis output files. This feature is currently supported only in the local version of the application.

As for the utility options, they become available to the user after the completion of the clustering operation (Fig. 39). Initially, users often experiment with different clustering resolutions, so it's important for them to select the active clustering column. This column is utilized for subsequent analysis steps such as functional enrichment and trajectory inference and many more. Another utility option is cluster renaming, which is useful for annotating clusters with cell type identities or custom names. It's also helpful in merging clusters that showcase similar gene expression patterns. The last utility operation is cluster

deletion. It is commonly used when unexpected cell types are found in the dataset or when a cluster is identified as a poor-quality cluster, according to the QC metrics. Users may also choose to delete clusters to focus on specific cell type categories for sub-clustering analysis (e.g. subclustering of fibroblasts).

Do you need help with the upload of 10x files?

Upload your data

Gene count matrix (scRNA-seq) | 10x input files (scRNA-seq)

RDS format object input (scRNA-seq)

Annex input files (scATAC-seq)

Load PBMC 10x dataset (example scRNA-seq)

Load example

By pressing this button, the example dataset is loaded

OR

To insert your own data:

- 1) Write the name of your project
- 2) Load the three required files, which can be found in the output folders of cellranger
- 3) Pre-filtering options that can exclude cells or genes of low information
- 4) Select between human or mouse
- 5) Press this button to upload your data!

Upload your files

Project name:

1. Choose barcodes.tsv.gz file

2. Choose features.tsv.gz file

3. Choose matrix.mtx.gz file

Include features detected in at least this many cells:

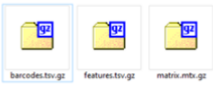
Include cells where at least this many features are detected:

Select organism:

Mus musculus (Mouse)

Homo sapiens (Human)

Upload



PCA estimated time in web server for a scRNA-seq dataset of 6,000 cells ~ 8sec (quick version), ~25min (slow version)

LSI estimated time in web server for a scATAC-seq dataset of 6,000 cells ~ 38 sec

*For datasets containing more than 10,000 cells the slow version of PCA is not suggested

(The execution times were measured in the web version of the tool. However, improved performance can be achieved by using the stand-alone version on PCs with appropriate CPU and RAM specifications.)

Figure 38. Examples of instructions for data input (at the top of the figure) and a banner showing estimated execution time for a dataset containing 6,000 cells (bottom of the figure).

Edit/export working object

Rename cluster

Cluster to be renamed (old name):

New name of the cluster:

[Rename](#)

Delete cluster

Cluster to be deleted:

[Delete](#)

Active clusters

Set active clustering column:

[Change clustering column](#)

Command history

[View command history!](#)

```

Command: NormalizeData(seurat_object, normalization.method = normalize_normMethod, scale.factor = as.numeric(normalize_normScaleFactor))
Time: 2024-05-06 20:36:18
assay : RNA
normalization.method : LogNormalize
scale.factor : 10000
margin : 1
verbose : TRUE
[1] "-----"
Command: FindVariableFeatures(seurat_object, selection.method = normalize_hvgMethod, nfeatures = as.numeric(normalize_hvgNgenes))
Time: 2024-05-06 20:36:19
assay : RNA
selection.method : vst
lods.span : 0.3
clip.max : zero
num.bin : 20
binning.method : equal_width
nfeatures : 2000
mean.cutoff : 0.1 8
dispersion.cutoff : 1 Inf
verbose : TRUE
[1] "-----"
Command: ScaleData(seurat_object, features = all.genes)

```

Figure 39. Utility options and code history functionality are shown.

3 Results

3.1 Analysis of synovial fibroblasts in *hTNFtg* arthritis mouse model

To demonstrate the capabilities of our application, we utilized two previously published datasets (Marietta Armaka et al. 2022), encompassing both scRNA-seq and scATAC-seq data. These datasets were originally generated to investigate the dynamics of single-cell transcriptomes and chromatin in Synovial Fibroblasts as they transition from homeostasis to pathology in a TNF-driven arthritis model. We specifically employed the Tg197 mouse model of arthritis (Keffer et al. 1991), and compared it to healthy wild type (Wt) mice.

For cell isolation, non-hematopoietic stromal cells (Cd45-, Cd31-, Ter119-, Pdpn+) were sorted from the synovium of whole ankle joints and used to prepare 10x Genomics scRNA-seq libraries. These libraries, sequenced with a depth of 400 million reads using an Illumina NextSeq 500 machine, comprised 6,667 single cells. Similarly, scATAC-seq libraries were generated following 10x Genomics guidelines, profiling 6,679 single nuclei.

In both experiments, cells were sourced from tissues of three healthy mice (WT, 4 weeks old) and six diseased *hTNFtg* mice, with three at an early disease stage (*hTNFtg/4*, 4 weeks old) and three at an established pathological stage (*hTNFtg/8*, 8 weeks old). As discussed in the introduction section, the Tg197 mouse model, characterized by the overexpression of the human TNF (hTNF) transgene, exhibits an arthritic phenotype marked by cartilage destruction and bone erosion, ultimately leading to joint function impairment.

Our developed application was utilized to reanalyze the transcriptomes of 5,903 synovial fibroblasts (SFs) and epigenomes of 6,046 cells from healthy mice (control sample) and arthritic mice at 4 and 8 weeks of age (early and established disease states). Additionally, any custom analysis steps performed outside the SCALA application environment will also be described in the current section of this dissertation.

3.2 Analysis using SCALA's scRNA-seq pipeline

For the scRNA-seq quality control step, cells with fewer than 500 detected features (genes) or with more than 10% of their reads mapped to the mitochondrial genome were excluded from further analysis.

Subsequently, downstream analysis of scRNA-seq proceeded with the following operations: the most highly variable features were identified using the mean-variance-plot (MVP) method (offered by the Seurat package), resulting in the identification of 1535 variable genes. The gene counts of each cell were normalized by the total cell counts, multiplied by 10,000, and then subjected to natural-log transformation. Normalized expression values for all genes were scaled by "regressing out" the mitochondrial content effect.

The scaled gene-by-cell expression matrix of the most variable genes was used as input for Principal Component Analysis (PCA). To determine the dataset's dimensionality, and thus the most informative principal components reflecting cell heterogeneity, Singular Value Decomposition (SVD) k-fold cross-validation was conducted using the *dismo* R library. This operation suggested that 25 principal components (PCs) could capture the most relevant aspects of cell diversity. These 25 PCs were then utilized for both cell clustering and non-linear dimensionality reduction analysis.

Specifically, to delineate distinct fibroblast subsets, graph-based clustering analysis was performed using Seurat's Louvain algorithm, with the resolution parameter set to 0.6. Furthermore, the 25 most informative PCs were employed for non-linear dimensionality reduction analysis, including both t-SNE and UMAP plots, enabling visualization of the newly identified cell clusters in 2D/3D space.

SF clustering resulted in the delineation of 10 distinct SF clusters, each presenting unique transcriptional profiles that embody homeostatic, inflammatory, and destructive properties that can be observed in healthy and arthritic joints. These distinguishing features were elucidated through marker gene identification analysis conducted for each SF cluster. Specifically, the transcriptomes of each cluster were compared against those of all other cells using the Wilcoxon rank sum test on normalized gene expression values. After the analysis was completed, genes meeting the criteria of an average log Fold Change (avg. logFC) > 0.25, a percentage of expression (% of cells in the cluster in which the gene is detected) > 25%, and a p-value < 0.01 were retained for additional modes of analysis.

Initially, genes showing up-regulation were employed as input for functional enrichment analysis. More specifically, GO enrichment analysis was conducted for each fibroblast (SF) cluster using *g:Profiler* package. By scrutinizing similarities and disparities among SF clusters in terms of markers and enriched functional terms, two clusters, labeled 0 and 9, were merged. Consequently, the resulting nine clusters were denoted as S1, S2a, S2b, S2c, S2d, S3, S4a, S4b, and S5. It is worth noting that these identified clusters demonstrated

changes in their relative abundances between healthy and diseased states. Interestingly, while some clusters are observed to diminish, others display expansion during disease progression.

Thy1 + clusters (S1, S2a, S2b, S2c, S3, and S5) were further categorized as “sublining”. Notably, their transcriptional and functional characteristics reflect features of tissue homeostasis preservation, apart from S5, which exhibits an immuno-regulatory role under healthy conditions. Enriched Gene Ontology (GO) terms for these populations encompass biological processes such as BMP, WNT, TGFbeta, and SMAD signaling, as well as responses to TNF and IFN-beta/gamma. Key markers for these clusters include Smoc2, Thbs1, Vwa, Rgma, Dkk2, Sfrp1, Ecrq4, Osr1, Nr2f2, Klf5, Clu, Id1, Meox1, Pi16, Sema3c, Efemp1, Ccl7, Il6, and Notch3.

Likewise, the Prg4High S4a cluster was designated as "lining" and associated with functions characterizing an inflammatory and destructive profile specific to this SF subpopulation. The lining phenotype is characterized by markers such as Tspan15, Hbegf, Htra4, and Clic5. In terms of enriched biological processes, we observed terms such as inflammatory response and class I antigen presentation.

Finally, clusters S2d and S4b exhibited a mixed expression profile of both Prg4 and Thy1 (Prg4+ Thy1+) and were thus labeled as "intermediate" subpopulations. Marker genes such as Fbln7, Thbs4, Cthrc1, Lrrc15, Dkk3, Mki67, Pdgfa, Birc5, Aqp1, Acta2, and Cxcl5, which were predominantly upregulated in the intermediate and lining compartments, have been previously implicated as contributors to fibroblast pathogenicity or associated with potential pathogenic roles. Correspondingly, terms such as regulation of immune response, redox response, fibroblast proliferation, cell division, and apoptosis were found to be enriched in S2d and S4b. In conclusion, the intermediate group of SFs exhibits a pro-inflammatory and proliferative character.

Next, cell cycle phase analysis was conducted, categorizing each cell into S, G1, or G2/M phase. Intriguingly, among the three SF populations demonstrating pathogenic characteristics, S4b exhibited the highest proportion of cells in the G2/M phase. This discovery was further reinforced by inspecting cycling markers obtained from the literature, which were mainly expressed in the S4b cluster. The mixed expression signature of Prg4 and Thy1 (Prg4 + Thy1 +), characteristic of this “intermediate” cell group, can be considered as a robust indicator of disease state.

Cellular trajectories were computed in the pooled dataset by employing the first 25 most informative principal components as input for the slingshot algorithm. To identify the clusters designated as the initial and final states of the trajectory, insights from current literature (Wei et al. 2020; Buechler et al. 2021)

were taken into consideration, along with analysis results from alternative external software applications such as scVelo (Bergen et al. 2020) and CellRank (Lange et al. 2022). The resulting minimum spanning tree highlighted the presence of a pathogenic branch comprised of clusters S2a, S2d, S4b, and S4a, with S4a identified as the terminal state and S1, S2b, S3, and S5 as potential starting points.

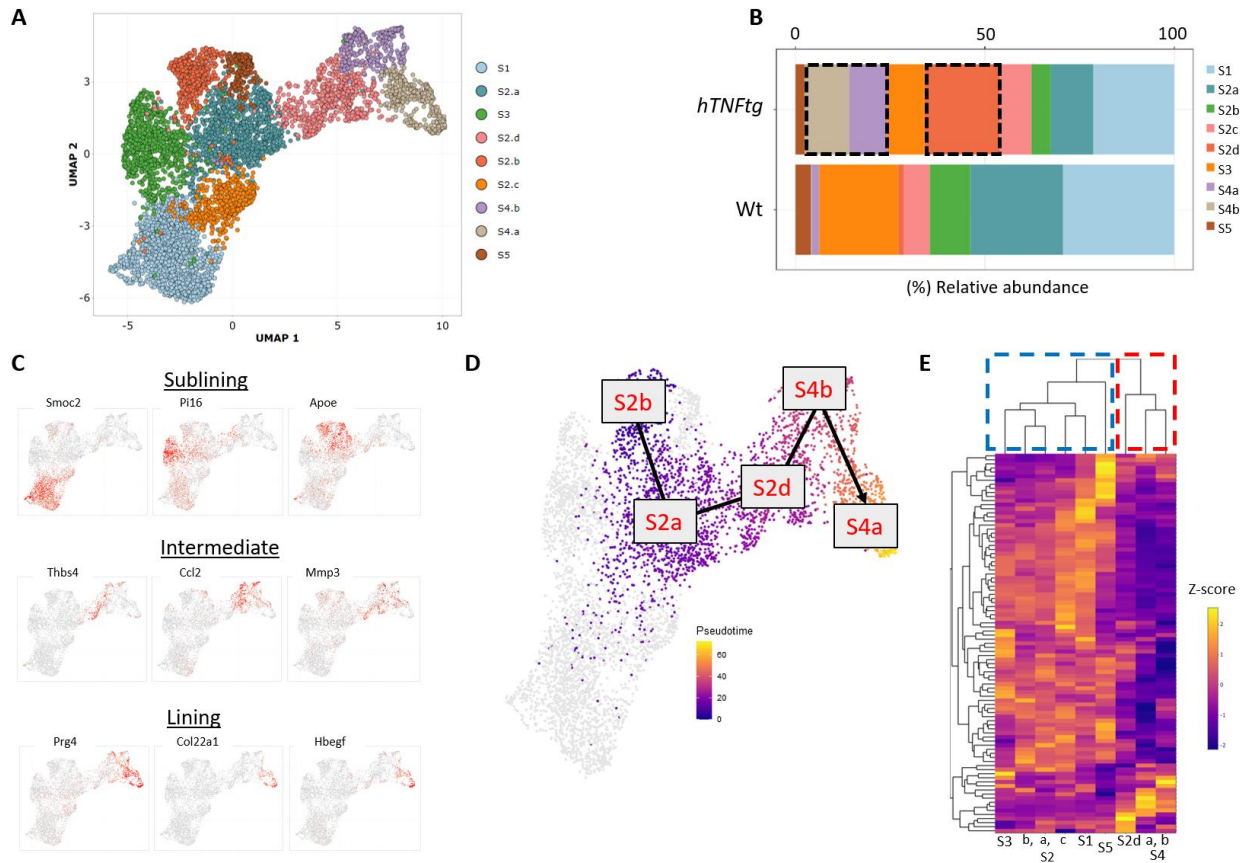


Figure 40. Results of the scRNA-seq data analysis with SCALA (A) UMAP plot depicting SF cells in 2D space. Cells are colored according to cluster identity. (B) Bar plot showcasing relative abundances of clusters in healthy and disease state. (C) Feature plots showing gene expression patterns of marker genes. (D) Trajectory analysis results shown as a UMAP overlay. Cells are colored by their pseudotime value in the lineage S2b – S2a – S2d – S4b – S4a. (E) Heatmap showing z-scores of regulons' activity across clusters. The column dendrogram divides clusters into two major groups: sublining clusters (blue) and intermediate & lining clusters (red) (Adopted from Tzaferis et Al., 2023).

We then proceeded to investigate ligand-receptor interactions between the sublining and intermediate compartments with the lining. Employing the nichnetR package we detected ligands and receptors exhibiting a percentage of expression > 10% in the clusters of interest and we uncovered both shared and specific interactions. More accurately, we identified 157 interactions between sublining-lining and 152 between intermediate-lining compartments. Among these, 126 interactions were shared, with 26 being specific to intermediate-lining and 31 to sublining-lining clusters. Interestingly, in sublining-lining interactions, we observed pairs of ligands and receptors involved in Wnt and BMP signaling pathways. On

the contrary, in intermediate-lining interactions, pairs associated with MMP13, IL-11, and RSPO2 signaling were detected.

As the final step in the scRNA-seq analysis pipeline, GRN analysis was conducted to identify regulons with preferential activation patterns at the cluster level, leading to the discovery of a total of 133 regulons. Notably, diverse activation patterns were evident across the various clusters, and hierarchical clustering of the top-80 regulons unveiled two distinct groups: the first comprising solely sublining clusters, and the second encompassing intermediate and lining clusters.

3.3 Analysis using SCALA's scATAC-seq pipeline

In the analysis of scATAC-seq data, quality control procedures were initially implemented, during this step cells with a Transcription Start Site (TSS) enrichment score below 4 and count-depth less than 1,000 unique nuclear fragments were excluded from downstream analysis. Subsequently, LSI was employed with a resolution set at 0.6, utilizing the first 30 dimensions, number of iterations equal to 4, and default settings otherwise. Moreover, a Uniform Manifold Approximation and Projection (UMAP) projection was created to enable visualization of cells in two-dimensional space.

Gene activity scores were computed as the summed local accessibility of promoter-associated count-tiles in the proximity of each gene, adopting a distance-weighted accessibility model. In detail, count-tiles in the range of 100,000 bp of a gene promoter were aggregated using the following distance weight formula:

$e^{(-|\frac{distance}{5000}|)} + e^{-1}$. An additional normalization step was implemented (multiplication by $\frac{1}{gene\ size}$, scaled linearly from 1 to 5), to account for gene length biases. Next, the above-weighted sum was multiplied by the aggregated Tn5 insertions in each tile. Gene scores were then scaled to 10,000 counts and log2-transformation was performed. To enhance the visualization of gene activity scores, a smoothing process was applied using the MAGIC algorithm (van Dijk et al. 2018).

Similar to the scRNA-seq analysis, clustering was conducted using the Louvain algorithm with a resolution of 0.6, resulting in the identification of 8 distinct clusters. Next, integration between the scATAC dataset and the previously analyzed scRNA dataset was executed. Our objective was to achieve "label transferring" between the annotated scRNA clusters and the newly emerged clusters identified during the scATAC clustering analysis. This integration process facilitated the labeling of scATAC-seq cells according to the 9 SF subpopulations designated during the scRNA analysis. Following integrative analysis, semi-supervised

trajectory inference with ArchR package, verified the existence of a pathogenic branch, which consists of S2a, S2d, S4b, and S4a clusters.

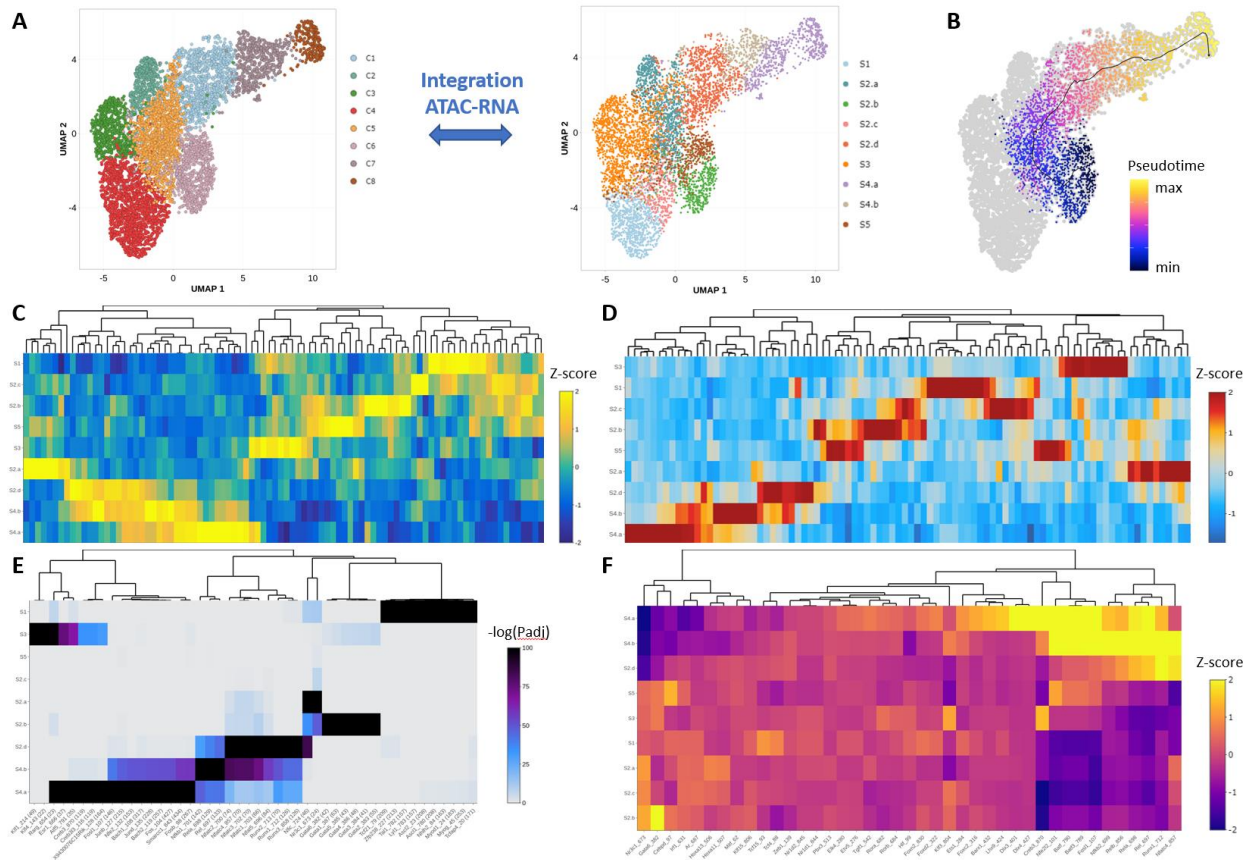


Figure 41. Results of scATAC-seq data analysis with SCALA. (A) UMAP plot of SFs after clustering (left) and label transferring from scRNA-seq data (right). (B) Trajectory analysis results shown as a UMAP overlay. Cells are colored by their pseudotime value in the lineage S2b – S2a – S2d – S4b – S4a. (C-F) Heatmaps showing top marker genes, marker peaks, enriched motifs and positive regulators, respectively. (Adopted from Tzaferis et Al., 2023)

Using previously calculated gene activity scores, the Wilcoxon test was employed to identify statistically significant marker features per cluster (applying $|\text{Log}_2\text{FC}| \geq 0.58$ and $\text{FDR} \leq 0.05$ thresholds). Additionally, a robust merged peak set was determined across SF clusters using MACS2 (Y. Zhang et al. 2008) software, which involved creating two pseudo-bulk replicates. Subsequently, iterative overlap peak merging (Corces et al. 2018) was applied to the pseudo-bulk replicates across SF subpopulations, resulting in a single merged peak set comprising 158,713 regions, each with a fixed length of 500 bps. Following this, differential accessibility analysis between cells was carried out to identify cluster-specific marker peaks (with $|\text{Log}_2\text{FC}| \geq 0.58$ and $\text{FDR} \leq 0.1$ thresholds). These marker peaks were then used as input to perform motif enrichment analysis using the CIS-BP database (applying $|\text{Log}_2\text{FC}| \geq 0.58$ and $\text{FDR} \leq 0.05$ thresholds). Collectively, these analytical approaches revealed distinct patterns of gene activity and peak/motif

accessibility across clusters. Moreover, hierarchical clustering based on z-scores further supported the classification of clusters into three main groups: sublining, intermediate, and lining.

