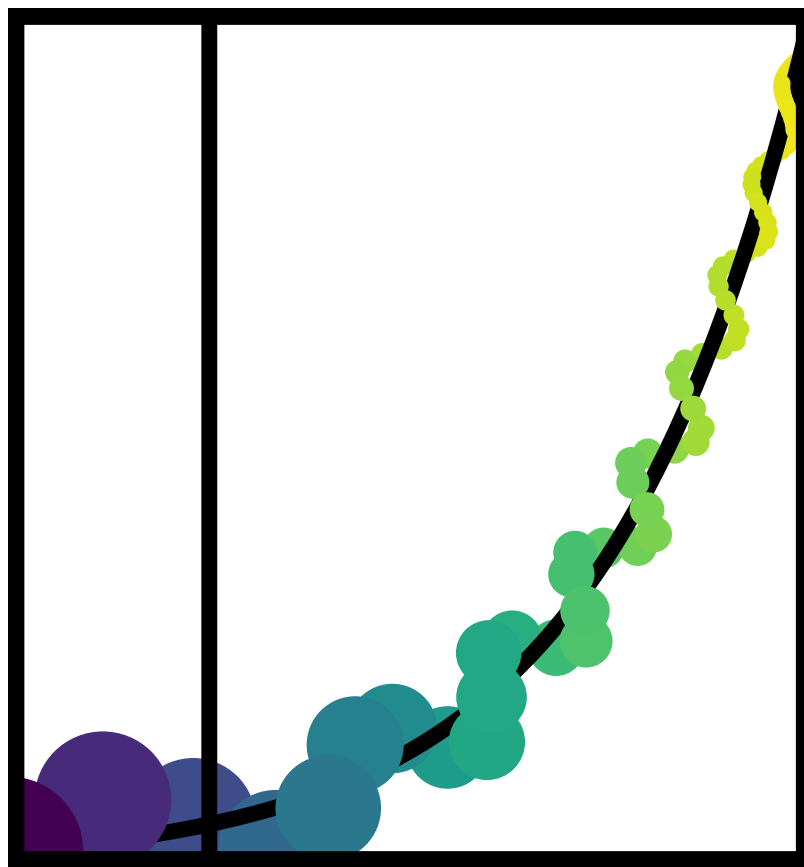


Design, Development and Evaluation of Single-cell Accessible Chromatin Sequencing Technologies



Ir. Florian De Rop

Under the supervision of

Dr. Suresh Poovathingal
Prof. Dr. Ir. Stein Aerts

KU Leuven
Groep Biomedische Wetenschappen
Faculteit Geneeskunde
Departement Menselijke Erfelijkheid



DOCTORAL SCHOOL
BIOMEDICAL SCIENCES

ONTWERP, ONTWIKKELING EN AFWEGING VAN SEQUENCINGTECHNOLOGIEËN VOOR CHROMATINE-TOEGANKELIJKHEID OP ENKELCELNIVEAU

Florian DE ROP

Jury:

Promotor: Prof. Dr. Ir. Stein Aerts
Copromotor: Dr. Suresh Poovathingal
Voorzitter leescommissie: Prof. Dr. Johan Van Lint
Voorzitter openbare verdediging: Prof. Dr. Joris Vermeersch
Juryleden: Prof. Dr. Céline Vallot, Prof. Dr. Klaas Mulder,
Prof. Dr. Ir. Jeroen Lammertyn, Prof. Dr. Thierry Voet

Proefschrift voorgedragen
tot het behalen van de
graad van Doctor in de
Biomedische
Wetenschappen

1 Oktober 2025

KU Leuven
Biomedical Sciences Group
Faculty of Medicine
Department of Human Genetics



DESIGN, DEVELOPMENT AND PERFORMANCE OF SINGLE-CELL ACCESSIBLE CHROMATIN SEQUENCING TECHNOLOGIES

Florian DE ROP

Jury:

Supervisor: Prof. Dr. Ir. Stein Aerts
Co-supervisor: Dr. Suresh Poovathingal
Chair examining committee: Prof. Dr. Johan Van Lint
Chair public defence: Prof. Dr. Joris Vermeersch
Jury members: Prof. Dr. Céline Vallot, Prof. Dr. Klaas Mulder,
Prof. Dr. Ir. Jeroen Lammertyn, Prof. Dr. Thierry Voet

Dissertation presented in
partial fulfilment of the
requirements for the
degree of Doctor in
Biomedical Sciences

October 1st, 2025

Contents

Abstract	ix
Layman's abstract	x
Vulgariserend abstract	xi
Contributions	xii
Author profile	xiv
Conflict of interest statement	xvi
Statement on use of generative AI	xvii
Acknowledgements	xviii
1 Introduction	1
1.1 Pioneers	1
1.2 The early molecular era	3
1.3 DNA-binding proteins	5
1.4 Sequencing the human genome	8
1.5 Quantitative sequencing assays	11
1.6 Single-cell RNA and ATAC-seq protocols	14
1.7 Scaling cell throughput	19
1.8 Perturbations	20
1.9 Single-cell data pre-processing	22
1.10 Single-cell data post-processing	25
1.11 Deep-learning models	27
1.12 Motivation and approach	29
2 HyDrop enables droplet-based single-cell ATAC-seq and single-cell RNA-seq using dis-	
solvable hydrogel beads	31
3 Systematic benchmarking of single-cell ATAC-sequencing protocols	59
4 Discussion	89
4.1 Development and evaluation of single-cell assays	89
4.2 Scaling omics data for predictive modelling	91
A Supplementary note: Design and development of HyDrop-ATAC and HyDrop-RNA	99
A.1 Method design and prior work	99
A.2 Execution and troubleshooting	101
A.3 Key results	103
B Supplementary note: Benchmarking single-cell chromatin accessibility sequencing meth-	
ods	105
B.1 Analysis	105
B.2 Key results	106
Bibliography	109

List of Figures

1.1	Spemann-Mangold organizer transplant	2
1.2	Waddington's epigenetic landscape	3
1.3	Jacob & Monod's gene regulatory network	4
1.4	Sanger sequencing	5
1.5	DNA condensation	7
1.6	Trends in sequencing platforms	10
1.7	FASTQ snippet.	10
1.9	Tn5 complexed with DNA	13
1.10	Trends in sequencing assays	14
1.11	Common single-cell perturbation vector designs	21
1.12	SAM snippet.	23
1.13	Barcode rank simulation	24
A.1	Article title page of "Hydrop enables droplet-based single-cell ATAC-seq and single-cell RNA-seq using dissolvable hydrogel beads"	98
A.2	Bead diameter optimisation	100
A.3	Droplet encapsulation chip design	100
A.5	Contaminated mixed-species count plot	102
B.1	Article title page of "Systematic benchmarking of single-cell ATAC sequencing protocols"	104

Abstract

Gene expression and chromatin accessibility profiling provide complementary views of gene regulation across tissues, developmental stages, and disease. While bulk sequencing assays average these signals across populations of cells, single-cell RNA-seq and single-cell ATAC-seq can resolve cellular heterogeneity and enable the characterization of individual cell types and states. The rapid development of single-cell technologies has highlighted two related challenges: developing methods that make high-quality single-cell profiling more accessible, and understanding how differences between technologies influence the biological conclusions drawn from the resulting data. This thesis addresses both challenges through the development of an open-source single-cell sequencing platform and a systematic benchmark of scATAC-seq technologies.

To make single-cell profiling more accessible, we developed HyDrop, a flexible, open-source droplet platform for scATAC-seq and scRNA-seq based on dissolvable polyacrylamide hydrogel beads. These beads carry split-pool barcodes which index RNA or accessible chromatin from single cells. We optimized the protocol by tuning primer load, adding a clathrin inhibitor, and titrating different enzymes to increase yield. HyDrop-ATAC generated 7996 high-quality single-cell profiles from flash-frozen mouse cortex. HyDrop-RNA produced 9508 transcriptomes from the same tissue. We also recovered a small population of rare, FACS-sorted *Drosophila* neurons, demonstrating applicability to low-input samples. Across mouse cortex, HyDrop-RNA reached a median 3389 UMIs and 1658 genes per cell, and HyDrop-ATAC recovered a median 4148 unique fragments per cell with high TSS enrichment. Chemistry, microfluidic chips, and protocols are open-source.

To assess how technology choices affect data quality and downstream analyses, we systematically compared eight scATAC-seq technologies, including HyDrop. To enable consistent processing across technologies, we developed PUMATAC, a containerized preprocessing framework that standardizes trimming, whitelist-aware barcode correction, and alignment across different scATAC-seq formats. The benchmark comprised eight scATAC-seq methods applied to human PBMCs: 10x v1, v1.1, v2, 10x Multiome, Bio-Rad ddSEQ, mtscATAC-seq, s3-ATAC, and HyDrop, for a total of 47 libraries and 169 227 high-quality cells. We detected various systematic differences across these technologies. Sequencing efficiency, defined as fraction of reads retained in the final cell-by-peak matrix, ranged from ~4% for s3-ATAC and HyDrop to ~28% for 10x v2. Per-cell sensitivity scaled with sequencing efficiency and impacted cell-type annotation accuracy, genotype demultiplexing, peak calling, differential accessibility, and transcription factor motif enrichment. Methods also differed in specificity: fragment promoter and TSS proximity as well as aggregate fraction of fragments in peaks varied markedly, with ddSEQ recovering more TSS-proximal fragments than 10x. This effect led to technique-dependent differentially accessible regions, and imperfect cross-technology integration even after batch correction.

Together, these studies address both the development and evaluation of single-cell sequencing technologies. HyDrop provides an accessible platform for generating single-cell RNA and chromatin accessibility profiles, while the benchmark demonstrates that technical differences between scATAC-seq methods can substantially affect sensitivity, specificity, and downstream biological analyses. These results highlight that improvements in single-cell genomics should be evaluated not only in terms of throughput, but also by the quality and biological informativeness of the signal recovered from each cell. HyDrop provides a platform on which such improvements can be developed, while the benchmarking framework provides a systematic way to assess their impact.

Layman's abstract

How can cells with identical DNA become neurons, muscle cells, or immune cells? The answer lies in gene regulation - the selective use of genes that defines cell identity, drives development, and fails in disease. Understanding this process has long been a central question in biology. Artificial intelligence is now transforming how we approach it. Models can accurately predict protein folding or forecast which genes are switched on or off in individual cells. These advances allow researchers to test ideas on a computer before turning to the lab, increasing the odds of success. Yet all such models depend on one crucial ingredient: large amounts of high-quality data.

One important source of such data comes from the field of *omics*, which broadly refers to systematic ways of measuring biology at scale. This thesis focuses on two such approaches at the single-cell level: single-cell RNA sequencing (scRNA-seq), which measures which genes are active in each cell, and single-cell ATAC-seq, which measures which region of DNA a cell is "using" at a given time. Together, these methods provide a small part of the puzzle of how cells work.

In the first part of this thesis, we developed HyDrop, an open-source method for generating single-cell data. The method uses tiny droplets of liquid to capture and read out information from thousands of individual cells in parallel. Unlike expensive commercial systems, HyDrop is transparent, flexible, and affordable, making cutting-edge single-cell experiments accessible to many more labs.

In the second part, we systematically compared 8 different methods for single-cell chromatin accessibility sequencing. Different technologies often give very different results, and without careful evaluation it is difficult to know which findings can be trusted. By benchmarking protocols side by side, we showed how trade-offs in cost, sensitivity, and data quality influence the biological insights one can draw. This helps researchers choose the right tool for their question, ensures resources are used efficiently, and improves the reliability of downstream analyses.

Taken together, these contributions meet two essential needs: creating new ways to generate high-quality biological data, and setting clear standards to evaluate the methods that produce it. Our results show that technology choices directly shape scientific conclusions from how cell types are identified to how subtle regulatory signals are detected.

Finally, we connect these advances to a broader trend. Just as AI in language and vision achieved breakthroughs by scaling data and computation, biology is entering a similar phase. Sequencing costs are falling, automation is rising, and massive datasets are becoming feasible. With open methods for data generation and rigorous frameworks for evaluation, we can move toward predictive models that link DNA to cellular behavior and ultimately to disease. This PhD thesis contributes to building that foundation.

Vulgariserend abstract

Hoe kan het dat cellen met identiek DNA toch uitgroeien tot bijvoorbeeld zenuwcellen, spiercellen of immuuncellen? Het antwoord ligt in genregulatie: de selectieve genexpressie die celidentiteit bepaalt, ontwikkeling stuurt en ziekten veroorzaakt wanneer zij faalt. Dit is al decennialang een kernvraag in de biologie. Kunstmatige intelligentie verandert nu hoe we dit vraagstuk benaderen. Modellen kunnen de 3D-structuur van eiwitten voorspellen of aangeven welke genen in afzonderlijke cellen op een bepaald moment aan- of uitgeschakeld zijn. Zulke modellen maken het mogelijk hypothesen eerst op de computer te toetsen voor we naar het lab gaan, waardoor de kans op succes aanzienlijk toeneemt. Maar alle modellen steunen op één cruciale bouwsteen: grote hoeveelheden data van hoge kwaliteit.

Een belangrijke bron daarvan komt uit de zogenaamde "omics", de verzamelnaam voor systematische metingen op grote schaal. Dit proefschrift richt zich op twee benaderingen op single-cell-niveau: single-cell RNA-sequencing (scRNA-seq), een techniek die meet welke genen per cel actief zijn, en single-cell ATAC-seq, een complementaire techniek die meet welke regio's van het DNA van een cel op een bepaald moment toegankelijk zijn. Samen leveren deze methoden een stukje van de puzzel van het functioneren van cellen.

In het eerste deel van dit project hebben we HyDrop ontwikkeld, een open-source methode om single-cell data te genereren. De methode gebruikt kleine druppels vloeistof om duizenden individuele cellen parallel te vangen, en hun RNA of chromatine uit te lezen. In tegenstelling tot dure commerciële systemen is HyDrop transparant, flexibel en betaalbaar, waardoor geavanceerde single-cell-experimenten voor veel meer laboratoria toegankelijk worden.

In het tweede deel hebben we acht methoden voor scATAC-seq systematisch vergeleken. Verschillende technologieën leveren vaak uiteenlopende resultaten op en zonder zorgvuldige evaluatie is het moeilijk te bepalen welke data betrouwbaar is. Door protocollen met elkaar te vergelijken, laten we zien hoe een balans tussen kosten, sensitiviteit en kwaliteit de biologische inzichten sturen. Dit helpt onderzoekers het juiste instrument voor hun vraag te kiezen, middelen efficiënt in te zetten en de betrouwbaarheid van analyses te verhogen.

Samen voorzien deze bijdragen in twee kernbehoeften: nieuwe manieren om hoogwaardige biologische data te genereren en duidelijke standaarden om de methoden die deze data opleveren te beoordelen. Onze resultaten laten zien dat technologiekeuzes wetenschappelijke conclusies direct vormgeven, van de identificatie van celtypen tot het detecteren van subtiele genregulatorische signalen.

Tot slot plaatsen we deze vooruitgang in een bredere ontwikkeling. Net als AI in taal en beeld doorbraken bereikte door data en rekenkracht fors op te schalen, staat de biologie aan het begin van een vergelijkbare fase. Sequencingkosten dalen, automatisering neemt toe en zeer grote datasets worden haalbaar. Met open methoden voor datageneratie en strikte kaders voor evaluatie kunnen we toewerken naar voorspellende modellen die DNA verbinden met cellulair gedrag en uiteindelijk met ziekte. Dit PhD-project helpt die basis te leggen.

Contributions

Scientific contributions

Within the Aerts lab, we gratefully acknowledge the contributions of Stein Aerts (conceptualization, methodology, project administration, supervision, funding acquisition, writing original draft, and writing review and editing), Carmen Bravo González-Blas (formal analysis, software, writing review and editing, supervision, and funding acquisition), Valerie M. Christiaens (investigation and resources), Christopher Campbell Flerin (conceptualization, methodology, software, formal analysis, writing original draft, and writing review and editing), Gert J. Hulselmans (data curation, software, and formal analysis), Joy N. Ismail (investigation and writing review and editing), Jasper Janssens (software), Suresh Poovathingal (conceptualization, methodology, investigation, validation, project administration, supervision, funding acquisition, writing original draft, and writing review and editing), Koen Theunis (methodology), and Jasper Wouters (writing review and editing).

We further thank our external collaborators for their essential support. These include Andrew C. Adey (supervision and funding acquisition), S. Tansu Bagdatli (investigation and resources), Theodore P. Braun (supervision and funding acquisition), Jason D. Buenrostro (supervision and funding acquisition), Zach Burkett (resources), Ginevra Caratù (investigation and resources), M. Ryan Corces (investigation and resources), Bart Deplancke (supervision and funding acquisition), Zain Y. Dossani (investigation and resources), Fabiana Duarte (investigation and resources), William J. Greenleaf (supervision and funding acquisition), Holger Heyn (conceptualization, methodology, writing original draft, supervision, and funding acquisition), Andrew Knights (investigation and resources), Camille Lucie Germaine Lambert (investigation and resources), Ronald Lebofsky (resources, supervision, and funding acquisition), Leif S. Ludwig (supervision and funding acquisition), Wendy Luo (investigation and resources), Lira Mamanova (investigation and resources), Domenica Marchese (investigation and resources), Gabriele Marcassa (resources), Patrick Maschmeyer (investigation and resources), Kerstin B. Meyer (investigation and resources), Christoph Muus (investigation and resources), Ryan Mulqueen (investigation and resources), Sebastián R. Najle (investigation and resources), Jörn Pezoldt (investigation and resources), Satish Pillai (supervision and funding acquisition), Albert Rafels (formal analysis), Aviv Regev (supervision and funding acquisition), Orit Rozenblatt-Rosen (investigation and resources), Rojesh Shrestha (investigation and resources), Michael Slyper (investigation and resources), Arnau Sebé-Pedrós (supervision and funding acquisition), Paula Soler-Vila (formal analysis), Sarah A. Teichmann (supervision and funding acquisition), Joris de Wit (resources), and Luciano G. Martelotto (resources).

Personal contributions

Florian V. De Rop contributed to conceptualization, methodology, data curation, formal analysis, investigation, validation, visualization, software, funding acquisition, writing original draft, writing review and editing, and resources.

Funding

Within Aerts lab, support was provided by the European Research Council H2020 (724226 "cis-CONTROL")

to Stein Aerts; 810287 "BCLLATLAS"; 874710 "ESPACE"), KU Leuven (C14/18/092 and C14/22/125 to Stein Aerts), and the Research Foundation Flanders (FWO) (G0B5619N, G0I2722N, G094121N to Stein Aerts; PhD fellowship to Florian V. De Rop). Additional support came from Aligning Science Across Parkinson's (ASAP-000430 to Stein Aerts and Christopher Campbell Flerin) and the Michael J. Fox Foundation for Parkinson's Research (ASAP-000430 associated support to Christopher Campbell Flerin). The Foundation Against Cancer, Belgium (2016-070 and F/2020/1396 to Stein Aerts; additional support to Jasper Wouters) and the Belgian Cancer Society (support to Jasper Wouters) contributed as well. VIB Tech Watch supported Suresh Poovathingal.

Externally, MCIN/AEI, Spain supported Holger Heyn (PID2020-115439GB-I00). The Wellcome Trust supported Kerstin B. Meyer and Sarah A. Teichmann (WT211276/Z/18/Z) and provided a Sanger core grant (WT206194). Support also came from the National Institute on Aging (NIH) (K99/R00 AG059918 to M. Ryan Corces) and further NIH grants to William J. Greenleaf (RM1-HG007735, UM1-HG009442, UM1-HG009436, R01-HG00990901, U19-AI057266), along with the Rita Allen Foundation, the Human Frontiers Science Program (RGY006S), the Chan Zuckerberg Initiative (2017-174468; 2018-182817), and the Emerson Collective. The Swiss National Science Foundation (310030_197082) and EPFL supported Bart Deplancke. The German Research Foundation (DFG) (Emmy Noether LU 2336/2-1), the NIH (UM1 HG012076), the Longevity Impetus Grant, and the Hector Research Career Development Award supported Leif S. Ludwig. Finally, Andrew C. Adey was supported by the NIH (RF1-MH128842; R35-GM124704; R01-DA047237) and the Silver Family Foundation Innovator Award.

Author profile

Florian De Rop (1996) is a bioscience engineer. He was a researcher at the VIB/KU Leuven Stein Aerts Lab of Computational Biology between 2018 and 2023 under Stein Aerts, Suresh Poovathingal, and Jasper Janssens. His students include Koenraad De Wispelaere and Natalie Roels. He currently lives and works in the San Francisco Bay Area.

Journal publications

1. Dickmanken, H., Wojno, M., Mahieu, L., Theunis, K., Ekşi, E. C., Christiaens, V., Kempynck, N., **De Rop, F. V.**, Roels, N., Spanier, K. I., Vandepoel, R., Hulselmans, G., Poovathingal, S., & Aerts, S. (2026). Evaluating single-cell ATAC-seq atlasing technologies using sequence-to-function modeling. *Nature Communications*, 17, 1951. <https://doi.org/10.1038/s41467-026-68742-4>
2. **De Rop, F. V.**, Hulselmans, G., Flerin, C., Soler-Vila, P., Rafels, A., Christiaens, V., González-Blas, C. B., Marchese, D., Caratù, G., Poovathingal, S., Rozenblatt-Rosen, O., Slyper, M., Luo, W., Muus, C., Duarte, F., Shrestha, R., Bagdatli, S. T., Corces, M. R., Mamanova, L., Knights, A., Meyer, K. B., Mulqueen, R., Taherinasab, A., Maschmeyer, P., Pezoldt, J., Lambert, C. L. G., Iglesias, M., Najle, S. R., Dossani, Z. Y., Martelotto, L. G., Burkett, Z., Lebofsky, R., Martin-Subero, J. I., Pillai, S., Sebé-Pedrós, A., Deplancke, B., Teichmann, S. A., Ludwig, L. S., Braun, T. P., Adey, A. C., Greenleaf, W. J., Buenrostro, J. D., Regev, A., Aerts, S., & Heyn, H. (2024). Systematic benchmarking of single-cell ATAC-sequencing protocols. *Nature Biotechnology*, 42, 916--926. <https://doi.org/10.1038/s41587-023-01881-x>
3. **De Rop, F. V.**, & Heyn, H. (2024). Benchmarking of single-cell ATAC sequencing tools. *Nature Biotechnology*, 42, 856--857. <https://doi.org/10.1038/s41587-023-01897-3>
4. **De Rop, F. V.**, Ismail, J. N., González-Blas, C. B., Hulselmans, G. J., Flerin, C. C., Janssens, J., Theunis, K., Christiaens, V. M., Wouters, J., Marcassa, G., de Wit, J., Poovathingal, S., & Aerts, S. (2022). HyDrop enables droplet-based single-cell ATAC-seq and single-cell RNA-seq using dissolvable hydrogel beads. *eLife*, 11, e73971. <https://doi.org/10.7554/eLife.73971>
5. Janssens, J., Aibar, S., Taskiran, I. I., Ismail, J. N., Estacio Gomez, A., Aughey, G., Spanier, K. I., **De Rop, F. V.**, González-Blas, C. B., Dionne, M., Grimes, K., Quan, X. J., Papasokrati, D., Hulselmans, G., Makhzami, S., De Waegeneer, M., Christiaens, V., Southall, T., & Aerts, S. (2022). Decoding gene regulation in the fly brain. *Nature*, 601(7894), 630--636. <https://doi.org/10.1038/s41586-021-04262-z>
6. Clark, E. C., **De Rop, F.**, Jimenez Garcia, I. A., Nogal Macho, A., Mannette, R. A., Nova Blanco, J. R., Ramaekers, K., & Vandermeulen, C. (2019). A systems map to elucidate the factors influencing vaccine coverage. *Transdisciplinary Insights*, 3(1), 115--150. <https://doi.org/10.1111/TDI2019.3.5>
7. Dal Dosso, F., Decrop, D., Pérez-Ruiz, E., Daems, D., Agten, H., Al-Ghezi, O., Bollen, O., Breukers, J., **De Rop, F.**, Katsafadou, M., Lepoudre, J., Lyu, L., Piron, P., Saesen, R., Sels, S., Soenen, R., Staljanssens, E., Tara-

porewalla, J., Kokalj, T., Spasic, D., & Lammertyn, J. (2018). Creasensor: SIMPLE technology for creatinine detection in plasma. *Analytica Chimica Acta*, 1000, 191--198. <https://doi.org/10.1016/j.aca.2017.11.026>

Published laboratory protocols

1. Poovathingal, S., Wojno, M., Theunis, K., **De Rop, F.**, & Aerts, S. (2025). *HyDrop v2 Bead Generation & Ligation Barcoding v2*. protocols.io. <https://doi.org/10.17504/protocols.io.8epv5n11dv1b/v2>
2. Poovathingal, S., **De Rop, F.**, Christiaens, V., Theunis, K., & Aerts, S. (2025). *HyDrop-v2-ATAC v1*. protocols.io. <https://doi.org/10.17504/protocols.io.x54v97mmpg3e/v1>
3. **De Rop, F.**, Aerts, S., & Poovathingal, S. (2023). *Microfluidic Chip Production v1.1*. protocols.io. <https://doi.org/10.17504/protocols.io.6qpvrq7pgmk/v4>
4. **De Rop, F.**, Poovathingal, S., & Aerts, S. (2023). *HyDrop-RNA v1.0*. protocols.io. <https://doi.org/10.17504/protocols.io.dm6gpwqjjlzp/v5>
5. **De Rop, F.**, Poovathingal, S., & Aerts, S. (2022). *HyDrop-ATAC v1.0*. protocols.io. <https://doi.org/10.17504/protocols.io.b4xvqxn6>
6. **De Rop, F.**, Poovathingal, S., & Aerts, S. (2022). *HyDrop Bead Generation & PCR Barcoding v1.0*. protocols.io. <https://doi.org/10.17504/protocols.io.n92ld9np9g5b/v11>

Talks and scientific presentations

- **2023** – Speaker, VIB Conference *Revolutionizing Next-Generation Sequencing*, fifth edition, Ghent, Belgium.
- **2023** – Speaker, Advances in Genome Biology and Technology General Meeting, Hollywood, Florida, United States.
- **2022** – Speaker, KU Leuven Institute for Single Cell Omics Kickoff Day, Leuven, Belgium.
- **2022** – Speaker, Single Cell Genomics, Utrecht, the Netherlands.
- **2021** – Visiting researcher and seminar speaker, Centre for Genomic Regulation, Barcelona, Spain.
- **2021** – Visiting researcher and seminar speaker, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.
- **2021** – Presenter, Aligning Science Across Parkinson's Celebration of Scientific Achievement, virtual meeting.
- **2021** – Short talk, LifeTime Conference 2.0: *HyDrop: a custom droplet microfluidics assay for single-cell ATAC-seq and single-cell RNA-seq*.

Conflict of interest statement

The author declares no conflict of interest.

Statement on use of generative AI

I did use the generative AI assistance tools Claude Sonnet 3.5, GPT-4, GPT-4.5, GPT-5, and Perplexity during the research/writing process of this thesis.

- Claude and GPT were used to generate all code for the new figures in this thesis, including the generation of the front page graphic.
- GPT was used to support the literature review by helping me understand or refresh concepts.
- Perplexity was used as an additional search engine for up-to-date literature discovery, complementing traditional methods such as review papers, bibliographies, and databases like PubMed.
- GPT was used extensively for $\LaTeX$ formatting of the thesis.
- This AI statement was entirely AI-generated.

Generative AI has only been used in accordance with the KU Leuven guidelines. I have reviewed and edited the content as needed, and I take full responsibility for the content of the thesis.

All writing in this thesis is original. The complete `git` commit history documents the progress of the thesis and can be used to verify that it was written by me.

Acknowledgements

Over the years, I have been asked a number of times if I would recommend the experience of going through a PhD program. My answer has always been a resounding "No!" followed by a brief pause and "...unless you can find a really good lab!" This answer is usually followed by the simple question: "Well, what makes a good lab?" First and foremost, the staff and students should be friendly and smart. You will draw on them for inspiration and companionship, and the effects of good colleagues on your own work will be profound. Second, there must be infrastructure and funding available to support good research. Both of these are easy for the would-be student to verify: go and visit, and take a look. Finally, the person in charge should be a visionary. He or she *must* have a clear and unique idea of what the future will look like, and how the lab will contribute to, and interact with it. Only then can truly novel and original output be produced consistently.

I was fortunate to have ended up, mostly by chance, in a lab where all three conditions were met exceedingly. At Aerts lab, I have met the kindest and most intelligent people who made research both fun and inspiring. During - and after - my tenure there, funding was ample, allowing us to invest in crazy, low-percentage ideas. Admittedly, most of the time we were just figuring out and doing things for the first time ourselves. A lab where this is possible can only exist under good leadership. Staying at the leading edge of a field requires the courage to commit to unique ideas, which are sometimes not clearly distinguishable from madness, and to effectively bet your continued existence on risky hypotheses over and over again. Keeping all members of an organization aligned during the stressful times that these projects invariably generate, is challenging in itself, let alone succeeding in the science to boot.

I would like to thank my advisor Stein, for building this environment. It is thanks to you that I find myself in the midst of an area of science that walks hand in hand with the most incredible revolution of our time. I would also like to thank my supervisor Suresh for his excellent daily support. We endured so many long nights in the lab building HyDrop together! Third, I want to thank Jasper for his supervision during my master thesis, which led to this work. All three of you have left an imprint on who I am today.

I want to thank Linda, for supporting me throughout. I have learned a lot from you.

Finally, I want to thank my parents, Dominique & Johan, for giving me everything.

Tôi xin dành bài luận văn này để tri ân tới cha mẹ đẻ của tôi, những người đã hy sinh tất cả vào năm 1996 để con trai họ có được một tương lai tốt đẹp hơn

Chapter 1

Introduction

The following introductory chapters will attempt to build the foundation for the explorations that follow by providing a historical overview of chromatin accessibility based on historical accounts¹⁻⁷ and frame it within the ongoing research on decoding gene regulation. Then, we illustrate the state of the art of single-cell chromatin sequencing technologies when I began this work, and delve into the capabilities and limitations of several breakthrough techniques. This context serves to highlight the importance of the subject matter, and to motivate our subsequent discussions and findings. Finally, machine learning advances in various -omics will be briefly discussed.

1.1 Pioneers

The word "cell" was first used in 1665 by Robert Hooke to describe the hexagonal walls he observed in a thin slice of cork through a self-made microscope.⁸ It took several more centuries of research until the concept of cells as the foundational building blocks of life began to take shape. Between 1817 and 1826, Baltic-German polymath and explorer Karl Ernst von Bär gave the first accounts of egg cells, development of the blastula, the notochord and the three germ layers. Only in 1839, Theodor Schwann proposed that animals, just like plants, were composed of cells,⁹ and Rudolf Virchow and Robert Remak later added that cells were the product of division of pre-existing cells.¹⁰ While studying this process of division in 1879, Walther Flemming first used the term "chromatin" to describe the filament-like threads between two dividing cells.¹¹ The term, derived from the Greek *χρῶμα* (*chroma*, color) referred to the filaments' tendency to absorb dyes. Flemming's observations did not differ much from his colleague Wilhelm Waldeyer's description of "chromosomes".⁴ Both of these early findings directly stemmed from the technological refinement of microscopic lenses, which had by now become accurate enough to permit detailed studies of developing embryos. This early area of cell biology produced a number of fundamental questions that would engage the minds of biologists to this day:

1. How can a single fertilised egg cell divide and self-organise into an entire organism?
2. Where is the information that steers this process encoded, and how is it interpreted?
3. How can this process go wrong, leading to disease, and can we manipulate it?

At that time, the concept of physical store and transfer of genetic information was not yet widely accepted, and the function and identity of Flemming and Waldeyer's chromosomes was still unknown. In search of answers,

¹Slack, Nat Rev Genet, 2002.

²Olins, Nat Rev Mol Cell Biol, 2003.

³Cobb, Curr Biol, 2015.

⁴Deichmann, Persp Biol Med, 2015.

⁵Deichmann, Dev Biol, 2016.

⁶Lis, Nat Struct Mol Biol, 2019.

⁷Roeder, Nat Struct Mol Biol, 2019.

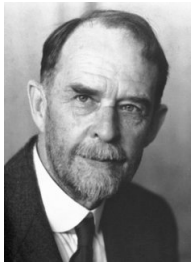
⁸Hooke, Micrographia, 1665.

⁹Schwann, Mikroskopische ... Pflanzen, 1839.

¹⁰Rudolf, Die Cellularpathologie ... Gewebelehre, 1858.

¹¹Flemming, Arch mikrosk Anat, 1879.

marine embryologist Thomas Hunt Morgan began studying inheritance in *Drosophila melanogaster* (fruit fly) in 1908. After years of breeding flies, a mutant with white eyes (aptly named *white*) appeared. Through careful tracking, Morgan discovered that these pigmentless eyes followed a sex-dependent inheritance pattern which could be reliably predicted across generations.¹² More than half a century earlier, Moravian friar Gregor Mendel had demonstrated the exact same principles in *Pisum sativum* (sweet pea), but his work was initially ignored and later disputed by the scientific community upon its rediscovery in 1900. By 1909, however, Danish botanist Wilhelm Johannsen had experimentally repeated much of Mendel's work, and coined the term "gene" to describe the abstract "unit of heredity" that is used to pass on traits between generations.¹³



T. Morgan

Morgan, together with his undergraduate student Alfred Sturtevant, discovered that co-inheritance of genes also followed reproducible patterns ("linkage analysis").¹⁴ This finding supported the idea that genes were located on a linear carrier. A second breakthrough came when Morgan demonstrated that the chromosomes observed by Flemming and Waldeyer were in fact the physical bodies responsible for the transfer of phenotypic traits,¹⁵ a discovery which yielded him a Nobel prize for physiology in 1933. Armed with experimental evidence gathered from thousands of generations of fruit flies, Morgan formally unified the theory of Mendelian inheritance with the physical transfer of genes on chromosomes in 1915.¹⁶

Together, Morgan's findings inaugurated the 20th century's era of genetics, and defined the context for many of the major discoveries that ensued. Hermann Muller, for example, discovered the mutagenic effects of irradiation in *Drosophila*,¹⁷ accelerating the study of mutations,¹⁸ and producing a Nobel prize for physiology in 1946. In fact, Muller showed that relocation of genes, inferred through Morgan and Sturtevant's linkage analysis, could alter their "variegation" (expression), a very early understanding of cis gene regulation!

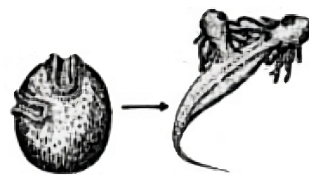


FIGURE. 1.1. Spemann-Mangold organizer transplantation. Reprod. from [19]

In parallel, the field of embryology began to consolidate. Early views such as "preformationism," the idea that all parts of the body were present in miniature form in the embryo at conception, were defeated when Hans Spemann began constricting dividing *Triturus* (newt) embryos. Both halves of the embryo could develop into full, though smaller, animals *if* part of the dorsal side of the embryo was present. Later, Spemann and his student Hilde Mangold, discovered that transplantation of the dorsal lip from one embryo to the ventral side of another embryo would induce the formation of a ventrally conjoined twin (fig. 1.1).²⁰ Importantly, the pair showed that the conjoined secondary embryo consisted of tissue from the *host*, indicating that the transplant merely instructed its host surroundings to change developmental fate. Spemann and Mangold later named this steering tissue the "organizer field," and multiple labs attempted (unsuccessfully) to purify and characterise the compounds responsible for these processes in an "organizer gold rush".¹

However, the link between embryonic development and the "new" insights in genetics was not readily made, even disputed by parts of the scientific community. This sentiment began to change after a young Conrad Waddington, who had previously been performing organizer grafting experiments in mammals and birds, visited Morgan's fly lab in 1935. Here, Waddington discovered that specific heritable genetic mutations could disrupt *Drosophila* wing development at specific stages.²¹ Based on these findings, Waddington wrote his seminal work *Organisers and Genes* in 1940, and introduced the notion that embryonic development was a large decision tree under the control of genes, which he metaphorically described as a "landscape".^{22,23} A ball, representing a cell's current state, metaphorically rolls down the furrows in the landscape, passing several bifurcations which ultimately determine its fate (fig. 1.2a). Beneath the surface, pegs symbolize genes, while tension wires represent

¹²Morgan, *Science*, 1910.

¹³Johannsen, *Elemente ... Vorlesungen*, 1909.

¹⁴Sturtevant, *J Exp Zool*, 1913.

¹⁵Morgan, *Science*, 1911.

¹⁶Morgan, *Mechanism of Mendelian Heredity*, 1915.

¹⁷Muller, *Science*, 1927.

¹⁸Muller, *J Genetics*, 1930.

²⁰Spemann, *Arch Mikrosk Anat*, 1924.

²¹Waddington, *J Genetics*, 1940.

²²Waddington, *Organisers and Genes*, 1940.

²³Waddington, *The Strategy of the Genes*, 1957.

their chemical effects and pull the surface into shape (fig. 1.2b). Disruptions in an organism's genetic context, for example by mutation, could alter the landscape and thus the cell's final state.

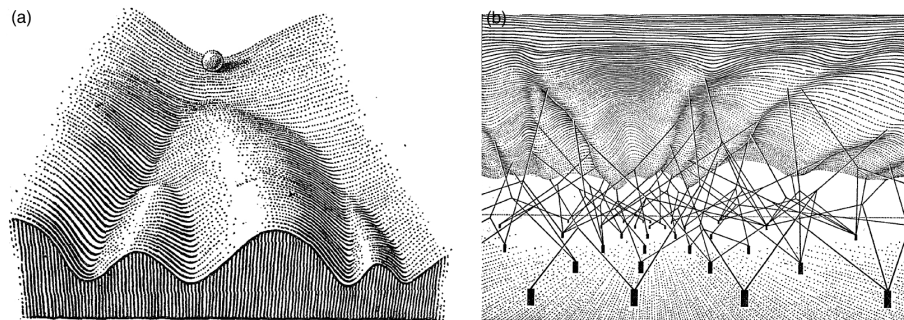


FIGURE. 1.2. Waddington's epigenetic landscape. a. A ball, representing a cell, rolls down the landscape into a stable state after passing several developmental bifurcations. b. Pegs represent genes, while guy wires under tension represent their chemical effects and define the shape of the landscape. Reproduced from [23].

1.2 The early molecular era

It must be noted that the genes Morgan and Waddington addressed were still somewhat vaguely-defined, as their exact physical nature beyond the chromosome was still unknown. In fact, the storage of genetic information was attributed to proteins at the time, as they made up a much larger portion of the chromosomal dry mass than nucleic acids (DNA, RNA), and were known to be more complex in composition. This notion was overturned in 1944, when Avery, MacLeod, and McCarty demonstrated that it was transfer of nucleic acids, and not proteins, purified from a non-virulent strain of bacteria that could "transform" a virulent strain to the non-virulent strain.^{24,25} Despite uncertainties about the physical nature of genetic information, awareness of the link between disease phenotypes, proteins and genes began to rise. For example, sickle cell anemia, which was known to follow a Mendelian, autosomal recessive inheritance pattern, was shown by Linus Pauling to result from an aberrant form of hemoglobin in red blood cells.²⁶ Using electrophoresis, Pauling and colleagues demonstrated that hemoglobin from affected individuals migrated differently in an electric field, and that people with the trait had a mixture of normal and sickle cell hemoglobin. Since hemoglobin is a protein, the most logical explanation was that the gene must encode the structure of the protein, for the first time showing "*a clear case of a change produced in a protein molecule by an allelic change in a single gene*". Pauling was later awarded the Nobel prize for Chemistry in 1954, and for peace in 1962. Another major breakthrough in the molecular understanding of genetic information came when Watson, Crick, and Franklin resolved DNA's double-helix structure using X-ray crystallography,²⁷ and for the first time suggested that "*the precise sequence of the bases is the code which carries the genetical information*".²⁸ It had thus become clear that life stores information in the nucleotide sequence of DNA, and Watson and Crick shared the Nobel prize for physiology for this discovery in 1962.



C. Waddington

Although Morgan and Waddington's visions of genes as decision makers for a cell's behaviour were largely theoretical, they were soon substantiated. Jacques Monod, also a former student of Thomas Hunt Morgan, observed that bacteria grown in a medium containing both glucose and lactose would preferentially feed on glucose, and would only begin feeding on lactose once the other had run out. Further research together with François Jacob showed that the production of β -galactosidase, the enzyme responsible for this metabolic switch,

²⁴Griffith, J Hyg, 1928.

²⁵Avery, J Exp Med, 1944.

²⁶Pauling, Science, 1949.

²⁷Watson, Nature, 1953.

²⁸Watson, Nature, 1953.

was under the control of a "repressor," which was inactivated by the presence of lactose. Together, the pair published these findings in 1961,²⁹ in an article which features the first diagram of a *gene regulatory network* (fig. 1.3). The Jacob-Monod diagram also explicitly showed the repressor as a protein originating from a gene. Genes thus (indirectly) controlled other genes, and therefore the behaviour of the cell - precisely how Conrad Waddington had prophesied twenty years prior! Jacob & Monod were jointly awarded the Nobel prize for physiology for the discovery of genetic control of enzymes in 1965. Earlier that month, Jacob, together with

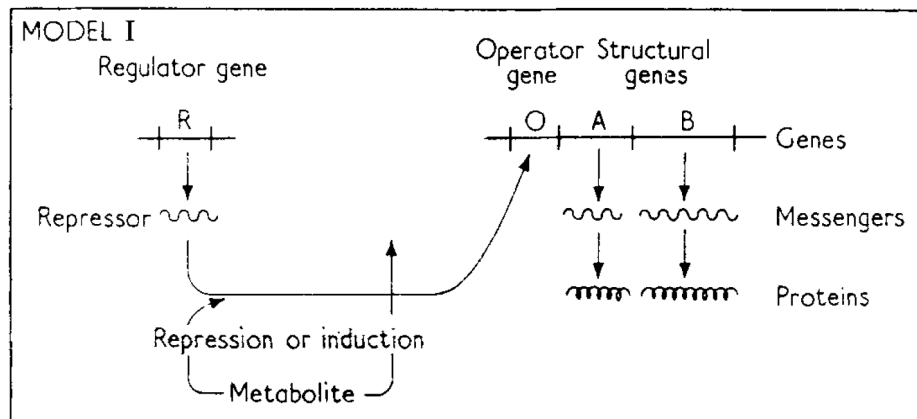


FIGURE. 1.3. Jacob & Monod's proposed model for the regulation of protein synthesis. Regulator gene *R* produces a repressor that acts at operator *O* to control expression of structural genes *A* and *B*. Specific metabolites activate or inactivate repression. Reproduced from [29].

Sidney Brenner, had also proposed the existence of an unstable form of RNA which acted as a "messenger" to transfer information from DNA-based genes to proteins.³⁰ This messenger RNA was already clearly depicted as the intermediate carrier of information between genes and proteins in Jacob & Monod's gene regulatory network. The exact code through which the sequence of mRNA translates to proteins was deciphered soon after using clever "forward" experiments: Marshall Nirenberg fed synthetic RNA homopolymer repeats (e.g. UUU_n, AAA_n, ...) to cell-free protein synthesis systems in which one of the 20 possible amino-acids were radioactively labeled. The detection of a radioactive protein product then revealed the identity of the amino-acid for which the input homopolymer coded.³¹ Har Gobind Khorana combined more complex RNA trinucleotide repeat polymers (e.g. ACA_n, UAG_n, ...) with chemical analysis of the resulting proteins to decipher the remaining possible 3-nucleotide codons, producing a finalised code cipher in 1966.³² These experiments together confirmed how genes encoded for proteins, and were awarded a joint Nobel prize for physiology for Nirenberg and Khorana in 1968.



W. Fiers

Precise extraction and purification of specific proteins or other macromolecules began to take center stage. For example, in 1966 Gilbert and Müller-Hill for the first time purified *lac* repressor protein fractions, confirming Jacob and Monod's theoretical model of proteins acting as repressors.³³ A year later, in a letter titled "*The Lac Operator is DNA*", the pair showed that the *lac* repressor specifically binds to DNA extracted from λ80 phage carrying the *lac* operon, and that this interaction could be disrupted by exposing the complex to IPTG.³⁴

While codons had been de-coded, a "reverse" experiment in which the complete sequence of a protein-coding gene could be determined had not been performed. Walter Fiers (Ghent University) was the first to do so in 1972 by reporting a complete sequence of the bacteriophage MS2 coat protein gene.³⁵ Fiers first digested the phage's ³²P-labeled RNA genome using ribonuclease T₁, which cuts RNA at G positions, and iteratively separated the resulting fragments based on size and charge through multiple rounds of two-dimensional fractionation. Joël Vandekerckhove and Marc

²⁹Jacob, J Mol Bio, 1961.

³⁰Brenner, Nature, 1961.

³¹Nirenberg, PNAS, 1961.

³²Khorana, CSH Symp Quant Biol, 1966.

³³Gilbert, PNAS, 1966.

³⁴Gilbert, PNAS, 1967.

³⁵Jou, Nature, 1972.

Van Montagu (both also Ghent University) had simultaneously been working on determining the amino-acid sequence of the MS2 coat protein.³⁶ Under the guidance of Van Montagu's preliminary amino-acid map and the Nirenberg-Khorana codon dictionary, Fiers and colleagues were able to puzzle together their fragments into a near-complete base sequence of the MS2 coat protein gene.

Finally, in 1973, Maxam & Gilbert sequenced the *lac* operator DNA, fully cementing Jacob & Monod's early model of gene regulation.³⁷ Restriction enzymes with known cut sites were used to digest a purified RNA copy of the operon, and the fragments were separated on a two-dimensional PAGE developed by Fiers and De Wachter.³⁸ Fragment lengths generated by different restriction enzymes could be used to infer the 27 bp-long sequence: 5'-TGG AATTGTGAGCGGATAACAATT-3'. Gilbert and Maxam remarked that the palindromic symmetry in this short motif could imply that the sequence is bound by a dimer or tetramer of *lac* repressor proteins. Later, they published an improved DNA sequencing method which relied on comparing the 4 electrophoresis ladders generated by chemically cleaving the DNA of interest at G, C, G+A or C+T sites.³⁹ The first exhaustive DNA sequencing method was thus a direct consequence of gene regulation research. Maxam and Gilbert were awarded the Nobel prize for chemistry for their work in 1980.

In 1977, Sanger and colleagues reported the complete genome sequence of bacteriophage ϕ X174 using the plus/minus sequencing method.⁴¹ Later that year, Sanger, Nicklen and Coulson introduced DNA sequencing with chain-terminating inhibitors.⁴⁰ Earlier work had shown that dideoxy variants of the four common nucleotides (A, C, G & T) would inhibit further polymerization when included into the newly synthesised DNA strand due to the absence of a reactive 3'-OH. Sanger devised that a polymerization reaction performed in a mixture of the four common nucleotides combined with a single dideoxy variant will produce a mixture of strands terminated at each position of the dideoxy variant. He further reasoned that performing this reaction for all four dideoxy variants in parallel should yield all the positions of each of the four bases (fig. 1.4). Initially, Maxam-Gilbert sequencing was more popular than Sanger's method because it could use purified DNA directly without cloning for template amplification, but the toxicity of hydrazine, dimethyl sulfate and radioactive ³²P ultimately prevented commercialisation and adoption. Sanger sequencing was commercialised throughout the 80s, first as a service, and later packaged into benchtop devices. Sanger was awarded a second Nobel prize for chemistry for this work in 1980.

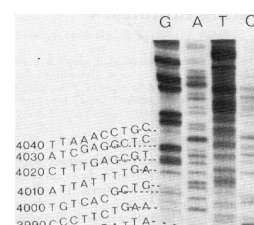


FIGURE. 1.4. Sanger sequencing. Reprod. from [40].

1.3 DNA-binding proteins

While the function of chromatin's DNA component as the primary store of information had become clear, the relationship between its protein content and nucleic acids was still under dispute. Alfred and Ellen Stedman investigated different histone fractions between different species and tissues, and proposed that "*the basic proteins of cell nuclei are gene inhibitors, each histone or protamine being capable of suppressing the activities of specific groups of genes*".⁴² This idea was confirmed experimentally when purified, protein-free DNA reconstituted with histone proteins was shown to be inactive in RNA synthesis, despite presence of RNA polymerase.⁴³ Further study of this inhibitory capacity found that "*relatively minor modifications of histone structure (acetylation), taking place on the intact protein molecule, offer a means of switching on or off RNA synthesis at different loci along the chromosome*".⁴⁴

By the early 1970s, biochemists had chemically identified five different histone proteins (H1, H2A, H2B, H3, H4), present at ratios of 0.5:1:1:1:1. Histones H2A and H2B were found to be rich in lysine, and H3 in arginine, positively charged amino acids known to bind DNA through electrostatic attraction.⁴⁵ X-ray diffraction research suggested a DNA "super-helix" structure stabilized by histone proteins,⁴⁶ while others found "thin fibrils"

³⁶Vandekerckhove, Arch Int Phys Biochim, 1973.

³⁷Gilbert, PNAS, 1973.

³⁸Wachter, An Biochem, 1972.

³⁹Maxam, PNAS, 1977.

⁴¹Sanger, Nature, 1977.

⁴²Stedman, PTRSL, 1951.

⁴³Huang, PNAS, 1962.

⁴⁴Allfrey, PNAS, 1964.

⁴⁵DeLange, Acc Chem Res, 1972.

⁴⁶Pardon, J Mol Biol, 1972.

and "nodular elements".⁴⁷ Hewish and Burgoyne found that an endogenous nuclear deoxyribonuclease cleaved chromatin DNA at regularly spaced sites, supporting a repeating chromatin substructure.⁴⁸ This hypothesis was strengthened by early electron microscopy, which directly visualized chromatin fibers as linear arrays of particles $\sim 70 \text{ \AA}$ in diameter resembling "beads on a string".⁴⁹

Finally, in 1974, Kornberg and Thomas showed that histone proteins H2A and H2B form dimers, and H3 and H4 formed stable tetramers.⁵⁰ When mixed with DNA, these dimers and tetramers returned X-ray diffraction spectra identical to native chromatin. Later research with enhanced resolution X-ray diffraction clearly showed a repeating pattern with a periodicity of $\sim 100 \text{ \AA}$.⁵¹ Kornberg remarked that "*the occurrence of multiples rather than just 200 base pair pieces is presumably due to some cleavage sites being blocked by nonhistone proteins*" - a prescient vision of transcription factors. Histone protein H1 was later identified as a stabilizer, binding to 20-80 bp long DNA links between histone octamers, leading to supercoiling and further chromatin compaction.⁵² This new "nucleosome" model of chromatin compacts DNA ~ 7 -fold at the nucleosome level, and further supercoiling explains how mitotic chromosomes achieve a 10 000-fold compaction. Later, Weintraub and Groudine showed that actively expressed genes are preferentially digested by DNase I.⁵³ Moreover, digestion with staphylococcal nuclease did *not* show preference for active genes, suggesting that active genes are still nucleosome-bound, although more loosely. Kornberg was awarded the Nobel prize for physiology for his work on chromatin structure in 1987.

By the late 70s, several pieces of the gene regulatory puzzle had been discovered:

1. Genetic information is stored in the base sequence of a cell's DNA.
2. This information is transferred to proteins according to a well-defined codon "dictionary" through messenger RNA.
3. Some proteins, such as the *lac* repressor, are capable of influencing transcription and thus the behaviour of the cell.

However, mechanistic and molecular insights in precisely *how* genes might be regulated molecularly were still sparse. Several breakthroughs came from the lab of Robert Roeder, who had previously purified the polymerase enzymes that catalyze the transcription of messenger RNA from DNA in 1969.⁵⁴ His lab showed that purified RNA polymerases could only initiate the transcription of DNA templates in the presence of uncharacterised cellular protein extracts. This finding confirmed earlier ideas that additional, then-unknown factors were necessary to initiate the transcription reaction.⁵⁵ Several such initiation factors were purified and characterised within the next year,⁵⁶ and it soon transpired that a large number of different (co-)factors are necessary for transcription to occur. Parallel research led to the discovery of a protein which induced expression of the 5S ribosomal RNA gene in *Xenopus* (clawed frog) egg cells. This factor, later named Transcription Factor III (*TFIIIA*), was the first protein shown to initiate transcription of a *specific* gene, rather than genome wide, in eukaryotes. Roeder suggested that *TFIIIA* might "*recognize a unique region of DNA*", a hypothesis which was explored and demonstrated soon after when *TFIIIA* was shown to bind precisely on position 45 to 96 of the 5S gene.⁵⁷ 5S rRNA, the gene enhanced by Roeder's original transcription factor, *TFIIIA*, is ubiquitously present in all cells across all known organisms. Transcription factors involved in more specialised, cell-type specific processes were identified and characterised soon after. Examples include the glucocorticoid receptor *Gr*,⁵⁸ *Sp1*,⁵⁹ and *Gal4*,⁶⁰ factors which are still under scrutiny to this day. It also became clear that some of these factors were extremely powerful: their ectopic expression could convert one cell type to another,⁶¹ and later specific combinations were found to even revert fully differentiated skin cells back to stem cells.⁶² The molecules responsible for Spemann & Mangold's organizer were hypothesized to be transcription factors, and several of them were identified molecularly shortly after.^{63,64}

In parallel, understanding of gene regulation was shifting from an exclusively promoter-proximal process (e.g. *TFIIIA*, TATA-box) to also include promoter-distal elements. Reporter studies by Banerji and colleagues

⁴⁷Slayter, Biochem, 1972.

⁴⁸Hewish, Biochem Biophys Res Comm, 1973.

⁴⁹Olins, Science, 1974.

⁵⁰Kornberg, Science, 1974.

⁵¹Kornberg, Science, 1974.

⁵²Finch, PNAS, 1976.

⁵³Weintraub, Science, 1976.

⁵⁴Roeder, Nature, 1969.

⁵⁵Ng, PNAS, 1979.

⁵⁶Matsui, J Biol Chem, 1980.

⁵⁷Engelke, Cell, 1980.

⁵⁸Payvar, PNAS, 1981.

⁵⁹Dynan, Cell, 1983.

⁶⁰Bram, PNAS, 1985.

⁶¹Davis, Cell, 1987.

⁶²Takahashi, Cell, 2006.

⁶³Cho, Cell, 1991.

⁶⁴Lamb, Science, 1993.

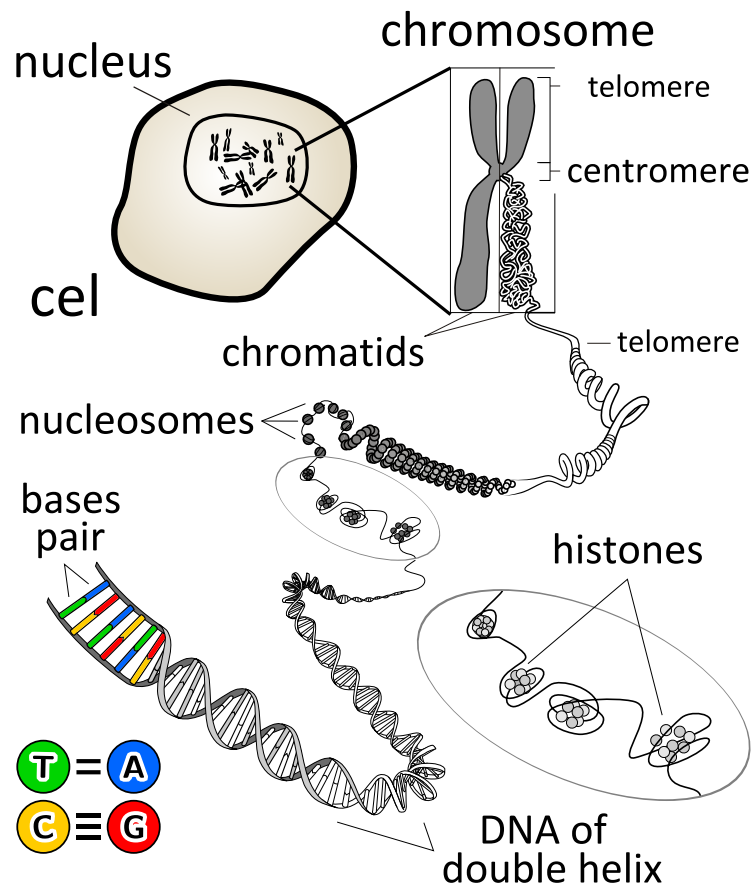


FIGURE 1.5. Levels of DNA compaction. Adapted from KES47, CC BY 3.0.

showed that a 72 bp tandem repeat sequence located 200 bp upstream of the SV40 early promoter, which was known to reduce nearby gene expression when mutated, could drive expression at large distances.⁶⁵ Constructs containing the 72 bp repeat, "for convenience referred to as the enhancer", showed a 200-fold increase in reporter (β -globin) expression, and removal of this specific sequence ablated the effect. Moreover, inversion of the region relative to the reporter gene conserved expression, and this effect remained when inserted at distances of up to 3.3 kb. Finally, co-transfection of separate plasmids - one with the SV40 72 bp enhancer, the other containing the β -globin reporter gene - did not show expression, confirming that the SV40 enhancer works in *cis* only. Further research on the immunoglobulin (*Ig*) enhancer, which specifically drives expression of *recombined* V(D)J genes only, showed that this enhancer was active only in B-lymphocyte myeloma cells, not in HeLa, 3T6, or mink lung cells - cell-type specific enhancer activity!⁶⁶ The group also correctly proposed that the cell-type specific nature of the *Ig* enhancer owes its effect to cell-type dependent expression of its binding transcription factors, and that chromosomal translocations could lead to oncogenic expression such as observed in Burkitt's lymphoma.

Together with the new discoveries of genomic *cis*-regulatory elements, an explosion in discovery of the physics behind protein-DNA-binding mechanisms occurred. Crystallography of λ phage (*cI*), *Cro* and *E. coli cap*, three known DNA-binding proteins, showed the first generalized mechanism of binding: two α -helices joined by a tight turn (helix-turn-helix motif) generate a sharp nick that fits into the major groove of DNA, leading to sequence-specific - though degenerate - binding.⁶⁷ Simultaneous studies in *D. melanogaster* led to the discovery of the homeodomain, a 60 amino acid DNA-binding domain highly conserved across metazoans, especially those

⁶⁵Banerji, *Cell*, 1981.

⁶⁶Banerji, *Cell*, 1983.

⁶⁷Pabo, *Ann Rev Biophys*, 1984.

with segmented body plans.⁶⁸ Later, it was shown that the homeodomain also uses a helix-turn-helix structure, but is significantly longer and complexes DNA differently than the λ HTH.⁶⁹ Zinc-binding repeating Cys2His2 motifs were further identified in *TFIIIA*, and these zinc-protein complexes were also shown to bind specific DNA motifs.^{70,71} Next, an amino acid motif conserved across transcription factors *CEBP*, *Myc*, *Fos*, *Jun* and the yeast transcription factor *GCN4* were shown to interdigitate through leucine residues every seventh position ("basic leucine zipper"). This interdigitation allows these factors to dimerize, and bind palindromic DNA motifs through adjacent DNA-recognizing residues.⁷² Finally, basic helix-loop-helix motifs were shown to similarly dimerize, allowing both monomers and hetero/homodimers to recognize and bind the palindromic E-box motif (consensus CANNTG, but ideally CACGTG).^{73,74} Structural studies also confirmed that the *lac* repressor is a tetramer protein capable of binding two operator sites simultaneously, forming DNA loops that repress expression.⁷⁵

1.4 Sequencing the human genome

While mechanistic insights into gene regulation began to form in the 1980s, precise knowledge of gene sequences in human was still sparse. Cystic fibrosis, the most common lethal autosomal recessive disease among Caucasians, was *suspected* to be monogenic based on its Mendelian inheritance pattern, but the precise gene responsible for the disease had not been identified. By 1985, the location of the gene was narrowed down to a region "1 percent of the human genome, about 30 million base pairs" in size.⁷⁶ This result was achieved using the linkage analysis methods previously invented by Morgan & Sturtevant, combined with restriction fragment length polymorphisms (RFLPs),⁷⁷ requiring thousands of clone library Southern blots. Sequencing the *CFTR* gene by Sanger's method took four more years,^{78,79} revealing the deletion of three nucleotides ($\delta F508$) was the disease's most common causal mutation. This marked the first time a disease's causal gene was identified by bottom-up genomics before its protein was known. HTT, the gene responsible for Huntington's disease, was identified in a similar way and timeframe.^{80,81}

The decade-long efforts required to identify CFTR and HTT led to widespread consensus in the academic community: a complete reference sequence of the human genome was necessary. With a human genome available, researchers would be able to immediately compare any extracted DNA to a healthy reference and locate disease polymorphisms rapidly. Moreover, it had become canon that genomes contained information beyond protein-coding genes, and that understanding the relationship between these different genomic elements and disease would lead to transformative medicine. Francis Collins, who had been heavily involved in the CFTR gene hunt, became a leading advocate for the Human Genome Project,⁸² which was officially kicked off in 1990. Originally a joint academic effort between the US National Institutes of Health, several international collaborators (United Kingdom, Germany, Japan, France) soon followed.⁸³ Interestingly, the US Department of Energy also played a major role, with Lawrence Livermore, Los Alamos and Lawrence Berkeley labs providing extensive support in computational analysis and data storage.

The project was first expected to take 15 years and cost 3 billion dollars. Early efforts focused on a hierarchical clone-by-clone approach: the genome was broken into fragments (~200 kb) and cloned into over 100 000 bacterial artificial chromosomes (BACs). These BACs were then linked into large contigs using a variety of techniques (mostly sequence tagging sites PCR, which was somewhat novel at the time,^{84,85} and restriction enzyme fingerprinting), which were in turn mapped to entire chromosomes. A minimum tile path map, dividing the genome into regions that were distinct enough to sequence by Sanger, was made. Individual BAC clones were then further fragmented into smaller ~1000 bp pieces which could be sequenced individually. The first several years of the project were spent on the minimum-tiling BAC generation and archival. Repetitive DNA regions and coverage gaps complicated the assembly of contigs, and Sanger's dideoxy method was incredibly labour

⁶⁸McGinnis, *Cell*, 1984.

⁶⁹Kissinger, *Cell*, 1990.

⁷⁰Miller, *EMBO J*, 1985.

⁷¹Pavletich, *Science*, 1991.

⁷²Landschulz, *Science*, 1988.

⁷³Murre, *Cell*, 1989.

⁷⁴Ellenberger, *Gen Dev*, 1994.

⁷⁵Lewis, *Science*, 1996.

⁷⁶Tsui, *Science*, 1985.

⁷⁷Botstein, *AJHG*, 1980.

⁷⁸Riordan, *Science*, 1989.

⁷⁹Rommens, *Science*, 1989.

⁸⁰Gusella, *Nature*, 1983.

⁸¹MacDonald, *Cell*, 1993.

⁸²Collins, *NEJM*, 1999.

⁸³Birney, *HMG*, 2021.

⁸⁴Saiki, *Science*, 1985.

⁸⁵Mullis, *CSHS*, 1986.

intensive. Luckily, automation soon followed. One company played a particularly important early role: ABI. First by releasing the ABI Prism 373 and 377, then the ABI Prism 310 capillary gel electrophoresis machine, and finally the ABI 3700 96-plex capillary gel electrophoresis analyzer, which became the workhorse of the human genome project. Ensuing throughput advances were staggering: whereas automated slab-gel instruments produced 1000 bp/day in 1985, the ABI 3700 capillary array reached 1000 bp/s by 1998, and in 1999, HGP centers were sequencing ~90 Gbp per month.

Directly shearing the genome into sequencing-ready fragments, the "shotgun" strategy, had previously been used to manually map the kilobase-sized genomes of several bacteriophages. This method was deemed unsuitable for larger genomes due to the sheer number of reads required, an amount of data that would not have been possible to align on early 90s computing infrastructure. However, in 1995, Craig Venter's group at the Institute for Genomic Research managed to sequence the 1.8 MB *H. influenzae* genome using the shotgun approach.⁸⁶ By cloning randomly sheared chromosomal fragments directly, and sequencing the resulting plasmids, the virus's entire genome was sequenced in just 14 months. 19 687 clones generated 24 304 reads (6.3× coverage), taking 30 hours to assemble on the 200 MHz SPARC supercomputer. The assembled genome revealed ~1750 genes, a third of which lacked any known function. Improvements in computing infrastructure and sequencing hardware had made the shotgun approach viable for larger genomes, and in 1998, Venter launched Celera Genomics in partnership with Perkin-Elmer/ABI. The group announced a plan for whole-genome shotgun sequencing of the 3.2 Gb human genome using 300 ABI 3700 capillary sequencers and a 500-processor Compaq Alpha compute cluster.⁸⁷ In parallel, Celera further refined their strategy, completing the *D. melanogaster* genome in 2000 - although heavily leaning on a partial (29 Mb of the 120 Mb) public reference genome published by international consortia.⁸⁸ By this time, the HGP consortium also began applying shotgun sequencing to its continuing effort. Both the HGP and Celera assemblies of the human genome - each approximately 90% complete - were released in 2001.^{89,90} Alignment took several months of compute. Of course, the Human Genome Project had been releasing intermittent drafts which Celera could freely download from GenBank. Celera's final assembly was composed of 27M shotgun reads plus 18M reads sourced from shredded public HGP data, and was accessible only under license. By 2005, improvements in the freely available public human genome sequence had reduced the commercial appeal of Celera's sequence database.⁹¹ Customer demand declined, and it was placed in the public domain in 2005.⁹¹

The human genome race cemented shotgun WGS as the default strategy for large genomes, and began an era of comparative genomics and high-throughput "omics" driven by rapidly falling sequencing costs. Its economic return was later estimated at nearly \$800 billion on a \$3 billion public investment, largely through the birth of sequencing-based diagnostics, bioinformatics, and high-throughput sequencing assays.⁹² The now completed assembly allowed researchers to shift focus from recording the genome to annotating it, and finally deciphering its logic. The new fields of transcriptomics, epigenomics, metagenomics, population genetics, ... would require sequencing not just one genome, but thousands of genomes and other nucleic acid-derived samples across many species, individuals and tissues. This demand could not be met by the Sanger sequencing method, stimulating the development of a variety of new sequencing technologies.

Several NGS platforms were launched almost simultaneously: Roche/454's pyrosequencing,⁹³ ABI's SOLiD, Helicos's single-molecule chemistry, and Solexa's sequencing-by-synthesis (SBS)⁹⁴ each promised sequencing at a fraction of the Sanger method's cost. The competitive landscape narrowed after 2007, when Illumina acquired Solexa and began to scale SBS to industrial capacity. Massively parallel sequencing by synthesis enabled large numbers of DNA templates to be sequenced simultaneously.⁹⁵ By 2012, benchtop instruments such as the MiSeq delivered data for ~\$0.10 per megabase,⁹⁶ compared with ~\$1000 per megabase in 2005.⁹⁷ Competing short-read platforms faded, and long-read entrants (PacBio SMRT, Oxford Nanopore) would not appear for another decade (fig. 1.6). A second wave of SBS providers (BGI/MGI, Ultima Genomics, Complete Genomics, Singular Genomics) has recently appeared, following expiry of key Illumina patents.

In Illumina sequencing by synthesis, adapter-bearing DNA fragments are captured on a flow-cell surface

⁸⁶Fleischmann, *Science*, 1995.

⁸⁷Venter, *Science*, 1998.

⁸⁸Adams, *Science*, 2000.

⁸⁹Lander, *Nature*, 2001.

⁹⁰Venter, *Science*, 2001.

⁹¹Kling, *EMBO Reports*, 2005.

⁹²Tripp, 2011.

⁹³Margulies, *Nature*, 2005.

⁹⁴Bennett, *Pharmacogen*, 2005.

⁹⁵Bentley, *Nature*, 2008.

⁹⁶Shendure, *Nat Biotech*, 2012.

⁹⁷Shendure, *Science*, 2005.

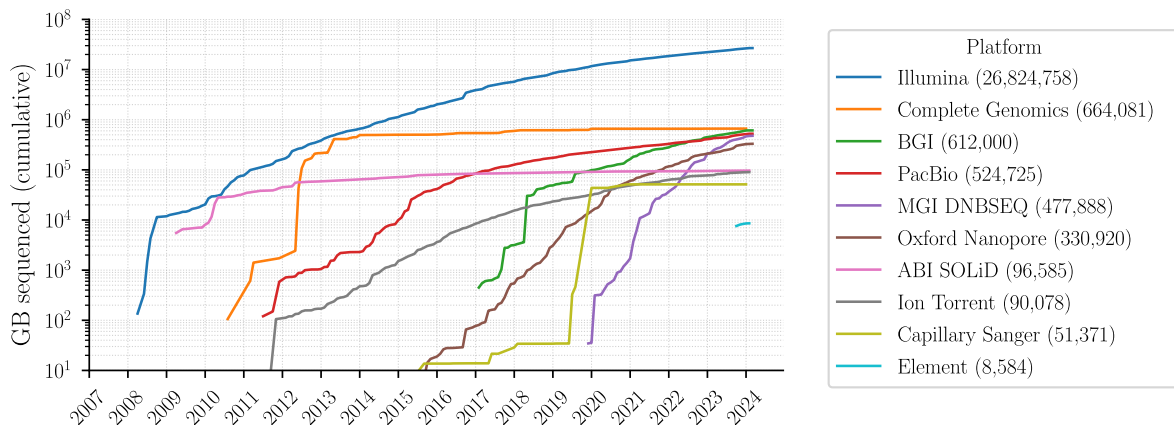


FIGURE 1.6. Trends in sequencing platforms: total number of gigabases submitted to GEO per platform.

and amplified into clusters, which are sequenced using fluorescent reversible-terminator nucleotides.⁹⁵ Each incorporation cycle is imaged, generating imaging data of the oligo lawn. Image processing is then used to determine the position of oligo clusters present on the lawn, and these are stored in `.locs` (locations) files. The basecalling algorithm will combine these `.locs` files together with the imaging data to extract fluorescence intensities for all colours used per cluster. Dominant fluorescent signals then determine the base called for each cluster for every cycle. When no base can be called, N is used. When a strand does not incorporate a base during a given cycle, or if it leads ahead by one base, it becomes out of phase with the other strands of that cluster for the rest of the sequencing run. This phasing/pre-phasing is responsible for signal deterioration further down the read, and is one of the limiting factors to maximum read length (originally ~30, now up to 300 bp). Theoretical phasing rates (approximately 0.1% of strands in a given cluster per cycle) can be used to estimate the contribution of phasing noise and deconvolve true signal from phased noise. Each basecall is assigned a "Phred" quality score Q based on signal-to-noise ratio, phasing, and separation of dense clusters: $Q = -10 \cdot \log_{10}(P)$ where P is the probability of the basecall being correct. A Q score of 30 corresponds to an erroneous basecall probability of 1 in 1000. The result is a binary `.bcl` file, which contains the basecalls for each cluster, and can then be converted to the ASCII `.fastq` format using `bc12fastq`.

```
@NS2000:001:HKYYLDSXX:1:1101:10000:1000 2:N:0:ATCACG
CGTACGATCGTAGCTAGCTGATCGTAGGTTGCTAGCTAGGA
+
HHHHGGGFFFEEDDDCCCB BBBBAAAAAAAAA@@@@@@@
@NS2000:001:HKYYLDSXX:1:1101:10001:1000 2:N:0:ATCACG
ACGTACGATCGTAGCTAGCTGATCGTAGGTTGCTAGCTAGA
+
HHHHGGGFFFEEDDDCCCB BBBBAAAAAAAAA@@@@@@@
```

FIGURE 1.7. FASTQ snippet.