In the ATAC assay, gene regulatory reconstruction was also carried out. Precisely, peak-to-gene linkages were identified by analyzing the correlation between enhancer peak accessibility and integrated gene expression values. Furthermore, TF motif accessibility was correlated with integrated TF gene expression on a cell-by-cell basis, identifying TFs with a Pearson $R^2 \geq 0.5$ and an adjusted p-value ≤ 0.05 , thus identifying 41 “positive regulators”.

3.4 Subclustering of Lining fibroblasts

During TNF-mediated arthritis, lining SFs maintain some of their homeostatic marker gene identity while also displaying an increased diversity in their transcriptome, suggesting potential impairment of their reparative functions post-disease onset. We observed markers associated with inflammatory response (Ccl2, Ccl5, Hmox1, Saa3), class I antigen presentation (H2-K1, B2m, H2-Q7), and ECM remodeling (Mmp3, Timp1, Cd44), which aligns with previous findings on arthritic lining synovial fibroblasts (LSFs) (Croft et al. 2019; Zhang et al. 2019). The expansion of LSFs during disease progression is linked to a decline in certain homeostatic functions, such as ER calcium homeostasis and oxygen level response. Interestingly, a detailed sub-clustering analysis of the S4a cluster revealed the existence of two cell groups, subclusters hS4a (homeostatic) and iS4a (inflammatory), with the inflammatory state iS4a becoming predominant during disease, overshadowing the homeostatic state hS4a. Top marker genes and enriched pathways characterizing the two identified sub-clusters are shown in Fig. 42.

3.5 Comparison of *hTNFtg* and STIA single cell RNA-seq data

After completing the scRNA-seq data analysis of the *hTNFtg* mouse, a chronic arthritis model, we compared it with previously published data from a different mouse model exhibiting acute inflammatory arthritis. To achieve this, we employed integration analysis, a technique useful for comparative or combinatorial analysis among different datasets and modalities. This method has been successfully applied to analyze biological replicates, mitigating potential technical differences, comparing data across species, and combining datasets from various modalities such as RNA and ATAC or CITE-seq.

In our case we leveraged a publicly available SC dataset from serum transfer-induced arthritis (STIA) model uploaded in the Gene Expression Omnibus (GEO) (accession code GSE129087) (Croft et al. 2019). For the generation of the STIA dataset, CD45- synovial cells from the hind limb joints were isolated and sort purified at the ninth day (3 biological replicates, each comprised of cells from the joints of three animals) and captured with the 10X Genomics Chromium system. The integration strategy that is implemented in Seurat package was employed. More specifically WT, *hTNFtg*, and STIA datasets were processed by applying normalization and most-variable-genes detection using the function “normalizeData” with default settings and “FindVariableFeatures” (method set to vst and number of variable features to 2000) respectively. Anchors between samples were identified using the function “FindIntegrationAnchors” with dimensions parameter set to 30, and then the resulting anchors were utilized to integrate all the samples together using the function “IntegrateData”. The final object, containing cells from the control and both arthritic models, was processed in a standard way, performing the steps of dimensionality reduction and clustering. The integrated clusters were defined by using the “FindClusters” function with a 0.4 resolution. Finally, the top marker genes per clusters were selected for visualization purposes.

Interestingly, by inspecting the integration results of normal and *hTNFtg* samples with the respective data from the STIA mouse model, we observed a similar pattern of both expansion and shrinkage of SF clusters compared to the WT control. Furthermore, we performed Spearman correlation analysis between SF clusters, using the normalized expression values of the most highly variable genes of the two datasets. Additionally, exploration of gene expression patterns for the top marker genes per cluster across datasets aligned with the previous analysis results, revealing similarities between particular clusters. Specifically, cluster F3 from STIA aligned with the sublining SFs of the *hTNFtg* mouse, while F5 matched the lining SFs. Finally, clusters F1, F2, and F4 exhibited more similarities with the intermediate SFs from the *hTNFtg* mouse model (Fig. 43).

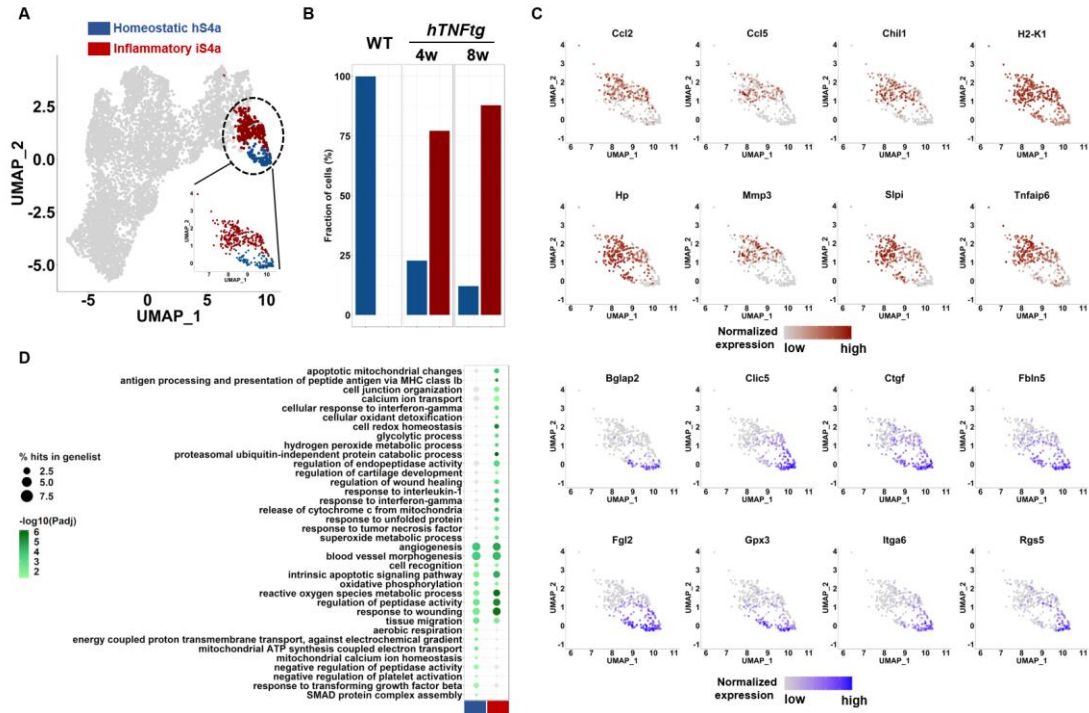


Figure 42. Subclustering of LSFs (A) UMAP showcasing the two identified sub-clusters (hS4a and iS4a are colored in blue and red respectively). (B) Barplot showing the relative abundances of the two lining subclusters across samples. (C) Featureplots showing expression of top marker genes per sub-cluster. (D) Dotplot summarizing the enriched GO terms of each subcluster. (Adopted from Armaka et Al., 2022)

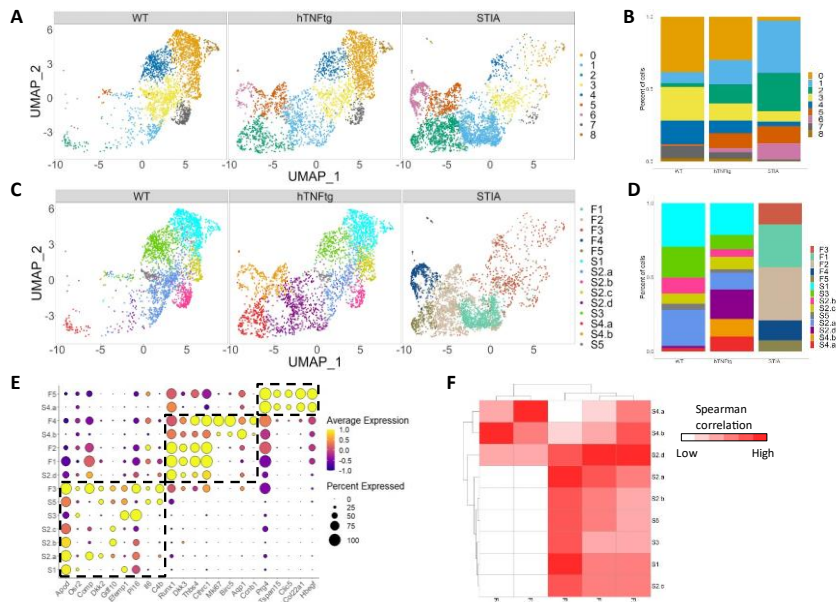


Figure 43. Integration analysis of hTNFtg and STIA mouse models. (A) UMAP plots showing cells of wt, hTNFtg and STIA samples post integration analysis. Cells are colored according to the integrated clusters (B) Barplot showcasing relative abundance of integrated clusters across samples. (C), (D) The same as A and B, but the color code is in alignment with the original cluster annotation of the datasets. (E) Dotplot showing normalized expression of top marker genes across samples. (F) Spearman correlation analysis between clusters of hTNFtg and STIA samples. (Adopted from Armaka et Al., 2022)

3.6 Cross species integration of mouse and human patients single cell data

In the preceding section, we concentrated on the integrated analysis of two distinct mouse models. Subsequently, we aimed to elucidate potential similarities and differences between single-cell gene expression profiles of *hTNFtg* mice and human RA patients. To this end, we integrated previously generated scRNA-seq data from synovial biopsies of RA patients (H) (F. Zhang et al. 2019; Stephenson et al. 2018) with the *hTNFtg* scRNA-seq dataset (M).

Initially, human genes were mapped to their mouse homologs using the Ensembl Biomart and MGI database, resulting in 17,594 homologous pairs. From the mouse dataset, only cells from pooled *hTNFtg* samples (3,051 cells) were processed, while from the three human datasets, only cells from RA patients (24,042 cells) were included. The integration strategy was implemented using the Seurat package. Specifically, all datasets were normalized, and the most variable genes were identified using the “normalizeData” function with default settings and the “FindVariableFeatures” function (with the method parameter set to “VST” and the number of variable features to 2000). Integration anchors between datasets were identified using the “FindIntegrationAnchors” function with the “dimensions” parameter set to 30, and these anchors were utilized to integrate the datasets using the “IntegrateData” function. The final integrated dataset, containing cells from both species, underwent dimensionality reduction, clustering, and marker gene identification. Integrated clusters were defined using the “FindClusters” function with a resolution of 0.3. Marker gene identification was performed using the “findAllMarkers” function with thresholds set to p-value < 0.01 and avgLFC \geq 0.25.

During functional enrichment analysis, upregulated genes from both human and mouse datasets were inserted into Metascape (Zhou et al. 2019). Significant terms and pathways (p-value < 0.05) were used to evaluate the similarities and differences across the datasets.

Interestingly, we observed that cells from both species align well within the newly integrated UMAP space. Employing unbiased graph-based clustering, we identified seven distinct sub-populations denoted as H1-H7 and M1-M7 (Fig. 44). Examination of correlation measurements among the most variable genes (MVGs) between human (H) and mouse (M) clusters unveiled significant similarities in synovial fibroblast (SF) expression profiles across the two species, with the exception of cluster 2. This cluster predominantly comprises human sublining synovial fibroblasts (SLSFs), with only a few mouse cells stemming from the SLSF category.

The synovial fibroblast (SF) populations in mice, including S1, S2a, S2b, S2c, S3, and S5, primarily co-localized with clusters 3 and 4, aligning well with previously annotated human sublining cell expression profiles. Human and murine lining Prg4-high cells are mainly found within cluster 1, with a lesser presence in cluster 7. Furthermore, we identified a previously underappreciated proliferative mixed lining/sublining SF state within these clusters. Cluster 5 predominantly consists of mouse S2d cells, with M5 associated with human cells in both clusters 5 and 6. This suggests that human clusters H5 and H6 could acquire the “intermediate” arthritis-specific profile previously delineated in mouse data analysis (Fig. 45).

Functional inter-species similarities were validated through Gene Ontology (GO) and pathway enrichment analyses of marker genes, as well as co-clustering of human (H) and mouse (M) groups. We identified conserved functions and processes of SLSFs in regulating vasculogenesis, muscle tissue development, and bone and tissue renewal, corresponding to clusters H3, M3, H4, and M4. Our analysis revealed that clusters M5 and H5 are characterized by pathogenic RA features such as metalloproteinase secretion, collagen catabolic processes, and bone destruction signaling pathways, aligning with the S2d SFs identified in the *hTNFtg* model. Clusters 1, 6, and 7, which contain SFs from the lining synovial compartment previously noted for their destructive properties, exhibit pro-proliferative pathways and regulate immune-related, cell adhesion and migration pathways. Furthermore, key marker genes exhibit significant conservation between mouse and human datasets. As anticipated, the analysis of the human-specific cluster 2 showed fewer shared features but highlighted common functions related to translation and ribosome assembly. Human H2 SFs are associated with the regulation of ossification, epithelial cell proliferation, and autophagy. Conversely, the gene expression of mouse M2 SFs is linked to post-translational modifications and apoptotic cell death, differentiating them from H2 SFs.

At level of gene regulation, analysis of human and mouse data using the SCENIC algorithm enabled the inference of common TF regulons across species. Specifically, we retained all conserved TFs identified in both datasets. We identified mouse regulatory modules by performing pairwise correlation analysis between motif deviations of conserved mouse and human TFs, followed by hierarchical clustering, as described in a previous publication. This methodology identified three primary regulatory modules corresponding to lining, intermediate, and sublining states, demonstrating considerable overlap between the two species.

The regulatory modules are governed by the activities of the TFs *Ar*, *Dlx3*, and *Runx1*. Gene Ontology (GO) enrichment analysis of TFs and their downstream target genes revealed that these modules share similar functions in both species. Module one (*Ar*) controls multipotent functions of the core SLSFs; module two

(Runx1) regulates functions associated with an inflammatory profile, consistent with the intermediate profile of SLSFs in the *hTNFtg* model. Notably, up to 25 of the 107 core mouse genes were identified as target genes in human cells, highlighting the translational potential for genes such as *Tnfaip3*, *Tnfaip6*, *Tlr2*, *Lrrc15*, and *Bmp2*. Additionally, the third module (*Dlx3*) exhibited functions that are less well-characterized, likely related to the lining SF profile in both human and mouse SFs.

Conclusively, the aforementioned analysis of mouse and human data at different levels enabled identification of similarities and differences that could help in the prioritization of potential novel targets in future therapeutic interventions.

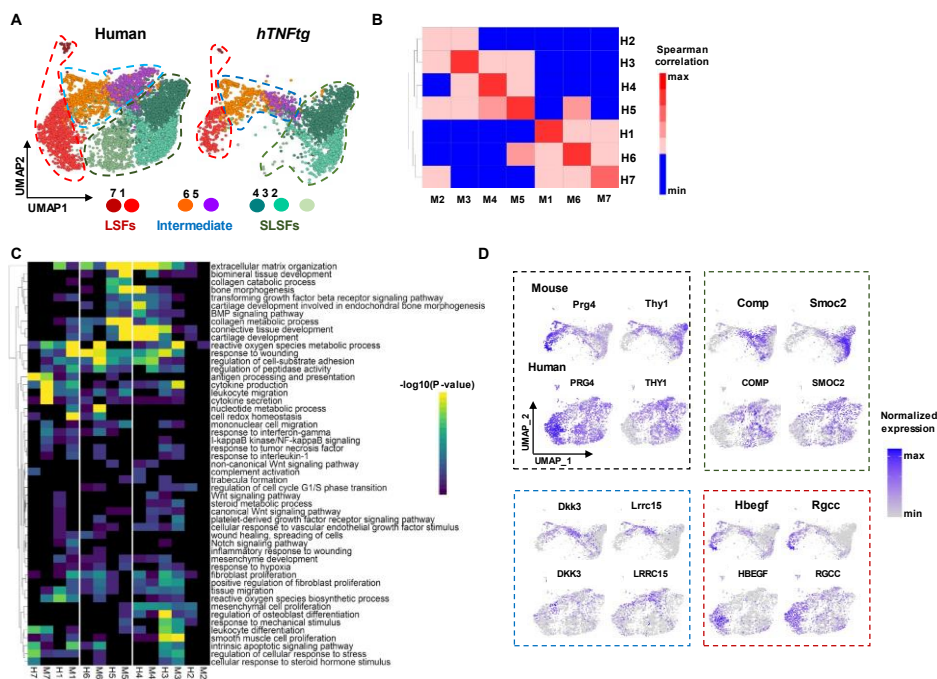


Figure 44. Mouse and Human data integration. (A) Cells of human and mouse datasets depicted in UMAP space after integration. Cells are colored by cluster identity. (B) Spearman correlation analysis between mouse and human clusters utilizing normalized expression data from the most highly variable genes of the datasets. (C) Heatmap showing enriched functional terms for every cluster. (D) Feature plots showing normalized expression for a set of fibroblast marker genes.

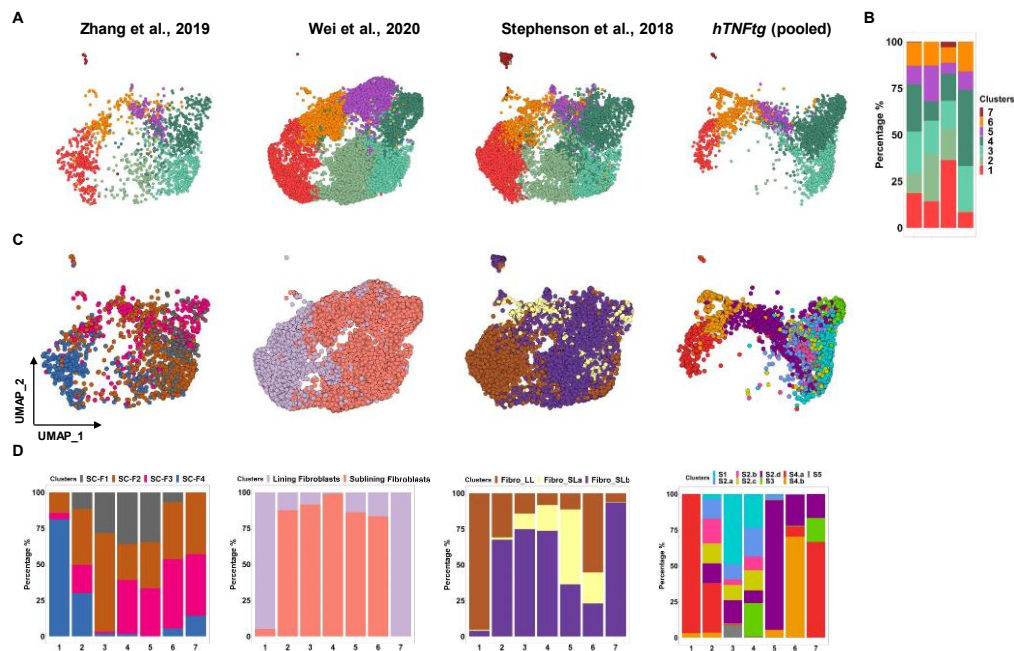


Figure 45. Mouse and Human data integration. (A) (C) Cells plotted in UMAP space after integration. A separate plot is employed for each dataset. In (A) cells are colored according to the integrated cluster while in (C) they are colored according to the original annotation of each dataset. (B) (D) Bar plots showing the relative abundances of integrated clusters or original annotation clusters across the four datasets.

3.7 Analysis with alternative workflows

For the single-cell datasets analyzed in this dissertation, all libraries were generated using the same chip and sequenced in a single run to minimize potential batch effect issues that could interfere with downstream analysis. To prevent skewing biological differences among the various conditions, samples were aggregated using the “cellranger aggr” option. Despite this, we also applied integration between conditions and batch correction. For batch correction Harmony (Korsunsky et al. 2019) was employed across both modalities, which allowed us to identify populations exhibiting expansion. Based on cell annotation from our initial analysis, we confirmed that the majority of cells in the expanding populations were annotated as S2.d, S4.b, and S4.a (91.27% in scRNA-seq and 84.68% in scATAC-seq), corresponding to the expected intermediate and lining clusters (Fig. 46).

To further ensure the robustness of the analysis output (which was based on the Seurat package pipeline), we performed a re-analysis using an alternative toolkit, Monocle3. This workflow comprised the following

steps: (i) Data pre-processing, (ii) Non-linear dimensionality reduction, (iii) Cell clustering, (iv) Inter-cluster comparisons, and (v) Trajectory analysis (Fig. 47).

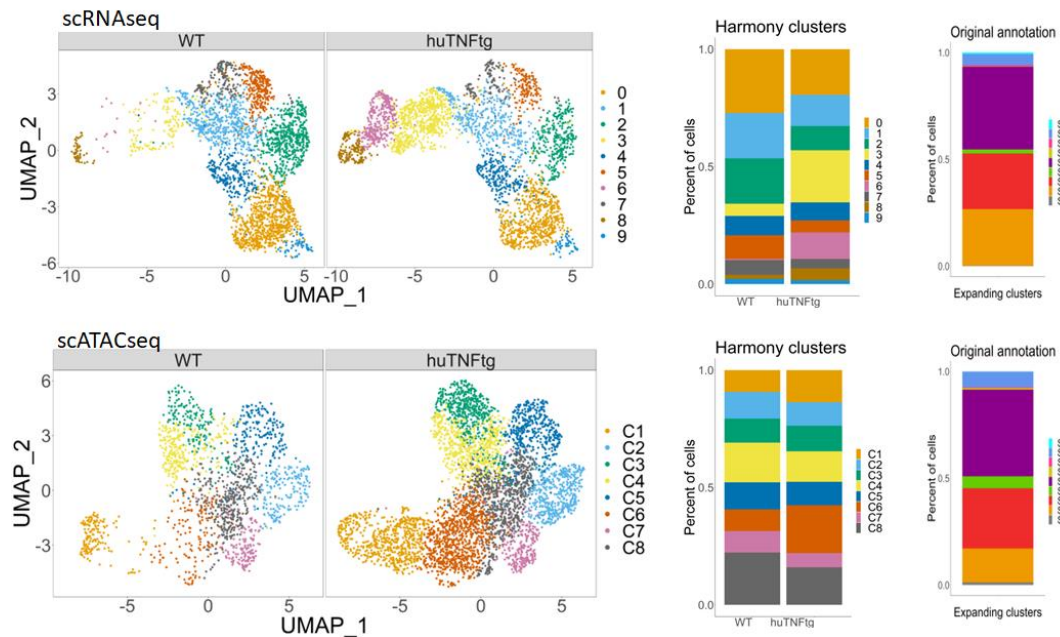


Figure 46. Batch correction analysis. (Left panel) UMAP plots of scRNA-seq and scATAC-seq data after batch correction with Harmony. (Right panel) Utilizing the original cluster annotations, the bar plots in both modalities confirm that the clusters which show an expansion during disease belong to the pathogenic populations.

Monocle3 workflow

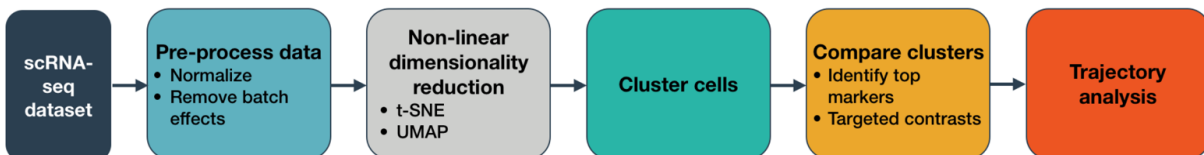


Figure 47. Flow diagram depicting the main steps of analysis incorporated in Monocle3 pipeline.

After loading the data according to the instructions, we tested for potential batch effects. Consistent with previous results from the Harmony package, no batch effects were detected between the samples. Next, normalization and PCA analysis were performed. The UMAP algorithm was used for cell visualization and the Leiden algorithm for clustering. Moreover, we utilized an additional clustering module offered in Monocle, which divides cells into large, well-separated groups called partitions. In our analysis, two main partitions were detected: the first included the sublining SFs, and the second contained the intermediate and lining SFs. By applying the original cell annotations and performing marker gene analysis, we observed that the top markers identified in Seurat were retained (Fig. 48).

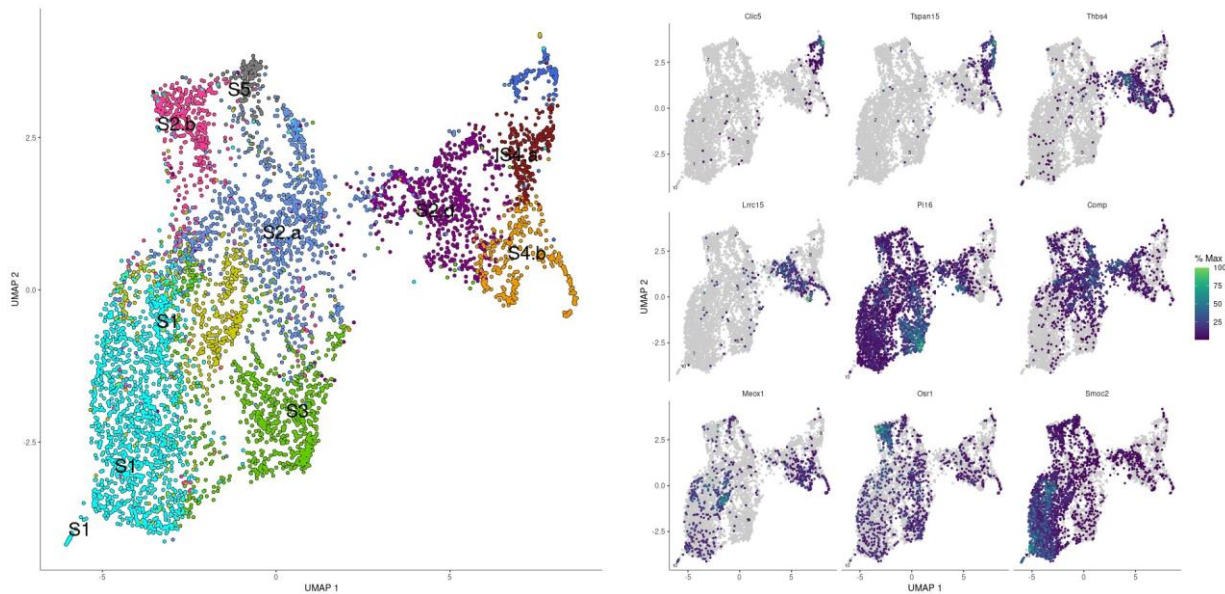


Figure 48. Monocle3 analysis results. (Left panel) UMAP plot showing cells after Monocle3 analysis has been performed. (Right panel) Feature plots depicting gene expression patterns of top marker genes.

Finally, we conducted trajectory analysis and pseudotime ordering of the cells. Monocle utilizes an algorithm that learns the sequence of gene expression changes each cell undergoes during a dynamic biological process. Once the overall trajectory is established, Monocle positions each cell accordingly within this trajectory. We employed four possible roots, as previously described, S1, S2b, S3, and S5. An additional analysis mode enables users to search for genes that change as a function of pseudotime. Specifically, Monocle identifies genes that vary between groups of cells along the trajectory graph using “Moran’s I”, a statistic measure from spatial autocorrelation analysis, which is also effective in single-cell RNA-seq datasets. After identifying the final set of genes that exhibit significant variation across clusters, Monocle groups these genes into modules. This is achieved by applying UMAP on the genes (instead of cells) and then using the Louvain community analysis algorithm to form modules. In our case, this procedure yielded 49 modules, with some showing preferential activity in the initial, intermediate, or the final states of the trajectory (Fig. 49).

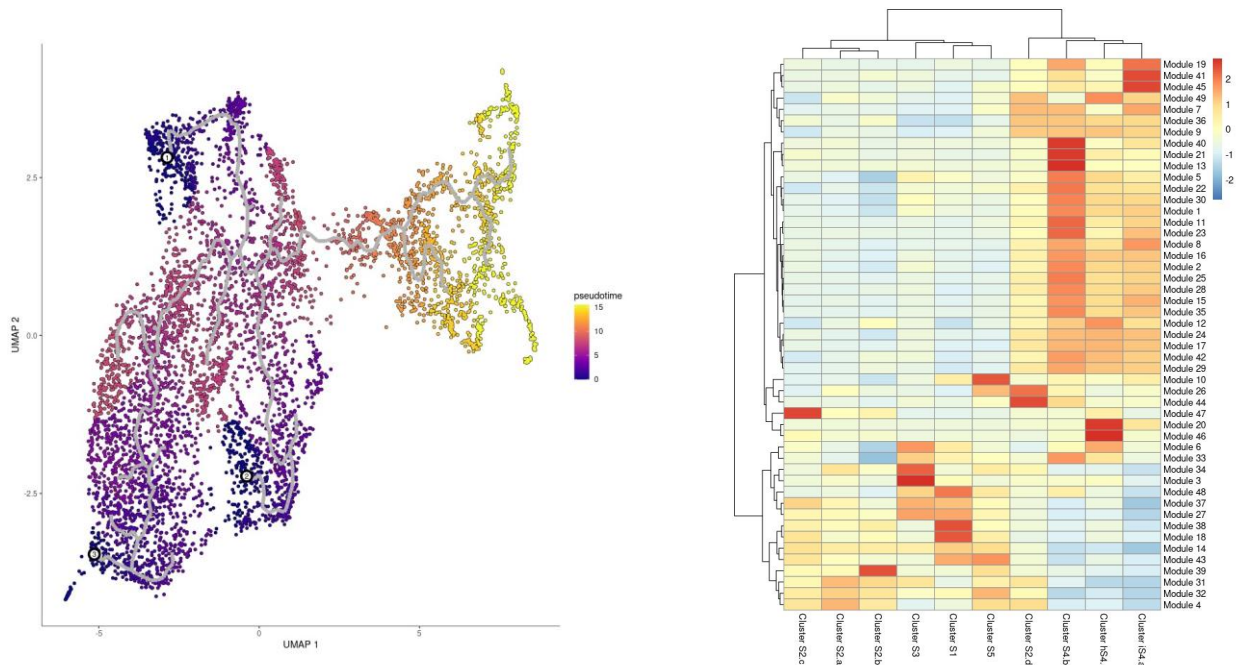


Figure 49. Monocle3 analysis results. (Left panel) UMAP plot depicting trajectory results. Different lineages starting from 3 root points are drawn while cells are colored by predicted pseudotime values. (Right panel) Heatmap showing the activity of gene modules varying between clusters along the trajectory.

3.8 Benchmarking with other similar tools

One of our final objectives was to perform a comparative analysis between SCALA and similar tools. Hence, we present a compilation of alternative tools (e.g., pagoda2 (“GitHub - Kharchenkolab/Pagoda2: R Package for Analyzing and Interactively Exploring Large-Scale Single-Cell RNA-Seq Datasets” n.d.), SingleCellAnalyzer (Prieto, Barrios, and Villaverde 2022), Bingle-seq (Dimitrov and Gu 2020), iCellR (K. H. Tang et al. 2022), cerebro (Hillje, Pelicci, and Luzi 2020), Is-CellR (Patel 2018), SeuratWizard (Yousif et al. 2020), ICARUS (Jiang et al. 2022), SC1 (Moussa and Mandoiu 2021), alona (Franzén and Björkegren 2020), WASP (Hoek et al. 2021), CHIPSTER (Kallio et al. 2011), Asc-Seurat (Pereira et al. 2021), GenePattern (Mah et al. 2019), PIVOT (Zhu et al. 2018)) designed to provide a user-friendly graphical interface for individuals with limited experience in bioinformatics. We emphasize the analytical capabilities offered by these tools and their complementarity to our application. It is noteworthy that, to our knowledge, only iCellR offers scATAC-seq analysis, and only six applications are available as web services. Additionally, SCALA stands out as one of the few tools offering modes for L-R and GRN analysis (Fig. 50).

Tool	link	WebServer	Data input supported format	Quality control	Normalization	Most Variable Genes detection	Linear dimensionality reduction	Non linear dimensionality reduction	Feature inspection	Clustering	Cluster annotation	Marker gene detection	Functional enrichment analysis	Cell cycle phase analysis	Signature scoring	Trajectory analysis	Gene regulatory network analysis	Ligand Receptor analysis	Integrations	Batch correction	ATAC analysis
SCALA	https://scala.fleming.gr	YES	10x data Cell – gene matrix RDS Seurat object	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
pagoda2	https://github.com/kharchenkolab/pagoda2	NO	10x data Cell – gene matrix	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
SingleCAnalyzer	https://singlecanalyzer.eu/home	YES	FASTQ HDF5	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗
Bingle-seq	https://github.com/dbdimitrov/BingleSeq	NO	10x data Cell – gene matrix	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
iCellR	https://github.com/rezakj/iCellR	NO	10x data Cell – gene matrix	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
cerebro	https://github.com/romanhaa/Cerebro	NO	RDS CRB	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Is-CellR	https://github.com/immcore/IS-CellR	NO	10x data Cell – gene matrix RDS Seurat object	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓
SeuratWizard	https://nasqar.abudhabi.nyu.edu/SeuratWizard/	YES	10x data Cell – gene matrix	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
ICARUS	https://launch.icarus-scRNAseq.cloud.edu.au/app/ICARUS	YES	10x data Cell – gene matrix RDS Seurat object	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
SC1	https://cd8.engr.uconn.edu/	YES	Cell – gene matrix	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗
alona	https://alona.panglaodb.se/	YES	Cell – gene matrix	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓
WASP	https://github.com/andreashoek/wasp	NO	Cell – gene matrix	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
CHIPSTER	https://chipster.csc.fi/	YES	10x data Cell – gene matrix	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
Asc-Seurat	https://asc-seurat.readthedocs.io/en/latest/installation.html	NO	10x data Cell – gene matrix	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
GenePattern	https://github.com/genepattern/single_cell_clustering_notebook	NO	10x data Cell – gene matrix	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
PIVOT	https://github.com/kimpenn/PIVOT	NO	10x data Cell – gene matrix	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗
Partek	https://www.partek.com/single-cell-gene-expression/	YES	FASTQ, 10x data Cell – gene matrix, RDS Seurat object, h5Ad	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓

Figure 50. Comparison with other tools. In this table our application is compared to other applications which offer a GUI or are accessible as a web service.

3.9 Performance and scalability of the application

In terms of performance, including execution times and RAM consumption, we conducted benchmarking tests using eight single-cell datasets (four scRNA-seq and four scATAC-seq), varying in cell numbers. The results demonstrated that our application effectively handled datasets containing hundreds of thousands of cells. However, we recommend users to utilize the desktop version of SCALA for larger datasets (> 50,000 cells), as certain analysis steps may require more than 64 GB of RAM memory. Additionally, the desktop version offers improved execution times for scATAC datasets by enabling multiprocessing, which facilitates parallel execution during computationally intensive processes (Fig. 51).

scRNA-seq datasets					
Operations	PBMC 3k	Joint Sfs 6k	Liver Fibro 60k	Ileum Immune 200k	Time format
QC	< 1 sec	< 1 sec	< 1 sec	< 1 sec	HH:MM:SS
Filtering	< 1 sec	< 1 sec	< 1 sec	< 1 sec	
Normalization and Scaling	< 1 sec	< 1 sec	00 00 07	00 00 12	
PCA quick	00 00 07	00 00 12	00 01 26	00 04 25	
PCA slow	00 05 51	00 21 31	06 02 05	10 38 54	
Clustering	< 1 sec	< 1 sec	00 00 20	00 00 56	
SNN	< 1 sec	< 1 sec	Not enough RAM	Not enough RAM	
UMAP	00 00 08	00 00 09	00 00 42	00 01 20	
Marker genes	00 00 17	00 00 36	00 18 24	00 19 41	
Signatures	00 00 03	00 00 13	00 01 06	00 02 57	
Enrichment (for one cluster)	00 00 01	00 00 02	00 00 02	00 00 02	
Cell cycle	< 1 sec	< 1 sec	00 00 01	00 00 02	
Annotation	00 00 01	00 00 01	00 00 02	00 00 03	
Doublets	00 00 22	00 01 00	Not enough RAM	Not enough RAM	
Trajectory	00 00 01	00 00 03	00 05 34	00 06 03	
LR (for a pair of clusters)	00 00 03	00 00 05	00 00 11	00 00 16	
GRNs	00 00 19	00 01 45	00 11 24	Not enough RAM	
Number of Cells	2700	5903	58358	201073	
Input size (RDS format)	21.4 MB	127.5 MB	364.9 MB	718.6 MB	
Machine Specs	CPU: 12th Gen Intel® Core™ i7-12700KF × 20 RAM: 64 GB OS: Ubuntu 20.04.6 LTS				
Max memory consumption	3.6 GB	6.9 GB	31 GB	17 GB (*without TF activity inference step)	
scATAC-seq datasets					
Operations	BMMCs 5k	SFs 6k	Skin 80k	COVID 200k	Time format
QC	< 1 sec	< 1 sec	00 00 03	00 00 06	HH:MM:SS
LSI	00 00 19	00 00 35	00 01 08	00 02 47	
Clustering	00 00 08	00 00 12	00 06 10	04 32 58	
UMAP and tSNE	00 00 24	00 00 31	00 06 52	00 32 11	
Marker genes	00 01 06	00 01 03	00 05 31	00 11 38	
Marker peaks	00 02 23	00 04 23	00 19 46	00 17 15	
Doublets	00 01 21	00 02 07	00 18 33	00 40 06	
Motif enrichment	00 03 20	00 02 34	00 05 45	00 10 35	
Integration	00 04 09	00 09 15	00 29 26	01 16 25	
Trajectory	00 00 04	00 00 02	00 02 38	00 05 11	
GRNs	00 08 12	00 12 38	00 12 40	00 37 43	
Number of cells	4932	6046	78996	200762	
Input size (Arrow format)	584.3 MB	3.2 GB	7.8 GB	11.5 GB	
Specs	CPU: 12th Gen Intel® Core™ i7-12700KF × 20 RAM: 64 GB OS: Ubuntu 20.04.6 LTS				
Max RAM consumption	3.5 GB	8 GB	10.3 GB	43.2 GB	

Figure 51. Benchmark of our application with different datasets as input. In this table datasets of various sizes have been utilized to test execution time and memory consumption for the different modes of analysis offered in our application for scRNA-seq and scATAC-seq data. The hardware specifications of the PC used for the benchmark are shown.

4 Discussion

The application that was developed in the context of this PhD project is a comprehensive bioinformatics pipeline offered both as a web-service and a stand-alone application. It performs end-to-end SC analysis, by using the current best practices of the field. It currently enables the analysis of scRNA-seq and scATAC-seq datasets, which comprise the vast majority of the available SC data to date, facilitating both independent and integrative analysis of the two modalities.

The architecture of our application enables a seamless integration between various software packages offering many modes of analysis to the end user. Both R programming language and web technologies have aided us to develop an interactive application that could be useful to both novice and advanced users.

To showcase its full capabilities, we utilized single cell data from *hTNFtg* mouse model of arthritis, using as input data from both scRNA-seq and scATAC-seq modalities. Moreover, to test whether our results were robust across different pipelines, we used two alternative workflows that resulted in similar results confirming our original findings. Additionally, to achieve comparison between the data from our use case scenario and other publicly available datasets we performed integration analysis.