Each sequenced strand is represented by four lines in the resulting `.fastq` file (fig. 1.7). The first line is a header, which stores metadata such as sequencer and cluster coordinates. The second line is the sequence itself, followed by a separator (+) on the third line, and the ASCII-encoded quality score on the fourth line.

1.5 Quantitative sequencing assays

Before the availability of next-generation sequencing technology, studying the genome and gene expression at scale relied heavily on microarrays. Microarrays were, and still are, used to quantify RNA transcripts,⁹⁸ genotype samples using SNP arrays,⁹⁹ and even measure DNA-protein interactions using ChIP-on-chip.¹⁰⁰ While high-density microarrays allow millions of probes to be interrogated in a single experiment, their limited dynamic range, high false-positive hybridisation and heterogeneous probe hybridisation capacities skew absolute quantification. The probe set also requires intelligent design and thus cannot discover novel sequences. Probeset size is particularly troubling for ChIP-on-Chip, since genome binning limits resolution to around 500-1000 bp - too low for novel TF motif discovery. Nonetheless, large scale, distributed application of microarray technologies culminated in a first draft of the Encyclopedia of DNA Elements (ENCODE) in 2007,¹⁰¹ a careful annotation of 30 MBs - around 1% - of the human genome.

The now upcoming availability of NGS made it possible to bypass microarrays: specific fractions of nucleic acids extracted from tissues could be sequenced, aligned to the reference genome, and counted. This revolution led to an explosion of quantitative sequencing assays being developed. One of the first such assays was histone methylation ChIP-seq, first published in 2007.¹⁰² By sequencing immunoprecipitated DNA directly rather than using microarrays, methylation could now be assessed genome-wide and with fewer quantitative biases. Performing this technique for a spectrum of 20 different lys/arg histone methylation ChIP-seq variants on CD4⁺ T cells, and correlating the annotations with GNF SymAtlas gene expression microarray data allowed genome-wide classification and discovery of various classes of functional elements. The distribution of methylation signatures could now be systematically correlated to phenotypes, helping understand disease and development. For example, a comparison of histone methylation in mouse embryonic stem cells and embryonic fibroblasts revealed transitory methylation changes in the de-differentiation of ES.¹⁰³

In parallel research, ChIP-seq was also being applied to individual transcription factors at a genomic resolution of around 50 base pairs, a welcome improvement compared to microarrays. Application of ChIP-seq on neuron-restrictive silencer factor (NRSF) in Jurkat T-cells showed that the factor can also bind "weakly" to altered, non-canonical motifs dependent on local chromatin state (methylation, accessibility) and co-factor motif context.¹⁰⁴ These findings contradicted earlier ideas of stringent binding to only canonical motifs, a rejection made possible only through the broadness of NGS. NGS also improved dynamic range compared to microarrays, showing for example that STAT1 motif enrichment was strongly correlated to absolute ChIP-seq peak height.¹⁰⁵

The increased genomic resolution of ChIP-seq was rapidly refining models of gene regulation, but the technique still required a strong antibody and prior knowledge of the protein or histone modification of interest. In contrast, DNase I, which had been used to digest chromatin since the late 70s,^{53,106} generates genome-wide fragments of regions not protected by histones in condensed chromatin. The main advantage of this approach over ChIP-based methods was that it did not require prior knowledge of the transcription factor or histone modification of interest: chromatin changes could be surveyed genome-wide. While fragments were originally assessed by Southern blot,¹⁰⁶ then capillary Sanger sequencing,¹⁰⁷ and later microarrays,¹⁰⁸ DNase hypersensitivity assays were soon translated to NGS,^{109,110} rapidly contributing to even higher resolution (1 base pair) understanding of gene regulation. For example, matched DNase-seq and glucocorticoid receptor (GR) ChIP-seq revealed that out of 8236 active GR binding sites in murine mammary adenocarcinoma, 95% were already accessible *prior* to receptor activation.¹¹¹ These findings rejected that GR might act as a pioneer factor, remodeling inaccessible chromatin and exposing its preferred motif after receptor activation. Moreover, pituitary gland cells shared only 11.4% of the binding sites accessible in adenocarcinoma, suggesting that differential chromatin context underlies differential responses to the same transcription factor. Unsurprisingly, ENCODE's second draft included DNase-seq data from 125 tissues, along with bisulfite sequencing, and H3K4me3 and TF ChIP-seq.^{112,113} In to-

⁹⁸Schena, *Science*, 1995.

⁹⁹Wang, *Science*, 1998.

¹⁰⁰Ren, *Science*, 2000.

¹⁰¹ENCODE, *Nature*, 2007.

¹⁰²Barski, *Cell*, 2007.

¹⁰³Mikkelsen, *Nature*, 2007.

¹⁰⁴Johnson, *Science*, 2007.

¹⁰⁵Robertson, *Nat Meth*, 2007.

¹⁰⁶Wu, *Nature*, 1980.

¹⁰⁷Crawford, *PNAS*, 2004.

¹⁰⁸Crawford, *Nat Meth*, 2006.

¹⁰⁹Boyle, *Cell*, 2008.

¹¹⁰Hesselberth, *Nat Meth*, 2009.

¹¹¹John, *Nat Genet*, 2011.

¹¹²Thurman, *Nature*, 2012.

¹¹³ENCODE, *Nature*, 2012.

tal, 2.9 million DNase I hypersensitivity sites were identified and annotated genome-wide, covering almost all previously known regulatory element candidates: 94.4% of previously known TF ChIP-seq regions were found to overlap with these DNase I sites, and only 3692 such sites were found in all cell types assessed, confirming that chromatin accessibility and downstream gene regulation are highly tissue and cell-type specific processes. Combination of these new regulatory readouts with GWAS further confirmed that many disease-related SNPs are located in regulatory regions, accelerating therapeutic target discovery.

By 2008, development of NGS began to displace microarrays in transcriptomics as well, first in yeast,¹¹⁴ and later in mouse,¹¹⁵ and human.¹¹⁶ Poly(A)⁺ mRNA was extracted and purified from brain, liver, and skeletal muscle from C57BL/6 mice, hydrolized and reverse-transcribed to cDNA using random hexamer priming.¹¹⁵ Illumina/Solexa was used to sequence the resulting libraries. This particular study, performed by Mortazavi et al., delivered the first quantitative whole transcriptome in mouse:

- ~145 000 distinct splicing events detected, with ~3500 genes showing alternative splicing.
- ~3000 newly identified or extended 5' exons.
- ~3000 novel extended 3' UTRs.
- ~17 000 newly annotated transcribed regions.
- 596 entirely novel candidate transcripts, including a previously unknown mir-124-1 precursor in the brain.

In addition, quantification of phage λ RNA spike-ins remained linear ($R^2 = 0.99$) over 5 orders of magnitude, and correlation between technical replicates was high ($R^2 = 0.96$). A technical comparison of RNA-seq with microarray transcriptomics showed that RNA-seq was technically superior in every aspect: RNA-seq identified 30% more differentially expressed genes than microarrays at the same false discovery rate, and the sequencing approach was better at detecting low-abundance genes.¹¹⁷ qPCR validation also confirmed more RNA-seq-specific findings than microarray-specific findings. Finally, ~200 000 RNA-seq reads mapped to possible exon-exon junctions, revealing novel splicing patterns. Each RNA extract was sequenced 7 times, unsurprisingly with a Spearman correlation of 0.96 across lanes, but extraction itself was not replicated.¹¹⁸ Several major findings emerged from RNA-seq. Eukaryotic transcriptomes turned out to be more heterogeneous than anticipated. Individual papers revealed large numbers of novel transcribed regions, widespread antisense transcription, and extensive variation in 3' ends.¹¹⁶

Studies combining different sequencing assays also rapidly led to new insights. Application of RNA-seq, FAIRE-seq and H3K27ac/H3K27me3 ChIP-seq on melanoma directly led to understanding the gene regulatory programs underlying proliferative and invasive states.¹¹⁹ Differential gene analysis between patient biopsies from these two melanoma states showed high MITF, SOX10, and PAX3 in proliferative, and High ZEB1, SOX9 in invasive states. Similarly, FAIRE-seq showed open chromatin regions near MITF and SOX10 target genes in proliferative, and SOX9 targets in invasive states, and ChIP-seq showed high H3K27 acetylation at AP-1/TEAD target enhancers in invasive cells. Repression of TEAD in invasive cells via siRNA showed decreased cell invasion in matrigel assays and reduction in SOX9 expression, confirming TEAD as a master regulator in invasive melanoma.

By 2010, next-generation sequencing itself had become ubiquitous, but library prep still required sonication, end-repair, and adapter ligation. A breakthrough on this front came from Tn5, an enzyme first studied decades earlier. Barbara McClintock had first observed "mutable loci" in maize arise from mobile DNA inserts,^{120,121} the first observation of position-effect variegation and cis-regulatory chromatin influence outside of *Drosophila*.¹⁸ Two decades later, foreign DNA insertions were observed to disrupt the *E. coli gal* operon.¹²² Entire resistance plasmids were found to integrate into phage λ , flanked by 1.5 kb inverted repeats.¹²³ The transposable element was finally named Tn5, a member of an entire class of transposons.¹²⁴ By 1982, both inverted repeats, named IS50-L and IS50-R, were sequenced.^{125,126} IS50-R encodes the 461 amino-acid primary transposase as well as a truncated version of the protein. This smaller protein is expressed from a second, downstream transcription start site and inhibits the complete transposase by competitive binding, repressing total transposition activity. A conserved stop

¹¹⁴Nagalakshmi, *Science*, 2008.

¹¹⁵Mortazavi, *Nat Meth*, 2008.

¹¹⁶Wang, *Nature*, 2008.

¹¹⁷Marioni, *Gen Res*, 2008.

¹¹⁸Wang, *Nat Rev Gen*, 2009.

¹¹⁹Verfaillie, *Nat Comm*, 2015.

¹²⁰McClintock, *PNAS*, 1950.

¹²¹Ravindran, *PNAS*, 2012.

¹²²Shapiro, *J Mol Bio*, 1969.

¹²³Berg, *PNAS*, 1975.

¹²⁴Bukhari, *Insertion ... Episomes*, 1977.

¹²⁵Reznikoff, 1981.

¹²⁶Reznikoff, *Cell*, 1982.

codon point mutation found in the nearly identical IS50-L repeat disrupts complete expression of the transposase there, but simultaneously introduces a promotor sequence for the embedded antibiotic resistance genes. The Tn5 transposon is thus capable of transferring antibiotic resistance genes and expressing them independently of the host's endogenous mechanisms. It also self-regulates both its transposon and resistance activities. Soon, Tn5's potential as a molecular tool was recognised. For example, Ruvkun and colleagues used Tn5 insertion mutagenesis to investigate symbiotic nitrogen-fixation genes in *Rhizobium meliloti*.¹²⁷ A first full crystallography of Tn5 protein (Tnp) in complex with DNA showed that it shares mechanistic features with HIV-1 integrase and RAG1/2 V(D)J recombinases.¹²⁸

Owing to its use as a lab tool, Tn5 protein mutants and their molecular mechanisms were quickly studied.^{131,132} A triple mutant, E54K (enhances OE binding) - M56A (disrupts inhibitor production) - L372P (increases activity by altering dimerization), was engineered to increase catalytic activity, and is the major variant in use to date ("hyperactive Tn5").¹³³ The transposase can bind to two sites on the ends of each IS50 sequence: the inside or outside end (IE, OE). A synthetic "mosaic end" (ME), engineered by combining elements of the IE and the OE showed enhanced binding capacity, improving dimerization and complexation of Tnp with DNA.^{134,135} The transposase's exact catalytic mechanism was unknown until Bhasin and colleagues showed how the Tn5 transposase nicks at the transposon ends, forms a hairpin, releases the donor dna, and opens the hairpin and integrates at the acceptor site.¹³⁶ A staggered cut is made at the acceptor site, and after insertion of the blunt-end transposon, the missing 9 base pairs are repaired, leading to a duplication. A classic DDE motif, known from RNaseH enzymes, in Tn5 transposase's active site coordinates magnesium ions critical for catalysis by stabilizing the DNA-protein complexes.^{129,137}

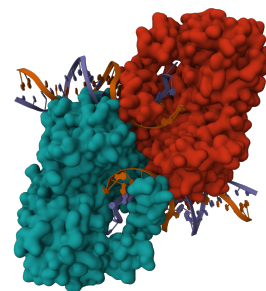


FIGURE. 1.9. Tn5 complexed with DNA.¹²⁹ Generated with mol*.¹³⁰

By the mid 2000s, Tn5 saw routine lab use to tag cloned DNA with selection markers, primers, or even regulatory elements.¹³⁸ This could be achieved without in vivo expression of the transposon by loading transposase protein onto pre-cleaved DNA, and introducing the complex into electroporated bacteria. The transposase was first used by Adey and Shendure to construct shotgun sequencing libraries in 2010.¹³⁹ This change greatly simplified the old workflow of fragmentation, end repair, and ligation, a process that spanned 6 to 10 hours. With Tn5 transposase "tagmentation", these pre-PCR steps were reduced to only 15 to 20 minutes. Moreover, the Adey protocol could use inputs as low as 10 pg as opposed to the 1 µg required by the ligation methods, a sensitivity increase that would later allow library extraction from single cells.

Buenrostro and colleagues realized around 2013 that a hyper-active Tn5 dimer pre-loaded with Illumina adapters could be used not only to construct sequencing libraries, but also to interrogate the accessibility of native chromatin similarly to DNase-seq.^{140,141} Transposition was performed on intact, unfixed nuclei, with Tn5 dimers sterically hindered by occupant nucleosomes and other DNA-bound proteins. Every insertion event marks an open stretch of chromatin, and, since Tn5 already ligates sequencing adapters, subsequent PCR generates a sequencing-ready library. The resulting protocol, named ATAC-seq (assay for transposase-accessible chromatin by sequencing), required just 500 to 50 000 cells - compared to the 1 to 100 million cells needed for DNase-seq, FAIRE-seq and ChIP-seq. The workflow itself also eliminated multi-day, loss-prone workflows such as adapter ligation and gel purification, and produced results within hours. Signal-to-noise ratio and sensitivity matched DNase-seq despite orders-of-magnitude lower input, with strong reproducibility ($R \sim 0.98$) and high correlation to DNase-seq peaks ($R \sim 0.8$). Unlike FAIRE-seq and DNase-seq, which assay bulk chromatin, ATAC-seq simultaneously revealed open chromatin, nucleosome positions, and transcription-factor footprints from minimal material. Omni-ATAC, an improved protocol, introduced the triple-detergent lysis buffer (NP-40 + Tween-20 + digitonin), and optimized magnesium concentrations, dramatically reducing mitochondrial contamination and

¹²⁷Ruvkun, *Cell*, 1982.

¹²⁸Davies, *Science*, 2000.

¹³¹Wiegand, *J Bact*, 1992.

¹³²Weinreich, *Gen Dev*, 1994.

¹³³Goryshin, *J Biol Chem*, 1998.

¹³⁴Zhou, *J Mol Biol*, 1998.

¹³⁵Steiniger-White, *J Mol Biol*, 2002.

¹³⁶Bhasin, *J Biol Chem*, 1999.

¹³⁷Reznikoff, *Mol Microbiol*, 2003.

¹³⁸Reznikoff, *Meth in Mol Biol*, 2004.

¹³⁹Adey, *Genome Biol*, 2010.

¹⁴⁰Buenrostro, *Nat Meth*, 2013.

¹⁴¹Buenrostro, *Curr Prot in Mol Biol*, 2015.

extending the assay to frozen tissue sections.¹⁴² These optimisations resulted in a 13-fold drop in mitochondrial reads and 3-fold rise in fraction of reads in peaks, and up to 15-fold increased number of unique fragments at similar sequencing depths.

1.6 Single-cell RNA and ATAC-seq protocols

The development of next-generation sequencing methods and quantitative -omics assays in the early 2000s - most notably RNA-seq, ChIP-seq, DNase-seq, and later ATAC-seq - allowed researchers to probe gene expression and regulatory elements in high-throughput, at genome-wide scales (fig. 1.10). Application of these technologies revealed tissue- and cell type-specific enhancer landscapes, complex transcription factor networks, novel transcript isoforms and splicing patterns, and altered chromatin states across disease and development. Together, these bulk methods laid conceptual frameworks for understanding the genome, and significantly advanced our understanding of development, immunity, and cancer.

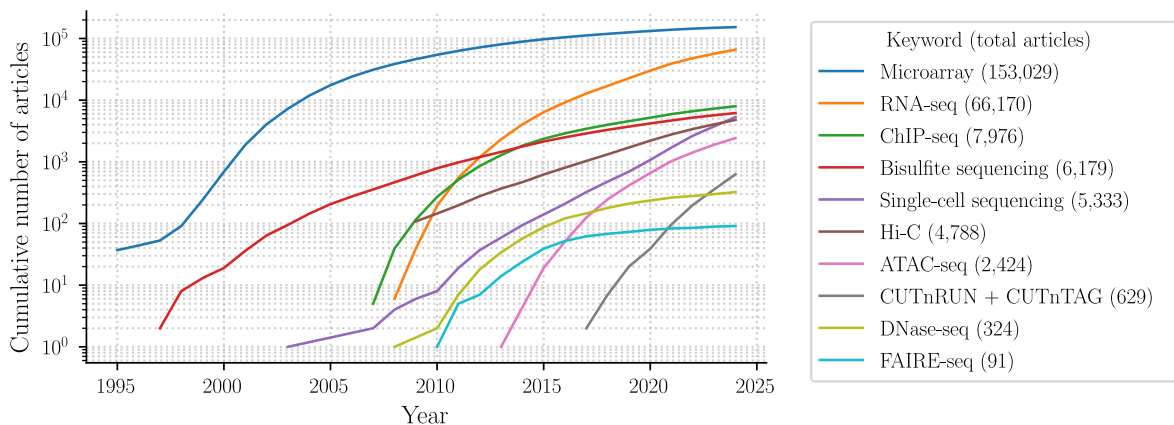


FIGURE 1.10. Trends in sequencing assays: number of PubMed hits to technique names

However, bulk assays cannot distinguish signals from different populations within the input sample. Unexpected heterogeneity, rare cell states, and stochastic dynamics are masked by the aggregation of thousands to millions of cells. The combination of FAC sorting and quantitative bulk sequencing methods can help, but is costly, laborious, and requires prior knowledge. Even in combination with cell-sorting, bulk assays remained unsuitable for a number of key open problems:¹⁴³

- Profiling of rare or transient cell-types, such as circulating tumor cells and metastatic cancer subclones.
- Profiling the full complexity of systems composed of potentially thousands of cell types, such as brains and the repertoire of immune cells.
- Reconstructing cell lineages across development, differentiation and disease progression in high temporal and potentially spatial resolution.
- Systematically defining cell types and identifying consensus markers, both in expression, but also in active genomic elements and modifications.
- Quantifying stochastic differences between cells sampled from the same type.
- High-throughput genetic or chemical perturbation studies.

Some of these considerations drove early exploration of gene expression at the single-cell level in the 90s, starting a decade-long quest for methods to amplify RNA material from individual cells. Eberwine and colleagues

¹⁴²Corces, Nat Meth, 2017.

¹⁴³Shapiro, Nat Rev Gen, 2013.

first injected an oligo-dT T7 primer and T7 reverse transcriptase directly into single rat hippocampal neurons to reverse transcribe their mRNA in situ.¹⁴⁴ The resulting cDNA was amplified using in vitro transcription, and relative expression levels of 10 known neuronal genes (mostly ion channels) were quantified by Southern blot. Even among *morphologically similar neurons*, significant differences in gene expression were observed, recording for the first time transcriptomic heterogeneity within the same tissue. About a decade later, a double T7-based amplification method was applied to profile 12 laser microdissected rat hippocampal neurons on cDNA microarrays, revealing two new marker genes for a novel pyramidal neuron subtype.¹⁴⁵ Soon, PCR-based cDNA amplification strategies,^{146,147} rather than IVT, was applied to single cells from mouse embryonic stem cells and blastomeres.¹⁴⁸ Mature mRNA was extracted by poly(dT) primers, reverse-transcribed and PCR-amplified, and the resulting single-cell libraries were each hybridised to *individual* RNA microarrays. This revealed that the blastocyst's inner cell mass contains at least two populations: some express genes typical of primitive endoderm (*Gata4*, *Gata6*) and others express epiblast markers (*Fgf4*, *Nanog*, *Klf2*).

"True" whole transcriptome single-cell sequencing was first demonstrated in 2009 by Tang and Surani at the Gurdon Institute.¹⁴⁹ By increasing reverse transcription and PCR extension times/cycles, the group was able to amplify enough material for sequencing cDNA libraries from 6 individual mouse blastomeres. This sequencing-based approach provided a number of immediate advantages over microarray analysis:

1. Comparison of single-blastomere mRNA-Seq data with bulk (320 pooled cells) Affymetrix microarray data showed that 94.3% of microarray-detected genes had at least 5 transcripts in the mRNA-Seq data, but mRNA-Seq detected 5270 (up to 75%) more genes in total.
2. Sensitivity and dynamic range (0.2 RPKM to 12 000 RPKM) were decidedly higher for mRNA-Seq than for analog microarrays.
3. Since near-whole transcripts were sequenced, novel splice junctions could be determined. The group discovered 1753 novel isoforms, demonstrated for the first time that different isoforms of the same gene can be expressed in the same single cell.

Tang et al. used the ABI SOLiD sequencer, which had become commercially available only two years prior. Since the sample extraction and amplification protocol itself were only slightly different from existing microarray protocols, it can be posed that development of the Tang method was an effect of sequencing hardware improvements. However, the new sequencing-based approach also came with a new set of challenges. Cost of the novel sequencing technology was very high - each cell in the Tang paper, sequenced at around 24 to 85 million reads, must have cost around \$10 000 to \$50 000 in SOLiD sequencing alone *per cell*. At the time, one Affymetrix microarray sample cost in the order of \$800 - significantly cheaper. Despite longer extension times, the protocol was also unable to capture transcripts longer than 3 kb, and was limited to poly(A)⁺ mRNA. These factors contributed to mRNA-Seq's status as a proof-of-concept technique rather than an immediate research tool.

One of the first protocols to be published after the Tang protocol was STRT-seq (Single-cell Tagged Reverse Transcription), developed by the Linnarsson lab at Karolinska Institute.^{150,151} The protocol applied template-switching reverse transcription (use of MMLV reverse transcriptase, which tags 5' transcripts via template switching),¹⁵² to cells picked into individual wells, followed by pooled single-primer PCR for cDNA amplification. A total of 92 mouse cells were profiled: 48 embryonic stem cells and 44 embryonic feeder cells. Libraries were sequenced on the Illumina Genome Analyzer Iix, for a total of 110M raw reads. mRNA capture efficiency was approximately 10%, as measured by RNA spike-in panels (ERCC). 80% of filtered reads aligned to exonic and splice-junction features in the mouse genome, which is high even by today's standards. Introduction of degenerate bases (NNNNNNNNNN) in the 5' adapter allowed multiple counts of PCR duplicates to be collapsed into transcript counts.¹⁵³ Addition of these unique molecular adapters (UMIs) to the STRT-seq protocol, plus parallel automated cell capture on microfluidic cell capture arrays (Fluidigm C1), and oligo design optimisation improved the capture efficiency of STRT-seq from ~10% to ~50%.¹⁵⁴

¹⁴⁴Eberwine, PNAS, 1992.

¹⁴⁵Kamme, J Neuroscience, 2003.

¹⁴⁶Brady, Meth Enzym, 1993.

¹⁴⁷Brady, Yeast, 2000.

¹⁴⁸Kurimoto, Nucl Ac Res, 2006.

¹⁴⁹Tang, Nat Met, 2009.

¹⁵⁰Islam, Genome Res, 2011.

¹⁵¹Islam, Nat Prot, 2012.

¹⁵²Zhu, Biotech, 2001.

¹⁵³Kivioja, Nat Meth, 2011.

¹⁵⁴Islam, Nat Meth, 2014.

In parallel research, also at the Karolinska Institute, Ramsköld and Sandberg developed Smart-seq (Switching Mechanism at the 5' end of RNA Template).¹⁵⁵ Similarly to STRT-seq, a combination of MMLV reverse-transcriptase followed by cDNA PCR was used. However, fragmentation of these full-length transcripts using Tn5 (or Covaris shearing) allowed for shotgun sequencing of the entire transcript. 42 cells from several human cancer cell lines, circulating tumor cells, and mouse oocytes were profiled. Improved isoform coverage showed that exon 13 of the NEDD4L gene, a prognosis marker, was differentially expressed between LNCaP prostate cancer and T24 bladder cancer cells. A year later, an updated Smart-seq protocol with increased sensitivity was published.¹⁵⁶ Here, the addition of a locked nucleic acid (LNA) at the TSO's 3' end and addition of betaine led to 2-4 fold cDNA yield increases. The workflow was also simplified and reduced in cost by switching to KAPA HiFi polymerase, which reduced GC-biases. The use of in-house produced recombinant Tn5¹⁵⁷ further reduced costs for Smart-seq2.¹⁵⁸ Researchers could now bypass Illumina's Nextera Tn5, which to this day is still enormously expensive (~\$3600 per 160 μ L). A quantitative comparison of Smart-seq and bulk RNA-seq revealed that aggregate Smart-seq data closely resembles bulk RNA-seq data from 1 million cells, and correlates strongly (Pearson R = 0.9) to ERCC spike-ins.¹⁵⁹ Moreover, nanoliter volume reactions used in the Fluidigm C1 consistently identified more genes and had more uniform transcript coverage than microliter volume reactions due to higher enzyme, template and substrate concentrations and lower loss in washes. Continued development led to further sensitivity improvements in Smart-seq3.¹⁶⁰ CEL-seq, developed in 2012, elected to use linear amplification by in-vitro transcription after reverse transcription rather than PCR-based approaches - effectively translating the 1992 Eberwine protocol to sequencing.¹⁶¹ The protocol generated 280 000 mRNA counts from single *C. elegans* blastomeres. An updated protocol later increased low capture efficiency from ~6% to ~22% and increased throughput to 96-format using the now trending Fluidigm C1.¹⁶²

In 2013, Sasagawa et al. argued that Smart-seq, CEL-seq and STRT-seq suffered from amplification biases caused by 3' enrichment, template switching, and pseudo-random Tn5 fragmentation, resulting in poor correlation between technical replicates (Pearson R between 0.7 and 0.85).¹⁶³ Their method, Quartz-seq, showed several sensitivity and specificity improvements: by employing Exo I digestion after reverse transcription and cDNA Covaris instead of Tn5 fragmentation, the amount of adapter sequences present in the data was reduced from 21% to 7%, and technical reproducibility was increased (Pearson R of ~0.93). Later, Quartz-seq2 would improve throughput by using FAC sorting into 1536-wells, and reduce costs from \$23 to \$0.50 per cell by earlier pooling of the reactions.¹⁶⁴ Stringent live-cell selection with FACS, addition of a UMI to the RT primer, optimized poly(dA) tailing of cDNA, and use of Y-shaped Illumina adapters further reduced biases and increased sensitivity, achieving a "UMI conversion rate" of 30-50%, compared to the original Quartz-seq and Drop-seq's 10%, and CEL-seq's 22%.^{165,166}

At this point in time, the first plate-based scRNA-seq protocols - particularly Smart-seq - began to spread to other labs, and single-cell sequencing was elected Nature method of the year in 2013.¹⁶⁷ For example, analysis of 430 single glioblastoma cells with Smart-seq showed heterogeneous expression of several receptor tyrosine kinases (EGFR, PDGFRA, FGFR1), potentially leading to redundancy and therapeutic resistance.¹⁶⁸ The existence of CD133+, therapy-resistant "glioblastoma stem-like cells" had been shown before,¹⁶⁹ but single-cell technologies made deeper characterisation possible. However, due to the high reagent and labour costs per cell (approximately \$30 per cell plus cost of sequencing), almost all studies from this period were limited to few (hundreds) of cells, mainly sourced from few-cell stage embryonic samples.¹⁷⁰ Profiling samples with inherently large cell counts and diversity, such as immune systems and neuron clusters, would require a different approach.

In 2014, researchers introduced a strategy to increase the throughput of plate-based methods by reducing costs through early pooling of reactions. In MARS-seq,¹⁷¹ cells were FAC sorted into 384-well plates and subjected to cell-barcoded reverse transcription. Reactions were then pooled, and second-strand synthesis and subsequent steps occurred in a shared microtube. cDNA was amplified by IVT, fragmented, reverse transcribed, further amplified by PCR, and sequenced. Applied to ~5500 mouse spleen cells, this approach revealed novel

¹⁵⁵Ramsköld, Nat Biotech, 2012.

¹⁵⁶Picelli, Nat Met, 2013.

¹⁵⁷Picelli, Gen Res, 2014.

¹⁵⁸Picelli, Nat Prot, 2014.

¹⁵⁹Wu, Nat Meth, 2014.

¹⁶⁰Hagemann-Jensen, Nat Biotech, 2020.

¹⁶¹Hashimshony, Cell Rep, 2012.

¹⁶²Hashimshony, Gen Biol, 2016.

¹⁶³Sasagawa, Gen Biol, 2013.

¹⁶⁴Sasagawa, Gen Biol, 2018.

¹⁶⁵Ziegenhain, Mol Cell, 2017.

¹⁶⁶Mereu, Nat Biotech, 2020.

¹⁶⁷Nature, Nat Meth, 2014.

¹⁶⁸Patel, Science, 2014.

¹⁶⁹Bao, Nature, 2006.

¹⁷⁰Kolodziejczyk, Mol Cell, 2015.

¹⁷¹Jaitin, Science, 2014.

dendritic cell subpopulations with distinct activation of *Irf4*, *TNF*, and *TGF- β* pathways in response to LPS stimulation.

Several scaling breakthroughs were made in 2015, the "annus mirabilis" for single-cell sequencing methods. First, droplet microfluidics¹⁷² was applied to single-cell sequencing. Researchers at David Weitz' lab had been working on using nanoliter droplets to perform antibody screening and rare template PCR,¹⁷³ referring to a hypothetical system of encapsulating molecules with a nucleic acid barcode that "*indicates which element is encapsulated in the droplet*".¹⁷⁴ This system was realised for single-cell RNA-seq in 2015 by two separate labs, both at Harvard in collaboration with Weitz lab: Macosko and colleagues (McCarroll lab) developed Drop-seq,¹⁷⁵ and Klein and colleagues (Kirschner lab) developed inDrop.^{176,177}

Both methods used droplet microfluidics to encapsulate single-cells in nanoliter volume droplets together with a bead carrying barcoded poly(dT) primers, but their amplification and library preparation strategies differed. Drop-seq used a Smart-seq-like approach: once encapsulated, cells are lysed, and their polyA⁺ mRNA is captured by the beads, which carry a poly(dT) oligonucleotide primer containing a UMI and a cell barcode. The emulsion is then broken, and beads are pooled and washed in saline sodium citrate to minimize cross-bead hybridizations. The beads, still with mRNA hybridized, are resuspended in a reverse transcriptase mix, and RT proceeds on-bead to produce full-length cDNA. After RT, exonuclease I is applied to degrade any unextended primers, reducing noise from unused barcode sequences. The resulting cDNA, which is covalently linked to the bead, is then PCR-amplified. Fragmentation proceeds using Nextera XT Tn5 to generate a sequencing library. All steps involving enzymes happen in a pooled manner, saving reagent cost. Drop-seq's cell purity (absence of cell doublets) was assessed using a mixture of HEK (human) and 3T3 (mouse) cell lines, and ranged between 90.4% to 98.8% based on cell concentration used. The method was demonstrated on 44 808 cells from the P14 mouse retina generated in 7 runs across 4 days, producing 39 cell clusters. *PPP1R17*, a new marker for the poorly-characterized nGnG (non-glycinergic/non-GABAergic) amacrine cells, was identified and validated by histochemistry. Transcript capture efficiency was estimated at ~10.7 to 12.8% using ddPCR validation. Drop-seq marked the first single-cell dataset spanning tens of thousands of cells, and reported a >100-fold reduction in cost and time compared to prior scRNA-seq methods (~6.5 cents per cell, 10 000 cells in 12 hours). Later, a microfluidics-free version of the protocol, using bead arrays rather than droplets, was developed.¹⁷⁸

inDrop used a different approach to both cell loading and reaction chemistry: cells are encapsulated with soft, deformable hydrogels rather than solid resin beads like Drop-seq, which can be stacked in the microfluidic channel. This ordered stack of beads allows almost every droplet to receive a bead, rather than ~2% for the poissonian loading used in Drop-seq.¹⁷⁹ Reaction chemistry was also different. After encapsulation, UV light releases the barcoded poly(dT) primers from the bead into the droplet, and reverse transcription happens within the droplet. Next, the CEL-seq IVT strategy is used to amplify the barcoded, single-cell transcriptomes in a pooled manner. The protocol was demonstrated on a total of 11 209 mouse embryonic stem cells, including a time series of leukemia inhibitory factor (LIF) withdrawal-induced differentiation over 7 days. 5 distinct ES sub-populations were identified, including rare lineages marked by epiblast (*Krt8/18*), primitive endoderm (*Col4a1/2*, *Lama1*), and heat shock markers (*Hsp90*, *Hspa5*). Counts agreed strongly with smFISH data ($R = 0.92$ for mean, 0.90 for CV).

Both Drop-seq and inDrop boasted a significant jump in throughput over pure plate-based sorting methods, from hundreds to several thousands, potentially tens of thousands, of cells per day. Higher cell counts, traded for per-cell sensitivity, now allowed for the profiling of highly varied samples such as immune cells. However, both methods also required fabrication of complex reagents (notably, barcoded beads and microfluidic chips) and complex microfluidic operation. In 2016, the first commercial droplet-microfluidic single-cell sequencing solution was released: 10x Genomics' 3' Single Cell Gene Expression kit and accompanying Chromium microfluidic controller combined barcoded hydrogel beads, enzymes, primers and microfluidics in a relatively simple package. The method drew on concepts from both inDrop and Drop-seq, opting for deformable hydrogel beads for higher cell capture, and a SMART-seq-like, PCR + Tn5 amplification and library preparation strategy. Importantly, the Chromium could load up to 8 samples in parallel, simplifying arrayed experiments. The method was

¹⁷²Thorsen, Phys Rev L, 2001.

¹⁷³Mazutis, Nat Prot, 2013.

¹⁷⁴Guo, Lab on Chip, 2012.

¹⁷⁵Macosko, Cell, 2015.

¹⁷⁶Klein, Cell, 2015.

¹⁷⁷Zilionis, Nat Prot, 2017.

¹⁷⁸Gierahn, Nat Met, 2017.

¹⁷⁹Abate, Lab on a Chip, 2009.

first demonstrated on 68 000 PBMCs sampled from 8 donors, identifying all major PBMC subtypes (T, NK, B, myeloid) and minor populations (e.g. megakaryocytes).¹⁸⁰ Approximately 4500 genes and 27 000 UMIs were detected on average per cell at a depth of 100k reads/cell, an order of magnitude higher than inDrop and Drop-seq. Mixed-species loading revealed a doublet rate of approximately 3%.

2015's second leap came when researchers in Shendure lab applied combinatorial indexing to single cells in-situ,¹⁸¹ marking the first single-cell ATAC sequencing protocol (sci-ATAC-seq, single-cell combinatorial indexing ATAC sequencing). In sci-ATAC, nuclei are tagged using Tn5 loaded with an array of barcoded adapters in 96-well plates and re-distributed into a second 96-well plate containing PCR mixes with unique primer combinations. Each nucleus is thus indexed by a combination of its Tn5 and PCR barcode, and individual cells are never physically separated. The method was applied to 14 533 cells from 4 different cell lines. Between 50 and 60% of reads were located in ENCODE-defined DNase I hypersensitivity sites, confirming assay fidelity. Approximately 5000 accessible fragments are captured per cell on average. Cross-cell collision rate was estimated at 11% using species identities in the cell lines, and individual cell barcodes contained ~90% species-pure accessible fragments. Interestingly, 4 different states were observed in GM12878 cells based on accessible enhancers. The same combinatorial indexing strategy was applied to scRNA-seq in 2017, this time profiling 49 360 *C. elegans* L2 larvae, approximately 50-fold coverage over its 762 somatic cells at this stage.¹⁸² Instead of 96 × 96, split-pooling was performed over 576 × 960 possible wells (first during in-situ reverse transcription, then Tn5 fragmentation and PCR). Median UMIs and genes detected in *C. elegans* was 575 UMIs and 431 respectively, identifying 27 major cell types, including rare neurons with only 1 or 2 cells per larva. Reagent costs were estimated at \$0.03-\$0.20 per cell. Updated protocols were published in 2022, offering higher sensitivity and lower costs (\$0.01 per cell).¹⁸³ SPLiT-seq (split-pool ligation-based transcriptome sequencing),¹⁸⁴ an independently developed combinatorial indexing protocol, introduced additional barcode ligation steps to increase barcode complexity to 96⁴. The method was first applied to generate 156 049 single-cell transcriptomes from P2 and P11 mouse brains and spinal cords,¹⁸⁵ and was recently commercialized into kits by Parse Biosciences - offering up to 10 million cells indexed per run.¹⁸⁶

Buenrostro and Greenleaf also released a protocol for ATAC-seq at single-cell level in 2015.¹⁸⁷ In the protocol, cells are loaded onto 96-plex chips on the Fluidigm C1, and lysed in the nanoliter reaction chamber. Nuclei are then incubated with Tn5 and washed with an EDTA or SDS buffer, releasing Tn5 in-situ. Individual nuclei are PCR-cycled on-chip with barcoded primers, indexing cell identity. Libraries are then collected for further amplification, and sequenced. The C1 scATAC-seq protocol was applied to a consortium of GM12878 (human lymphoblastoid), K562 (human leukemia), H1 hESCs, V6.5 mouse ESCs, EML1 (mouse early hematopoietic), TF1 (human erythroblast), HL-60 (human promyeloblast), and BJ (fibroblast) for a total of 1632 cells profiled. On average, 73 000 fragments aligned to the genome, and around 35% of reads were located in called peak regions. While the aggregate data was approximately 500-fold more sensitive than bulk ATAC-seq (normalized to cell counts), individual aggregate-accessible peaks contained an average of 0.10 fragments per cell - showing extreme sparsity. Pearson correlation between samples and correlation with ENCODE DNase-seq data was high ($R = 0.86$ and 0.80).

The sci-ATAC-seq and the Fluidigm C1 protocol require either a custom barcoded Tn5 (not widely available at the time), or an expensive, specialized microfluidics controller. Moreover, the C1 protocol lacked scale, and the sci protocol was low in sensitivity. A relatively simple and robust protocol, leaning on FAC sorting of bulk pre-tagmented nuclei into individual 96/384-well plates, and amplifying their accessible chromatin individually, was published in 2018, providing intermediate scale and sensitivity.¹⁸⁸ In 2019, two commercial solutions, both using high-throughput droplet-microfluidics, were released by BioRad and then 10x Genomics. BioRad's product, the ddSEQ SureCell ATAC, was a hard, resin bead-based protocol developed in conjunction with Buenrostro lab.¹⁸⁹ Bulk pre-tagmented nuclei are encapsulated with barcoded hard resin beads on the ddSEQ microfluidic platform. Since super-poissonian loading cannot be achieved without hydrogel beads, only a small percentage of droplets would contain a bead under normal circumstances, resulting in a low cell capture rate of around 2% - similar to Drop-seq. However, smart application of Tn5 tracking allowed the low capture rate to be circumvented:

¹⁸⁰Zheng, Nat Comm, 2017.

¹⁸¹Cusanovich, *Science*, 2015.

¹⁸²Cao, *Science*, 2017.

¹⁸³Martin, Nat Prot, 2023.

¹⁸⁴Rosenberg, *Science*, 2018.

¹⁸⁵Kuijpers, BMC Gen, 2024.

¹⁸⁶Parse Biosciences, *Press Release*, 2024.

¹⁸⁷Buenrostro, *Nature*, 2015.

¹⁸⁸Chen, Nat Comm, 2018.

¹⁸⁹Lareau, Nat Biotech, 2019.

beads are "overloaded", leading to 2.5 beads rather than < 0.05 encapsulated per droplet on average. Each nucleus' material is thus split over multiple single-cell barcodes. Cell barcodes that are statistically significantly enriched for overlap between their captured fragments can be detected and merged in downstream analysis. The resulting cell capture rate for dscATAC-seq is reported as 95% (5324 cells recovered from 5600 loaded). Genomic alignment and sensitivity were also high, reporting 95% alignment to reference, and a median of 34 046 unique reads per cell with 58.8% in peaks. In addition, barcoded Tn5 ("dsci-ATAC") was demonstrated to further increase throughput to up to 100 000 cells per device, for a total of 510 123 single cells profiled in the study. In contrast to the BioRad product, the 10x Genomics solution used hydrogel beads, just like their gene expression kit.¹⁹⁰ Bulk pre-tagmented are encapsulated in droplets, where the bead dissolves and releases barcoded primers. In-situ PCR indexes single-nucleus accessible fragments, and a second PCR generates a sequencing-ready library. The method was applied to 61 806 cells from blood and bone marrow across 16 donors and 37 818 basal cell carcinoma cells from 7 patients, confirming tumor-specific variation in accessibility and copy numbers. An average of 27 800 unique fragments were detected per cell, and between 31.5 and 62.5% of reads were located within peaks. Based on both articles, it thus appeared that the BioRad's dscATAC was more sensitive than the 10x solution.

For both RNA-seq and ATAC-seq, higher sensitivity of extraction and amplification methods allowed researchers to translate these protocols to the single-cell level. A known problem with classic ATAC-seq methods is that only fragments tagged with an i5 adapter on one end and an i7 adapter on the other are PCR-amplified. Since adapter loading on the Tn5 dimer is a random process, and each insertion site is equally likely to receive either possible combination of i5 and i7 insertions, 50% of all fragments generated are either i5-i5 or i7-i7 species. Symmetrical strand sci-ATAC (s3-ATAC) solves this problem by introducing a uracil base after Tn5's mosaic end adapters.¹⁹¹ Uracil-intolerant polymerase is then used in the gap fill step, generating fragments with incomplete overhang. These overhangs can be reliably targeted with a universal i5 adapter, resulting in fragments which are always i5- and i7-ended - effectively doubling capture. Application on 2175 human and 837 mouse brain cells yielded between 100 280 and 178 069 median unique reads per cell, approximately 6-fold higher than 10x Genomics scATAC-seq, with between 48.1% and 78.2% of reads falling in peaks.

1.7 Scaling cell throughput

The release of the 10x Chromium scRNA-seq kit marked a milestone in single-cell sequencing research. It was user friendly enough, and struck a good balance between cell throughput and single-cell quality. Large single-cell atlases included Tabula Muris, which combined droplet-based sequencing with plate-based Smart-seq2,¹⁹² and a *Drosophila* brain atlas generated using 10x Chromium.¹⁹³ The Human Cell Atlas was proposed as an international initiative in 2017.¹⁹⁴ In the fly brain atlas, 157 000 high-quality single-cell *D. melanogaster* transcriptomes were generated in 26 runs sampled across 8 timepoints between 0 and 50 days old. CEL-seq2 and SMART-seq2 data supplemented the medium-depth 10x Chromium cells with high-quality "anchors". Novel cell-type-specific gene regulatory networks were inferred from olfactory projection neurons, and glia using SCENIC. Participant TFs were validated for motif enrichment in these cell types using FACS-assisted ATAC-seq.

Meanwhile, plate-based combinatorial indexing strategies like SPLiT-seq and the sci-family of protocols have in general skirted further the extreme end of the sensitivity/throughput scale, leading to their use in several million-cell scale atlasing projects (Table 1.1).

Scaling droplet-based methods to million cell numbers is also possible. Cells or nuclei can be distributed over 96-well plates and pre-indexed by reverse transcription with a barcoded primer akin to plate-based combinatorial indexing protocols ("scifi-RNA-seq").²⁰⁰ The pre-indexed cells are then pooled and run through 10x Chromium droplet generation together at extremely overloaded concentrations: whereas the 10x Chromium was designed for ~10 000 cells per channel, scifi-RNA-seq increases this to up to 1.53 million cells per channel without apparent reduction in per-cell data quality. The same principle can be applied to ATAC-seq: scifi-ATAC-seq and txci-ATAC-seq use pre-indexing in 96 wells with barcoded adapter Tn5, enabling the profiling of up to

¹⁹⁰Satpathy, Nat Biotech, 2019.

¹⁹¹Mulqueen, Nat Biotech, 2021.

¹⁹²Consortium, Nature, 2018.

¹⁹³Davie, Cell, 2018.

¹⁹⁴Regev, eLife, 2017.

²⁰⁰Datlinger, Nat Meth, 2021.

TABLE 1.1 Large-scale single-cell atlases using combinatorial indexing techniques

Year	Cells	Sample	Technique
2019 ¹⁹⁵	2 072 011	E9.5 to E13.5 mouse embryos	sci-RNA-seq3
2020 ¹⁹⁶	4 062 980	Human embryos	sci-RNA-seq3
2023 ¹⁹⁷	2 583 967 + 1 587 880	Macaque brain	sci-RNA-seq3 + snATAC-seq3
2024 ¹⁹⁸	12.4 × 10 ⁶	E8 to P0 mouse embryos	sci-RNA-seq3
2025 ¹⁹⁹	100.6 × 10 ⁶	50 human cancer cell lines	Parse GigaLab

200 000 single nuclei per 10x scATAC-seq lane.^{201,202} Sensitivity also remained high, sampling between 21 362 and 45 622 unique fragments per cell on average across tissues assessed. Interestingly, these findings suggest that the chemical and physical reactions within the nanoliter droplets are not rate-limited by concentration of enzymes and reagents.

Similar approaches have been used not to increase pure throughput, but to scale number of samples per droplet run. Again, the main strategy here is to pre-index or "hash" samples with oligonucleotide barcodes. Hashed samples can then be co-processed through droplet generation together, and the sample hash can computationally separate cells into sample bins post hoc. Hashing can be done by attaching sample-barcoded oligonucleotides to their membranes, either by antibody-antigen binding²⁰³ or lipid/cholesterol embedding.²⁰⁴ Within each droplet, the membrane-bound oligonucleotides co-captured together with the associated cell's mRNA or tagged genome. Antibodies have demonstrated slightly better accuracy, while lipid hashing can be used when high-strength antibodies are not available.²⁰⁵ The increased precision of antibody-based methods could be attributed to the many successive washes involved, but which may also reduce cell viability and integrity.

Instead of surface antibodies, oligonucleotides can also directly index generic (not custom-barcoded) Tn5. SNUBar (single-nucleotide barcoding) introduces a single oligonucleotide barcode (SBO) during tagmentation.²⁰⁶ This SBO contains an 18 bp sample barcode and hybridises to Tn5 transposomes through a universal adapter sequence. During tagmentation these Tn5-SBO complexes remain stable as Tn5 integrates itself into accessible regions. Nuclei are then washed, removing unbound Tn5 and excess SBOs, and samples are pooled. After droplet encapsulation, the SBO is released from Tn5 and captured by the cell barcode during PCR amplification, linking cell barcodes with sample barcodes.

1.8 Perturbations

Large screens that investigate the effect of silencing specific genes using RNA interference emerged in the early 2000s.^{207,208} Such perturbation screens follow either an arrayed design, where perturbations are individually plated and tracked in physical space, or a pooled design, where cell lines are transduced with pooled shRNA vector libraries at a given MOI, and survivors are sequenced. When CRISPR was developed as a molecular tool for rapid and accurate gene editing in 2013,²⁰⁹ full genetic knockouts became widely available, producing a pooled genome-wide screen involving 64 000 sgRNAs for 18 000 gene knockouts just a year later.^{210–212} In addition to genetic knockouts, specific genes can also be up- or downregulated using catalytically dead *Cas9* fused to transcription activators such as *VP64* or *p65*, or repressors such as *KRAB*. Both arrayed and pooled

²⁰¹Zhang, Gen Biol, 2024.

²⁰²Zhang, Gen Biol, 2024.

²⁰³Stoeckius, Gen Biol, 2018.

²⁰⁴McGinnis, Nat Meth, 2019.

²⁰⁵Mylka, Gen Biol, 2022.

²⁰⁶Wang, Mol Cell, 2021.

²⁰⁷Berns, Nature, 2004.

²⁰⁸Boutros, Gen Biol, 2006.

²⁰⁹Ran, Nat Prot, 2013.

²¹⁰Shalem, Science, 2014.

²¹¹Behan, Nature, 2019.

²¹²Pacini, Nat Comm, 2021.

CRISPR/RNAi screens have drawbacks: arrayed screens require vast amounts of labour and reagents, and pooled CRISPR-screens are limited to very simple readouts such as cell survival or marker expression.

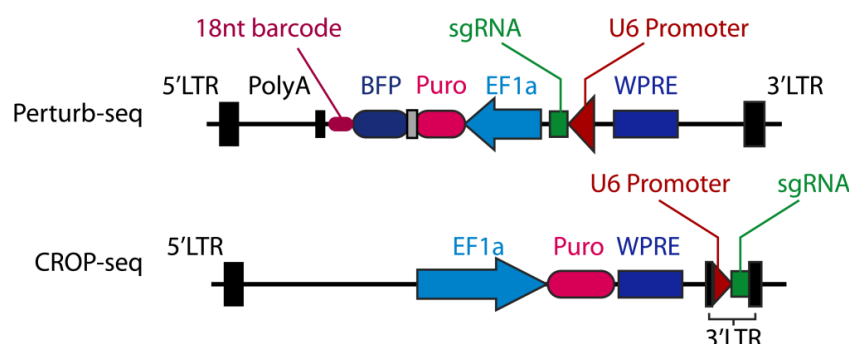


FIGURE 1.11. Common single-cell perturbation vector designs. Adapted from [213].

In 2016, researchers at the Broad institute devised that CRISPR edits could be performed in a pooled manner with a read-out at the single-cell level - leveraging the newly developed scRNA-seq technologies to also read perturbation identity at the single-cell level.²¹⁴ In this method, Perturb-seq, each vector is designed to express two transcripts after integrating in the host genome: a first transcript containing the sgRNA, driven by U6, and a second transcript containing a puromycin selection cassette, BFP, and a 20 nucleotide long guide barcode (GBC), driven by EF1 α . The GBC transcript is polyadenylated, allowing it to be captured and counted through standard 3' scRNA-seq protocols (10x GEX in this specific case). The {sgRNA:GBC} dictionary can then be used to identify the genetic edit present in each captured cell.²¹⁴ This pooled approach allows rapid parallelisation of perturbation screens with limited labour. Using this system, researchers observed that even within the same perturbation, individual cells can exhibit different responses. In a separate Perturb-seq study, Adamson and colleagues found heterogeneous PERK/ATF4 activation after depletion of BiP (*HSPA5*), despite similar levels of *HSPA5* knockdown.²¹⁵ TAP-seq (Targeted Perturb-seq) improves on the original protocol by using targeted gene panel amplification rather than whole-transcriptome analysis.²¹⁶ 96% of sequenced reads mapped to target genes compared to ~10% in typical Perturb-seq experiments, achieving the same analytical power as Perturb-seq at 19- to 49-fold lower cost. Application to a dCas9-KRAB vector with sgRNAs targeting 1788 candidate enhancers at MOI of 1 identified 81 new enhancer-target gene pairs. These 81 enhancers affected 24% of the genes in the amplification panel, a hit rate far higher than previous Perturb-seq screens (3.5 - 6.0%). The assay's costs can be further scaled down by multiplexing perturbations or cells: if gene regulatory networks are assumed sparse in number of participants, then the differential expression of a cell responding to multiple perturbations (high MOI), or a droplet containing multiple single-perturbation cells, can be considered a linear combination of the effect of its components. Pooling cells or perturbations in this manner and demultiplexing the individual terms using PCA can reduce cost ten-fold and quantify interaction terms between different perturbations.²¹⁷ CROP-seq, a different perturbation platform, uses a poly-adenylated gRNA which is directly captured by 3' scRNA-seq rather than guide barcodes expressed on the selection cassette (fig. 1.11).²¹⁸ The method was demonstrated using Drop-seq instead of 10x Chromium, showing that the open-source method has sufficient sensitivity to detect the gRNA transcripts. A small screen of Jurkat T cells, knocking out 6 genes and 23 TFs involved in T-cell receptor signaling pathways (87 gRNAs total), confirmed that *LCK*, *ZAP70*, and *LAT* knockouts blocked TCR induction. A total of 5,798 high-quality single-cell transcriptomes were analyzed, with a median of 80 cells per targeted gene.

Perturbation screens have also been coupled with scATAC-seq readouts. In Perturb-ATAC,²¹⁹ cells are transduced in pools with Perturb-seq-like vectors. Single cells were then captured and tagged on the Fluidigm C1, limiting the size of the study to ~4300 cells total. The GBCs are reverse-transcribed on-chip, allowing them to be amplified and captured through the ensuing ATAC PCR to generate perturbation-coupled scATAC-seq profiles. Dual knockdown of *EBF1* and *EZH2* (in B lymphoblasts), and *KLF4* and *ZNF750* (in keratinocytes), showed stronger repression than individual perturbations, and this relationship was observed to be cell-type specific.

²¹⁴Dixit, *Cell*, 2016.

²¹⁵Adamson, *Cell*, 2016.

²¹⁶Schraivogel, *Nat Meth*, 2020.

²¹⁷Yao, *Nat Biotech*, 2024.

²¹⁸Datlinger, *Nat Meth*, 2017.

²¹⁹Rubin, *Cell*, 2019.

Spear-ATAC is able to overcome Perturb-ATAC's throughput limitations by using droplet encapsulation on 10x Chromium.²²⁰ After pooled lentiviral transduction, cells undergo the regular 10x scATAC-seq workflow, but with an additional primer added to the droplet encapsulation step. This primer anneals to the integrated lentiviral cassette, amplifying its barcode. By using droplet-based methods, 414 sgRNA knockdowns could be profiled across ~105 000 cells (sampled between 3 and 21 days post transduction). Analyzing the resulting time series revealed time-dependent effects of the knockdowns. For example, *GATA1* knockdown was found to first upregulate *STAT5* motif accessibility, but later *RUNX* motifs. As expected, these GRNs were also observed to be cell-type specific between K562, MCF-7 and GM12878 cells.

One disadvantage of the CRISPR screening methods above is that most of the vectors described do not contain *Cas9*, limiting the studies to transgenic and immortal cell lines. Most perturbation studies so far also deal with CRISPRa/i, CRISPR knockouts, but not yet with gene expression constructs (for example: ectopic TF expression). Such studies are considerably more difficult to implement in a pooled manner, since transcript length has an effect on transfection, packaging and transduction efficiency, which in turn biases MOI and dosage. These effects call for the titration of viral concentrations and arrayed screening, which is expensive and laborious.

Single-cell RNA sequencing has recently also been used in chemical screens. For example, application of sci-RNA-seq on 5000 wells with 188 different small molecules $\times$ four doses $\times$ duplicates $\times$ 3 cell lines showed that *MEK* inhibitors affect *KRAS/ABL*-mutant A549 and K562 cells, but not MCF-7.²²¹ 17 different *HDAC* (histone deacetylase) inhibitors each shared a transcriptional trajectory, arresting cell-cycling signaling. Moreover, the large number of cells assessed (649 340) allowed per-condition cell counts to be aggregated into pseudo-viability scores. More recently, an extremely large study applied Parse Biosciences' GigaLab SPLiT-seq to profile 379 drugs across 50 genetically diverse human cancer cell lines, capturing over 95.6 million single-cell transcriptomes (median 1287 cells per condition, median 1890 transcripts per cell) across 52 886 unique cell line-drug-dose combinations. All cell lines were grown together to form "mosaic tumor" spheroids in mixed cultures. The screen, a collaboration between Arc institute and Vevo Therapeutics, revealed pronounced transcriptional shifts in response to proteasome, *HDAC*, *PI3K/AKT*, and *CDK* inhibitors, and confirmed genotype-specific responses to *RAS/RAF* pathway inhibitors like Dabrafenib and RMC-6236.

1.9 Single-cell data pre-processing

Once data is sequenced, and raw .fastq data is available, the first step in data analysis is alignment to a reference genome. A number of algorithms exist for this task, and the optimal choice depends on several factors. For genomic alignment, bowtie2 and bwa-mem are commonly used for Illumina short reads. Minimap2 is generally better suited to long reads from Nanopore and PacBio. Sequencing errors differ in frequency and type between platforms, and alignment parameters need to account for these differences.

Whereas genomic reads, such as generated by ChIP-seq, can be directly aligned to the genome, mRNA transcripts overlap discontinuously with the genome due to splicing, and are thus more challenging to align. Aligning short-read RNA-seq data to the genome requires splice site annotations, which required expressed sequence tags (ESTs) or full-length cDNA sequencing to generate. TopHat ended the need for ESTs by introducing computational splice site identification.²²² The algorithm identifies exons by looking for splice site motifs near the flanks of the reads that only partially map ("initially unmapped") to the reference genome with bowtie. Multiple reads are then aggregated to statistically validate a splice site. The resulting read alignments can then be passed to Cufflinks,²²³ which assembles them into possible transcript isoforms and can provide a statistical estimate of their relative quantities without long-read cDNA sequencing data. Like many classes of algorithms, a trade-off between memory and compute can also be made, for example STAR requires large amounts of memory due to its use of an uncompressed genome index, but is faster,²²⁴ while HISAT2 is more memory-efficient, but takes longer.

The diversity in sequencing technologies and alignment algorithms led to the use of several independent alignment formats, each tied to specific sequencing technologies or alignment algorithms. Analyses on different

²²⁰Pierce, Nat Meth, 2021.

²²¹Srivatsan, Science, 2020.

²²²Trapnell, Bioinf, 2009.

²²³Trapnell, Nat Biotech, 2010.

²²⁴Dobin, Bioinf, 2013.

datasets (e.g. variant calling, genotype analysis) required cross parsing or conversion. The introduction of the sequence alignment map (SAM) format as part of the 1000 Genomes Project in 2009,²²⁵ provided a unified standard for storing aligned sequencing data which has since become the de facto standard. Each read is stored in a single line containing 11 mandatory fields including: read name, reference name (chromosome or contig), position, mapping quality, the actual sequence, several other alignment quality parameters as well as optional tags (fig. 1.12).

```
@HD      VN:1.6  SO:coordinate
@SQ      SN:chr1 LN:248956422
@SQ      SN:chr2 LN:242193529

read001  99  chr1  1000  60  50M  =  1150  200
          ACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACG
          IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII  NM:i:0  MD:Z:50
read001  147 chr1  1150  60  50M  =  1000 -200
          ACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACG
          IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII  NM:i:0  MD:Z:50
```

FIGURE 1.12. SAM snippet.

An indexed binary representation (BAM) saves storage space and speed up searching and processing with random access. This also allows users to quickly inspect desired genomic regions using visualisation tools like *igv*.²²⁶ Alignment quality itself is also a useful metric for assessing an experiment, where low genomic or transcriptomic alignment rates are indicative of low quality. When this occurs independently of per-base sequencing quality (i.e. sequencing run issues), either sample quality (degradation, viability) or library preparation (low input DNA, low reverse transcription efficiency, over-amplification or over-fragmentation) are usually at fault. Additionally, metrics such as reads mapped to the mitochondrial genome can indicate sample damage. These cases require optimization searches of better extraction and library prep parameters.

	barcode_1	barcode_2	...	barcode_N
feature_1	3	2	...	13
feature_2	2	3	...	1
feature_3	1	14	...	18
⋮	⋮	⋮	⋮	⋮
feature_M	25	0	...	0

TABLE 1.2 Example feature count matrix across different cells.

After alignment, the next step is to vectorize the data by counting aligned reads in genomic regions. This generates a cell by feature count matrix, unlocking linear algebra (Table 1.2). The region sets used for counting differ by assay used. scRNA-seq requires gene annotations, whereas for scATAC-seq several strategies exist: ideally, previously validated sample-specific peak regions are used to limit data sparsity. These peak sets can be called from previous local data with peak callers like MACS.²²⁷ MACS uses a modeling-based approach: reads are shifted by half the learned fragment length to center to true binding sites (ChIP-seq bias). A sliding window scans the genome and uses a Poisson background (scaled to control when available) to call regions whose enrichment is statistically significant. It also removes PCR duplicates beyond what random sampling predicts. MACS outperformed contemporary tools (ChIPSeq Peak Finder, FindPeaks, QuEST) by calling more true peaks, showing higher motif enrichment and better summit-to-motif resolution while keeping lower false positives. MACS2

²²⁵Li, Bioinf, 2009.²²⁶Thorvaldsdóttir, Brief Bioinf, 2013.²²⁷Zhang, Gen Biol, 2008.

supports paired-end peak calling. Gaspar subsequently described a modified version that addressed artifacts in paired-end processing²²⁸. When sample-specific peak sets are not available, public data or genomic bins can be used. Particularly, the ENCODE SCREEN regions is a set of 926 535 human and 339 815 mouse candidate cis-regulatory elements (cCREs) identified based on ATAC-seq, DNase hypersensitivity, ChIP-seq, histone modifications (H3K4me3, H3K27ac), and CTCF binding aggregated from over 9000 efforts²²⁹. These regions represent potential enhancers, promoters, and insulators. Finally, the selected region set should be filtered for problematic regions to remove false peaks.²³⁰

Raw count matrices contain all counts in all barcodes sequenced in the dataset. Filtering true cells from background noise barcodes is a dataset-defining step for any single-cell dataset. The barcode-rank plot (also referred to as "knee" or "elbow plot") visualizes the distribution of reads across all sequenced barcodes in the sequencing dataset (fig. 1.13). It is obtained by counting the number of reads per barcode, and ranking each barcode in a log-log figure - similarly in concept to a Zipf or Lorenz curve.

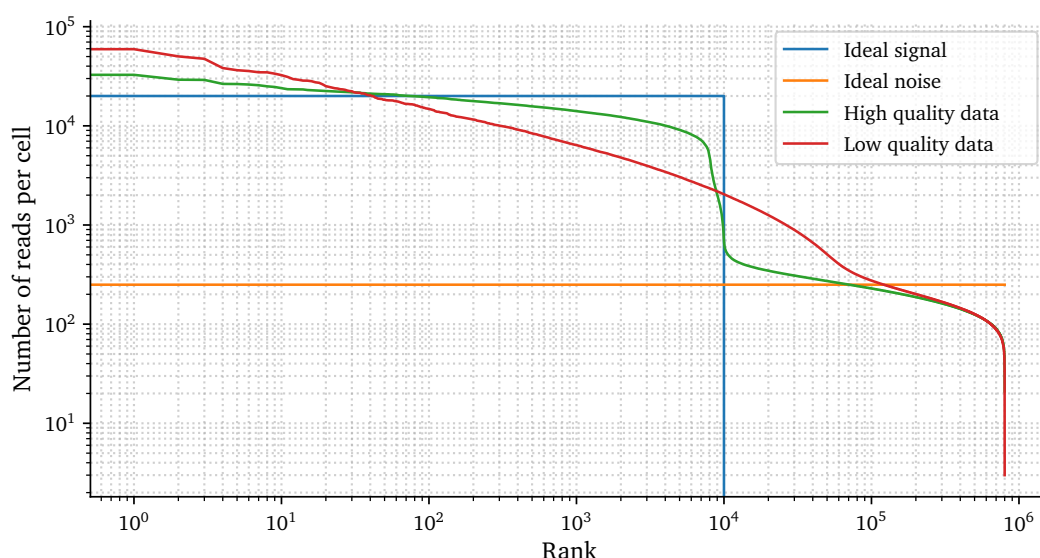


FIGURE 1.13. The barcode rank plot. In this simulation, 200 million reads are distributed across 10 000 cells, and the whitelist is 800 000 barcodes in length. The ceiling for mean number of reads per barcode is thus 200M reads/10k cell = 20k reads/cell (perfect signal), with a floor of 200M reads/800k barcodes = 250 reads/barcode (perfect noise).

In the "ideal signal" case, all reads are evenly distributed across a small number of barcodes that matches the number of input cells (e.g. 10 000 in the figure). This means that there are no quality differences between real cells, and there are no barcodes with low read counts. The opposite scenario ("ideal noise") has an even distribution of all reads across *all* barcodes, and we cannot distinguish true cell barcodes from noise barcodes. Real datasets fall between these two extremes. A high quality dataset exhibits a clear cut-off, where the number of reads per barcode sharply drops off in the transition between real cells and background noise. A dataset of lower quality will exhibit a more gradual transition between real cells and background noise. Based on this simple plot, we can estimate at a glance the rough quality characteristics of any single-cell dataset:

- The number of cells captured in the dataset is given by the barcode rank where true cells are separated from background noise ("knee"). If the dataset is sequenced to near-saturation, the total number of barcodes can even exceed the whitelist, as sequencing/PCR/synthesis errors introduce non-whitelisted species.
- The approximate number of reads associated with each true cell can be estimated from the y-axis.

²²⁸Gaspar, 2018.

²²⁹Consortium, *Nature*, 2020.

²³⁰Amemiya, *Sci Rep*, 2019.

- Overall data quality can be estimated by comparing the area under real cells and area under noise, or simply the "sharpness" of the cutoff.
- Counts assigned to barcodes below the cell-calling threshold do not, by themselves, determine the contamination level within retained cells. Barcode multiplets are a separate source of error.²³¹

1.10 Single-cell data post-processing

Conventional bulk RNA-seq analysis tools modeled gene counts using Poissonian (edgeR)²³² or negative binomial (DESeq)²³³ distributions. Statistical tests can then be used to identify statistically significant differentially expressed genes. Directly visualising correlations or variance between many RNA-seq or microarray samples is difficult due to the high-dimensionality of the data. Traditionally, this was done by first reducing the dimensionality of the data using PCA, and then visualising the first 2 principal components in a 2D space. Alternatively, heatmaps with dendrograms were used to order correlated samples.

Compared to bulk RNA-seq and microarray data, single-cell sequencing data is highly zero-inflated: when sampling transcripts in a single cell, the probability of weakly-expressed genes not being detected is high, leading to large amounts of "dropouts" (false negative gene expression). This sparsity, coupled with the more fine-grained nature of cell states in single-cell data further introduces non-linearity, since gradients of expression in the population are no longer averaged out. This problem was partially solved by the introduction of t-SNE, a technique first designed to visualize the MNIST dataset of handwritten digits.²³⁴ The earliest implementations of t-SNE were used for single-cell mass cytometry data (45 protein levels per cell),²³⁵ then single-cell qPCR (175 genes),²³⁶ and later single-cell whole transcriptomics (after PCA on ~20 000 genes).¹⁷⁵ t-SNE preserves local structure (close neighborhood of cells) at the cost of some concessions to global structure: clusters of single cells can appear closer in the t-SNE embedding than they are in high-dimensional gene space. UMAP is a slightly newer dimensionality reduction algorithm which aims to better preserve global structure.²³⁷ Some common criticisms of both t-SNE and UMAP is that they still distort global topologies.²³⁸ In addition, both methods are sensitive to hyperparameters (perplexity, number of neighbours) and stochasticity (e.g. initialisation). These tools are thus only used for exploration and quick visualisation of the data.

By 2019, several single-cell sequencing technologies had emerged. Each had their own strengths and weaknesses, sometimes leading to technical artifacts in datasets even when assessing the same biological sample. Even the same technology used on different days could lead to batch effects due to sample degradation or other technical artifacts between experiments. This translates to cells clustering by dataset batch rather than biological state. Seurat aims to facilitate integration of multiple single-cell datasets across different technologies and conditions.²³⁹ The package uses canonical correlation analysis (CCA)²⁴⁰ to calculate linear combinations of genes (canonical correlation vectors, in this case "metagenes") that capture common information between two datasets. Cells are then co-projected in this shared metagene space to create a new "joint embedding" where clustering and differential expression analysis can be performed. The canonical vectors can be used to re-transform differential metagenes back to the original gene set used. Seurat's CCA implementation could also be used for label transfer: if one dataset is labeled or annotated, proximity in the joint embedding space can be used to transfer labels from one dataset to another. Similarly to Seurat, Scanpy unifies preprocessing, highly variable gene selection, PCA, graph-based clustering, t-SNE/UMAP embedding, differential expression in a python package.²⁴¹ The cells by gene count matrix is stored along unstructured metadata in the AnnData format.

Whereas scRNA-seq is limited to counts in ~22 000 genes (human, mouse), the accessible genome can span hundreds of thousands of open chromatin regions which can each be sampled a maximum of two times per cell. These effects lead to highly-dimensional, but sparse count matrices. Latent semantic indexing (LSI), an approach borrowed from text analysis, can be used to perform cell clustering and dimensionality reduction.²⁴²

²³¹Lareau, Nat Comm, 2020.

²³²Robinson, Bioinf, 2010.

²³³Anders, Gen Biol, 2010.

²³⁴Maaten, J Mach Learn, 2008.

²³⁵Amir, Nat Biotech, 2013.

²³⁶Saadatpour, Gen Biol, 2014.

²³⁷Becht, Nat Biotech, 2018.

²³⁸Chari, PLoS Comp Biol, 2023.

²³⁹Stuart, Cell, 2019.

²⁴⁰Hotelling, *Biometrika*, 1936.

²⁴¹Wolf, Gen Biol, 2018.

²⁴²Cusanovich, Cell, 2018.

Briefly, single-cell accessibility is first counted in either annotated regions or 5 kb genome bins, binarized, and scaled according to inverse document frequency (peaks that are not commonly accessible are weighted higher). Singular value decomposition (SVD), is then used to decompose the resulting weighted cell-accessibility matrix into latent features across cells:

$$M = U\Sigma V^T = \begin{bmatrix} \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \end{bmatrix}_{(m \times n)} = \begin{bmatrix} \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \vdots & \vdots & \ddots & \vdots \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \end{bmatrix}_{(m \times r)} \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix}_{(r \times r)} \begin{bmatrix} \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \end{bmatrix}_{(r \times n)}$$

Here, m is the number of peaks, n is the number of cells, and $r = \min(m, n)$ is the rank of the decomposition. U represents the peak-composition of the latent features, Σ is diagonal and scales each latent feature, and V^T represents the contribution of each latent feature to each cell. Based on Σ , we can select the top n most important features and cluster cells in this reduced space using t-SNE. Differentially accessible regions (DARs) between clusters can then be calculated on the raw binary accessibility matrix using logistic regression, which is suited to model the odds of a binary event like chromatin accessibility.

cisTopic, a probabilistic framework for scATAC-seq analysis, was used extensively to generate and post-process all scATAC-seq count matrices.²⁴³ LDA²⁴⁴ is used to calculate two sets of distributions: the region-topic distribution (regulatory regions grouped into "topics") and the topic-cell distribution (probability of a topic being associated with a cell) which should approximate the original data when multiplied.²⁴⁵

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}_{m \times n} \approx N \cdot \begin{bmatrix} p_{r_1 t_1} & p_{r_1 t_2} & \cdots & p_{r_1 t_k} \\ p_{r_2 t_1} & p_{r_2 t_2} & \cdots & p_{r_2 t_k} \\ \vdots & \vdots & \ddots & \vdots \\ p_{r_m t_1} & p_{r_m t_2} & \cdots & p_{r_m t_k} \end{bmatrix}_{m \times k} \cdot \begin{bmatrix} p_{t_1 c_1} & p_{t_1 c_2} & \cdots & p_{t_1 c_n} \\ p_{t_2 c_1} & p_{t_2 c_2} & \cdots & p_{t_2 c_n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{t_k c_1} & p_{t_k c_2} & \cdots & p_{t_k c_n} \end{bmatrix}_{k \times n}$$

LDA parameters (probabilities) are estimated using a Markov chain monte carlo approach: regions are initially randomly assigned to topics, and topic-region assignment is then updated iteratively in each cell based on two factors: (1) the topic's occurrence in the cell and (2) the region's occurrence in each topic across all other cells. Sampling is only performed in this latent topic-assignment vector for each region in each cell (collapsed Gibbs sampling), which reduces total computational complexity compared to sampling in both region-topic and topic-cell distribution. This approach is more efficient than other Bayesian approaches (e.g. variational inference) due to the extreme sparsity of the data. t-SNE or UMAP can then be used to visualize the resulting normalized topic-cell distributions. LDA's topics are also interpretable,²⁴⁶ a property that is exploited to extract transcription factor binding motifs from the cis-regulatory topics found. Motif enrichment analyses can be performed using cisTarget.²⁴⁷ Regions and combinations of regions can also be annotated using nearest genes using ChIPseeker.²⁴⁸

cisTopic also introduces a region score which rewards regions that are more specific to a topic, and penalises regions that are ubiquitous across topics:

$$\text{Region score} = \beta_{r,t} \left(\log \beta_{r,t} - \frac{1}{T} \sum_{t=1}^T \log \beta_{r,t} \right)$$

²⁴³Bravo González-Blas, Nat Meth, 2019.

²⁴⁴Pritchard, *Genetics*, 2000.

²⁴⁵Blei, *J. Mach. Learn. Res.* 2003.

²⁴⁶Steyvers, 2007.

²⁴⁷Imrichová, *Nuc Ac Res*, 2015.

²⁴⁸Yu, *Bioinf*, 2015.

Consider Shannon entropy:

$$H(X) = - \sum_i p(x_i) \log p(x_i)$$

Then the term:

$$\beta_{r,t} \log \beta_{r,t}$$

can be considered the local information content of the region in topic context, while:

$$\frac{\beta_{r,t}}{T} \sum_{t=1}^T \log \beta_{r,t}$$

...is the average log-probability of the region across all topics.

The region score is thus an average-normalized measure of a given region's information content, indicating how specific it is to each topic. Topic-associated regions can be tested for transcription-factor motif enrichment using RcisTarget and region-based motif-ranking databases.²⁴³ cisTopic's LDA approach was benchmarked against other popular tools such as chromVAR, scABC, BROCKMAN, SCRAT, Cicero, and LSI, using both simulated single-cell ATAC-seq datasets and real data from FAC-sorted hematopoietic cells. cisTopic typically yielded higher scores for cell-type clustering, especially in low-coverage scenarios, confirming that the method can more robustly capture biologically meaningful patterns in sparse single-cell ATAC-seq data.

1.11 Deep-learning models

Recently, the availability of massive public databases and availability of compute resources have led to widespread application of deep learning in biology. Advances in algorithms trace back to 1959, when Hubel and Wiesel showed that specific neurons in a cat's visual cortex respond to common spatial patterns such as edges, corners, and orientations, in specific locations within the cat's visual field.²⁴⁹ These patterns can overlap when layers of neurons pass information onwards, forming broader structures which ultimately allow the cat to recognise complex objects. This idea was borrowed to implement convolutional neural networks (CNNs) for handwritten digit recognition.^{250,251} Rasterized (vectorized) images are scanned by a "filter", converting the original data into a new representation. This new representation is then passed on to the next layer, where it is again scanned by a new filter, and so on until the final layer. CNNs are thus well suited for detecting spatial patterns, and integrating these patterns into larger, more global structures. As a result, CNNs have been successfully used in enhancer prediction models. An early example was DeepSEA,²⁵² a CNN trained on ENCODE and Roadmap Epigenomics data (690 TF ChIP-seq + 125 DHS + 104 H3K27ac/H3K4Me3 datasets). This model could simultaneously predict these 919 chromatin features at 1 base resolution with higher accuracy than previous tools (CADD, GWAFA). Saturation mutagenesis analyses showed that DeepSEA could de novo predict known allelic imbalances and eQTLs (using chromatin features as a proxy).

Several track-predicting foundation models have also been developed. In 2018 Basenji2 was published, a CNN trained on 4229 human epigenomic (DNase-seq, histone-ChIP) and transcriptional (CAGE) tracks across more than 500 cell and tissue types. With a large receptive field of 131 kb,²⁵³ the model could predict 128-bp binned coverage across the genome: users feed any 131-kb reference or mutant sequence to the pretrained model (or fine-tune on new tracks) and obtain coverage predictions for each of the 4229 tissue-specific tracks

²⁴⁹Hubel, J Physiol, 1959.

²⁵⁰Fukushima, Biol Cyber, 1980.

²⁵¹Lecun, Proc IEEE, 1998.

²⁵²Zhou, Nat Meth, 2015.

²⁵³Kelley, Gen Res, 2018.

it was trained on. In 2021, Enformer replaced dilated convolutions with global self-attention and pushed the receptive field to 100 kb with a 200 kb input window.²⁵⁴ Trained now on 5313 human and 1643 mouse tracks (TF and histone-mark ChIP, DNase I, ATAC and CAGE), it predicts these tracks at 128 bp resolution, including long-range interactions. Borzoi is another CNN-attention hybrid, trained on thousands of CAGE, DNase, ATAC, and ChIP-seq datasets similarly to Basenji2 and Enformer, but also adding 866 human and 279 mouse RNA-seq tracks.²⁵⁵ The model predicts these tracks now at 32 bp-resolution. Kundaje lab introduced ChromBPNet, a CNN trained on ENCODE ATAC-seq and DNase-seq tracks for the five tier 1 ENCODE cell lines (GM12878, HepG2, K562, IMR90, H1-hESC).²⁵⁶ A smaller sub-net learned each enzyme's sequence biases based on non-peak regions signal, allowing ChromBPNet to isolate each motif's true footprint independent from assay used. This allows the model to predict, at single base resolution, the accessibility read profile (shape + total counts) that either ATAC-seq or DNase-seq would produce.

Recent work has also shown that single-cell datasets can be mined directly to learn enhancer grammar and design new regulatory elements. scRNA-seq of nine patient-derived cultures uncovered an intermediate state that co-expresses melanocytic factors (*SOX10*, *MITF*) and mesenchymal regulators (*SOX6*, *EGR3*).²⁵⁷ Follow-up bulk ATAC-seq at four time points after *SOX10* knock-down showed gradual chromatin shifts rather than abrupt switches. Expanding to 26 melanoma samples from six species, Minnoye et al. found *SOX10*/*MITF*/TFAP2A-rich enhancers in melanocytic cells and AP-1/TEAD motifs in mesenchymal cells, with motif usage conserved in dog, horse, pig and zebrafish.²⁵⁸ The DeepMEL model, a CNN-RNN trained on 339 099 human ATAC peaks, predicted state-specific enhancer activity from sequence alone and guided the synthesis of new MES and MEL enhancers. A similar approach has been used to decode *Drosophila* brain development. Janssens and colleagues profiled 240 919 scATAC-seq nuclei from the developing *Drosophila* brain across nine time points.²⁵⁹ Peaks were clustered with cisTopic to obtain topic sets of co-accessible, cell-type-specific regions. DeepFlyBrain was trained to predict each topic's aggregated accessibility from 2 kb of sequence. SHAP DeepExplainer scores highlighted the nucleotides most responsible for each prediction, revealing, for example, that *eyeless* and *mef2* motifs act synergistically when placed in close proximity, whereas repressor motifs can override activators. In-silico saturation mutagenesis confirmed the minimal length of active enhancers (as short as 49 bp), and reporter studies showed that 41 of 63 synthesized candidates drove the expected cell-type expression in transgenic flies. Building on these workflows, the CREsted pipeline trains compact CNNs (e.g. DeepBICCN2, 6M parameters) directly on scATAC-seq peaks, interprets motif logic and performs de novo enhancer design. Starting from topic-assigned peaks or pseudobulk heights, models are first pretrained on the full peak set and then fine-tuned for cell-type specificity. In mouse motor cortex, CREsted reached Pearson $r \sim 0.82$ on held-out peaks and retrieved the motifs underlying 171 in-vivo-validated enhancers. A human PBMC version recovered every known TF footprint at the CD79A and TCR α enhancers and captured 34% more informative bases than Borzoi. DeepZebrafish, trained on 793 273 peaks across 639 zebrafish cell-type/time-point combinations, served as an oracle for greedy in-silico evolution: five of seven 2 kb designs drove reporter expression specifically in cardiac muscle, somatic muscle or endothelial cells at 48 hours post fertilisation. Domain or sample-focused models like DeepMEL, DeepFlyBrain, DeepZebraFish complement large foundation models such as Basenji2, Enformer and scFoundation: they require far less compute, yet excel at highly specific questions like cell-type enhancer logic.

Introduced in 2017, the transformer model architecture abandoned recurrence and convolutions in favour of multi-head self-attention blocks.²⁶⁰ Every token attends to every other, so gradients propagate more directly and training parallelises across all sequence positions. This paper triggered the field-wide shift to attention-only networks that now dominate NLP, vision and sequence biology. For example, meta's Evolutionary Scale Modeling (ESM) applies transformer architecture (specifically, BERT) to protein language to predict masked tokens in amino acid sequences. The log-likelihood of proposed sequences serves as an estimate for how disruptive given mutations will be, allowing researchers to triage synthetic designs. ESM-3 also introduced multimodal masked token modeling: joint tokenized amino-acid sequence, atomic structure, and region annotations can be passed partially masked, and the model will jointly reconstruct missing tokens. This led to the generation of esmGFP, a completely synthetically designed fluorescent protein.

In a similar vein, the original DNABERT was an 86M parameter transformer trained on the 3.2 Gbp human

²⁵⁴Avsec, Nat Meth, 2021.

²⁵⁵Linder, Nat Gen, 2025.

²⁵⁶Pampari, bioRxiv, 2025.

²⁵⁷Wouters, Nat Cell Biol, 2020.

²⁵⁸Minnoye, Gen Res, 2020.

²⁵⁹Janssens, Nature, 2022.

²⁶⁰Vaswani, arXiv, 2017.

reference genome to classify promoters, TF-binding sites, and other regulatory features.²⁶¹ The model was competitive with ClinVar for SNV scoring, but had a relatively small context window of 512 bp, so could not compete in long range tasks such as enhancer-promoter pairing. Nucleotide Transformer (NT) extended the BERT approach to multi-species data: adding both genomic and evolutionary variety by using 3300 individual human genomes and 850 other organisms, totaling 262B tokens. NT consistently outperformed BERT and DeepSEA on promoter, enhancer, chromatin, and histone mark prediction, and enabled zero-shot scoring of SNPs and indels - achieving better accuracy than previous models on ClinVar benchmarks.²⁶² Recently, Evolutionary Scale Modeling for Genomes (Evo 2) pushed this scaling further with a 40B-parameter model trained on 9.3T tokens spanning all domains of life.²⁶³ The model scores variant effects across taxa, and supports up to 1 million tokens of context, enabling the generation of entire synthetic mitochondrial or bacterial genomes. Loss decreased smoothly with more data and parameters, with no saturation observed so far.

AlphaFold 1 was a CNN trained on protein structures (PDB) and multiple sequence alignments (MSAs) to predict residue-residue distances.²⁶⁴ AlphaFold 2 switched architecture, passing MSAs through Evoformer, a model with transformer-based self-attention, to predict spatial contacts and ultimately complete 3D protein structure.²⁶⁵ These changes led it to achieve a median backbone accuracy of 0.96 Å in CASP14 - surpassing the next best method by ~2 Å. AlphaFold 3 included a generative diffusion model to handle atomic coordinates directly rather than relying on MSAs. As a result, the model is capable of predicting binding interactions across protein, DNA/RNA, and small molecules, and performs better than AlphaFold 2 on other tasks.

Geneformer applies the transformer playbook to a collection of 30M human scRNA-seq cells across various tissues, with cells tokenized as a ranked list of their top 2048 genes sorted by expression (no count information).²⁶⁶ A transformer architecture is then trained to fill in masked gene positions, implicitly learning patterns of co-expression and regulation across tissues and conditions present in the data. The model is designed to fine-tune for specific tasks such as dosage-sensitivity classification, regulatory range prediction, and chromatin-state inference. scFoundation, another masked-prediction transformer, was trained on 100M single cells.²⁶⁷ Varying size from 3M to 100M parameters showed that its loss curves followed the same token vs FLOPs power laws seen in language models.

1.12 Motivation and approach

At the start of our project in 2018, single-cell sequencing methods had become established research tools, but were still expensive and complicated enough to preclude routine use. At the time, academic scRNA-seq protocols were either low in throughput,^{150,155,161,163} or low in sensitivity,^{175,176,178,182} and reserved for use by specialised labs (see section 1.6). The majority of cells were sequenced using the 10x Chromium, a commercial device/kits which packaged inDrop's droplet microfluidic cell isolation strategy with aspects of Drop-seq and SMART-seq's RNA library preparation into a user-friendly package.¹⁸⁰ The Chromium provided a convenient - and more importantly *reliable* - way to generate high-quality single-cell RNA-seq libraries²⁶⁸ and struck a comfortable balance between throughput and sensitivity.

We made three observations:

1. At a cost of €10 000 per 8 samples of 10 000 cells, the 10x scRNA kit was ill-suited for exploratory research.
2. The kit operated as a black box, with limited possibility for modification.
3. The kit's sensitivity was around 10 times higher than inDrop and Drop-seq's, implying that the sensitivity ceiling for droplet-based methods had not yet been reached by open-source methods.

In addition, a droplet-based single-cell ATAC sequencing method was not available yet, again constraining researchers to either low-sensitivity²⁶⁹ or low-throughput^{187,188} methods. Since droplet methods had provided

²⁶¹ Ji, Bioinf, 2021.

²⁶² Dalla-Torre, Nat Meth, 2025.

²⁶³ Brixi, bioRxiv, 2025.

²⁶⁴ Senior, Nature, 2020.

²⁶⁵ Jumper, Nature, 2021.

²⁶⁶ Theodoris, Nature, 2023.

²⁶⁷ Hao, Nat Meth, 2024.

²⁶⁸ Mereu, Nat Biotech, 2020.

²⁶⁹ Cusanovich, Science, 2015.

a balanced trade-off between throughput and sensitivity that fit our lab's scRNA-seq needs,¹⁹³ we set out to develop an equivalent for scATAC-seq. We also understood that combined scRNA/ATAC-seq would be necessary to deconvolve the individual and (cross-) combined effects of gene expression and chromatin accessibility in future experiments. A better understanding of how to design and build a droplet-based scATAC-seq method would be a good stepping stone to building multiomics technologies. Specifically, we wanted to build an open-source droplet platform capable of both scRNA-seq and scATAC-seq, at higher sensitivities than inDrop and Drop-seq. This work had already been initiated before my arrival at the lab, and I spent my first year as a master thesis student contributing to this project. Further researching and developing this method, which we later named HyDrop, occupied the first half of this PhD project.

Throughout the development of HyDrop-ATAC, we were often confronted with questions surrounding the method's sensitivity (the number of fragments detected per cell) and specificity (the enrichment of these fragments in enhancers and near transcription start sites, collectively forming chromatin accessibility "peaks"). Such questions arose both internally and externally, and echoed previous calls for an in-depth method performance exploration since the release of Bio-Rad¹⁸⁹ and 10x Genomics¹⁹⁰ scATAC-seq products in 2019. At the time, a number of open-source scATAC-seq protocols existed,^{188,242} but required significant levels of expertise (e.g. automated liquid handling for scaling) or non-commercially available components (e.g. indexed Tn5). Both BioRad and 10x's scATAC-seq solutions were more user-friendly and handled the most technically challenging aspect, microfluidic droplet generation, on benchtop devices (10x's Chromium Controller and Bio-Rad's ddSEQ). However, the Chromium controller already enjoyed a wide presence in large research centres due to the success of 10x's scRNA-seq products, and the Fluidigm C1 had fallen out of use due to throughput and per-cell price considerations. As a result, 10x dominated the commercial scATAC-seq market from 2016 onward, leading to ~1000 publications compared to ddSEQ's ~70 by 2025.

We reasoned that an exact quantification of the performance differences among commercial and non-commercial protocols could stimulate informed decision making. Several such benchmarks had been performed for scRNA-seq methods: the earliest by Wu and Quake,²⁷⁰ one by Ziegenhain and Enard,²⁷¹ and a larger, more recent study by Mereu and Heyn.²⁶⁸ In the Mereu benchmark, a reference sample was distributed to collaborating labs, which then each generated at least one scRNA-seq dataset. This reference consisted of a mixture of PBMCs, mouse colon cells, HEK293T, 3T3 and dog MDCK cells. This approach complicated genome alignment and further post-processing, and we therefore opted to limit the reference sample to a mixture of human PBMC extracted from one male and one female donor. The use of only liquid tissue also simplified sample handling: PBMC can be distributed frozen and thawed slowly before nuclei extraction. Finally, a large reference dataset of single-cell PBMC chromatin accessibility did not yet exist, ensuring that our benchmark's final dataset could also provide value as a public resource.

²⁷⁰Wu, Nat Meth, 2013.

²⁷¹Ziegenhain, Mol Cell, 2017.

Chapter 2

**HyDrop enables droplet-based single-cell
ATAC-seq and single-cell RNA-seq using
dissolvable hydrogel beads**

Hydrop enables droplet-based single-cell ATAC-seq and single-cell RNA-seq using dissolvable hydrogel beads

Florian V De Rop^{1,2}, Joy N Ismail^{1,2}, Carmen Bravo González-Blas^{1,2}, Gert J Hulselmans^{1,2}, Christopher Campbell Flerin^{1,2,3}, Jasper Janssens^{1,2}, Koen Theunis^{1,2,3}, Valerie M Christiaens^{1,2}, Jasper Wouters^{1,2}, Gabriele Marcassa^{1,4}, Joris de Wit^{1,4}, Suresh Poovathingal^{1*†}, Stein Aerts^{1,2*†}

¹VIB-KU Leuven/VIB Center for Brain & Disease Research, Leuven, Belgium;

²Laboratory of Computational Biology, Department of Human Genetics, KU Leuven, Leuven, Belgium; ³Aligning Science Across Parkinson's (ASAP) Collaborative Research Network, Chevy Chase, United States; ⁴Laboratory of Synapse Biology, Department of Neurosciences, KU Leuven, Leuven, Belgium

Abstract Single-cell RNA-seq and single-cell assay for transposase-accessible chromatin (ATAC-seq) technologies are used extensively to create cell type atlases for a wide range of organisms, tissues, and disease processes. To increase the scale of these atlases, lower the cost and pave the way for more specialized multiome assays, custom droplet microfluidics may provide solutions complementary to commercial setups. We developed HyDrop, a flexible and open-source droplet microfluidic platform encompassing three protocols. The first protocol involves creating dissolvable hydrogel beads with custom oligos that can be released in the droplets. In the second protocol, we demonstrate the use of these beads for HyDrop-ATAC, a low-cost noncommercial scATAC-seq protocol in droplets. After validating HyDrop-ATAC, we applied it to flash-frozen mouse cortex and generated 7996 high-quality single-cell chromatin accessibility profiles in a single run. In the third protocol, we adapt both the reaction chemistry and the capture sequence of the barcoded hydrogel bead to capture mRNA, and demonstrate a significant improvement in throughput and sensitivity compared to previous open-source droplet-based scRNA-seq assays (Drop-seq and inDrop). Similarly, we applied HyDrop-RNA to flash-frozen mouse cortex and generated 9508 single-cell transcriptomes closely matching reference single-cell gene expression data. Finally, we leveraged HyDrop-RNA's high capture rate to analyze a small population of fluorescence-activated cell sorted neurons from the *Drosophila* brain, confirming the protocol's applicability to low input samples and small cells. HyDrop is currently capable of generating single-cell data in high throughput and at a reduced cost compared to commercial methods, and we envision that HyDrop can be further developed to be compatible with novel (multi) omics protocols.

Editor's evaluation

De Rop et al. introduce a flexible microfluidics-based single-cell genomics technology that expands and improves previously existing custom droplet-based scRNA-seq protocols (inDrops and Drop-seq) in interesting directions: better data quality, simplified workflow, high-cell recovery, and flexibility towards other single-cell applications.

***For correspondence:**
suresh.poovathingal@kuleuven.be (SP);
stein.aerts@kuleuven.be (SA)

[†]These authors contributed equally to this work

Competing interest: The authors declare that no competing interests exist.

Funding: See page 21

Preprinted: 06 June 2021

Received: 16 September 2021

Accepted: 21 February 2022

Published: 23 February 2022

Reviewing Editor: Naama Barkai, Weizmann Institute of Science, Israel

© Copyright De Rop et al. This article is distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use and redistribution provided that the original author and source are credited.

eLife digest Scientists are now able to determine the order of chemical blocks, or nucleic acids, that make up the genetic code. These sequencing tools can be used to identify which genes are active within a biological sample. They do this by extracting and analysing open chromatin (regions of DNA that are accessible to the cell's machinery), or sequences of RNA (the molecular templates cells use to translate genes into working proteins).

Initially, most sequencing tools could only provide an 'averaged-out' profile of the genes activated in bulk pieces of tissue which contain multiple types of cell. However, advances in technology have led to new methods that can extract and analyse open chromatin or RNA from individual cells.

First, the cells are separated, via a technique called microfluidics, into tiny droplets of water along with a single bead that carries a unique barcode. The cell is then broken apart inside the droplet and the barcode within the bead gets released and attaches itself to the genetic material extracted from the cell. All the genetic material inside the droplets is then pooled together and sequenced. Researchers then use the barcode tags to identify which bits of RNA or DNA belong to each cell.

Single-cell sequencing has many advantages, including being able to pinpoint precise genetic differences between healthy and abnormal cells, and to create cell atlases of whole organisms, tissues and microbial communities. But existing methods for extracting chromatin are very expensive, and there were no openly available tools for processing thousands of cells at speed. Furthermore, while several single-cell RNA sequencing tools are already freely available, they are not very sensitive or practical to use.

Here, De Rop et al. have developed a new open-source platform called HyDrop that overcomes these barriers. The method entails a new type of barcoded bead and optimised elements of existing microfluidics protocols using open-source reagents. These changes created a more user-friendly workflow and increased sensitivity of sequencing at no additional cost.

De Rop et al. used their new platform to screen the RNA and open chromatin of thousands of individuals cells from the brains of mice and flies. HyDrop outperformed other open-source methods when working in RNA-sequencing mode. It also provides the first open-source tool for sequencing open chromatin in single cells. Further improvements are expected as researchers tweak the platform, which for now provides an affordable alternative to existing methods.

Introduction

In the past 5 years, droplet microfluidics have been applied extensively to partition single cells and sequence their nucleic acid content. Two methods that pioneered the field, inDrop (*Klein et al., 2015*) and Drop-seq (*Macosko et al., 2015*), both rely on the same working principle: individual cells are rapidly encapsulated into a nanoliter droplet together with a barcoded bead. Barcoding primers carried by the beads inside the emulsion are then used to index each individual cell's mRNA. This process occurs either inside the droplet, where the cell's mRNA is reverse transcribed using barcoded primers released by the barcoded bead (inDrop) or after emulsion breaking, where the cellular mRNA is anchored onto barcodes carried by resin beads (Drop-seq). The thousands of single-cell transcriptomes can then be processed for next-generation sequencing in a pooled manner. The high throughput and low per-cell reagent consumption associated with this approach have allowed researchers to profile gene expression of tens of thousands of single cells (*Kalish et al., 2020; Yap et al., 2021; Karaiskos et al., 2017*) at an affordable cost. However, the labor requirements of current open-source droplet microfluidic protocols combined with their limited sensitivity have hindered widespread adoption. Furthermore, academic development of droplet-microfluidic single-cell sequencing technology after inDrop and Drop-seq has been limited. To our knowledge, only one noncommercial droplet-based single-cell assay for transposase accessible chromatin (scATAC-seq) protocol has been published so far (*Chen et al., 2019*). Despite its elegant conceptual solution to droplet-based combined scATAC and scRNA-seq, the SNARE-seq protocol is still labor intensive, and the use of resin beads leads to reduced cell capture rates. These resin beads are loaded at dilute concentrations to prevent microfluidic obstruction, and as a result, many droplets are empty, leading to a low cell capture rate (~2%), and reagent waste (*Zhang et al., 2019*). Instead, deformable hydrogel beads can be stacked and loaded into droplets at a fixed rate without the risk of microfluidic failure, thereby increasing cell capture rate

to >50% (Zhang et al., 2019; Abate et al., 2009). Several commercial solutions have emerged since inDrop and Drop-seq (Zheng et al., 2017; Sarkar et al., 2018), and their application has been used to generate hundreds of thousands (Davie et al., 2018; Svensson et al., 2020) - and recently millions (Datlinger et al., 2021; Ren et al., 2021) - of single-cell transcriptomes at a high sensitivity. However, the high cost of these commercial protocols remains prohibitive for many research applications, and their fixed nature limits custom protocol development.

In order to increase the sensitivity and user-friendliness of open-source microfluidic protocols, and to provide a more flexible and open platform, we developed HyDrop, a new hydrogel-based droplet microfluidic method for high-throughput scRNA-seq or scATAC-seq. We adapted inDrop's original isothermal extension bead barcoding protocol (Zilionis et al., 2017) to a linear amplification workflow to generate more uniformly barcoded beads. Next, we applied a custom hydrogel bead production process similar to a recently published protocol (Wang et al., 2020) to generate dissolvable beads, improving barcoded primer release and diffusion. We also optimized Drop-seq's pooled template-switching reverse transcription strategy for application inside the cell/bead emulsion, and optimized the assay's sensitivity by testing several different cDNA library preparation strategies. The combination of these adaptations resulted in a significantly increased sensitivity for HyDrop-RNA compared to inDrop and Drop-seq at no additional cost, and in a more user-friendly workflow. Additionally, the dissolution of the hydrogel beads stabilizes the cell/bead emulsion during linear amplification thermocycling. This change allowed us to implement to our knowledge the first open-source single-cell ATAC-seq in droplets using hydrogel beads. We applied both technologies to mouse cerebral cortex and generated single-cell ATAC-seq and single-cell RNA-seq data that is highly concordant with reference data.

Results

Generation of dissolvable hydrogel beads with barcodes for scATAC-seq and scRNA-seq

We generated barcoded hydrogel beads that can dissolve and release their embedded barcoded oligonucleotide. Polyacrylamide beads incorporating disulfide cross-linkers and short oligonucleotide polymerase chain reaction (PCR) priming sites were generated by droplet microfluidics similar to a recently published method (Wang et al., 2020). A custom droplet microfluidic chip (Figure 3—figure supplement 1) was employed to produce beads of approximately 50 μm diameter. These hydrogel beads were then barcoded using a modified three-round split-pool linear amplification synthesis method (Klein et al., 2015; Zilionis et al., 2017), resulting in $96 \times 96 \times 96$ (884,736) barcode possibilities (see Supplementary file 1 – 'Molecular sequence description of HyDrop bead barcoding' for an in-depth visualization of the nucleic acid sequences in every step). The terminal sequence used in the final round of barcoding can be varied depending on the assay the beads will be used for (see Methods, Figure 1a). A sequence complementary to the Tn5 transposase adapter was used to capture tagmented chromatin fragments in scATAC-seq and a unique molecular identifier (UMI) (Kivioja et al., 2011)+ poly (dT) sequence was used to capture and count poly (A)+ mRNA in scRNA-seq (protocols described further). The barcoded beads were stored in a glycerol-based freezing buffer at -80°C in order to prevent loss of primers over time (Figure 1—figure supplement 1).

We validated the extension of the hydrogel bead primers using fluorescent probes complementary to the beads 3-prime terminal sequence (Zilionis et al., 2017; Figure 1b and Figure 1—figure supplement 2) and the sub-barcode purity using fluorescent probes complementary to 1 of the 96 sub-barcode possibilities (Figure 1c, Figure 1—figure supplements 2 and 3, Tables S12). These experiments showed that there was no significant loss of primers or mixing of barcodes throughout the barcoding process, and that the beads are uniform in size and primer content. Additional testing revealed that our modified linear amplification barcoding method produced more uniformly barcoded beads compared to the isothermal amplification protocol described in inDrop (Figure 1—figure supplement 2). Furthermore, the disulfide moieties incorporated in both the bead's polymer matrix and oligonucleotide linker are cleaved when exposed to reducing conditions, such as dithiothreitol (DTT). This chemical method of release is more user-friendly compared to the UV-mediated (Klein et al., 2015; Zilionis et al., 2017) primer release as the beads do not have to be shielded from light. In addition to improved primer release compared to nondissolvable beads (Figure 1—figure

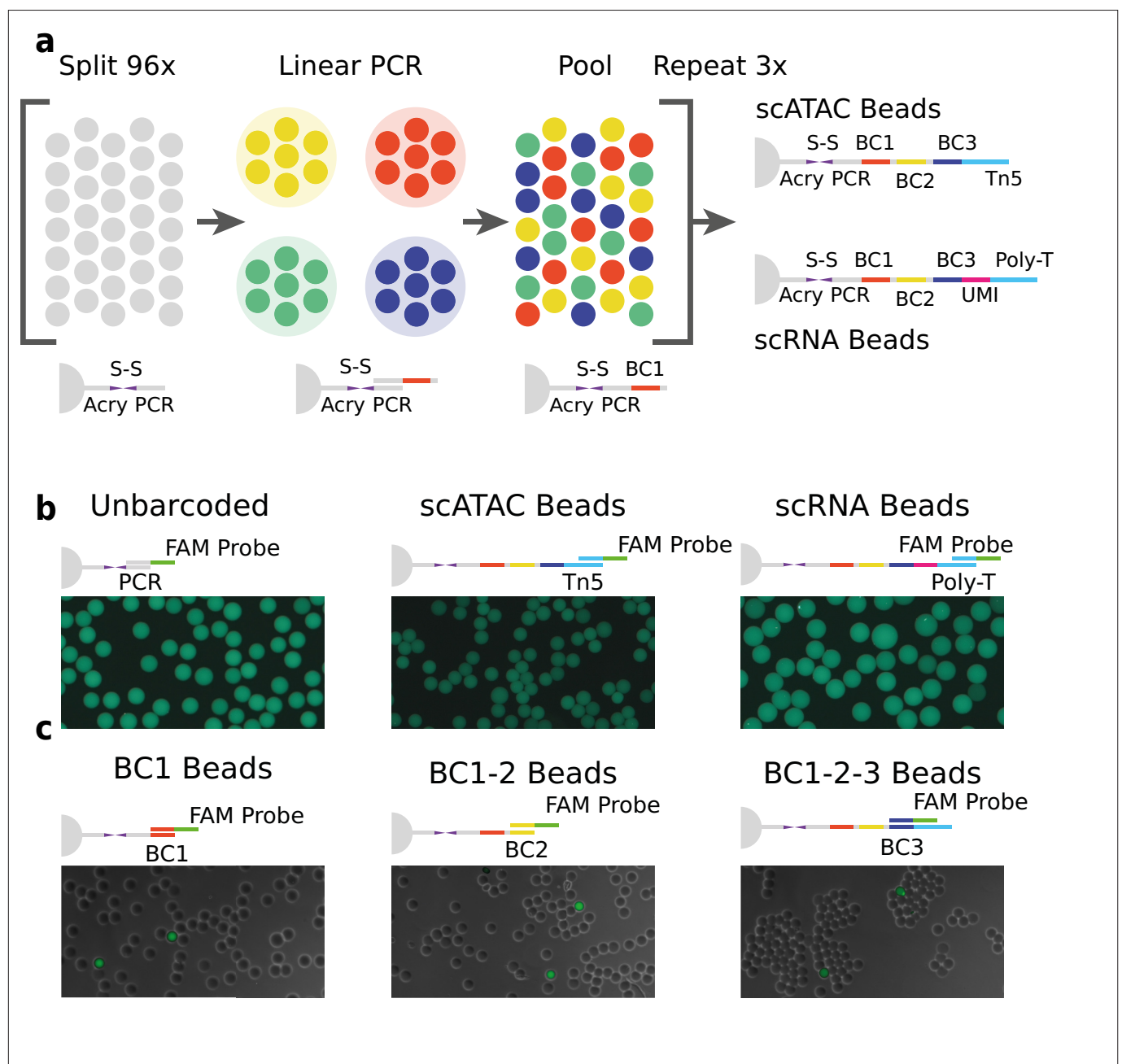


Figure 1. Technical overview of HyDrop barcoded bead production and quality control. **(a)** Split-pool process for barcoding of dissolvable hydrogel beads for single-cell RNA-sequencing (scRNA-seq) and single-cell assay for transposase-accessible chromatin (scATAC-seq). Unbarcoded hydrogel beads are sequentially distributed over 96 wells, sub-barcoded using linear amplification polymerase chain reaction (PCR), repooled, and re-distributed two more times to generate $96 \times 96 \times 96$ (884,736) possible barcode combinations. Different 3-prime terminal capture sequences are possible depending on the oligonucleotide sequence appended in the last step. **(b)** Semiquantitative assessment of bead primer incorporation by fluorescence in-situ hybridisation (FISH) using a fluorescein amidite (FAM) probe after every sub-barcoding step shows that bead fluorescence uniformity is retained throughout the barcoding process. **(c)** FISH with FAM probes complementary to only one of 96 sub-barcode possibilities shows that approximately 1/96 beads exhibit fluorescence for a selected sub-barcode probe. Fluorescence signal is overlaid with a brightfield image at 50% transparency to indicate positions of nonfluorescent beads (see **Figure 1—figure supplements 2–6** for additional quality control experiments and full images).

The online version of this article includes the following source data and figure supplement(s) for figure 1:

Figure supplement 1. Hydrogel bead integrity after being frozen and thawed.

Figure 1 continued on next page

Figure 1 continued

Figure supplement 2. Fluorescence signal of freshly barcoded beads.

Figure supplement 2—source data 1. Count, mean diameter, and mean intensity of linear amplification and isothermal extension beads.

Figure supplement 3. Brightfield images of barcoded beads incubated with FAM oligonucleotide probe complementary to one out of 96 possible sub-barcodes.

Figure supplement 3—source data 1. Total counts and fluorescence-positive counts of beads incubated with fluoresceine amidite (FAM) oligonucleotide probe complementary to 1 out of 96 barcode possibilities.

Figure supplement 4. Fluorescence signal of nondissolvable 50 μ M HyDrop-RNA beads.

Figure supplement 4—source data 1. Count, mean diameter, and mean background-adjusted intensity of beads incubated in 50 mM of DTT for 15 and 50 min as well as negative control.

Figure supplement 5. HyDrop-ATAC emulsion before and after PCR, with dissolvable and non-dissolvable barcoded hydrogel beads.

Figure supplement 6. HyDrop-RNA bead fluorescence intensity and barnyard plots for acrydite primer concentrations of 50, 12 and 6 μ M.

supplement 4, Table S3), dissolvable beads did not disrupt the emulsion during the thermocycling (**Figure 1—figure supplement 5**). We hypothesize that this effect is due to the lower physical stress generated by the dissolved hydrogel as opposed to the solid hydrogel. This added benefit allowed us to implement scATAC-seq (see further) on HyDrop. Finally, by varying the concentration of the acrydite primer during bead synthesis, lower or higher amounts of cleavable barcoded primers could be generated. We found that when the concentration of acrydite primer incorporated in the hydrogel matrix was high (50 μ M, similar to InDrop), excess unreacted barcodes could not be sufficiently filtered out in further downstream steps. These primers were then carried over to subsequent reactions, leading to random barcoding of free fragments after droplet merging, and subsequently to cell-mixed expression or chromatin accessibility profiles. The bead primer concentration with an optimal balance between sensitivity and library purity was found to be 12 μ M for both scATAC-seq and scRNA-seq (**Figure 1—figure supplement 6**).

Implementation and accuracy assessment of HyDrop-ATAC

We implemented a new open-source protocol for single-cell ATAC-seq using HyDrop's dissolvable barcoded hydrogel beads. Nuclei were Tn5 tagmented in bulk and coencapsulated with HyDrop-ATAC beads. Pitstop, a selective small molecule clathrin inhibitor, was supplemented during nuclei extraction and tagmentation to increase nucleus permeability to Tn5 (**Mulqueen et al., 2019**). Inside the droplet, the hydrogel beads dissolve and release their uniquely barcoded primers inside the droplet due to the presence of DTT carried by the nuclei/PCR mix. Subsequent thermocycling of the emulsion denatures the Tn5 protein complex and releases accessible chromatin fragments within the droplet. These fragments were then linearly amplified and cell indexed by the bead's barcoded primers after which the emulsion was broken and the indexed ATAC fragments were pooled, PCR amplified, and sequenced (**Figure 2**, see – **Supplementary file 2** 'Molecular sequence description of HyDrop-ATAC' for an in-depth visualization of the nucleic acid sequences in every step).

In order to coencapsulate beads and nuclei with a high capture rate and minimal microfluidic complexity, we developed a custom microfluidic chip (**Figure 3a and b**). The chip design features one inlet for beads, one inlet for cells or nuclei, and one inlet for the emulsion carrier oil. This configuration is slightly more convenient to operate compared to InDrop's four channel setup. Several layers of passive filters near the inlet ports mitigate dust and debris buildup during droplet generation to prevent obstruction of the channels. Beads and nuclei were loaded via a tip reservoir to reduce nonlinear flow behavior and the potential accumulation of cells/nuclei and hydrogel beads associated with narrow tubes (**Sinha et al., 2019; Hur et al., 2011; Figure 3b**). Due to the stability of all flows and the deformable nature of the hydrogel beads, >90% occupancy of hydrogel beads in droplets could be achieved (**Abate et al., 2009; Figure 3c**), resulting in a final cell recovery of ~65%.

To assess the purity of scATAC-seq libraries generated by HyDrop-ATAC, we performed two mixed species experiments. First, we generated single-nuclei ATAC-seq libraries from a 50:50 mixture of human breast cancer (MCF-7) and a mouse melanoma cell line generated previously (**Dankort et al., 2009**). We developed a custom preprocessing and mapping pipeline for HyDrop-ATAC data (**Dankort**

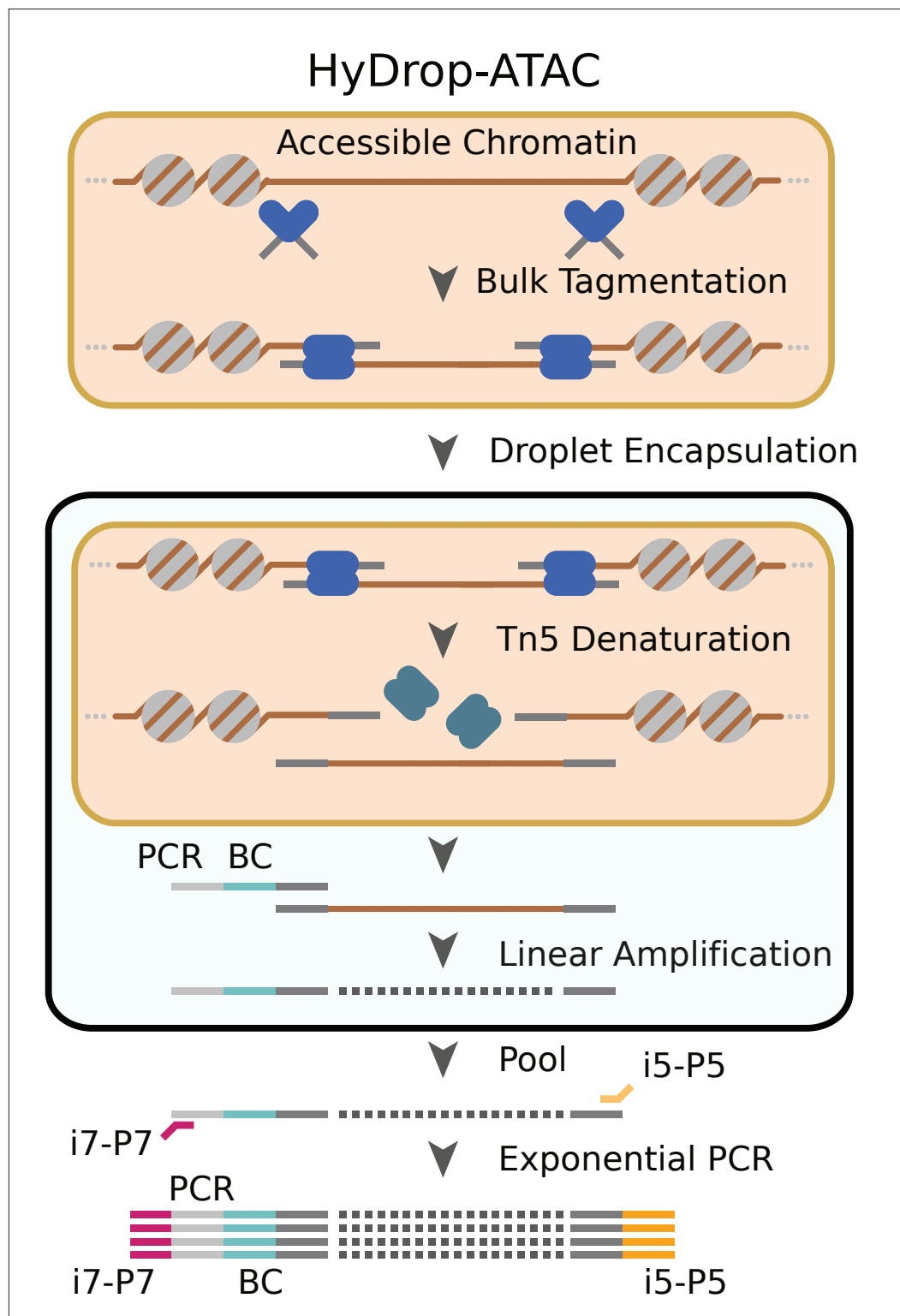


Figure 2. Schematic overview of HyDrop-ATAC. Nuclear membrane is visualized in salmon, water droplet is visualized in blue. Nuclei were Tn5 tagged in bulk and co-encapsulated with HyDrop-ATAC beads, where the hydrogel beads dissolve and release their uniquely barcoded primers. Thermocycling of the emulsion releases accessible chromatin fragments which are then linearly amplified and cell indexed by the bead's barcoded primers within the droplet. The emulsion is then broken and the indexed ATAC fragments are pooled, PCR amplified, and sequenced.

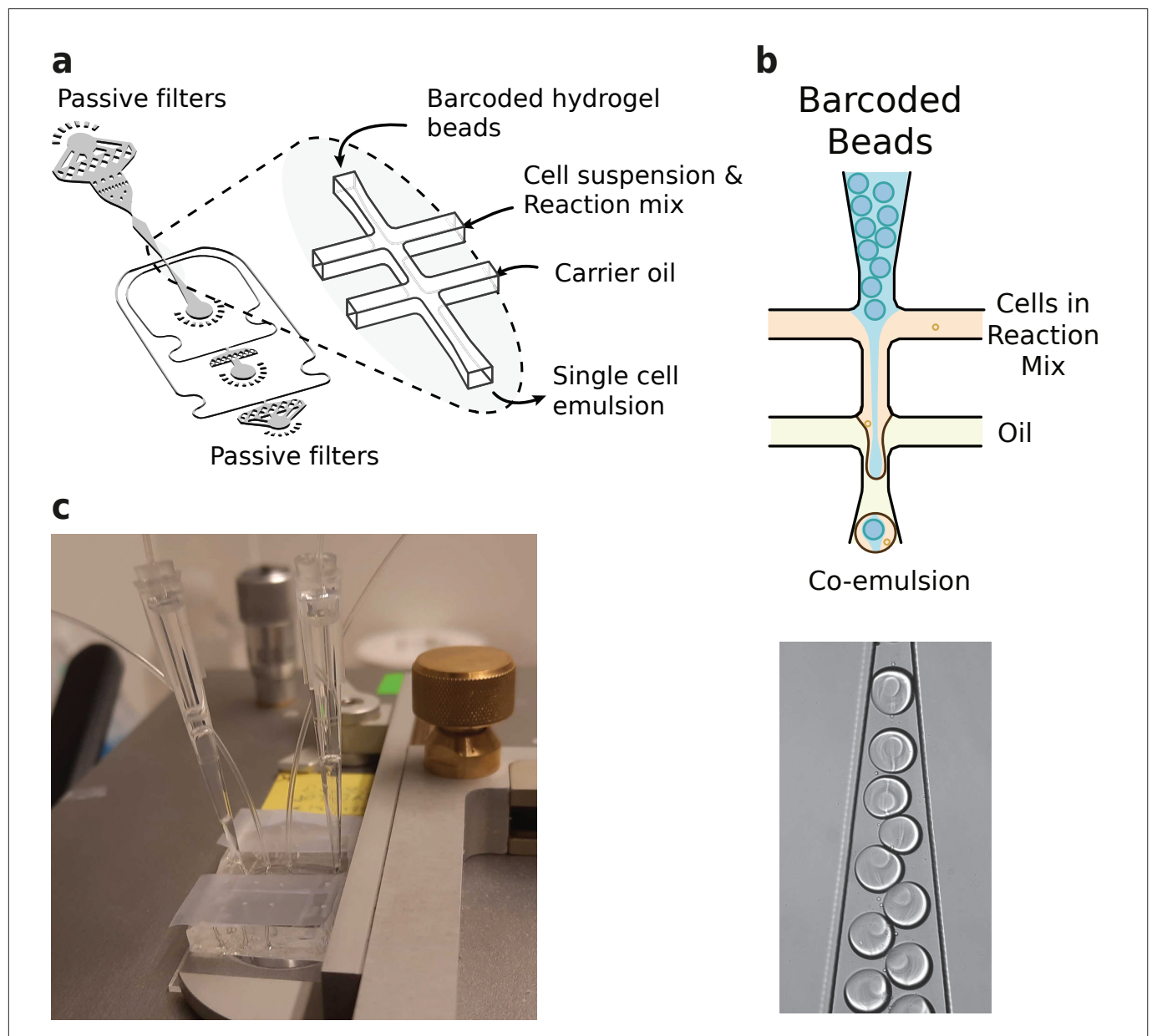


Figure 3. Microfluidics setup employed in HyDrop emulsion generation. **(a)** 3d rendering of HyDrop chip design. The design has three inlets: one each for carrier oil, barcoded hydrogel beads and cell/reaction mix (see **Figure 3—figure supplement 1** for full design). Passive filters at each inlet prevent dust and debris from entering the droplet generating junction. **(b)** Diagram and snapshot of cell/bead encapsulation into microdroplets. **(c)** Microfluidic chip setup on the Onyx integrated microfluidic instrument. Cells and beads are loaded into pipette tips and plugged into a HyDrop chip. Flow of oil and aqueous phases is achieved by Onyx displacement syringe pumps.

The online version of this article includes the following figure supplement(s) for figure 3:

Figure supplement 1. Microfluidic chip designs for cell/bead encapsulation and hydrogel bead generation - oil, monomer mix, cell suspensions, and beads can be pumped in through their respective inlets.

et al., 2009). After filtering the cell barcodes for a minimum transcription start sites (TSS) enrichment score of 7 and unique fragment count of 1000, we recovered 1353 cells from an input of 2000, and a median of 2705 unique fragments per cell. About 88% of all barcode reads were paired with the whitelist when one mismatch was allowed along the entire length of the 30 bp barcode. An additional 1.16% and 1.13% of HyDrop barcodes could be assigned to the whitelist when two and three

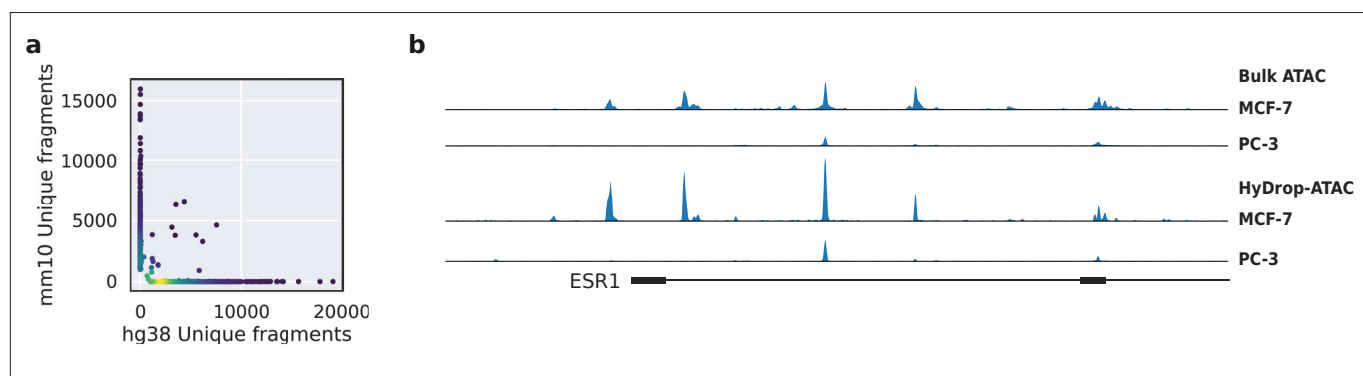


Figure 4. Validation of HyDrop-ATAC on mixed species cell lines. (a) Scatterplot of the number of unique fragments detected in a 50:50 mixture of human Michigan Cancer Foundation-7 (MCF-7) and mouse melanoma cells colored by local density estimation. (b) reads per genome coverage (RPGC)-normalized aggregate genome tracks comparing HyDrop-ATAC and bulk ATAC-seq profiles of human MCF-7 and prostate cancer-3 (PC-3) cell lines around the estrogen receptor 1 (ESR1) locus, scaled to maximum of all four samples. Aggregate enrichment profile of reads around transcription start site, see **Figure 4—figure supplement 1** for clustering of these cells and count correlation with public data. Supplementary source data files available for figure (a).

The online version of this article includes the following figure supplement(s) for figure 4:

Figure supplement 1. UMAP and count correlations within bulk ATAC-seq data for MCF-7 and PC-3 HyDrop-ATAC.

mismatches were allowed, respectively. The remaining 9% of barcode sequences could not be identified and are most likely a result of frameshift due to incorporation errors in the synthetic oligonucleotide production process. Notably, only 39.7% of barcode reads could be matched with the whitelist in a public inDrop dataset (Kalish et al., 2020), when one mismatch was allowed. When prefiltered cell barcode sequences were used as a whitelist for a public mouse retina Drop-seq dataset (Macosko et al., 2015), 49.5% of reads could be matched. We identified 98.4% of cells as either human or mouse at a minimum purity of 95% fragments mapping to either species (Figure 4a). Next, we generated libraries from a mixture of MCF-7/PC-3/Mouse cortex (45:45:10) to evaluate whether two human cell types can be distinguished. A spike-in of 10% mouse cells was used as an internal control. We recovered 2602 human cells, 466 mouse cells, and 93 species doublets after filtering for 95% species purity. Clustering human cells (together with the MCF-7 cells from the first species mixing experiment to evaluate batch effects) recovered two distinct populations, each exhibiting specific ATAC-seq peaks near MCF-7 or PC-3 marker genes (Figure 4—figure supplement 1a). Aggregated reads per cluster showed typical ATAC-seq profiles concordant with public bulk ATAC-seq data (Consortium, 2012; Figure 4b, Figure 4—figure supplement 1).

Application of HyDrop-ATAC on flash-frozen mouse cortex recapitulates cerebral cellular heterogeneity and cell-type specific accessibility profiles

To evaluate the performance of HyDrop-ATAC on primary tissue, we generated single-cell libraries from snap-frozen, dissected adult mouse brain cortex. Libraries were sequenced to approximately 75% duplication rate. After filtering for a minimum of 1000 unique nuclear fragments, a TSS enrichment score of 5, and removing 506 cells (6%) detected as doublets by Scrublet (Wolock et al., 2019), we recovered a total of 7996 single nuclei. Cells passing the filters had a median of 4148 fragments per cell, a median TSS enrichment score of 13, and a median of 53% of fragments in peaks, reflecting high-quality cells and low levels of background signal (Figure 5a–d). Even though the number of unique fragments per cell (~4 K) is lower than that of commercial methods (e.g. 17–20 K per cell for 10x Genomics, see Methods), HyDrop-ATAC achieves comparable results in terms of TSS enrichment and fraction of reads in peaks scores. We used cisTopic (Bravo González-Blas et al., 2019) to reduce the dimensionality of the dataset and the Leiden algorithm (Traag et al., 2019) to cluster cells (Figure 5e). Cell annotation using the aggregated ATAC signal around several neocortex markers (Yao et al., 2021; Zeisel et al., 2018) recovered 19 distinct cell types, similar to previously published scATAC-seq mouse cortex data (Preissl et al., 2018; Li et al., 2020; Figure 5—figure supplement

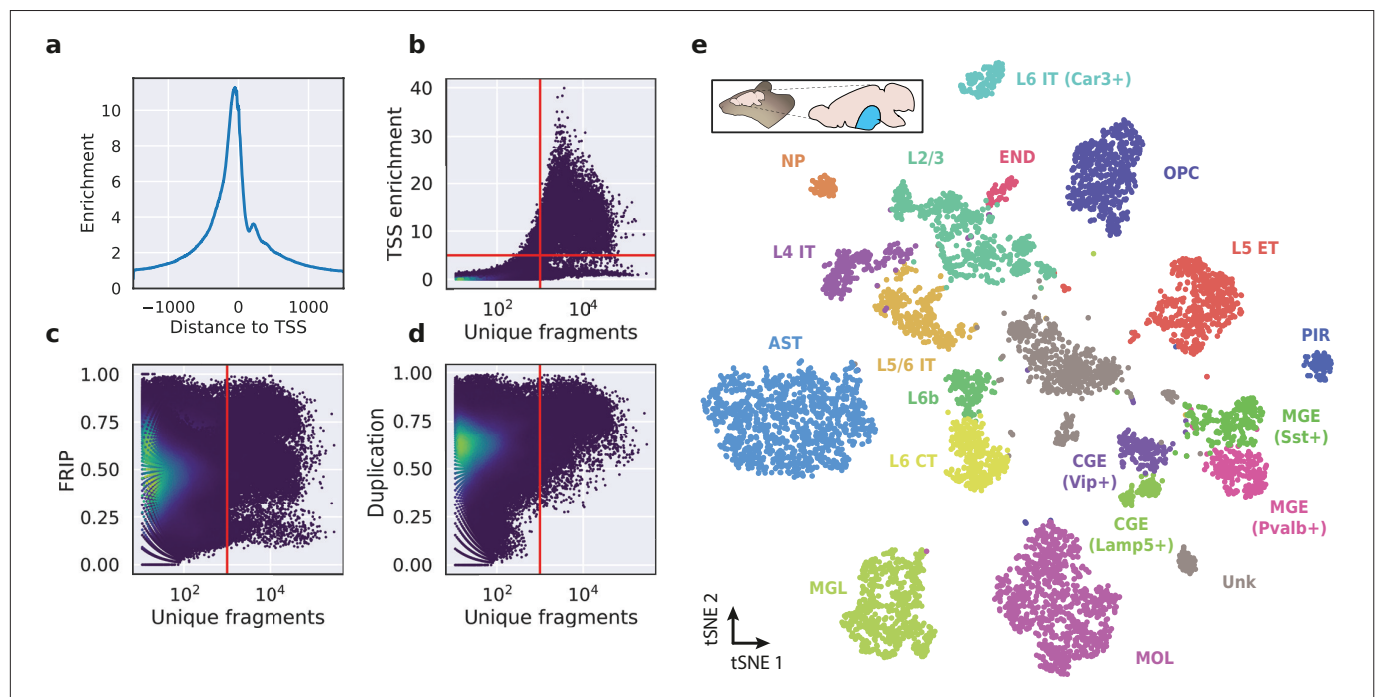


Figure 5. Application of HyDrop-ATAC on flash-frozen mouse cortex recapitulates cellular heterogeneity. Aggregate enrichment of ATAC fragments near transcription start sites (TSS) (a), TSS enrichment per barcode (b), fraction of reads in peaks (FRIP) per barcode (c) and duplication rate per barcode (d) in mouse cortex HyDrop-ATAC data. A minimum TSS enrichment of 5 and a unique number of fragments of 1000 are used as cut-off values to separate cells from background (red lines). Cells are colored by local density estimation. (e) UMAP projection of 7996 mouse cortex nuclei annotated with cell type inferred by accessibility near marker genes. Abbreviations: microglia (MGL), mature oligodendrocytes (MOL), oligodendrocyte precursors (OPC), astrocytes (AST), endothelial cells (END), piriform cortex neurons (PIR), caudal and medial ganglionic eminence derived neurons (CGE, MGE), layers 2–6 intratelencephalic (IT), L5 extratelencephalic (ET), L5/6 near projecting excitatory neurons (NP), L6 corticoencephalic (CT), and deep L6 excitatory neurons (L6b). See **Figure 5—figure supplement 1** for cluster marker gene activities. Supplementary source data files available for figures b, c, and d.

The online version of this article includes the following figure supplement(s) for figure 5:

Figure supplement 1. Heatmap of mouse cortex HyDrop-ATAC gene activity – gene activity was imputed by normalized accessibility within a 10 kb window around the gene.

1). For example, we identified oligodendrocyte precursors and mature oligodendrocytes, marked by exclusive accessibility nearby *Sox10* and *Pdgfra2*, respectively. Within ganglionic eminence-derived interneurons, we were able to further distinguish medial ganglionic eminence-derived subtypes with specific ATAC-seq signal near either *Vip* or *Lamp5*, and caudal ganglionic eminence-derived subtypes with accessibility near either *Sst* or *Pvalb*. Finally, HyDrop-ATAC data revealed distinct cell-type specific differentially accessible regions (**Figure 6a and b**).

Implementation of HyDrop-RNA as a hybrid method between inDrop and Drop-Seq

We next implemented a new scRNA-seq assay using barcoded bead primers carrying a 3-prime poly (dT) sequence. Single cells or nuclei were resuspended in a reverse transcriptase mix and coencapsulated into microdroplets with 3-prime poly (dT) HyDrop beads. The same microfluidic chip design was used for both HyDrop-RNA and HyDrop-ATAC. Cells are lysed inside the droplets upon contact with the lysis buffer in which the barcoded beads were suspended. Simultaneously, barcoded primers were released from the hydrogel bead after exposure to DTT present in the reverse transcriptase mix. Reverse transcription inside the emulsion generates thousands of barcoded single-cell cDNA libraries in parallel. The emulsion was then broken and the single-cell transcriptome libraries were processed further in a pooled manner (**Figure 7**, see **Supplementary file 3**).

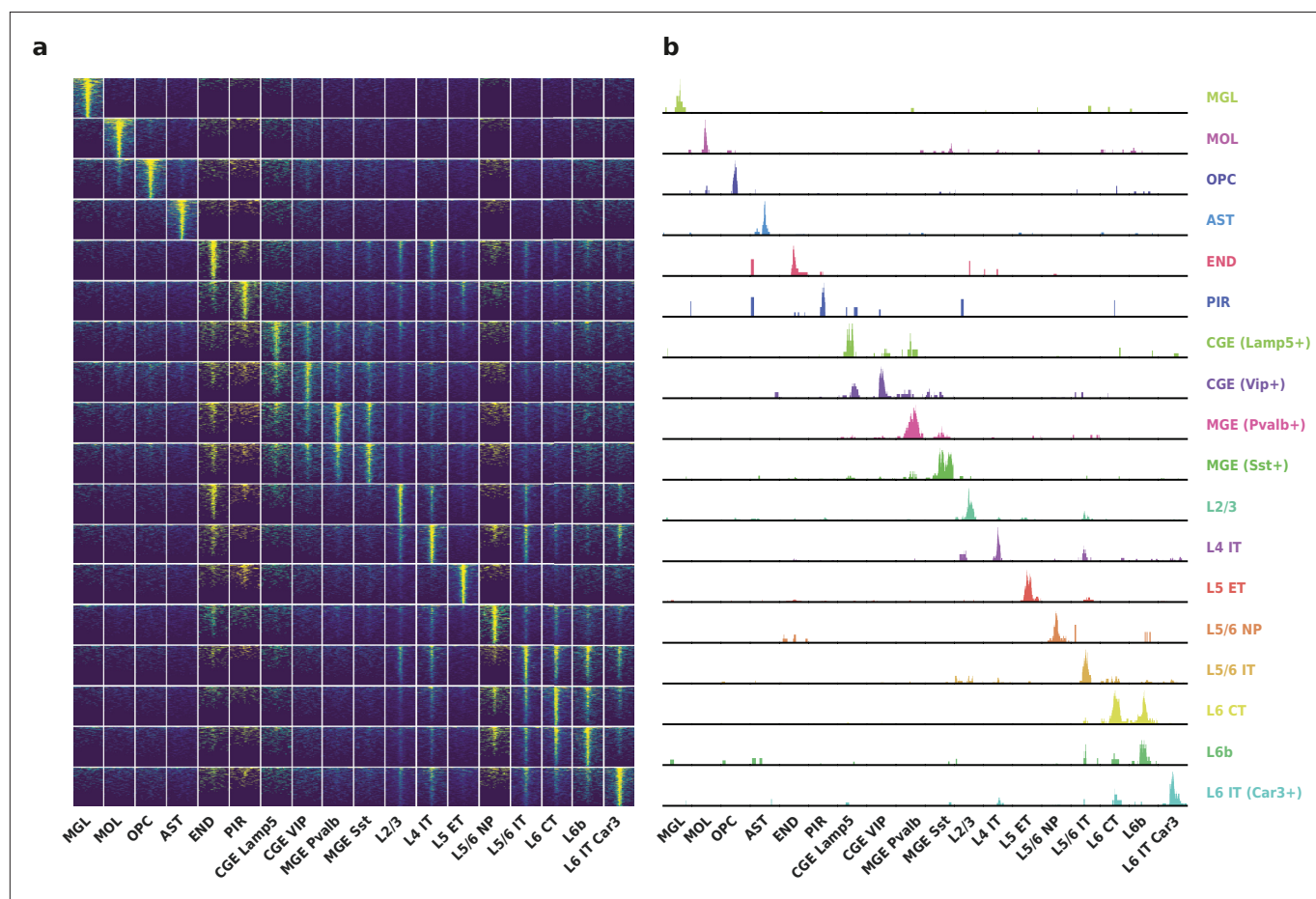


Figure 6. Differentially accessible regions (DAR) between cell types recovered by HyDrop-ATAC on mouse cerebral cortex. **(a)** Aggregate accessibility of top 1000 DAR per cluster. **(b)** Row-scaled, counts-per-million (CPM)-normalized aggregate genome track covering the top one DAR for each cluster.

'Molecular sequence description of HyDrop-RNA' for an in-depth visualization of the nucleic acid sequences in every step, similarly to the InDrop protocol (Zilionis *et al.*, 2017). Although both assays are based on hydrogel beads, HyDrop-RNA differs significantly from InDrop. HyDrop-RNA employs a template switching oligo (TSO) reverse transcription technique (similar to Drop-seq), rather than an in vitro transcription/random hexamer priming workflow. This change simplifies and speeds up the protocol significantly with no reduction in sensitivity.

To improve the assay's sensitivity, we interrogated the impact of several alterations to the protocol's reaction chemistry by testing them on a 50:50 human-mouse (human melanoma, mouse melanoma) mixture. We first investigated the use of pooled exonuclease I treatment after reverse transcription to remove unused barcode primers. We reasoned that these unused barcoded primers could potentially prime transcripts during the subsequent bulk inverse-suppressive polymerase chain reaction (IS-PCR), leading to a loss of purity of transcripts associated with a given barcode. As evident from the increase in pure cells detected, we found that exonuclease I treatment indeed improved assay purity (Figure 8—figure supplement 1). We then tested the implementation of a locked nucleic acid (LNA) in the TSO, as it has been shown to increase assay sensitivity due to increased stability of the TSO-cDNA complex (Picelli *et al.*, 2014). We also investigated the addition of guanosine triphosphate (GTP) and polyethylene glycol (PEG) to the in-droplet reverse transcription reaction. Both the addition of PEG as a molecular crowding agent and GTP as a stabilizer of the TSO complex has been shown to improve assay sensitivity in SMART-seq3 (Hagemann-Jensen *et al.*, 2020). Finally, we wondered whether a second strand synthesis library construction approach could outperform the TSO/IS-PCR approach. In order to test this, we compared both alkaline hydrolysis and enzymatic treatment (RNase H) to

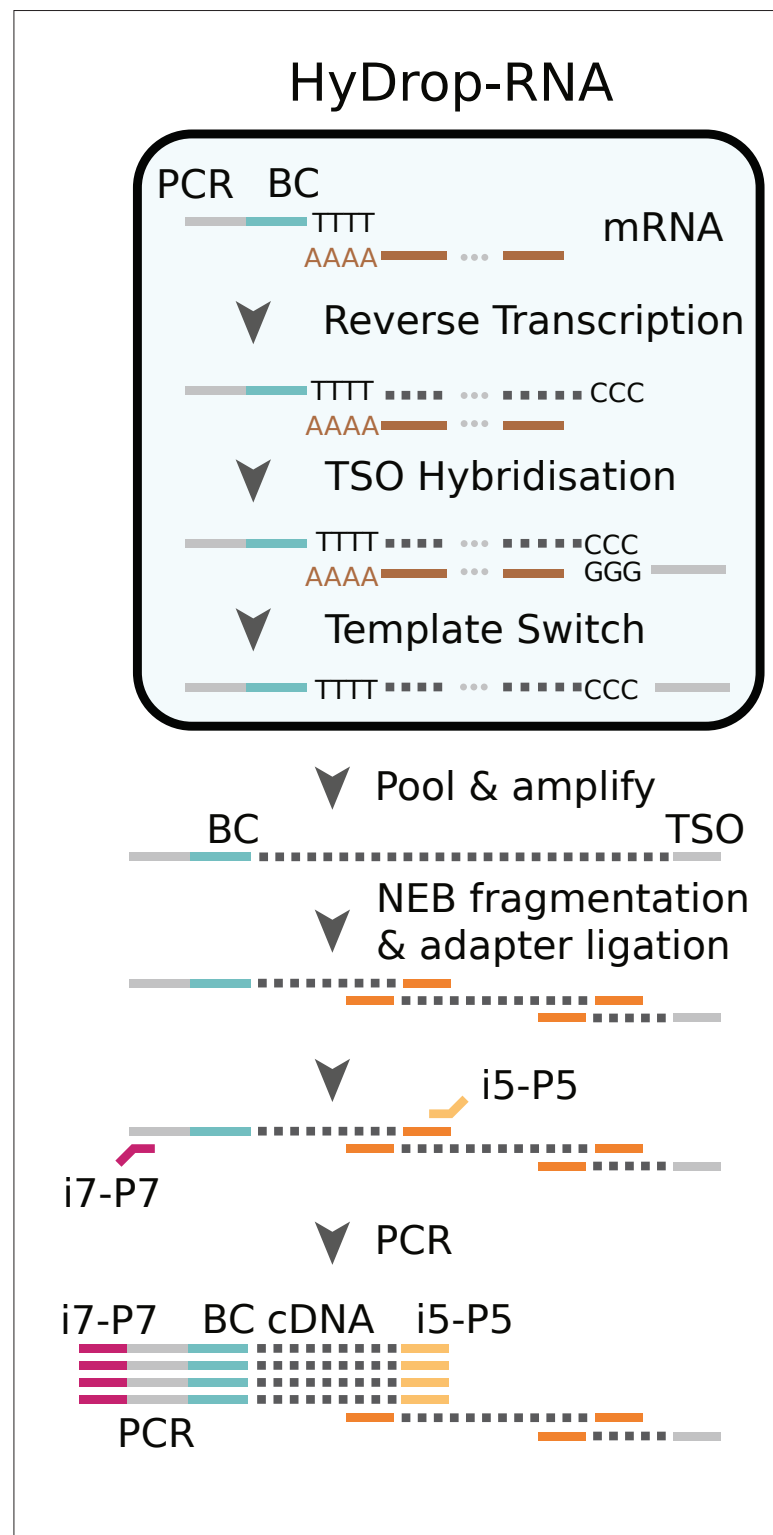


Figure 7. Schematic overview of HyDrop-ATAC. Single cells or nuclei are resuspended in a reverse transcriptase mix and coencapsulated into microdroplets with 3-prime poly (dT) HyDrop beads. Cells are lysed inside the droplets and barcoded primers (BC) are released from the hydrogel bead. Reverse transcription inside the emulsion using a template-switching oligo (TSO) generates thousands of barcoded single-cell cDNA libraries in parallel which are processed further in a pooled manner after breaking the emulsion.