One important aspect of this dissertation was the study of synovial fibroblast subsets and their properties in both homeostasis and TNF-mediated chronic arthritis at a single-cell resolution. More precisely, we aimed to delineate their transcriptomic profiles, chromatin accessibility, spatial distribution, and the regulatory networks governing the transition from a healthy state to arthritic pathology with precision. Marker gene detection and functional enrichment analysis facilitated the identification of three principal fibroblast super-clusters. Lining synovial fibroblasts (Thy1⁻ LSFs) were observed to modulate the size of the lining layer through apoptotic and migratory mechanisms. Conversely, sublining synovial fibroblasts (Thy1⁺ SLSFs) were responsive to growth factors and differentiation signals, including WNT, BMP, and TGFbeta. However, during the progression of arthritis, distinct subtypes of SLSFs undergo a phenotypic transition, relinquishing their homeostatic functions and acquiring activated characteristics. Consequently, novel arthritis-specific subpopulations emerge, two examples are *Dkk3/Lrrc15*⁺ and *Birc5/Aqp1*⁺ subsets, which demonstrate elevated inflammatory and destructive attributes. These findings underscore the intricate networks orchestrating the pathogenesis of arthritis. Notably, the combinatorial analysis of both modalities played a pivotal role in the prioritization of TFs and regulatory networks, revealing distinct activity patterns across various synovial fibroblast (SF) groups and conditions. Furthermore, a noteworthy section of this study was the integration of SF profiles from the *hTNFtg* mouse model with SFs of the STIA

mouse model, as well as SF profiles derived from RA human patients. Through meticulous examination of gene expression patterns, correlation analyses, and enriched biological pathways, coupled with regulatory network analysis, this integrative approach unveiled both commonalities and disparities among SFs across diverse datasets.

Closing this section, it is important to acknowledge several limitations and prospective future directions for the developed application. Firstly, the application comprises various software packages, and there is no automated mechanism to update all packages simultaneously without risking compatibility issues. Consequently, updating the packages used in SCALA necessitates extensive testing, both in the context of official updates from the development team and independent user updates in local installations. Secondly, a potential future limitation is scalability. Although the application performed effectively with datasets ranging from 6,000 to over 200,000 cells, we observed diminished responsiveness of the graphical components when handling datasets exceeding 50,000 cells. Enhancing the multi-threading implementation could mitigate these issues by delegating the visualization and plotting tasks to one or more dedicated threads.

To enhance and expand the functionality of our application in future releases, several additional features could be considered. Firstly, the integration analysis between different datasets, as well as spatial transcriptomics analysis could be introduced as additional modes. These enhancements can be seamlessly incorporated into our application as they fall within the same framework of the Seurat package. Furthermore, alternative workflows for cell-to-cell communication analysis and trajectory analysis could be incorporated by adding the implementations from the CellChat and Monocle packages. Additionally, the aesthetic aspects of the application could be improved by adding more customizable plotting options, such as different color palettes and additional export format options.

5 Conclusions

This dissertation had two primary objectives. The first was to develop a user-friendly application that integrates various software packages to provide a comprehensive pipeline for the analysis, exploration, and visualization of single-cell RNA sequencing (scRNA-seq) and single-cell assay for transposase-accessible chromatin sequencing (scATAC-seq) data. The second aim was to employ the developed pipeline, along with custom bioinformatics analysis steps, to study and characterize the *hTNFtg* mouse model of rheumatoid arthritis at single-cell resolution.

This study has significantly advanced our understanding of the cellular heterogeneity of synovial fibroblasts in the chronic RA model *hTNFtg* by examining their properties in both healthy and diseased contexts. Utilizing various bioinformatics analysis steps, we identified distinct gene expression programs and regulatory networks that may drive the initiation and progression of pathology. Furthermore, through integration analysis with other datasets, including the STIA mouse model of RA and human data from RA patients, we identified similarities and differences at the level of genes, biological pathways, and cell population dynamics.

Additionally, we developed an application named SCALA, which takes advantage of the R programming language and web technologies such as HTML, JavaScript, CSS, and R/Shiny. SCALA enables automated and interactive analysis of single-cell transcriptomics and epigenomics data. The application is built upon two widely used software packages, Seurat and ArchR, which support key analytical steps, ranging from quality control to the identification of distinct cell populations. SCALA is designed for interoperability with other software packages, thereby facilitating the execution of more complex analytical tasks.

A notable advantage of SCALA is its user-friendly interface, making it accessible to a broad audience within the biomedical community. SCALA is available both online as a web service (<http://scala.fleming.gr/>) and offline through local installation via GitHub (<https://github.com/PavlopoulosLab/SCALA>) or Docker (<https://hub.docker.com/r/pavlopouloslab/scala>). As an open-access R package, SCALA allows advanced users to modify and improve the source code according to their needs.

In conclusion, the bioinformatics methodologies described in this dissertation, alongside the development of a computational pipeline offered as a user-friendly web application, could be highly beneficial for biomedical scientists seeking to analyze and explore their data in an interactive manner. The analysis of a use case scenario involving the *hTNFtg* mouse model of arthritis yielded significant biological findings, enhancing our understanding of the cellular heterogeneity of fibroblast populations. These findings could

aid biologists and researchers from other disciplines in the identification of important genes and pathways that could be useful for diagnostic purposes or therapeutic intervention in the future.

6 Summary

Single-cell technologies have revolutionized biomedical research by enabling the study of the genome, transcriptome, proteome, and epigenome at unprecedented resolution. These experimental assays are further enhanced by spatial transcriptomics methods, such as Visium and CosMx, as well as imaging techniques like CODEX (CO-Detection by indEXing) and RNA-FISH (RNA Fluorescence In Situ Hybridization). These methods facilitate the validation of marker genes identified through bioinformatics analysis by allowing the visualization of RNA transcripts and proteins within complex tissue architectures.

The boom of single-cell assays in the last decade has led to the development of numerous software packages dedicated to data analysis, visualization, and exploration. To date, more than 1,700 tools are available. Both R and Python programming languages are favored by developers, while some tools are written in C++, Matlab, or other languages. Among the most widely used tools in the single-cell community are Seurat, Scanpy, Cicero, and ArchR. The first two are utilized for the analysis of scRNA-seq data, while the latter two are used for the analysis of scATAC-seq data. These tools, in combination with other methodologies, offer various modes of analysis to the end users, including quality control, dimensionality reduction, clustering of cells, identification of marker genes, annotation of cell populations, trajectory analysis, integration of RNA and ATAC assays, motif enrichment analysis, and identification of regulatory networks, to name just a few.

To alleviate potential technical difficulties in installing, managing, and combining various software packages into one pipeline, we developed SCALA. Our application offers seamless integration of different tools in a user-friendly and interactive environment, available both online and as a stand-alone application. To achieve this, we leveraged the R programming language and web technologies such as HTML, JavaScript, CSS, and R/Shiny.

We utilized SCALA for an end-to-end analysis of scRNA-seq and scATAC-seq data originated from synovial fibroblasts (SFs) of the *hTNFtg* mouse model of arthritis. This analysis enabled us to characterize the heterogeneity of SFs during homeostasis, early, and established disease states. Specifically, we observed cell population dynamics and transitions between different cellular states. Additionally, integrating both modalities facilitated the identification of regulatory networks with preferential activity across different cell sub-populations. Moreover, we performed a comparative analysis between SFs in the *hTNFtg* mouse model, the STIA mouse model, and SFs from human RA patients, highlighting similarities and differences at different levels such as in gene expression patterns and enriched biological pathways.

Conclusively, we consider that the scientific material presented in the current dissertation can have a positive impact on the biomedical community in several ways. First, the pipeline we developed is freely available, enabling users to analyze their data or modify the source code and adjust the pipeline to their specific needs since it is an open-source project. Additionally, the biological findings may be of significant interest to researchers specializing in RA disease, as the genes, biological pathways, and master regulators identified by the bioinformatics analysis could be potentially targeted for diagnostic or therapeutic purposes.

7 Περίληψη

Οι τεχνολογίες αλληλούχισης σε επίπεδο μοναδιαίου κυττάρου έχουν φέρει επανάσταση στο πεδίο της ιατροβιολογικής έρευνας, καθώς επιτρέπουν την μελέτη του γονιδιώματος, του μεταγραφώματος, του πρωτεώματος και του επιγονιδιώματος με πρωτοφανή ακρίβεια. Αυτές οι πειραματικές τεχνικές ενισχύονται περαιτέρω από μεθόδους μεταγραφωμικής σε ιστούς, όπως το Visium και το CosMx, καθώς και τεχνικές απεικόνισης όπως το CODEX (CO-Detection by indEXing) και το RNA-FISH (RNA Fluorescence In Situ Hybridization). Αυτές οι μέθοδοι διευκολύνουν την επικύρωση γονιδίων που χαρακτηρίζουν κυτταρικούς πληθυσμούς (και έχουν ταυτοποιηθεί μέσω της βιοπληροφορικής ανάλυσης), επιτρέποντας την οπτικοποίηση μορίων RNA και των πρωτεϊνών μέσα σε ιστούς.

Η ραγδαία ανάπτυξη των τεχνικών μοναδιαίου κυττάρου την τελευταία δεκαετία οδήγησε στην ανάπτυξη εκατοντάδων πακέτων λογισμικού που επικεντρώνονται στην ανάλυση, την οπτικοποίηση και διερεύνηση των δεδομένων. Μέχρι σήμερα, είναι διαθέσιμα περισσότερα από 1.700 υπολογιστικά εργαλεία. Οι γλώσσες προγραμματισμού R και Python προτιμώνται από τους προγραμματιστές έναντι των υπολοίπων, ενώ υπάρχουν και κάποια υπολογιστικά εργαλεία είναι υλοποιημένα σε C ++, Matlab και άλλες γλώσσες προγραμματισμού. Ανάμεσα στα πιο δημοφιλή πακέτα λογισμικού στην επιστημονική κοινότητα είναι το Seurat, Scanpy, Cicero και ArchR. Τα δύο πρώτα χρησιμοποιούνται για την ανάλυση δεδομένων scRNA-seq, ενώ τα άλλα δύο χρησιμοποιούνται για την ανάλυση δεδομένων scATAC-seq. Αυτά τα υπολογιστικά εργαλεία, σε συνδυασμό με άλλες μεθοδολογίες, προσφέρουν στους χρήστες την δυνατότητα να εκτελέσουν βήματα της ανάλυσης με διαφορετική λειτουργικότητα. Μια σύντομη περιγραφή αυτών περιλαμβάνει τα βήματα του ελέγχου ποιότητας των δεδομένων, της μείωσης των αρχικών διαστάσεων των δεδομένων (αναφέρεται στην μείωση των γονιδίων που λαμβάνονται υπόψη για την τελική ανάλυση), της ομαδοποίησης των κυττάρων, της αναγνώρισης γονιδίων που χαρακτηρίζουν κυτταρικούς πληθυσμούς και την απόδοση αυτών σε γνωστούς κυτταρικούς τύπους, την εύρεση σχέσεων κυτταρικής διαφοροποίησης μεταξύ των κυτταρικών πληθυσμών, την συνδυαστική ανάλυση των δεδομένων τύπου RNA και ATAC, της ανάλυσης εμπλουτισμού μοτίβων στο DNA που αναγνωρίζονται από μεταγραφικούς παράγοντες και της εύρεσης ρυθμιστικών δικτύων.

Η ανάπτυξη της εφαρμογής SCALA (που παρουσιάζεται σε αυτή την διατριβή) έγινε με σκοπό να αντιμετωπίσει πιθανές τεχνικές δυσκολίες που θα μπορούσαν να εμφανιστούν σε χρήστες κατά την εγκατάσταση, την διαχείριση ή την συνδυαστική χρήση πολλαπλών πακέτων λογισμικού ανάλυσης δεδομένων αλληλούχισης μοναδιαίων κυττάρων. Η εφαρμογή που αναπτύχθηκε προσφέρει μια απρόσκοπτη διασύνδεση πολλαπλών υπολογιστικών εργαλείων σε ένα διαδραστικό και φιλικό προς τον

χρήστη περιβάλλον. Είναι διαθέσιμη προς χρήση δωρεάν τόσο μέσω Διαδικτύου όσο και σαν εφαρμογή σε προσωπικό υπολογιστή μετά από τοπική εγκατάσταση. Για την επίτευξη του παραπάνω εγχειρήματος αξιοποιήσαμε την γλώσσα προγραμματισμού R και τις τεχνολογίες ιστού HTML, JavaScript, CSS, και R/Shiny.

Χρησιμοποιήσαμε την εφαρμογή μας για να επιτύχουμε μια ολοκληρωμένη ανάλυση των δεδομένων scRNA-seq και scATAC-seq από ινοβλάστες της άρθρωσης του αστραγάλου ποντικών στο μοντέλο αρθρίτιδας *hTNFtg*. Η ανάλυση αυτή μας επέτρεψε να χαρακτηρίσουμε την κυτταρική ετερογένεια των ινοβλαστών κατά τη διάρκεια της ομοιόστασης, και στα στάδια της πρώιμης και μεταγενέστερης φάσης της νόσου. Συγκεκριμένα, παρατηρήσαμε τις μεταβολές στην εκπροσώπηση συγκεκριμένων κυτταρικών πληθυσμών και την διαδικασία μετάβασης μεταξύ διαφορετικών κυτταρικών καταστάσεων. Επιπλέον, η συνδυαστική ανάλυση δεδομένων και από τις δύο πειραματικές τεχνικές διευκόλυνε τον εντοπισμό δικτύων ρύθμισης με δραστηριότητα σε διαφορετικές ομάδες κυττάρων. Επιπρόσθετα, πραγματοποιήσαμε συγκριτική ανάλυση μεταξύ των ινοβλαστών από το μοντέλο ποντικού *hTNFtg*, από το μοντέλο ποντικού STIA και από ινοβλάστες που προέρχονται από ασθενείς με ρευματοειδή αρθρίτιδα, επισημαίνοντας τις ομοιότητες και τις διαφορές τους σε πολλαπλά επίπεδα όπως στο κομμάτι της γονιδιακής έκφρασης και το κομμάτι των ενεργοποιημένων βιολογικών μονοπατιών (μετά από ανάλυση εμπλουτισμού).

Συμπερασματικά, θεωρούμε ότι το επιστημονικό υλικό που παρουσιάζεται στην παρούσα διδακτορική διατριβή μπορεί να έχει θετικό αντίκτυπο στη βιοϊατρική κοινότητα σε πολλαπλά επίπεδα. Αρχικά, η υπολογιστική εφαρμογή που αναπτύξαμε είναι διαθέσιμη δωρεάν, επιτρέποντας στους τελικούς χρήστες να αναλύσουν τα δεδομένα τους ή ακόμα και να τροποποιήσουν τον πηγαίο κώδικα και να τον προσαρμόσουν στις δικές τους ανάγκες, καθώς πρόκειται για μια εφαρμογή ανοικτού λογισμικού. Επιπλέον, τα βιολογικά ευρήματα μπορούν να αποτελέσουν αντικείμενο ενδιαφέροντος για ερευνητές που εξειδικεύονται στη μελέτη της νόσου της ρευματοειδούς αρθρίτιδας, καθώς κάποια από τα γονίδια, τα βιολογικά μονοπάτια και τους ρυθμιστικούς παράγοντες που εντοπίστηκαν μέσω της βιοπληροφορικής ανάλυσης θα μπορούσαν να χρησιμοποιηθούν ενδεχομένως για διαγνωστικούς ή θεραπευτικούς σκοπούς.

8 Acronyms

ACPAs	anti-citrullinated protein antibodies
Acta2	Actin alpha 2
Adamdec1	ADAM-like, decysin 1
AI	artificial intelligence
AIA	Adjuvant-induced arthritis model
AP1	Activator protein 1
Aqp1	Aquaporin 1
ARE	AU-rich elements
ATAC	Assay for Transposase-Accessible Chromatin
AUC	Area Under the Curve
AXIN1	Axin 1
B2m	Beta-2 microglobulin
Birc5	Baculoviral IAP repeat-containing 5
Bmp4	Bone morphogenetic protein 4
bp	base pairs
CAFs	cancer associated fibroblasts
CAIA	Collagen antibody-induced arthritis model
Ccl	C-C motif chemokine ligand
CCR	C-C motif chemokine receptor
CD	Crohn's disease
CD45	Cluster of Differentiation 45
cDNA	Complementary DNA
CIA	Type II collagen induced arthritis
circRNAs	circular RNAs
Clic5	Chloride intracellular channel 5
Clu	Clusterin
Coch	Coagulation factor C homolog
Col15a1	Collagen, type XV, alpha 1
Comp	Cartilage Oligomeric Matrix Protein
CPU	Central Processing Unit
cRNA	Complementary RNA
CSS	Cascade Style Sheet
Cthrc1	Collagen triple helix repeat containing protein 1
Cxcl	Chemokine (C-X-C motif) ligand
Dkk	Dickkopf WNT Signaling Pathway Inhibitor
Dlx3	Distal-less homeobox 3
DNA	Deoxyribonucleic Acid
ECM	Extracellular Matrix
Ecr4	ECRG4 augurin precursor
Efemp1	EGF containing fibulin extracellular matrix protein 1

EPCAM	Epithelial Cell Adhesion Molecule
FAP	Fibroblast Activation Protein Alpha
Fbln1	Fibulin 1
FLS	fibroblast like synoviocytes
G6PI	glucose-6-phosphate isomerase protein
GBs	Gigabytes
GM-CSF	Granulocyte- Macrophage Colony- Stimulating Factor
GO	Gene Ontology
GRN	Gene Regulatory Network
GUI	Graphic User Interface
H&E	Hematoxylin and Eosin
H2-K1	Histocompatibility 2, K1, K region
H2-Q7	Histocompatibility 2, Q region locus 7
Hbegf	Heparin-binding EGF-like growth factor
Hhip	Hedgehog-interacting protein
HLA-DRA	Major histocompatibility complex, class II, DR alpha
Hmox1	Heme oxygenase 1
HTML	Hyper Text Markup Language
hTNFtg	human TNF transgene
Htra4	HtrA serine peptidase 4
IBD	Inflammatory Bowel Disease
Id1	Inhibitor of DNA binding 1, HLH protein
IFNs	Interferons
IL	Interleukin
ISREs	IFN-stimulated response elements
JNK	c-Jun N-terminal kinase JNK
K/BxN	a cross between non-obese diabetic mice (NOD) and mice with a KRN T-cell receptor transgene (K/B)
KEGG	Kyoto Encyclopedia of Genes and Genomes
Klf5	Kruppel like factor 5
lncRNAs	long noncoding RNAs
L-R	Ligand - Receptor
Lrrc15	Leucine rich repeat containing 15
LSFs	Lining Synovial Fibroblasts
LSI	Latent Semantic Indexing
MAP	Mitogen-Activated Protein
MDS	Multi-Dimensional Scaling
Meox1	Mesenchyme Homeobox 1
MEX	Market Exchange Format
Mki67	Antigen identified by monoclonal antibody Ki 67
ML	machine learning

MMPs	matrix metalloproteinases
mRNA	messenger RNA
MST	Minimal Spanning Tree
mTNF	transmembrane TNF
MVP	Mean-Variance Plot
ncRNA	noncoding RNA
NFκB	Nuclear factor kappa-light-chain-enhancer of activated B cells
NGS	Next-Generation Sequencing
Notch3	Notch receptor 3
Npnt	Nephronectin
Nr2f2	Nuclear receptor subfamily 2, group F, member 2
Osr1	Odd-skipped related transcription factor 1
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
Pdgfa	Platelet derived growth factor subunit A
PDGFRA	Platelet Derived Growth Factor Receptor Alpha
PDPN	Podoplanin
PHATE	Potential of Heat-diffusion for Affinity-based Trajectory Embedding
Pi16	Peptidase Inhibitor 16
PI3K	Phosphoinositide 3-kinase
PIA	pristane-induced arthritis
Prg4	Proteoglycan 4
QC	Quality Control
RA	Rheumatoid arthritis
RAG	recombination activation gene
RAM	Random Access Memory
RF	rheumatoid factor
Rgma	Repulsive guidance molecule family member A
RNA	Ribonucleic Acid
RNA-seq	RNA sequencing
RSPO2	R-spondin 2
Saa3	Serum amyloid A 3
SC	single cell
scATAC-seq	single cell ATAC sequencing
SCID	severe combined immunodeficiency disease
scRNA-seq	single cell RNA sequencing
Sema3c	Sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C
Sfrp1	Secreted frizzled-related protein 1
SLE	Systemic Lupus Erythematosus
SLSFs	Sublining Synovial Fibroblasts
SMAD	SMA and MAD-related protein