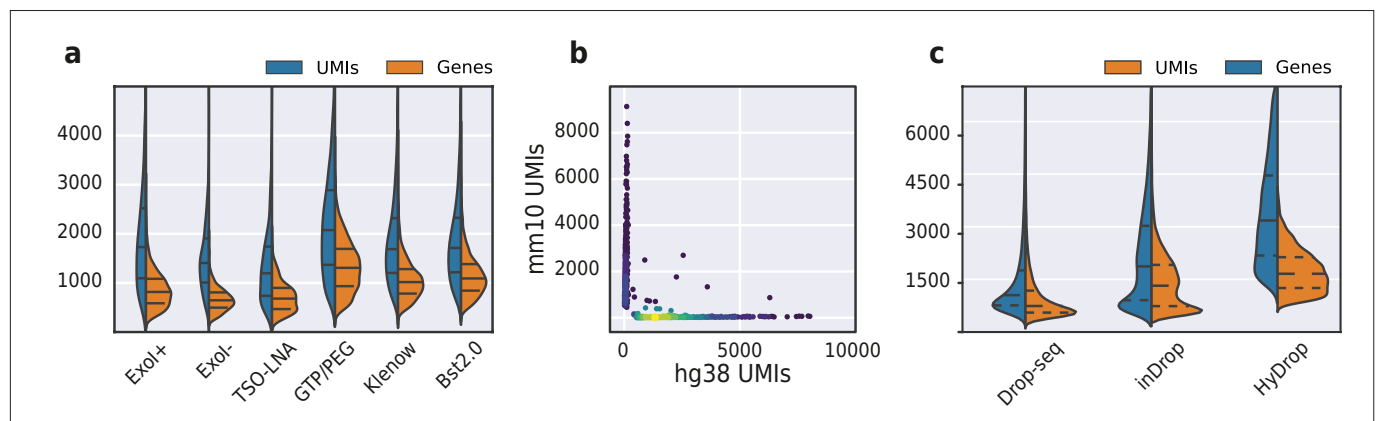


Figure 8. Validation and benchmarking of HyDrop-RNA on species mixed cell line samples and mouse cerebral cortex. **(a)** Comparison of unique molecular identifier (UMI) and gene count of HyDrop-RNA with and without Exo I treatment postdroplet merging, with the use of a locked nucleic acid (LNA) template switching oligo (TSO) and with GTP/PEG, BST2.0, and Klenow fragment library preparation. Inner lines represent Q1, median, and Q3. See **Figure 8—figure supplement 1** for species purity plots of these experiments. **(b)** Scatterplot of human and mouse UMIs detected in a 50:50 mixture of human MCF-7 and mouse melanoma cells colored by local density estimation. **(c)** Comparison of UMI and gene count of public inDrop mouse cortex data, public Drop-seq mouse retina data, and HyDrop-RNA mouse cortex data. Inner lines represent Q1, median, and Q3. See **Figure 8—figure supplement 2** for additional quality comparison including commercial methods. Supplementary source data files available for figures a, b, and c. The online version of this article includes the following figure supplement(s) for figure 8:

Figure supplement 1. HyDrop-RNA species-mixing purity plots for several different protocol versions.

Figure supplement 2. Collection of quality control metrics for inDrop, Drop-seq, HyDrop-RNA, and 10× datasets on mouse brain cells.

remove the RNA strand from the first strand product, and evaluated the performance of both Bst 2.0 DNA polymerase and Klenow (exo-) (Hughes et al., 2020; Stickels et al., 2021) fragment for second strand synthesis. We found that the classical TSO and IS-PCR protocol supplemented with GTP/PEG performed best, yielding a median of 2110 UMIs and 1325 genes per cell with a species purity of 90.1% (Figure 8a and Figure 8—figure supplement 1). Accordingly, the GTP/PEG method was used in all further HyDrop-RNA experiments. Applying this protocol to a 50:50 human-mouse (MCF-7, mouse melanoma) mixture recovered 1235 human and 846 mouse cells with 169 species doublets at a cutoff of 95% species purity, with a median of 1439 UMIs and 904 genes per cell (Figure 8b).

HyDrop-RNA on flash-frozen mouse cortex recovers cerebral cell types

We then used HyDrop-RNA to generate 9508 single-nuclei transcriptomes from snap-frozen mouse cortex in a single experiment. At a sequencing saturation of approximately 60% duplicates, we achieve a median of 3389 UMIs and 1658 genes per cell before, and 3404 UMIs and 1662 genes per cell after doublet filtering (Figure 8c), compared to the median of 1920 UMIs and 1321 genes detected by inDrop snRNA-seq on mouse auditory cortex neurons (Kalish et al., 2020) and the median of 1071 UMIs and 763 genes detected by Drop-seq on mouse retina neurons (Macosko et al., 2015). Both datasets also exhibited a lower read alignment efficiency compared to HyDrop, with Drop-seq mapping 52%/21%, and inDrop mapping 48%/28% of reads to the mouse genome/transcriptome, while HyDrop reaches 88%/55%. HyDrop's 'sequencing efficiency' is higher than both inDrop and Drop-seq's: 7.44% of HyDrop's sequenced reads end up as mapped transcripts with a UMI, whereas this metric is 3.24% for the inDrop dataset and 4.46% for the Drop-seq dataset (Figure 8—figure supplement 2). 10× Chromium v2 gene expression reference data reports a median number of genes of 775–2679 and a median number of UMIs of 1127–6570 on E18 and adult mouse brain nuclei, and a sequencing efficiency of 2.54%–20.9% (see methods). Comparing the top per cluster differentially expressed genes with markers from the Allen Brain Atlas SMART-seq data (Yao et al., 2021) revealed 30 distinct populations corresponding to previously identified cell types (Figure 9a–d, Figure 9—figure supplement 1). In addition to the major neuronal and glial populations previously detected in our HyDrop-ATAC experiment, we detect a small population of vascular leptomeningeal cells and layer two intratelencephalic neurons from the medial entorhinal area (L2 IT ENTm). We also detect both D1 and D2 medium spiny neurons as a result of residual striatal tissue and layer 3 *Scnn1a* + neurons from

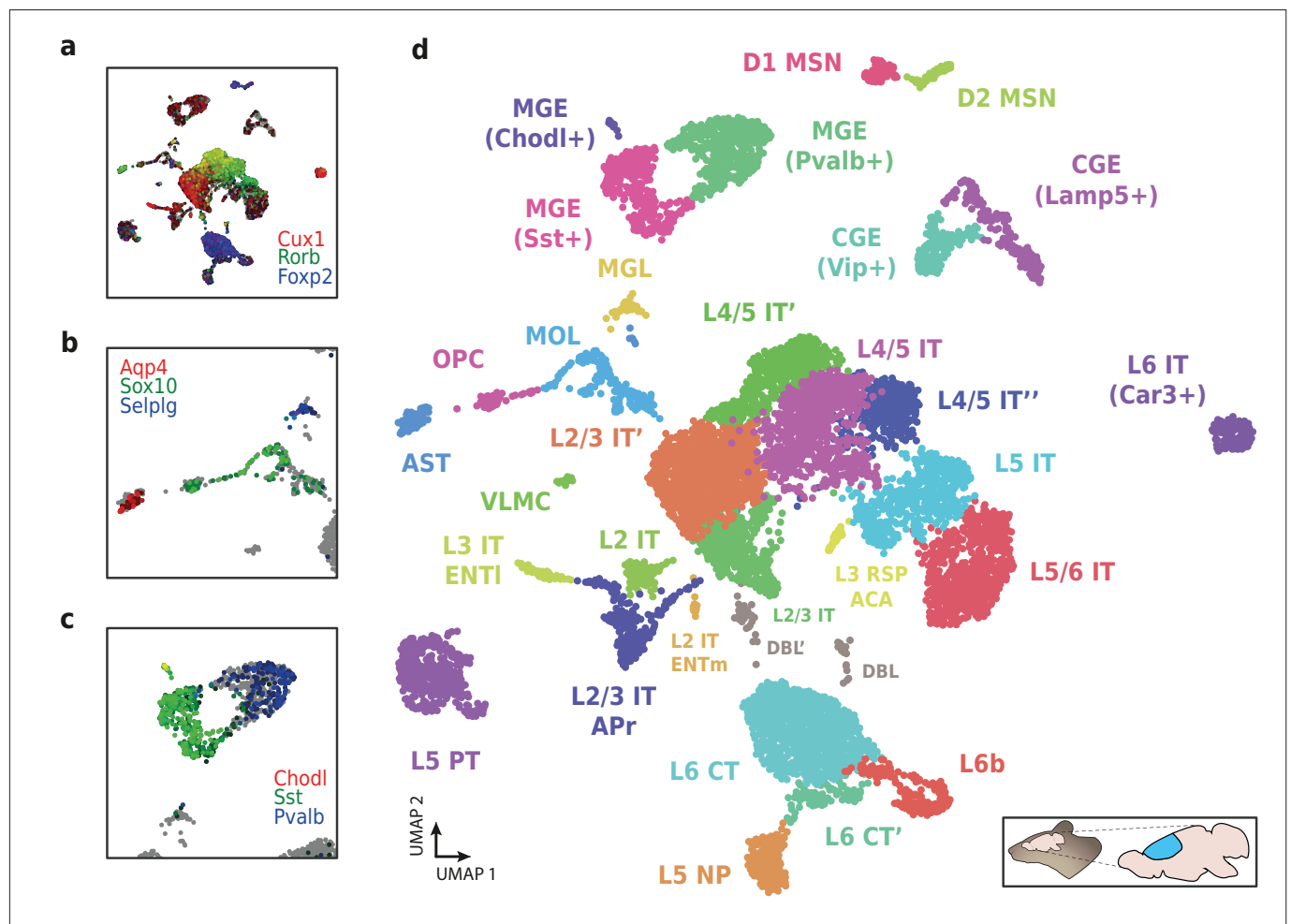


Figure 9. Application of HyDrop-RNA on flash-frozen mouse cerebral cortex. Mouse cortex UMAP colored by log-scaled unique molecular identifier counts of Cux1, Rorb, Foxp2 (a), Aqp4, Sox10, Selplg (b), Chodl, Sst, and Pvalb (c). Colors are scaled to minimum and maximum values. (d) UMAP projection of 9507 mouse cortex nuclei annotated with cell type inferred by marker gene expression. Abbreviations: microglia (MGL), mature oligodendrocytes (MOL), oligodendrocyte precursors (OPC), astrocytes (AST), endothelial cells (END), piriform cortex neurons (PIR), caudal and medial ganglionic eminence derived neurons (CGE, MGE), layers 2–6 intratelencephalic (IT), pyramidal tract (PT), near projecting excitatory neurons (NP) and corticoencephalic (CT) neurons, layer two intratelencephalic medial entorhinal area neurons (L2 IT ENTm), L2/3 intratelencephalic area prostriata neurons (L2/3 IT APPr), layer three intratelencephalic entorhinal neurons (L3 IT ENTl), layer retrosplenial and anterior cingulate area neurons (L3 RSP ACA), deep L6 excitatory neurons (L6b), D1 and D2 medium spiny neurons (MSN), and vascular leptomeningeal cells (VLMC). See **Figure 9—figure supplement 1** for expression of top differentially accessible genes within these clusters.

The online version of this article includes the following figure supplement(s) for figure 9:

Figure supplement 1. Expression of HyDrop-RNA mouse cortex top three differentially expressed genes from each cluster.

the retrosplenial and anterior cingulate area (L3 RSP ACA), concordant with Atlas SMART-seq data (Yao et al., 2021).

HyDrop-RNA on FAC-sorted *Drosophila* neurons confirms high capture rate even on small input samples

To assess HyDrop-RNA's performance on low cell input samples, we also performed the protocol on approximately 1500 FAC-sorted neurons from the *Drosophila* brain. We dissected brains in which mCherry expression was driven in specific cell populations by a Gal4 driver line (R74G01-Gal4) and used mCherry-positive cells as input for HyDrop-RNA (Figure 10a). Of the 1500 cells obtained after FACS sorting, we recovered 973 fly brain cells with a median of 1307 UMIs and 640 genes

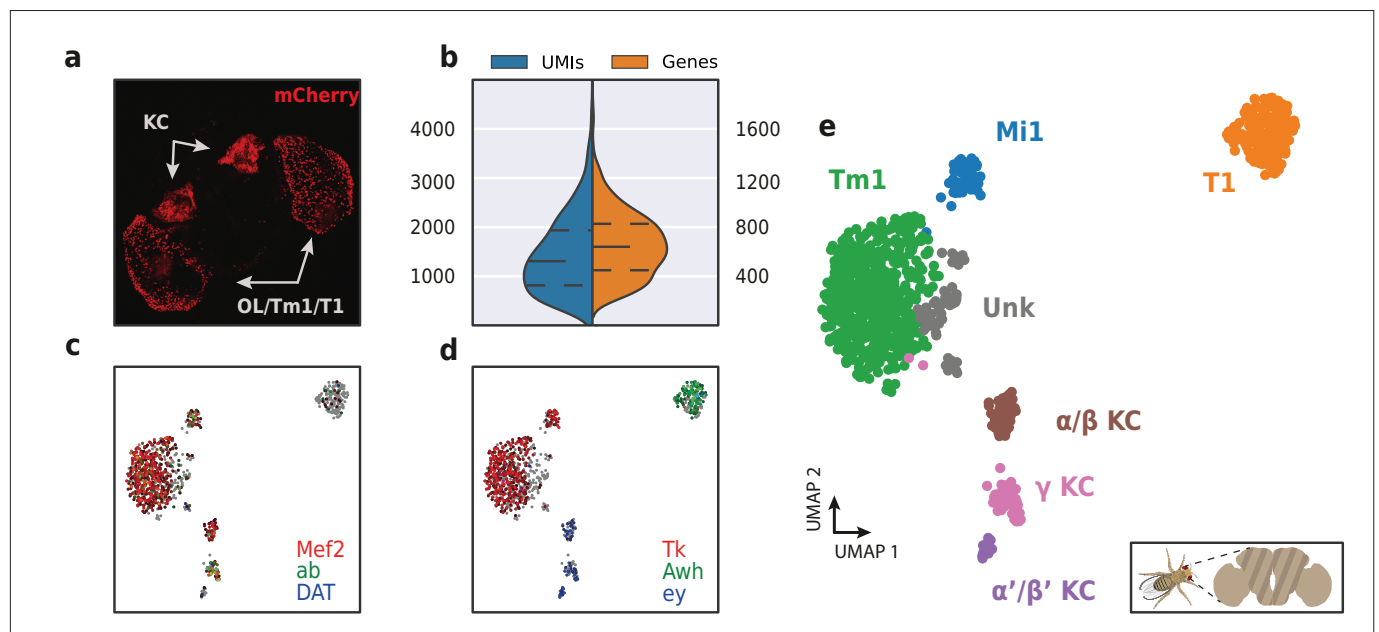


Figure 10. Application of HyDrop-RNA on FAC-sorted *Drosophila* neurons. (a) Confocal maximum intensity projection of R74G01-Gal4> UAS-mCherry brain. (b) Violin plot of unique molecular identifiers (UMIs) and genes detected in nuclei derived from FAC-sorted fly neurons. Inner lines represent Q1, median, and Q3. Fly neuron UMAP colored by log-scaled UMI counts of Mef2, ab, DAT (c) and Tk, Awh, ey (d). Colors are scaled to minimum and maximum values. (e) UMAP of 973 FAC-sorted *Drosophila* neurons annotated with cell types inferred by marker gene expression (KC, Kenyon cells; Tm1, transmedullary neuron; Mi1, medullary intrinsic neuron). See **Figure 10—figure supplement 1** for expression of marker genes within these clusters. Supplementary source data file available for figure b.

The online version of this article includes the following figure supplement(s) for figure 10:

Figure supplement 1. Expression of marker genes of HyDrop-RNA on FAC-sorted fly brain cells.

(**Figure 10b**). In comparison, in-house Drop-seq performed on the entire fly brain recovered a median of 579 UMIs and 289 genes per cell at a deeper sequencing saturation (**Davie et al., 2018**). Annotation of the 973 single-cell transcriptomes obtained by HyDrop confirmed the presence of all three Kenyon cell subtypes alongside T1 and Tm1 neurons, as expected from our stainings and previous reports (**Konstantinides et al., 2018**). Surprisingly, we also detected a small population of Mi1 neurons (**Figure 10c-e**, **Figure 10—figure supplement 1**).

Discussion

We developed HyDrop, an open-source platform for single-cell RNA- or single-cell ATAC-seq using hydrogel beads. We applied HyDrop to generate thousands of mouse, human, and *Drosophila* single-cell gene expression and chromatin accessibility profiles to demonstrate the protocol's applicability to a variety of different biological samples. HyDrop-ATAC and HyDrop-RNA experiments on mouse and fly tissues recapitulated the cellular heterogeneity of these complex samples and strongly agreed with reference datasets (**Davie et al., 2018**; **Yao et al., 2021**; **Preissl et al., 2018**; **Li et al., 2020**). We found HyDrop-RNA to outperform its open-source predecessors (**Klein et al., 2015**; **Macosko et al., 2015**; **Chen et al., 2019**) both in terms of sensitivity and user-friendliness. Moreover, at a per-cell library cost of <\$0.03 (**Supplementary file 5**) it does so at a significantly lower cost compared to commercial droplet-microfluidic alternatives (**Zheng et al., 2017**; **Satpathy et al., 2019**; **Lareau et al., 2019**).

Optimization and modification of HyDrop's reaction chemistry and bead composition may lead to further improvements in sensitivity. For example, we consider that benchmarking additional bead barcoding strategies, such as direct on-bead DNA synthesis or barcode ligation (**Delley and Abate, 2020**) instead of linear amplification may further improve overall barcode quality. Ligation strategies would also allow the incorporation of oligonucleotide modifications in the barcode's capture sequence,

a possible necessity for adaptation of HyDrop to novel assays. For example, this change would permit HyDrop's extension to protocols such as s3-ATAC (Mulqueen et al., 2021), which leverages the use of uracil bases and uracil-intolerant polymerases to double the amount of usable fragments generated by Tn5. The current 96 Kalish et al., 2020 design could also be changed to 384 Macosko et al., 2015 (like inDrop) or even 768 Macosko et al., 2015. This two-step barcoding approach would result in shorter barcode sequences, allowing for cost-saving in sequencing and increased flexibility when pooling HyDrop libraries together with other library types such as 10× Genomics 3' gene expression libraries. However, rescaling the barcodes in this manner would also necessitate the acquisition of a larger pool of sub-barcodes - from three 96 well plates to 8 and 16 plates respectively - and reduce the total barcode complexity from 884,736 to 147,456 and 589,824. Finally, one of the main challenges of implementing custom microfluidics systems is the considerable price of accurate stepper motor syringe pumps (~\$2000 per independent channel) and the need for trained personnel. However, several simplified alternatives have recently emerged (Langer et al., 2018; Kim et al., 2016), and we believe their application with HyDrop could further reduce the start-up costs and complexity of single-cell experiments.

We envision that HyDrop's reduction in both cost and labor will accelerate the scaling of large-scale atlasing efforts and bring the benefits of single-cell sequencing to smaller projects. Additionally, HyDrop's dissolvable bead synthesis and linear amplification barcoding toolkit could potentially be exploited to produce more complex beads incorporating multiple capture sequences, leading to implementation of single-cell (multi) omics assays such as the capture of accessible chromatin, (m) RNA, and surface epitopes (Stoeckius et al., 2017) or intracellular proteins (Rivello et al., 2020) from single cells.

Materials and methods

Microfluidic droplet generator manufacturing

Microfluidic droplet generators were produced using standard SU-8 lithography and polydimethylsiloxane (PDMS) lithography according to well-established protocols (Xia and Whitesides, 1998). Briefly, the designs for droplet generators were made in AutoCAD R2014 and the designs are printed onto a chrome mask using a laser writer. The SU-8 lithography is performed on a 4-inch silicon wafer using SU 8-2050 (Microchem) negative photoresist using UV aligner (EVG-620). As per manufacturers' recommendation, spin coating of the wafer with SU8 was performed at 500 rpm (ramp 100 rpm/s) - 10 s and 2000 rpm (ramp: 300 rpm/s) - 30 s, to achieve a feature height of 70–80 µm. For preparing the PDMS chip, a mixture of PDMS monomer and cross-linker (Dow Corning SYLGARD 184) was prepared at a ratio of 10:1 and mixed thoroughly. The mixture was degassed in vacuum for 45 min and poured onto an SU-8 master template and baked at 80°C for 4 hr. Inlet ports were cut using a 1-mm biopsy needle after which the chips were exposed to high-voltage plasma for 30 s and bonded onto a glass slide. 5 µL of 2% Trichloro (1 H, 1H, 2H, 2H-perfluorooctyl)silane in hydrofluoroether (HFE) was injected into each channel, incubated for 10 min at room temperature and excess oil was removed by applying pressurized air. Chips were finally baked at 100°C for 2 hr (more detailed methods for photolithography and PDMS lithography in <https://doi.org/10.17504/protocols.io.bvpin5ke>).

Barcoded hydrogel bead manufacturing and storage

Dissolvable hydrogel beads are synthesized similar to a previously published protocol Ren et al., 2021 and barcoded according to a modified inDrop protocol (Zilionis et al., 2017). For synthesizing 2–3 mL batch of beads, 2 mL of Bead Monomer Mix (6% acrylamide, 0.55% bisacryloylcystoamine, 10% Tris-buffered saline with EDTA and Triton X-100 (TBSET) (10 mM Tris-HCl pH 8, 137 mM NaCl, 2.7 mM KCl, 10 mM EDTA, 0.1% Triton X-100), 12 µM acrydite primer, 0.6% ammonium persulfate) was encapsulated into 50 µm diameter droplets in HFE-7500 Novac oil with EA-008 surfactant (RAN Biotech). 1 mL aliquots of the resulting emulsion was layered with 400 µL of mineral oil and incubated at 65°C for 14 hr. Excess mineral oil and the emulsion oil were removed and 2–3 washes with 1 mL of droplet breaking solution (20% PFO in HFE) were performed. Beads were pelleted at 5000 xg, 4°C for 30 s and washed twice in 1 mL of 1% SPAN-80 in hexane. Beads were sequentially washed in TBSET until all hexane phase was removed.

Beads were washed twice in Bead Wash Buffer (10 mM Tris-HCl pH 8, 0.1% Tween-20), twice in PCR Wash Buffer (10 mM Tris-HCl pH8, 50 mM KCl, 1.5 mM MgCl₂, 0.1% Tween-20). The subsequent liquid handling in the 96-well plate is performed using Hamilton microlab STAR robot. 22.5 µL of beads were distributed to a 96-well plate. 2.5 µL of 100 µM sub-barcode primer and 25 µL of KAPA HiFi Hotstart master mix (Roche) were added to each well and the plate was thermocycled (95°C 3 min, 5 cycles of 98°C 20 s, 38°C 4 min, 72°C 2 min, 1 cycle of 98°C 1 min, 38°C 10 min, 72°C 4 min, followed by a final hold on 4°C) with intermittent vortexing during every annealing step. 50 µL of STOP-25 (10 mM Tris-HCl pH 8, 25 mM EDTA, 0.1% Tween-20, 100 mM KCl) was added to each well to deactivate the polymerase and its contents were pooled. Remaining beads in wells were washed out with 100 µL of STOP-25 and the beads were rotated at room temperature for 30 min. Beads were then washed with STOP-10 (10 mM Tris-HCl pH 8, 10 mM EDTA, 0.1% Tween-20, 100 mM KCl) and rotated for 10 min in denaturation solution (150 mM NaOH, 85 mM BRIJ-35). Beads were washed twice in denaturation solution and twice more in neutralization solution (100 mM Tris-HCl pH 8, 10 mM EDTA, 0.1% Tween-20, 100 mM NaCl). The sub-barcoding step was repeated twice more for a total of three sub-barcodes.

Hydrogel beads were sequentially filtered using a 70 µm strainer (Falcon). For both the HyDrop-ATAC and RNA beads were pelleted at 300 xg, 4°C and resuspended in 5 mL of Bead Freezing Buffer (150 mM NaCl, 125 mM Tris-HCl pH 7, 10 mM MgCl₂, 4% Tween-20, 0.75% Triton X-100, 30% glycerol, 0.3% BSA). Beads were pelleted at 300 xg, 4°C and resuspended in 5 mL of Bead Freezing Buffer a second time and incubated at 4°C for at least 3 hr. Beads were pelleted at 1000 xg, 4°C and the pellet was aliquoted into 35 µL aliquots and stored at -80°C for long-term storage (further method details in <https://doi.org/10.17504/protocols.io.b4cyqsxw>).

Hydrogel bead fluorescence in situ hybridization quality control

Bead QC was performed as described previously (Zilionis *et al.*, 2017). Briefly, 10 µL of hydrogel beads were resuspended in 1 mL of hybridization buffer (5 mM Tris-HCl pH 8.0, 5 mM EDTA, 0.05% Tween-20, 1 M KCl) and centrifuged for 1 min at 1000 xg. The wash was repeated once more, and 960 µL of the supernatant was removed. 4 µL of 200 µM specific FAM probe was added depending on which part of the barcode needed testing (see supplementary oligonucleotide table). The beads were incubated at room temperature in the dark for 30 min. Beads were washed thrice in QC buffer and visualized under a Zeiss Axioplan two microscope and a Zeiss Colibri seven light source (300 ms exposure time, 80% light source intensity). Occasional punctate intensities (diameter ~1 µm) were observed in quality control performed on 10x beads and are thought to be dye precipitates or crystals.

Fluorescence images were analyzed using opencv2 (Bradski, 2015). Briefly, foreground and background were determined using Otsu's method. Circles were detected in the foreground using Hough's algorithm. The mean grayscale intensity of the background was subtracted from the grayscale intensity at the center of each detected circle. The center intensity method was favored over a volumetric intensity measurement over the entire bead as slight differences in focus between images can affect intensity near the edge of the bead. For the 1/96 sub-barcode experiment, both fluorescence and brightfield images were taken of the same beads incubated with 1 out of 96 possible sub-barcodes. Total number of beads was counted manually from the brightfield image, while positive beads were counted using the image analysis tool from the fluorescence images. A mask of the positive beads from the fluorescence image was overlaid onto the brightfield image to mark positive beads for visualization. Image analysis tool code and raw data of all used images available at https://github.com/aertslab/hydrop_data_analysis, (copy archived at [swh:1:rev:059bf5a7779dc2894670ecf5f820c14bceb68493](https://www.swh.io/rev/059bf5a7779dc2894670ecf5f820c14bceb68493); Rop, 2022 and <https://doi.org/10.5281/zenodo.6415968>).

Cell culture and cell dissociation

MCF-7 cells (NCI-DTP Cat# MCF7, RRID:CVCL_0031) were cultured in RPMI1640 (ThermoFisher 11875093) medium supplemented with 10% FBS (ThermoFisher 10270-106), 1% penicillin/streptomycin (Life Technologies 15140122), and 10 µg/mL insulin (Sigma Aldrich I9278) and passaged twice per week. PC-3 cells (NCI-DTP Cat# PC-3, RRID:CVCL_0035) were cultured in RPMI1640 medium supplemented with 10% FBS and 1% penicillin/streptomycin and passaged twice per week. Mouse melanoma cells were cultured in DMEM (ThermoFisher 13345364) supplemented with 10% FBS and 1% penicillin/streptomycin and passaged once per week. MM087 melanoma cells were cultured in F-10 Nutrient mix supplemented with 10% FBS and 1% penicillin/streptomycin and passaged once

per week. All cell lines tested negative for mycoplasma prior to use. The identity of PC-3, MCF-7, and mouse melanoma cell lines were correctly identified since their aggregate scATAC-seq data correlated highly with public data (see **Figure 4b**, **Figure 4—figure supplement 1** for PC-3 and MCF-7, mouse melanoma data not shown). The identity of our MM087 was confirmed by SNP comparison with the MM087 genome for our manuscript 'Robust gene expression programs underlie recurrent cell states and phenotype switching in melanoma' (Wouters et al., <https://doi.org/10.1038/s41556-020-0547-3>).

Cells were washed in PBS and dissociated into single-cell suspensions by adding 1.5 mL of 0.05% trypsin (Life Technologies 25300054) and waiting for 5–7 min. The single-cell suspension was centrifuged at 500 rcf for 5 min at 4°C and the resulting pellet was resuspended in PBS. This PBS wash was repeated once more and the single-cell suspension was processed further.

Fly rearing and cell dissociation

GMR74G01-Gal4 (BDSC Cat# 39868, RRID:BDSC_39868) and UAS-mCherry (BL#38425) flies were obtained from Bloomington *Drosophila* Stock Center. The resulting cross-strain was raised on standard cornmeal-agar medium at 25°C at a 12 hr light/dark cycle. Fifty adult brains were dissected in DPBS and transferred to a tube containing 100 µL of cold DPBS solution. Samples were centrifuged at 500 rcf for 1 min and the supernatant was replaced by 50 µL of dispase (3 mg/mL, Sigma-Aldrich, D4818, 2 mg) and 75 µL of collagenase I (100 mg/mL, Invitrogen, 17100–017). Brains were dissociated in a Thermoshaker (Grant Bio PCMT) at 500 rpm for 2 hr at 25°C, with pipette mixing every 15 min. Cells were subsequently washed with 1000 µL of cold DPBS solution and centrifuged at 500 rcf for 5 min at 4°C and resuspended in 250 µL of DPBS with 0.04% BSA. Cell suspensions were passed through a 10 µm pluriStrainer (ImTec Diagnostics, 435001050). Cells were sorted based on viability and mCherry positivity using the Sony MA900 cell sorter. Sorted cells were collected into Eppendorf tubes precoated with 1% BSA.

Cell line nuclei extraction

A pellet of one million dissociated cells or fewer was incubated on ice in 200 µL of ATAC lysis buffer (1% BSA, 10 mM Tris-HCl pH 7.5, 10 mM NaCl, 0.1% Tween-20, 0.1% NP-40, 3 mM MgCl₂, 70 µM Pitstop in DMSO, 0.01% digitonin) for 5–7 min. 1 mL of ATAC nuclei wash buffer (1% BSA, 10 mM Tris-HCl pH 7.5, 0.1% Tween-20, 10 mM NaCl, 3 mM MgCl₂) was added and the nuclei were pelleted at 500 xg, 4°C for 5 min. The resulting pellet was resuspended in 100 µL of ice-cold PBS and filtered with a 40 µm strainer (Flowmi).

Mouse cortex dissection

All animal experiments were conducted according to the KU Leuven ethical guidelines and approved by the KU Leuven Ethical Committee for Animal Experimentation (approved protocol numbers ECD P183/2017). Mice were maintained in a specific pathogen-free facility under standard housing conditions with continuous access to food and water. Mice used in the study were 57 days old and were maintained on 14 hr light, 10 hr dark light cycle from 7 to 21 hr. In this study, cortical brain tissue from female P57 BL/6Jax was used. Animals were anesthetized with isoflurane, and decapitated. Cortices were collected, divided in four equal quadrants along the dorsoventral and anterior-posterior axis, and immediately snap frozen in liquid nitrogen. For HyDrop-ATAC, a ventral/posterior quadrant from the left hemisphere was used. For HyDrop-RNA, a dorsal/anterior quadrant was used from the left hemisphere of a second mouse.

Snap-frozen mouse cortex nuclei extraction

For the preparation of nuclei for scRNA-seq, we used a modified protocol from the recently published single-nuclei preparation toolbox **Delley and Abate, 2020** to extract nuclei from snap-frozen mouse cortex samples. Briefly, a ~1 cm³ frozen piece of mouse cortex tissue was transferred to 0.5 mL of ice-cold homogenization buffer (Salt-tris solution - 146 mM NaCl, 10 mM Tris 7.5, 1 mM CaCl₂, 21 mM MgCl₂, 250 mM sucrose, 0.03% Tween-20, 0.01% BSA, 25 mM KCl, 1 mM 2-mercaptoethanol, 1× complete protease inhibitor, 0.5 U/ul of RNase In Plus (Promega)) in a Dounce homogenizer mortar and thawed for 2 min. The tissue was homogenized with 10 strokes of pestle A and 5 strokes of pestle B until a homogeneous nuclei suspension was achieved. The resulting homogenate was filtered through a 70-µm cell strainer (Corning). The homogenizer and the filter are rinsed with an additional

500 μ L of homogenization buffer. The tissue material was pelleted at 500 xg and the supernatant was discarded. The tissue pellet was resuspended in a homogenization buffer without Tween-20. An addition 1.65 ml of homogenization buffer was topped up and mixed with 2.65 ml of gradient medium (75 mM sucrose, 1 mM CaCl_2 , 50% Optiprep, 5 mM MgCl_2 , 10 mM Tris 7.5, 1 mM 2-mercaptoethanol 1 $\times$ complete protease inhibitor, 0.5 U/ μ L of RNase In Plus (Promega)). 4 mL of 29% iodoxanol cushion was prepared with a diluent medium (250 mM sucrose, 150 mM KCl, 30 mM MgCl_2 , 60 mM Tris 8) and was loaded into an ultracentrifuge tube. 5.3 mL of sample in homogenization buffer + gradient medium was gently layered on top of the 29% iodoxanol cushion. Sample was centrifuged at 7700 xg, 4°C for 30 min and the supernatant was gently removed without disturbing the nuclei pellet. Nuclei were resuspended in 100 μ L of nuclei buffer (1% BSA in PBS + 1 U/ μ L of RNase Inhibitor).

For the preparation of nuclei for ATAC-seq, we used a slightly modified protocol to extract nuclei from snap-frozen mouse cortex samples. Briefly, a $\sim 1 \text{ cm}^3$ frozen piece of mouse cortex tissue was transferred to 1 mL of ice-cold homogenization buffer (320 mM sucrose, 10 mM NaCl, 3 mM Mg (OAc), 10 mM Tris 7.5, 0.1 mM EDTA, 0.1% IGEPAL-CA360, 1 $\times$ complete protease inhibitor and 1 mM DTT) in a Dounce homogenizer mortar and thawed for 2 min. The tissue was homogenized with 10 strokes of pestle A and 5 strokes of pestle B until a homogeneous nuclei suspension was achieved. The resulting homogenate was filtered through a 70 μ m cell strainer (Corning). 2.65 mL of ice-cold gradient medium was added to 2.65 mL of homogenate and mixed well. 4 mL of 29% iodoxanol cushion (129.2 mM sucrose, 77.5 mM KCl, 15.5 mM MgCl_2 , 31 mM Tris-HCl pH 7.5, 29% iodoxanol) was loaded into ultracentrifuge tube. 5.3 mL of sample in homogenization buffer + gradient medium was gently layered on top of the 29% iodoxanol cushion. Sample was centrifuged at 7700 xg, 4°C for 30 min and the supernatant was gently removed without disturbing the nuclei pellet. Nuclei were resuspended in 100 μ L of nuclei buffer (1% BSA in PBS).

HyDrop-ATAC library preparation

A total of 50,000 nuclei were resuspended in 50 μ L of ATAC reaction mix (10% DMF, 10% Tris-HCl pH 7.4, 5 mM MgCl_2 , 5 ng/ μ L Tn5, 70 μ M Pitstop in DMSO, 0.1% Tween-20, 0.01% digitonin) and incubated at 37°C for 1 hr. 100 μ L of 0.1% BSA in PBS was added and the nuclei were pelleted at 500 xg, 4°C for 5 min and resuspended in 40 μ L of 0.1% BSA in PBS.

Tagmented nuclei were added to 100 μ L of PCR mix (1.3 $\times$ Phusion HF Buffer, 15% Optiprep, 1.3 mM dNTPs, 39 mM DTT, 0.065 U/ μ L Phusion HF polymerase, 0.065 U/ μ L Deep Vent polymerase, 0.013 U/ μ L ET SSB). PCR mix was coencapsulated with 35 μ L of freshly thawed HyDrop-ATAC beads in HFE-7500 Novac oil with EA-008 surfactant (RAN Biotech) on the Onyx microfluidics platform (Droplet Genomics). The resulting emulsion was collected in aliquots of 50 μ L total volume and thermocycled according to the linear amplification program (72°C 15 min, 98°C 3 min, 13 amplification cycles of [98°C 10 s, 63°C 30 s, 72°C 1 min], followed by a final hold on 4°C). 125 μ L of recovery agent (20% PFO in HFE), 55 μ L of GITC buffer (5 M GITC, 25 mM EDTA, 50 mM Tris-HCl pH 7.4) and 5 μ L of 1 M DTT was added to each separate aliquot of 50 μ L thermocycled emulsion and incubated on ice for 5 min. 5 μ L of Dynabeads was added to the aqueous phase and incubated for 10 min. Dynabeads were pelleted on a Nd magnet and washed twice with 80% EtOH. Elution was performed in 50 μ L of EB-DTT-Tween (10 mM DTT, 0.1% Tween-20 in EB (10 mM Tris-HCl, pH 8.5)). A 1 $\times$ Ampure bead purification was performed according to manufacturer's recommendations. Elution was performed in 30 μ L of EB-DTT (10 mM DTT in EB). Eluted library was further amplified in 100 μ L of PCR mix (1 $\times$ KAPA HiFi, 1 μ M index i7 primer, 1 μ M index i5 primer). Final library was purified in a 0.4 $\times$ –1.2 $\times$ double-sided Ampure purification and eluted in 25 μ L of EB-DTT (10 mM DTT in EB). User-friendly protocol available on <https://www.protocols.io/view/hydrop-atac-v1-0-bxsbpnan>. A detailed cost analysis of HyDrop-ATAC can be found in supplementary sheet 2.

HyDrop-RNA single-cell library preparation

For a recovery of 2000 cells, 3795 cells were resuspended in 85 μ L of RT mix (1 $\times$ Maxima RT Buffer, 0.9 mM dNTPs, 25 mM DTT, 1.3 mM GTP, 15% Optiprep, 1.3 U/ μ L RNase inhibitor, 15 U/ μ L Maxima hRT, 12.5 μ M TSO, 4.4% PEG-8000). RT mix was co-encapsulated with 35 μ L of freshly thawed HyDrop-RNA beads in RAN oil on the Onyx microfluidics platform. The resulting emulsion was collected in aliquots of 50 μ L total volume and thermocycled according to the RT program (42°C for 90 min, 11 cycles of [50°C for 2 min, 42°C for 2 min], 85°C for 5 min, followed by a final hold on

4°C). 125 µL of recovery agent (20% PFO in HFE), 55 µL of GITC Buffer (5 M GITC, 25 mM EDTA, 50 mM Tris-HCl pH 7.4) and 5 µL of 1 M DTT was added to each separate aliquot of 50 µL thermocycled emulsion and incubated on ice for 5 min. 99 µL of Ampure XP beads was added to the aqueous phase and incubated for 10 min. Ampure beads were pelleted on a Nd magnet and washed twice with 80% EtOH. Elution was performed in 30 µL of EB-DTT-Tween (10 mM DTT, 0.1% Tween-20 in EB). Exonuclease treatment was performed by adding 4 µL of 10× NEBuffer 3.1, 4 µL of Exo I, and 2 µL of dH₂O to 30 µL of eluted library. The Exo I reaction mix was incubated at 37°C for 5 min, 80°C for 1 min for heat inactivation followed by a final hold at 4°C. 2 µL of 1 M DTT was added and a 0.8× Ampure XP purification was performed according to manufacturer's recommendations. cDNA was eluted in 40.5 µL of EB-DTT (10 mM DTT in EB) and added to ISPCR mix (40 µL library, 50 µL 2× KAPA HiFi, 10 µL 10 µM TSO-P primer). PCR cycling was performed according to the ISPCR program (95°C for 3 min, 13 cycles of [98°C for 20 s, 63°C for 20 s, 72°C for 3 min.], 72 °C for 5 min) followed by a final hold at 4°C. 2 µL of 1 M DTT was added and a 0.6× Ampure XP purification was performed according to manufacturer's recommendations. cDNA was eluted in 28.5 µL of EB-DTT. Final sequencing library was prepared according to the following customized NEB Ultra II FS protocol (NEB E7805S). 80 ng of amplified cDNA was fragmented in Ultra II fragmentation mix (26 µL of amplified cDNA, 7 µL of NEBNext Ultra II FS Reaction Buffer, 2 µL of NEBNext Ultra II FS Enzyme Mix) on the following thermocycling program: 37°C for 10 min, 65°C for 30 min, and a final hold at 4°C. 15 µL of EB was added and a 0.8× Ampure purification was performed according to manufacturer's recommendation and eluted in 35 µL. Fragmented library was adapter ligated in NEBNext Ultra II adapter ligation mix (35 µL of fragmented library, 30 µL of NEBNext Ultra II Ligation Master Mix, 1 µL of NEBNext Ligation Enhancer, 2.5 µL of NEBNext Adapter for Illumina) at 20°C for 15 min, with 4°C final hold. 28.5 µL of EB was added and a 0.8× Ampure purification was performed according to manufacturer's recommendation and eluted in 30 µL. Eluted library was amplified in PCR master mix (50 µL 2× KAPA HiFi, 10 µL 10 µM Hy-i7 primer, 10 µL 10 µM Hy-i5 primer, 30 µL eluted library) in the following thermocycling program: 95°C for 3 min, 13 cycles of (98°C for 20 s, 64°C for 30 s, 72°C for 30 s), 72°C for 5 min, and a final hold at 4°C. Sequencing-ready library was purified using a 0.8× Ampure purification and eluted in 30 µL of EB. User-friendly protocol available on <https://doi.org/10.17504/protocols.io.b4xwqxpe>.

HyDrop-RNA optimization trials

We performed six trials on a 50:50 mixture of human melanoma (MM087) and mouse melanoma (MMel). Trials were performed as described in the general HyDrop-RNA protocol, but with the following changes. All trials, except for the GTP/PEG trial, were performed using the following RT reaction mix (1.6× Maxima h-RT buffer, 1.6 mM dNTPs, 47 mM DTT, 15% Optiprep, 1.6 U/µL RNase Inhibitor, 15.7 U/µL Maxima hRT, 12.5 µM TSO). For the exocondition, the exonuclease I treatment was skipped. For all other conditions the exonuclease I treatment was performed as described above. For the TSO-LNA trial, an LNA TSO was used instead of the regular TSO. For the GTP/PEG trial, all steps were performed as described in the main protocol.

For the Klenow fragment second strand synthesis trial, the purified first strand product was treated with 1 µL of *Escherichia coli* RNase H (NEB M0297S). The mixture was incubated at 37°C for 30 min after which the enzyme was inactivated using 10 mM EDTA. The single stranded product was purified using 1.2× Ampure XP bead purification (BD sciences) and eluted in 25 µL of EB buffer. dN-SMRT primer was added to the single strand product to a final concentration of 2.5 µM and the mixture was denatured by incubation at 95°C for 5 min. The sample was then allowed to cool to room temperature and incorporated in the Klenow enzyme mix (1× Maxima h-RT buffer, 1 mM dNTP, 1 U/µL of Klenow Exo-; NEB M0212L) was added to the single strand library. The Klenow enzyme mix was incubated at 37°C for 60 min. The second strand reaction was stopped by heating the product at 85°C for 5 min. The sample was purified using 1× Ampure XP and eluted in 40 µL of EB buffer. The purified second strand product was amplified with ISPCR primers as described above.

For the BST 2.0 polymerase second strand synthesis trial, the purified first strand product was treated with 1 µL of *E. coli* RNase H (NEB M0297S). The mixture was incubated at 37°C for 30 min after which the enzyme was inactivated using 10 mM EDTA. The single stranded product was purified using 1.2× Ampure XP bead purification (BD sciences) and eluted in 25 µL of EB buffer. dN-SMRT primer was added to the single strand product to a final concentration of 2.5 µM and the mixture was denatured by incubation at 95°C for 5 min. The sample was then allowed to cool to room temperature

and incorporated in the Bst 2.0 enzyme mix (1× Isothermal amplification buffer, 1 mM dNTP, 1 U/μL of Bst 2.0 DNA polymerase; NEB M0537L) was added to the denatured library and the mixture was incubated at 55°C for 10 min and 60°C for 45 min. The second strand reaction was stopped by heating the product at 85°C for 5 min. The sample was purified using 1× Ampure XP and eluted in 40 μL of EB buffer. The purified second strand product was amplified with ISPCR primers as described above.

Sequencing

HyDrop-ATAC libraries were sequenced on Illumina NextSeq500 or NextSeq2000 systems using 50 cycles for read 1 (ATAC paired-end mate 1), 52 cycles for index 1 (barcode), 10 cycles for index 2 (sample index), and 50 cycles for read 2 (ATAC paired-end mate 2).

HyDrop-RNA libraries were sequenced on Illumina NextSeq2000 systems using 50 cycles for read 1 (3-prime cDNA), 10 cycles for index 1 (sample index, custom i7 read primer), 10 cycles for index 2 (sample index), and 58 cycles for read 2 (barcode+ UMI, custom read two primer).

HyDrop-ATAC data processing

Barcode reads were trimmed to exclude the intersub-barcode linear amplification adapters using a mawk script. Next, the VSN scATAC-seq pre-processing pipeline (*Waegeneer et al., 2021*) was used to map the reads to the reference genome and generate a fragments file for downstream analysis. Here, barcode reads were compared to a whitelist (of 884,736 valid barcodes), and corrected, allowing for a maximum 1 bp mismatch. Uncorrected and corrected barcodes were appended to the fastq sequence identifier of the paired end ATAC-seq reads. Reads were mapped to the reference genome using bwa-mem with default settings, and the barcode information was added as tags to each read in the bam file. Duplicate-marking was performed using samtools markdup. In the final step of the pipeline, fragment files were generated using Sinto (<https://github.com/timoast/sinto>). For mixed species data, cells were filtered for a minimum of 1000 unique fragments and a minimum TSS enrichment of 7. For mouse cortex data, higher level analysis such as clustering and differential accessibility were performed using cisTopic (*Bravo González-Blas et al., 2019*). In brief, cells were filtered for a minimum of 1000 unique fragments and a minimum TSS enrichment of 5. Fragments overlapping mouse candidate cis-regulatory regions (*ENCODE Project Consortium et al., 2020*) were counted, and the resulting matrix was filtered for potential cell doublets using a Scrublet (*Wolock et al., 2019*) threshold of 0.35. Cells were Leiden clustered based on the cell-topic probability matrix generated by an initial cisTopic LDA incorporating 51 topics, at a resolution of 0.9 with 10 neighbours. A consensus peak set was generated from per-cluster peaks and used to recount fragments. Cells were filtered using the same filtering parameters and a new model with 50 topics was trained. Cells were again Leiden clustered based on the cell-topic probability matrix generated by the second LDA, at a resolution of 0.9 with 10 neighbours. Region accessibility was imputed based on binarized topic-region and cell-topic distributions. Gene activity was imputed based on Gini index-weighted imputed accessibility in a 10 kb up/downstream decaying window around each gene including promoters. Leiden clusters were annotated based on imputed gene accessibility around marker genes (*Yao et al., 2021; Zeisel et al., 2018*). Differentially accessible regions were called using one-versus-all Wilcoxon rank-sum tests for each cell type, with an adjusted p-value of 0.05 and log₂FC of 1.5. RPGC-normalized aggregate genome coverage bigwigs were generated from BAM files using DeepTools (*Ramírez et al., 2016*). Per-cluster genome coverage tracks were generated using pyBigWig.

HyDrop-RNA data processing

Barcode reads were trimmed to exclude the intersub-barcode linear amplification adapters using a mawk script. Reads were then mapped and cell-demultiplexed using STARsolo (*Kaminow et al., 2021*) in CB_UMI_Complex mode. The resulting STARsolo-filtered count matrices were further analyzed using Scanpy (*Wolf et al., 2018*). In short, cells were filtered on expression of a maximum of 4000 genes, and a maximum of 1% UMIs from mitochondrial genes. Genes were filtered on expression in a minimum of three cells. Potential cell doublets were filtered out using a Scrublet (*Wolock et al., 2019*) threshold of 0.25. The filtered expression matrix was scaled to total counts and log-normalized. Total counts and mitochondrial reads were regressed out and UMAP embedding was performed after PCA. Cells were annotated and fine tuned based on differential gene expression of marker genes sourced

from either the Davie et al. *Drosophila* brain atlas (Davie et al., 2018) or the Allen Brain RNA-seq Database (Yao et al., 2021).

Raw inDrop (SRR10545068 to SRR10545079) and Drop-seq (SRR1853178 to SRR1853184) sequencing data was downloaded from SRA. Data was sampled to the HyDrop-RNA mouse cortex sample sequencing depth (52,738 reads per cell) and mapped to mouse reference genome using STARsolo, allowing one mismatch in the cell barcodes (like in our HyDrop-RNA and HyDrop-ATAC analyses). Full analysis process is documented on the Hydrop GitHub repository (https://github.com/aertslab/hydrop_data_analysis/tree/main/HyDrop-RNA_publicdata_comparison). Public reference 10× single-cell ATAC-seq data was sourced from <https://support.10xgenomics.com/single-cell-atac/datasets> ('Flash frozen cortex, hippocampus, and ventricular zone from embryonic mouse brain (E18)', 'Fresh cortex from adult mouse brain (P50)'). Public reference 10× single-cell gene expression data was sourced from <https://support.10xgenomics.com/single-cell-gene-expression/datasets> ('1 k Brain Nuclei from an E18 Mouse', '2 k Brain Nuclei from an Adult Mouse (> 8 weeks)'). Public PC-3 and MCF-7 ATAC-seq data was sourced from ENCODE (ENCFF772EFK: doi:10.17989/ENCSR422SUG, ENCFF024FNF: doi:10.17989/ENCSR499ASS).

Data was visualized using a combination of Seaborn (Waskom, 2021) and Matplotlib (Hunter, 2007). A vector image representing mouse head and cortex was sourced from SciDraw (Kennedy, 2020).

Acknowledgements

We thank Andrew Adey for his great advice on Tn5 and scATAC-seq. We also thank Sebastián Najle, Céline Vallot, and Arnau Sebé-Pedrós for many discussions on droplet microfluidics, Frederik Ceysens and the KU Leuven Nanocenter for their support with microfabrication, Ghanem Ghanem for his kind donation of the MM087 melanoma lines, Jean-Christophe Marine for his kind donation of the mouse melanoma lines, and Koen De Wispelaere for his assistance during his master internship.

Additional information

Funding

Funder	Grant reference number	Author
H2020 European Research Council	724226_cis- CONTROL	Stein Aerts
KU Leuven	C14/18/092	Stein Aerts
Fonds Wetenschappelijk Onderzoek	G0B5619N	Stein Aerts
Michael J. Fox Foundation for Parkinson's Research	ASAP-000430	Christopher Campbell Flerin
Aligning Science Across Parkinson's	ASAP-000430	Christopher Campbell Flerin
Foundation Against Cancer	2016-070	Stein Aerts
Stichting Tegen Kanker		Jasper Wouters
Belgian Cancer Society		Jasper Wouters
Fonds Wetenschappelijk Onderzoek		Florian V De Rop Carmen Bravo González-Blas Jasper Janssens
VIB Tech Watch		Suresh Poovathingal

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Florian V De Rop, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review and editing; Joy N Ismail, Investigation, Writing – review and editing; Carmen Bravo González-Blas, Formal analysis, Software, Writing – review and editing; Gert J Hulselmans, Data curation, Software; Christopher Campbell Flerin, Software, Writing – original draft, Writing – review and editing; Jasper Janssens, Software; Koen Theunis, Methodology; Valerie M Christiaens, Gabriele Marcassa, Joris de Wit, Resources; Jasper Wouters, Writing – review and editing; Suresh Poovathingal, Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Writing – original draft, Writing – review and editing; Stein Aerts, Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – original draft, Writing – review and editing

Author ORCIDs

Florian V De Rop  <http://orcid.org/0000-0001-5241-924X>

Jasper Wouters  <http://orcid.org/0000-0002-7129-2990>

Suresh Poovathingal  <http://orcid.org/0000-0002-3236-2255>

Stein Aerts  <http://orcid.org/0000-0002-8006-0315>

Ethics

Ethical approval All animal experiments were conducted according to the KU Leuven ethical guidelines and approved by the KU Leuven Ethical Committee for Animal Experimentation (approved protocol numbers ECD P037/2016, P014/2017, and P062/2017). All use of cell lines was approved by the KU Leuven Ethical Committee for Research under project number S63316.

Decision letter and Author response

Decision letter <https://doi.org/10.7554/eLife.73971.sa1>

Author response <https://doi.org/10.7554/eLife.73971.sa2>

Additional files

Supplementary files

- Supplementary file 1. "Molecular sequence description of HyDrop bead barcoding": visual description of progression of nucleotide sequences involved in every step of the barcoding process.
- Supplementary file 2. "Molecular sequence description of HyDrop-ATAC": visual description of progression of nucleotide sequences involved in every step of HyDrop-ATAC.
- Supplementary file 3. "Molecular sequence description of HyDrop-RNA": visual description of progression of nucleotide sequences involved in every step of HyDrop-RNA.
- Supplementary file 4. "Reagents and oligonucleotide list": list of all reagents and oligonucleotides used in HyDrop.
- Supplementary file 5. "Price calculation for HyDrop bead barcoding and HyDrop-ATAC": detailed price calculation for HyDrop bead barcoding and HyDrop-ATAC.
- Transparent reporting form
- Source data 1. Source data files for **Figure 4a**, **Figure 5b–d**, **Figure 8a–c**, **Figure 8—figure supplement 2**, and **Figure 10b**.

Data availability

The data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE175684 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE175684>) and on SCoPe (https://scope.aertslab.org/#/HyDrop*/welcome). Source data files have been provided for Figures 2a, 3a, 3b, 3c and 3i, and can be regenerated data analysis tutorials for HyDrop. Data analysis tutorials for HyDrop are available on GitHub (https://github.com/aertslab/hydrop_data_analysis; copy archived at [swh.io/rev:059bf5a7779dc2894670ecf5f820c14bceb68493](https://www.swh.io/rev/059bf5a7779dc2894670ecf5f820c14bceb68493)).

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
De Rop F, Ismail J, Bravo González-Blas C, Hulselmans G, Flerin C, Janssens J, Christiaens V, Poovathingal S, Aerts S	2021	HyDrop: droplet-based scATAC-seq and scRNA-seq using dissolvable hydrogel beads	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE175684	NCBI Gene Expression Omnibus, GSE175684

The following previously published datasets were used:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Macosko EZ	2015	Drop-Seq analysis of P14 mouse retina single-cell suspension	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63472	NCBI Gene Expression Omnibus, GSE63472
Kalish BT, Greenberg ME	2020	inDrop scRNA-seq of mouse primary auditory cortex	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE140883	NCBI Gene Expression Omnibus, GSE140883
10x Genomics	2019	10x scATAC-seq on flash frozen cortex, hippocampus, and ventricular zone from embryonic mouse brain (E18)	https://support.10xgenomics.com/single-cell-atac/datasets/1.2.0/atac_v1_E18_brain_flash_5k	10x Genomics Resources, atac_v1_E18_brain_flash_5k
10x Genomics	2019	10x scATAC-seq on fresh cortex from adult mouse brain (P50)	https://support.10xgenomics.com/single-cell-atac/datasets/1.2.0/atac_v1_adult_brain_fresh_5k	10x Genomics Resources, atac_v1_adult_brain_fresh_5k
10x Genomics	2017	10x scRNA-seq on 1k brain nuclei from an E18 mouse	https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/nuclei_900	10x Genomics Resources, nuclei_900
10x Genomics	2017	10x scRNA-seq on 2k brain nuclei from an adult mouse (>8 weeks)	https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/nuclei_2k	10x Genomics Resources, nuclei_2k
ENCODE Project Consortium	2021	ATAC-seq from MCF-7 (ENCSR422SUG)	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE169929	NCBI Gene Expression Omnibus, GSE169929
ENCODE Project Consortium	2021	ATAC-seq from PC-3 (ENCSR499ASS)	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE170337	NCBI Gene Expression Omnibus, GSE170337

References

- Abate AR, Chen CH, Agresti JJ, Weitz DA. 2009. Beating Poisson encapsulation statistics using close-packed ordering. *Lab on a Chip* **9**:2628–2631. DOI: <https://doi.org/10.1039/b909386a>, PMID: 19704976
- Bradski G. 2015. Automated Calibration of RF On-Wafer Probing and Evaluation of Probe Misalignment Effects Using a Desktop Micro-Factory. *Journal of Computer and Communications* **4**:122–125. DOI: <https://doi.org/10.4236/jcc.2016.43009>
- Bravo González-Blas C, Minnoye L, Papasokrati D, Aibar S, Hulselmans G, Christiaens V, Davie K, Wouters J, Aerts S. 2019. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nature Methods* **16**:397–400. DOI: <https://doi.org/10.1038/s41592-019-0367-1>, PMID: 30962623

- Chen S**, Lake BB, Zhang K. 2019. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature Biotechnology* **37**:1452–1457. DOI: <https://doi.org/10.1038/s41587-019-0290-0>, PMID: 31611697
- Consortium EP**. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**:57–74. DOI: <https://doi.org/10.1038/nature11247>, PMID: 22955616
- Dankort D**, Curley DP, Cartlidge RA, Nelson B, Karnezis AN, Damsky WE Jr, You MJ, DePinho RA, McMahon M, Bosenberg M. 2009. Braf(V600E) cooperates with Pten loss to induce metastatic melanoma. *Nature Genetics* **41**:544–552. DOI: <https://doi.org/10.1038/ng.356>, PMID: 19282848
- Datlinger P**, Rendeiro AF, Boenke T, Senekowitsch M, Krausgruber T, Barreca D, Bock C. 2021. Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. *Nature Methods* **18**:635–642. DOI: <https://doi.org/10.1038/s41592-021-01153-z>, PMID: 34059827
- Davie K**, Janssens J, Koldere D, De Waegeneer M, Pech U, Kreft L, Aibar S, Makhzami S, Christiaens V, Bravo González-Blas C, Poovathingal S, Hulselmans G, Spanier KI, Moerman T, Vanspauwen B, Geurs S, Voet T, Lammertyn J, Thienpont B, Liu S, et al. 2018. A Single-Cell Transcriptome Atlas of the Aging *Drosophila* Brain. *Cell* **174**:982–998. DOI: <https://doi.org/10.1016/j.cell.2018.05.057>, PMID: 29909982
- Delley CL**, Abate AR. 2020. Modular barcode beads for microfluidic single cell genomics. *Cell Biology* **1**:e292326. DOI: <https://doi.org/10.1101/2020.09.10.292326>
- ENCODE Project Consortium**, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shores N, Adrian J, Kawi L, Davis CA, Dobin A, Kaul R, Halow J, Van Nostrand EL, Freese P, Gorkin DU, Shen Y, He Y, Mackiewicz M, Pauli-Behn F, Williams BA, et al. 2020. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**:699–710. DOI: <https://doi.org/10.1038/s41586-020-2493-4>, PMID: 32728249
- Hagemann-Jensen M**, Ziegenhain C, Chen P, Ramsköld D, Hendriks G-J, Larsson AJM, Faridani OR, Sandberg R. 2020. Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nature Biotechnology* **38**:708–714. DOI: <https://doi.org/10.1038/s41587-020-0497-0>, PMID: 32518404
- Hughes TK**, Wadsworth MH, Gierahn TM, Do T, Weiss D, Andrade PR, Ma F, de Andrade Silva BJ, Shao S, Tsoi LC, Ordoñez-Montanes J, Gudjonsson JE, Modlin RL, Love JC, Shalek AK. 2020. Second-Strand Synthesis-Based Massively Parallel scRNA-Seq Reveals Cellular States and Molecular Features of Human Inflammatory Skin Pathologies. *Immunity* **53**:878–894. DOI: <https://doi.org/10.1016/j.immuni.2020.09.015>, PMID: 33053333
- Hunter JD**. 2007. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **9**:90–95. DOI: <https://doi.org/10.1109/MCSE.2007.55>
- Hur SC**, Henderson-MacLennan NK, McCabe ERB, Di Carlo D. 2011. Deformability-based cell classification and enrichment using inertial microfluidics. *Lab on a Chip* **11**:912–920. DOI: <https://doi.org/10.1039/c0lc00595a>, PMID: 21271000
- Kalish BT**, Barkat TR, Diel EE, Zhang EJ, Greenberg ME, Hensch TK. 2020. Single-nucleus RNA sequencing of mouse auditory cortex reveals critical period triggers and brakes. *PNAS* **117**:11744–11752. DOI: <https://doi.org/10.1073/pnas.1920433117>, PMID: 32404418
- Kaminow B**, Yunusov D, Dobin A. 2021. GenomicSuperSignature: interpretation of RNA-seq experiments through robust, efficient comparison to public databases. *Bioinformatics (Oxford, England)* **1**:e55. DOI: <https://doi.org/10.1101/2021.05.05.442755>
- Karaiskos N**, Wahle P, Alles J, Boltengagen A, Ayoub S, Kipar C, Kocks C, Rajewsky N, Zinzen RP. 2017. The *Drosophila* embryo at single-cell transcriptome resolution. *Science (New York, N.Y.)* **358**:194–199. DOI: <https://doi.org/10.1126/science.aan3235>, PMID: 28860209
- Kennedy A**. 2020. mouse brain silhouette. Zenodo. <https://zenodo.org/record/3925919>
- Kim B**, Hahn YK, You D, Oh S, Choi S. 2016. A smart multi-pipette for hand-held operation of microfluidic devices. *The Analyst* **141**:5753–5758. DOI: <https://doi.org/10.1039/c6an00681g>, PMID: 27478886
- Kivioja T**, Vähäurto A, Karlsson K, Bonke M, Eng M, Linnarsson S, Taipale J. 2011. Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods* **9**:72–74. DOI: <https://doi.org/10.1038/nmeth.1778>, PMID: 22101854
- Klein AM**, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. 2015. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**:1187–1201. DOI: <https://doi.org/10.1016/j.cell.2015.04.044>, PMID: 26000487
- Konstantinides N**, Kapuralin K, Fadil C, Barboza L, Satija R, Desplan C. 2018. Phenotypic Convergence: Distinct Transcription Factors Regulate Common Terminal Features. *Cell* **174**:622–635. DOI: <https://doi.org/10.1016/j.cell.2018.05.021>, PMID: 29909983
- Langer K**, Bremond N, Boitard L, Baudry J, Bibette J. 2018. Micropipette-powered droplet based microfluidics. *Biomicrofluidics* **12**:044106. DOI: <https://doi.org/10.1063/1.5037795>, PMID: 30034569
- Lareau CA**, Duarte FM, Chew JG, Kartha VK, Burkett ZD, Kohlway AS, Pokholok D, Aryee MJ, Steemers FJ, Lebofsky R, Buenrostro JD. 2019. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nature Biotechnology* **37**:916–924. DOI: <https://doi.org/10.1038/s41587-019-0147-6>, PMID: 31235917
- Li YE**, Preissl S, Hou X, Zhang Z, Zhang K, Fang R, Qiu Y, Poirion O, Li B, Liu H, Wang X, Han JY, Lucero J, Yan Y, Kuan S, Gorkin D, Nunn M, Mukamel EA, Behrens MM, Ecker J, et al. 2020. An Atlas of Gene Regulatory Elements in Adult Mouse Cerebrum. *Neuroscience* **1**:e87585. DOI: <https://doi.org/10.1101/2020.05.10.087585>
- Macosko EZ**, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Mathersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA. 2015. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**:1202–1214. DOI: <https://doi.org/10.1016/j.cell.2015.05.002>, PMID: 26000488

- Mulqueen RM, DeRosa BA, Thornton CA, Sayar Z, Torkenczy KA, Fields AJ, Wright KM, Nan X, Ramji R, Steemers FJ, O'Roak BJ, Adey AC. 2019. Improved Single-Cell ATAC-Seq Reveals Chromatin Dynamics of in Vitro Corticogenesis. *bioRxiv*. DOI: <https://doi.org/10.1101/637256>
- Mulqueen RM, Pokholok D, O'Connell BL, Thornton CA, Zhang F, O'Roak BJ, Link J, Yardımcı GG, Sears RC, Steemers FJ, Adey AC. 2021. High-content single-cell combinatorial indexing. *Nature Biotechnology* **39**:1574–1580. DOI: <https://doi.org/10.1038/s41587-021-00962-z>, PMID: 34226710
- Picelli S, Faridani OR, Björklund AK, Winberg G, Sagasser S, Sandberg R. 2014. Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols* **9**:171–181. DOI: <https://doi.org/10.1038/nprot.2014.006>, PMID: 24385147
- Preissl S, Fang R, Huang H, Zhao Y, Raviram R, Gorkin DU, Zhang Y, Sos BC, Afzal V, Dickel DE, Kuan S, Visel A, Pennacchio LA, Zhang K, Ren B. 2018. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nature Neuroscience* **21**:432–439. DOI: <https://doi.org/10.1038/s41593-018-0079-3>, PMID: 29434377
- Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research* **44**:W160–W165. DOI: <https://doi.org/10.1093/nar/gkw257>, PMID: 27079975
- Ren X, Wen W, Fan X, Hou W, Su B, Cai P, Li J, Liu Y, Tang F, Zhang F, Yang Y, He J, Ma W, He J, Wang P, Cao Q, Chen F, Chen Y, Cheng X, Deng G, et al. 2021. COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell* **184**:1895–1913. DOI: <https://doi.org/10.1016/j.cell.2021.01.053>, PMID: 33657410
- Rivello F, van Buijtenen E, Matuła K, van Buggenum J, Vink P, van Eenennaam H, Mulder KW, Huck WTS. 2020. Single-cell intracellular epitope and transcript detection revealing signal transduction dynamics. *Molecular Biology* **1**:e08120. DOI: <https://doi.org/10.1101/2020.12.02.408120>
- Rop FD. 2022. hydrop_data_analysis. swb:1:rev:059bf5a7779dc2894670ecf5f820c14bceb68493. Software Heritage. https://archive.softwareheritage.org/swb:1:dir:28fb90ef390d0028cdc38102244567996855e4fd;origin=https://github.com/aertslab/hydrop_data_analysis;visit=swb:1:snp:8c8b341705393049010ae1d01233a0ae531d82b;anchor=swb:1:rev:059bf5a7779dc2894670ecf5f820c14bceb68493
- Sarkar A, Mei A, Paquola ACM, Stern S, Bardy C, Klug JR, Kim S, Neshat N, Kim HJ, Ku M, Shokhirev MN, Adamowicz DH, Marchetto MC, Jappelli R, Erwin JA, Padmanabhan K, Shtrahman M, Jin X, Gage FH. 2018. Efficient Generation of CA3 Neurons from Human Pluripotent Stem Cells Enables Modeling of Hippocampal Connectivity In Vitro. *Cell Stem Cell* **22**:684–697. DOI: <https://doi.org/10.1016/j.stem.2018.04.009>, PMID: 29727680
- Satpathy AT, Granja JM, Yost KE, Qi Y, Meschi F, McDermott GP, Olsen BN, Mumbach MR, Pierce SE, Corces MR, Shah P, Bell JC, Jhutti D, Nemec CM, Wang J, Wang L, Yin Y, Giresi PG, Chang ALS, Zheng GXY, et al. 2019. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nature Biotechnology* **37**:925–936. DOI: <https://doi.org/10.1038/s41587-019-0206-z>, PMID: 31375813
- Sinha N, Subedi N, Wimmers F, Soennichsen M, Tel J. 2019. A Pipette-Tip Based Method for Seeding Cells to Droplet Microfluidic Platforms. *Journal of Visualized Experiments* **1**:e48. DOI: <https://doi.org/10.3791/57848>, PMID: 30799837
- Stickels RR, Murray E, Kumar P, Li J, Marshall JL, Di Bella DJ, Arlotta P, Macosko EZ, Chen F. 2021. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nature Biotechnology* **39**:313–319. DOI: <https://doi.org/10.1038/s41587-020-0739-1>, PMID: 33288904
- Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, Smibert P. 2017. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods* **14**:865–868. DOI: <https://doi.org/10.1038/nmeth.4380>, PMID: 28759029
- Svensson V, da Veiga Beltrame E, Pachter L. 2020. A curated database reveals trends in single-cell transcriptomics. *Database* **2020**:baaa073. DOI: <https://doi.org/10.1093/database/baaa073>, PMID: 33247933
- Traag VA, Waltman L, van Eck NJ. 2019. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* **9**:5233. DOI: <https://doi.org/10.1038/s41598-019-41695-z>, PMID: 30914743
- Waegeneer MD, Flerin CC, Davie K. 2021. vib-singlecell-nf/vsn-pipelines. 1.01. Zenodo. <https://zenodo.org/record/4468513>
- Wang Y, Cao T, Ko J, Shen Y, Zong W, Sheng K, Cao W, Sun S, Cai L, Zhou YL, Zhang XX, Zong C, Weissleder R, Weitz D. 2020. Dissolvable Polyacrylamide Beads for High-Throughput Droplet DNA Barcoding. *Advanced Science (Weinheim, Baden-Wuerttemberg, Germany)* **7**:1903463. DOI: <https://doi.org/10.1002/advs.201903463>, PMID: 32328429
- Waskom ML. 2021. seaborn: statistical data visualization. *Journal of Open Source Software* **6**:3021. DOI: <https://doi.org/10.21105/joss.03021>
- Wolf FA, Angerer P, Theis FJ. 2018. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology* **19**:15. DOI: <https://doi.org/10.1186/s13059-017-1382-0>, PMID: 29409532
- Wolock SL, Lopez R, Klein AM. 2019. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Systems* **8**:281–291. DOI: <https://doi.org/10.1016/j.cels.2018.11.005>, PMID: 30954476
- Xia Y, Whitesides GM. 1998. SOFT LITHOGRAPHY. *Annual Review of Materials Science* **28**:153–184. DOI: <https://doi.org/10.1146/annurev.matsci.28.1.153>
- Yao Z, van Velthoven CTJ, Nguyen TN, Goldy J, Sedeno-Cortes AE, Baftizadeh F, Bertagnoli D, Casper T, Chiang M, Crichton K, Ding SL, Fong O, Garren E, Glandon A, Gouwens NW, Gray J, Graybuck LT, Hawrylycz MJ, Hirschstein D, Kroll M, et al. 2021. A taxonomy of transcriptomic cell types across the isocortex

- and hippocampal formation. *Cell* **184**:3222–3241. DOI: <https://doi.org/10.1016/j.cell.2021.04.021>, PMID: [34004146](https://pubmed.ncbi.nlm.nih.gov/34004146/)
- Yap E-L**, Pettit NL, Davis CP, Nagy MA, Harmin DA, Golden E, Dagliyan O, Lin C, Rudolph S, Sharma N, Griffith EC, Harvey CD, Greenberg ME. 2021. Bidirectional perisomatic inhibitory plasticity of a Fos neuronal network. *Nature* **590**:115–121. DOI: <https://doi.org/10.1038/s41586-020-3031-0>, PMID: [33299180](https://pubmed.ncbi.nlm.nih.gov/33299180/)
- Zeisel A**, Hochgerner H, Lönnerberg P, Johnsson A, Memic F, van der Zwan J, Häring M, Braun E, Borm LE, La Manno G, Codeluppi S, Furlan A, Lee K, Skene N, Harris KD, Hjerling-Leffler J, Arenas E, Ernfors P, Marklund U, Linnarsson S. 2018. Molecular Architecture of the Mouse Nervous System. *Cell* **174**:999–1014.. DOI: <https://doi.org/10.1016/j.cell.2018.06.021>, PMID: [30096314](https://pubmed.ncbi.nlm.nih.gov/30096314/)
- Zhang X**, Li T, Liu F, Chen Y, Yao J, Li Z, Huang Y, Wang J. 2019. Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. *Molecular Cell* **73**:130–142.. DOI: <https://doi.org/10.1016/j.molcel.2018.10.020>, PMID: [30472192](https://pubmed.ncbi.nlm.nih.gov/30472192/)
- Zheng GXY**, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Gregory MT, Shuga J, Montesclaros L, Underwood JG, Masquelier DA, Nishimura SY, Schnall-Levin M, Wyatt PW, Hindson CM, Bharadwaj R, et al. 2017. Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8**:14049. DOI: <https://doi.org/10.1038/ncomms14049>, PMID: [28091601](https://pubmed.ncbi.nlm.nih.gov/28091601/)
- Zilionis R**, Nainys J, Veres A, Savova V, Zemmour D, Klein AM, Mazutis L. 2017. Single-cell barcoding and sequencing using droplet microfluidics. *Nature Protocols* **12**:44–73. DOI: <https://doi.org/10.1038/nprot.2016.154>, PMID: [27929523](https://pubmed.ncbi.nlm.nih.gov/27929523/)

Chapter 3

Systematic benchmarking of single-cell ATAC-sequencing protocols

Systematic benchmarking of single-cell ATAC-seq protocols

Received: 18 January 2022

Accepted: 22 June 2023

Published online: 3 August 2023

 Check for updates

Florian V. De Rop^{1,2}, Gert Hulselmans^{1,2}, Chris Flerin^{1,2}, Paula Soler-Vila³, Albert Rafels⁴, Valerie Christiaens^{1,2}, Carmen Bravo González-Blas^{1,2}, Domenica Marchese⁴, Ginevra Caratù⁴, Suresh Poovathingal¹, Orit Rozenblatt-Rosen⁵, Michael Slyper⁶, Wendy Luo⁶, Christoph Muus⁶, Fabiana Duarte^{7,8}, Rojesh Shrestha^{7,8}, S. Tansu Bagdatli⁹, M. Ryan Corces¹⁰, Lira Mamanova¹¹, Andrew Knights¹¹, Kerstin B. Meyer¹¹, Ryan Mulqueen¹², Akram Taherinasab^{13,14}, Patrick Maschmeyer^{15,16}, Jörn Pezoldt^{17,18}, Camille Lucie Germaine Lambert^{17,18}, Marta Iglesias^{4,19}, Sebastián R. Najle⁴, Zain Y. Dossani^{20,21}, Luciano G. Martelotto^{22,23}, Zach Burkett²⁴, Ronald Lebofsky²⁴, José Ignacio Martín-Subero^{3,25,26,27}, Satish Pillai^{20,21}, Arnau Sebé-Pedrós^{4,19,28}, Bart Deplancke^{17,18}, Sarah A. Teichmann^{11,29}, Leif S. Ludwig^{6,15,16}, Theodore P. Braun^{13,14}, Andrew C. Aday¹², William J. Greenleaf^{9,30}, Jason D. Buenrostro^{7,8}, Aviv Regev^{5,31,32}, Stein Aerts^{1,2,33}✉ & Holger Heyn^{4,19,33}✉

Single-cell assay for transposase-accessible chromatin by sequencing (scATAC-seq) has emerged as a powerful tool for dissecting regulatory landscapes and cellular heterogeneity. However, an exploration of systemic biases among scATAC-seq technologies has remained absent. In this study, we benchmark the performance of eight scATAC-seq methods across 47 experiments using human peripheral blood mononuclear cells (PBMCs) as a reference sample and develop PUMATAC, a universal preprocessing pipeline, to handle the various sequencing data formats. Our analyses reveal significant differences in sequencing library complexity and tagmentation specificity, which impact cell-type annotation, genotype demultiplexing, peak calling, differential region accessibility and transcription factor motif enrichment. Our findings underscore the importance of sample extraction, method selection, data processing and total cost of experiments, offering valuable guidance for future research. Finally, our data and analysis pipeline encompasses 169,000 PBMC scATAC-seq profiles and a best practices code repository for scATAC-seq data analysis, which are freely available to extend this benchmarking effort to future protocols.

Data quality in single-cell sequencing studies directly influences successful interpretation. Technologies that generate high and accurate molecule counts allow for a precise characterization of cells and can yield deep insights into the underlying tissue biology. To facilitate

an informed decision-making process on the choice of technology, systematic benchmarking efforts have been performed for sample preparation protocols¹ and single-cell RNA-sequencing (scRNA-seq) technologies². However, such efforts are still lacking for chromatin

A full list of affiliations appears at the end of the paper. ✉e-mail: stein.aerts@kuleuven.be; holger.heyne@cng.org.eu

accessibility profiling technologies³, such as single-cell assay for transposase-accessible chromatin by sequencing (scATAC-seq)⁴. Recent technological advances enable large-scale studies and have established scATAC-seq as a major pillar in systematic profiling efforts^{5–7}. While transcriptomic profiles based on scRNA-seq provide information to infer cellular phenotypes, scATAC-seq detects accessible chromatin sites that pinpoint genomic regions involved in gene regulation. The latter is particularly important to derive mechanistic insights into cell-type development and differentiation or to identify drivers of cell state dynamics following a stimulus, perturbations or disease (Table 1).

This work benchmarked eight scATAC-seq methods across 47 experiments, including technical and center replicates for Bio-Rad ddSEQ, HyDrop, s3-ATAC and different variants of the 10x Genomics scATAC-seq assay. Peripheral blood mononuclear cells (PBMCs) served as a reference sample to minimize technical variability related to sample preparation, allowing the systematic evaluation of method performance across multiple quality control metrics. PUMATAC⁸, our pipeline for universal mapping of ATAC-seq data, further reduced variability in data preprocessing and enabled systematic benchmarking of the presented data, also allowing the extension to future technologies. Differences between methods were driven by sequencing library complexity and tagmentation specificity, with consequences on the performance of key features for data analysis and interpretation and the integration of datasets into joint cellular atlases.

Results

We performed a systematic, multicenter benchmarking study of eight different scATAC-seq protocols. Our benchmark includes all variants of 10x Genomics scATAC-seq (v1 (ref. 9), v1.1, v2, multiome and mitochondrial scATAC (mtscATAC)¹⁰) as well as Bio-Rad ddSEQ¹¹, HyDrop¹² and s3-ATAC¹³. A reference sample of PBMCs from two adult donors (male and female) mixed at a 1:1 ratio was used to simulate complex sample composition (containing multiple cell types and conditions) and minimize sample preparation complexity. This reference sample was distributed for a multicenter benchmarking study and was used for all experiments, with the exception of the indicated replicate datasets (Fig. 1a). Each experiment was performed in technical replicates across centers with a target of 3,000 cells per sample to recover all major PBMC cell types, such as T and B cell subtypes, natural killer (NK) cells, monocytes and dendritic cells (DCs). In total, we generated 47 datasets (Fig. 1b), including replicates across at least three centers with two technical replicate experiments for all methods, except s3-ATAC and 10x v1.

PUMATAC: a generic and automated analysis pipeline

To compare different protocols in a unified manner, all sequencing data were analyzed using PUMATAC (pipeline for universal mapping of ATAC-seq data)⁸, a newly developed scATAC-seq preprocessing pipeline (Fig. 1c). Briefly, PUMATAC takes scATAC-seq data and applies a set of uniform preprocessing steps, including cell barcode error correction, adapter trimming, reference genome alignment and mapping quality filtering (Methods). We chose bwa-mem2 for read alignment as bwa-mem was used in the original manuscripts describing the eight technologies. PUMATAC then records aligned chromatin fragments in the ubiquitous bed-like ‘fragments file’ format, a tab-separated text file providing the start and end positions of each fragment and its corresponding cell barcode. On average, across samples, 97% of aligned fragments by PUMATAC shared their barcodes and coordinates with their counterparts aligned by CellRanger (10x Genomics; Supplementary Fig. 1a). PUMATAC currently includes workflows for the eight technologies described in this benchmark study but allows for the modular addition of new scATAC-seq sequencing methods by modifying existing templates¹⁴. PUMATAC also features a reimplement of bap2 (ref. 15), which detects and merges barcodes with significant fragment identity overlap. This step is necessary to merge bead doublets

Table 1 | Estimated costs per experiment of 5,000 cells

	10x v2	10x multiome	Bio-Rad ddSEQ	s3-ATAC	HyDrop
Sequenced reads per cell at saturation	55,000	68,000	19,000	1,467,000	10,000
Expected unique fragments per cell	22,427	10,155	5,249	66,130	1,884
Expected unique fragments in peaks per cell	13,680	6,398	2,992	12,565	716
Assay price per 5,000 cells	\$1,565	\$2,843	\$1,100	\$800	\$100
Sequencing cost	\$791	\$978	\$273	\$21,088	\$144
Total cost per cell	\$0.471	\$0.764	\$0.275	\$3.80	\$0.049

Saturation sequencing depth is defined as the depth at which 50% of fragments in cells are duplicates. The expected number of unique fragments was calculated as the expected number of unique fragments at saturation sequencing depth, interpolated or extrapolated using a Langmuir model, and multiplied by the median FRIP score per technique to achieve the expected in peaks count. The price per 5,000 cells is for one 10x lane, one ddSEQ lane, four s3-ATAC plates (1,440 cells each, for 5,760 cells in total) or one HyDrop ATAC run. Sequencing cost is defined as the price of sequencing 5,000 cells to saturation depth, calculated at a cost of \$2.875 per 1 million reads, using NovaSeq S2 100 cycles (10x v2, 10x multiome and HyDrop) or 200 cycles (ddSEQ and s3-ATAC; 4,000 million reads sequenced at \$10,000 to \$11,000 total cost). The sequencing cost for 10x multiome RNA component was not included.

(inherent to Bio-Rad ddSEQ data) and was performed on all techniques for uniformity (Supplementary Fig. 1c). As expected, between 55% and 92% of Bio-Rad-filtered cell barcodes were the result of such merging events due to the method’s bead overloading strategy, as opposed to a median of 1.5% for non-Bio-Rad samples (Supplementary Fig. 1b).

After preprocessing, the fragments files were postprocessed using cisTopic¹⁶. High-quality cells were separated from background noise barcodes and low-quality cells using sample-specific, algorithmically defined minimum thresholds on the number of unique fragments and transcription start site (TSS) enrichment (Fig. 1d, Extended Data Fig. 2 and Methods). Background noise barcodes are barcodes enriched with trace amounts of fragments through three mechanisms: (1) ambient accessible chromatin fragments may initiate barcoding reactions inside cell-free droplets, (2) unbound barcodes in the bead stock solutions may contaminate legitimate barcoding reactions, or (3) barcode impurities on the beads may lead to ambivalent labeling of chromatin fragments from a single nucleus¹⁵. Nucleosome-free fragments from viable cells show enrichment around TSSs, a hallmark used to exclude low-quality cells from subsequent analysis^{4,17}. All original sequencing datasets were then downsampled to 40,796 reads per cell, the highest common number of reads available across all samples (hereafter referred to as the 40k dataset). Finally, the downsampled datasets were processed again using PUMATAC and cisTopic and subset to the original set of filtered cell barcodes, ensuring that the number of reads per cell remained constant. All downstream analyses were performed on the 40k dataset or further downsampled sets derived thereof.

To generate a preliminary count matrix, fragments were first counted in the ENCODE SCREEN regions¹⁸. Cell doublets were detected and removed based on cell-type identity using Scrublet¹⁹ and donor genotype identity using Freemuxlet²⁰ (Methods). Next, cells were clustered using cisTopic and annotated through an automated label transfer from independently annotated scRNA-seq PBMC reference data²¹ using Seurat²². For each sample, cell-type-specific chromatin accessibility peaks were detected using MACS2 (ref. 23) and aggregated into a high-quality set of sample-specific consensus peaks. This strategy was used to discover peaks specific to smaller cell populations. Each sample’s fragments were then recounted in its own consensus peak set, generating a consensus peak count matrix before a second round of cisTopic clustering and Seurat label transfer was performed.

The resulting 47 annotated and clustered datasets were used for all downstream analyses and comparisons. Finally, a third round of consensus peak sets was generated from all samples and aggregated into one master set of PBMC consensus peaks, which was used to generate a merged count matrix containing cells from all experiments. This strategy was used to generate three distinct benchmarking datasets containing (1) all cells at original sequencing depth (Extended Data Fig. 3), (2) all cells at 40k sequencing depth (Fig. 4b) and (3) an equal number of cells per cell type at 40k sequencing depth (Extended Data Fig. 4a). We also generated randomly sampled sets (2,500, 2,000, 1,500, 1,000 and 500 cells) to investigate potential effects of total cell counts on label transfer and differentially accessible region (DAR) calling. All merged benchmarking datasets and all individual samples are available in SScope, our public single-cell viewer (https://scope.aertslab.org/#/scATAC-seq_Benchmark/scATAC-seq_Benchmark/welcome).

Chromatin fragment capture and sequencing efficiency

The PUMATAC analysis workflow discarded sequencing reads at several filtering steps throughout preprocessing. The fraction of reads that ultimately remains in the filtered count matrix is an important measure of assay quality as it is inversely related to sequencing costs. Furthermore, understanding the causes of read loss can shed light on how such losses affect downstream analyses and how to improve reaction chemistries or sample preparation workflows. We therefore calculated the fraction of total reads lost at each filtering stage and quantified differences in these metrics between samples (Fig. 1e). The first loss occurred when barcode reads were corrected based on a barcode sequence whitelist and when mapped to the reference genome. More specifically, reads without a valid cell barcode or those that could not be aligned with a Phred mapping quality score of >30 were discarded. Both steps combined accounted for relatively small losses, ranging from 10.4% for 10x v2 to 22.7% for HyDrop. Filtering true cells from background noise and low-quality cells resulted in significantly larger losses for some methods. Between 7% (10x v2) and 60% (s3-ATAC) of mapped fragments were discarded at this stage. Notably, fluorescence-activated cell sorting (FACS) of live cells before nuclei extraction reduced such losses to below 6% for two mtscATAC-seq experiments (samples mt* Br1 and mt* Br2) compared to 36% in mtscATAC-seq without FACS (samples mt M1, mt M2, mt C1 and mt C2), suggesting that the additional FACS step removes significant amounts of ambient chromatin and damaged cells. Because diploid cells can produce a maximum of two unique fragments originating from a given accessible region and the fact that scATAC-seq data are sparse compared to all possible accessible sites, accessibility in peak regions is usually binarized within a single cell. Out of the quality-filtered fragments, a large fraction includes duplicates resulting from fragment amplification steps during sample preparation, which are discarded instead of quantified. The fraction of duplicate reads in our datasets ranged from 5% for s3-ATAC to more than 70% for HyDrop when sequenced at 40,000 reads per cell. Finally, a significant portion of unique fragments in each cell did not overlap with peak regions and was thus not considered in analysis methods based on peak count matrices. The proportion of such fragments varied between 39%

(10x v2) and 82% (s3-ATAC) of all unique reads in cells. As a consequence of the above filtering steps, the fraction of original sequencing reads that are ultimately associated with cells, not duplicated and located within peak regions, can be surprisingly low, ranging between 28% (10x v2) and below 4% (s3-ATAC and HyDrop). These findings suggest that sequencing scATAC-seq experiments are generally highly inefficient, and protocol optimization steps should be performed to maximize cell quality and library complexity and minimize ambient chromatin contamination and PCR duplication. While both s3-ATAC and HyDrop attain a markedly lower sequencing efficiency than commercial assays, this reduced sensitivity was caused by different mechanisms; s3-ATAC samples contained many fragments outside of peak regions, whereas HyDrop fragments were highly duplicated.

Sensitivity and specificity

After initial filtering steps, the remaining cells exhibited stark differences in quality metrics across techniques. In terms of sensitivity, 10x v2 performed best, recovering, on average, 10,021 unique fragments in peaks per cell, which was significantly higher than Bio-Rad ddSEQ (4,228), HyDrop (1,180) and s3-ATAC (1,203; Fig. 1f and Extended Data Fig. 1d). TSS enrichment was also significantly stratified across methods with 10x v1.1, mtscATAC and s3-ATAC scoring 21.7 or lower, 10x v1, v2, multiome and HyDrop samples scoring between 25.2 and 27.6 and Bio-Rad ddSEQ scoring, on average, 32.6 (Fig. 1g and Extended Data Fig. 1f). The fraction of reads in peaks (FRIP) was significantly lower in HyDrop (38.5%), mtscATAC-seq (37.3%) and s3-ATAC (18.9%) than in Bio-Rad ddSEQ and the other 10x methods (57.3–62.7%; Fig. 1h and Extended Data Fig. 1e). Whereas the same method of nuclei extraction was used for all 10x and HyDrop samples (except for mtscATAC where Tween 20 and digitonin were omitted from the cell lysis buffer to retain mitochondrial chromatin), Bio-Rad ddSEQ and s3-ATAC experiments used a lysis buffer without NP-40 or Dounce homogenization in NIB-HEPES with 0.1% Tween and 0.1% NP-40 to extract nuclei, respectively. To investigate whether differences in FRIP, TSS enrichment or number of unique fragments between 10x/HyDrop, Bio-Rad ddSEQ and s3-ATAC could be attributed to differences in nuclei extraction, we performed the following three additional 10x v1.1 experiments: (1) in control experiment 1 (sample 10x v1.1c C1), we used a custom lysis buffer without NP-40 (as in Bio-Rad ddSEQ); (2) in control experiment 2 (sample 10x v1.1c C2), we extracted nuclei using Dounce homogenization (as described in s3-ATAC); and (3) in control experiment 3 (sample 10x v1.1 C3), we used the standard 10x v1.1 nuclei extraction protocol. Both control runs (C1 and C2) generated cells with FRIP and TSS enrichment scores on par with standard 10x v1.1 runs but retrieved a higher than average number of unique fragments. This suggests that the nuclei extraction method was not the sole driving factor causing reduced performance (for example, fragment numbers, FRIP and TSS enrichment) in Bio-Rad ddSEQ and s3-ATAC samples. Finally, we found that Bio-Rad ddSEQ samples had a higher median fragment length than all other techniques (Extended Data Fig. 1h).

With increasing sequencing depth, the number of unique fragments increased but saturated at different levels across techniques.

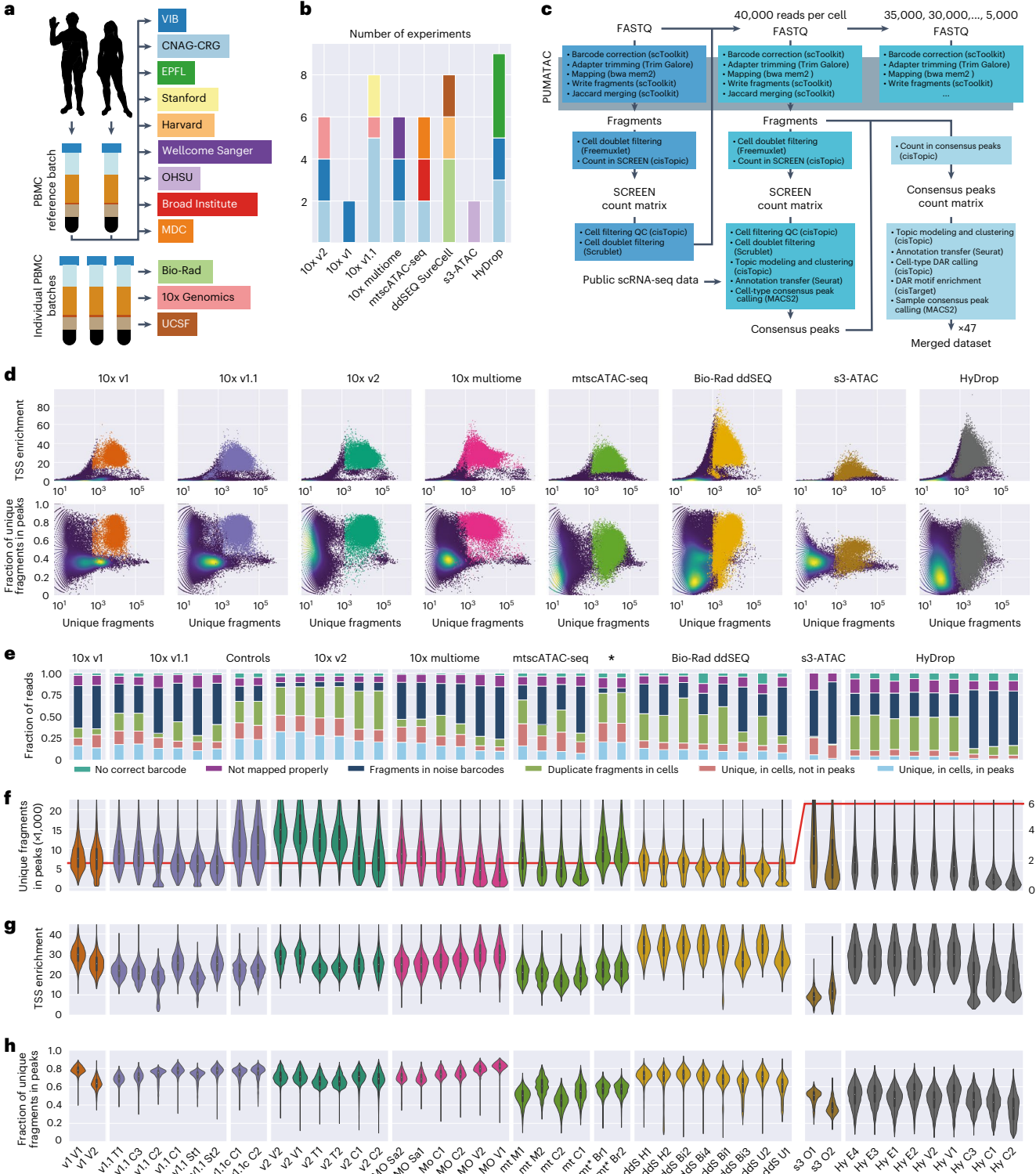
Fig. 1 | Overview of experimental design and low-level quality control metrics. **a**, Schematic overview of the experimental design; CNAG-CRG, Centro Nacional de Análisis Genómico; EPFL, École Polytechnique Fédérale de Lausanne; OHSU, Oregon Health & Science University; MDC, Max Delbrück Center for Molecular Medicine in the Helmholtz Association; UCSF, University of California San Francisco; VIB, Vlaams Instituut voor Biotechnologie. **b**, Bar chart of the number of experiments performed per technology colored by institute of origin. **c**, Diagram of the universal PUMATAC data analysis pipeline and further downstream analyses; QC, quality control. **d**, Distribution of TSS enrichment, FRIP and total unique fragment counts for all barcodes across all technologies. The blue, green and yellow color scale denotes local density. Saturated colors mark barcodes identified as cells. The distributions for individual samples

are shown in Extended Data Fig. 2. **e**, Stacked bar plot showing the fraction of reads lost across each step of data processing. 'Unique, in cells, in peaks' is the final fraction of sequencing reads retained in count matrices. Asterisk among technology names indicates mtscATAC-seq samples performed on PBMC that were viability FAC-sorted prior to tagmentation. **f–h**, Distributions of unique fragments in peaks (**f**), TSS enrichment (**g**) and fraction of unique fragments in peaks (**h**) in filtered cell barcodes. The scale was shifted to accommodate lower fragment counts in s3-ATAC and HyDrop, indicated by a red line denoting a value of 6,000; $n = 178,453$ cells (before doublet filtering) examined over 47 independent experiments. Median values are indicated by central white dots, quartiles are indicated by black boxes, and minima/maxima/centers are not indicated.

For example, at lower sequencing depths, ddSEQ samples performed better than 10x v1, v1.1 and multiome, an effect that was reversed at a higher sequencing depth (Fig. 2a). TSS enrichment was also dependent on sequencing depth but saturated rapidly (Fig. 2b). Sequencing efficiency decreased with higher sequencing depths due to the increase in duplicate reads (Fig. 2c).

We also investigated the performance of the scRNA-seq component of the six 10x multiome experiments. Therefore, we first analyzed

the unfiltered gene expression count matrices to define optimal thresholds of unique molecular identifiers (UMIs) for cell filtering (Extended Data Fig. 5a). We then downsampled the RNA-seq data to a common depth (28,417 reads per cell), realigned the downsampled data and compared the number of genes and unique fragments in peaks in the superset of barcodes identified in either the chromatin accessibility or gene expression component (Extended Data Fig. 5b). In four of six samples, scRNA-seq and scATAC-seq components correlated strongly



and identified the same barcodes. However, in two replicates (samples MO C1 and MO C2), almost 4,000 barcodes were exclusively identified in the RNA component (Extended Data Fig. 5c), suggesting that cell filtering based on both scRNA-seq and scATAC-seq metrics can generate false-negative cell calls in at least one modality.

Doublet counts and cell-type label transfer efficiency

We identified doublets (two cells with the same cell barcode) using simulated data or genotype information with Scrublet and Freemuxlet, respectively (Methods). Scrublet's doublet score linearly decreased with median unique fragments in peaks (Fig. 2f), and Freemuxlet's confidence (that is, difference in log likelihood between doublet and singlet assignments) increased with these metrics (Fig. 2g), suggesting that the number of fragments is a critical factor behind doublet detection. Outlier results for s3-ATAC O1, mtscATAC M1 and mtscATAC M2 could be explained by the increased read lengths (and thus increased single-nucleotide polymorphism (SNP) recovery) of 74 base pairs (bp) and 150 bp, respectively, as opposed to 50 bp for the other samples.

Next, we annotated cell types by inferring gene activities using chromatin accessibility in a genomic window around the TSS and subsequent label transfer from annotated scRNA-seq reference datasets using Seurat²². When assigning cell-type annotations, Seurat returns a confidence score to each assignment representing the majority percentage of a weighted vote based on cell types present in the reference. On average, a Seurat label transfer score of 0.7 corresponded to an 82% concordance between cell-type identities assigned to ATAC and RNA components of multiome samples, although this effect was cell-type dependent (Extended Data Fig. 6a,b). Similar to Scrublet and Freemuxlet, Seurat's confidence was strongly dependent on sequencing depth (Fig. 2d and Extended Data Fig. 7a) and the number of unique fragments per cell (Fig. 2h), confirming the number of fragments to be a strong predictor for the quality of downstream analyses. Consequently, both 10x and Bio-Rad ddSEQ methods attained high median label transfer scores, while HyDrop and s3-ATAC scored markedly lower (Fig. 2k and Extended Data Fig. 1g). The most common PBMC types, including B cells, T cells and CD14⁺ monocytes, were recovered by all techniques. However, the differences between methods became evident when focusing on rare cell types, such as DCs and NK cells, showing the lowest assignment scores across techniques (Fig. 2l and Extended Data Fig. 6). For the latter cell types, s3-ATAC and HyDrop recovered a lower fraction or no cells, suggesting that increased sensitivity is required for comprehensive cell-type annotation. Following cell-type assignment, we aggregated high-quality cell-type-specific peaks into consensus peaks. Across all samples and techniques, the fraction of reads within these consensus peaks was strongly correlated with TSS enrichment, affirming that TSS enrichment can be used as a predictor of signal-to-noise ratio before and independent of peak calling (Fig. 2i).

DARs

Similar to scRNA-seq, scATAC-seq data interpretation largely relies on the identification of signature features between cell types and states. Therefore, we calculated DARs between cell types within each of the 47 samples and evaluated each method's performance by comparing the number and median fold enrichment strength of DARs recovered across methods. These metrics strongly varied by cell type. In CD14⁺ monocytes, a cell type that was identified in all 47 samples, 10x methods recovered between 26,000 and 30,000 DARs and HyDrop recovered 29,000, whereas s3-ATAC and ddSEQ only recovered 17,000 and 15,000, respectively (Fig. 2m). A similar contrast was observed in the strength of these DARs; the top 2,000 CD14⁺ monocyte DARs recovered by 10x methods were a median of 17.7- to 23.5-fold enriched in CD14⁺ monocytes compared to the other cell types, while this fold enrichment was 10.6-fold, 7.5-fold and 4.9-fold in ddSEQ, HyDrop and s3-ATAC, respectively (Fig. 2n). Of note, DAR strength was positively

dependent on sequencing depth (Fig. 2e and Extended Data Fig. 7c), while the number of DARs recovered was not (Extended Data Fig. 7b).

Total cell counts and cell counts within each cell type were not equal across all 47 samples (Fig. 2l and Supplementary Fig. 2c). To investigate the effect of total cell count on Seurat label transfer scores and DAR strength, we downsampled each experiment to 2,500, 2,000, 1,500, 1,000 and 500 total cells. Here, both Seurat label transfer scores and DAR strength were positively affected by higher cell counts (Fig. 3a and Extended Data Fig. 8a,c), but the total number of DARs recovered remained largely unaffected (Extended Data Fig. 8b). To mitigate count-dependent biases, we recalculated cell-type-specific DARs in eight technology-exclusive sets sampled from the initially merged 169,000 cells. Each technology-exclusive set contained the lowest number of cells available for each cell type across technologies: 555 B cells, 747 CD14⁺ monocytes, 1,008 naive T cells, 1,769 cytotoxic T cells, 83 DCs, 126 NK cells and 144 CD16⁺ monocytes (Extended Data Fig. 4a). In this cell count-balanced analysis, 10x methods recovered more DARs than ddSEQ and HyDrop across all cell types (Fig. 3b). s3-ATAC also recovered a large number of DARs but of strongly decreased strength compared to other techniques (Fig. 3a). In terms of DAR strength, 10x v1 and v2 performed best, with ddSEQ performing on par with the remaining 10x methods.

Additionally, we calculated cell-type-specific DARs on a merged dataset containing all eight aforementioned cell-type-balanced subsets. A comparison of the scATAC-seq signal at the strongest DAR (Fig. 3d) or across the top 2,000 strongest DARs per cell type (Supplementary Fig. 3) showed general agreement in signals between the 10x methods and Bio-Rad ddSEQ. Both s3-ATAC and HyDrop showed weaker signals around DARs. Specifically, the signal was more broadly distributed around the DAR periphery in s3-ATAC, whereas the absolute signal was weaker but more specific in HyDrop. This observation agreed with earlier quality metrics indicating higher fragment count and lower FRIP for s3-ATAC and the inverse for HyDrop. We then calculated the overlap between these consensus DARs and the DARs recovered by each of the eight cell-type-balanced sets. Both 10x and Bio-Rad ddSEQ largely agreed with the strongest consensus peaks, sharing between 85% and 97% of these regions (Fig. 3e and Extended Data Fig. 4b). s3-ATAC and HyDrop recovered 60% and 70%, respectively. However, ddSEQ recovered a significantly lower number of DARs from the consensus DAR set (40%; Fig. 3f). Across all eight individual ddSEQ samples, DARs were also positioned closer to TSSs than in all other techniques (Fig. 2j and Extended Data Fig. 4c). Combined with ddSEQ's significantly higher TSS enrichment across all fragments, this suggests that ddSEQ recovers TSS-proximal DARs in agreement with 10x but lacks enrichment in TSS-distal fragments. In the merged set, we also calculated DARs between cell populations of the same cell type but originating from different techniques. Here, only s3-ATAC returned significantly enriched regions, finding a total of 193,913 regions that were more accessible in s3-ATAC than in the other techniques.

To evaluate the biological relevance of the DARs identified by each protocol, we performed a motif enrichment analysis on each cell type's top 2,000 strongest DARs by using pyCisTarget^{24–26}. Across all cell types, DARs from 10x methods scored highest in normalized enrichment scores of transcription factor motifs, with mean scores of higher than 10 across the top 50 most enriched motifs. ddSEQ, s3-ATAC and HyDrop scored significantly lower, with mean scores of 6–8. This stratification was conserved when examining cell-type-specific enriched motifs (Fig. 3g).

Interpretation, integration and validation

Above, we calculated DARs between cell types within a sample. For biological discoveries, often different samples are compared to identify genomic regions with differential accessibility in certain conditions (for example, disease versus healthy and knockout versus control). To compare each method's ability to detect variation between samples,

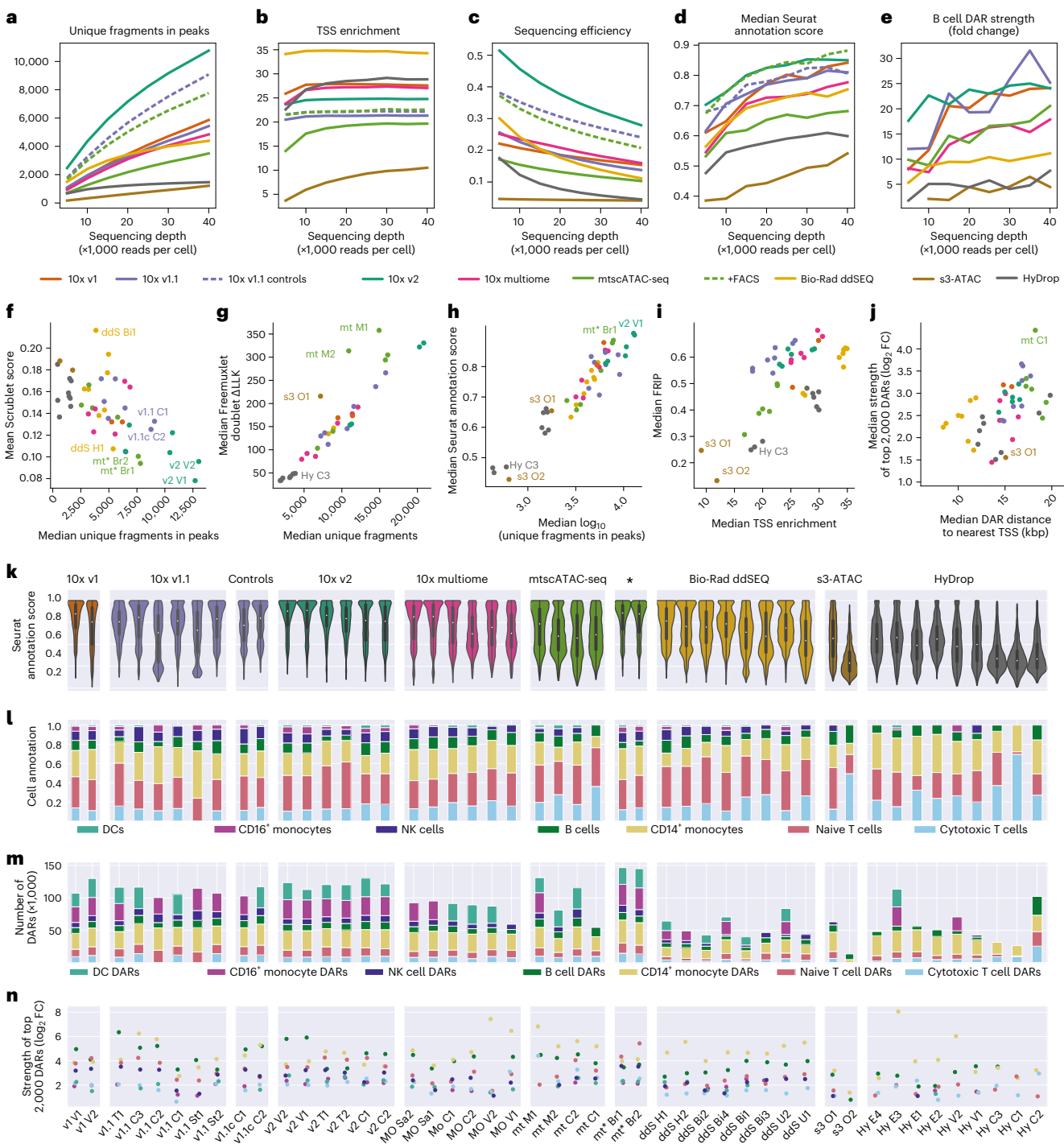


Fig. 2 | Differences in automated cell-type annotation accuracy and differential region calling between techniques. a–e. Line graphs showing the effect of sequencing depth (×1,000 reads per cell) on unique fragments in peak regions (a), TSS enrichment (b), sequencing efficiency (fraction of reads that are associated with filtered cells, not duplicated and located within a peak region; c), median Seurat label transfer score (d) and fold change enrichment of B cell DARs (e). **f.** Scatter plot of mean Scrublet score and median unique number of fragments in peaks across cells. **g.** Scatter plot of the median LLK (log likelihood) of Freemuxlet’s per cell doublet classification and median unique number of fragments across cells. **h.** Scatter plot of median Seurat score and median log₁₀ of unique fragments in peaks in cells. **i.** Scatter plot of median FRIP and median TSS enrichment across cells in each sample. **j.** Scatter plot

of log₂ (fold change) (log₂ (FC)) enrichment of the top 2,000 DARs across cell types and median distance of DARs to the nearest TSS (bp). Sample s3 O2 is omitted due to it being an outlier with a median distance of >300,000 bp; kbp, kilobase pairs. In f–j, each point represents one experiment, and points are colored by technique. **k.** Distributions of Seurat scores across samples. **l.** Fraction of cell types recovered in cells from individual samples; *n* = 169,156 cells (after doublet filtering) examined over 47 independent experiments. Median values are indicated by central white dots, quartiles are indicated by black boxes, and minima/maxima/centers are not indicated. **m.** Number of DARs recovered per cell type across samples. **n.** log₂ (fold change enrichment) of the top 2,000 DARs colored by cell type.

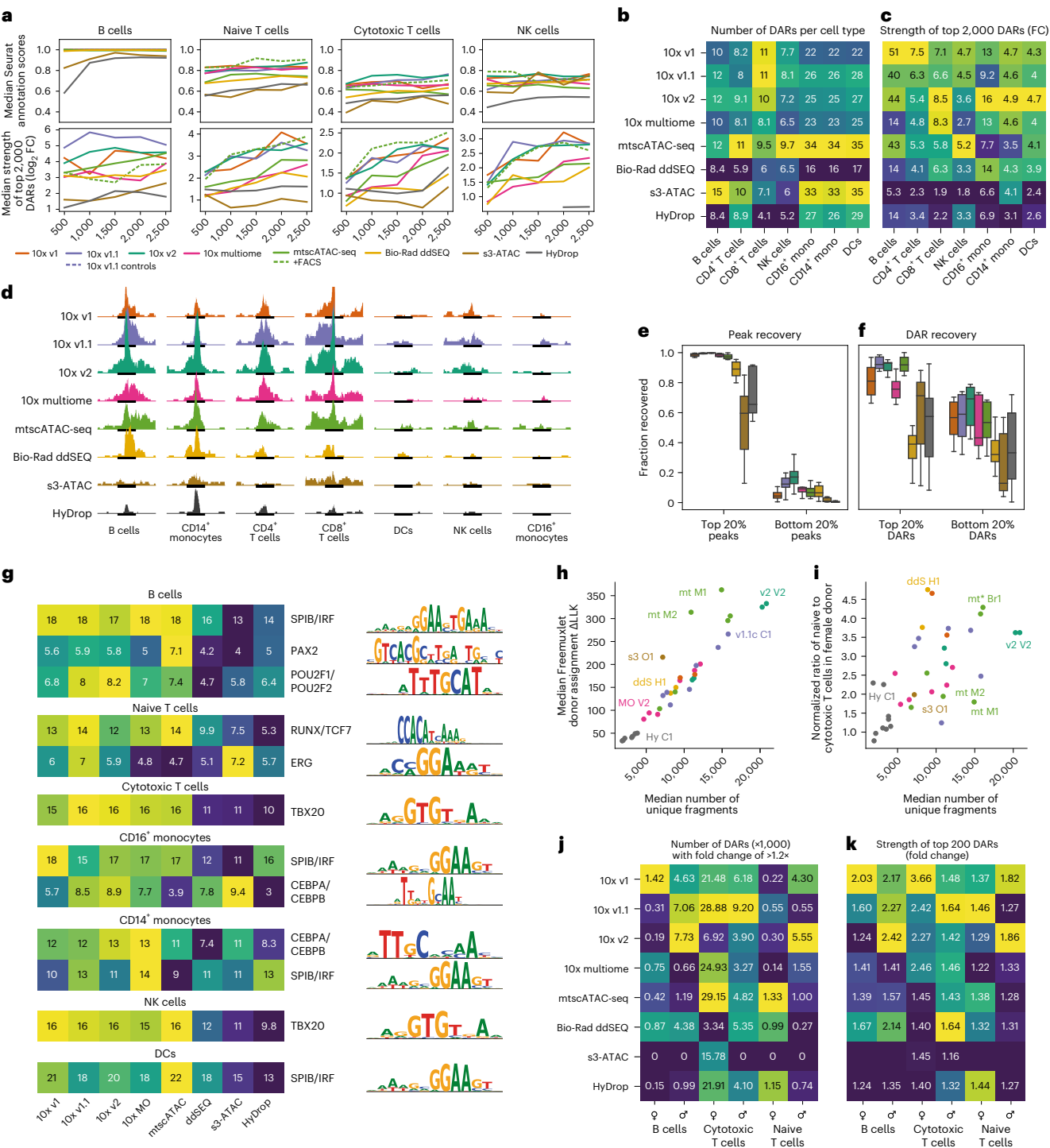


Fig. 3 | Performance differences in detection of motif enrichment and sexual dimorphisms. a, Dependency of Seurat label transfer scores and average $\log_2$ (fold change) of the top 2,000 DARs in selected cell types on the total number of cells in selected cell types. **b, c**, Heat map of the number of DARs (**b**) and heat map of the fold change enrichment (**c**) of the top 2,000 DARs sourced from the cell-type fair merged sets for every technique across cell types. Colors are scaled per column; mono, monocytes. **d**, Fragment coverage within each cell type's strongest common DAR found in the merged set. Each cell type contains an equal number of cells across technologies. All tracks are scaled to the same absolute coverage. The 10x v1 track is slightly truncated to accommodate **j**. **e, f**, Fraction of the top 20% and bottom 20% peaks and DARs found by the

merged cell-type fair set recovered in subsets from individual technologies. **g**, Heat map of the normalized enrichment score of cell-type-specific transcription factor motifs. Colors are scaled per row. **h**, Scatter plot of median Freemuxlet donor assignment log likelihood difference to second-best guess and median number of unique fragments. **i**, Scatter plot of the ratio of naive T cells to cytotoxic T cells in male and female subpopulations and median number of unique fragments. In **h** and **i**, each point represents one experiment, and points are colored by technique. **j**, Heat map of the number of sex-specific DARs. Colors are scaled per column. **k**, Heat map of fold change enrichment of the 200 strongest sex-specific DARs. Colors are scaled per column.



Fig. 4 | Overview of the merged dataset. a, Genome tracks of sex-specific DARs in B cells and cytotoxic and naive T cells showing coverage in male and female subpopulations. Coverage uses a common scale across techniques within each cell type and is cell count normalized (cell counts are equal for each sex within techniques); Chr, chromosome. **b**, Top, t-distributed stochastic neighbor embedding (t-SNE) projection of all 167,000 filtered cells across all 47 experiments colored by technology and cell type. Bottom, batch-corrected t-SNE using technology of origin as the batch variable. Data points were randomly shuffled before plotting. **c**, First two principal components of all experiments based on 15 basic quality metrics. Red links indicate same day replicate experiments. Caption denotes technique followed by center of sample origin (B for BioRad, Br for Broad, C for CNAG-CRG, E for EPFL, H for Harvard, M for MDC, O for OHSU, Sa for Sanger, St for Stanford, T for 10x Genomics, U for UCSF, and V for VIB), and finally an index (either A or B) to denote the replicate set if more than one set of replicates was performed at the same center. **d**, Local inverse Simpson's index values before and after Harmony batch correction.

where differences can be more subtle than between cell types, we focused on differences observed between male and female samples. For donor identification, we first applied Freemuxlet for genotype demultiplexing²⁰ (Supplementary Fig. 2a–c) before assigning each sample's sex through count numbers on the X and Y chromosomes (Extended Data Fig. 2b). Similar to the previously applied doublet calling, Freemuxlet's donor assignment confidence was strongly correlated with the number of fragments (Fig. 3h). From a biological perspective, the ratio of naive to cytotoxic T cells was higher in the female sample than in male cells across all techniques, an effect dependent on the number of unique fragments recovered per cell (Fig. 3i). Sensitivity will thus directly impact studies that assume correct donor identification. We then used the aforementioned cell count-balanced strategy to calculate DARs between male and female samples for each cell type. Here, differences were markedly more subtle than between cell types, with median DAR enrichments of around 2-fold (Fig. 3k) as opposed to 5-fold in DCs or 50-fold in B cells (Fig. 3c). The strongest sexual dimorphism was observed in cytotoxic T cells, where sex-specific DARs were strongly enriched and highest in number (Figs. 3j,k and 4a). In agreement with these findings, sex differences in T cell abundance and chromatin accessibility have been previously reported in the context of naive/cytotoxic T cell counts and distinct response to extrinsic stimulation²⁷. Similar to results obtained for cell-type-specific DARs, ddSEQ captured fewer and weaker DARs than the 10x methods, although the differences were less pronounced. Both s3-ATAC and HyDrop recovered even fewer and weaker or no sex-specific DARs (Fig. 3j,k).

Next, we tested each method's ability to be integrated into joint datasets, a key requirement for decentralized multicenter projects, such as the Human Cell Atlas project. From the peak sets of the 47 individual downsampled samples, we derived a consensus set of 198,421 peaks and recounted all downsampled data in this common peak set to generate a complete merged dataset of 178,502 cells. After filtering for doublets, we performed dimensionality reduction on the remaining 169,227 cells using cisTopic (Fig. 4b). We then quantified each technology's ability to cocluster with other techniques using the local inverse Simpson's index, depicting the average number of technologies in a cell's neighborhood (Fig. 4d)²⁸. Before batch effect correction, cells clustered by technology, even across the different 10x variants.

After data integration using Harmony, differences were partially remediated, although ddSEQ and HyDrop integrated worse than the other methods (Fig. 4b–d). Additionally, we performed a principal-component analysis on 15 key quality metrics (Fig. 4c). This dimension-reduced representation exhibited a radial axis of ‘quality’ centered around the bottom lefthand corner and showed that technical replicate experiments produced highly similar datasets.

In addition to the systematic benchmarking using PBMCs, we applied our universal data analysis strategy to publicly available adult mouse cortex scATAC-seq data for each of the eight technologies. Therefore, the raw sequencing data were aligned to the mouse reference genome, downsampled to equal cell numbers and clustered using cisTopic before generating consensus peak sets for each sample. We then calculated basic quality metrics, such as unique fragment numbers in peaks, and repeated these steps for several sequencing depth downsampling levels. In all metrics, 10x and ddSEQ performed markedly better than HyDrop and s3-ATAC (Extended Data Fig. 9a,b). However, in this tissue context, s3-ATAC recovered more unique fragments in peaks than HyDrop at all levels of downsampling. Surprisingly, considering previous results, 10x v1 recovered more unique fragments in peaks than 10x v2, the latter of which now performed on par with the best Bio-Rad ddSEQ samples. While our PBMC benchmark sought to systematically eliminate sample preparation effects by using cells in suspension as a reference sample, we cannot entirely disentangle the effect of sample (nuclei) preparation differences on technology performance in more complex tissues, such as the mouse cortex. Here, tissue excision and nuclei extraction require more complex procedures and expertise, which may be a more defining factor for data quality than the intrinsic sensitivity of each technology.

Protocol economics

To compare costs per sample across the tested methods, we iteratively downsampled the fragments files for each sample and calculated the fraction of duplicate fragments per iteration. We then calculated the average sequencing depth at which a duplication rate of 50% was attained and quantified the number of fragments at this depth (including non-peak fragments; Extended Data Fig. 10). These findings provided an estimated total cost for a 5,000-cell experiment for each technique. In this regard, the costs per cell were fivefold and tenfold cheaper for HyDrop than for Bio-Rad ddSeq and 10x v2, respectively; 10x multiome was more expensive (1.5-fold) than the stand-alone assay 10x v2 but was markedly less sensitive and less efficient in sequencing. Thus, the matched gene expression information of the multiome assay results in a considerable additional cost, which has to be taken into account in designs for which scATAC-seq is the primary focus of a study. The cost per experiment for s3-ATAC (\$800) ranged between HyDrop (\$100) and Bio-Rad ddSEQ (\$1,100), but the high library complexity resulted in high sequencing costs and the highest overall costs per cell. However, several variables can be tuned to reduce cost or to increase efficiency. For example, 10,000 cells (instead of 5,000 cells) can be loaded on the same microfluidic lane to double cost efficiency in the droplet-based methods (10x, Bio-Rad ddSEQ and HyDrop) or s3-ATAC samples can be sequenced at lower depths to reduce sequencing costs. All technologies can be sequenced at increased depths to improve sensitivity at the expense of cost efficiency.

Discussion

Technology benchmarking studies with multicenter designs require thorough experimental planning to limit variability to the tested protocols, while keeping other factors constant. In this regard, we used PBMCs as a reference sample, which is ideal for multicenter benchmarking studies as cells can be aliquoted, stored, shipped and thawed without major losses in sample quality and composition. However, although such unified reference samples allowed us to include expert laboratories and companies around the world, we cannot exclude

different performances of the methods in other tissues. Although the analysis of mouse brain datasets pointed to largely consistent method performance, a protocol could outperform other methods depending on the cell type and tissue context. Data analysis represents an additional variable in benchmarking studies, as protocol-specific pipelines harbor quality control and filtering steps that challenge comparative results if not synchronized. Therefore, we developed PUMATAC, a unified data preprocessing pipeline, which was used to process data types from all methods tested here. Importantly, its modular design allows the extension to future technologies and assay variants, making it a valuable software resource to benchmark the next generation of scATAC-seq methods.

Our reference sample resource and unified data processing pipeline allowed a systematic comparison of different scATAC-seq methods. Methods broadly agreed on cell-type identity and transcription factor activities but showed stark differences in sequencing library quality and tagmentation specificity to open chromatin sites. In general, HyDrop and s3-ATAC performed markedly lower in most quality control metrics. HyDrop captured significantly fewer fragments than 10x and Bio-Rad methods. s3-ATAC fragments were less likely to be enriched around TSSs, and high duplicate rates were observed in HyDrop, suggesting room for optimization in PBMC samples for these non-commercial technologies. Although the 10x protocols we tested apply similar chemistries, we see variable performance between variants and replicates. While the former points to the potential impact of buffer composition (for example, specific to mtscATAC and 10x multiome), the latter suggests differences due to sample handling. Nevertheless, despite the difference in number of unique fragments detected, 10x v1 and v2 scored equally across higher-level quality metrics, such as label transfer and motif enrichment scores, attesting to the high specificity and signal-to-noise ratio of the 10x v1 scATAC-seq kits. Bio-Rad ddSEQ samples returned weaker and fewer DARs, which in turn recovered weaker transcription factor motif signatures. Bio-Rad ddSEQ fragments were also less specific to TSS-distal accessible sites, which could also contribute to reduced integration capacity with other methods, an effect that could only partially be resolved using batch effect correction algorithms.

We found sequencing efficiency to be generally low for scATAC-seq experiments. Two strategies could mitigate this issue: optimized sample preparation and nuclei extraction protocols to minimize the amount of ambient chromatin in samples, potentially applying FACS for sample cleanup, and sequencing below library saturation to limit the number of duplicate reads. Generally, before embarking into large-scale production, the higher reagent costs of commercialized methods have to be considered in light of lower complexity and accuracy for non-commercial variants. The aforementioned factors plus a method’s accessibility and dataset integrability across studies should eventually drive the technology selection process. It is important to state that non-commercialized methods have not gone through a rigorous optimization process. Time and budget constraints of academic labs often limit excessive protocol optimization procedures. However, their open-source character and low reagent costs make them available for community-driven improvements and low-budget data generation effort, respectively. Therefore, we encourage researchers to continue striving for new creative solutions as the driving force for next-generation technologies.

In addition to evaluating different scATAC-seq methods, this work provides resources for the single-cell genomics community. Our PUMATAC pipeline is openly available to be used as an open-source alternative to commercial software and is flexible to analyze current and future data types. All code to reproduce all of our analyses and graphs is included with the pipeline. All benchmarking datasets can be downloaded in raw and processed formats for the testing and development of computational algorithms, for example, data integration tools. Finally, the here-derived set of consensus DARs across different

technologies represents a high-confidence set of candidate enhancers and promoters underlying cell identity as a resource for further exploration.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41587-023-01881-x>.

References

- Massoni-Badosa, R. et al. Sampling time-dependent artifacts in single-cell genomics studies. *Genome Biol.* **21**, 112 (2020).
- Mereu, E. et al. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat. Biotechnol.* **38**, 747–755 (2020).
- Minnoye, L. et al. Chromatin accessibility profiling methods. *Nat. Rev. Methods Primer* **1**, 11 (2021).
- Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
- Cusanovich, D. A. et al. A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell* **174**, 1309–1324 (2018).
- Domcke, S., et al. A human cell atlas of fetal chromatin accessibility. *Science* **370**, eaba7612 (2020).
- Zhang, K. et al. A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184**, 5985–6001 (2021).
- Hulselmans, G., De Rop, F. & Flerin, C. Pipeline for universal mapping of ATAC-seq. *Zenodo* <https://doi.org/10.5281/zenodo.7764884> (2023).
- Satpathy, A. T. et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019).
- Lareau, C. A. et al. Massively parallel single-cell mitochondrial DNA genotyping and chromatin profiling. *Nat. Biotechnol.* **39**, 451–461 (2021).
- Lareau, C. A. et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.* **37**, 916–924 (2019).
- De Rop, F. V. et al. Hydrop enables droplet-based single-cell ATAC-seq and single-cell RNA-seq using dissolvable hydrogel beads. *eLife* **11**, e73971 (2022).
- Mulqueen, R. M. et al. High-content single-cell combinatorial indexing. *Nat. Biotechnol.* **39**, 1574–1580 (2021).
- Flerin, C. C., Davie, K., Hulselmans, G. & Waegeneer, M. D. vib-singlecell-nf/vsn-pipelines: v0.27.0. *Zenodo* <https://zenodo.org/record/5751297> (2021).
- Lareau, C. A., Ma, S., Duarte, F. M. & Buenrostro, J. D. Inference and effects of barcode multiplets in droplet-based single-cell assays. *Nat. Commun.* **11**, 866 (2020).
- Bravo González-Blas, C. et al. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods* **16**, 397–400 (2019).
- Ou, J. et al. ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data. *BMC Genomics* **19**, 169 (2018).
- The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
- Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst.* **8**, 281–291 (2019).
- Zhang, F., Kang, H. M. & Yun, Y. popsicle. *GitHub* <https://github.com/statgen/popsicle> (2019).
- Ding, J. et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* **38**, 737–746 (2020).
- Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
- Zhang, Y. et al. Model-based analysis of ChIP-seq (MACS). *Genome Biol.* **9**, R137 (2008).
- Herrmann, C., Van de Sande, B., Potier, D. & Aerts, S. i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules. *Nucleic Acids Res.* **40**, e114 (2012).
- Imrichová, H., Hulselmans, G., Kalender Atak, Z., Potier, D. & Aerts, S. i-cisTarget 2015 update: generalized cis-regulatory enrichment analysis in human, mouse and fly. *Nucleic Acids Res.* **43**, W57–W64 (2015).
- González-Blas, C. B. et al. SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat. Methods* <https://doi.org/10.1038/s41592-023-01938-4> (2023).
- Klein, S. L. & Flanagan, K. L. Sex differences in immune responses. *Nat. Rev. Immunol.* **16**, 626–638 (2016).
- Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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 license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license 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 license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023

¹VIB Center for Brain and Disease Research, Leuven, Belgium. ²Department of Human Genetics, KU Leuven, Leuven, Belgium. ³Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain. ⁴CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. ⁵Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA. ⁶Broad Institute of MIT and Harvard, Cambridge, MA, USA. ⁷Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, USA. ⁸Gene Regulation Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA. ⁹Department of Genetics, Stanford University, Stanford, CA, USA. ¹⁰Gladstone Institute of Neurological Disease, San Francisco, CA, USA. ¹¹Wellcome Sanger Institute, Cambridge, UK. ¹²Department of Molecular and Medical Genetics, Oregon Health & Science University, Portland, OR, USA. ¹³Division of Hematology & Medical Oncology, Knight Cancer Institute, Oregon Health & Sciences University, Portland, OR, USA. ¹⁴Division of Oncologic Sciences, Knight Cancer Institute, Oregon Health & Sciences University, Portland, OR, USA. ¹⁵Berlin Institute of Health at Charité–Universitätsmedizin Berlin, Berlin, Germany. ¹⁶Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin Institute for Medical Systems Biology (BIMSB), Berlin, Germany. ¹⁷Laboratory of Systems Biology and Genetics, Institute of Bioengineering, School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. ¹⁸Swiss Institute of Bioinformatics (SIB), Lausanne,

Switzerland. ¹⁹Universitat Pompeu Fabra (UPF), Barcelona, Spain. ²⁰Vitalant Research Institute, San Francisco, CA, USA. ²¹Department of Laboratory Medicine, University of California, San Francisco, CA, USA. ²²Adelaide Centre for Epigenetics and the South Australian Immunogenomics Cancer Institute, Faculty of Health and Medical Sciences, The University of Adelaide, Adelaide, South Australia, Australia. ²³University of Melbourne Centre for Cancer Research, Victoria Comprehensive Cancer Centre, Melbourne, Victoria, Australia. ²⁴Digital Biology Group, Bio-Rad, Pleasanton, CA, USA. ²⁵Departament de Fonaments Clínics, Facultat de Medicina, Universitat de Barcelona, Barcelona, Spain. ²⁶Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, Spain. ²⁷Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. ²⁸ICREA, Barcelona, Spain. ²⁹Department of Physics/Cavendish Laboratory, University of Cambridge, Cambridge, UK. ³⁰Chan Zuckerberg Biohub, San Francisco, CA, USA. ³¹Koch Institute of Integrative Cancer Research, Cambridge, MA, USA. ³²Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA. ³³These authors jointly supervised this work: Stein Aerts, Holger Heyn. ✉e-mail: stein.aerts@kuleuven.be; holger.heyne@cnag.crg.eu

Methods

Sample preparation

Human PBMC thawing and nuclei isolation. Cryopreserved human PBMCs from one male donor and one female donor were purchased from AllCells and distributed across institutes to generate the following samples: v1 V1, v1 V2, v1.1 C1–v1.1 C3, v1.1 St1, v1.1 St2, v1.1c C1, v1.1c C2, v2 V1, v2 V2, v2 C1, v2 C2, MO Sa1, MO Sa2, MO C1, MO C2, MO V1, MO V2, mt M1, mt M2, mt C1, mt C2, mt* Br1, mt* Br2, ddS H1, ddS H2, s3 O1, Hy E1–Hy E4, Hy V1, Hy V2 and Hy C1–Hy C3. For the remaining samples (v1.1 T1, v2 T1, v2 T2, ddS B1–ddS B4, ddS U1, ddS U2 and s3 O2), locally available cryopreserved PBMCs were used. In these short identifiers, the first part indicates the technology used, the second part indicates the first one or two letters from the center where the experiment was performed, and the number identifies each technical replicate.

Unless specified otherwise in technology-specific methods sections (for Bio-Rad ddSEQ, mtscATAC-seq, 10x v1.1 control runs and s3-ATAC), cryopreserved PBMCs were thawed according to the 10x Genomics demonstrated protocol CG00039 ('Fresh Frozen Human Peripheral Blood Mononuclear Cells for Single Cell RNA Sequencing'). Briefly, 1 ml of frozen cells was rapidly thawed in a water bath at 37 °C and transferred to a 50-ml tube using a 1,000- μ l wide-bore tip. Next, 1 ml of medium prewarmed to 37 °C and supplemented with 10% fetal bovine serum (FBS; Thermo Fisher Scientific) was added dropwise with gentle swirling of the sample. After 1 min of incubation at room temperature, 2, 4, 8 and 16 ml of medium with 10% FBS were added dropwise with 1 min of incubation at room temperature in between. The cell suspension was then centrifuged at 300g for 5 min at room temperature. The pellet was resuspended in 10 ml of medium supplemented with 10% FBS, and cells were counted. Unless specified otherwise in the technology-specific methods sections, the isolation of nuclei was performed according to the 10x Genomics demonstrated protocol 'Nuclei Isolation for Single Cell ATAC Sequencing'. Briefly, 1 million cells from the cell mix were transferred to a 1.5-ml microcentrifuge tube and centrifuged at 500g for 5 min at 4 °C. The supernatant was removed without disrupting the cell pellet, and 100 μ l of chilled lysis buffer (10 mM Tris-HCl (pH 7.4), 10 mM NaCl, 3 mM MgCl₂, 0.1% Tween 20, 0.1% NP-40 substitute, 0.01% digitonin and 1% bovine serum albumin (BSA)) was added and mixed by pipetting ten times. Samples were then incubated on ice for 3 min. Following lysis, 1 ml of chilled wash buffer (10 mM Tris-HCl (pH 7.4), 10 mM NaCl, 3 mM MgCl₂, 0.1% Tween 20 and 1% BSA) was added and mixed by pipetting. Nuclei were centrifuged at 500g for 5 min at 4 °C, and the supernatant was removed without disrupting the nuclei pellet. Based on the starting number of cells and assuming a 50% loss during the procedure, nuclei were resuspended into the appropriate volume of chilled diluted Nuclei Buffer (10x Genomics) to achieve a concentration of 925–2,300 nuclei per μ l, suitable for a target recovery of 3,000 nuclei. This combination of PBMC thawing and nuclei isolation was used for all 10x samples (except mtscATAC-seq protocols, v1.1 control runs and v1.1 St1 and v1.1 St2 samples) and all HyDrop samples, but not for s3-ATAC and Bio-Rad ddSEQ samples. The method of cell counting was performed differently depending on the center of origin. For all samples generated in VIB, cells and nuclei were counted using a LUNA automated cell counter (Logos Biosystems). For Stanford and Sanger samples, cells and nuclei were counted manually using a hemocytometer. For all Bio-Rad ddSEQ and s3-ATAC and all CNAG samples, cells and nuclei were counted using a TC20 cell counter (Bio-Rad). For Broad samples and samples generated by the company 10x Genomics, cells and nuclei were counted using a Countess II or III FL automated cell counter (Thermo Fisher).

10x ATAC v1 (short identifiers v1 V1 and v1 V2). PBMCs were thawed, and nuclei were isolated as described above. Two technical replicates were generated on the same day starting from the same freshly thawed nuclei suspension. scATAC-seq libraries were prepared according to the Chromium Single Cell ATAC reagent kits v1.0 user guide (10x Genomics,

CG000001 Rev D). Briefly, the transposition reaction was prepared by mixing the desired number of nuclei with ATAC Buffer (10x Genomics) and ATAC Enzyme (10x Genomics) and incubated for 60 min at 37 °C; 4,590 nuclei were loaded with the goal of recovering 3,000 nuclei. Nuclei were partitioned into Gel Bead-in-Emulsions (GEMs) by using the Chromium Controller (Chip E). DNA linear amplification was then performed by incubating the GEMs under the following thermal cycling conditions: 72 °C for 5 min, 98 °C for 30 s and 12 cycles of 98 °C for 10 s, 59 °C for 30 s and 72 °C for 1 min. GEMs were broken using Recovery Agent (10x Genomics), and the resulting DNA was purified by sequential Dynabeads and SPRIselect reagent beads cleanups. Libraries were indexed by PCR using a Single Index kit (Plate N) and incubated under the following thermal cycling conditions: 98 °C for 45 s and ten cycles of 98 °C for 20 s, 67 °C for 30 s and 72 °C for 20 s with a final extension of 72 °C for 1 min. Sequencing libraries were subjected to a final bead cleanup with SPRIselect reagent.

Samples v1 V1 and v1 V2 were sequenced on a NovaSeq 6000 using a NovaSeq SP kit (100 cycles; 20028401, Illumina), and sequencing was performed using the following read protocol: 50 cycles (read 1), 8 cycles (i7 index read), 16 cycles (i5 index read) and 49 cycles (read 2).

10x ATAC v1.1 (short identifiers v1.1 C1–v1.1 C3, v1.1 T1, v1.1 St1 and v1.1 St2). PBMCs were thawed, and nuclei were isolated as described above for samples v1.1 C1–v1.1 C3 and v1.1 T1. For samples v1.1 St1 and v1.1 St2, a different thawing/isolation protocol was used. Here, each cryopreserved PBMC sample was thawed in 50 ml of thaw medium (IMDM, 10% FBS and 200 Kunitz U ml⁻¹ DNase) preheated to 37 °C and incubated for 15 min at 37 °C. DNase was ordered from Worthington Biochem (LS002007) and resuspended in HBSS at 20,000 U ml⁻¹ (100 $\times$ stock). Cells were pelleted at 300g (1,200 r.p.m.), resuspended in 5 ml of thaw medium and layered over 5 ml of Ficoll in a 15-ml conical tube. Cells were then spun at 500g (1,500 r.p.m.) with no brake for 30 min at room temperature in a swinging-bucket centrifuge. 2 mL of the mononuclear cell layer was collected and diluted with 10 ml of room temperature PBS. Cells were put on ice and maintained at 4 °C until use.

Technical replicates were generated on the same day starting from the same freshly thawed nuclei suspension. scATAC-seq libraries were prepared according to the Chromium Single Cell ATAC reagent kit v1.1 user guide (10x Genomics, CG000209 Rev D). Briefly, the transposition reaction was prepared by mixing the desired number of nuclei with ATAC Buffer (10x Genomics) and ATAC Enzyme (10x Genomics) and was then incubated for 60 min at 37 °C; 4,590 nuclei were loaded with the goal of recovering 3,000 nuclei. For sample '10x v1.1 V2', 9,180 nuclei were loaded instead of 4,590 due to a counting error. Nuclei were partitioned into GEMs by using a Chromium Controller with Chip H. DNA linear amplification was then performed by incubating the GEMs under the following thermal cycling conditions: 72 °C for 5 min, 98 °C for 30 s and 12 cycles of 98 °C for 10 s, 59 °C for 30 s and 72 °C for 1 min. GEMs were broken using Recovery Agent (10x Genomics), and the resulting DNA was purified by sequential Dynabeads and SPRIselect reagent beads cleanups. Libraries were indexed by PCR using a Single Index kit N set A (10x Genomics, PN-1000212) and incubated under the following thermal cycling conditions: 98 °C for 45 s and ten cycles of 98 °C for 20 s, 67 °C for 30 s and 72 °C for 20 s with a final extension of 72 °C for 1 min. Sequencing libraries were subjected to a final bead cleanup with SPRIselect reagent.

Samples v1.1 St1 and v1.1 St2 were sequenced on an Illumina Next-Seq 500 machine using a high-output flow cell with 34 bp paired-end reads. Samples v1.1 C1–v1.1 C3 and v1.1 T1 were sequenced on an Illumina NovaSeq 6000 with the following sequencing conditions: 50 bp (read 1), 8 bp (i7 index), 16 bp (i5 index) and 49 bp (read 2).

10x ATAC v2 (short identifiers v2 V1, v2 V2, v2 T1, v2 T2, v2 C1 and v2 C2). PBMCs were thawed, and nuclei were isolated as described above. Technical replicates were generated on the same day starting

from the same freshly thawed nuclei suspension. scATAC-seq libraries were prepared according to the Chromium Single Cell ATAC reagent kits v2 user guide (10x Genomics, CG000496 Rev B). Briefly, the transposition reaction was prepared by mixing the desired number of nuclei with ATAC Buffer (10x Genomics) and ATAC Enzyme (10x Genomics) and was then incubated for 30 min at 37 °C; 4,590 nuclei were loaded with a goal of recovering 3,000 nuclei. Nuclei were partitioned into GEMs by using a Chromium Controller with Chip H. Sample v2 T1 was the only sample for which Chromium X was used. DNA linear amplification was then performed by incubating the GEMs under the following thermal cycling conditions: 72 °C for 5 min, 98 °C for 30 s and 12 cycles of 98 °C for 10 s, 59 °C for 30 s and 72 °C for 1 min. GEMs were broken using Recovery Agent (10x Genomics), and the resulting DNA was purified by sequential Dynabeads and SPRIselect reagent beads cleanups. Libraries were indexed by PCR using a Single Index kit N set A and incubated under the following thermal cycling conditions: 98 °C for 45 s and eight cycles of 98 °C for 20 s, 67 °C for 30 s and 72 °C for 20 s with a final extension of 72 °C for 1 min. Sequencing libraries were subjected to a final bead cleanup with SPRIselect reagent.

Samples v2 v1 and v2 v2 were sequenced on an Illumina NextSeq 2000 under the following sequencing conditions: 50 bp (read 1), 8 bp (i7 index), 16 bp (i5 index) and 50 bp (read 2). Samples v2 C1, v2 C2, v2 T1 and v2 T2 were sequenced on an Illumina NovaSeq 6000 under the following sequencing conditions: 50 bp (read 1), 8 bp (i7 index), 16 bp (i5 index) and 49 bp (read 2).

10x multiome (short identifiers MO Sa1, MO Sa2, MO C1, MO C2, MO V1 and MO V2). PBMCs were thawed as described above. The isolation of nuclei was slightly different, including the use of RNase inhibitors to ensure RNA quality. Briefly, two pools of cells (technical replicates) were generated from the two donors by mixing 500,000 cells per donor, totaling 1 million cells per pool. Cells were pelleted for 5 min at 300g and 4 °C and were washed twice in 1 ml of wash buffer (10 mM Tris-HCl (pH 7.4), 10 mM NaCl, 3 mM MgCl₂, 1% BSA, 0.1% Tween 20, 1 mM DTT and 1 U μl⁻¹ RNase inhibitor). After the second wash and final centrifugation, cells were resuspended in 0.1 ml of chilled lysis buffer (10 mM Tris-HCl (pH 7.4), 10 mM NaCl, 3 mM MgCl₂, 0.1% Tween 20, 0.1% NP-40, 0.01% digitonin, 1% BSA and 1 mM DTT) and incubated for 3 min on ice. Nuclei were washed three times in 1 ml of wash buffer (10 mM Tris-HCl (pH 7.4), 10 mM NaCl, 3 mM MgCl₂, 1% BSA, 0.1% Tween 20, 1 mM DTT and 1 U μl⁻¹ RNase inhibitor) by centrifuging at 500g for 5 min.