Smoc2	SPARC related modular calcium binding 2
SNN	Shared-Nearest Neighbor
SNP	Single Nucleotide Polymorphism
STIA	Serum Transfer Induced arthritis
sTNF	soluble TNF
SVD	Singular Value Decomposition
TACE	TNF- α converting enzyme
Ter119	Mouse TER-119 Erythroid Antigen MAb (Clone TER-119)
TF	transcription factor
TGF- β	transforming growth factor beta
Th17	T helper 17 cells
Thbs1	Thrombospondin 1
THY1	Thy-1 cell surface antigen
Tlr2	Toll-like receptor 2
TNF	Tumor Necrosis Factor
Tnfaip	Tumor necrosis factor alpha induced protein 6
TNFR1	TNF receptor I
TNFR2	TNF receptor II
TRAF2	TNF receptor-associated factor 2
Tregs	regulatory T cells
tSNE	t-Distributed Stochastic Neighbor Embedding
Tspan15	Tetraspanin 15
TSS	transcription Start Site
UMAP	uniform manifold approximation and projection
UMI	unique molecular identifier
VEGF	vascular endothelial growth factor
VST	Variance Stabilizing Transformation
Vwa	von Willebrand factor A domain
WNT	Wingless and Int-1
wt	Wild Type
ZAP-70	Zeta chain associated protein kinase 70kDa
ZIA	Zymosan induced arthritis

9 Παράρτημα

9.1 Προϋποθέσεις απόκτησης διδακτορικού

Δημοσιεύσεις σχετικές με την διδακτορική διατριβή σε επιστημονικά περιοδικά της βάσεως του PubMed

Armaka M, Konstantopoulos D, Tzaferis C, Lavigne MD, Sakkou M, Liakos A, Sfikakis PP, Dimopoulos MA, Fousteri M, Kollias G. Single-cell multimodal analysis identifies common regulatory programs in synovial fibroblasts of rheumatoid arthritis patients and modeled TNF-driven arthritis. *Genome Med.* 2022 Jul 26;14(1):78. doi: 10.1186/s13073-022-01081-3. **PMID: 35879783**; PMCID: PMC9316748. [Ισοδύναμος πρώτος συγγραφέας]

Tzaferis C, Karatzas E, Baltoumas FA, Pavlopoulos GA, Kollias G, Konstantopoulos D. SCALA: A complete solution for multimodal analysis of single-cell Next Generation Sequencing data. *Comput Struct Biotechnol J.* 2023 Oct 20;21:5382-5393. doi: 10.1016/j.csbj.2023.10.032. **PMID: 38022693**; PMCID: PMC10651449. [Πρώτος συγγραφέας]

Παρουσίαση της μεθοδολογίας ή μέρος των αποτελεσμάτων της διδακτορικής διατριβής σε επιστημονικό συνέδριο στην Ελλάδα ή το εξωτερικό

16th Conference of the Hellenic Society for Computational Biology and Bioinformatics (HSCBB22, 2022) [Τίτλος ομιλίας: “SCALA: A web application for multimodal analysis of single-cell next-generation sequencing data”]

Διαθεσιμότητα του πηγαίου κώδικα και της υπολογιστικής πλατφόρμας που αναπτύχθηκε

<https://scala.fleming.gr>

<https://github.com/PavlopoulosLab/SCALA>

<https://hub.docker.com/r/pavlopouloslab/scala>

Διαθεσιμότητα των βιολογικών δεδομένων που χρησιμοποιήθηκαν

<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA778928>



Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

SCALA: A complete solution for multimodal analysis of single-cell Next Generation Sequencing data

Christos Tzaferis^a, Evangelos Karatzas^b, Fotis A. Baltoumas^b, Georgios A. Pavlopoulos^{b,c,*}, George Kollias^{a,c,d,**}, Dimitris Konstantopoulos^{a,***}^a Institute for Bioinnovation, Biomedical Sciences Research Center "Alexander Fleming", Vari, Greece^b Institute for Fundamental Biomedical Research, Biomedical Sciences Research Center "Alexander Fleming", Vari, Greece^c Research Institute of New Biotechnologies and Precision Medicine, National and Kapodistrian University of Athens, Greece^d Department of Physiology, Medical School, National and Kapodistrian University of Athens, Greece

ARTICLE INFO

Keywords:

Single-cell RNA sequencing analysis
 Single-cell ATAC-seq analysis
 Automated analysis of single-cell Next Generation Sequencing data
 Integrative analysis of single-cell Next Generation Sequencing data

ABSTRACT

Analysis and interpretation of high-throughput transcriptional and chromatin accessibility data at single-cell (sc) resolution are still open challenges in the biomedical field. The existence of countless bioinformatics tools, for the different analytical steps, increases the complexity of data interpretation and the difficulty to derive biological insights. In this article, we present SCALA, a bioinformatics tool for analysis and visualization of single-cell RNA sequencing (scRNA-seq) and Assay for Transposase-Accessible Chromatin using sequencing (scATAC-seq) datasets, enabling either independent or integrative analysis of the two modalities. SCALA combines standard types of analysis by integrating multiple software packages varying from quality control to the identification of distinct cell populations and cell states. Additional analysis options enable functional enrichment, cellular trajectory inference, ligand-receptor analysis, and regulatory network reconstruction. SCALA is fully parameterizable, presenting data in tabular format and producing publication-ready visualizations. The different available analysis modules can aid biomedical researchers in exploring, analyzing, and visualizing their data without any prior experience in coding. We demonstrate the functionality of SCALA through two use-cases related to TNF-driven arthritic mice, handling both scRNA-seq and scATAC-seq datasets. SCALA is developed in R, Shiny and JavaScript and is mainly available as a standalone version, while an online service of more limited capacity can be found at <http://scala.pavlopouloslab.info> or <https://scala.fleming.gr>.

1. Introduction

Single-cell RNA sequencing (scRNA-seq) and ATAC sequencing (scATAC-seq) are both Next Generation Sequencing (NGS) techniques that have enabled the study of the transcriptome and epigenome, respectively, at an unprecedented resolution [1–5]. Exploitation of these two modalities allows researchers to observe the heterogeneity of cell populations in more depth compared to established bulk RNA-sequencing techniques.

Since the first scRNA-seq publication [6], advances in technology and equipment have led to an exponential increase in the number of cells

(from hundreds to millions) that can be simultaneously sequenced in one run. Widely used technologies that have been introduced over the past ten years include Fluidigm C1 [7], Smart-seq2 [8], Drop-seq [9] and 10x Genomics [10], whereas new protocols such as the 10x multiome and spatial transcriptomics [11] have also emerged. Both scRNA-seq and scATAC-seq techniques have been used in various experimental settings such as the investigation of different tissues, developmental timepoints, disease states and organisms. ScRNA-seq for example, has played a crucial role in the comprehensive annotation of cell types in multiple organisms (e.g., Human Cell Atlas for *Homo sapiens* [12], Tabula Muris for *Mus musculus* [13]), as well as in the identification of novel cell

* Corresponding author at: Institute for Fundamental Biomedical Research, Biomedical Sciences Research Center "Alexander Fleming", Vari, Greece.

** Corresponding author at: Institute for Bioinnovation, Biomedical Sciences Research Center "Alexander Fleming", Vari, Greece.

*** Corresponding author.

E-mail addresses: pavlopoulos@fleming.gr (G.A. Pavlopoulos), kollias@fleming.gr (G. Kollias), konstantopoulos@fleming.gr (D. Konstantopoulos).¹ Present Address: Georgios A. Pavlopoulos; George Kollias, Dimitris Konstantopoulos; Biomedical Sciences Research Center "Alexander Fleming", 34 Fleming Street, Vari, 16672, Greece.<https://doi.org/10.1016/j.csbj.2023.10.032>

Received 11 June 2023; Received in revised form 16 October 2023; Accepted 16 October 2023

Available online 20 October 2023











2001-0370/© 2023 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

RESEARCH

Open Access



Single-cell multimodal analysis identifies common regulatory programs in synovial fibroblasts of rheumatoid arthritis patients and modeled TNF-driven arthritis

Marietta Armaka^{1*†}, Dimitris Konstantopoulos^{1†}, Christos Tzaferis^{2,3†}, Matthieu D. Lavigne^{1,4†},
Maria Sakkou^{2,3,5†}, Anastasios Liakos¹, Petros P. Sfikakis^{3,6,7}, Meletios A. Dimopoulos^{3,8},
Maria Fousteri^{1*†} and George Kollias^{2,3,5,7*†}

Abstract

Background: Synovial fibroblasts (SFs) are specialized cells of the synovium that provide nutrients and lubricants for the proper function of diarthrodial joints. Recent evidence appreciates the contribution of SF heterogeneity in arthritic pathologies. However, the normal SF profiles and the molecular networks that govern the transition from homeostatic to arthritic SF heterogeneity remain poorly defined.

Methods: We applied a combined analysis of single-cell (sc) transcriptomes and epigenomes (scRNA-seq and scATAC-seq) to SFs derived from naïve and *hTNFg* mice (mice that overexpress human TNF, a murine model for rheumatoid arthritis), by employing the Seurat and ArchR packages. To identify the cellular differentiation lineages, we conducted velocity and trajectory analysis by combining state-of-the-art algorithms including scVelo, Slingshot, and PAGA. We integrated the transcriptomic and epigenomic data to infer gene regulatory networks using ArchR and custom-implemented algorithms. We performed a canonical correlation analysis-based integration of murine data with publicly available datasets from SFs of rheumatoid arthritis patients and sought to identify conserved gene regulatory networks by utilizing the SCENIC algorithm in the human arthritic scRNA-seq atlas.

Results: By comparing SFs from healthy and *hTNFg* mice, we revealed seven homeostatic and two disease-specific subsets of SFs. In healthy synovium, SFs function towards chondro- and osteogenesis, tissue repair, and immune surveillance. The development of arthritis leads to shrinkage of homeostatic SFs and favors the emergence of SF profiles

[†]Marietta Armaka, Dimitris Konstantopoulos, Christos Tzaferis, Matthieu D. Lavigne, and Maria Sakkou contributed equally to this work.

[†]Marietta Armaka, Maria Fousteri, and George Kollias jointly supervised the study.

*Correspondence: armaka@fleming.gr; fousteri@fleming.gr; kollias@fleming.gr

¹ Institute for Fundamental Biomedical Research, Biomedical Sciences Research Center "Alexander Fleming", Vari, Greece

² Institute for Bioinnovation, Biomedical Sciences Research Center "Alexander Fleming", Vari, Greece

Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Research Article

Discovery of the First-in-Class Inhibitors of Hypoxia Up-Regulated Protein 1 (HYOU1) Suppressing Pathogenic Fibroblast Activation

Dimitra Papadopoulou, Vasiliki Mavrikaki, Filippos Charalampous, Christos Tzaferis, Martina Samiotaki, Konstantinos D. Papavasilelou, Antreas Afantitis, Niki Karagianni, Maria C. Denis ... See all authors

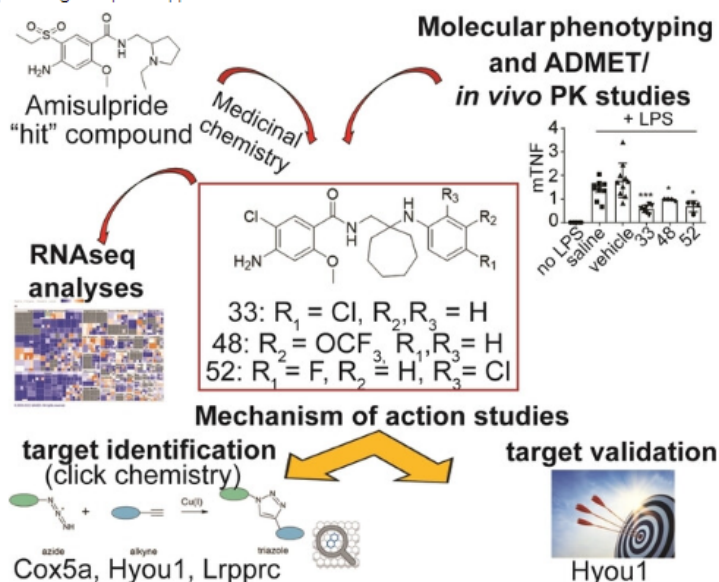
First published: 10 February 2024 | <https://doi.org/10.1002/anie.202319157>

Read the full text >

PDF TOOLS SHARE

Graphical Abstract

A multidisciplinary pipeline consisting of medicinal chemistry, molecular phenotyping, chemoproteomics, RNA-sequencing, short hairpin RNAs (shRNAs)-mediated gene silencing and *in vivo* studies led to a novel series of compounds suppressing pathogenic fibroblast activation via Hypoxia up-regulated protein 1 (HYOU1) inhibition. The first reported HYOU1 inhibitors are presented as a promising therapeutic approach in fibroblast-related diseases.



Abstract

Fibroblasts are key regulators of inflammation, fibrosis, and cancer. Targeting their activation in these complex diseases has emerged as a novel strategy to restore tissue homeostasis. Here, we present a multidisciplinary lead discovery approach to identify and optimize small molecule inhibitors of pathogenic fibroblast activation. The study encompasses medicinal chemistry, molecular phenotyping assays, chemoproteomics, bulk RNA-sequencing analysis, target validation experiments, and chemical absorption, distribution, metabolism, excretion and toxicity (ADMET)/pharmacokinetic (PK)/*in vivo* evaluation. The parallel synthesis employed for the production of the new benzamide derivatives enabled us to a) pinpoint key structural elements of the scaffold that provide potent fibroblast-deactivating effects in cells, b) discriminate atoms or groups that favor or disfavor a desirable ADMET profile, and c) identify metabolic "hot spots". Furthermore, we report the discovery of the first-in-class inhibitor leads for hypoxia up-regulated protein 1 (HYOU1), a member of the heat shock protein 70 (HSP70) family often associated with cellular stress responses, particularly under hypoxic conditions. Targeting HYOU1 may therefore represent a potentially novel strategy to modulate fibroblast activation and treat chronic inflammatory and fibrotic disorders.

Repurposing the antipsychotic drug amisulpride for targeting synovial fibroblast activation in arthritis

Dimitra Papadopoulou,¹ Fani Roumelioti,¹ Christos Tzaferis,^{1,2} Panagiotis Chouvardas,¹ Anna-Kathrine Pedersen,³ Filippos Charalampous,¹ Eleni Christodoulou-Vafeiadou,⁴ Lydia Ntari,⁴ Niki Karagianni,⁴ Maria C. Denis,⁴ Jesper V. Olsen,³ Alexios N. Matralis,¹ and George Kollias^{1,2,5}

¹Institute for Bioinnovation, Biomedical Sciences Research Centre Alexander Fleming, Vari, Greece. ²Department of Physiology, School of Medicine, National and Kapodistrian University of Athens, Athens, Greece. ³Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. ⁴Biomedcode Hellas SA, Vari, Greece. ⁵Center of New Biotechnologies & Precision Medicine, School of Medicine, National and Kapodistrian University of Athens, Athens, Greece.

Synovial fibroblasts (SFs) are key pathogenic drivers in rheumatoid arthritis (RA). Their *in vivo* activation by TNF is sufficient to orchestrate full arthritic pathogenesis in animal models, and TNF blockade proved efficacious for a high percentage of patients with RA albeit coindicating rare but serious side effects. Aiming to find new potent therapeutics, we applied the L1000CDS² search engine, to repurpose drugs that could reverse the pathogenic expression signature of arthritogenic human TNF-transgenic (*hTNFtg*) SFs. We identified a neuroleptic drug, namely amisulpride, which reduced SFs' inflammatory potential while decreasing the clinical score of *hTNFtg* polyarthritis. Notably, we found that amisulpride function was neither through its known targets dopamine receptors D2 and D3 and serotonin receptor 7 nor through TNF-TNF receptor I binding inhibition. Through a click chemistry approach, potentially novel targets of amisulpride were identified, which were further validated to repress *hTNFtg* SFs' inflammatory potential *ex vivo* (*Ascc3* and *Sec62*), while phosphoproteomics analysis revealed that treatment altered important fibroblast activation pathways, such as adhesion. Thus, amisulpride could prove beneficial to patients experiencing RA and the often-accompanying comorbid dysthymia, reducing SF pathogenicity along with its antidepressive activity, serving further as a "lead" compound for the development of novel therapeutics against fibroblast activation.

Introduction

Rheumatoid arthritis (RA) is a chronic inflammatory disease characterized by swelling and gradual destruction of the joints. Synovial fibroblasts (SFs) are one of the main cell types that hyperproliferate during RA progression, producing inflammatory cytokines/chemokines and degrading metalloproteinases, leading progressively to increased joint inflammation, stiffness, and pain (1). We have shown that mice overexpressing TNF, by carrying a human TNF transgene (*hTNFtg* mice) (2) or by deletion of ARE elements in the endogenous *Tnf* gene (*TNF^{ΔARE}* mice) (3), spontaneously develop chronic polyarthritis, with histological manifestations fully resembling human RA. Notably, arthritic pathology in both models develops independently of the adaptive immune response highlighting the dominant role of the innate/stromal compartment in the development of disease (3, 4). Most importantly, TNF signaling via TNF receptor I (TNFR1) in SFs was found to be both required and necessary for the orchestration of full RA-like pathology (5, 6). Interestingly, *hTNFtg* SFs have been found to highly correlate with RA human fibroblast-like synoviocytes (FLS) both at the bulk (7) and at the single-cell RNA levels (8–10). Together, these findings established in principle the dominant *in vivo* role of SFs in the initiation and progression of chronic polyarthritis and suggested a mechanism that may also explain, at least in part, the development of joint pathology in the human disease.

Current first-line therapies against RA, including disease-modifying antirheumatic drugs (DMARDs), such as methotrexate, and targeted synthetic DMARDs inhibiting several kinases, such as Janus kinases or mitogen-activated protein kinase (MAPK), offer significant clinical benefits (11). However, although

Conflict of interest: The authors have declared that no conflict of interest exists.

Copyright: © 2023, Papadopoulou et al. This is an open access article published under the terms of the Creative Commons Attribution 4.0 International License.

Submitted: August 31, 2022

Accepted: March 28, 2023

Published: May 8, 2023

Reference information: *JCI Insight*. 2023;8(9):e165024.
<https://doi.org/10.1172/jci.insight.165024>.

ARTICLE

Lung tumor MHCII immunity depends on in situ antigen presentation by fibroblasts

Dimitra Kerdidani¹, Emmanouil Aerakis¹, Kleio-Maria Verrou², Ilias Angelidis¹, Katerina Douka¹, Maria-Anna Maniou¹, Petros Stamoulis¹, Katerina Goudevenou¹, Alejandro Prados¹, Christos Tzaferis^{1,2}, Vasileios Ntafis³, Ioannis Vamvakaris⁴, Evangelos Kaniaris⁵, Konstantinos Vachlas⁶, Evangelos Sepsas⁶, Anastasios Koutsopoulos⁷, Konstantinos Potaris⁶, and Maria Tsoumakidou^{1,2}

A key unknown of the functional space in tumor immunity is whether CD4 T cells depend on intratumoral MHCII cancer antigen recognition. MHCII-expressing, antigen-presenting cancer-associated fibroblasts (apCAFs) have been found in breast and pancreatic tumors and are considered to be immunosuppressive. This analysis shows that antigen-presenting fibroblasts are frequent in human lung non-small cell carcinomas, where they seem to actively promote rather than suppress MHCII immunity. Lung apCAFs directly activated the TCRs of effector CD4 T cells and at the same time produced C1q, which acted on T cell C1qbp to rescue them from apoptosis. Fibroblast-specific MHCII or C1q deletion impaired CD4 T cell immunity and accelerated tumor growth, while inducing C1qbp in adoptively transferred CD4 T cells expanded their numbers and reduced tumors. Collectively, we have characterized in the lungs a subset of antigen-presenting fibroblasts with tumor-suppressive properties and propose that cancer immunotherapies might be strongly dependent on in situ MHCII antigen presentation.

Introduction

The series of immunological events that takes place between tumors and tumor-draining LNs forms a cyclic trajectory that is being referred to as the cancer-immunity cycle (Chen and Mellman, 2013). In the first step of these events, tumor antigens are carried to the tumor-draining LNs and are partly transferred to resident dendritic cells (DCs; Ruhland et al., 2020). In LNs, migratory and resident DCs present their antigenic cargo to antigen-inexperienced (naive) T cells, which become differentiated effector cells that egress from LNs and enter tumors. In tumors, CD8 cells exhibit direct killing activity against cancer cells, but they are seriously dependent on CD4 T cells for function and transition to memory cells (Ahrends et al., 2019; Binnewies et al., 2019; Śledzińska et al., 2020; Zander et al., 2019; Bos and Sherman, 2010; Schietinger et al., 2010). Although our current understanding of the functional space in the cancer-immunity cycle is that MHCII cancer antigen presentation primarily occurs in LNs, the contribution of in situ cancer antigen presentation has not been ruled out (Dammeijer et al., 2020; Oh et al., 2020).