After the last centrifugation, cells were resuspended in chilled Nuclei Buffer (1× Nuclei Buffer, 1 mM DTT and 1 U μl⁻¹ RNase inhibitor) calculated and loaded according to the Chromium Next GEM Single Cell Multiome ATAC + GEX user guide (protocol CG000338 Rev A); 4,590 nuclei were loaded with the goal of recovering 3,000 nuclei. For loading onto 10x chips, we sought to recover 3,000 nuclei. Following the isolation of nuclei and transposition, GEMs were generated using GEM Chip J. GEM cleanup and preamplification PCR were performed as per the user guide. For the ATAC-seq library, eight cycles of PCR were run, while seven cycles of PCR were performed for cDNA amplification. Of the amplified cDNA, 25% of the material was used for gene expression library construction with 15 cycles of PCR for both technical replicates.

For samples MO Sa1, MO Sa2, MO C1 and MO C2, ATAC libraries were sequenced on an Illumina NovaSeq 6000 using the following read protocol: 50 cycles (read 1), 8 cycles (i7 index read), 24 cycles (i5 index read) and 49 cycles (read 2). ATAC libraries from MO V1 and MO V2 were sequenced according to the same parameters but on a NextSeq 2000. For samples MO Sa1, MO Sa2, MO C1 and MO C2, RNA libraries were sequenced on an Illumina NovaSeq 6000 using the following read protocol: 28 cycles (read 1), 10 cycles (i7 index read), 10 cycles (i5 index read) and 90 cycles (read 2). RNA libraries from MO V1 and MO V2 were sequenced according to the same parameters but on a NextSeq 2000.

10x mtscATAC (short identifiers mt M1, mt M2, mt C1, mt C2, mt* Br1 and mt* Br2). Cryopreserved PBMCs were thawed as described above. For samples mt* Br1 and mt* Br2, cells were also washed, and 250,000 live cells were sorted using SytoxBlue at a 1:1,000 dilution as a live/dead cell stain. Samples mt M1, mt M2, mt C1 and mt C2 were not sorted. Cells from each donor were subsequently pooled at a 1:1 ratio, and, after washing, cells were fixed in 1% formaldehyde (Thermo Fisher, 28906) in PBS for 10 min at room temperature, quenched with glycine solution to a final concentration of 0.125 M and washed twice in PBS via centrifugation at 400g for 5 min at 4 °C. Cells were subsequently treated with lysis buffer (10 mM Tris-HCl (pH 7.4), 10 mM NaCl, 3 mM MgCl₂, 0.1% NP-40 and 1% BSA) for 3 min on ice, followed by the addition of 1 ml of chilled wash buffer and inversion (10 mM Tris-HCl (pH 7.4), 10 mM NaCl, 3 mM MgCl₂ and 1% BSA) before centrifugation at 500g for 5 min at 4 °C. The supernatant was discarded, and cells were diluted in 1× diluted Nuclei Buffer before counting using trypan blue and a Countess II FL automated cell counter. Subsequently, mtscATAC-seq libraries were generated using the Chromium Next GEM Single Cell ATAC Library & Gel Bead kit (v1.1, 1000175) according to the manufacturer's instructions (CG000209); 4,590 nuclei were loaded with the goal of recovering 3,000 nuclei. Briefly, following tagmentation, cells were loaded onto a Chromium Controller Single Cell instrument to generate single-cell GEMs, followed by linear PCR, as described in the protocol using a C1000 Touch thermal cycler with the 96-Deep Well Reaction Module (Bio-Rad). After breaking the GEMs, barcoded tagmented DNA was purified and further amplified to enable sample indexing (11 cycles of PCR) and enrichment of mtscATAC-seq libraries. The final libraries were quantified using a Qubit double-stranded DNA high-sensitivity assay kit (Invitrogen) and a high-sensitivity DNA chip run on a Bioanalyzer 2100 system (Agilent).

Samples mt C1 and mt C2 were sequenced on an Illumina NovaSeq 6000 under the following sequencing conditions: 50 bp (read 1), 8 bp (i7 index), 16 bp (i5 index) and 49 bp (read 2). Samples mt M1 and mt M2 were sequenced on an Illumina NovaSeq 6000 under the following sequencing conditions: 150 bp (read 1), 8 bp (i7 index), 16 bp (i5 index) and 150 bp (read 2). Samples mt* Br1 and mt* Br2 were sequenced on an Illumina Nextseq 550 with paired-end reads (2 × 72 cycles), 8 cycles for index 1 and 16 cycles for index 2.

Bio-Rad SureCell ATAC (short identifiers ddS Bi1–ddS Bi4, ddS H1, ddS H2, ddS U1 and ddS U2). Cryopreserved PBMCs were quickly thawed in a water bath at 37 °C, rinsed with culture medium (RPMI supplemented with 15% FBS) and treated with 0.2 U μl⁻¹ DNase I (Thermo Fisher Scientific) in 5 ml of culture medium at 37 °C for 30 min. After DNase I treatment, cells were washed once with medium and twice with ice-cold 1× PBS supplemented with 0.1% BSA. Cells were then filtered with a 35-μm cell strainer (Corning), and cell viability and concentration were measured with trypan blue on a TC20 automated cell counter (Bio-Rad).

For a detailed description of tagmentation protocols and buffer formulations, refer to the SureCell ATAC-Seq Library Prep kit user guide (17004620, Bio-Rad). Collected cells and tagmentation buffers were chilled on ice. Lysis was performed simultaneously with tagmentation. After washing, equal numbers of cells from each donor were mixed with Whole-Cell Tagmentation Mix containing 0.1% Tween 20, 0.01% digitonin and 1× PBS supplemented with 0.1% BSA, ATAC Tagmentation Buffer and ATAC Tagmentation Enzyme (ATAC Tagmentation Buffer and ATAC Tagmentation Enzyme are both included in the SureCell ATAC-Seq Library Prep kit (17004620, Bio-Rad)). The mix was split into two technical replicates, and cells were then mixed and agitated on a ThermoMixer (5382000023, Eppendorf) for 30 min at 37 °C. Tagmented cells were kept on ice before encapsulation.

Tagmented cells were loaded onto a ddSEQ Single-Cell Isolator (12004336, Bio-Rad). For samples ddS H1, ddS H2, ddS U1 and ddS U2, 5,000 nuclei were loaded with the goal of recovering 3,000 nuclei.

scATAC-seq libraries were prepared using a SureCell ATAC-Seq Library Prep kit (17004620, Bio-Rad) and SureCell ddSEQ Index kit (12009360, Bio-Rad). Bead barcoding and sample indexing were performed in a C1000 Touch thermal cycler with a 96-Deep Well Reaction Module (1851197, Bio-Rad). The following PCR conditions were used: 37 °C for 30 min; 85 °C for 10 min; 72 °C for 5 min; 98 °C for 30 s; eight cycles of 98 °C for 10 s, 55 °C for 30 s and 72 °C for 60 s and a single 72 °C extension for 5 min to finish. Emulsions were broken, and products were cleaned up using Ampure XP beads (A63880, Beckman Coulter). Barcoded amplicons were further amplified using a C1000 Touch thermal cycler with a 96-Deep Well Reaction Module. The following PCR conditions were used: 98 °C for 30 s and seven cycles of 98 °C for 10 s, 55 °C for 30 s and 72 °C for 60 s and a single 72 °C extension for 5 min to finish. PCR products were purified using Ampure XP beads and quantified on an Agilent Bioanalyzer (G2939BA, Agilent) using a high-sensitivity DNA kit (5067-4626, Agilent).

For samples ddS H1, ddS H2, ddS Bi3, ddS Bi4, ddS U1 and ddS U2, libraries were sequenced on a NextSeq 550 (SY-415-1002, Illumina) using a NextSeq High-Output kit (150 cycles; 20024907, Illumina) and the following read protocol: 118 cycles (read 1), 8 cycles (i7 index) and 40 cycles (read 2). For ddS Bi1 and ddS Bi2, samples were sequenced on a NovaSeq according to the same protocol. A custom sequencing primer was required for read 1 (16005986, Bio-Rad; included in the kit).

HyDrop ATAC (short identifiers Hy E1–Hy E4, Hy V1, Hy V2 and Hy C1–Hy C3). PBMCs were thawed, and nuclei were isolated as described above. HyDrop was performed as previously described¹² but with an updated barcoded hydrogel bead design and minor improvements in nuclei handling. Briefly, barcoded hydrogel beads were produced as described previously but using 384 × 384 combinations of primers instead of the original method using 96 × 96 × 96 combinations, resulting in a barcode sequence of 30 bp instead of 50 bp. One million PBMCs were counted, pelleted and resuspended in 200 µl of ATAC Lysis Buffer (1% BSA, 10 mM Tris-HCl (pH 7.5), 10 mM NaCl, 0.1% Tween 20, 0.1% NP-40, 3 mM MgCl₂, 70 µM Pitstop in DMSO and 0.01% digitonin) for 5 min on ice. One milliliter of ATAC Nuclei Wash Buffer (1% BSA, 10 mM Tris-HCl (pH 7.5), 0.1% Tween 20, 10 mM NaCl and 3 mM MgCl₂) was added, and nuclei were pelleted at 500g at 4 °C for 5 min. The resulting pellet was resuspended in 100 µl of ice-cold PBS and filtered with a 40-µm strainer (Flowmi); 25,000 PBMC nuclei were resuspended in 25 µl of ATAC Reaction Mix (10% dimethylformamide, 10% Tris-HCl (pH 7.4), 5 mM MgCl₂, 5 ng µl⁻¹ Tn5, 70 µM Pitstop in DMSO, 0.1% Tween 20 and 0.01% digitonin) and incubated at 37 °C for 1 h without shaking. To recover a target of 3,000 nuclei, 5,625 tagmented nuclei were added to 48 µl of PCR mix (1.3× Phusion HF buffer, 15% OptiPrep, 1.3 mM dNTPs, 39 mM DTT, 0.065 U µl⁻¹ Phusion HF polymerase, 0.065 U µl⁻¹ Deep Vent polymerase and 0.013 U µl⁻¹ ET SSB). PCR mix was coencapsulated with 35 µl of freshly thawed HyDrop ATAC beads in hydrofluoroether 7500 Novac oil with EA-008 surfactant (RAN Biotech) on an Onyx microfluidics platform (Droplet Genomics). The resulting emulsion was collected in aliquots of 25 µl in total volume and thermocycled according to the linear amplification program (72 °C for 15 min; 98 °C for 3 min; 12 amplification cycles of 98 °C for 10 s, 63 °C for 30 s and 72 °C for 1 min and a final hold at 4 °C). One hundred and twenty-five microliters of Recovery Agent (20% perfluorooctanol in hydrofluoroether 7500), 55 µl of guanidinium thiocyanate buffer (5 M guanidinium thiocyanate, 25 mM EDTA and 50 mM Tris-HCl (pH 7.4)) and 5 µl of 1 M DTT were added to each separate aliquot of 50 µl of thermocycled emulsion and incubated on ice for 5 min. Five microliters of Dynabeads was added to the aqueous phase and incubated for 10 min. Dynabeads were pelleted on a neodymium magnet and washed twice with 80% ethanol. Elution was performed in 50 µl of elution buffer (10 mM Tris-HCl, pH 8.5) supplemented with 10 mM DTT and 0.1% Tween 20. A 1× Ampure bead purification was performed according to manufacturer's recommendations. Elution was performed in 30 µl of elution buffer supplemented

with 10 mM DTT. Eluted library was further amplified in 100 µl of PCR mix (1× KAPA HiFi, 1 µM index i7 primer and 1 µM index i5 primer). The final library was purified in a 0.4–1.2× double-sided Ampure purification, eluted in 25 µl of elution buffer supplemented with 10 mM DTT and quality controlled on an Agilent Bioanalyzer high-sensitivity chip (Agilent Technologies).

Samples Hy V1, Hy V2 and Hy E1–Hy E4 were loaded at 750 pM on a NextSeq 2000 using a NextSeq 2000 P2 kit (100 cycles; 20046811, Illumina), and sequencing was performed using the following read protocol: 49 cycles (read 1), 10 cycles (i7 index read), 31 cycles (i5 index read) and 48 cycles (read 2). Samples Hy C1–Hy C3 were sequenced on a NovaSeq 6000 using the same parameters.

s3-ATAC (short identifiers s3 O1 and s3 O2). Samples s3 O1 and s3 O2 were generated on different days according to the following protocol. Only sample s3 O1 was performed on the reference PBMC sample of two donors. The PBMC pellet was thawed and suspended in NIB-HEPES (pH 7.2; 10 mM HEPES-KOH (BP310-500 (Fisher Scientific) and 1050121000 (Sigma-Aldrich), respectively), 10 mM NaCl, 3 mM MgCl₂ (Fisher Scientific, AC223210010), 0.1% (vol/vol) IGEPAL CA-630 (Sigma-Aldrich, I3021) and 0.1% (vol/vol) Tween (Sigma-Aldrich, P-7949)) before Dounce homogenization. s3-ATAC was then performed as described previously¹³. Two plates were prepared for a total of 2,880 nuclei per sample. Briefly, nuclei were flow sorted via a Sony SH800 to remove debris and attain an accurate count per well before PCR in 1× TD buffer. Immediately following sorting completion, the plate was sealed and centrifuged for 5 min at 500g and 4 °C to ensure that nuclei were within the buffer. Nucleosomes and remaining transposases were then denatured with the addition of 1 µl of 0.1% SDS (roughly 0.01% final concentration) per well. Then, 4 µl of NPM (Nextera XT kit, Illumina) per well was subsequently added to perform gap-fill on tagmented genomic DNA, with an incubation at 72 °C for 10 min. Next, 1.5 µl of 1 µM A14-LNA-ME oligonucleotides was added to supply the template for adapter switching. The polymerase-based adapter switching was then performed under the following conditions: initial denaturation at 98 °C for 30 s and ten cycles of 98 °C for 10 s, 59 °C for 20 s and 72 °C for 10 s. The plate was then held at 10 °C. After adapter switching, 1% (vol/vol) Triton X-100 in ultrapure water (Sigma, 93426) was added to quench persisting SDS. The following was then combined per well for PCR: 16.5 µl of sample, 2.5 µl of indexed i7 primer at 10 µM, 2.5 µl of indexed i5 primer at 10 µM, 3 µl of ultrapure water, 25 µl of NEBNext Q5U 2× master mix (New England Biolabs, M0597S) and 0.5 µl of 100× SYBR Green I (Thermo Scientific, S7563) for a total of 50 µl of reaction per well. Real-time PCR was performed on a Bio-Rad CFX under the following conditions measuring SYBR fluorescence every cycle: 98 °C for 30 s and 16–18 cycles of 98 °C for 10 s, 55 °C for 20 s and 72 °C for 30 s, fluorescent reading and 72 °C for 10 s. After fluorescence passed an exponential growth and began to inflect, the samples were held at 72 °C for another 30 s and stored at 4 °C. Amplified libraries were then cleaned by pooling 25 µl per well into a 15-ml conical tube and cleaning via a QIAquick PCR purification column following the manufacturer's protocol (Qiagen, 28106). The pooled sample was eluted in 50 µl of 10 mM Tris-HCl (pH 8.0). Library molecules then went through size selection via SPRI selection beads (Mag-Bind TotalPure NGS Omega Biotek, M1378-01). Next, 50 µl of vortexed and fully suspended room temperature SPRI beads was combined with the 50-µl library (one cleanup) and incubated at room temperature for 5 min. The reaction was then placed on a magnetic rack, and, once cleared, the supernatant was removed. The remaining pellet was rinsed twice with 100 µl of fresh 80% ethanol. After the ethanol was pipetted out, the tube was spun down and placed back on the magnetic rack to remove any lingering ethanol. Next, 31 µl of 10 mM Tris-HCl (pH 8.0) was used to resuspend the beads off the magnetic rack, followed by an incubation for 5 min at room temperature. The tube was again placed on the magnetic rack, and, once cleared, the full volume of supernatant was moved

to a clean tube. DNA was then quantified by Qubit double-stranded DNA high-sensitivity assay following the manufacturer's instructions (Thermo Fisher, Q32851). Libraries were diluted to 2 ng μL^{-1} and run on an Agilent TapeStation 4150 D5000 tape (Agilent, 5067-5592). Library molecule concentration within the range of 100 to 1,000 bp was then used for final library dilution of 1 nM.

Samples s3 O1 and s3 O2 were sequenced on a NovaSeq S2 flow cell following the manufacturer's recommendations (Illumina, 20028315) as paired-end libraries with 10 cycle index reads and 85 cycles (O1) or 90 cycles (O2) for reads 1 and 2.

10x v1.1 control runs (short identifiers v1.1c C1 and v1.1c C2). Two additional control runs were performed on the same day as v1.1 C3. Control run v1.1c C1 was performed using the standard 10x nuclei extraction lysis buffer with the omission of NP-40 to simulate the whole-cell protocol used in Bio-Rad ddSEQ experiments. Control run v1.1c C2 was performed using the Dounce homogenization protocol as described in the s3-ATAC experiments but without FACS. Starting from permeabilized cells or Dounce-extracted nuclei, both control runs were performed exactly according to the standard 10x v1.1 protocol simultaneously with v1.1 C3.

Samples v1.1c C1 and v1.1c C2 were sequenced on a NovaSeq 6000 with 50 cycles for read 1, 49 cycles for read 2, 8 cycles for index 1 and 16 cycles for index 2.

10x scRNA-seq. Cryopreserved PBMCs were thawed as described above, and equal numbers of cells from each donor were mixed. The cell mix was partitioned into GEMs by using the Chromium Controller system (10x Genomics), with a target recovery of 5,000 total cells. We generated three technical replicates by loading three channels of Chip G with the same cell mix. cDNA sequencing libraries were prepared using the Next GEM Single Cell 3' reagent kit v3.1 (10x Genomics, PN-1000268), following the manufacturer's instructions. Briefly, after GEM-RT cleanup, cDNA was amplified during 12 cycles, and cDNA quality control and quantification were performed on an Agilent Bioanalyzer high-sensitivity chip (Agilent Technologies). cDNA libraries were indexed by PCR using the PN-220103 Chromium i7 Sample Index Plate. Size distribution and concentration of 3' cDNA libraries were verified on an Agilent Bioanalyzer high-sensitivity chip (Agilent Technologies).

Sequencing of cDNA libraries was performed on an Illumina NovaSeq 6000 using the following sequencing conditions to obtain approximately 40,000 reads per cell: 28 bp (read 1), 8 bp (i7 index), 0 bp (i5 index) and 89 bp (read 2).

Data preprocessing

Unified scATAC-seq data analysis pipeline (PUMATAC). We developed PUMATAC, a unified Nextflow v21.04.3 (ref. 29) pipeline, to align samples from multiple technologies to the reference genome and write fragments files from these reference genome alignments (<https://github.com/aertslab/PUMATAC>). The steps implemented in PUMATAC are described briefly in the text below and in detail with examples at https://github.com/aertslab/scATAC-seq_benchmark. All code necessary to reproduce our analyses and graphics is present in notebooks in this repository.

Barcode correction and FASTQ processing (singlecelltoolkit in PUMATAC). Each barcode was compared to the whitelist barcodes and kept (with bam tag 'CB') if it was a perfect match or if changing any of the bases by 1 bp resulted in a match (Hamming distance of 1). Barcodes that were unable to be corrected were retained with the 'CR' bam tag. The barcode tag information, including the original barcode quality scores ('CY'), was added to the comments field in each of the two paired-end FASTQ files. Adapter trimming was then performed using TrimGalore (version 0.6.6)³⁰ with the '-paired' option, which in turn runs Cutadapt³¹.

Reference genome alignment and fragments writing (bwa-mem in PUMATAC). We first aligned full sequencing datasets of all samples to the GRCh38 or mm10 reference genome using PUMATAC. We then filtered cells (described later) and downsampled all sequencing data to a common sequencing depth of 40,796 reads per cell and realigned these downsampled FASTQ files. In PUMATAC, alignment was performed using bwa-mem2 (v2.2.1)³² with the 'mem' method and default mapping parameters. The '-C' option was used to copy the barcode tag information from the FASTQ file to the resulting bam file. Read group information was taken from the FASTQ name field in the first line of each input file and added with the '-R' option in bwa-mem2. The 'fixmate' tool from SAMtools (version 1.12)³³ was used to add mate coordinates and insert sizes to the file. Reads were aligned to the GRCh38 reference for the PBMC samples and to mm10 for mouse public data. From the resulting aligned reads in .bam format, fragments were written in the bed-like fragments.tsv.gz format using a combination of SAMtools³⁴ and AWK, according to the base pair shift rules described in the Cell Ranger manual (<https://support.10xgenomics.com/single-cell-atac/software/pipelines/latest/output/fragments>).

Barcode multiplet detection (barcard in PUMATAC). For each sample, we detected barcode multiplets using barcard, our own reimplementation of bap (<https://github.com/caleblareau/bap>). Similar to bap, barcard subsets fragments files to barcodes associated with at least 1,000 unique fragments. For each remaining barcode, the number of unique fragments that share their beginning and end coordinates between the two barcodes divided by the total number of barcodes found in both barcodes combined is calculated with every other barcode. The Jaccard indices for these barcode pairs are then ranked and thresholded using Otsu's algorithm to identify barcode multiplets. Following the identification of barcode multiplets in each sample, a new tag ('DB') was added to the bam file to represent droplet barcodes. This tag contained either the original corrected barcode from the CB tag (in the case of singlets) or an underscore-separated concatenation of each corrected barcode that forms the multiplet. This step, and others, was parallelized using GNU Parallel³⁵. Similarly, fragments.tsv.gz files were rewritten to merge detected barcode multiplets. While this step is only necessary for Bio-Rad ddSEQ samples, we detected and merged multiplets in all samples.

PUMATAC validation using CellRanger. We realigned all 10x v1, v1.1, v2 and multiome data using CellRanger-arc. We then subset fragments files generated by CellRanger and PUMATAC on barcodes identified as cell barcodes. For each pair of barcodes, we then calculated the number of unique fragments that were attributed to that barcode by both CellRanger and PUMATAC based on beginning and ending coordinates. The Jaccard index was calculated based on this number.

Downstream analyses. Starting from the fragments files generated by PUMATAC and merged using barcard, we then performed further analyses such as clustering, cell-type annotation, differential region calling and transcription factor motif analysis using a combination of bioinformatics packages.

Single-cell-level quality control and barcode filtering (pyscTopic). We used the Python implementation of cisTopic¹⁶ (pyscTopic; <https://github.com/aertslab/pyscTopic>) to collect single-cell-level quality control statistics and filter barcodes starting from the PUMATAC fragments files. For quality control purposes, we considered all barcodes with at least ten unique fragments. We used the GRCh38 or mm10 BioMart³⁶ gene annotation to calculate TSS enrichment. We followed the current ENCODE recommendations in calculating TSS enrichment by examining read depth in a 2,000-bp window on each side of the TSS (<https://www.encodeproject.org/data-standards/terms/#enrichment>). Cells were then filtered from barcodes using Otsu algorithm-defined

thresholds on TSS enrichment and number of unique fragments per cell. In the first pass, we counted fragments in SCREEN regions¹⁸. This count matrix was then used to annotate cells with known cell types, after which consensus peaks could be called over all cell types present (see later). We then used this consensus peak set to recount the fragments and perform further downstream analyses in the second pass.

Donor identification (Freemuxlet in VSN/PUMATAC). We used Freemuxlet to identify barcodes belonging to each of the two mixed individuals in each sample and simultaneously identify doublets on the basis of barcodes with mixed genotypes. We first ran a prefiltering step to filter the bam file for only the selected cells after the initial quality control. We then used the popsicle suite of tools (<https://github.com/statgen/popsicle>), first 'dsc-pileup' to quantify reads overlapping known variants and then Freemuxlet to call sample and doublet identity in each barcode. Freemuxlet requires a list of known variants in the genome along with their allele frequencies. To obtain this, we used the 1000 Genomes Phase 3 dataset³⁷ and applied filtering steps to keep only SNPs with a minor allele frequency of at least 10%. This step was automated using VSN/PUMATAC¹⁴.

Doublet identification (Scrublet). We used the Python implementation of Scrublet¹⁹ to identify doublets among the barcodes selected based on the fragments/regions count matrix during the initial quality control steps. Doublet thresholds for each sample were set manually, and doublets were removed for downstream analysis.

Sequencing depth downsampling (seqtk). After the initial quality control steps to select cells, we identified the sample with the fewest number of reads per cell and used this sample as the reference to which the others were downsampled. After identifying the target number of reads per cell in the reference sample (ddS Bi3, with 40,796 reads per cell), each of the FASTQ files for the other samples were downsampled to that depth. Downsampling was performed with seqtk (<https://github.com/lh3/seqtk>, version 1.3-r106). The seed parameter ('-s') was set to the same value for all files to ensure that the reads remained paired across the paired-end and barcode files. Following downsampling of each FASTQ file, the mapping procedure was repeated to produce new downsampled fragments and bam files. This was repeated for 35,000, 30,000, 25,000, 20,000, 15,000, 10,000 and 5,000 reads per cell, and these FASTQ files were further processed as described earlier and later.

Cell-type identification (Seurat). Label transfer was performed using an annotated PBMC reference dataset²¹ consisting of nine independent technology types and batches. We used Seurat (v4.0.3) to perform the label transfer steps in an R (v4.1.0) environment. Label transfer was performed using methods outlined in the Seurat vignettes and associated²². In brief, each of the nine already annotated PBMC reference datasets was compared pairwise to find cells that serve as anchors between them and then used to generate an integrated reference that minimized technical differences.

For the scATAC-seq data, we used this integrated PBMC reference to predict cell types. A gene activity matrix was first estimated, and label transfer was performed by assigning query cells based on the local neighborhood around each anchor in the integrated reference, with the highest scoring cell type being assigned. Following prediction of the scATAC-seq cell types, we refined these classifications by using the clusters identified in the scATAC-seq data. Clustering was first performed on cells with the Leiden algorithm using a high resolution to generate many fine-grained clusters. For each cluster, we then assigned a consensus cell-type identity to the entire cluster based on the majority cell type identified by label transfer. In this way, the ATAC-based clusters were labeled with the most likely cell type, while peak information was retained for later analysis. Where multiple clusters existed for one cell

type, these were merged and used to generate cell-type-specific peak sets in downstream steps.

Two-pass dimensionality reduction (pycisTopic). Fragments were first counted in ENCODE SCREEN regions to generate a preliminary count matrix. This count matrix was used to filter cells based on TSS enrichment and number of unique fragments. The SCREEN regions count matrix was then used to train cisTopic's latent Dirichlet allocation models, and the model with an optimal number of topics was selected as described in ref. 16. Based on Seurat cell-type identification and high-resolution Leiden clustering, a consensus cell type was assigned to each cell. Cell-type-specific peaks (see later) were called based on these consensus cell types and aggregated into a new consensus peak set. The fragments were recounted in this new peak set to generate a consensus peak count matrix. This count matrix was used to retrain cisTopic's latent Dirichlet allocation models, and an optimal model was chosen for the second pass and used to reduce the dimensionality of the data.

Consensus peak calling (pycisTopic). In pyCisTopic, the 'export_pseudobulk' function was used to create cell-type-specific fragments and bigwig files using the consensus cell types. These were in turn used to generate cell-type-specific consensus peaks for each sample by recalling the subset of cells with MACS2 (ref. 23). Peak calling for quality control purposes was performed using MACS2 with settings specific to ATAC-seq experiments ('genome_size = hs', 'shift = 73', 'ext_size = 146' and 'q_value = 0.01'). We used 'Version 2' of the ENCODE candidate cis-Regulatory Elements with blacklisted regions removed³⁸ (Supplementary Fig. 1b). These regions were used after the quality control steps to create the cisTopic object, perform the first-pass clustering and obtain consensus peaks. Duplicate rates were calculated for each barcode by dividing the number of unique fragments by the total.

Cell-type-specific peak sets were generated for each sample. For each sample, the cell-type-specific peak sets were then merged into a final sample-specific consensus set. Each sample's FRIP was calculated using these final consensus peaks. The consensus sets of all 47 samples were then merged into one master set, in which all data were counted to form the merged datasets.

Region overlap calculation (HOMER). The mergePeaks function of the HOMER suite (v4.11)³⁹ with default parameters (adding -d given -venn options) was used to find overlap between several region sets, notably the overlap between consensus DARs and peaks recovered in the merged cell-type fair set and in the individual cell-type fair sets.

DAR calling (pycisTopic). A Wilcoxon rank-sum test was used to calculate significance of enrichment of regions (fold change) between each specified contrast using cisTopic's imputed region accessibility based on cell topic and topic region probabilities. We contrasted cell types (type 1 versus all), technologies (cells from cell type 1 from technique A versus cells from cell type 1 from technique B) and donors (cells from cell type 1 from donor A versus cells from cell type 1 from donor B). Because high-quality samples produce many DARs that can skew the distribution of DAR enrichment scores to lower numbers, we chose to show the distributions of the top 2,000 DARs for each cell type contrast and the top 200 for male–female contrasts. For cell-type contrasts, differential accessibility was thresholded at a minimum of 1.5× fold change enrichment. For male/female contrasts, a minimum threshold of 1.2× fold change enrichment was used.

Transcription factor motif enrichment analysis (cisTarget). Cell-type-specific and male/female-specific DARs were analyzed for transcription factor motif enrichment using cisTarget^{24,25} and standard parameters and settings for the human genome.

Price calculation and sequencing saturation. We calculated the price of a hypothetical 5,000-cell experiment based on US list prices for commercial methods and original manuscripts for open-source methods as follows. The 10x scATAC-seq v2 assay was quoted as \$1,750 for 48 Next Gem Chip H (1000161), \$930 for 96 indices (1000212) and \$24,300 for 16 Chromium Next GEM Single Cell ATAC v2 (1000390) for a weighted total of \$1,565 per lane. For the 10x multiome assay, the price was \$1,750 for 48 Next Gem Chip J (1000234), \$930 for 96 indices (1000215) and \$44,760 for 16 Chromium Next GEM Single Cell Multiome ATAC + Gene Expression (1000283) for a weighted total of \$2,843 per lane. For Bio-Rad ddSEQ SureCell ATAC, the price was \$8,800 for a complete SureCell ATAC-Seq library prep kit (17004620), which accommodates eight samples. For s3-ATAC, the cost per plate is ~\$200, and each plate can accommodate 1,440 cells. For HyDrop, the cost per run is ~\$100 and can recover 8,000–10,000 cells.

Bio-Rad ddSEQ reports a doublet rate of 3.76% at a recovery of 5,000 cells (https://www.bio-rad.com/sites/default/files/webroot/web/pdf/lsr/literature/ATAC-Seq_Poster.pdf). 10x Chromium supports a recovery of up to 10,000 cells but at a doublet rate of 8%. At an expected doublet rate of 4%, 5,000 cells can be recovered (<https://kb.10xgenomics.com/hc/en-us/articles/360001378811-What-is-the-maximum-number-of-cells-that-can-be-profiled->). HyDrop reports 6% doublets on 8,000 recovered cells¹². These three microfluidic methods use the same microfluidic concepts to encapsulate single cells, and doublet rates are similar when correcting for the number of cells recovered. We therefore reasoned that the most fair comparison would be to assume a recovery of 5,000 cells per 10x and ddSEQ lane or HyDrop run. For s3-ATAC, we assumed 1,440 cells per plate, for a total of 5,760 cells across four plates in s3-ATAC.

Full sequencing depth fragments files were subset to filtered cell barcodes (before doublet filtering and minimum TSS enrichment threshold, that is, only filtered by Otsu thresholding on minimum number of reads). We then subsampled these fragments files using a range of fractions and fitted a Michaelis–Menten kinetic model on the resulting duplication rate by the number of reads per cell. We defined the saturation sequencing depth at which each technology is expected to reach 50% duplicate fragments after sequencing.

scRNA-seq analysis (scanpy). After aligning scRNA-seq data to the reference genome using CellRanger or CellRanger-arc, scanpy⁴⁰ was used to calculate single-cell quality control metrics for each sample. True cells were filtered from noise using Otsu-derived cutoffs on minimum number of UMIs.

Public mouse brain data reanalysis. Public mouse scATAC-seq data were downloaded from the following sources: 10x Genomics scATAC-seq v1.0 on Chromium (https://s3-us-west-2.amazonaws.com/10x.files/samples/cell-atac/1.2.0/atac_v1_adult_brain_fresh_5k/atac_v1_adult_brain_fresh_5k_fastqs.tar), 10x Genomics scATAC-seq v1.1 on Chromium (https://s3-us-west-2.amazonaws.com/10x.files/samples/cell-atac/2.1.0/8k_mouse_cortex_ATACv1p1_nextgem_Chromium_X/8k_mouse_cortex_ATACv1p1_nextgem_Chromium_X_fastqs.tar), 10x Genomics scATAC-seq v2 on Chromium X (https://s3-us-west-2.amazonaws.com/10x.files/samples/cell-atac/2.1.0/8k_mouse_cortex_ATACv2_nextgem_Chromium_X/8k_mouse_cortex_ATACv2_nextgem_Chromium_X_fastqs.tar), 10x Genomics scATAC-seq v2 on Chromium (https://s3-us-west-2.amazonaws.com/10x.files/samples/cell-atac/2.1.0/8k_mouse_cortex_ATACv2_nextgem_Chromium_Controller/8k_mouse_cortex_ATACv2_nextgem_Chromium_Controller_fastqs.tar), 10x Genomics Multiome ATAC (https://s3-us-west-2.amazonaws.com/10x.files/samples/cell-arc/2.0.0/e18_mouse_brain_fresh_5k/e18_mouse_brain_fresh_5k_fastqs.tar), Bio-Rad ddSEQ (SRA accession number [SRR1449447](https://www.ncbi.nlm.nih.gov/sra/SRR1449447)), s3-ATAC (SRA accession number [SRX10841853](https://www.ncbi.nlm.nih.gov/sra/SRX10841853)) and HyDrop (SRA accession number [PRJNA733185](https://www.ncbi.nlm.nih.gov/sra/PRJNA733185)).

These data were reanalyzed in a similar manner as the PBMC datasets. Briefly, cells were filtered from noise as described above. All datasets were then downsampled to the highest common depth and further intervals of 5,000 reads per cell. These downsampled sets were realigned to the mm10 reference genome using PUMATAC and counted in the SCREEN regions. Cells were clustered, and per cluster peaks were called and aggregated into a consensus peak set per sample. Datasets were recounted in these consensus peak sets to produce the quality metrics shown in our manuscript.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Single-cell ATAC accessibility and gene expression data can be viewed at https://scope.aertslab.org/#/scATAC-seq_Benchmark/scATAC-seq_Benchmark. Single-cell ATAC coverage bigwigs and DAR/peak BEDs can be downloaded at https://ucscctracks.aertslab.org/papers/scatac_benchmark/ and viewed using University of California Santa Cruz's custom track hub. Sequencing data, fragments files and count matrices are freely available at Gene Expression Omnibus under accession number [GSE194028](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE194028) (ref. 41). Summary quality metrics for all samples can be found in Supplementary Table 1. Source data are provided with this paper.

Code availability

All data analysis scripts can be found at https://github.com/aertslab/scATAC-seq_benchmark ref. 42. PUMATAC can be found at <https://github.com/aertslab/PUMATAC>.

References

- Di Tommaso, P. et al. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
- Krueger, F., James, F., Ewels, P., Afyounian, E. & Schuster-Boeckler, B. FelixKrueger/TrimGalore: v0.6.7. *Zenodo* <https://zenodo.org/record/5127899> (2021).
- Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*. **17**, 10–12 (2011).
- Md, V., Misra, S., Li, H. & Aluru, S. Efficient architecture-aware acceleration of BWA-MEM for multicore systems. Preprint at <http://arxiv.org/abs/1907.12931> (2019).
- Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
- Danecek, P. et al. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
- Tange, O. GNU Parallel 2018. *Zenodo* <https://zenodo.org/record/1146014> (2018).
- Durinck, S. et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).
- Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
- Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: identification of problematic regions of the genome. *Sci. Rep.* **9**, 9354 (2019).
- Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
- Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
- De Rop, F. et al. Datasets supplementary to systematic benchmarking of single-cell ATAC sequencing protocols. *Gene Expression Omnibus* <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE194028> (2023).

42. De Rop, F. aertslab/scATAC-seq_benchmark. Zenodo <https://doi.org/10.5281/zenodo.8034473> (2023).

Acknowledgements

H.H. received support for the project PID2020-115439GB-I00 funded by MCIN/AEI/10.13039/501100011033. This publication is also supported as part of a project (BCLLATLAS and ESPACE) that has received funding from the European Research Council under the European Union's Horizon 2020 Research and Innovation Programme (grant agreement numbers 810287 and 874710). This work was supported by a European Research Council Consolidator grant to S.A. (724226_cis-CONTROL), KU Leuven (C14/22/125 to S.A.), Foundation Against Cancer (F/2020/1396 to S.A.), F.W.O. (grants GOI2722N, GOB5619N and G094121N to S.A. and a PhD fellowship to F.D.) and Aligning Science Across Parkinson's (grant number ASAP-000430 to S.A.). K.B.M. and S.A.T. are supported by Wellcome (WT211276/Z/18/Z and Sanger core grant WT206194). Computing was performed at the Vlaams Supercomputer Center and high-throughput sequencing at the Genomics Core Leuven. M.R.C. is supported by the National Institutes on Aging K99/RO0AG059918. This work was supported by funding from the Rita Allen Foundation (W.J.G.) and the Human Frontiers Science Program (RGY006S; W.J.G.). W.J.G. is a Chan Zuckerberg Biohub investigator and acknowledges grants 2017-174468 and 2018-182817 from the Chan Zuckerberg Initiative and National Institutes of Health grants RM1-HG007735, UM1-HG009442, UM1-HG009436, RO1-HG00990901 and U19-AI057266 (to W.J.G.). W.J.G. acknowledges funding from Emerson Collective. B.D. received financial support by Swiss National Science Foundation 310030_197082 and the EPFL. L.S.L. receives support from an Emmy Noether fellowship by the German Research Foundation (LU 2336/2-1), a National Institutes of Health grant UM1HG012076, a Longevity Impetus grant and a Hector Research Career Development Award by the Hector Fellow Academy. A.C.A. is supported by National Institutes of Health grants RF1-MH128842, R35-GM124704 and RO1-DA047237 as well as a Silver Family Foundation Innovator Award.

Author contributions

H.H., S.A., F.D.R. and C.F. conceptualized the work and designed the methodological approach. G.H., C.F. and F.D.R. developed PUMATAC software. F.D.R., C.F., G.H., P.S.-V. and A.R. performed data analyses. F.D.R., V.C., D.M., G.C., W.L., F.D., R.S., S.T.B., L.M., A.K., K.B.M., R.M., A.T., P.M., J.P., C.L.G.L., M.I., S.R.N., Z.Y.D., O.R.-R., M.S., C.M. and M.R.C. provided samples and/or performed experiments. H.H., S.A. and F.D.R. wrote the manuscript. F.D.R. visualized the data. Z.B., R.L. and L.G.M. provided access to existing data. H.H., S.A., S.P., R.L., J.I.M.-S., S.P.,

A.S.-P., B.D., S.A.T., L.S.L., C.B.G.-B., T.P.B., A.C.A., W.J.G., J.D.B. and A.R. provided supervision and/or funding.

Competing interests

H.H. is a cofounder and equity holder of Omniscope, a Scientific Advisory Board member of MiRXES and a consultant to Moderna. A.R. is a founder and equity holder of Celsius Therapeutics, an equity holder in Immunitas Therapeutics and, until 31 August 2020, was a Scientific Advisory Board member of Syros Pharmaceuticals, Neogene Therapeutics, Asimov and Thermo Fisher Scientific. A.R. is an employee of Genentech and has equity in Roche. O.R.-R. is an employee of Genentech. A.R. and O.R.-R. are co-inventors on patent applications filed by the Broad Institute relating to single-cell genomics. In the past 3 years, S.A.T. has received remuneration for Scientific Advisory Board Membership from Sanofi, GlaxoSmithKline, Foresite Labs and Qiagen. S.A.T. is a cofounder of and holds equity in Transition Bio. 10x Genomics has licensed intellectual property on which W.J.G. is listed as an inventor. W.J.G. holds options in 10x Genomics and is a consultant for Ultima Genomics, Guardant Health. W.J.G. is a scientific cofounder of Protillion Biosciences. J.D.B. holds patents related to ATAC-seq and SHARE-seq and is a Scientific Advisory Board member of Camp4 and seqWell. A.A. is an author on licensed patents that cover the nucleosome disruption and indexed tagmentation design. L.L. holds a patent for lineage tracing using mitochondrial genome mutations and single-cell genomics and is a consultant to Cartography Biosciences. B.D. is cofounder and shareholder of Alitheia Genomics. The remaining authors declare no competing interests.

Additional information

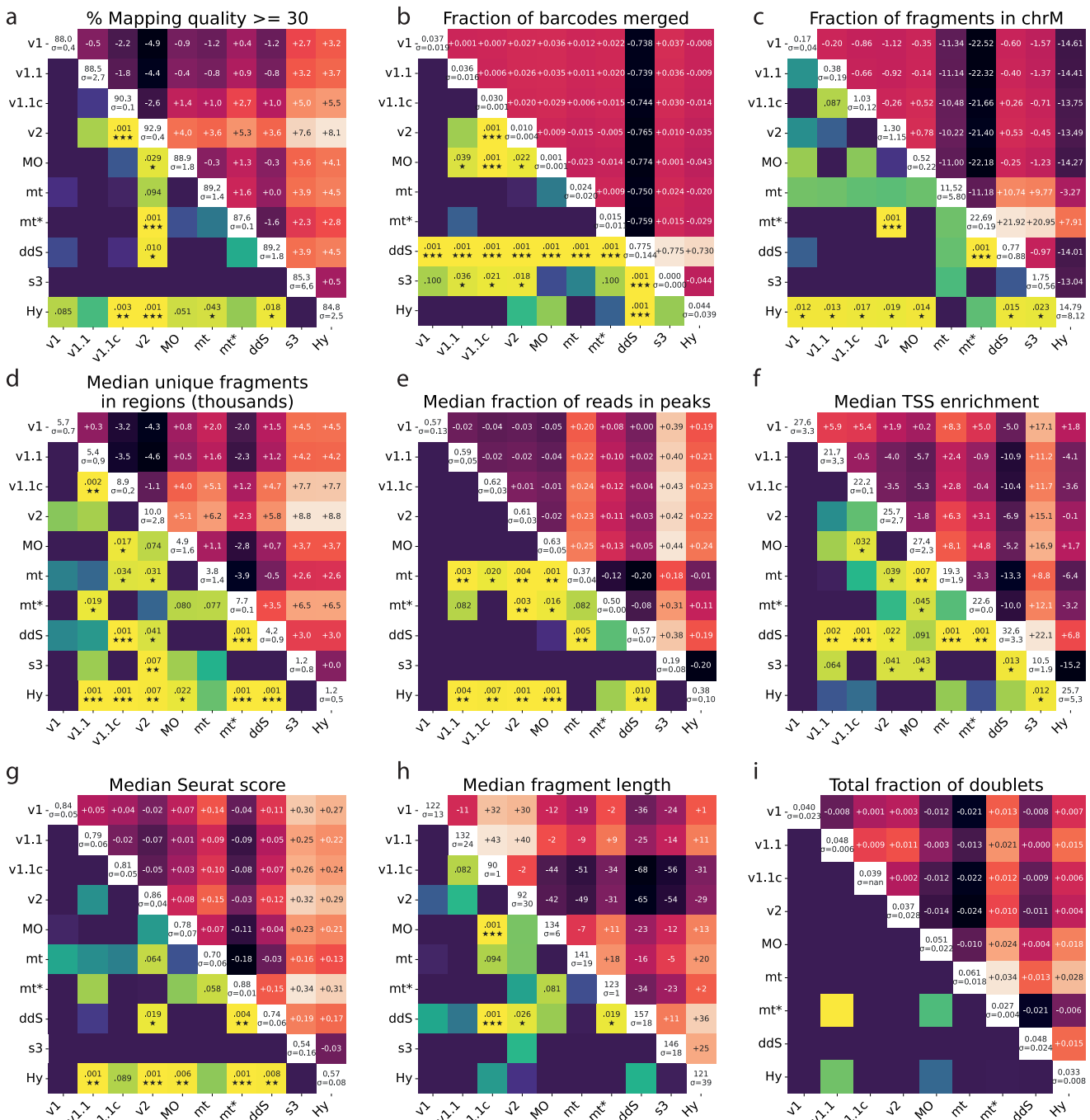
Extended data is available for this paper at <https://doi.org/10.1038/s41587-023-01881-x>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41587-023-01881-x>.

Correspondence and requests for materials should be addressed to Stein Aerts or Holger Heyn.

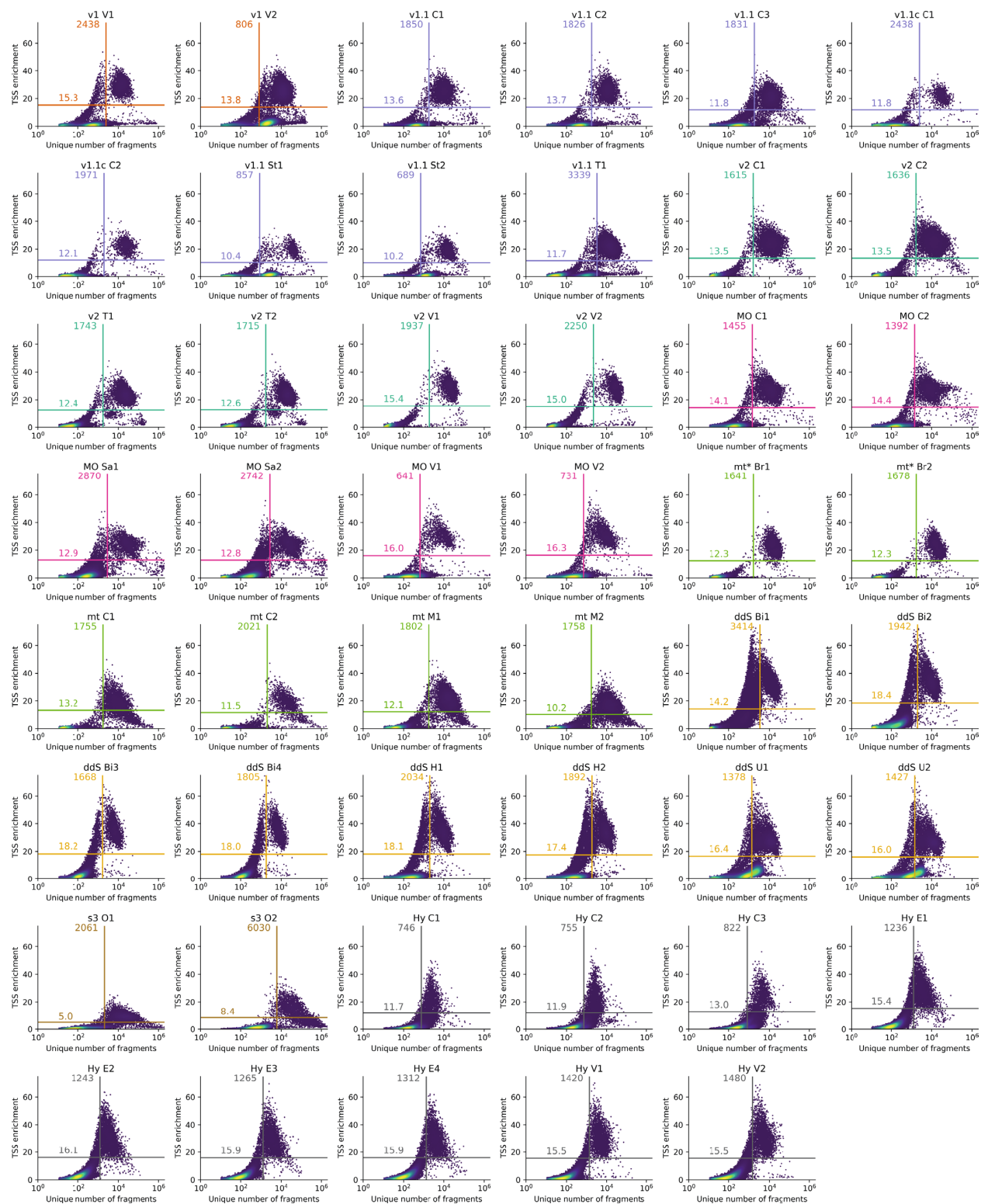
Peer review information *Nature Biotechnology* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

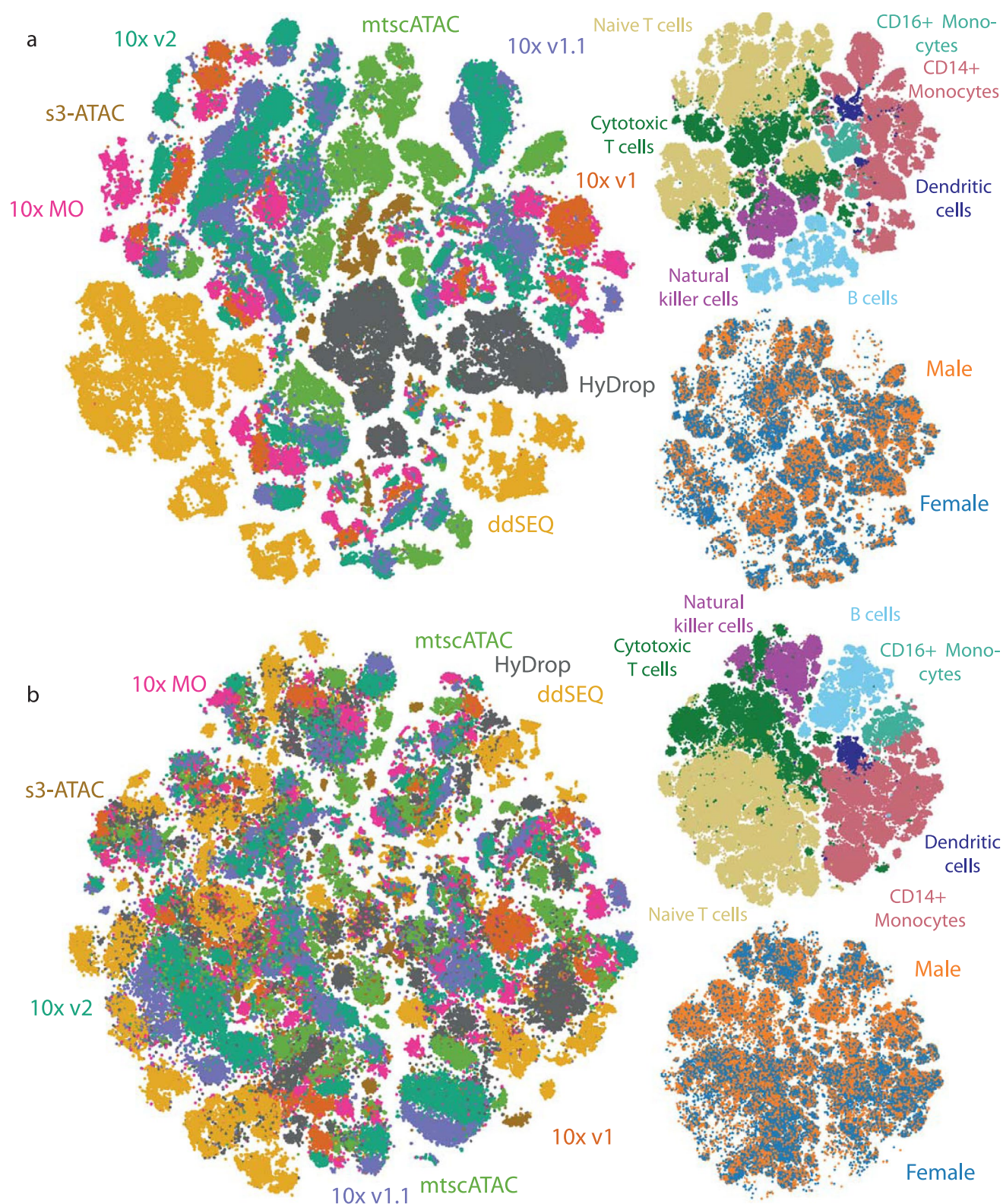


Extended Data Fig. 1 | Games-Howell heatmaps of several quality control metrics. Diagonal shows mean values across samples from each technology. Upper triangle shows pairwise differences between means. Lower triangle indicates significance of expected pairwise differences. mt* indicates FAC sorting for these mtscATAC-seq samples. **a**, % of all reads with Phred mapping quality score ≥ 30 . **b**, Fraction of barcodes being the result of a jaccard merging event (= being a barcode multiplet). **c**, Fraction of fragments aligned to mitochondrial

genome. Figures d-g are cell-level statistics, meaning first a median is taken within each sample's cells, and these medians are compared between techniques. **d**, Median unique fragments in peaks, in cells (x1000). **e**, Median fraction of reads in peaks in cells. **f**, Median TSS enrichments of reads within cells. **g**, Median Seurat scores across cells. **h**, Median fragment length in fragments files. **i**, Total number of doublets (both Freemuxlet + Scrublet-identified) filtered out, expressed as a fraction of total cells before doublet filtering.

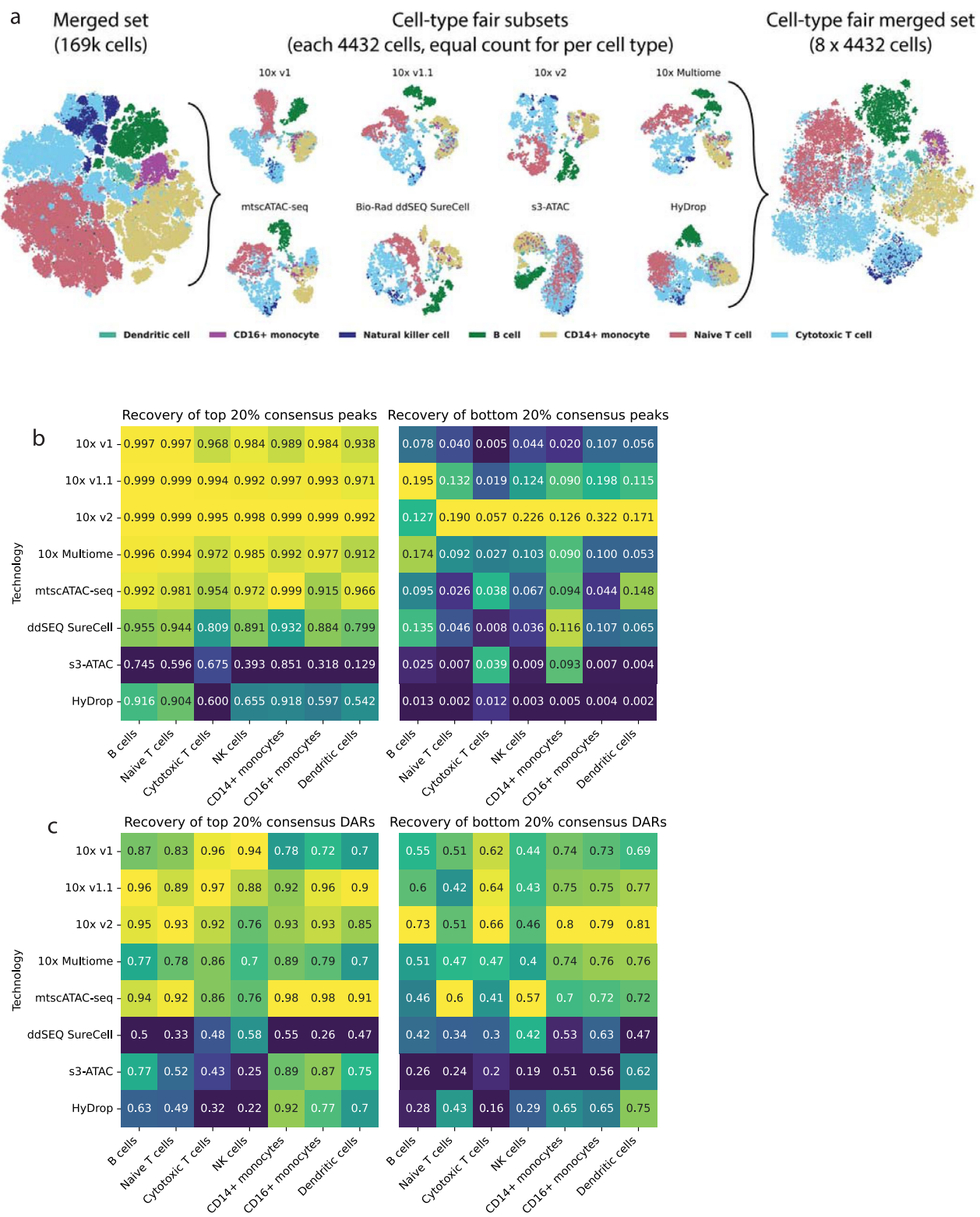


Extended Data Fig. 2 | Individual TSS enrichment and Unique number of fragments distributions for each sample. Thresholds on both axes calculated using Otsu's algorithm. Barcodes meeting both thresholds (upper right quadrant) are identified as cells in further analysis.

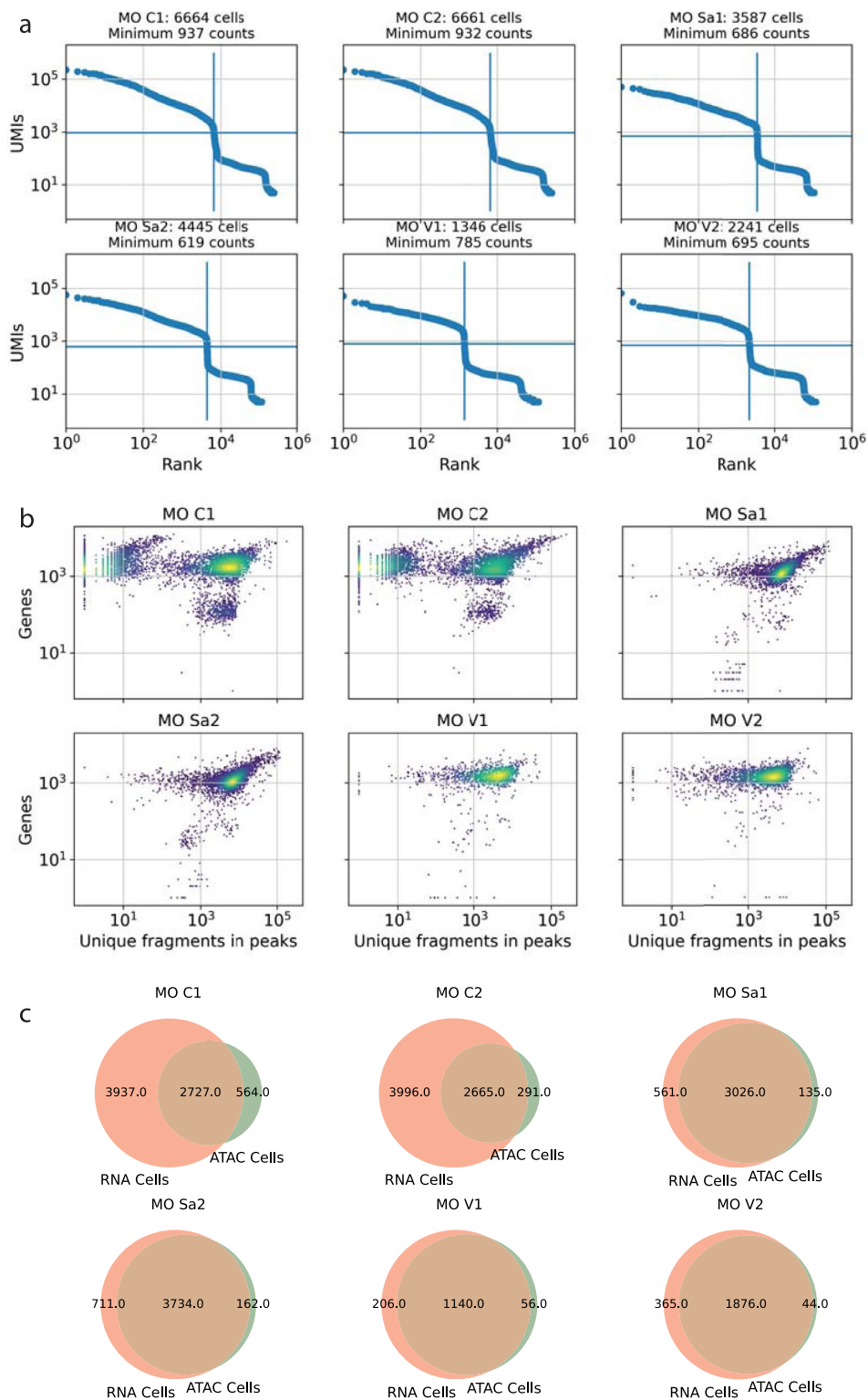


Extended Data Fig. 3 | tSNE of full, non-downsampled data. Datapoints were randomly shuffled before plotting. **a**, Non-harmony-corrected tSNE of non-downsampled data, colored by technology of origin, cell type and donor identity.

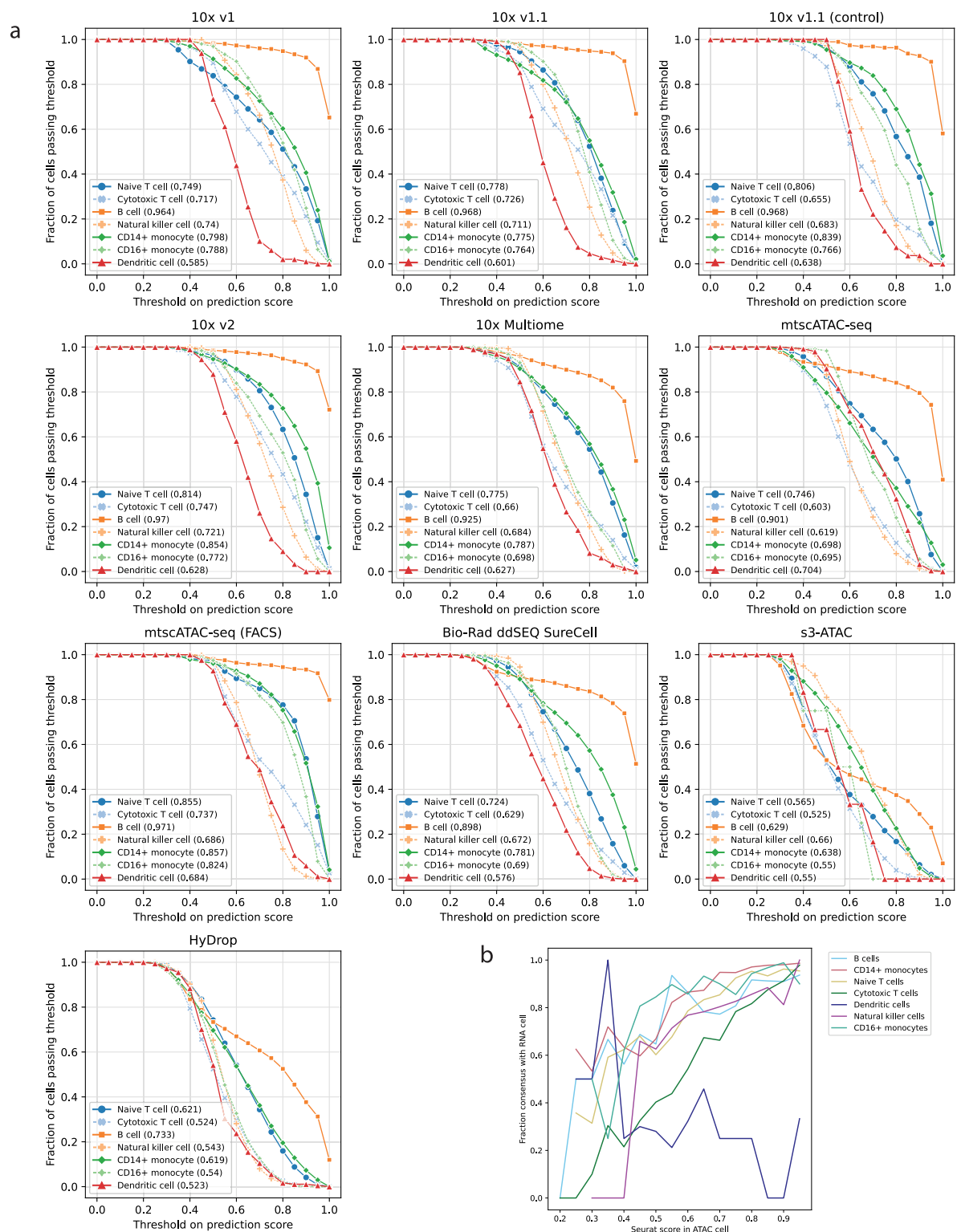
b, Harmony-corrected (by technology, considering 10x v1, 1.1, 2 as a single technology) tSNE of non-downsampled data, colored by technology of origin, cell type and donor identity.



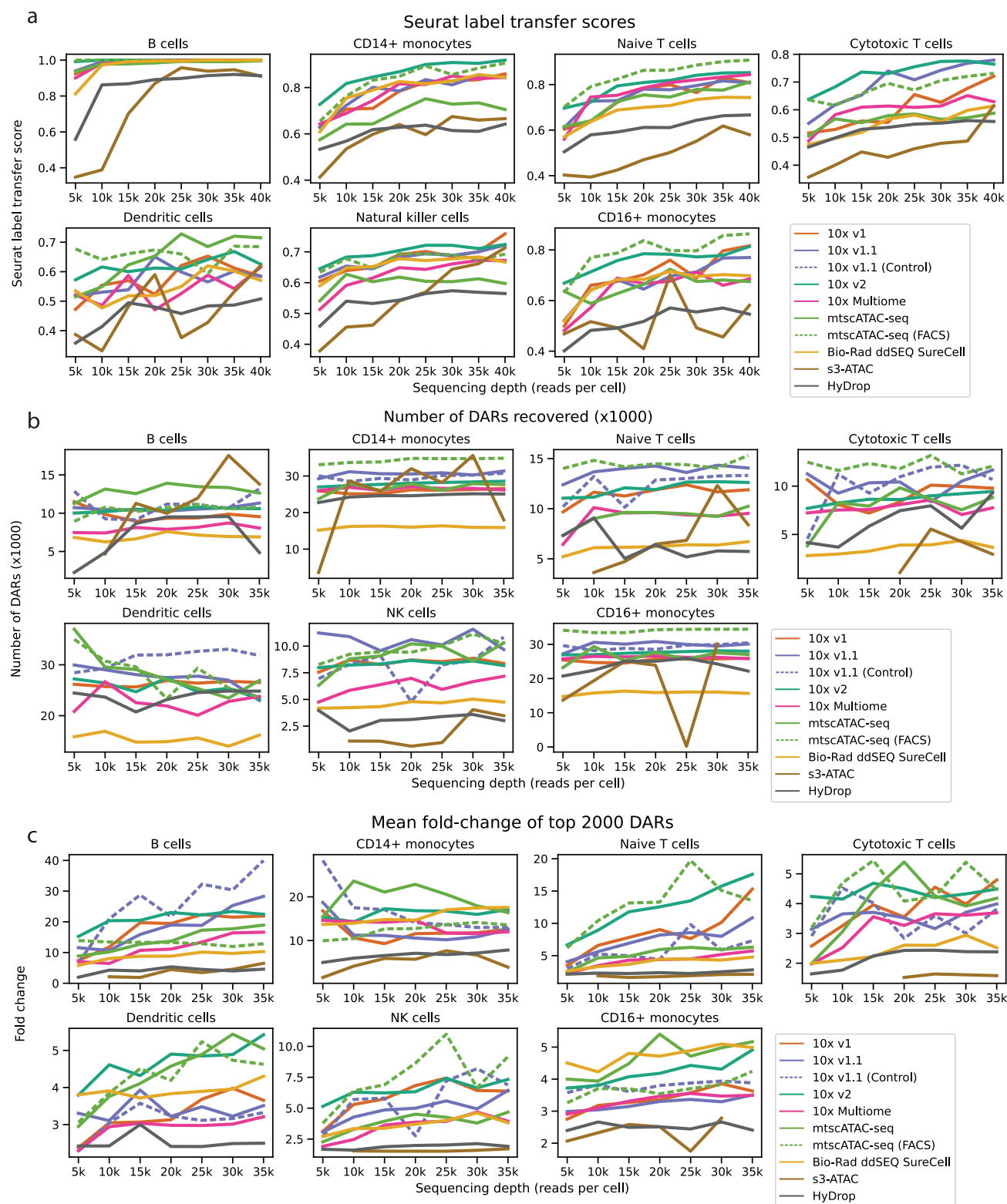
Extended Data Fig. 4 | Cell-type fair subsetting and peak/DAR calculation. **a**, Diagram showing cell-type fair subsetting strategy. **b**, Heatmap showing the fraction of cell-type-fair peaks recovered by the individual technologies' subsets, separated by cell type. **c**, Heatmap showing the fraction of cell-type-fair DARs recovered by the individual technologies' subsets, separated by cell type.

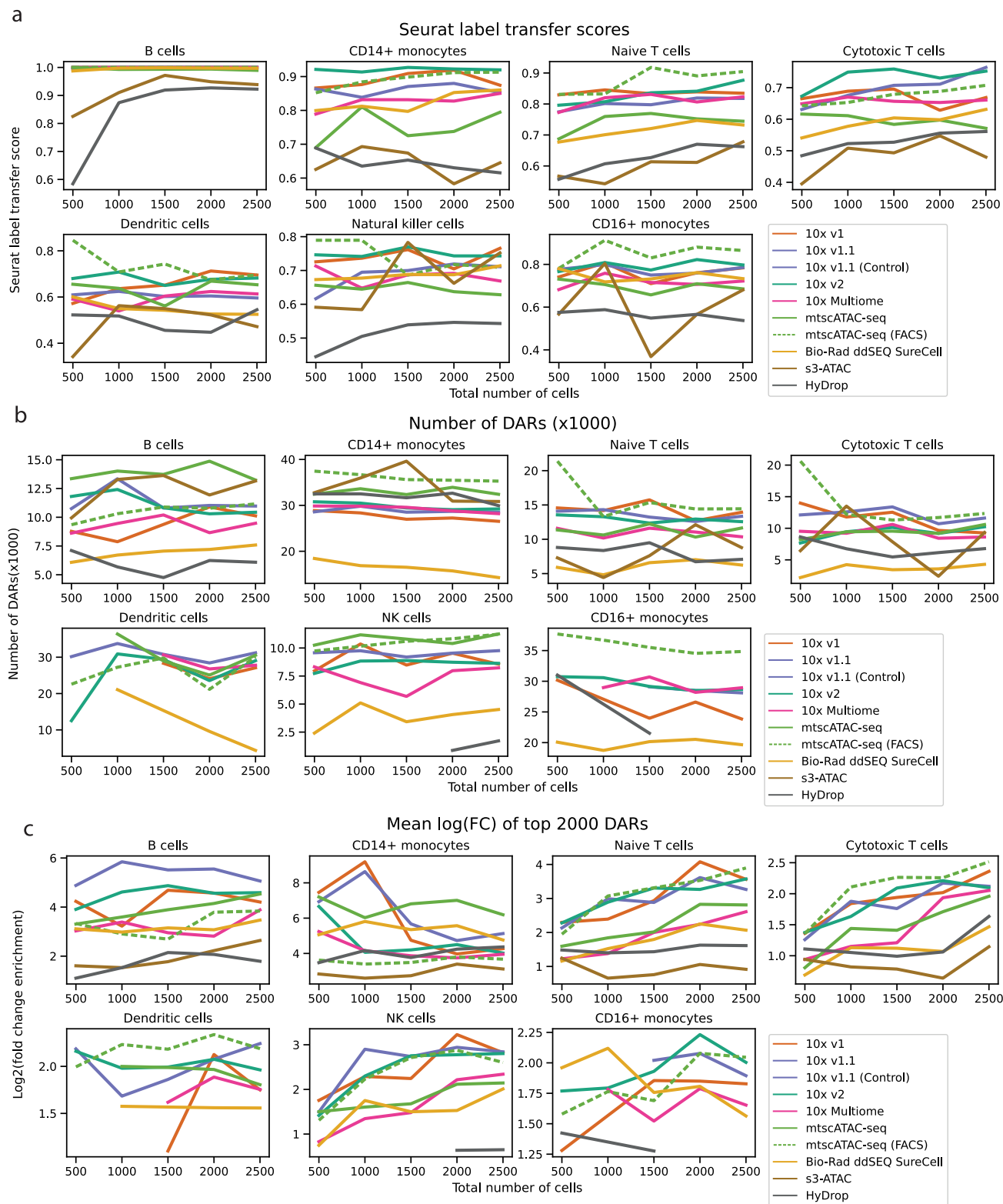


Extended Data Fig. 5 | 10x multiome RNA performance. **a**, Knee plots of RNA component of 10x multiome samples. Minimum UMI thresholds were automatically defined using Otsu's algorithm. **b**, Scatterplots of number of genes versus unique fragments in peaks detected in the same multiome cells. **c**, Venn diagrams showing overlap between cell barcodes identified as cells based on ATAC or RNA components.



Extended Data Fig. 6 | Threshold prediction curves for Seurat label transfer scores. a, Lineplots showing fraction of reads passing the threshold with increasing thresholds. Area under curve (AUC) values for each cell type in figure legends. **b**, Lineplot showing label transfer agreement between RNA and ATAC modalities of 10x Multiome samples on y axis, and Seurat scores in the ATAC cell.

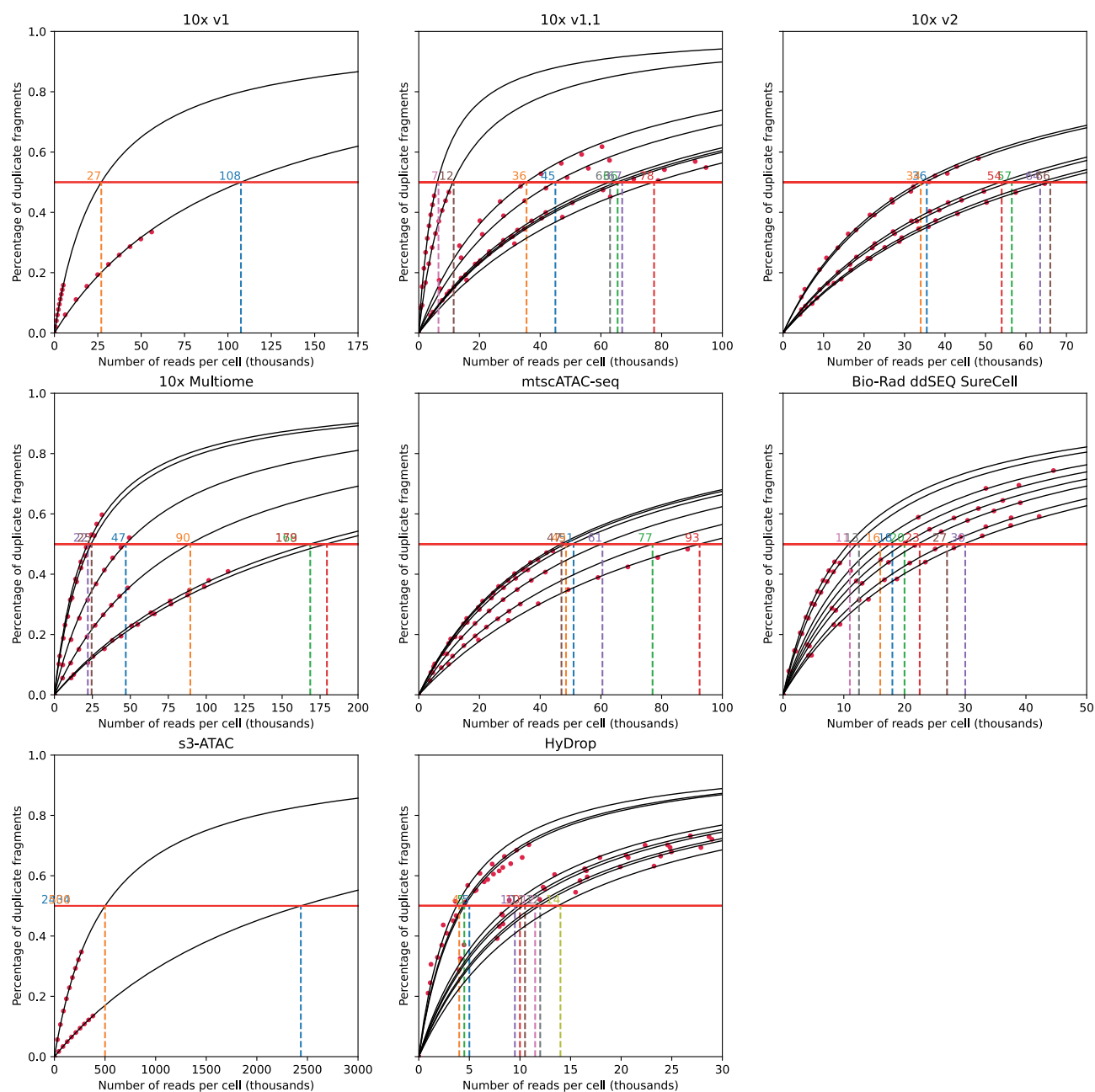




Extended Data Fig. 8 | Effect of number of cells on cell type annotation and DAR calling. For these figures, all samples with at least 2,500 cells were subset to 2,500, 2,000, ... random cells, and de-novo annotated with Seurat. **a**, Dependency of Seurat label transfer scores on total number of cells.

b, Dependency of number of DARs detected in each cell type on total number of cells. **c**, Dependency of mean log₂ fold-change enrichment of top 2,000 DARs detected in each cell type on total number of cells.

Article <https://doi.org/10.1038/s41587-023-01881-x>



Extended Data Fig. 10 | Duplication rate in fragments files. X-axis shows the number of reads per cell (thousands), y-axis shows the percentage of fragments in cells that are duplicates. Red dots indicate data points sampled from fragments files, black lines are a fitted Michaelis-Menten curve. Red line indicates a 50% duplicate rate.

Chapter 4

Discussion

The work described in this thesis encompasses two parts. First, we developed HyDrop, an open-source droplet microfluidic platform for scATAC-seq and scRNA-seq, which is competitive with commercial solutions. Second, we developed PUMATAC, a method-agnostic framework for pre-processing and evaluating scATAC-seq datasets, and used it to perform a systematic benchmark of current methods. We open-sourced all protocols and software, opening possibilities for further improvement. HyDrop has since been reproduced in other labs, and derivative products are available commercially.²⁷² Together, the work touches on all low-level aspects of omics data generation: we optimized reaction chemistries, engineered hardware, wrote software, managed large data structures, and coordinated high-power compute.

4.1 Development and evaluation of single-cell assays

Despite our advancements in open-sourcing methods, microfluidic droplet-based assays still require expertise to implement. Automating flow focusing droplet microfluidics is generally challenging, especially with standard automation hard- and software. The 10x Chromium Connect, an OEM Hamilton Nimbus liquid handler paired with an on-deck Chromium microfluidic controller, offers fully automated droplet generation and library prep, but is serial in nature: one batch of up to 8 samples is processed at a time. This approach thus does not increase sample throughput without human intervention, and is best-suited for small experimental scales. Conversely, Tahoe-100M was generated mostly by one person using 11 well-plates of Parse Bio's Evercode WT Mega kits,¹⁹⁹ which are also amenable to automation since they are plate-native.

Droplet generation itself is also a confounding process that affects the entire sample. If the process fails by the hand of dust, cell or DNA aggregates, air bubbles, chip imperfections, pump controller or gasket failures, etc., the entire sample is affected or lost completely. Manual droplet generation on an Onyx or Chromium followed by partial automation of the library prep is also impractical for many (96+) individual samples due to staggered delivery of samples and general sensitivity to delays. Plate-based single-cell methods can deal with sample multiplexing natively, for example by pre-sorting cells into plates using FACS, freezing the plates, and thawing the plates for processing on demand - a workflow which can be automated end to end. Moreover, the most recent SMART-seq assays have scaled down reagent cost through acoustic nanoliter dispensing and 1536 well-plate formats, and combinatorial indexing methods have come closer to 10x Chromium in terms of sensitivity.

Does this mean that droplet-based assays have no practical application for large scale data generation? Of course not. Tahoe-100M's cells were filtered on a minimum of 700 UMIs and 250 genes: 10x scRNA-seq undoubtedly still offers higher per-cell quality metrics. Furthermore, since the release of the 10x scATAC-seq v2

²⁷²RAN Biotechnologies, 2025.

¹⁹⁹Zhang, *bioRxiv*, 2025.

kits, which can now reach up to 100 000 unique fragments per cell, it has become clear that the sensitivity ceiling for droplet-based methods has not been attained by open-source droplet or plate-based scATAC-seq methods. For scRNA-seq, Smart-seq remains the sensitivity gold standard, but has yet to be applied to the same scales as even modest droplet-generated atlases - likely also owing to the high sequencing burden associated. Given only a small slowdown in 10x Genomics' annual kit sales²⁷³ relative to the general capital slump in biotech and research post-ZIRP, it appears that end-users are still willing to make the sensitivity/convenience/cost trade-off in favour of commercial droplet-microfluidics. Thus, how should we further develop the technology? Let us examine outstanding problems and their potential solutions:

1. **Sample throughput is limited, even on parallel droplet generation chips.** To address this issue, antibody, lipid or oligonucleotide hashes or pre-indexing have been shown to scale throughput by at least 15x,^{200,205} and can scale both samples and absolute number of cells. These multiplexing methods are generally directly compatible with HyDrop's oligonucleotide sequence designs, but none of them are currently compatible with multiomics assays.
2. **Emulsion-making is difficult to automate.** PIP-seq, a recent physical sample partitioning development, could offer a more generalizable sample throughput solution.²⁷⁴ Here, droplet partitioning is achieved by particle-templated emulsification: an aqueous phase (cells, enzyme mix and a hydrogel bead template) is vortexed with a fluorinated oil phase containing surfactant, and each bead templates one droplet of defined volume as the emulsion forms. This process can be performed in 96-well plates, and can be fully automated! The original PIP-seq article only shows scRNA-seq through this method, and implementing scATAC-seq would pose some challenges: droplet PCR does not pair well with solid or hydrogel beads - the beads must dissolve to prevent emulsion coalescence. HyDrop and 10x achieve this by disulfide bonds which are reduced by contact with the cell and enzyme mix. PIP-seq requires cell and reagent mix to be pre-mixed and vortexed together, so the disulfide approach would not work. However, inDrop has previously used o-nitrobenzyl-mediated UV cleavage to release primers, an idea which could potentially be translated to PIP-ATAC-seq: UV-cleavable beads could be vortexed together with single-cells and enzymes, and the resulting emulsion could then be exposed to UV post-hoc. While this is impractical for manual lab processing, it should not pose a problem for a lights-out automated workflow. Fluent BioSciences, which owns the IP to PIP-seq, and commercialized the assay, was recently acquired by Illumina - promising further development.²⁷⁵
3. **In-house bead production requires microfluidics and expertise.** After the HyDrop eLife paper was published, we were contacted by several labs who wanted to produce in-house beads, and correspondence revealed that bead production was the largest hurdle to implementation. One common issue was difficulty producing stable emulsions for bead generation, which may possibly be addressed by using PIP-seq-style emulsion generation for bead production as well. Aside from emulsion making, the current barcoding protocol is already fully automatable on existing hardware.

The issues presented above are related mostly to distribution and convenience rather than physical limits of droplet technology. Sensitivity differences between the open-source assay compared to the commercial alternative also play a major role in adoption, and it would thus pay off to tip the balance closer towards the open-source alternative. Since the original publication of HyDrop, an improved version has been developed and open-sourced,²⁷⁶ offering vastly improved sensitivity as well as more robust bead barcoding protocols. In this updated protocol, bead production was optimized for purity, strongly improving sequencing efficiency and reaching quality parity with 10x. Additional optimisation of extraction and amplification procedures is likely to further push the sensitivity needle in favor of open-source protocols.

Awareness of the sensitivity and specificity biases described here is especially important in scATAC-seq studies compared to scRNA-seq studies, since peak region sets are sample-dependent rather than fixed gene sets. Therefore, high-specificity methods should be prioritised when generating reference peak region sets if public references (e.g. ENCODE ChIP-seq or DNase-seq) are not available a priori. Cost-efficient but lower sensitivity/specificity open-source methods can be used for data generation at scale. Sequencing efficiency of these

²⁷³10x Genomics, Inc., 2025.
²⁰⁰Datlinger, Nat Meth, 2021.

²⁰⁵Mylka, Gen Biol, 2022.
²⁷⁴Clark, Nat Biotech, 2023.

²⁷⁵Illumina, Inc., 2024.
²⁷⁶Dickmanken, bioRxiv, 2025.

open-source methods can be improved by sequencing less deeply, since fewer reads will be lost to duplicates, and available resources can be allocated to sampling more cells rather than sequencing more. This strategy will lead to lower per-cell sensitivity, but increasing the total number of cells may be able to partially compensate. For example, both cell type annotation quality (seurat label transfer scores) and DAR strength scaled with higher cell counts (fig. 3a). We hypothesize that the benefit of sampling additional cells diminishes when per-cell information becomes insufficient for the analysis of interest.

4.2 Scaling omics data for predictive modelling

Overall, our project feeds into a broader, cross-disciplinary trend of scaling large data explicitly for model training. This trend first began to show results in games like chess and Go,²⁷⁷ but has recently manifested in more tangible applications such as medical imaging, self-driving cars,²⁷⁸ and language models.²⁷⁹ For example, whereas Google's BERT was pre-trained on BookCorpus and English Wikipedia, some 16 GB of uncompressed text, in 2019,^{280,281} OpenAI's GPT-3 was trained just a year later on 300 billion tokens drawn from a mixture totalling 499 billion tokens: filtered Common Crawl (410B), two internet-based book corpora (12B and 55B), WebText2 (19B) and Wikipedia (3B).²⁷⁹ Compute was scaled by roughly three orders of magnitude over the same interval, from 5.3×10^{20} FLOPs for BERT-Large to 3.14×10^{23} FLOPs for GPT-3 175B.²⁷⁹ The resulting performance improvement, which was primarily the result of a large commitment of data and compute, has allowed GPT-3 to penetrate into consumer markets, eventually becoming the most rapidly adopted technology of all time.²⁸² Exploring the parameter sensitivity ("scaling laws") of transformer models has shown that their performance continues, and is likely to continue, to scale sublinearly with model dimensions, training dataset size, and compute.^{283,284}

Biology is almost certainly at an inflection point in world-total data generation capacity. The cost of sequencing and quantitative sequencing assays has been falling rapidly at predictable rates.²⁸⁵ What was at the start of this work considered a large atlas of 100 000 cells can now be routinely generated in a few days, and 10 to 100 million cell datasets are being released. A question that is much debated today is **whether the data + compute scaling paradigm will generalize to biological models**. Increasing model and dataset scale has improved performance on several evaluated biological tasks. Gains depend on the task, data composition, and evaluation design.^{263,286} Especially transformer-based protein-language models, notably Meta's ESM family, seem to obey the same scaling laws as their natural language counterparts: between 2020 and 2025, the model family iterated from 650M²⁸⁷ to 15B²⁸⁸ and recently 98B²⁸⁹ parameters, adding generative features.

However, deeper scrutiny reveals more nuance. First, architectural novelty has in some cases outshined pure scaling, especially in regimes where data is scarce. For example, AlphaFold 1 was already trained on a large fraction of available PDB and MSA data, and AlphaFold 3's weights are only 1 GB in size. Introduction of attention with Evoformer blocks and end to end diffusion led to the largest performance jumps. Recent benchmarks have also shown that protein/ligand binding models suffer performance drops when re-trained with more stringent train-test leakage filtering.²⁹⁰ Performance on ligands that were sparsely represented in the training data was especially weak, suggesting that "*data scarcity remains a major challenge for these methods*".²⁹¹

Second, smarter data engineering can rival or exceed sheer volume. Nucleotide Transformer improved more on human tasks by adding 850 non-human genomes than by increasing model size from 500M to 2.5B parameters. Evo 2 was released as a 7B checkpoint trained on 2.4T tokens and a 40B checkpoint trained on 9.3T tokens, so the smaller model consumed roughly one twentieth of the training compute. Both were pretrained with data weighting focused on genic windows. On clinical variant benchmarks the two perform competitively: the 7B and 40B models rank together ahead of unsupervised baselines on coding SNVs, on BRCA1 saturation mutagenesis and on SpliceVarDB, and the 40B pulls clearly ahead mainly on noncoding variants.²⁶³ Scale showed

²⁷⁷ Sutton, 2019.

²⁷⁸ Waymo, 2025.

²⁷⁹ Brown, arXiv, 2020.

²⁸⁰ Devlin, arXiv, 2019.

²⁸¹ Liu, 2019.

²⁸² Parveen, Sci Rep, 2024.

²⁸³ Kaplan, arXiv, 2020.

²⁸⁴ Hoffmann, 2022.

²⁸⁵ Wetterstrand, 2023.

²⁶³ Brixi, bioRxiv, 2025.

²⁸⁶ Adduri, Arc Institute, 2025.

²⁸⁷ Rives, PNAS, 2021.

²⁸⁸ Lin, Science, 2023.

²⁸⁹ Hayes, Science, 2025.

²⁹⁰ Durairaj, bioRxiv, 2024.

²⁹¹ Škrinjar, 2025.

up more visibly in generation quality than in raw capability: both models produced coherent mitochondrial-scale sequence, but tetranucleotide usage of generated sequence agreed more closely with natural genomes for the 40B model.

Third, smaller, application-specific models routinely outperform large foundation models in constrained contexts. Bespoke, domain-focused networks trained from scratch on single-cell data such as DeepMEL for melanoma, DeepFlyBrain for *Drosophila*, and DeepZebrafish for embryogenesis, reach higher precision than Enformer and ChromBPNet on enhancer logic while only a few million parameters in size.

Given the high labor and financial cost associated with (single-cell) sequencing assays, and the open question whether the scaling strategy will pay off proportionally, naively launching large omics scaling operations has not been entirely derisked. Considering the nuances above, it might not be the optimal path forward with our current knowledge today, and perhaps focusing on deeper sampling of individual samples might be a better strategy.

When comparing quantitative omics with DNA language modeling, we must note that the amount of pure genomic data available is large: GenBank contains 34 Tbp from 4.7 billion nucleotide sequences representing 581 000 formally described species.²⁹² Conversely, the quantitative omics space has been sampled much more sparsely. At the time of its most recent description, GEO held over 200 000 studies and 6.5 million samples,²⁹³ an Entrez query in September 2026 returns 295 798 series across 8.7 million samples, of which 144 954 series are expression profiling by high-throughput sequencing, 50 953 series are genome binding or occupancy profiling, and fewer than 10 000 series mention ATAC-seq at all. High-quality perturbation datasets are even fewer in number. This is especially problematic considering that the effect space is also highly combinatorial for both genetic and chemical perturbations. Gigabytes of .fastq-formatted sequencing data also often reduce to sparse count matrices of fewer than 50 MB in size. In this regard, quantitative omics might be too sparse yet to derive scaling curves.

Assuming a scaling strategy will hold for quantitative omics, several challenges lie ahead still. Because of the large amount of capital expenditure associated with even modestly-sized omics projects, the main problem is identifying which data types to commit to. Assuming a fixed budget, and a goal of "predict gene regulation in a given disease" (essentially, current virtual cell efforts), several factors affect how we should design a data collection effort:

1. **Which data type(s) should we collect?** Within sequencing omics assays, we would need to decide on modality (DNA versus accessible chromatin, RNA, Hi-C, methylation, ...), bulk versus single-cell, long versus short reads and so on. Beyond sequencing, there are also dozens of imaging techniques, intracellular and surface protein measurements, functional assays, ... etc. available. There are certainly interaction terms between these different options - for example, ATAC + RNA might be more informative than either alone. We currently do not have a good quantification of "information content per money spent" for various omics techniques. In LLM pre-training, smaller training runs are used to calculate expected information gain and marginal loss reduction, but text acquisition is approximately costless compared to omics data.
2. **On which sample?** Omics measurements are generally destructive (although live-cell imaging methods can now monitor selected RNA species).²⁹⁴ As a result, studies in humans have concentrated around blood, tissue biopsies, and banked samples. While the translational benefit of testing on human samples is directly evident, these samples are not only scarce in absolute terms, but also difficult to obtain due to regulatory and privacy concerns. From a pure information gathering perspective, it is unclear if focusing exclusively on human tissue is the correct call today. For a given budget, substantially deeper sampling of model organisms may prove to be valuable: regulatory programs are conserved across surprisingly distant species.²⁵⁸ Even simple yeast retains nucleosomes and histone modifications, including methylation, although it lacks cytosine DNA methylation system found in mammals.^{295,296} Cultivating yeast and rearing flies, across many strains and mutants, is more practical than clinical sample collection at scale, and well-controlled genetic backgrounds are a significant benefit. It is thus possible that model organism data at scale could yield models that transfer well to human genomes. Patient-derived organoids could

²⁹²Sayers, Nuc Ac Res, 2025.

²⁹³Clough, Nuc Ac Res, 2024.

²⁹⁴Zuo, Nature Protocols, 2026.

²⁵⁸Minnoye, Gen Res, 2020.

²⁹⁵Carrozza, Cell, 2005.

²⁹⁶Capuano, Analytical Chemistry, 2014.

offer a compromise between clinical relevance and experimental scalability. Culture variability remains a limitation, with organoid differentiation efficiency being highly dependent on input iPSC.²⁹⁷

3. **Investigating which perturbation(s), if any?** Some of the largest current data collection efforts use massively parallel perturbation sampling with single-cell transcriptomic readouts, including both drug and genetic perturbations.^{298–300} However, parallel perturbations have come at the cost of sampling depth and precise control over dosage. Intelligent, iterative perturbation library design might be a source of optimizations.³⁰¹
4. **Using which type of automation?** Even if up-front funding were limitless, scaling up data collection will require more than just parallelizing existing personnel and machines. Today, manual research campaigns require different machines, used over various days, and these combinations are generally unique to a project. Moreover, there is certainly a lot of skill involved in collecting and processing samples. Committing to a single scaled data pipeline, where a tissue or other sample is entered one end, and data comes out the other, will require at least the re-designing of existing machines and their close integration with liquid handlers. Humanoid robots might assist during a transitory period, but for large scaling projects, they will most likely not be deployed beyond maintenance work. Similarly to what droplet microfluidics did for single-cell, automated capillary gel-electrophoresis did for DNA sequencing, flow cytometry did for cell sorting, ... we require additional development of specialized monotask machines. Especially in parallel treatment of many samples, more work is needed. For example, a machine that could ingest an arbitrary tissue, and return a highly viable single-cell suspension *reliably* still seems beyond reach for now.
5. **Sampling how many cells, at what per-cell depth?** Even the most sensitive plate-based protocols recover under half of a cell's mRNA molecules,¹⁶⁰ and the droplet and split-pool chemistries used for atlas-scale data recover roughly 5–15% (3.4% for CEL-seq, 7.1% for inDrop, 12.8% for Drop-seq, 6.7–8.1% for 10x).^{175,176,180,302,303} We can thus tune sequencing and assay sampling depth, with number of cells sequenced. For genomic assays, a similar trade-off exists, and single-cell sparsity is even more severe.

Table 4.1 places the largest current data-generation efforts on these five dimensions.

The Human Cell Atlas (HCA), founded in 2016 and still ongoing, prioritizes broader coverage of human cellular diversity. By 2024, the project spanned 100 million cells from over 10 000 individuals, using single-cell and spatial approaches to construct reference maps across tissues.^{304,305} The BRAIN Initiative Cell Atlas Network (BICAN), launched by NIH in 2022 with \$500 million over five years, profiles single-nucleus transcriptome, chromatin accessibility, DNA methylation and spatial transcriptomics from several hundred human brains, and releases data publicly.^{306,307} HuBMAP profiles healthy human organs with multiplexed imaging, spatial transcriptomics and single-nucleus assays.³⁰⁸

Large efforts within Aligning Science Across Parkinson's (ASAP) investigate genetic variation and cell-type-specific regulatory mechanisms in Parkinson's disease.^{309,310} The work reports long-read whole-genome sequencing alongside snATAC-seq profiles from 3.1 million nuclei and snRNA-seq profiles from 1.1 million nuclei, from two brain regions across 190 donors, including 75 individuals with Parkinson's.

Similar projects are launching in China. BGI-Research and collaborators founded the 10 Billion Cells Alliance in August 2025, proposing reference, disease and perturbation-response atlases using Stereo-cell (cellular imaging + transcriptomics).³¹¹ The Chinese Immune Multi-Omics Atlas, published in 2026, uses single-cell RNA and chromatin accessibility measurements from more than 10 million circulating immune cells across 428 adults.³¹²

CIMA's OneK1K profiled 1.27 million blood cells from 982 people and mapped cell-type-specific eQTLs to autoimmune risk loci.³¹³ Its follow-up project, TenK10K, launched in 2024 by Garvan with Illumina and 10x, targets 50 million cells and genomes from 10 000 people.³¹⁴

²⁹⁷ Andersson-Rolf, *Nature Reviews Molecular Cell Biology*, 2026.

²⁹⁸ Tahoe Therapeutics, 2026.

²⁹⁹ Huang, *bioRxiv*, 2025.

³⁰⁰ Zhu, *Cell*, 2026.

³⁰¹ Huang, *bioRxiv*, 2023.

¹⁶⁰ Hagemann-Jensen, *Nat Biotech*, 2020.

¹⁷⁵ Macosko, *Cell*, 2015.

¹⁷⁶ Klein, *Cell*, 2015.

¹⁸⁰ Zheng, *Nat Comm*, 2017.

³⁰² Grün, *Nature Methods*, 2014.

³⁰³ Svensson, *Nature Methods*, 2017.

³⁰⁴ Human Cell Atlas, 2024.

³⁰⁵ Chan Zuckerberg Initiative,

³⁰⁶ Kaiser, 2022.

³⁰⁷ Ecker, *bioRxiv*, 2026.

³⁰⁸ HuBMAP Consortium, *Nature*, 2019.

³⁰⁹ Aligning Science Across Parkinson's,

³¹⁰ Sigalova, *bioRxiv*, 2026.

³¹¹ BGI Group, 2025.

³¹² Yin, *Science*, 2026.

³¹³ Yazar, *Science*, 2022.

³¹⁴ Garvan Institute of Medical Research,

A number of startups have also oriented explicitly on a "data-first, model next" path to revenue. Noetik, founded in 2022, generates microarrays from clinical tissues, and profiles them using H&E histology, spatial transcriptomics, spatial protein measurements and DNA sequencing.³¹⁵ Their private model is trained on paired histology and spatial transcriptomic data, and predicts spatial gene expression from H&E alone.³¹⁶ Recursion, another startup, also focuses heavily on imaging, using immunofluorescence plus chemical perturbations. Their public RxRx3 resource includes more than 17 000 gene knockouts and 1600 compounds.³¹⁷ Xaira's X-Atlas/Orion (2025) applied genome-wide CRISPRi to two cell lines and sequenced eight million cells to more than 16 000 UMIs each.²⁹⁹ X-Atlas/Pisces extends this to 25.6 million cells across seven cellular contexts, currently the largest Perturb-seq dataset.³¹⁸ Owkin's MOSAIC, a \$50 million consortium with five European and US cancer centres, profiles 7000 tumour samples across seven cancer types with GeoMx, CosMx, Visium and Chromium.³¹⁹ Sixty patients are open access, the rest is held private.

IGVF (NHGRI, 2021-2026) maps chromatin accessibility and transcriptome with single-nucleus multiome in cell lines, iPSC-derived and primary cells, perturbs variants, elements and genes with CRISPR screens, MPRAs and saturation mutagenesis, and trains models of variant effect on the result.^{320,321} It is the only consortium that pairs perturbation with a regulatory readout, at 10⁴-10⁵ cells per experiment.

Tahoe-100M contains 100 million single-cell transcriptomes of 50 cancer cell lines treated with 379 small molecules: 1135 drug-dose conditions, 52 886 line-drug conditions, median 1287 cells per condition.¹⁹⁹ After Tahoe's initial proof of concept, Arc, Biohub and Tahoe announced a joint effort of more than 120 million cells spanning 225 000 drug-patient interactions, with more than four times the perturbation richness of Tahoe-100M.²⁹⁸

The Billion Cells Project, launched between CZI, 10x and Ultima Genomics in 2025, targets one billion single-cell profiles including genetic perturbations, in mouse, zebrafish and primary human cell models.³²² Illumina's competing Billion Cell Atlas, in partnership with Eli Lilly, AstraZeneca and Merck, now aims to measure CRISPR-perturbation responses using scRNA-seq in 200 disease cell lines.³²³ The work relies heavily on the PIP-seq IP that Illumina re-packaged into its Illumina Single Cell 3' RNA Prep kit. In July 2026, Illumina reported that more than 350 million cells had been sequenced for its Billion Cell Atlas.³²⁴

TABLE 4.1 Large-scale omics data-generation efforts mapped onto the five design dimensions, grouped by strategy.

Effort	Readout	System	Perturbation	Production	Scale
<i>Observational: natural variation in native tissues</i>					
Human Cell Atlas, 2024 status ^{304,305}	scRNA-seq, spatial transcriptomics	Native human tissue, mostly healthy	None (reference)	Distributed across many laboratories; droplet	100M cells, > 10 000 donors
BICAN, 2022-2027 ^{306,307}	snRNA-seq, snATAC-seq, single-cell methylation, spatial transcriptomics	Whole human brain across the lifespan; macaque, marmoset, mouse	None (reference)	Consortium of 11 centres; droplet and plate	\$500M over 5 years; several hundred human brains planned
ASAP Parkinson's, 2026 ^{309,310}	Long-read WGS, snATAC-seq, snRNA-seq	Post-mortem human brain, two regions; PD and control	None (donor variation)	Consortium; droplet	3.1M snATAC-seq nuclei; 1.1M snRNA-seq nuclei; 190 donors (75 PD)
CIMA, 2026 ³¹²	scRNA-seq, scATAC-seq	Human blood	None (donor variation)	Consortium; droplet	10M cells, 428 donors

continued on next page

2024.	³¹⁸ Xaira Therapeutics, 2026.	³²² Chan Zuckerberg Initiative, 2025.
³¹⁵ Noetik,	³¹⁹ Owkin, 2023.	³²³ Illumina, 2026.
³¹⁶ Noetik,	³²⁰ IGVF Consortium, <i>Nature</i> , 2024.	³²⁴ Illumina, 2026.
³¹⁷ Recursion Pharmaceuticals, Inc.,	³²¹ Li, <i>Nucleic Acids Research</i> , 2026.	

Table 4.1 (continued)

Effort	Readout	System	Perturbation	Production	Scale
TenK10K, 2024 ^{313,314}	scRNA-seq (10x GEM-X), WGS	Human blood; autoimmune, cardiovascular and cancer cohorts	None (donor variation)	Single centre (Garvan) with Illumina and 10x support	50M cells, 10 000 donors (target); predecessor OneK1K, 1.2M cells from 982 donors
HuBMAP; HTAN (NIH), 2018- ^{308,325}	Spatial proteomics and transcriptomics, snRNA-seq, snATAC-seq, imaging	Healthy human organs (HuBMAP); tumours across progression (HTAN)	Natural variation; disease progression	Consortia; mixed platforms	-
Noetik, 2022- ^{315,316}	H&E, spatial transcriptomics, spatial proteomics, DNA-seq	Clinical tumour tissue (microarrays)	Natural variation, treatment history	In-house pipeline	-
Owkin MOSAIC, 2023- ³¹⁹	H&E, spatial transcriptomics, scRNA-seq, WES	Clinical tumour tissue, seven cancer types, five hospitals	Natural variation, treatment history	Consortium with 10x	~7000 samples (target)
BGI 10 Billion Cells Alliance, 2025 ³¹¹	Imaging + transcriptomics (Stereo-cell)	Reference, disease and perturbed samples (proposed)	Proposed	In-house platform	1×10^{10} cells (target)
<i>Hybrid: mapping combined with perturbation</i>					
IGVF (NHGRI), 2021-2026 ^{320,321}	snATAC-seq, snRNA-seq (multiome), MPRA, saturation mutagenesis, single-cell CRISPR screens	Human cell lines, iPSC-derived and primary cells; mouse	Variants, elements and genes: CRISPR, MPRA, deep mutational scanning	Consortium; distributed	-
<i>Interventional: genetic or chemical perturbations</i>					
CZI Billion Cells Project, 2025 ³²²	scRNA-seq (10x; Ultima sequencing)	Mouse, zebrafish, primary human cells	Genetic, plus observational	Distributed; droplet	1×10^9 cells (target)
Recursion RxRx3 ³¹⁷	Fluorescence imaging	HUVEC	17 000 knockouts, 1600 compounds	Fully robotic, in-house	~2.2M images
Tahoe-100M, 2025 ¹⁹⁹	scRNA-seq (split-pool)	50 cancer cell lines	379 compounds, 1135 drug-dose conditions	Plate-native split-pool indexing (Mosaic)	100M cells, 52 886 line-drug conditions, median 1287 cells per condition
X-Atlas/Orion, 2025 ²⁹⁹	scRNA-seq (10x Flex)	Two cell lines	Genome-wide CRISPRi, all protein-coding genes	Fix-cryopreserve (FiCS) batching	8M cells at >16 000 UMIs per cell
X-Atlas/Pisces, 2026 ³¹⁸	scRNA-seq (10x Flex)	Seven cellular contexts	Genome-wide CRISPRi	FiCS batching	25.6M cells
Arc-Biohub-Tahoe, planned 2026 ²⁹⁸	scRNA-seq	Patient-derived disease models	Compounds; >4× the perturbation richness of Tahoe-100M	Contract production by Tahoe (Mosaic)	120M cells, 225 000 drug-patient pairs
Recursion-Roche-Genentech, 2026 ^{326,327}	Imaging + transcriptomics	iPSC-derived neurons and microglia	Genome-wide genetic	Robotic, in-house	-

continued on next page

Table 4.1 (continued)

Effort	Readout	System	Perturbation	Production	Scale
Illumina Billion Cell Atlas, 2026 ³²³	scRNA-seq (PIP-seq)	200 disease cell lines	CRISPR	Vortex emulsification, plate-native	350M of 1×10^9 cells (July 2026)

We can see trends emerge in current mega-projects. Generally speaking, academic consortia more often work with native tissues, measure several modalities per donor, and publish their findings and data as public reference resources. Industry opts for a broad perturbation approach, often in cell lines, a decision which stems from the industry’s aim to ultimately predict the effects of future therapeutic interventions. Perturbation efforts read out either the transcriptome or images, none reads out a second sequencing modality. Chromatin accessibility appears only in the observational projects (ASAP, CIMA). Perturb-ATAC, CRISPR-sciATAC and Spear-ATAC established this is technically possible,^{328–330} but have so far only been demonstrated at scales of up to 100,000 cells. Mouse and zebrafish appear just once, yeast and *Drosophila* not at all. Arc, Tahoe and Roche discuss future plans in organoids, but as of 2026, no data has been published yet. The largest cell counts come from plate-native, internal strategies: SPLiT-seq (Tahoe), PIP-seq (Illumina) and fixed-cell batching (X-Atlas). The observational consortia scale instead by distributing conventional droplet workflows across many laboratories, and as a result achieve better quality depth per sample.

Together, this cohort of collection projects suggests both industry and academic confidence in omics data scaling strategies has grown significantly. It is also likely that the process of data scaling will self-reinforce. Better understanding of proteins and gene regulation will lead to development of better molecular tools such as synthetic enzymes, optimal expression systems, and better perturbation delivery methods. Improved models will also improve experimental design through more accurate hypothesis (and thus informative data) generation, and may even generate synthetic data for finetuning. Advances in lab automation coupled with automated protocol generation through LLM support (PyLabRobot) will further accelerate this process.³³¹

It remains yet to be seen which strategy will contribute positively to drug discovery and therapy development. If all goes well, the emerging data ecosystem might contain several complementary types of resource: a mixture of broad reference atlases, characterized patient cohorts, and large private perturbation models. As model predictive performance increases, downstream medicine and consumer products will further stimulate capital attention and model scale. In that sense, the biotech and pharmaceutical industry may soon become a data economy where information becomes the primary currency of discovery.

³²⁸Rubin, *Cell*, 2019.

³²⁹Liscovitch-Brauer, *Nature Biotechnol-*

ogy, 2021.

³³⁰Pierce, *Nature Communications*, 2021.

³³¹Wierenga, *bioRxiv*, 2023.



TOOLS AND RESOURCES



Hydrop enables droplet-based single-cell ATAC-seq and single-cell RNA-seq using dissolvable hydrogel beads

Florian V De Rop^{1,2}, Joy N Ismail^{1,2}, Carmen Bravo González-Blas^{1,2},
Gert J Hulselmans^{1,2}, Christopher Campbell Flerin^{1,2,3}, Jasper Janssens^{1,2},
Koen Theunis^{1,2,3}, Valerie M Christiaens^{1,2}, Jasper Wouters^{1,2}, Gabriele Marcassa^{1,4},
Joris de Wit^{1,4}, Suresh Poovathingal^{1*†}, Stein Aerts^{1,2*†}

¹VIB-KU Leuven/VIB Center for Brain & Disease Research, Leuven, Belgium;

²Laboratory of Computational Biology, Department of Human Genetics, KU Leuven, Leuven, Belgium; ³Aligning Science Across Parkinson's (ASAP) Collaborative Research Network, Chevy Chase, United States; ⁴Laboratory of Synapse Biology, Department of Neurosciences, KU Leuven, Leuven, Belgium

Abstract Single-cell RNA-seq and single-cell assay for transposase-accessible chromatin (ATAC-seq) technologies are used extensively to create cell type atlases for a wide range of organisms, tissues, and disease processes. To increase the scale of these atlases, lower the cost and pave the way for more specialized multiome assays, custom droplet microfluidics may provide solutions complementary to commercial setups. We developed HyDrop, a flexible and open-source droplet microfluidic platform encompassing three protocols. The first protocol involves creating dissolvable hydrogel beads with custom oligos that can be released in the droplets. In the second protocol, we demonstrate the use of these beads for HyDrop-ATAC, a low-cost noncommercial scATAC-seq protocol in droplets. After validating HyDrop-ATAC, we applied it to flash-frozen mouse cortex and generated 7996 high-quality single-cell chromatin accessibility profiles in a single run. In the third protocol, we adapt both the reaction chemistry and the capture sequence of the barcoded hydrogel bead to capture mRNA, and demonstrate a significant improvement in throughput and sensitivity compared to previous open-source droplet-based scRNA-seq assays (Drop-seq and inDrop). Similarly, we applied HyDrop-RNA to flash-frozen mouse cortex and generated 9508 single-cell transcriptomes closely matching reference single-cell gene expression data. Finally, we leveraged HyDrop-RNA's high capture rate to analyze a small population of fluorescence-activated cell sorted neurons from the *Drosophila* brain, confirming the protocol's applicability to low input samples and small cells. HyDrop is currently capable of generating single-cell data in high throughput and at a reduced cost compared to commercial methods, and we envision that HyDrop can be further developed to be compatible with novel (multi) omics protocols.

***For correspondence:**
suresh.poovathingal@kuleuven.be (SP);
stein.aerts@kuleuven.be (SA)

[†]These authors contributed equally to this work

Competing interest: The authors declare that no competing interests exist.

Funding: See page 21

Preprinted: 06 June 2021

Received: 16 September 2021

Accepted: 21 February 2022

Published: 23 February 2022

Reviewing Editor: Naama Barkai, Weizmann Institute of Science, Israel

© Copyright De Rop et al. This article is distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use and redistribution provided that the original author and source are credited.

Editor's evaluation

De Rop et al. introduce a flexible microfluidics-based single-cell genomics technology that expands and improves previously existing custom droplet-based scRNA-seq protocols (inDrops and Drop-seq) in interesting directions: better data quality, simplified workflow, high-cell recovery, and flexibility towards other single-cell applications.

FIGURE A.1. Article title page of "Hydrop enables droplet-based single-cell ATAC-seq and single-cell RNA-seq using dissolvable hydrogel beads", chapter 2

Appendix A

Supplementary note: Design and development of HyDrop-ATAC and HyDrop-RNA

In this note, we will walk through the design and development of HyDrop, our cost-efficient and transparent single-cell ATAC (and single-cell RNA) sequencing platform. We published this work in eLife in 2022 (fig. A.1, chapter 2).³³²

A.1 Method design and prior work

One of the first design choices was whether to use inDrop or Drop-seq style beads. Drop-seq's beads are made out of solid resin surface-coated with oligonucleotides whereas inDrop uses a soft, deformable polyacrylamide hydrogel with oligonucleotide barcodes loaded throughout. Both ideas are superficially similar, but inDrop's soft hydrogel beads have one major advantage: they can be densely stacked in a microfluidic channel, leading to deterministic loading, enabling a higher cell capture rate of $> 80\%$ versus $< 5\%$ (see also section 1.6). We also considered that manufacturing and barcoding polyacrylamide hydrogels would be more straightforward than resin beads, the latter which we would need to source from commercial suppliers. Based on these considerations, the choice was made to use polyacrylamide hydrogel beads.

Next, we considered the bead barcoding reaction. In Drop-seq, oligonucleotide barcodes are synthesized directly on the resin bead in 12 rounds of split-pool oligonucleotide synthesis (4^{12}). In inDrop, oligonucleotide sub-barcodes are annealed and copied to a primer stub by non-cycled isothermal extension in 96^3 split-pooling. We reasoned that a cycled linear barcode amplification reaction rather than an isothermal extension could lead to higher bead barcoding reaction efficiency, subsequent higher density of barcodes within the bead, and higher in-droplet target capture as a result.

We also wanted to replace inDrop's photocleavable oligonucleotide linker by a disulfide bond, permitting the release of barcodes by reduction with DTT. This change would remove inDrop's requirement to work in the dark and possibly increase sensitivity. After all, ambient light exposure to light could cause the beads to prematurely release their barcodes - reducing the assay's specificity. By the end of the year, the state of our development had reached the following milestones:³³³

1. We designed and manufactured a microfluidic chip for the generation of the stable acrylamide monomer

³³²De Rop, eLife, 2022.

³³³De Rop, KU Leuven, 2019.

emulsions needed to produce polyacrylamide hydrogel beads using soft lithography.³³⁴ We characterised the relationship between oil and water flow rates on this chip, and predicted flow rates that would generate 50 μm diameter droplets with minimal size variance using iterative linear model optimisation (fig. A.2). The resulting beads were indeed of a stable 53 μm diameter with a standard deviation of 5 μm .

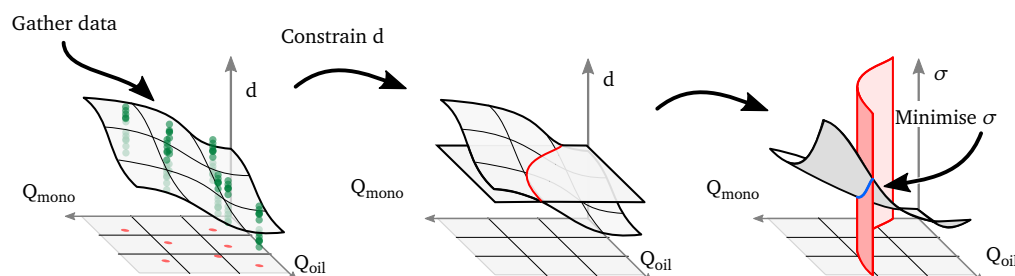


FIGURE. A.2. Bead diameter optimisation strategy. From [333].

2. We barcoded these polyacrylamide hydrogel beads using split-pooling, producing 384×384 possible cell barcodes across all beads.
3. We replaced inDrop's photocleavable oligonucleotide linker with a disulfide bond, and we successfully validated that our beads, after incubation with DTT, released primers into the supernatant.
4. We designed a new cell encapsulation microfluidic chip which included several layers of passive filters due to prior experiences with flow-obstructing dust particles. Using this updated design, we were able to reliably encapsulate single cells with reagent mixes across several experiments.

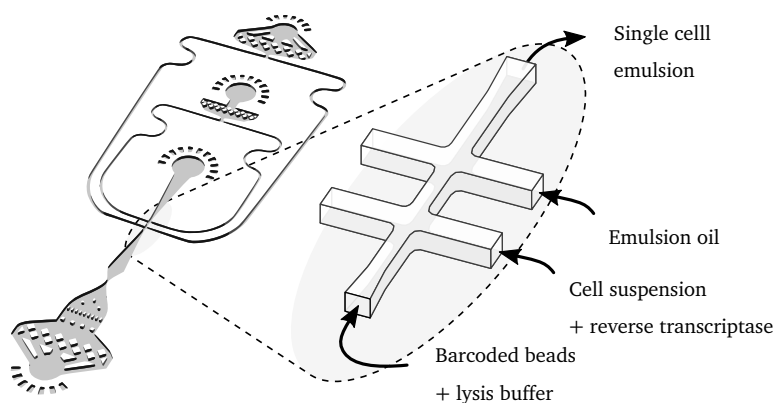


FIGURE. A.3. Droplet encapsulation chip design. Reproduced from [333].

5. To implement scRNA-seq on our now functional droplet platform, we tested several different cDNA generation and subsequent sequencing library preparation strategies, focusing on total cDNA and library yield as optimisation targets. Ultimately, we settled on a protocol that was based on in-droplet Maxima H Minus reverse transcription, PCR amplification of cDNA by KAPA polymerase, and NEB Next Ultra II FS DNA sequencing library preparation. This protocol was different from both inDrop, which uses a more involved reverse transcription $\rightarrow$ second-strand synthesis $\rightarrow$ in-vitro transcription $\rightarrow$ random hexamer reverse transcription $\rightarrow$ PCR workflow, and Drop-seq, which relies on Tn5-based sequencing library prep from cDNA.

In short, we had set up a workflow for producing hydrogel beads, barcoding those beads using linear amplification, and using them in scRNA-seq experiments that produced sequencing-ready cDNA libraries. In parallel to the scRNA-seq research, we had also begun preliminary scATAC-seq trials. This required the production of a new type of beads, this time with oligos that annealed to the Tn5 insertion adapter sequence instead of to poly(A)

³³⁴Xia, Angew Chem, 1998.

tails. Our scATAC-seq protocol relied on encapsulating pre-tagmented single nuclei into droplets together with the new ATAC-seq compatible beads and a variety of polymerases and commercial PCR master mixes. PCR cycling these emulsions should then lead to pre-amplified scATAC-seq libraries.

A.2 Execution and troubleshooting

What followed was a long series of troubleshooting issues which were often shared between the scRNA-seq and scATAC-seq protocols. In rough chronological order, we dealt with the following:

Problem: The bulk data characteristics of our earliest scRNA-seq libraries were very promising: expected cDNA and final library fragment size distributions were as expected, ~40% of reads aligned to reference transcriptome and these reads were 3' enriched. However, we were experiencing difficulties sequencing the cell barcodes: between 67% and 70% of the barcode reads contained poly-G repeats. Out of the remaining 30%, only 0.2% matched our barcode whitelist, compared to the > 95% expected for 10x Genomics scRNA-seq datasets. This failure made it impossible to assign transcripts to cell of origin.

Solution: At first, we thought that the custom sequencing primer used to sequence the cell barcodes might be at fault. We knew that our NextSeq500 used Illumina's 2-channel laser technology. G bases are encoded by the absence of signal, whereas the other three bases are encoded by positive fluorescence in either or both of the two dyes used. Poly-G base calls in the final data could thus be interpreted as the absence of any signal from the sequenced molecule, indicating that the sequencing primer may not have annealed properly - precluding the generation of any signal. We attempted to test this hypothesis by making a batch of mock beads in which the custom priming site was replaced by Illumina's standard primers - which are highly optimized for specificity. The resulting dataset yielded barcode reads of which now only 33% had a poly-G problem, but still only around 23% matched the whitelist - still far below what was necessary to do a successful single-cell sequencing experiment. Next, we redesigned the barcode structure so that the cell barcode was now read in one go, and did not require the use of a custom sequencing primer. More than 90% of the new barcode reads matched our whitelist. The remaining 10% could be attributed to incorporation errors during vendor oligonucleotide synthesis, PCR, and sequencing errors.

Later on, this change also had an important effect on our scATAC-seq assay: it allowed us to sequence both ends of the chromatin fragments as opposed to only a single end. Since ATAC-seq quantifies Tn5 insertion events, this increases the counting sensitivity and positional accuracy of called peaks. Note that insertion events within the same peak region are correlated due to proximity, so total information content is not doubled by sequencing both ends of all fragments. Another welcome effect of paired-end sequencing is that it improves cross-contamination estimation: when two cell barcodes are statistically enriched in identical fragments (exact same insertion sites with base-pair accuracy), they are highly likely to be cross-contaminated, a feature which can be used to estimate bead purity.²³¹

Problem: Our first generation of beads had barcode oligos linked to their hydrogel matrix using a cleavable disulfide linker. We hypothesized that diffusion of barcodes from the center of the polyacrylamide hydrogel bead into the droplet would be a slow and reaction rate-limiting process. To test diffusion, we measured the beads' total primer content before and after DTT incubation, using probe fluorescence as a proxy. A strong residual fluorescent signal after even 50 minutes of continuous incubation with 50 mM DTT confirmed that much of the initial primer population was retained within the bead. Since DTT has a molecular weight comparable to a single nucleotide, and has good water solubility, we did not expect its diffusion to be the rate-limiting factor. Of course, 10x Genomics' beads were dissolvable, and we reasoned that this contributed to their assay's increased sensitivity.

Solution: We included N,N'-Bis(acryloyl)cystamine (BAC) in the hydrogel polymerization reaction, replacing classic bisacrylamide and making the gel polymer's crosslinks reversibly cleavable under reducing conditions with DTT. These beads dissolved in under 10 minutes of exposure to 10 mM DTT at room temperature.

²³¹Lareau, Nat Comm, 2020.

Problem: Cell-plus-bead-in-oil emulsions did not survive PCR thermocycling, which was necessary for amplifying accessible genome fragments for our scATAC-seq efforts. During PCR, microdroplets coalesced into large drops, meaning that bulk rather than single-cell libraries were be generated.

Solution: We tested various changes to the protocol, including different PCR master mixes (NEB, KAPA, Terra, Q5), PCR programs (number of cycles, temperatures) and density regulators (PEG-8000, Iodoxanol), but this issue was ultimately solved as a welcome side-effect of making the beads dissolvable. We then realised that the thermocycling induced convective flows in the emulsions, leading to bead collisions which broke the emulsion. When the beads were dissolved within the droplets, this effect disappeared and emulsions remained stable throughout PCR cycling.

Milestone: At this point, we were able to sequence cell barcodes, align them to a whitelist, align the genomic reads to a reference (both transcriptomic and genomic), and keep stable emulsions throughout the process. Our scRNA-seq data was 3' transcriptome enriched, and our scATAC-seq data aggregated into peaks and exhibited typical ATAC-seq fragment length distribution, with nucleosome and helical pitch. However, data in both our scATAC-seq and scRNA-seq libraries did not aggregate into pure cells in mixed-species experiments: all "cells" contained a mixture of human and mouse reads. In addition, rather than sharp cutoffs between cell barcodes and low-signal noise barcodes, we observed a gradual transition from cells to noise. This would prove to be the most challenging issue during the entire HyDrop research campaign.

Solution attempt 1: We reasoned that the poor separation between true cells and background noise barcodes, and the observed cross-species contamination both practically meant that any cell barcode could associate with any cell during our assay. One hypothesis was that somewhere along the process, unused cell barcodes were associating with un-cell-barcoded (or potentially even cell-barcoded) DNA/RNA. In the droplet protocol, emulsion reverse transcription (scRNA) or emulsion PCR (scATAC) was followed by emulsion breaking and a bead cleanup to remove primers before subsequent PCR. We decided to run the pre-PCR library on Agilent BioAnalyzer, both before and after bead cleanup, and immediately observed very strong primer peaks that exactly matched the size of our expected cell barcodes released by the beads. This meant that after emulsion breaking, unused barcodes were free to associate randomly with un-barcoded DNA/RNA present, that bead cleanup was not sufficient, and that a large amount of free barcodes were passed into the amplification PCR. Here, these free barcodes could randomly capture un-tagged cDNA or ATAC fragments, or even out-compete PCR primers annealing to correctly barcoded DNA molecules.

We thus introduced a second magnetic purification after emulsion breaking, which strongly reduced free cell barcode contamination, but only mildly improved the species mixing we observed in our libraries. For scATAC-seq, we also tried using exponential in-droplet PCR by adding a reverse primer to the reaction, hoping to improve barcode consumption at the risk of introducing PCR biases, but this did also not improve purity.

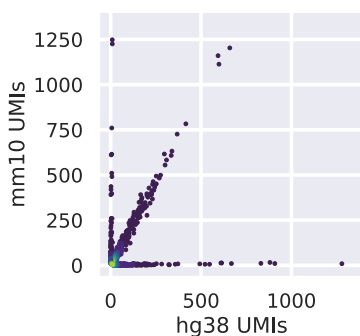


FIGURE. A.5. Contaminated mixed-species count plot. Adapted from [332].

Solution attempt 2: Thus far, we had not yet confirmed that our PCR-produced beads were pure: cross-contamination could also occur during bead production, leading to mixed-barcode beads rather than clonal beads. These contaminated beads would generate randomly barcoded fragments, explaining the observed noise downstream. This idea was strongly supported by evidence in Sanger sequencing, where we observe that the first sub-barcode locus was very pure, but the second was not, and the third was more pure again. This could mean that throughout barcoding, there is cross-contamination of the individual wells, leading to a mixture of primers from round one reaching the second barcoding stage and causing secondary barcoding contamination there. In order to test this, we performed a pilot run with two rounds (384^2) of barcoding rather than three (96^3). This change improved species purity in our species-mixing experiments, but still not to usable levels, and did not convince us to stick with this approach.

Solution attempt 3: At this point in time, we were very close to giving up. It had been more than a year

since we began tackling the contamination problem, and, the diagonal we observed on the species-mixing plots began resembling a large middle finger (fig. A.5).

Finally, we reduced the acrydite primer load during bead generation from 50 μM to 1.5, 6.25, and 12.5 μM , drastically reducing the total number of available sites for barcoding. Beads made with the 6 and 12.5 μM primer concentrations immediately produced species-pure scRNA-seq, and later scATAC-seq libraries, whereas the 1.5 μM condition did not produce detectable cDNA.

Most likely, this change affected purity in two ways: first, the polymerase reaction is now more likely to saturate the number of available barcode sites, reducing the number of available "previous-round" priming sites in the next round. Second, a lower total primer load released into the droplet is more likely to be purified out completely. Later research (post-publication) also indicated that washing bead wells individually during barcoding rather than in a pooled manner further increased bead purity and thus sensitivity. Individual dilutive washing of the sub-barcoding primers before pooling prevents these barcodes from annealing with remaining un-templated sites from the previous round of barcoding, improving barcode purity.

In parallel we had also implemented a number of small novelties. For example, we added Pitstop 2, a selective clathrin-mediated endocytosis inhibitor which also permeabilizes the nuclear membrane by destabilizing nucleoporin complexes,³³⁵ to the tagmentation reaction. This change slightly increased total yield and peak fidelity. We also added guanidinium isothiocyanate, a strong chaotrope known to denature proteins and solubilize nucleic acids, to the emulsion after PCR, and this further improved library yield. Next, we titrated several different sources of Tn5 and found no discernible difference in performance or fragment biases between commercial (Illumina, diagenode) and open-source Tn5.¹⁵⁷ Since Nextera Tn5 comprises 46% of the droplet reaction cost, savings made by using in-house Tn5 are considerable. In addition, we tested a number of optimizations for HyDrop-RNA. Exonuclease I (Exo I) has been used in single-cell RNA-seq workflows to remove excess primers and barcodes that could lead to non-specific priming.¹⁷⁵ Addition or exclusion of Exo I did not appear to make a difference in sensitivity. Introduction of two riboguanosines in the template-switching oligo, which improve thermal annealing stability to 3' CCC motif of the untemplated cDNA, also did not appear to increase sensitivity.¹⁵⁸ Second-strand synthesis through Bst2.0 or Klenow fragment showed improvements,³³⁶ but the best cDNA amplification strategy proved to be PCR thermocycling supplemented with GTP.¹⁶⁰

A.3 Key results

After the above troubleshooting, HyDrop achieved functional scRNA-seq and scATAC-seq: mixed-species experiments returned pure ($\sim 90\%$) single-cell profiles, and we were able to assign cell types to our first experiments on mixed-cancer cell lines (MCF-7 and PC3). We applied HyDrop-ATAC to flash-frozen mouse cortex, generating 7996 high-quality single-nucleus ATAC profiles in one run. We also applied HyDrop-RNA to the same cohort of mouse cortex samples, and to FAC-sorted *Drosophila* neurons, capturing 7996 and 973 cells respectively. The latter was particularly exciting, since this data contributed directly to ongoing research at the time by profiling rare cells not observed in 10x scATAC-seq data on the same sample.²⁵⁹ HyDrop-RNA achieved median 3389 UMIs and 1658 genes per cell in mouse cortex. HyDrop-ATAC recovered 4148 median fragments per cell, with high TSS enrichment scores. These sensitivities were sufficient for confident cell-type annotation in both modalities, revealing 19 and 30 distinct cell types in the scATAC and scRNA-seq dataset respectively.

³³⁵Mulqueen, *bioRxiv*, 2020.

¹⁵⁷Picelli, *Gen Res*, 2014.

¹⁷⁵Macosko, *Cell*, 2015.

¹⁵⁸Picelli, *Nat Prot*, 2014.

³³⁶Stickels, *Nat Biotech*, 2020.

¹⁶⁰Hagemann-Jensen, *Nat Biotech*, 2020.

²⁵⁹Janssens, *Nature*, 2022.



Systematic benchmarking of single-cell ATAC-seq protocols

Received: 18 January 2022

Accepted: 22 June 2023

Published online: 03 August 2023

Check for updates

Florian V. De Rop^{1,2}, Gert Hulselmans^{1,2}, Chris Flerin^{1,2}, Paula Soler-Vila³, Albert Rafels⁴, Valerie Christiaens^{1,2}, Carmen Bravo González-Blas^{1,2}, Domenica Marchese⁴, Ginevra Caratù⁴, Suresh Poovathingal¹, Orit Rozenblatt-Rosen⁵, Michael Slyper⁶, Wendy Luo⁶, Christoph Muus⁶, Fabiana Duarte^{7,8}, Rojesh Shrestha^{7,8}, S. Tansu Bagdatli⁹, M. Ryan Corces¹⁰, Lira Mamanova¹¹, Andrew Knights¹¹, Kerstin B. Meyer¹¹, Ryan Mulqueen¹², Akram Taherinasab^{13,14}, Patrick Maschmeyer^{15,16}, Jörn Pezoldt^{17,18}, Camille Lucie Germaine Lambert^{17,18}, Marta Iglesias^{4,19}, Sebastián R. Najle⁴, Zain Y. Dossani^{20,21}, Luciano G. Martelotto^{22,23}, Zach Burkett²⁴, Ronald Lebofsky²⁴, José Ignacio Martín-Subero^{3,25,26,27}, Satish Pillai^{20,21}, Arnau Sebé-Pedrós^{4,19,28}, Bart Deplancke^{17,18}, Sarah A. Teichmann^{11,29}, Leif S. Ludwig^{6,15,16}, Theodore P. Braun^{13,14}, Andrew C. Adey¹², William J. Greenleaf^{9,30}, Jason D. Buenrostro^{7,8}, Aviv Regev^{5,31,32}, Stein Aerts^{1,2,33} & Holger Heyn^{4,19,33}✉

Single-cell assay for transposase-accessible chromatin by sequencing (scATAC-seq) has emerged as a powerful tool for dissecting regulatory landscapes and cellular heterogeneity. However, an exploration of systemic biases among scATAC-seq technologies has remained absent. In this study, we benchmark the performance of eight scATAC-seq methods across 47 experiments using human peripheral blood mononuclear cells (PBMCs) as a reference sample and develop PUMATAC, a universal preprocessing pipeline, to handle the various sequencing data formats. Our analyses reveal significant differences in sequencing library complexity and tagmentation specificity, which impact cell-type annotation, genotype demultiplexing, peak calling, differential region accessibility and transcription factor motif enrichment. Our findings underscore the importance of sample extraction, method selection, data processing and total cost of experiments, offering valuable guidance for future research. Finally, our data and analysis pipeline encompasses 169,000 PBMC scATAC-seq profiles and a best practices code repository for scATAC-seq data analysis, which are freely available to extend this benchmarking effort to future protocols.

FIGURE. B.1. Article title page of "Systematic benchmarking of single-cell ATAC sequencing protocols", chapter 3

Appendix B

Supplementary note: Benchmarking single-cell chromatin accessibility sequencing methods

The following note will describe how we benchmarked several of the most prominent scATAC-seq techniques available in 2023. We performed 47 experiments across 8 different scATAC-seq methods to generate 169k single-cell chromatin accessibility profiles. Our benchmark quantified reproducible technical biases in not only the sensitivity, but also the specificity of these methods, produced a scATAC-seq pre-processing pipeline with best practices tutorials, and yielded the largest dataset of scATAC-seq in PBMCs to date. We published this work in Nature Biotechnology in 2023 (fig. B.1, chapter 3).³³⁷

B.1 Analysis

We included a number of different techniques in the initial phase of our benchmark:

- 10x Genomics scATAC-seq v1 and v1.1: The most widely used scATAC-seq method. While v1 reagent kits were technically still available for purchase, the microfluidic chip required for this kit had been de-commissioned due to patent infringement litigation by BioRad. An updated version (v1.1) had been released together with the nextGEM chip, which circumvented the BioRad intellectual property by generating droplets in a reservoir deposition style instead of using flow focusing. During the benchmark, 10x Genomics released their v2 kit, which was also included.
- 10x Multiome ATAC + Gene Expression: Introduced in 2020, the 10x multiome kit generated both scRNA-seq and scATAC-seq data from the same cells.
- Bio-Rad ddSEQ SureCell ATAC: 10x's direct competitor. The Bio-Rad scATAC-seq method also used droplet microfluidics, but employed hard resin beads (like Drop-seq) and an ingenious overloading strategy to reach cell capture rates on par with 10x.¹⁸⁹
- mtscATAC-seq: A modification of the 10x scATAC-seq protocol where omission of Tween-20 and digitonin from nuclei extraction buffers keeps the cell membrane intact but permeabilized to Tn5. As a result, the cell's mitochondria are retained, and their genomes are also sequenced, allowing lineage tracing.
- s3-ATAC: This plate-based combinatorial indexing method uses a uracil-modified Tn5 adapter to circumvent homotypic-ended fragments, doubling fragment capture.¹⁹¹
- HyDrop: Similar to 10x scATAC-seq, but using commercially available reagents and open-source microfluidics.³³²

³³⁷De Rop, Nat Biotech, 2023.

¹⁹¹Mulqueen, Nat Biotech, 2021.

³³²De Rop, eLife, 2022.

¹⁸⁹Lareau, Nat Biotech, 2019.

Data originating from the different techniques strongly differed in sequencing format. Moreover, among the several method-specific data analysis pipelines, there was no consensus as to how this data should be pre-processed and aligned to the reference genome. We had faced a similar problem during the development of HyDrop: despite the structural similarities between 10x Genomics' and HyDrop's sequencing format, 10x's Cell Ranger software could not be used to pre-process HyDrop data since several internal parameters (such as read lengths and barcode whitelist) were hard-coded to match 10x data. To solve this problem, we developed PUMATAC, a Nextflow³³⁸ pre-processing pipeline capable of aligning sequencing data agnostic of sequencing format. PUMATAC takes raw sequencing data (.fastq), trims adapters, performs barcode error correction using a user-specified whitelist, aligns genomic reads to a reference genome, and converts these genome alignments to .bed-like .fragments files. The use of Nextflow provides several advantages:

1. Intelligent interruption procedures: per step in the pipeline, an md5 hash of the sample names, input files, output files and parameters is computed. If the workflow is interrupted and re-started, these hashes are re-computed. If no change in the hash is detected, it can be assumed that no run parameters have changed, and the pipeline can resume where it left off.
2. Management of parallel processing: samples and steps that are not dependent on each other can be scheduled independently from each other's completion status, facilitating co-processing of small and large datasets. Each step in the pipeline has definitions for maximum CPU and memory usage, and available resources are utilised maximally.
3. Built-in support for software containers such as Docker and singularity, improving environment-independent compatibility.

PUMATAC-generated fragments files were then counted in ENCODE SCREEN regions using pycisTopic.²⁴³ Cells were then filtered on fragment and TSS-enrichment cutoffs determined with Otsu's algorithm.³³⁹ Once cell counts were known for all samples, data was downsampled to matching read depths and re-processed to account for differences in sequencing depths. Cells were then doublet-filtered (scrublet)³⁴⁰ assigned to donors (freemuxlet)³⁴¹ and broad PBMC cell types (seurat)²³⁹. Cell-type specific peak sets were called (MACS2)³⁴² and aggregated, and the entire process was repeated once more on these consensus peak sets. Finally, differentially accessible regions were called between cell types and donors, and motif enrichment in these DARs was performed using cisTarget.³⁴³

B.2 Key results

We gathered a total of 47 individual scATAC-seq datasets across the 8 methods assessed. Each library was sequenced to sufficient depth (at least 42 000 reads/cell), data deposited centrally, and analysed with PUMATAC. We targeted a recovery of ~3000 cells for each dataset. The complete, filtered dataset spans 169 227 cells with counts in 198 421 consensus peaks and is freely available.

Standard scATAC-seq pre-processing steps discard sequenced reads for various reasons: no valid barcode detected (5-10%), poor genomic alignment (10-20%), not within a peak region (5-15%), not associated with a called cell (7-60%). In most cases, this last step of filtering cells from noise contributed greatest to losses. FAC sorting live cells before nuclei extraction reduced these losses from 32% to 7% in mtscATAC-seq experiments, suggesting that the issue is at least partially caused by contaminating DNA from damaged cells. For all techniques, the fraction of total reads remaining in the final cell by peak array was small, ranging between 28% for 10x v2 to below 4% for s3-ATAC and HyDrop v1. This "sequencing efficiency" directly scaled overall sensitivity (absolute median number of fragments in peaks detected per cell) at a given sequencing depth (fig. 2a), and is thus a key metric in experimental design as it relates information gained to resources spent. Finally, the sequencing depth

³³⁸Di Tommaso, Nat Biotech, 2017.

²⁴³Bravo González-Blas, Nat Meth, 2019.

³³⁹Otsu, IEEE Sys Man Cyb, 1979.

³⁴⁰Wolock, Cell Sys, 2019.

³⁴¹Zhang, GitHub, 2019.

²³⁹Stuart, Cell, 2019.

³⁴²Zhang, Gen Biol, 2008.

³⁴³Aibar, Nat Meth, 2017.

at which the number of unique fragments detected begins to saturate varied greatly across methods (extended data fig. 10), suggesting that technique-specific optima between cell scaling versus depth scaling may be needed.

Raw fragment sensitivity differences directly affected all downstream analyses: doublet detection, genotype demultiplication, cell type annotation and significance of DARs all positively depended on median number of unique fragments detected per cell (fig. 2f-j), and these effects became more pronounced when investigating subtle differences such as differential motif enrichment studies between donors (fig. 3j-k). Naively, most of these effects could be explained by statistical power scaling with sensitivity, but technologies also differed qualitatively: both fragment TSS enrichment scores and raw FRIP values varied strongly (fig. 1h). For example, Bio-Rad ddSEQ fragments were significantly more TSS-proximal than 10x fragments, which was also reflected in the TSS-proximity of ddSEQ-derived DARs (fig. 2j) and subsequent motif analyses (fig. 3g). In general, high-signal DARs were easily recovered by all techniques, but weak DARs were recovered in a more technique-dependent manner (fig. 3e-f), further implying technique-specific fragment capture bias. This bias was also evident in the difficulty at which samples integrated across technologies, even after computational batch-effect correction (*harmony*;³⁴⁴ fig. 4c-d). These technique-specific biases could be attributed to differences in Tn5 sources, concentrations, barcode purity, and amplification and extraction protocols across technologies.