A few studies have directly addressed the role of peripheral antigen presentation in CD4 T cell responses (Dammeijer et al., 2020; Doebis et al., 2011; Low et al., 2020; McLachlan and

Jenkins, 2007; Schøller et al., 2019). In cancer, three lines of evidence support that the TCRs are stimulated in situ within solid tumors. First, the CD4 TCR repertoire is regionally shaped by the local neoantigen load (Joshi et al., 2019). Second, stemlike CD8 T cells reside in dense MHCII-expressing cell niches within tumors (Jansen et al., 2019). Third, right flank tumors that differ only in one MHCII neoantigen with left flank tumors are infiltrated by higher numbers of neoantigen-specific CD4⁺ T cells (Alspach et al., 2019). DCs are scarce and immature within solid tumors and are generally considered to exert their primary effects in LNs (Dammeijer et al., 2020; Ferris et al., 2020; Maier et al., 2020; Ruhland et al., 2020; Oh et al., 2020). Because structural tissue cells greatly outnumber professional antigen-presenting cells, express immune genes (Krausgruber et al., 2020), and can be induced to present antigens (Koyama et al., 2019; Low et al., 2020), we hypothesized that they are required for local antigen presentation and anti-tumor immunity.

Fibroblasts are largely considered to be immunosuppressive cells (Biffi and Tuveson, 2021). Accordingly, the recently identified MHCII⁺ antigen-presenting cancer-associated fibroblasts (apCAFs) in pancreatic adenocarcinoma (PDAC) and breast carcinoma (BC) are presumed to induce immune tolerance

¹Institute of Bioinnovation, Biomedical Sciences Research Center "Alexander Fleming," Vari, Greece; ²Greek Research Infrastructure for Personalized Medicine, Medical School, National and Kapodistrian University of Athens, Athens, Greece; ³Animal House Facility, Biomedical Sciences Research Center "Alexander Fleming," Vari, Greece; ⁴Department of Pathology, Sotiria Chest Hospital, Athens, Greece; ⁵Department of Respiratory Medicine, Sotiria Chest Hospital, Athens, Greece; ⁶Department of Thoracic Surgery, Sotiria Chest Hospital, Athens, Greece; ⁷Department of Pathology, Medical School, University of Crete, Crete, Greece.

Correspondence to Maria Tsoumakidou: tsoumakidou@fleming.gr.

© 2022 Kerdidani et al. This article is distributed under the terms of an Attribution–Noncommercial–Share Alike–No Mirror Sites license for the first six months after the publication date (see <http://www.rupress.org/terms/>). After six months it is available under a Creative Commons License (Attribution–Noncommercial–Share Alike 4.0 International license, as described at <https://creativecommons.org/licenses/by-nc-sa/4.0/>).



Col6a1⁺/CD201⁺ mesenchymal cells regulate intestinal morphogenesis and homeostasis

Maria-Theodora Melissari¹ · Ana Henriques^{1,6} · Christos Tzaferis² · Alejandro Prados^{2,6} · Michalis E. Sarris¹ · Niki Chalkidi¹ · Dimitra Mavroei¹ · Panagiotis Chouvardas^{2,3,4} · Sofia Grammenoudi¹ · George Kollias^{2,5} · Vasiliki Koliaraki¹

Received: 29 April 2021 / Revised: 26 November 2021 / Accepted: 1 December 2021 / Published online: 15 December 2021
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2021

Abstract

Intestinal mesenchymal cells encompass multiple subsets, whose origins, functions, and pathophysiological importance are still not clear. Here, we used the *Col6a1*^{Cre} mouse, which targets distinct fibroblast subsets and perivascular cells that can be further distinguished by the combination of the CD201, PDGFR α and α SMA markers. Developmental studies revealed that the *Col6a1*^{Cre} mouse also targets mesenchymal aggregates that are crucial for intestinal morphogenesis and patterning, suggesting an ontogenic relationship between them and homeostatic PDGFR α ^{hi} telocytes. Cell depletion experiments in adulthood showed that *Col6a1*⁺/CD201⁺ mesenchymal cells regulate homeostatic enteroendocrine cell differentiation and epithelial proliferation. During acute colitis, they expressed an inflammatory and extracellular matrix remodelling gene signature, but they also retained their properties and topology. Notably, both in homeostasis and tissue regeneration, they were dispensable for normal organ architecture, while CD34⁺ mesenchymal cells expanded, localised at the top of the crypts, and showed increased expression of villous-associated morphogenetic factors, providing thus evidence for the plasticity potential of intestinal mesenchymal cells. Our results provide a comprehensive analysis of the identities, origin, and functional significance of distinct mesenchymal populations in the intestine.

Keywords Fibroblasts · Colitis · Tissue damage · Cell plasticity

Maria-Theodora Melissari and Ana Henriques contributed equally to this work.

✉ Vasiliki Koliaraki
koliaraki@fleming.gr

- ¹ Institute for Fundamental Biomedical Research, Biomedical Sciences Research Center (B.S.R.C.) “Alexander Fleming”, 16672 Vari, Greece
- ² Institute for Bioinnovation, Biomedical Sciences Research Center (B.S.R.C.) “Alexander Fleming”, 16672 Vari, Greece
- ³ Department of Medical Oncology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland
- ⁴ Department for BioMedical Research, University of Bern, Bern, Switzerland
- ⁵ Department of Physiology, Medical School, National and Kapodistrian University of Athens, 11527 Athens, Greece
- ⁶ Present Address: Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain

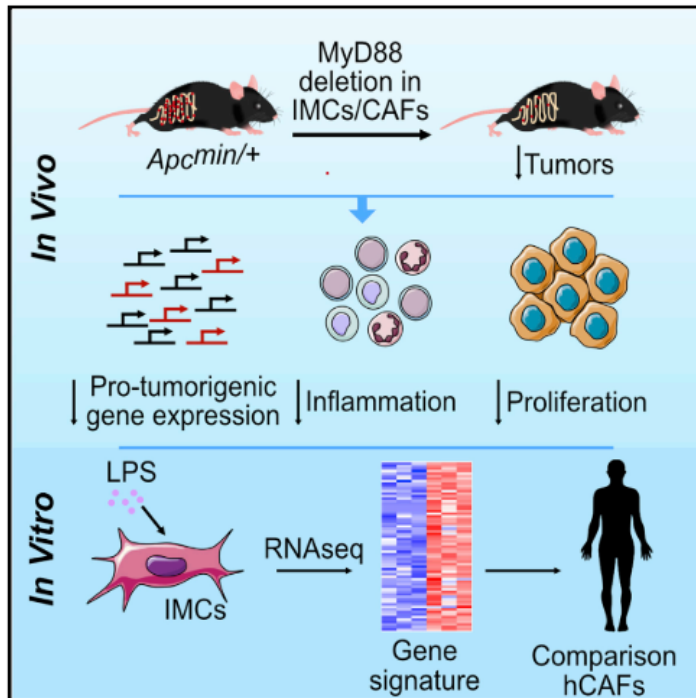
Introduction

The mammalian intestine is characterized by a unique architecture, which ensures both efficient nutrient and water absorption and rapid self-renewal of the intestinal epithelium. Self-renewal is mediated by Lgr5⁺ multi-potent crypt-base stem cells (CBCs) that progressively give rise to transit amplifying (TA) progenitor cells and differentiated epithelial cell populations with specific absorptive or secretive functions [1, 2]. The tight regulation of this architectural organization is mediated by a gradient of factors produced both by epithelial and stromal cells. Among stromal cells, intestinal mesenchymal cells (IMCs) have emerged as an important cell type for the development and homeostasis of the intestine, by providing both structural support and regulatory elements [3]. Of particular interest is their contribution to the maintenance of the stem cell niche via the production of soluble mediators [4]. Notably, in the absence of epithelial Wnts, production of stromal Wnts is sufficient for the

Cell Reports

Innate Sensing through Mesenchymal TLR4/MyD88 Signals Promotes Spontaneous Intestinal Tumorigenesis

Graphical Abstract



Authors

Vasiliki Koliaraki, Niki Chalkidi, Ana Henriques, ..., David J. Hackam, Manolis Pasparakis, George Kollias

Correspondence

koliaraki@fleming.gr (V.K.),
kollias@fleming.gr (G.K.)

In Brief

Koliaraki et al. show that MyD88 in mesenchymal cells is responsible for its tumor-promoting role in the *Apc^{min/+}* model. They further show that this is a TLR4-mediated mechanism that leads to the production of pro-tumorigenic molecules, also identified in human CAFs.

Highlights

- Deletion of MyD88 or TLR4 in IMCs and/or CAFs leads to reduced intestinal tumorigenesis
- The phenotype of the IMC-specific MyD88 mice is similar to the complete knockouts
- MyD88^{-/-} IMCs show a reduced pro-tumorigenic and/or inflammatory gene expression profile
- Human CAFs show upregulation of a similar MyD88-specific gene expression signature



Koliaraki et al., 2019, Cell Reports 26, 536–545
January 15, 2019 © 2018 The Author(s).
<https://doi.org/10.1016/j.celrep.2018.12.072>

CellPress

Cellular complexity and crosstalk in murine TNF-dependent ileitis: Different fibroblast subsets propel spatially defined ileal inflammation through TNFR1 signalling

Lida Iliopoulou,^{*1} Christos Tzaferis,^{*1} Alejandro Prados,^{*2} Fani Roumelioti,¹ Vasiliki Koliaraki,³ and George Kollias^{1,4}

¹Institute for Bioinnovation, Biomedical Sciences Research Center "Alexander Fleming", Vari, Greece

²Institute for Research in Biomedicine (IRB Barcelona), Barcelona Institute of Science and Technology, Barcelona, Spain

³Institute for Fundamental Biomedical Research, Biomedical Sciences Research Center "Alexander Fleming", Vari, Greece

⁴Department of Physiology, Medical School, National and Kapodistrian University of Athens, Athens, Greece

***Contributed equally**

Abstract:

Crohn's disease represents a persistent inflammatory disorder primarily affecting the terminal ileum. Through the application of single-cell RNA sequencing, we unveil the intricate cellular complexities within murine TNF-dependent ileitis, developing in *Tnf^{ΔARE}* mice. Detailed immune cell analysis highlights B cell expansion, T cell effector reprogramming, and macrophage lineage shifts during inflammation. Focusing on stromal cells, we reveal a strong pro-inflammatory character, acquired by all fibroblast subsets, which exhibit complex communication patterns with the infiltrating immune and surrounding stromal cells. Interestingly, we identify that *Tnf^{ΔARE}*-induced ileitis is initiated in the lamina propria via TNFR1 pathway activation in villus-associated fibroblasts (Telocytes and *Pdgfra^{low}* cells). Furthermore, we unveil separate spatial subsets of fibroblasts acting as exclusive responders to TNF, each orchestrating inflammation in different intestinal layers. Additionally, manipulating the *Tnfrsf1a* gene exclusively in fibroblast subsets suggests that inflammation is initiated by telocytes and *Pdgfra^{low}* cells, while trophocytes drive its progression. This introduces novel evidence of spatial regulation of inflammation by fibroblast subsets, inciting and advancing disease in different layers of the gut. These findings underscore the pivotal role of fibroblasts in the inception and advancement of ileitis, proposing that targeting different fibroblast populations could impede the disease development and chronicity of inflammation.

miR-221/222 drive synovial fibroblast expansion and pathogenesis of TNF-mediated arthritis

Fani Roumelioti^{1,2}, Christos Tzaferis^{1,3}, Dimitris Konstantopoulos¹, Dimitra Papadopoulou^{1,3}, Alejandro Prados^{1,*}, Maria Sakkou^{1,4}, Anastasios Liakos⁵, Panagiotis Chouvardas^{1,#}, Theodore Meletakos⁵, Yiannis Pandis^{1,**}, Niki Karagianni⁶, Maria Denis⁶, Maria Fousteri⁵, Marietta Armaka⁵, and George Kollias^{1,3,4}

¹ Institute for Bioinnovation, Biomedical Sciences Research Centre (B.S.R.C.) "Alexander Fleming", Vari, 16672, Greece;

² Department of Pathophysiology, Medical School, National and Kapodistrian University of Athens, Athens, 11527, Greece

³ Department of Physiology, Medical School, National and Kapodistrian University of Athens, Athens, 11527, Greece

⁴ Center of New Biotechnologies & Precision Medicine, National and Kapodistrian University of Athens Medical School, Athens, Greece

⁵ Institute for Fundamental Biomedical Research, Biomedical Sciences Research Center "Alexander Fleming", Vari, Greece.

⁶ Biomedcode Hellas SA, Vari, 16672, Greece

Correspondence to:

George Kollias, Biomedical Sciences Research Center (BSRC), 'Alexander Fleming',
34 Alexander Fleming Street, Vari, 16672 Greece
(contact details: kollias@fleming.gr, tel. +302109656507).

Abstract

MicroRNAs (miRNAs) constitute fine tuners of gene expression and are implicated in a variety of diseases spanning from inflammation to cancer. miRNA expression is deregulated in rheumatoid arthritis (RA), however, their specific role in key arthritogenic cells such as the synovial fibroblast (SF) remains elusive. We have shown in the past that the expression of the miR-221/222 cluster is upregulated in RA SFs. Here, we demonstrate that miR-221/222 activation is downstream of major inflammatory cytokines, such as TNF and IL-1 β , which promote miR-221/222 expression independently. miR-221/222 expression in SFs from the *huTNFtg* mouse model of arthritis correlates with disease progression. Targeted transgenic overexpression of miR-221/222 in SFs of the *huTNFtg* mouse model led to further expansion of synovial fibroblasts and disease exacerbation. miR-221/222 overexpression altered the transcriptional profile of SFs igniting pathways involved in cell cycle progression and ECM regulation. Validated targets of miR-221/222 included p27 and p57 cell cycle inhibitors, as well as Smarca1 (a chromatin remodeling component). In contrast, complete genetic ablation of miR-221/222 in arthritic mice led to decreased proliferation of fibroblasts, reduced synovial expansion and attenuated disease. scATAC-seq data analysis revealed increased miR-221/222 gene activity in the pathogenic and activated clusters of the intermediate and lining compartment. Taken together, our results establish an SF-specific pathogenic role of the miR-221/222 cluster in arthritis and suggest that its therapeutic targeting in specific subpopulations should inform the design of novel fibroblast-targeted therapies for human disease.

10 References

- Aibar, Sara, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, et al. 2017. "SCENIC: Single-Cell Regulatory Network Inference and Clustering." *Nature Methods* 14 (11): 1083–86. <https://doi.org/10.1038/NMETH.4463>.
- Aidinis, Vassilis, David Plows, Sylva Haralambous, Maria Armaka, Petros Papadopoulos, Maria Zambia Kanaki, Dirk Koczan, Hans Juergen Thiesen, and George Kollias. 2003. "Functional Analysis of an Arthritogenic Synovial Fibroblast." *Arthritis Research & Therapy* 5 (3): R140. <https://doi.org/10.1186/AR749>.
- Aran, Dvir, Agnieszka P Looney, Leqian Liu, Esther Wu, Valerie Fong, Austin Hsu, Suzanna Chak, et al. 2019. "HHS Public Access" 20 (2): 163–72. <https://doi.org/10.1038/s41590-018-0276-y>. Reference-based.
- Armaka, Maria, Maria Apostolaki, Peggy Jacques, Dimitris L. Kontoyiannis, Dirk Elewaut, and George Kollias. 2008. "Mesenchymal Cell Targeting by TNF as a Common Pathogenic Principle in Chronic Inflammatory Joint and Intestinal Diseases." *The Journal of Experimental Medicine* 205 (2): 331–37. <https://doi.org/10.1084/JEM.20070906>.
- Armaka, Marietta, Dimitris Konstantopoulos, Christos Tzaferis, Matthieu D. Lavigne, Maria Sakkou, Anastasios Liakos, Petros P. Sfikakis, Meletios A. Dimopoulos, Maria Fousteri, and George Kollias. 2022. "Single-Cell Multimodal Analysis Identifies Common Regulatory Programs in Synovial Fibroblasts of Rheumatoid Arthritis Patients and Modeled TNF-Driven Arthritis." *Genome Medicine* 14 (1): 1–25. <https://doi.org/10.1186/S13073-022-01081-3>/FIGURES/8.
- Armaka, Marietta, Caroline Ospelt, Manolis Pasparakis, and George Kollias. 2018. "The P55TNFR-IKK2-Ripk3 Axis Orchestrates Arthritis by Regulating Death and Inflammatory Pathways in Synovial Fibroblasts." *Nature Communications* 9 (1). <https://doi.org/10.1038/S41467-018-02935-4>.
- Atrekhany, Kamar Sulu N., Violetta S. Gogoleva, Marina S. Drutskaya, and Sergei A. Nedospasov. 2020. "Distinct Modes of TNF Signaling through Its Two Receptors in Health and Disease." *Journal of Leukocyte Biology* 107 (6): 893–905. <https://doi.org/10.1002/JLB.2MR0120-510R>.
- Badia-I-Mompel, Pau, Jesús Vélez Santiago, Jana Braunger, Celina Geiss, Daniel Dimitrov, Sophia Müller-Dott, Petr Taus, et al. 2022. "DecoupleR: Ensemble of Computational Methods to Infer Biological Activities from Omics Data." *Bioinformatics Advances* 2 (1). <https://doi.org/10.1093/BIOADV/VBAC016>.

- Baysoy, Alev, Zhiliang Bai, Rahul Satija, and Rong Fan. 2023. "The Technological Landscape and Applications of Single-Cell Multi-Omics." *Nature Reviews Molecular Cell Biology* 2023 24:10 24 (10): 695–713. <https://doi.org/10.1038/s41580-023-00615-w>.
- Becht, Etienne, Leland McInnes, John Healy, Charles Antoine Dutertre, Immanuel W.H. Kwok, Lai Guan Ng, Florent Gehroux, and Evan W. Newell. 2018. "Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP." *Nature Biotechnology* 37 (1): 38–47. <https://doi.org/10.1038/NBT.4314>.
- Bergen, Volker, Marius Lange, Stefan Peidli, F. Alexander Wolf, and Fabian J. Theis. 2020. "Generalizing RNA Velocity to Transient Cell States through Dynamical Modeling." *Nature Biotechnology* 38 (12): 1408–14. <https://doi.org/10.1038/S41587-020-0591-3>.
- Blanco-Carmona, Enrique. 2022. "Generating Publication Ready Visualizations for Single Cell Transcriptomics Using SCpubr." *BioRxiv*, March, 2022.02.28.482303. <https://doi.org/10.1101/2022.02.28.482303>.
- Blondel, Vincent D., Jean Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. "Fast Unfolding of Communities in Large Networks." *Journal of Statistical Mechanics: Theory and Experiment* 2008 (10): P10008. <https://doi.org/10.1088/1742-5468/2008/10/P10008>.
- Bravo González-Blas, Carmen, Liesbeth Minnoye, Dafni Papisokrati, Sara Aibar, Gert Hulselmans, Valerie Christiaens, Kristofer Davie, Jasper Wouters, and Stein Aerts. 2019. "CisTopic: Cis-Regulatory Topic Modeling on Single-Cell ATAC-Seq Data." *Nature Methods* 16 (5): 397–400. <https://doi.org/10.1038/S41592-019-0367-1>.
- Browaeys, Robin, Wouter Saelens, and Yvan Saeys. 2020. "NicheNet: Modeling Intercellular Communication by Linking Ligands to Target Genes." *Nature Methods* 17 (2): 159–62. <https://doi.org/10.1038/s41592-019-0667-5>.
- Buechler, Matthew B., Rachana N. Pradhan, Akshay T. Krishnamurthy, Christian Cox, Aslihan Karabacak Calviello, Amber W. Wang, Yeqing Angela Yang, et al. 2021. "Cross-Tissue Organization of the Fibroblast Lineage." *Nature* 593 (7860): 575–79. <https://doi.org/10.1038/S41586-021-03549-5>.
- Bumgarner, Roger. 2013. "Overview of DNA Microarrays: Types, Applications, and Their Future." *Current Protocols in Molecular Biology* Chapter 22 (SUPPL.101). <https://doi.org/10.1002/0471142727.MB2201S101>.

- Bunis, Daniel G., Jared Andrews, Gabriela K. Fragiadakis, Trevor D. Burt, and Marina Sirota. 2021. "DittoSeq: Universal User-Friendly Single-Cell and Bulk RNA Sequencing Visualization Toolkit." *Bioinformatics (Oxford, England)* 36 (22–23): 5535–36. <https://doi.org/10.1093/BIOINFORMATICS/BTAA1011>.
- Chan Zuckerberg Initiative. n.d. "CZ CELLxGENE Discover." <https://cellxgene.cziscience.com/>.
- Chen, Geng, Baitang Ning, and Tieliu Shi. 2019. "Single-Cell RNA-Seq Technologies and Related Computational Data Analysis." *Frontiers in Genetics* 10 (APR): 441123. <https://doi.org/10.3389/FGENE.2019.00317/BIBTEX>.
- Chen, Xin, Monika Bäuml, Daniela N. Männel, O. M. Zack Howard, and Joost J. Oppenheim. 2007. "Interaction of TNF with TNF Receptor Type 2 Promotes Expansion and Function of Mouse CD4+CD25+ T Regulatory Cells." *Journal of Immunology (Baltimore, Md. : 1950)* 179 (1): 154–61. <https://doi.org/10.4049/JIMMUNOL.179.1.154>.
- Conesa, Ana, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michal Wojciech Szczesniak, et al. 2016. "A Survey of Best Practices for RNA-Seq Data Analysis." *Genome Biology* 17 (1). <https://doi.org/10.1186/S13059-016-0881-8>.
- Conigliaro, Paola, Carlo Perricone, Robert A. Benson, Paul Garside, James M. Brewer, Roberto Perricone, and Guido Valesini. 2010. "The Type I IFN System in Rheumatoid Arthritis." *Autoimmunity* 43 (3): 220–25. <https://doi.org/10.3109/08916930903510914>.
- Corces, M. Ryan, Jeffrey M. Granja, Shadi Shams, Bryan H. Louie, Jose A. Seoane, Wanding Zhou, Tiago C. Silva, et al. 2018. "The Chromatin Accessibility Landscape of Primary Human Cancers." *Science (New York, N.Y.)* 362 (6413). <https://doi.org/10.1126/SCIENCE.AAV1898>.
- Croft, Adam P., Joana Campos, Kathrin Jansen, Jason D. Turner, Jennifer Marshall, Moustafa Attar, Loriane Savary, et al. 2019. "Distinct Fibroblast Subsets Drive Inflammation and Damage in Arthritis." *Nature* 570 (7760): 246–51. <https://doi.org/10.1038/S41586-019-1263-7>.
- Crowson, Cynthia S., Katherine P. Liao, John M. Davis, Daniel H. Solomon, Eric L. Matteson, Keith L. Knutson, Mark A. Hlatky, and Sherine E. Gabriel. 2013. "Rheumatoid Arthritis and Cardiovascular Disease." *American Heart Journal* 166 (4): 622. <https://doi.org/10.1016/J.AHJ.2013.07.010>.
- Cui, Haotian, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. 2024.

“ScGPT: Toward Building a Foundation Model for Single-Cell Multi-Omics Using Generative AI.” *Nature Methods*. <https://doi.org/10.1038/S41592-024-02201-0>.

Danese, Anna, Maria L. Richter, Kridsakorn Chaichoompu, David S. Fischer, Fabian J. Theis, and Maria Colomé-Tatché. 2021. “EpiScanpy: Integrated Single-Cell Epigenomic Analysis.” *Nature Communications* 12 (1). <https://doi.org/10.1038/S41467-021-25131-3>.

Davidson, Sarah, Mark Coles, Tom Thomas, George Kollias, Burkhard Ludewig, Shannon Turley, Michael Brenner, and Christopher D. Buckley. 2021. “Fibroblasts as Immune Regulators in Infection, Inflammation and Cancer.” *Nature Reviews. Immunology* 21 (11): 704–17. <https://doi.org/10.1038/S41577-021-00540-Z>.

Dijk, David van, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J. Carr, Cassandra Burdziak, et al. 2018. “Recovering Gene Interactions from Single-Cell Data Using Data Diffusion.” *Cell* 174 (3): 716-729.e27. <https://doi.org/10.1016/J.CELL.2018.05.061>.

Dimitrov, Daniel, and Quan Gu. 2020. “BingleSeq: A User-Friendly R Package for Bulk and Single-Cell RNA-Seq Data Analysis.” *PeerJ* 8 (December). <https://doi.org/10.7717/PEERJ.10469>.

Dinarello, Charles Anthony. 2019. “The IL-1 Family of Cytokines and Receptors in Rheumatic Diseases.” *Nature Reviews. Rheumatology* 15 (10): 612–32. <https://doi.org/10.1038/S41584-019-0277-8>.

Douni, Eleni, Katerina Akassoglou, Lena Alexopoulou, Spiros Georgopoulos, Sylva Haralambous, Sally Hill, George Kassiotis, et al. 1995. “Transgenic and Knockout Analyses of the Role of TNF in Immune Regulation and Disease Pathogenesis.” *Journal of Inflammation* 47 (1–2): 27–38. <https://europepmc.org/article/med/8913927>.

Efremova, Mirjana, Miquel Vento-Tormo, Sarah A Teichmann, and Roser Vento-Tormo. 2020. “CellPhoneDB: Inferring Cell-Cell Communication from Combined Expression of Multi-Subunit Ligand-Receptor Complexes.” *Nature Protocols* 15 (4): 1484–1506. <https://doi.org/10.1038/s41596-020-0292-x>.

Ekiz, H. Atakan, Christopher J. Conley, W. Zac Stephens, and Ryan M. O’Connell. 2020a. “CIPR: A Web-Based R/Shiny App and R Package to Annotate Cell Clusters in Single Cell RNA Sequencing Experiments.” *BMC Bioinformatics* 21 (1): 191. <https://doi.org/10.1186/S12859-020-3538-2>.

Ekiz, H Atakan, Christopher J Conley, W Zac Stephens, and Ryan M O’Connell. 2020b. “CIPR: A Web-

Based R/Shiny App and R Package to Annotate Cell Clusters in Single Cell RNA Sequencing Experiments." *BMC Bioinformatics* 21 (1): 191. <https://doi.org/10.1186/s12859-020-3538-2>.

Finak, Greg, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K. Shalek, Chloe K. Slichter, et al. 2015. "MAST: A Flexible Statistical Framework for Assessing Transcriptional Changes and Characterizing Heterogeneity in Single-Cell RNA Sequencing Data." *Genome Biology* 16 (1). <https://doi.org/10.1186/S13059-015-0844-5>.

Franzén, Oscar, and Johan L.M. Björkegren. 2020. "Alona: A Web Server for Single-Cell RNA-Seq Analysis." *Bioinformatics (Oxford, England)* 36 (12): 3910–12. <https://doi.org/10.1093/BIOINFORMATICS/BTAA269>.

Frasnelli, Matthias E, David Tarussio, Veronique Chobaz-Péclat, Nathalie Busso, and Alexander So. 2005. "TLR2 Modulates Inflammation in Zymosan-Induced Arthritis in Mice." *Arthritis Research & Therapy* 7 (2): R370. <https://doi.org/10.1186/AR1494>.

Furman, David, Judith Campisi, Eric Verdin, Pedro Carrera-Bastos, Sasha Targ, Claudio Franceschi, Luigi Ferrucci, et al. 2019. "Chronic Inflammation in the Etiology of Disease across the Life Span." *Nature Medicine* 25 (12): 1822–32. <https://doi.org/10.1038/S41591-019-0675-0>.

García-Alonso, Carlos R., Leonor M. Pérez-Naranjo, and Juan C. Fernández-Caballero. 2014. "Multiobjective Evolutionary Algorithms to Identify Highly Autocorrelated Areas: The Case of Spatial Distribution in Financially Compromised Farms." *Annals of Operations Research* 219 (1): 187–202. <https://doi.org/10.1007/s10479-011-0841-3>.

Geiler, Thomas, JÖUrg Kriegsmann, Gernot M. Keyszer, Renate E. Gay, and Steffen Gay. 1994. "A New Model for Rheumatoid Arthritis Generated by Engraftment of Rheumatoid Synovial Tissue and Normal Human Cartilage into SCID Mice." *Arthritis and Rheumatism* 37 (11): 1664–71. <https://doi.org/10.1002/ART.1780371116>.

Gelder, Russell N. Van, Mark E. Von Zastrow, Andrea Yool, William C. Dement, Jack D. Barchas, and James H. Eberwine. 1990. "Amplified RNA Synthesized from Limited Quantities of Heterogeneous CDNA." *Proceedings of the National Academy of Sciences of the United States of America* 87 (5): 1663–67. <https://doi.org/10.1073/PNAS.87.5.1663>.

"GitHub - Kharchenkolab/Pagoda2: R Package for Analyzing and Interactively Exploring Large-Scale Single-Cell RNA-Seq Datasets." n.d. Accessed June 6, 2024. <https://github.com/kharchenkolab/pagoda2>.

"GitHub - Xmc811/Scillus: R Package for Single-Cell Dataset Processing and Visualization." n.d. Accessed June 6, 2024. <https://github.com/xmc811/Scillus>.

Goel, Niti, and Sue Stephens. 2010. "Certolizumab Pegol." *MABs* 2 (2): 137–47. <https://doi.org/10.4161/MABS.2.2.11271>.

Granja, Jeffrey M., M. Ryan Corces, Sarah E. Pierce, S. Tansu Bagdatli, Hani Choudhry, Howard Y. Chang, and William J. Greenleaf. 2021a. "ArchR Is a Scalable Software Package for Integrative Single-Cell Chromatin Accessibility Analysis." *Nature Genetics* 53 (3): 403–11. <https://doi.org/10.1038/S41588-021-00790-6>.

Granja, Jeffrey M., M. Ryan Corces, Sarah E. Pierce, S. Tansu Bagdatli, Hani Choudhry, Howard Y. Chang, and William J. Greenleaf. 2021b. "Author Correction: ArchR Is a Scalable Software Package for Integrative Single-Cell Chromatin Accessibility Analysis." *Nature Genetics* 53 (6): 935. <https://doi.org/10.1038/s41588-021-00850-x>.

Haghverdi, Laleh, Florian Buettner, and Fabian J. Theis. 2015. "Diffusion Maps for High-Dimensional Single-Cell Analysis of Differentiation Data." *Bioinformatics (Oxford, England)* 31 (18): 2989–98. <https://doi.org/10.1093/BIOINFORMATICS/BTV325>.

Hao, Yuhan, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, et al. 2021. "Integrated Analysis of Multimodal Single-Cell Data." *Cell* 184 (13): 3573–3587.e29. <https://doi.org/10.1016/J.CELL.2021.04.048>.

Hayden, Matthew S., and Sankar Ghosh. 2014. "Regulation of NF- κ B by TNF Family Cytokines." *Seminars in Immunology* 26 (3): 253–66. <https://doi.org/10.1016/J.SMIM.2014.05.004>.

He, Di, Di Wang, Ping Lu, Nan Yang, Zhigang Xue, Xianmin Zhu, Peng Zhang, and Guoping Fan. 2021. "Single-Cell RNA Sequencing Reveals Heterogeneous Tumor and Immune Cell Populations in Early-Stage Lung Adenocarcinomas Harboring EGFR Mutations." *Oncogene* 40 (2): 355–68. <https://doi.org/10.1038/S41388-020-01528-0>.

He, Xinwei, Kun Qian, Ziqian Wang, Shirou Zeng, Hongwei Li, and Wei Vivian Li. 2023. "ScAce: An Adaptive Embedding and Clustering Method for Single-Cell Gene Expression Data." *Bioinformatics (Oxford, England)* 39 (9). <https://doi.org/10.1093/BIOINFORMATICS/BTAD546>.

- He, Zhen, Shuofeng Hu, Yaowen Chen, Sijing An, Jiahao Zhou, Runyan Liu, Junfeng Shi, et al. 2024. "Mosaic Integration and Knowledge Transfer of Single-Cell Multimodal Data with MIDAS." *Nature Biotechnology*. <https://doi.org/10.1038/S41587-023-02040-Y>.
- Heumos, Lukas, Anna C. Schaar, Christopher Lance, Anastasia Litinetskaya, Felix Drost, Luke Zappia, Malte D. Lücken, et al. 2023. "Best Practices for Single-Cell Analysis across Modalities." *Nature Reviews Genetics* 2023 24:8 24 (8): 550–72. <https://doi.org/10.1038/s41576-023-00586-w>.
- Hillje, Roman, Pier Giuseppe Pelicci, and Lucilla Luzi. 2020. "Cerebro: Interactive Visualization of ScRNA-Seq Data." *Bioinformatics (Oxford, England)* 36 (7): 2311–13. <https://doi.org/10.1093/BIOINFORMATICS/BTZ877>.
- Hoek, Andreas, Katharina Maibach, Ebru Özmen, Ana Ivonne Vazquez-Armendariz, Jan Philipp Mengel, Torsten Hain, Susanne Herold, and Alexander Goesmann. 2021. "WASP: A Versatile, Web-Accessible Single Cell RNA-Seq Processing Platform." *BMC Genomics* 22 (1). <https://doi.org/10.1186/S12864-021-07469-6>.
- Holmdahl, R., L. Jansson, M. Andersson, and R. Jonsson. 1992. "Genetic, Hormonal and Behavioural Influence on Spontaneously Developing Arthritis in Normal Mice." *Clinical and Experimental Immunology* 88 (3): 467. <https://doi.org/10.1111/J.1365-2249.1992.TB06473.X>.
- Holmdahl, Rikard, Kristofer Rubin, Lars Klareskog, Erik Larsson, and Hans Wigzell. 1986. "Characterization of the Antibody Response in Mice with Type II Collagen-Induced Arthritis, Using Monoclonal Anti-Type II Collagen Antibodies." *Arthritis and Rheumatism* 29 (3): 400–410. <https://doi.org/10.1002/ART.1780290314>.
- Horai, Reiko, Shinobu Saijo, Hidetoshi Tanioka, Susumu Nakae, Katsuko Sudo, Akihiko Okahara, Toshimi Ikuse, Masahide Asano, and Yoichiro Iwakura. 2000. "Development of Chronic Inflammatory Arthropathy Resembling Rheumatoid Arthritis in Interleukin 1 Receptor Antagonist-Deficient Mice." *The Journal of Experimental Medicine* 191 (2): 313–20. <https://doi.org/10.1084/JEM.191.2.313>.
- Horak, Christine E., and Michael Snyder. 2002. "ChIP-Chip: A Genomic Approach for Identifying Transcription Factor Binding Sites." *Methods in Enzymology* 350: 469–83. [https://doi.org/10.1016/S0076-6879\(02\)50979-4](https://doi.org/10.1016/S0076-6879(02)50979-4).
- Hwang, Byungjin, Ji Hyun Lee, and Duhee Bang. 2018. "Single-Cell RNA Sequencing Technologies and Bioinformatics Pipelines." *Experimental & Molecular Medicine* 2018 50:8 50 (8): 1–14. <https://doi.org/10.1038/s12276-018-0071-8>.

- Jiang, Andrew, Klaus Lehnert, Linya You, and Russell G. Snell. 2022. "ICARUS, an Interactive Web Server for Single Cell RNA-Seq Analysis." *Nucleic Acids Research* 50 (W1): W427–33. <https://doi.org/10.1093/NAR/GKAC322>.
- Jin, Suoqin, Christian F Guerrero-Juarez, Lihua Zhang, Ivan Chang, Raul Ramos, Chen-Hsiang Kuan, Peggy Myung, Maksim V Plikus, and Qing Nie. 2021. "Inference and Analysis of Cell-Cell Communication Using CellChat." *Nature Communications* 12 (1): 1088. <https://doi.org/10.1038/s41467-021-21246-9>.
- Kallio, M. Aleks, Jarno T. Tuimala, Taavi Hupponen, Petri Klemelä, Massimiliano Gentile, Ilari Scheinin, Mikko Koski, Janne Käki, and Eija I. Korpelainen. 2011. "Chipster: User-Friendly Analysis Software for Microarray and Other High-Throughput Data." *BMC Genomics* 12 (October). <https://doi.org/10.1186/1471-2164-12-507>.
- Keffer, J., L. Probert, H. Cazlaris, S. Georgopoulos, E. Kaslaris, D. Kioussis, and G. Kollias. 1991. "Transgenic Mice Expressing Human Tumour Necrosis Factor: A Predictive Genetic Model of Arthritis." *The EMBO Journal* 10 (13): 4025–31. <https://doi.org/10.1002/J.1460-2075.1991.TB04978.X>.
- Kesimoglu, Ziyne Nesibe, and Serdar Bozdog. 2023. "SUPREME: Multiomics Data Integration Using Graph Convolutional Networks." *NAR Genomics and Bioinformatics* 5 (2). <https://doi.org/10.1093/NARGAB/LQAD063>.
- Koliaraki, Vasiliki, Alejandro Prados, Marietta Armaka, and George Kollias. 2020. "The Mesenchymal Context in Inflammation, Immunity and Cancer." *Nature Immunology* 21 (9): 974–82. <https://doi.org/10.1038/S41590-020-0741-2>.
- Kollias, George, Piyi Papadaki, Florence Apparailly, Margriet J. Vervoordeldonk, Rikard Holmdahl, Vera Baumans, Christian Desaintes, et al. 2011. "Animal Models for Arthritis: Innovative Tools for Prevention and Treatment." *Annals of the Rheumatic Diseases* 70 (8): 1357–62. <https://doi.org/10.1136/ARD.2010.148551>.
- Kontoyiannis, Dimitris, Manolis Pasparakis, Theresa T. Pizarro, Fabio Cominelli, and George Kollias. 1999. "Impaired on/off Regulation of TNF Biosynthesis in Mice Lacking TNF AU-Rich Elements: Implications for Joint and Gut-Associated Immunopathologies." *Immunity* 10 (3): 387–98. [https://doi.org/10.1016/S1074-7613\(00\)80038-2](https://doi.org/10.1016/S1074-7613(00)80038-2).
- Korsunsky, Ilya, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko,

- Michael Brenner, Po ru Loh, and Soumya Raychaudhuri. 2019. "Fast, Sensitive and Accurate Integration of Single-Cell Data with Harmony." *Nature Methods* 16 (12): 1289–96. <https://doi.org/10.1038/S41592-019-0619-0>.
- Kukurba, Kimberly R., and Stephen B. Montgomery. 2015. "RNA Sequencing and Analysis." *Cold Spring Harbor Protocols* 2015 (11): 951. <https://doi.org/10.1101/PDB.TOP084970>.
- Lange, Marius, Volker Bergen, Michal Klein, Manu Setty, Bernhard Reuter, Mostafa Bakhti, Heiko Lickert, et al. 2022. "CellRank for Directed Single-Cell Fate Mapping." *Nature Methods* 19 (2): 159–70. <https://doi.org/10.1038/S41592-021-01346-6>.
- Lefèvre, Stephanie, Anette Knedla, Christoph Tennie, Andreas Kampmann, Christina Wunrau, Robert Dinsler, Adelheid Korb, et al. 2009. "Synovial Fibroblasts Spread Rheumatoid Arthritis to Unaffected Joints." *Nature Medicine* 15 (12): 1414–20. <https://doi.org/10.1038/NM.2050>.
- Lewis, Myles J., Michael R. Barnes, K. Blighe, Katriona Goldmann, Sharmila Rana, Jason A. Hackney, Nandhini Ramamoorthi, et al. 2019. "Molecular Portraits of Early Rheumatoid Arthritis Identify Clinical and Treatment Response Phenotypes." *Cell Reports* 28 (9): 2455-2470.e5. <https://doi.org/10.1016/J.CELREP.2019.07.091>.
- Li, Xinmin, and Cun Yu Wang. 2021. "From Bulk, Single-Cell to Spatial RNA Sequencing." *International Journal of Oral Science* 2021 13:1 13 (1): 1–6. <https://doi.org/10.1038/s41368-021-00146-0>.
- Luecken, Malte D, and Fabian J Theis. 2019. "Current Best Practices in Single-cell RNA-seq Analysis: A Tutorial." *Molecular Systems Biology* 15 (6). <https://doi.org/10.15252/MSB.20188746>.
- Maaten, L.J.P. van der, and G.E. Hinton. 2008. "Visualizing High-Dimensional Data Using t-SNE." *Journal of Machine Learning Research* 9 (nov): 2579–2605. <https://research.tilburguniversity.edu/en/publications/visualizing-high-dimensional-data-using-t-sne>.
- Madsen, Pernille M., Haritha L. Desu, Juan Pablo de Rivero Vaccari, Yoleinny Florimon, Ditte G. Ellman, Robert W. Keane, Bettina H. Clausen, Kate L. Lambertsen, and Roberta Brambilla. 2020. "Oligodendrocytes Modulate the Immune-Inflammatory Response in EAE via TNFR2 Signaling." *Brain, Behavior, and Immunity* 84 (February): 132–46. <https://doi.org/10.1016/J.BBI.2019.11.017>.
- Madsen, Pernille M., Dario Motti, Shaffiat Karmally, David E. Szymkowski, Kate Lykke Lambertsen, John

- R. Bethea, and Roberta Brambilla. 2016. "Oligodendroglial TNFR2 Mediates Membrane TNF-Dependent Repair in Experimental Autoimmune Encephalomyelitis by Promoting Oligodendrocyte Differentiation and Remyelination." *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience* 36 (18): 5128–43. <https://doi.org/10.1523/JNEUROSCI.0211-16.2016>.
- Mah, Clarence K., Alexander T. Wenzel, Edwin F. Juarez, Thorin Tabor, Michael M. Reich, and Jill P. Mesirov. 2019. "An Accessible, Interactive Genepattern Notebook for Analysis and Exploration of Single-Cell Transcriptomic Data." *F1000Research* 7. <https://doi.org/10.12688/F1000RESEARCH.15830.2/DOI>.
- Martin, Jerome C., Christie Chang, Gilles Boschetti, Ryan Ungaro, Mamta Giri, John A. Grout, Kyle Gettler, et al. 2019. "Single-Cell Analysis of Crohn's Disease Lesions Identifies a Pathogenic Cellular Module Associated with Resistance to Anti-TNF Therapy." *Cell* 178 (6): 1493-1508.e20. <https://doi.org/10.1016/J.CELL.2019.08.008>.
- Mazumdar, Sohini, and David Greenwald. 2009. "Golimumab." *MAbs* 1 (5): 422–31. <https://doi.org/10.4161/MABS.1.5.9286>.
- McCarthy, Davis J, Kieran R Campbell, Aaron T L Lun, and Quin F Wills. 2017. "Scater: Pre-Processing, Quality Control, Normalization and Visualization of Single-Cell RNA-Seq Data in R." *Bioinformatics* 33 (8): 1179. <https://doi.org/10.1093/BIOINFORMATICS/BTW777>.
- McDavid, Andrew, Greg Finak, Pratip K. Chattopadhyay, Maria Dominguez, Laurie Lamoreaux, Steven S. Ma, Mario Roederer, and Raphael Gottardo. 2013. "Data Exploration, Quality Control and Testing in Single-Cell QPCR-Based Gene Expression Experiments." *Bioinformatics (Oxford, England)* 29 (4): 461–67. <https://doi.org/10.1093/BIOINFORMATICS/BTS714>.
- McGinnis, Christopher S, Lyndsay M Murrow, and Zev J Gartner. 2019. "DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors." *Cell Systems* 8 (4): 329-337.e4. <https://doi.org/10.1016/j.cels.2019.03.003>.
- McInnes, Iain B., Christopher D. Buckley, and John D. Isaacs. 2016. "Cytokines in Rheumatoid Arthritis - Shaping the Immunological Landscape." *Nature Reviews. Rheumatology* 12 (1): 63–68. <https://doi.org/10.1038/NRRHEUM.2015.171>.
- "Method of the Year 2013." 2014. *Nature Methods* 11 (1): 1. <https://doi.org/10.1038/NMETH.2801>.

- Meyer, Jesse G., Ryan J. Urbanowicz, Patrick C.N. Martin, Karen O'Connor, Ruowang Li, Pei Chen Peng, Tiffani J. Bright, et al. 2023. "ChatGPT and Large Language Models in Academia: Opportunities and Challenges." *BioData Mining* 16 (1). <https://doi.org/10.1186/S13040-023-00339-9>.
- Monach, Paul A., Diane Mathis, and Christophe Benoist. 2008. "The K/BxN Arthritis Model." *Current Protocols in Immunology* Chapter 15 (SUPPL. 81). <https://doi.org/10.1002/0471142735.IM1522S81>.
- Moussa, Marmar, and Ion I. Mandoiu. 2021. "SC1: A Tool for Interactive Web-Based Single-Cell RNA-Seq Data Analysis." *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology* 28 (8): 820–41. <https://doi.org/10.1089/CMB.2021.0051>.
- Müller-Dott, Sophia, Eirini Tsirvouli, Miguel Vázquez, Ricardo O Ramirez Flores, Pau Badia-i-Mompel, Robin Fallegger, Astrid Læg Reid, and Julio Saez-Rodriguez. 2023. "Expanding the Coverage of Regulons from High-Confidence Prior Knowledge for Accurate Estimation of Transcription Factor Activities." *BioRxiv*, January, 2023.03.30.534849. <https://doi.org/10.1101/2023.03.30.534849>.
- Müller-Ladner, Ulf, Jörg Kriegsmann, Barry N. Franklin, Shigeru Matsumoto, Thomas Geiler, Renate E. Gay, and Steffen Gay. 1996. "Synovial Fibroblasts of Patients with Rheumatoid Arthritis Attach to and Invade Normal Human Cartilage When Engrafted into SCID Mice." *The American Journal of Pathology* 149 (5): 1607. [/pmc/articles/PMC1865262/?report=abstract](https://pubmed.ncbi.nlm.nih.gov/149/5/1607/).
- Nandakumar, Kutty S.elva, and Rikard Holmdahl. 2006. "Antibody-Induced Arthritis: Disease Mechanisms and Genes Involved at the Effector Phase of Arthritis." *Arthritis Research & Therapy* 8 (6): 223. <https://doi.org/10.1186/AR2089>.
- Nandakumar, Kutty Selva, Lars Svensson, and Rikard Holmdahl. 2003. "Collagen Type II-Specific Monoclonal Antibody-Induced Arthritis in Mice: Description of the Disease and the Influence of Age, Sex, and Genes." *The American Journal of Pathology* 163 (5): 1827–37. [https://doi.org/10.1016/S0002-9440\(10\)63542-0](https://doi.org/10.1016/S0002-9440(10)63542-0).
- Ngo, S. T., F. J. Steyn, and P. A. McCombe. 2014. "Gender Differences in Autoimmune Disease." *Frontiers in Neuroendocrinology* 35 (3): 347–69. <https://doi.org/10.1016/J.YFRNE.2014.04.004>.
- Ntougkos, Evangelos, Panagiotis Chouvardas, Fani Roumelioti, Caroline Ospelt, Mojca Frank-Bertoncelj, Andrew Filer, Christopher D. Buckley, Steffen Gay, Christoforos Nikolaou, and George Kollias. 2017. "Genomic Responses of Mouse Synovial Fibroblasts During Tumor Necrosis Factor-Driven Arthritogenesis Greatly Mimic Those in Human Rheumatoid Arthritis." *Arthritis & Rheumatology (Hoboken, N.J.)* 69 (8): 1588–1600. <https://doi.org/10.1002/ART.40128>.

- Packer, Jonathan S., Qin Zhu, Chau Huynh, Priya Sivaramakrishnan, Elicia Preston, Hannah Dueck, Derek Stefanik, et al. 2019. "A Lineage-Resolved Molecular Atlas of *C. Elegans* Embryogenesis at Single-Cell Resolution." *Science (New York, N.Y.)* 365 (6459). <https://doi.org/10.1126/SCIENCE.AAX1971>.
- Papalexii, Efthymia, and Rahul Satija. 2018. "Single-Cell RNA Sequencing to Explore Immune Cell Heterogeneity." *Nature Reviews. Immunology* 18 (1): 35–45. <https://doi.org/10.1038/NRI.2017.76>.
- Patel, Mitulkumar V. 2018. "IS-CellR: A User-Friendly Tool for Analyzing and Visualizing Single-Cell RNA Sequencing Data." *Bioinformatics (Oxford, England)* 34 (24): 4305–6. <https://doi.org/10.1093/BIOINFORMATICS/BTY517>.
- Pereira, W. J., F. M. Almeida, D. Conde, K. M. Balmant, P. M. Triozzi, H. W. Schmidt, C. Dervinis, G. J. Pappas, and M. Kirst. 2021. "Asc-Seurat: Analytical Single-Cell Seurat-Based Web Application." *BMC Bioinformatics* 22 (1). <https://doi.org/10.1186/S12859-021-04472-2>.
- Pliner, Hannah A., Jonathan S. Packer, José L. McFaline-Figueroa, Darren A. Cusanovich, Riza M. Daza, Delasa Aghamirzaie, Sanjay Srivatsan, et al. 2018. "Cicero Predicts Cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data." *Molecular Cell* 71 (5): 858-871.e8. <https://doi.org/10.1016/J.MOLCEL.2018.06.044>.
- Polz, Johannes, Annika Remke, Sabine Weber, Dominic Schmidt, Dorothea Weber-Steffens, Anne Pietryga-Krieger, Nils Müller, Uwe Ritter, Sven Mostböck, and Daniela N. Männel. 2014. "Myeloid Suppressor Cells Require Membrane TNFR2 Expression for Suppressive Activity." *Immunity, Inflammation and Disease* 2 (2): 121–30. <https://doi.org/10.1002/IID3.19>.
- Prieto, Carlos, David Barrios, and Angela Villaverde. 2022. "SingleCAnalyzer: Interactive Analysis of Single Cell RNA-Seq Data on the Cloud." *Frontiers in Bioinformatics* 2 (May). <https://doi.org/10.3389/FBINF.2022.793309>.
- Qiu, Xiaojie, Qi Mao, Ying Tang, Li Wang, Raghav Chawla, Hannah A. Pliner, and Cole Trapnell. 2017. "Reversed Graph Embedding Resolves Complex Single-Cell Trajectories." *Nature Methods* 14 (10): 979–82. <https://doi.org/10.1038/NMETH.4402>.
- Raudvere, Uku, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. 2019. "G:Profiler: A Web Server for Functional Enrichment Analysis and Conversions of Gene Lists (2019 Update)." *Nucleic Acids Research* 47 (W1): W191–98. <https://doi.org/10.1093/NAR/GKZ369>.

Sakaguchi, Noriko, Takeshi Takahashi, Hiroshi Hata, Takashi Nomura, Tomoyuki Tagami, Sayuri Yamazaki, Toshiko Sakihama, et al. 2003. "Altered Thymic T-Cell Selection Due to a Mutation of the ZAP-70 Gene Causes Autoimmune Arthritis in Mice." *Nature* 426 (6965): 454–60. <https://doi.org/10.1038/NATURE02119>.

"Samuel-Marsh/ScCustomize: Version 2.1.2." n.d. Accessed June 6, 2024. <https://doi.org/10.5281/ZENODO.10724532>.

Sinha, Sarthak, Holly D. Sparks, Elodie Labit, Hayley N. Robbins, Kevin Gowing, Arzina Jaffer, Eren Kutluber, et al. 2022. "Fibroblast Inflammatory Priming Determines Regenerative versus Fibrotic Skin Repair in Reindeer." *Cell* 185 (25): 4717–4736.e25. <https://doi.org/10.1016/J.CELL.2022.11.004>.

Smillie, Christopher S., Moshe Biton, Jose Ordovas-Montanes, Keri M. Sullivan, Grace Burgin, Daniel B. Graham, Rebecca H. Herbst, et al. 2019. "Intra- and Inter-Cellular Rewiring of the Human Colon during Ulcerative Colitis." *Cell* 178 (3): 714–730.e22. <https://doi.org/10.1016/J.CELL.2019.06.029>.

Smolen, Josef S., Daniel Aletaha, Anne Barton, Gerd R. Burmester, Paul Emery, Gary S. Firestein, Arthur Kavanaugh, et al. 2018. "Rheumatoid Arthritis." *Nature Reviews. Disease Primers* 4 (February). <https://doi.org/10.1038/NRDP.2018.1>.

Sommarin, Mikael N.E., Rasmus Olofzon, Sara Palo, Parashar Dhapola, Shamit Soneji, Göran Karlsson, and Charlotta Böiers. 2023. "Single-Cell Multiomics of Human Fetal Hematopoiesis Define a Developmental-Specific Population and a Fetal Signature." *Blood Advances* 7 (18): 5325–40. <https://doi.org/10.1182/BLOODADVANCES.2023009808>.

Srirangan, Srinivasan, and Ernest H. Choy. 2010. "The Role of Interleukin 6 in the Pathophysiology of Rheumatoid Arthritis." *Therapeutic Advances in Musculoskeletal Disease* 2 (5): 247. <https://doi.org/10.1177/1759720X10378372>.

Stephenson, William, Laura T. Donlin, Andrew Butler, Cristina Rozo, Bernadette Bracken, Ali Rashidfarrokhi, Susan M. Goodman, et al. 2018. "Single-Cell RNA-Seq of Rheumatoid Arthritis Synovial Tissue Using Low-Cost Microfluidic Instrumentation." *Nature Communications* 9 (1). <https://doi.org/10.1038/S41467-017-02659-X>.

Street, Kelly, Davide Risso, Russell B. Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. 2018. "Slingshot: Cell Lineage and Pseudotime Inference for Single-Cell Transcriptomics." *BMC Genomics* 19 (1). <https://doi.org/10.1186/S12864-018-4772-0>.

- Stuart, Tim, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, Yuhao Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. 2019. "Comprehensive Integration of Single-Cell Data." *Cell* 177 (7): 1888-1902.e21. <https://doi.org/10.1016/J.CELL.2019.05.031>.
- Stuart, Tim, Avi Srivastava, Shaista Madad, Caleb A. Lareau, and Rahul Satija. 2021. "Single-Cell Chromatin State Analysis with Signac." *Nature Methods* 18 (11): 1333–41. <https://doi.org/10.1038/S41592-021-01282-5>.
- Sudoł-Szopińska, Iwona, Ewa Kontny, Włodzimierz Maśliński, Monika Prochorec-Sobieszek, Brygida Kwiatkowska, Katarzyna Zaniewicz-Kaniewska, and Agnieszka Warczyńska. 2012. "The Pathogenesis of Rheumatoid Arthritis in Radiological Studies. Part I: Formation of Inflammatory Infiltrates within the Synovial Membrane." *Journal of Ultrasonography* 12 (49): 202. <https://doi.org/10.15557/JOU.2012.0007>.
- Tang, Fuchou, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, et al. 2009. "MRNA-Seq Whole-Transcriptome Analysis of a Single Cell." *Nature Methods* 6 (5): 377–82. <https://doi.org/10.1038/NMETH.1315>.
- Tang, Kwan Ho, Shuai Li, Alireza Khodadadi-Jamayran, Jayu Jen, Han Han, Kayla Guidry, Ting Chen, et al. 2022. "Combined Inhibition of SHP2 and CXCR1/2 Promotes Antitumor T-Cell Response in NSCLC." *Cancer Discovery* 12 (1): 47–61. <https://doi.org/10.1158/2159-8290.CD-21-0369>.
- Tarhan, Leyla, Jon Bistline, Jean Chang, Bryan Galloway, Emily Hanna, and Eric Weitz. 2023. "Single Cell Portal: An Interactive Home for Single-Cell Genomics Data." *BioRxiv : The Preprint Server for Biology*. <https://doi.org/10.1101/2023.07.13.548886>.
- Thanati, Foteini, Evangelos Karatzas, Fotis A. Baltoumas, Dimitrios J. Stravopodis, Aristides G. Eliopoulos, and Georgios A. Pavlopoulos. 2021. "FLAME: A Web Tool for Functional and Literature Enrichment Analysis of Multiple Gene Lists." *Biology* 10 (7). <https://doi.org/10.3390/BIOLOGY10070665>.
- Tobón, Gabriel J., Pierre Youinou, and Alain Saraux. 2010. "The Environment, Geo-Epidemiology, and Autoimmune Disease: Rheumatoid Arthritis." *Journal of Autoimmunity* 35 (1): 10–14. <https://doi.org/10.1016/J.JAUT.2009.12.009>.
- Trapnell, Cole, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J. Lennon, Kenneth J. Livak, Tarjei S. Mikkelsen, and John L. Rinn. 2014. "The Dynamics and Regulators of Cell Fate Decisions Are Revealed by Pseudotemporal Ordering of Single Cells." *Nature Biotechnology* 32 (4): 381–86. <https://doi.org/10.1038/NBT.2859>.

- Trentham, David E., Alexander S. Townes, and Andrew H. Kang. 1977. "Autoimmunity to Type II Collagen an Experimental Model of Arthritis." *The Journal of Experimental Medicine* 146 (3): 857–68. <https://doi.org/10.1084/JEM.146.3.857>.
- Tzaferis, Christos, Evangelos Karatzas, Fotis A. Baltoumas, Georgios A. Pavlopoulos, George Kollias, and Dimitris Konstantopoulos. 2023. "SCALA: A Complete Solution for Multimodal Analysis of Single-Cell Next Generation Sequencing Data." *Computational and Structural Biotechnology Journal* 21 (January): 5382–93. <https://doi.org/10.1016/J.CSBJ.2023.10.032>.
- Vasilopoulos, Y., V. Gkretsi, M. Armaka, V. Aidinis, and G. Kollias. 2007. "Actin Cytoskeleton Dynamics Linked to Synovial Fibroblast Activation as a Novel Pathogenic Principle in TNF-driven Arthritis." *Annals of the Rheumatic Diseases* 66 (Suppl 3): iii23. <https://doi.org/10.1136/ARD.2007.079822>.
- Veroni, Caterina, Barbara Serafini, Barbara Rosicarelli, Corrado Fagnani, Francesca Aloisi, and Cristina Agresti. 2020. "Connecting Immune Cell Infiltration to the Multitasking Microglia Response and TNF Receptor 2 Induction in the Multiple Sclerosis Brain." *Frontiers in Cellular Neuroscience* 14 (July). <https://doi.org/10.3389/FNCEL.2020.00190/FULL>.
- Wei, Kevin, Ilya Korsunsky, Jennifer L. Marshall, Anqi Gao, Gerald F.M. Watts, Triin Major, Adam P. Croft, et al. 2020. "Notch Signalling Drives Synovial Fibroblast Identity and Arthritis Pathology." *Nature* 582 (7811): 259–64. <https://doi.org/10.1038/S41586-020-2222-Z>.
- Wicks, Ian P., and Andrew W. Roberts. 2016. "Targeting GM-CSF in Inflammatory Diseases." *Nature Reviews. Rheumatology* 12 (1): 37–48. <https://doi.org/10.1038/NRRHEUM.2015.161>.
- Williams, Cameron G., Hyun Jae Lee, Takahiro Asatsuma, Roser Vento-Tormo, and Ashraful Haque. 2022. "An Introduction to Spatial Transcriptomics for Biomedical Research." *Genome Medicine* 14 (1). <https://doi.org/10.1186/S13073-022-01075-1>.
- Willrich, Maria A.V., David L. Murray, and Melissa R. Snyder. 2015. "Tumor Necrosis Factor Inhibitors: Clinical Utility in Autoimmune Diseases." *Translational Research : The Journal of Laboratory and Clinical Medicine* 165 (2): 270–82. <https://doi.org/10.1016/J.TRSL.2014.09.006>.
- Wolf, F. Alexander, Philipp Angerer, and Fabian J. Theis. 2018. "SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis." *Genome Biology* 19 (1). <https://doi.org/10.1186/S13059-017-1382-0>.
- Yan, Feng, David R. Powell, David J. Curtis, and Nicholas C. Wong. 2020. "From Reads to Insight: A

Hitchhiker's Guide to ATAC-Seq Data Analysis." *Genome Biology* 21 (1).
<https://doi.org/10.1186/S13059-020-1929-3>.

Yousif, Ayman, Nizar Drou, Jillian Rowe, Mohammed Khalfan, Kristin C. Gunsalus, and Kristin C. Gunsalus. 2020. "NASQAR: A Web-Based Platform for High-Throughput Sequencing Data Analysis and Visualization." *BMC Bioinformatics* 21 (1). <https://doi.org/10.1186/S12859-020-03577-4>.

Zappia, Luke, and Fabian J. Theis. 2021. "Over 1000 Tools Reveal Trends in the Single-Cell RNA-Seq Analysis Landscape." *Genome Biology* 22 (1): 1–18. <https://doi.org/10.1186/S13059-021-02519-4/TABLES/1>.

Zhang, Fan, Kevin Wei, Kamil Slowikowski, Chamith Y. Fonseka, Deepak A. Rao, Stephen Kelly, Susan M. Goodman, et al. 2019. "Defining Inflammatory Cell States in Rheumatoid Arthritis Joint Synovial Tissues by Integrating Single-Cell Transcriptomics and Mass Cytometry." *Nature Immunology* 20 (7): 928–42. <https://doi.org/10.1038/S41590-019-0378-1>.

Zhang, Yong, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoutte, David S. Johnson, Bradley E. Bernstein, Chad Nussbaum, et al. 2008. "Model-Based Analysis of ChIP-Seq (MACS)." *Genome Biology* 9 (9). <https://doi.org/10.1186/GB-2008-9-9-R137>.

Zhou, Yingyao, Bin Zhou, Lars Paché, Max Chang, Alireza Hadj Khodabakhshi, Olga Tanaseichuk, Christopher Benner, and Sumit K. Chanda. 2019. "Metascape Provides a Biologist-Oriented Resource for the Analysis of Systems-Level Datasets." *Nature Communications* 10 (1). <https://doi.org/10.1038/S41467-019-09234-6>.

Zhu, Qin, Stephen A. Fisher, Hannah Dueck, Sarah Middleton, Mugdha Khaladkar, and Junhyong Kim. 2018. "PIVOT: Platform for Interactive Analysis and Visualization of Transcriptomics Data." *BMC Bioinformatics* 19 (1). <https://doi.org/10.1186/S12859-017-1994-0>.