Awareness of the sensitivity and specificity biases described here is especially important in scATAC-seq studies compared to scRNA-seq studies, since peak region sets are sample-dependent rather than fixed gene sets. Therefore, high-specificity methods should be prioritised when generating reference peak region sets if public references (e.g. ENCODE ChIP-seq or DNase-seq) are not available a priori. Cost-efficient but lower sensitive/specificity open-source methods can be used for data generation at scale. Sequencing efficiency of these open-source methods can be improved by sequencing less deeply, since fewer reads will be lost to duplicates, and available resources can be allocated to sampling more cells rather than sequencing more. This strategy will lead to lower per-cell sensitivity, but increasing the total number of cells may be able to partially compensate. For example, both cell type annotation quality (*seurat* label transfer scores) and DAR strength scaled with higher cell counts (fig. 3a). We hypothesize that the benefit of sampling additional cells diminishes when per-cell information becomes insufficient for the analysis of interest. Global optima for the trade-offs between cell counts, per-cell depth, and total experimental cost with respect to maximum predictive value remain an open problem, but this benchmark provides a strong indication of current performances, and a standardized evaluation framework for future datasets.

³⁴⁴Korsunsky, Nat Meth, 2019.

Bibliography

- [1] J. M. W. SLACK. Conrad Hal Waddington: The Last Renaissance Biologist? *Nature Reviews Genetics* **3**: 889–895, 2002. (see pp. 1, 2)
- [2] D. E. OLINS and A. L. OLINS. Chromatin History: Our View from the Bridge. *Nature Reviews Molecular Cell Biology* **4**: 809–814, 2003. (see p. 1)
- [3] M. COBB. Who Discovered Messenger RNA? *Current Biology* **25**: R526–R532, 2015. (see p. 1)
- [4] U. DEICHMANN. Chromatin: Its History, Current Research, and the Seminal Researchers and Their Philosophy. *Perspectives in Biology and Medicine* **58**: 143–164, 2015. (see p. 1)
- [5] U. DEICHMANN. Epigenetics: The Origins and Evolution of a Fashionable Topic. *Developmental Biology* **416**: 249–254, 2016. (see p. 1)
- [6] J. T. LIS. A 50 Year History of Technologies That Drove Discovery in Eukaryotic Transcription Regulation. *Nature Structural & Molecular Biology* **26**: 777–782, 2019. (see p. 1)
- [7] R. G. ROEDER. 50+ Years of Eukaryotic Transcription: An Expanding Universe of Factors and Mechanisms. *Nature Structural & Molecular Biology* **26**: 783–791, 2019. (see p. 1)
- [8] R. HOOKE. *Micrographia*. Great Britain: The Royal Society, Jan. 1665. (see p. 1)
- [9] T. SCHWANN. *Mikroskopische Untersuchungen über Die Uebereinstimmung in Der Struktur Und Dem Wachsthum Der Thiere Und Pflanzen*. Berlin: Sander, 1839. (see p. 1)
- [10] V. RUDOLF. *Die Cellularpathologie in Ihrer Begründung Auf Physiologische Und Pathologische Gewebelehre*. Berlin: Verlag von August Hirschwald, 1858. (see p. 1)
- [11] W. FLEMMING. Beiträge zur Kenntniss der Zelle und ihrer Lebenserscheinungen. *Archiv für mikroskopische Anatomie* **16**: 302–436, 1879. (see p. 1)
- [12] T. H. MORGAN. Sex Limited Inheritance in *Drosophila*. *Science* **32**: 120–122, 1910. (see p. 2)
- [13] W. JOHANNSEN. *Elemente der exakten Erblchkeitslehre*. Deutsche wesentlich erweiterte Ausgabe in fünfundzwanzig Vorlesungen. German Public domain. Digitized by the MBLWHOI Library for the Biodiversity Heritage Library Jena: G. Fischer, 1909. (see p. 2)
- [14] A. H. STURTEVANT. The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *Journal of Experimental Zoology* **14**: 43–59, 1913. (see p. 2)
- [15] T. H. MORGAN. Random Segregation Versus Coupling in Mendelian Inheritance. *Science* **34**: 384–384, 1911. (see p. 2)
- [16] T. H. MORGAN, A. H. STURTEVANT, and H. J. MULLER. *The Mechanism of Mendelian Heredity*. Harvard University, 1915. 276 pp. (see p. 2)
- [17] H. J. MULLER. Artificial Transmutation of the Gene. *Science* **66**: 84–87, 1927. (see p. 2)
- [18] H. J. MULLER. Types of visible variations induced by X-rays in *Drosophila*. *Journal of Genetics* **22**: 299–334, 1930. (see pp. 2, 12)
- [19] A. TAGARELLI, A. PIRO, P. LAGONIA, and G. TAGARELLI. Hans Spemann. One Hundred Years Before the Birth of Experimental Embryology. *Anatomia, Histologia, Embryologia* **33**: 28–32, 2004. (see p. 2)

- [20] H. SPEMANN and H. MANGOLD. über Induktion von Embryonenanlagen Durch Implantation Artfremder Organisatoren. *Arch. Mikrosk. Anat. Entw. Mech.* 599–638, 1924. (see p. 2)
- [21] C. H. WADDINGTON. The Genetic Control of Wing Development in *Drosophila*. *Journal of Genetics* **41**: 75–113, 1940. (see p. 2)
- [22] C. H. WADDINGTON. *Organisers and Genes*. Cambridge: The University Press, 1940. (see p. 2)
- [23] C. H. WADDINGTON. *The Strategy of the Genes*. London: George Allen and Unwin, 1957. (see pp. 2, 3)
- [24] F. GRIFFITH. The Significance of Pneumococcal Types. *The Journal of Hygiene* **27**: 113–159, 1928. (see p. 3)
- [25] O. T. AVERY, C. M. MACLEOD, and M. MCCARTY. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types. Induction of Transformation by a Desoxyribosenucleic Acid Fraction Isolated from *Pneumococcus* Type III. *Journal of Experimental Medicine* **79**: 137–158, 1944. (see p. 3)
- [26] L. PAULING, H. A. ITANO, S. J. SINGER, and I. C. WELLS. Sickle Cell Anemia, a Molecular Disease. *Science* **110**: 543–548, 1949. (see p. 3)
- [27] J. WATSON and F. CRICK. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**: 737–738, 1953. (see p. 3)
- [28] J. WATSON and F. CRICK. Genetical Implications of the Structure of Deoxyribonucleic Acid. *Nature* **171**: 964–967, 1953. (see p. 3)
- [29] F. JACOB and J. MONOD. Genetic Regulatory Mechanisms in the Synthesis of Proteins. *Journal of Molecular Biology* **3**: 318–356, 1961. (see p. 4)
- [30] S. BRENNER, F. JACOB, and M. MESELSOHN. An Unstable Intermediate Carrying Information from Genes to Ribosomes for Protein Synthesis. *Nature* **190**: 576–581, 1961. (see p. 4)
- [31] M. W. NIRENBERG and J. H. MATTHAEI. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proceedings of the National Academy of Sciences of the United States of America* **47**: 1588–1602, 1961. (see p. 4)
- [32] H. G. KHORANA, H. BÜUCHI, H. GHOSH, N. GUPTA, T. M. JACOB, H. KÖSSEL, R. MORGAN, S. A. NARANG, E. OHTSUKA, and R. D. WELLS. Polynucleotide Synthesis and the Genetic Code. *Cold Spring Harbor Symposia on Quantitative Biology* **31**: 39–49, 1966. (see p. 4)
- [33] W. GILBERT and B. MÜLLER-HILL. Isolation of the lac repressor. *Proceedings of the National Academy of Sciences of the United States of America* **56**: 1891–1898, 1966. (see p. 4)
- [34] W. GILBERT and B. MÜLLER-HILL. The lac operator is DNA. *Proceedings of the National Academy of Sciences of the United States of America* **58**: 2415–2421, 1967. (see p. 4)
- [35] W. M. JOU, G. HAEGEMAN, M. YSEBAERT, and W. FIERS. Nucleotide Sequence of the Gene Coding for the Bacteriophage MS2 Coat Protein. *Nature* **237**: 82–88, 1972. (see p. 4)
- [36] J. VANDEKERCKHOVE and M. VAN MONTAGU. The Use of Fluorescamine in the Sequence Analysis of the Bacteriophage MS2 Structural Proteins, Purified at the Nanomole Scale. *Archives Internationales De Physiologie Et De Biochimie* **81**: 984, 1973. (see p. 5)
- [37] W. GILBERT and A. MAXAM. The nucleotide sequence of the lac operator. *Proceedings of the National Academy of Sciences of the United States of America* **70**: 3581–3584, 1973. (see p. 5)
- [38] R. DE WACHTER and W. FIERS. Preparative two-dimensional polyacrylamide gel electrophoresis of 32 P-labeled RNA. *Analytical Biochemistry* **49**: 184–197, 1972. (see p. 5)
- [39] A. M. MAXAM and W. GILBERT. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America* **74**: 560–564, 1977. (see p. 5)
- [40] F. SANGER, S. NICKLEN, and A. R. COULSON. DNA Sequencing with Chain-Terminating Inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**: 5463–5467, 1977. (see p. 5)

- [41] F. SANGER, G. M. AIR, B. G. BARRELL, N. L. BROWN, A. R. COULSON, J. C. FIDDES, C. A. HUTCHISON III, P. M. SLOCOMBE, and M. SMITH. Nucleotide sequence of bacteriophage ϕ X174 DNA. *Nature* **265**: 687–695, 1977. (see p. 5)
- [42] E. STEDMAN and E. STEDMAN. The Basic Proteins of Cell Nuclei. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **235**: 565–595, 1951. (see p. 5)
- [43] R. C. HUANG and J. BONNER. Histone, a suppressor of chromosomal RNA synthesis. *Proceedings of the National Academy of Sciences of the United States of America* **48**: 1216–1222, 1962. (see p. 5)
- [44] V. G. ALLFREY, R. FAULKNER, and A. E. MIRSKY. Acetylation and Methylation of Histones and Their Possible Role in the Regulation of RNA Synthesis. *Proceedings of the National Academy of Sciences of the United States of America* **51**: 786–794, 1964. (see p. 5)
- [45] R. J. DELANGE and E. L. SMITH. Structures of histones. *Accounts of Chemical Research* **5**: 368–373, 1972. (see p. 5)
- [46] J. F. PARDON and M. H. WILKINS. A super-coil model for nucleohistone. *Journal of Molecular Biology* **68**: 115–124, 1972. (see p. 5)
- [47] H. S. SLAYTER, T. Y. SHIH, A. J. ADLER, and G. D. FASMAN. Electron microscopy and circular dichroism studies on chromatin. *Biochemistry* **11**: 3044–3054, 1972. (see p. 6)
- [48] D. R. HEWISH and L. A. BURGOYNE. Chromatin sub-structure: The digestion of chromatin DNA at regularly spaced sites by a nuclear deoxyribonuclease. *Biochemical and Biophysical Research Communications* **52**: 504–510, 1973. (see p. 6)
- [49] A. L. OLINS and D. E. OLINS. Spheroid chromatin units (v bodies). *Science* **183**: 330–332, 1974. (see p. 6)
- [50] R. D. KORNBERG and J. O. THOMAS. Chromatin Structure; Oligomers of the Histones. *Science (New York, N.Y.)* **184**: 865–868, 1974. (see p. 6)
- [51] R. D. KORNBERG. Chromatin Structure: A Repeating Unit of Histones and DNA. *Science* **184**: 868–871, 1974. (see p. 6)
- [52] J. T. FINCH and A. KLUG. Solenoidal model for superstructure in chromatin. *Proceedings of the National Academy of Sciences of the United States of America* **73**: 1897–1901, 1976. (see p. 6)
- [53] H. WEINTRAUB and M. GROUDINE. Chromosomal subunits in active genes have an altered conformation. *Science* **193**: 848–856, 1976. (see pp. 6, 11)
- [54] R. G. ROEDER and W. J. RUTTER. Multiple Forms of DNA-dependent RNA Polymerase in Eukaryotic Organisms. *Nature* **224**: 234–237, 1969. (see p. 6)
- [55] S. Y. NG, C. S. PARKER, and R. G. ROEDER. Transcription of Cloned *Xenopus* 5S RNA Genes by *X. laevis* RNA Polymerase III in Reconstituted Systems. *Proceedings of the National Academy of Sciences* **76**: 136–140, 1979. (see p. 6)
- [56] T. MATSUI, J. SEGALL, and R. G. ROEDER. Multiple Factors Are Required for the Accurate Transcription of Purified Genes by RNA Polymerase III. *Journal of Biological Chemistry* **255**: 11986–11991, 1980. (see p. 6)
- [57] D. R. ENGELKE, S.-Y. NG, B. S. SHASTRY, and R. G. ROEDER. Specific Interaction of a Purified Transcription Factor with an Internal Control Region of 5S RNA Genes. *Cell* **19**: 717–728, 1980. (see p. 6)
- [58] F. PAYVAR, O. WRANGE, J. CARLSTEDT-DUKE, S. OKRET, J. A. GUSTAFSSON, and K. R. YAMAMOTO. Purified Glucocorticoid Receptors Bind Selectively in Vitro to a Cloned DNA Fragment Whose Transcription Is Regulated by Glucocorticoids in Vivo. *Proceedings of the National Academy of Sciences of the United States of America* **78**: 6628–6632, 1981. (see p. 6)
- [59] W. S. DYNAN and R. TJIAN. Isolation of Transcription Factors That Discriminate between Different Promoters Recognized by RNA Polymerase II. *Cell* **32**: 669–680, 1983. (see p. 6)

- [60] R. J. BRAM and R. D. KORNBERG. Specific Protein Binding to Far Upstream Activating Sequences in Polymerase II Promoters. *Proceedings of the National Academy of Sciences* **82**: 43–47, 1985. (see p. 6)
- [61] R. L. DAVIS, H. WEINTRAUB, and A. B. LASSAR. Expression of a Single Transfected cDNA Converts Fibroblasts to Myoblasts. *Cell* **51**: 987–1000, 1987. (see p. 6)
- [62] K. TAKAHASHI and S. YAMANAKA. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126**: 663–676, 2006. (see p. 6)
- [63] K. W. Y. CHO, B. BLUMBERG, H. STEINBEISSER, and E. M. DE ROBERTIS. Molecular Nature of Spemann's Organizer: The Role of the *Xenopus* Homeobox Gene *Goosecoid*. *Cell* **67**: 1111–1120, 1991. (see p. 6)
- [64] T. M. LAMB, A. K. KNECHT, W. C. SMITH, S. E. STACHEL, A. N. ECONOMIDES, N. STAHL, G. D. YANCOPOLOUS, and R. M. HARLAND. Neural Induction by the Secreted Polypeptide Noggin. *Science* **262**: 713–718, 1993. (see p. 6)
- [65] J. BANERJI, S. RUSCONI, and W. SCHAFFNER. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**: 299–308, 1981. (see p. 7)
- [66] J. BANERJI, L. OLSON, and W. SCHAFFNER. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* **33**: 729–740, 1983. (see p. 7)
- [67] C. O. PABO and R. T. SAUER. Protein-DNA Recognition. *Annual Review of Biophysics* **53**: 293–321, 1984. (see p. 7)
- [68] W. MCGINNIS, R. L. GARBER, J. WIRZ, A. KUROIWA, and W. J. GEHRING. A homologous protein-coding sequence in *Drosophila* homeotic genes and its conservation in other metazoans. *Cell* **37**: 403–408, 1984. (see p. 8)
- [69] C. R. KISSINGER, B. S. LIU, E. MARTIN-BLANCO, T. B. KORNBERG, and C. O. PABO. Crystal structure of an engrailed homeodomain-DNA complex at 2.8 Å resolution: a framework for understanding homeodomain-DNA interactions. *Cell* **63**: 579–590, 1990. (see p. 8)
- [70] J. MILLER, A. D. MCLACHLAN, and A. KLUG. Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus* oocytes. *EMBO Journal* **4**: 1609–1614, 1985. (see p. 8)
- [71] N. P. PAVLETICH and C. O. PABO. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. *Science* **252**: 809–817, 1991. (see p. 8)
- [72] W. H. LANDSCHULZ, P. F. JOHNSON, and S. L. MCKNIGHT. The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins. *Science* **240**: 1759–1764, 1988. (see p. 8)
- [73] C. MURRE, P. S. MCCAW, H. VAESSIN, M. CAUDY, L. Y. JAN, Y. N. JAN, C. V. CABRERA, J. N. BUSKIN, S. D. HAUSCHKA, and A. B. LASSAR. Interactions between heterologous helix-loop-helix proteins generate complexes that bind specifically to a common DNA sequence. *Cell* **58**: 537–544, 1989. (see p. 8)
- [74] T. ELLENBERGER, D. FASS, M. ARNAUD, and S. C. HARRISON. Crystal structure of transcription factor E47: E-box recognition by a basic region helix-loop-helix dimer. *Genes & Development* **8**: 970–980, 1994. (see p. 8)
- [75] M. LEWIS, G. CHANG, N. C. HORTON, M. A. KERCHER, H. C. PACE, M. A. SCHUMACHER, R. G. BRENNAN, and P. LU. Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science* **271**: 1247–1254, 1996. (see p. 8)
- [76] L. C. TSUI, M. BUCHWALD, D. BARKER, J. C. BRAMAN, R. KNOWLTON, J. W. SCHUMM, H. EIBERG, J. MOHR, D. KENNEDY, N. PLAVSIC, and ET AL. Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. *Science* **230**: 1054–1057, 1985. (see p. 8)
- [77] D. BOTSTEIN, R. L. WHITE, M. SKOLNICK, and R. W. DAVIS. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* **32**: 314–331, 1980. (see p. 8)

- [78] J. R. RIORDAN, J. M. ROMMENS, B.-S. KEREM, N. ALON, R. ROZMAHEL, Z. GRZELCZAK, J. ZIELENSKI, S. LOK, N. PLAVSIC, J.-L. CHOU, M. L. DRUMM, M. C. IANNUZZI, F. S. COLLINS, and L.-C. TSUI. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245**: Erratum in: *Science* 1989 Sep 29;245(4925):1437, 1066–1073, 1989. (see p. 8)
- [79] J. M. ROMMENS, M. C. IANNUZZI, B. KEREM, M. L. DRUMM, G. MELMER, M. DEAN, R. ROZMAHEL, J. L. COLE, D. KENNEDY, N. HIDAKA, and ET AL. Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* **245**: 1059–1065, 1989. (see p. 8)
- [80] J. F. GUSELLA, N. S. WEXLER, P. M. CONNEALLY, S. L. NAYLOR, M. A. ANDERSON, R. E. TANZI, P. C. WATKINS, K. OTTINA, M. R. WALLACE, A. Y. SAKAGUCHI, and ET AL. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**: 234–238, 1983. (see p. 8)
- [81] M. E. MACDONALD, C. M. AMBROSE, M. P. DUYAO, R. H. MYERS, C. LIN, L. SRINIDHI, G. BARNES, S. A. TAYLOR, M. JAMES, N. GROOT, H. MACFARLANE, B. JENKINS, M. A. ANDERSON, N. S. WEXLER, J. F. GUSELLA, G. P. BATES, S. BAXENDALE, H. HUMMERICH, S. KIRBY, M. NORTH, S. YOUNGMAN, R. MOTT, G. ZEHETNER, Z. SEDLACEK, A. POUSTKA, A.-M. FRISCHAUF, H. LEHRACH, A. J. BUCKLER, D. CHURCH, L. DOUCETTE-STAMM, M. C. O'DONOVAN, L. RIBA-RAMIREZ, M. SHAH, V. P. STANTON, S. A. STROBEL, K. M. DRATHS, J. L. WALES, P. DERVAN, D. E. HOUSMAN, M. ALTHERR, R. SHIANG, L. THOMPSON, T. FIELDER, J. J. WASMUTH, D. TAGLE, J. VALDES, L. ELMER, M. ALLARD, L. CASTILLA, M. SWAROOP, K. BLANCHARD, F. S. COLLINS, R. SNELL, T. HOLLOWAY, K. GILLESPIE, N. DATSON, D. SHAW, and P. S. HARPER. A Novel Gene Containing a Trinucleotide Repeat That Is Expanded and Unstable on Huntington's Disease Chromosomes. *Cell* **72**: 971–983, 1993. (see p. 8)
- [82] F. S. COLLINS. Shattuck lecture—medical and societal consequences of the Human Genome Project. *New England Journal of Medicine* **341**: 28–37, 1999. (see p. 8)
- [83] E. BIRNEY. The International Human Genome Project. *Human Molecular Genetics* **30**: R161–R163, 2021. (see p. 8)
- [84] R. K. SAIKI, S. SCHARF, F. FALOONA, K. B. MULLIS, G. T. HORN, H. A. ERLICH, and N. ARNHEIM. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**: 1350–1354, 1985. (see p. 8)
- [85] K. MULLIS, F. FALOONA, S. SCHARF, R. SAIKI, G. HORN, and H. ERLICH. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor Symposia on Quantitative Biology* **51**: 263–273, 1986. (see p. 8)
- [86] R. D. FLEISCHMANN, M. D. ADAMS, O. WHITE, R. A. CLAYTON, E. F. KIRKNESS, A. R. KERLAVAGE, C. J. BULT, J. F. TOMB, B. A. DOUGHERTY, and J. M. MERRICK. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**: 496–512, 1995. (see p. 9)
- [87] J. C. VENTER, M. D. ADAMS, G. G. SUTTON, A. R. KERLAVAGE, H. O. SMITH, and M. HUNKAPILLER. Shotgun sequencing of the human genome. *Science* **280**: 1540–1542, 1998. (see p. 9)
- [88] M. D. ADAMS. The genome sequence of *Drosophila melanogaster*. *Science* **287**: 2185–2195, 2000. (see p. 9)
- [89] E. S. LANDER. Initial Sequencing and Analysis of the Human Genome. *Nature* **409**: 860–921, 2001. (see p. 9)
- [90] J. C. VENTER. The sequence of the human genome. *Science* **291**: Erratum in: *Science* 2001 Jun 5;292(5523):1838, 1304–1351, 2001. (see p. 9)
- [91] J. KLING. Where the future went. In response to market trends and patenting laws, genomics companies are adapting their strategies. *EMBO Reports* **6**: 1012–1014, 2005. (see p. 9)
- [92] S. TRIPP and M. GRUEBER. Economic Impact of the Human Genome Project tech. rep. Prepared for United States Department of Energy Office of Biological and Environmental Research Battelle Memorial Institute, May 2011 (see p. 9)

- [93] M. MARGULIES, M. EGHOLM, W. E. ALTMAN, S. ATTIYA, J. S. BADER, L. A. BEMBEN, J. BERKA, M. S. BRAVERMAN, Y. J. CHEN, Z. CHEN, S. B. DEWELL, L. DU, J. M. FIERRO, X. V. GOMES, B. C. GODWIN, W. HE, S. HELGESEN, C. H. HO, G. P. IRZYK, S. C. JANDO, M. L. ALLENQUER, T. P. JARVIE, K. B. JIRAGE, J. B. KIM, J. R. KNIGHT, J. R. LANZA, J. H. LEAMON, S. M. LEFKOWITZ, M. LEI, J. LI, K. L. LOHMAN, H. LU, V. B. MAKHIJANI, K. E. MCDADE, M. P. MCKENNA, E. W. MYERS, E. NICKERSON, J. R. NOBILE, R. PLANT, B. P. PUC, M. T. RONAN, G. T. ROTH, G. J. SARKIS, J. F. SIMONS, J. W. SIMPSON, M. SRINIVASAN, K. R. TARTARO, A. TOMASZ, K. A. VOGT, G. A. VOLKMER, S. H. WANG, Y. WANG, M. P. WEINER, P. YU, R. F. BEGLEY, and J. M. ROTHBERG. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: Epub 2005 Jul 31. Erratum in: *Nature*. 2006 May 4;441(7089):120. Ho, Chun He [corrected to Ho, Chun Heen]., 376–380, 2005. (see p. 9)
- [94] S. T. BENNETT, C. BARNES, A. COX, L. DAVIES, and C. BROWN. Toward the 1,000 dollars human genome. *Pharmacogenomics* **6**: 373–382, 2005. (see p. 9)
- [95] D. R. BENTLEY, S. BALASUBRAMANIAN, H. P. SWERDLOW, Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53–59, 2008. (see pp. 9, 10)
- [96] J. SHENDURE and E. LIEBERMAN AIDEN. The Expanding Scope of DNA Sequencing. *Nature Biotechnology* **30**: 1084–1094, 2012. (see p. 9)
- [97] J. SHENDURE, G. J. PORRECA, N. B. REPPAS, X. LIN, J. P. MCCUTCHEON, A. M. ROSENBAUM, M. D. WANG, K. ZHANG, R. D. MITRA, and G. M. CHURCH. Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Science* **309**: 1728–1732, 2005. (see p. 9)
- [98] M. SCHENA, D. SHALON, R. W. DAVIS, and P. O. BROWN. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science* **270**: 467–470, 1995. (see p. 11)
- [99] D. G. WANG, J. B. FAN, C. J. SIAO, A. BERNO, P. YOUNG, R. SAPOLSKY, G. GHANDOUR, N. PERKINS, E. WINCHESTER, J. SPENCER, L. KRUGLYAK, L. STEIN, L. HSIE, T. TOPALOGLOU, E. HUBBELL, E. ROBINSON, M. MITTMANN, M. S. MORRIS, N. SHEN, D. KILBURN, J. RIOUX, C. NUSBAUM, S. ROZEN, T. J. HUDSON, R. LIPSHUTZ, M. CHEE, and E. S. LANDER. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**: 1077–1082, 1998. (see p. 11)
- [100] B. REN, F. ROBERT, J. J. WYRICK, O. APARICIO, E. G. JENNINGS, I. SIMON, J. ZEITLINGER, J. SCHREIBER, N. HANNETT, E. KANIN, T. L. VOLKERT, C. J. WILSON, S. P. BELL, and R. A. YOUNG. Genome-wide location and function of DNA binding proteins. *Science* **290**: 2306–2309, 2000. (see p. 11)
- [101] ENCODE. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**: 799–816, 2007. (see p. 11)
- [102] A. BARSKI, S. CUDDAPAH, K. CUI, T.-Y. ROH, D. E. SCHONES, Z. WANG, G. WEI, I. CHEPELEV, and K. ZHAO. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129**: 823–837, 2007. (see p. 11)
- [103] T. S. MIKKELSEN, M. KU, D. B. JAFFE, B. ISSAC, E. LIEBERMAN, G. GIANNOUKOS, P. ALVAREZ, W. BROCKMAN, T. K. KIM, R. P. KOCH, W. LEE, E. MENDENHALL, A. O'DONOVAN, A. PRESSER, C. RUSS, X. XIE, A. MEISSNER, M. WERNIG, R. JAENISCH, C. NUSBAUM, E. S. LANDER, and B. E. BERNSTEIN. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**: Epub 2007 Jul 1, 553–560, 2007. (see p. 11)
- [104] D. S. JOHNSON, A. MORTAZAVI, R. M. MYERS, and B. WOLD. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science* **316**: 1497–1502, 2007. (see p. 11)
- [105] G. ROBERTSON, M. HIRST, M. BAINBRIDGE, M. BILENKY, Y. ZHAO, T. ZENG, G. EUSKIRCHEN, B. BERNIER, R. VARHOL, A. DELANEY, N. THIessen, O. L. GRIFFITH, A. HE, M. MARRA, M. SNYDER, and S. JONES. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods* **4**: 651–657, 2007. (see p. 11)
- [106] C. WU. The 5' ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. *Nature* **286**: 854–860, 1980. (see p. 11)

- [107] G. E. CRAWFORD, I. E. HOLT, J. C. MULLIKIN, D. TAI, R. BLAKESLEY, G. BOUFFARD, A. YOUNG, C. MASIELLO, E. D. GREEN, T. G. WOLFSBERG, F. S. COLLINS, and N. I. O. H. I. S. CENTER. Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 992–997, 2004. (see p. 11)
- [108] G. E. CRAWFORD, S. DAVIS, P. C. SCACHERI, G. RENAUD, M. J. HALAWI, M. R. ERDOS, R. GREEN, P. S. MELTZER, T. G. WOLFSBERG, and F. S. COLLINS. DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nature Methods* **3**: 503–509, 2006. (see p. 11)
- [109] A. P. BOYLE, S. DAVIS, H. P. SHULHA, P. MELTZER, E. H. MARGULIES, Z. WENG, T. S. FUREY, and G. E. CRAWFORD. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**: 311–322, 2008. (see p. 11)
- [110] J. R. HESSELBERTH, X. CHEN, Z. ZHANG, P. J. SABO, R. SANDSTROM, A. P. REYNOLDS, R. E. THURMAN, S. NEPH, M. S. KUEHN, W. S. NOBLE, S. FIELDS, and J. A. STAMATOYANNOPOULOS. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature Methods* **6**: 283–289, 2009. (see p. 11)
- [111] S. JOHN, P. J. SABO, R. E. THURMAN, M. H. SUNG, S. C. BIDDIE, T. A. JOHNSON, G. L. HAGER, and J. A. STAMATOYANNOPOULOS. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature Genetics* **43**: 264–268, 2011. (see p. 11)
- [112] R. THURMAN, E. RYNES, R. HUMBERT, J. VIERSTRA, M. MAURANO, E. HAUGEN, N. SHEFFIELD, A. STERGACHIS, H. WANG, B. VERNOT, K. GARG, S. JOHN, R. SANDSTROM, D. BATES, L. BOATMAN, T. CANFIELD, M. DIEGEL, D. DUNN, A. EBERSOL, T. FRUM, E. GISTE, A. JOHNSON, E. JOHNSON, T. KUTYAVIN, B. LAJOIE, B. LEE, K. LEE, D. LONDON, D. LOTAKIS, S. NEPH, F. NERI, E. NGUYEN, H. QU, A. REYNOLDS, V. ROACH, A. SAFI, M. SANCHEZ, A. SANYAL, A. SHAFER, J. SIMON, L. SONG, S. VONG, M. WEAVER, Y. YAN, Z. ZHANG, B. LENHARD, M. TEWARI, M. DORSCHNER, R. HANSEN, P. NAVAS, G. STAMATOYANNOPOULOS, V. IYER, J. LIEB, S. SUNYAEV, J. AKEY, P. SABO, R. KAUL, T. FUREY, J. DEKKER, G. CRAWFORD, and J. STAMATOYANNOPOULOS. The accessible chromatin landscape of the human genome. *Nature* **489**: 75–82, 2012. (see p. 11)
- [113] ENCODE. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74, 2012. (see p. 11)
- [114] U. NAGALAKSHMI, Z. WANG, K. WAERN, C. SHOU, D. RAHA, M. GERSTEIN, and M. SNYDER. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**: Epub 2008 May 1, 1344–1349, 2008. (see p. 12)
- [115] A. MORTAZAVI, B. A. WILLIAMS, K. MCCUE, L. SCHAEFFER, and B. WOLD. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**: Epub 2008 May 30, 621–628, 2008. (see p. 12)
- [116] E. T. WANG, R. SANDBERG, S. LUO, I. KHREBTUKOVA, L. ZHANG, C. MAYR, S. F. KINGSMORE, G. P. SCHROTH, and C. B. BURGE. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470–476, 2008. (see p. 12)
- [117] J. C. MARIONI, C. E. MASON, S. M. MANE, M. STEPHENS, and Y. GILAD. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research* **18**: Epub 2008 Jun 11, 1509–1517, 2008. (see p. 12)
- [118] Z. WANG, M. GERSTEIN, and M. SNYDER. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**: 57–63, 2009. (see p. 12)
- [119] A. VERFAILLIE, H. IMRICHOVA, Z. K. ATAK, M. DEWAELE, F. RAMBOW, G. HULSELMANS, V. CHRISTIAENS, D. SVETLICHNYY, F. LUCIANI, L. VAN DEN MOOTER, S. CLAERHOUT, M. FIERIS, F. JOURNE, G. E. GHANEM, C. HERRMANN, G. HALDER, J. C. MARINE, and S. AERTS. Decoding the regulatory landscape of melanoma reveals TEADS as regulators of the invasive cell state. *Nature Communications* **6**: 6683, 2015. (see p. 12)
- [120] B. MCCLINTOCK. The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences of the United States of America* **36**: 344–355, 1950. (see p. 12)
- [121] S. RAVINDRAN. Barbara McClintock and the discovery of jumping genes. *Proceedings of the National Academy of Sciences of the United States of America* **109**: 20198–20199, 2012. (see p. 12)

- [122] J. A. SHAPIRO. Mutations caused by the insertion of genetic material into the galactose operon of *Escherichia coli*. *Journal of Molecular Biology* **40**: 93–105, 1969. (see p. 12)
- [123] D. E. BERG, J. DAVIES, B. ALLET, and J. D. ROCHAIX. Transposition of R factor genes to bacteriophage lambda. *Proceedings of the National Academy of Sciences of the United States of America* **72**: 3628–3632, 1975. (see p. 12)
- [124] A. I. BUKHARI, J. A. SHAPIRO, and S. L. ADHYA. Insertion Phenomena: DNA Insertion Elements, Plasmids, and Episomes. Papers from a meeting, Cold Spring Harbor, May 1976 Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory, 1977. xiv, 782 (see p. 12)
- [125] W. S. REZNIKOFF, S. J. ROTHSTEIN, R. A. JORGENSEN, R. C. JOHNSON, and J. C. P. YIN. “The structure of Tn5” in: *Molecular Biology, Pathogenicity, and Ecology of Bacterial Plasmids* ed. by S. LEVY, R. CLOWES, and H. KOENIG. New York: Plenum Press, 1981. 371–380 (see p. 12)
- [126] W. S. REZNIKOFF. Tn5 transposition and its regulation. *Cell* **31**: 307–308, 1982. (see p. 12)
- [127] G. B. RUVKUN, V. SUNDARESAN, and F. M. AUSUBEL. Directed transposon Tn5 mutagenesis and complementation analysis of *Rhizobium meliloti* symbiotic nitrogen fixation genes. *Cell* **29**: 551–559, 1982. (see p. 13)
- [128] D. R. DAVIES, I. Y. GORYSHIN, W. S. REZNIKOFF, and I. RAYMENT. Three-dimensional structure of the Tn5 synaptic complex transposition intermediate. *Science* **289**: 77–85, 2000. (see p. 13)
- [129] S. LOVELL, I. Y. GORYSHIN, W. R. REZNIKOFF, and I. RAYMENT. Two-metal active site binding of a Tn5 transposase synaptic complex. *Nature Structural Biology* **9**: 278–281, 2002. (see p. 13)
- [130] D. SEHNAL, S. BITTRICH, M. DESHPANDE, R. SVOBODOVÁ, K. BERKA, V. BAZGIER, S. VELANKAR, S. BURLEY, J. KOČA, and A. ROSE. Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Research* **49**: W431–W437, 2021. (see p. 13)
- [131] T. W. WIEGAND and W. S. REZNIKOFF. Characterization of two hypertransposing Tn5 mutants. *Journal of Bacteriology* **174**: 1229–1239, 1992. (see p. 13)
- [132] M. D. WEINREICH, A. GASCH, and W. S. REZNIKOFF. Evidence that the cis preference of the Tn5 transposase is caused by nonproductive multimerization. *Genes & Development* **8**: 2363–2374, 1994. (see p. 13)
- [133] I. Y. GORYSHIN and W. S. REZNIKOFF. Tn5 in Vitro Transposition. *Journal of Biological Chemistry* **273**: 7367–7374, 1998. (see p. 13)
- [134] M. ZHOU, A. BHASIN, and W. S. REZNIKOFF. Molecular genetic analysis of transposase-end DNA sequence recognition: cooperativity of three adjacent base-pairs in specific interaction with a mutant Tn5 transposase. *Journal of Molecular Biology* **276**: 913–925, 1998. (see p. 13)
- [135] M. STEINIGER-WHITE, A. BHASIN, S. LOVELL, I. RAYMENT, and W. S. REZNIKOFF. Evidence for “Unseen” Transposase–DNA Contacts. *Journal of Molecular Biology* **322**: 971–982, 2002. (see p. 13)
- [136] A. BHASIN, I. Y. GORYSHIN, and W. S. REZNIKOFF. Hairpin formation in Tn5 transposition. *Journal of Biological Chemistry* **274**: 37021–37029, 1999. (see p. 13)
- [137] W. S. REZNIKOFF. Tn5 as a model for understanding DNA transposition. *Molecular Microbiology* **47**: 1199–1206, 2003. (see p. 13)
- [138] W. S. REZNIKOFF, I. Y. GORYSHIN, and J. J. JENDRISAK. Tn5 as a Molecular Genetics Tool: In Vitro Transposition and the Coupling of in Vitro Technologies with in Vivo Transposition. *Methods in Molecular Biology* **260**: 83–96, 2004. (see p. 13)
- [139] A. ADEY, H. G. MORRISON, ASAN, X. XUN, J. O. KITZMAN, E. H. TURNER, B. STACKHOUSE, A. P. MACKENZIE, N. C. CARUCCIO, X. ZHANG, and J. SHENDURE. Rapid, Low-Input, Low-Bias Construction of Shotgun Fragment Libraries by High-Density in Vitro Transposition. *Genome Biology* **11**: R119, 2010. (see p. 13)
- [140] J. D. BUENROSTRO, P. G. GIRESI, L. C. ZABA, H. Y. CHANG, and W. J. GREENLEAF. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* **10**: 1213–1218, 2013. (see p. 13)

- [141] J. D. BUENROSTRO, B. WU, H. Y. CHANG, and W. J. GREENLEAF. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current Protocols in Molecular Biology* **109**: 21.29.1–21.29.9, 2015. (see p. 13)
- [142] M. R. CORCES, A. E. TREVINO, E. G. HAMILTON, P. G. GREENSIDE, N. A. SINNOTT-ARMSTRONG, S. VESUNA, A. T. SATPATHY, A. J. RUBIN, K. S. MONTINE, B. WU, A. KATHIRIA, S. W. CHO, M. R. MUMBACH, A. C. CARTER, M. KASOWSKI, L. A. ORLOFF, V. I. RISCA, A. KUNDAJE, P. A. KHAVARI, T. J. MONTINE, W. J. GREENLEAF, and H. Y. CHANG. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nature Methods* **14**: 959–962, 2017. (see p. 14)
- [143] E. SHAPIRO, T. BIEZUNER, and S. LINNARSSON. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics* **14**: Epub 2013 Jul 30, 618–630, 2013. (see p. 14)
- [144] J. EBERWINE, H. YEH, K. MIYASHIRO, Y. CAO, S. NAIR, R. FINNELL, M. ZETTEL, and P. COLEMAN. Analysis of gene expression in single live neurons. *Proceedings of the National Academy of Sciences of the United States of America* **89**: 3010–3014, 1992. (see p. 15)
- [145] F. KAMME, R. SALUNGA, J. YU, D. T. TRAN, J. ZHU, L. LUO, A. BITTNER, H. Q. GUO, N. MILLER, J. WAN, and M. ERLANDER. Single-cell microarray analysis in hippocampus CA1: demonstration and validation of cellular heterogeneity. *Journal of Neuroscience* **23**: 3607–3615, 2003. (see p. 15)
- [146] G. BRADY and N. N. ISCOVE. Construction of cDNA libraries from single cells. *Methods in Enzymology* **225**: 611–623, 1993. (see p. 15)
- [147] G. BRADY. Expression profiling of single mammalian cells—small is beautiful. *Yeast* **17**: 211–217, 2000. (see p. 15)
- [148] K. KURIMOTO, Y. YABUTA, Y. OHINATA, Y. ONO, K. D. UNO, R. G. YAMADA, H. R. UEDA, and M. SAITOU. An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. *Nucleic Acids Research* **34**: e42, 2006. (see p. 15)
- [149] F. TANG, C. BARBACIORU, Y. WANG, E. NORDMAN, C. LEE, N. XU, X. WANG, J. BODEAU, B. B. TUCH, A. SIDDIQUI, K. LAO, and M. A. SURANI. mRNA-Seq Whole-Transcriptome Analysis of a Single Cell. *Nature Methods* **6**: 377–382, 2009. (see p. 15)
- [150] S. ISLAM, U. KJÄLLQUIST, A. MOLINER, P. ZAJAC, J.-B. FAN, P. LÖNNERBERG, and S. LINNARSSON. Characterization of the Single-Cell Transcriptional Landscape by Highly Multiplex RNA-seq. *Genome Research* **21**: 1160–1167, 2011. (see pp. 15, 29)
- [151] S. ISLAM, U. KJÄLLQUIST, A. MOLINER, P. ZAJAC, J.-B. FAN, P. LÖNNERBERG, and S. LINNARSSON. Highly multiplexed and strand-specific single-cell RNA 5' end sequencing. *Nature Protocols* **7**: 813–828, 2012. (see p. 15)
- [152] Y. Y. ZHU, E. M. MACHLEDER, A. CHENCHIK, R. LI, and P. D. SIEBERT. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques* **30**: 892–897, 2001. (see p. 15)
- [153] T. KIVIOJA, A. VÄHÄRAUTIO, K. KARLSSON, M. BONKE, M. ENGE, S. LINNARSSON, and J. TAIPALE. Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods* **9**: 72–74, 2011. (see p. 15)
- [154] S. ISLAM, A. ZEISEL, S. JOOST, G. LA MANNO, P. ZAJAC, M. KASPER, P. LÖNNERBERG, and S. LINNARSSON. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods* **11**: Epub 2013 Dec 22, 163–166, 2014. (see p. 15)
- [155] D. RAMSKÖLD, S. LUO, Y.-C. WANG, R. LI, Q. DENG, O. R. FARIDANI, G. A. DANIELS, I. KHREBTUKOVA, J. F. LORING, L. C. LAURENT, G. P. SCHROTH, and R. SANDBERG. Full-Length mRNA-Seq from Single-Cell Levels of RNA and Individual Circulating Tumor Cells. *Nature Biotechnology* **30**: 777–782, 2012. (see pp. 16, 29)

- [156] S. PICELLI, Å. K. BJÖRKLUND, O. R. FARIDANI, S. SAGASSER, G. WINBERG, and R. SANDBERG. Smart-Seq2 for Sensitive Full-Length Transcriptome Profiling in Single Cells. *Nature Methods* **10**: 1096–1098, 2013. (see p. 16)
- [157] S. PICELLI, Å. K. BJÖRKLUND, B. REINIUS, S. SAGASSER, G. WINBERG, and R. SANDBERG. Tn5 Transposase and Tagmentation Procedures for Massively Scaled Sequencing Projects. *Genome Research* **24**: 2033–2040, 2014. (see pp. 16, 103)
- [158] S. PICELLI, O. R. FARIDANI, Å. K. BJÖRKLUND, and R. SANDBERG. Full-Length RNA-seq from Single Cells Using Smart-seq2. *Nature Protocols* **9**: 171–181, 2014. (see pp. 16, 103)
- [159] A. R. WU, N. F. NEFF, T. KALISKY, P. DALERBA, B. TREUTLEIN, M. E. ROTHENBERG, F. M. MBURU, G. L. MANTALAS, S. SIM, M. F. CLARKE, and S. R. QUAKE. Quantitative assessment of single-cell RNA-sequencing methods. *Nature Methods* **11**: Epub 2013 Oct 20, 41–46, 2014. (see p. 16)
- [160] M. HAGEMANN-JENSEN, C. ZIEGENHAIN, P. CHEN, D. RAMSKÖLD, G.-J. HENDRIKS, A. J. M. LARSSON, O. R. FARIDANI, and R. SANDBERG. Single-Cell RNA Counting at Allele and Isoform Resolution Using Smart-seq3. *Nature Biotechnology* **38**: 708–714, 2020. (see pp. 16, 93, 103)
- [161] T. HASHIMSHONY, F. WAGNER, N. SHER, and I. YANAI. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Reports* **2**: 666–673, 2012. (see pp. 16, 29)
- [162] T. HASHIMSHONY, N. SENDEROVICH, G. AVITAL, A. KLOCHENDLER, Y. DE LEEUW, L. ANAVY, D. GENNERT, S. LI, K. J. LIVAK, O. ROZENBLATT-ROSEN, Y. DOR, A. REGEV, and I. YANAI. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biology* **17**: 77, 2016. (see p. 16)
- [163] Y. SASAGAWA, I. NIKAIDO, T. HAYASHI, H. DANNO, K. D. UNO, T. IMAI, and H. R. UEDA. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biology* **14**: Erratum in: *Genome Biol.* 2017 Jan 18;18(1):9. doi: 10.1186/s13059-017-1154-x, R31, 2013. (see pp. 16, 29)
- [164] Y. SASAGAWA, H. DANNO, H. TAKADA, M. EBISAWA, K. TANAKA, T. HAYASHI, A. KURISAKI, and I. NIKAIDO. Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. *Genome Biology* **19**: 29, 2018. (see p. 16)
- [165] C. ZIEGENHAIN, B. VIETH, S. PAREKH, B. REINIUS, A. GUILLAUMET-ADKINS, M. SMETS, H. LEONHARDT, H. HEYN, I. HELLMANN, and W. ENARD. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell* **65**: 631–643.e4, 2017. (see p. 16)
- [166] E. MEREU, A. LAFZI, C. MOUTINHO, C. ZIEGENHAIN, D. J. MCCARTHY, A. ÁLVAREZ-VARELA, E. BATLLE, D. SAGAR, D. GRÜN, J. K. LAU, S. C. BOUTET, C. SANADA, A. OOI, R. C. JONES, K. KAIHARA, C. BRAMPTON, Y. TALAGA, Y. SASAGAWA, K. TANAKA, T. HAYASHI, C. BRAEUNING, C. FISCHER, S. SAUER, T. TREFZER, C. CONRAD, X. ADICONIS, L. T. NGUYEN, A. REGEV, J. Z. LEVIN, S. PAREKH, A. JANJIC, L. E. WANGE, J. W. BAGNOLI, W. ENARD, M. GUT, R. SANDBERG, I. NIKAIDO, I. GUT, O. STEGLE, and H. HEYN. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nature Biotechnology* **38**: Epub 2020 Apr 6, 747–755, 2020. (see p. 16)
- [167] NATURE. Method of the Year 2013. *Nature Methods* **11**: 1–1, 2014. (see p. 16)
- [168] A. P. PATEL, I. TIROSH, J. J. TROMBETTA, A. K. SHALEK, S. M. GILLESPIE, H. WAKIMOTO, D. P. CAHILL, B. V. NAHED, W. T. CURRY, R. L. MARTUZA, D. N. LOUIS, O. ROZENBLATT-ROSEN, M. L. SUVÀ, A. REGEV, and B. E. BERNSTEIN. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**: 1396–1401, 2014. (see p. 16)
- [169] S. BAO, Q. WU, R. E. MCLENDON, Y. HAO, Q. SHI, A. B. HJELMELAND, M. W. DEWHIRST, D. D. BIGNER, and J. N. RICH. Glioma stem cells promote radioresistance by preferential activation of the DNA damage response. *Nature* **444**: Epub 2006 Oct 18, 756–760, 2006. (see p. 16)
- [170] A. A. KOŁODZIEJCZYK, J. K. KIM, V. SVENSSON, J. C. MARIONI, and S. A. TEICHMANN. The technology and biology of single-cell RNA sequencing. *Molecular Cell* **58**: 610–620, 2015. (see p. 16)

- [171] D. A. JAITIN, E. KENIGSBERG, H. KEREN-SHAUL, N. ELEFANT, F. PAUL, I. ZARETSKY, A. MILDNER, N. COHEN, S. JUNG, A. TANAY, and I. AMIT. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**: 776–779, 2014. (see p. 16)
- [172] T. THORSEN, R. W. ROBERTS, F. H. ARNOLD, and S. R. QUAKE. Dynamic Pattern Formation in a Vesicle-Generating Microfluidic Device. *Physical Review Letters* **86**: 4163–4166, 2001. (see p. 17)
- [173] L. MAZUTIS, J. GILBERT, W. L. UNG, D. A. WEITZ, A. D. GRIFFITHS, and J. A. HEYMAN. Single-cell analysis and sorting using droplet-based microfluidics. *Nature Protocols* **8**: 870–891, 2013. (see p. 17)
- [174] M. T. GUO, A. ROTEM, J. A. HEYMAN, and D. A. WEITZ. Droplet Microfluidics for High-Throughput Biological Assays. *Lab on a Chip* **12**: 2146, 2012. (see p. 17)
- [175] E. Z. MACOSKO, A. BASU, R. SATIJA, J. NEMESH, K. SHEKHAR, M. GOLDMAN, I. TIROSH, A. R. BIALAS, N. KAMITAKI, E. M. MARTERSTECK, J. J. TROMBETTA, D. A. WEITZ, J. R. SANES, A. K. SHALEK, A. REGEV, and S. A. MCCARROLL. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**: 1202–1214, 2015. (see pp. 17, 25, 29, 93, 103)
- [176] A. M. KLEIN, L. MAZUTIS, I. AKARTUNA, N. TALLAPRAGADA, A. VERES, V. LI, L. PESHKIN, D. A. WEITZ, and M. W. KIRSCHNER. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161**: 1187–1201, 2015. (see pp. 17, 29, 93)
- [177] R. ZILIONIS, J. NAINYS, A. VERES, V. SAVOVA, D. ZEMMOUR, A. M. KLEIN, and L. MAZUTIS. Single-Cell Barcoding and Sequencing Using Droplet Microfluidics. *Nature Protocols* **12**: 44–73, 2017. (see p. 17)
- [178] T. M. GIERAHN, M. H. WADSWORTH, T. K. HUGHES, B. D. BRYSON, A. BUTLER, R. SATIJA, S. FORTUNE, J. C. LOVE, and A. K. SHALEK. Seq-Well: Portable, Low-Cost RNA Sequencing of Single Cells at High Throughput. *Nature Methods* **14**: 395–398, 2017. (see pp. 17, 29)
- [179] A. R. ABATE, C. H. CHEN, J. J. AGRESTI, and D. A. WEITZ. Beating Poisson encapsulation statistics using close-packed ordering. *Lab on a Chip* **9**: 2628–2631, 2009. (see p. 17)
- [180] G. ZHENG, J. TERRY, P. BELGRADER, P. RYVKIN, Z. BENT, R. WILSON, S. ZIRALDO, T. WHEELER, G. MCDERMOTT, J. ZHU, M. GREGORY, J. SHUGA, L. MONTESCLAROS, J. UNDERWOOD, D. MASQUELIER, S. NISHIMURA, M. SCHNALL-LEVIN, P. WYATT, C. HINDSON, R. BHARADWAJ, A. WONG, K. NESS, L. BEPPU, H. DEEG, C. MCFARLAND, K. LOEB, W. VALENTE, N. ERICSON, E. STEVENS, J. RADICH, T. MIKKELSEN, B. HINDSON, and J. BIELAS. Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8**: 14049, 2017. (see pp. 18, 29, 93)
- [181] D. A. CUSANOVICH, R. DAZA, A. ADEY, H. A. PLINER, L. CHRISTIANSEN, K. L. GUNDERSON, F. J. STEEMERS, C. TRAPNELL, and J. SHENDURE. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**: 910–914, 2015. (see p. 18)
- [182] J. CAO, J. S. PACKER, V. RAMANI, D. A. CUSANOVICH, C. HUYNH, R. DAZA, X. QIU, C. LEE, S. N. FURLAN, F. J. STEEMERS, A. ADEY, R. H. WATERSTON, C. TRAPNELL, and J. SHENDURE. Comprehensive Single-Cell Transcriptional Profiling of a Multicellular Organism. *Science* **357**: 661–667, 2017. (see pp. 18, 29)
- [183] B. K. MARTIN, C. QIU, E. NICHOLS, M. PHUNG, R. GREEN-GLADDEN, S. SRIVATSAN, R. BLECHER-GONEN, B. J. BELIVEAU, C. TRAPNELL, J. CAO, and J. SHENDURE. Optimized single-nucleus transcriptional profiling by combinatorial indexing. *Nature Protocols* **18**: Epub 2022 Oct 19, 188–207, 2023. (see p. 18)
- [184] A. B. ROSENBERG, C. M. ROCO, R. A. MUSCAT, A. KUCHINA, P. SAMPLE, Z. YAO, L. T. GRAYBUCK, D. J. PEELER, S. MUKHERJEE, W. CHEN, S. H. PUN, D. L. SELLERS, B. TASIC, and G. SEELIG. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**: 176–182, 2018. (see p. 18)
- [185] L. KUIJPERS, B. HORNUNG, M. C. G. N. VAN DEN HOUT-VAN VROONHOVEN, W. F. J. VAN IJCKEN, F. GROSVELD, and E. MULUGETA. Split Pool Ligation-based Single-cell Transcriptome sequencing (SPLiT-seq) data processing pipeline comparison. *BMC Genomics* **25**: 361, 2024. (see p. 18)
- [186] PARSE BIOSCIENCES. Parse Launches GigaLab to Scale Single Cell Sequencing to Billions of Cells Annually. *Press Release* Accessed: 2025-01-23, 2024. (see p. 18)

- [187] J. D. BUENROSTRO, B. WU, U. M. LITZENBURGER, D. RUFF, M. L. GONZALES, M. P. SNYDER, H. Y. CHANG, and W. J. GREENLEAF. Single-Cell Chromatin Accessibility Reveals Principles of Regulatory Variation. *Nature* **523**: 486–490, 2015. (see pp. 18, 29)
- [188] X. CHEN, R. J. MIRAGAIA, K. N. NATARAJAN, and S. A. TEICHMANN. A Rapid and Robust Method for Single Cell Chromatin Accessibility Profiling. *Nature Communications* **9**: 5345, 2018. (see pp. 18, 29, 30)
- [189] C. A. LAREAU, F. M. DUARTE, J. G. CHEW, V. K. KARTHA, Z. D. BURKETT, A. S. KOHLWAY, D. POKHOLOK, M. J. ARYEE, F. J. STEEMERS, R. LEBOSKY, and J. D. BUENROSTRO. Droplet-Based Combinatorial Indexing for Massive-Scale Single-Cell Chromatin Accessibility. *Nature Biotechnology* **37**: 916–924, 2019. (see pp. 18, 30, 105)
- [190] A. T. SATPATHY, J. M. GRANJA, K. E. YOST, Y. QI, F. MESCHI, G. P. McDERMOTT, B. N. OLSEN, M. R. MUMBACH, S. E. PIERCE, M. R. CORCES, P. SHAH, J. C. BELL, D. JHUTTY, C. M. NEMEC, J. WANG, L. WANG, Y. YIN, P. G. GIRESI, A. L. S. CHANG, G. X. Y. ZHENG, W. J. GREENLEAF, and H. Y. CHANG. Massively Parallel Single-Cell Chromatin Landscapes of Human Immune Cell Development and Intratumoral T Cell Exhaustion. *Nature Biotechnology* **37**: 925–936, 2019. (see pp. 19, 30)
- [191] R. M. MULQUEEN, D. POKHOLOK, B. L. O'CONNELL, C. A. THORNTON, F. ZHANG, B. J. O'ROAK, J. LINK, G. G. YARDIMCI, R. C. SEARS, F. J. STEEMERS, and A. C. ADEY. High-Content Single-Cell Combinatorial Indexing. *Nature Biotechnology* **39**: 1574–1580, 2021. (see pp. 19, 105)
- [192] T. M. CONSORTIUM. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**: 367–372, 2018. (see p. 19)
- [193] K. DAVIE, J. JANSSENS, D. KOLDERE, M. DE WAEGENEER, U. PECH, L. KREFT, S. AIBAR, S. MAKHZAMI, V. CHRISTIAENS, C. BRAVO GONZÁLEZ-BLAS, S. POOVATHINGAL, G. HULSELMANS, K. I. SPANIER, T. MOERMAN, B. VANSIPAUNEN, S. GEURS, T. VOET, J. LAMMERTYN, B. THIENPONT, S. LIU, N. KONSTANTINIDES, M. FIERIS, P. VERSTREKEN, and S. AERTS. A Single-Cell Transcriptome Atlas of the Aging Drosophila Brain. *Cell* **174**: Epub 2018 Jun 18, 982–998.e20, 2018. (see pp. 19, 30)
- [194] A. REGEV, S. A. TEICHMANN, E. S. LANDER, I. AMIT, C. BENOIST, E. BIRNEY, B. BODENMILLER, P. CAMPBELL, P. CARNINCI, M. CLATWORTHY, H. CLEVERS, B. DEPLANCKE, I. DUNHAM, J. EBERWINE, R. EILS, W. ENARD, A. FARMER, L. FUGGER, B. GÖTTGENS, N. HACOEN, M. HANIFFA, M. HEMBERG, S. KIM, P. KLENERMAN, A. KRIEGSTEIN, E. LEIN, S. LINNARSSON, E. LUNDBERG, J. LUNDEBERG, P. MAJUMDER, J. C. MARIONI, M. MERAD, M. MHLANGA, M. NAWIJN, M. NETEA, G. NOLAN, D. PE'ER, A. PHILLIPAKIS, C. P. PONTING, S. QUAKE, W. REIK, O. ROZENBLATT-ROSEN, J. SANES, R. SATIJA, T. N. SCHUMACHER, A. SHALEK, E. SHAPIRO, P. SHARMA, J. W. SHIN, O. STEGLE, M. STRATTON, M. J. T. STUBBINGTON, F. J. THEIS, M. UHLEN, A. VAN OUDENAARDEN, A. WAGNER, F. WATT, J. WEISSMAN, B. WOLD, R. XAVIER, and N. YOSEF. The Human Cell Atlas. *Elife* **6**: 2017. (see p. 19)
- [195] J. CAO, M. SPIELMANN, X. QIU, X. HUANG, D. M. IBRAHIM, A. J. HILL, F. ZHANG, S. MUNDLOS, L. CHRISTIANSEN, F. J. STEEMERS, C. TRAPNELL, and J. SHENDURE. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**: 496–502, 2019. (see p. 20)
- [196] J. CAO, D. R. O'DAY, H. A. PLINER, P. D. KINGSLEY, M. DENG, R. M. DAZA, M. A. ZAGER, K. A. ALDINGER, R. BLECHER-GONEN, F. ZHANG, M. SPIELMANN, J. PALIS, D. DOHERTY, F. J. STEEMERS, I. A. GLASS, C. TRAPNELL, and J. SHENDURE. A human cell atlas of fetal gene expression. *Science* **370**: eaba7721, 2020. (see p. 20)
- [197] K. L. CHIOU, X. HUANG, M. O. BOHLEN, S. TREMBLAY, A. R. DECASIEN, D. R. O'DAY, C. H. SPURRELL, A. A. GOGATE, T. M. ZINTEL, M. G. ANDREWS, M. I. MARTÍNEZ, L. M. STARITA, M. J. MONTAGUE, M. L. PLATT, J. SHENDURE, and N. SNYDER-MACKLER. A single-cell multi-omic atlas spanning the adult rhesus macaque brain. *Science Advances* **9**: eadh1914, 2023. (see p. 20)
- [198] C. QIU, B. K. MARTIN, I. C. WELSH, R. M. DAZA, T. M. LE, X. HUANG, E. K. NICHOLS, M. L. TAYLOR, O. FULTON, D. R. O'DAY, A. R. GOMES, S. ILCISIN, S. SRIVATSAN, X. DENG, C. M. DISTECHE, W. S. NOBLE, N. HAMAZAKI, C. B. MOENS, D. KIMELMAN, J. CAO, A. F. SCHIER, M. SPIELMANN, S. A. MURRAY, C. TRAPNELL, and J. SHENDURE. A single-cell time-lapse of mouse prenatal development from gastrula to birth. *Nature* **626**: Epub 2024 Feb 14, 1084–1093, 2024. (see p. 20)

- [199] J. ZHANG, A. A. UBAS, R. DE BORJA, V. SVENSSON, N. THOMAS, N. THAKAR, I. LAI, A. WINTERS, U. KHAN, M. G. JONES, V. TRAN, J. PANGALLO, E. PAPALEXI, A. SAPRE, H. NGUYEN, O. SANDERSON, M. NIGOS, O. KAPLAN, S. SCHROEDER, B. HARIADI, S. MARRUJO, C. CURCA, A. SALVINO, G. GALLARETA OLIVARES, R. KOEHLER, G. GEISS, A. ROSENBERG, C. ROCO, D. MERICO, N. ALIDOUST, H. GOODARZI, and J. YU. Tahoe-100M: A Giga-Scale Single-Cell Perturbation Atlas for Context-Dependent Gene Function and Cellular Modeling Feb. 2025 DOI: 10.1101/2025.02.20.639398 (see pp. 20, 89, 94, 95)
- [200] P. DATLINGER, A. F. RENDEIRO, T. BOENKE, M. SENEKOWITSCH, T. KRAUSGRUBER, D. BARRECA, and C. BOCK. Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. *Nature Methods* **18**: Epub 2021 May 31, 635–642, 2021. (see pp. 19, 90)
- [201] X. ZHANG, A. P. MARAND, H. YAN, and R. J. SCHMITZ. scifi-ATAC-seq: massive-scale single-cell chromatin accessibility sequencing using combinatorial fluidic indexing. *Genome Biology* **25**: 90, 2024. (see p. 20)
- [202] H. ZHANG, R. M. MULQUEEN, N. IANNUZO, D. O. FARRERA, F. POLVERINO, J. J. GALLIGAN, J. G. LEDFORD, A. C. ADEY, and D. A. CUSANOVICH. txci-ATAC-seq: a massive-scale single-cell technique to profile chromatin accessibility. *Genome Biology* **25**: 78, 2024. (see p. 20)
- [203] M. STOECKIUS, S. ZHENG, B. HOUCK-LOOMIS, S. HAO, B. Z. YEUNG, W. M. R. MAUCK, P. SMIBERT, and R. SATIJA. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biology* **19**: 224, 2018. (see p. 20)
- [204] C. S. MCGINNIS, D. M. PATTERSON, J. WINKLER, D. N. CONRAD, M. Y. HEIN, V. SRIVASTAVA, J. L. HU, L. M. MURROW, J. S. WEISSMAN, Z. WERB, E. D. CHOW, and Z. J. GARTNER. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nature Methods* **16**: Epub 2019 Jun 17, 619–626, 2019. (see p. 20)
- [205] V. MYLKA, I. MATETOVICI, S. POOVATHINGAL, J. AERTS, N. VANDAMME, R. SEURINCK, K. VERSTAEN, G. HULSELMANS, S. VAN DEN HOECKE, I. SCHEYLTJENS, K. MOVAHEDI, H. WILS, J. REUMERS, J. VAN HOUDT, S. AERTS, and Y. SAEYS. Comparative analysis of antibody- and lipid-based multiplexing methods for single-cell RNA-seq. *Genome Biology* **23**: 55, 2022. (see pp. 20, 90)
- [206] K. WANG, Z. XIAO, Y. YAN, R. YE, M. HU, S. BAI, E. SEI, Y. QIAO, H. CHEN, B. LIM, S. H. LIN, and N. E. NAVIN. Simple oligonucleotide-based multiplexing of single-cell chromatin accessibility. *Molecular Cell* **81**: 4319–4332.e10, 2021. (see p. 20)
- [207] K. BERNIS, E. M. HIJMANS, J. MULLENDERS, T. R. BRUMMELKAMP, A. VELDS, M. HEIMERIKX, R. M. KERKHOVEN, M. MADIREDOJO, W. NIJKAMP, B. WEIGELT, R. AGAMI, W. GE, G. CAVET, P. S. LINSLEY, R. L. BEIJERSBERGEN, and R. BERNARDS. A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* **428**: 431–437, 2004. (see p. 20)
- [208] M. BOUTROS, L. P. BRÁS, and W. HUBER. Analysis of cell-based RNAi screens. *Genome Biology* **7**: R66, 2006. (see p. 20)
- [209] F. A. RAN, P. D. HSU, J. WRIGHT, V. AGARWALA, D. A. SCOTT, and F. ZHANG. Genome engineering using the CRISPR-Cas9 system. *Nature Protocols* **8**: Epub 2013 Oct 24, 2281–2308, 2013. (see p. 20)
- [210] O. SHALEM, N. E. SANJANA, E. HARTENIAN, X. SHI, D. A. SCOTT, T. MIKKELSON, D. HECKL, B. L. EBERT, D. E. ROOT, J. G. DOENCH, and F. ZHANG. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**: Epub 2013 Dec 12, 84–87, 2014. (see p. 20)
- [211] F. BEHAN, F. IORIO, G. PICCO, E. GONÇALVES, C. BEAVER, G. MIGLIARDI, R. SANTOS, Y. RAO, F. SASSI, M. PINNELLI, R. ANSARI, S. HARPER, D. JACKSON, R. MCRAE, R. POOLEY, P. WILKINSON, D. VAN DER MEER, D. DOW, C. BUSER-DOEPNER, A. BERTOTTI, L. TRUSOLINO, E. STRONACH, J. SAEZ-RODRIGUEZ, K. YUSA, and M. GARNETT. Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. *Nature* **568**: Epub 2019 Apr 10, 511–516, 2019. (see p. 20)
- [212] C. PACINI, J. M. DEMPSTER, I. BOYLE, E. GONÇALVES, H. NAJGEBAUER, E. KARAKOC, D. VAN DER MEER, A. BARTHORPE, H. LIGHTFOOT, P. JAAKS, J. M. MCFARLAND, M. J. GARNETT, A. TSHERNIAK, and F. IORIO. Integrated cross-study datasets of genetic dependencies in cancer. *Nature Communications* **12**: 1661, 2021. (see p. 20)

- [213] S. XIE, A. COOLEY, D. ARMENDARIZ, P. ZHOU, and G. C. HON. Frequent sgRNA-barcode recombination in single-cell perturbation assays. *PLoS One* **13**: e0198635, 2018. (see p. 21)
- [214] A. DIXIT, O. PARNAS, B. LI, J. CHEN, C. P. FULCO, L. JERBY-ARNON, N. D. MARJANOVIC, D. DIONNE, T. BURKS, R. RAYCHOWDHURY, B. ADAMSON, T. M. NORMAN, E. S. LANDER, J. S. WEISSMAN, N. FRIEDMAN, and A. REGEV. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**: 1853–1866.e17, 2016. (see p. 21)
- [215] B. ADAMSON, T. M. NORMAN, M. JOST, A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* **167**: 1867–1882.e21, 2016. (see p. 21)
- [216] D. SCHRAIVOGEL, A. R. GSCHWIND, J. H. MILBANK, D. R. LEONCE, P. JAKOB, L. MATHUR, J. O. KORBEL, C. A. MERTEN, L. VELTEN, and L. M. STEINMETZ. Targeted Perturb-seq enables genome-scale genetic screens in single cells. *Nature Methods* **17**: 629–635, 2020. (see p. 21)
- [217] D. YAO, L. BINAN, J. BEZNEY, B. SIMONTON, J. FREEDMAN, C. J. FRANGIEH, K. DEY, K. GEIGER-SCHULLER, B. ERASLAN, A. GUSEV, A. REGEV, and B. CLEARY. Scalable genetic screening for regulatory circuits using compressed Perturb-seq. *Nature Biotechnology* **42**: Epub 2023 Oct 23, 1282–1295, 2024. (see p. 21)
- [218] P. DATLINGER, A. F. RENDEIRO, C. SCHMIDL, T. KRAUSGRUBER, P. TRAXLER, J. KLUGHAMMER, L. C. SCHUSTER, A. KUCHLER, D. ALPAR, and C. BOCK. Pooled CRISPR screening with single-cell transcriptome read-out. *Nature Methods* **14**: Epub 2017 Jan 18, 297–301, 2017. (see p. 21)
- [219] A. J. RUBIN, K. R. PARKER, A. T. SATPATHY, Y. QI, B. WU, A. J. ONG, M. R. MUMBACH, A. L. JI, D. S. KIM, S. W. CHO, B. J. ZARNEGAR, W. J. GREENLEAF, H. Y. CHANG, and P. A. KHAVARI. Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks. *Cell* **176**: Epub 2018 Dec 20, 361–376.e17, 2019. (see p. 21)
- [220] S. E. PIERCE, J. M. GRANJA, and W. J. GREENLEAF. High-throughput single-cell chromatin accessibility CRISPR screens enable unbiased identification of regulatory networks in cancer. *Nature Communications* **12**: 2969, 2021. (see p. 22)
- [221] S. R. SRIVATSAN, J. L. MCFALINE-FIGUEROA, V. RAMANI, L. SAUNDERS, J. CAO, J. PACKER, H. A. PLINER, D. L. JACKSON, R. M. DAZA, L. CHRISTIANSEN, F. ZHANG, F. STEEMERS, J. SHENDURE, and C. TRAPNELL. Massively multiplex chemical transcriptomics at single-cell resolution. *Science* **367**: 45–51, 2020. (see p. 22)
- [222] C. TRAPNELL, L. PACTER, and S. L. SALZBERG. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: Epub 2009 Mar 16, 1105–1111, 2009. (see p. 22)
- [223] C. TRAPNELL, B. A. WILLIAMS, G. PERTEA, A. MORTAZAVI, G. KWAN, M. J. VAN BAREN, S. L. SALZBERG, B. J. WOLD, and L. PACTER. Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation. *Nature Biotechnology* **28**: 511–515, 2010. (see p. 22)
- [224] A. DOBIN, C. A. DAVIS, F. SCHLESINGER, J. DRENKOW, C. ZALESKI, S. JHA, P. BATUT, M. CHAISSON, and T. R. GINGERAS. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: Epub 2012 Oct 25, 15–21, 2013. (see p. 22)
- [225] H. LI, B. HANDSAKER, A. WYSOKER, T. FENNELL, J. RUAN, N. HOMER, G. MARTH, G. ABECASIS, and R. DURBIN. The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* **25**: 2078–2079, 2009. (see p. 23)
- [226] H. THORVALDSDÓTTIR, J. T. ROBINSON, and J. P. MESIROV. Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and Exploration. *Briefings in Bioinformatics* **14**: 178–192, 2013. (see p. 23)
- [227] Y. ZHANG, T. LIU, C. A. MEYER, J. EECKHOUTE, D. S. JOHNSON, B. E. BERNSTEIN, C. NUSBAUM, R. M. MYERS, M. BROWN, W. LI, and X. S. LIU. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**: Open access, 2008. (see p. 23)
- [228] J. M. GASPAR. Improved peak-calling with MACS2. View ORCID Profile, 2018. (see p. 24)

- [229] E. P. CONSORTIUM, J. E. MOORE, M. J. PURCARO, H. E. PRATT, C. B. EPSTEIN, N. SHORESH, J. ADRIAN, T. KAWLI, C. A. DAVIS, A. DOBIN, R. KAUL, J. HALOW, E. L. VAN NOSTRAND, P. FREESE, D. U. GORKIN, Y. SHEN, Y. HE, M. MACKIEWICZ, F. PAULI-BEHN, B. A. WILLIAMS, A. MORTAZAVI, C. A. KELLER, X. O. ZHANG, S. I. ELHAJJAJY, J. HUEY, D. E. DICKEL, V. SNETKOVA, X. WEI, X. WANG, J. C. RIVERA-MULIA, J. ROZOWSKY, J. ZHANG, S. B. CHHETRI, A. VICTORSEN, K. P. WHITE, A. VISEL, G. W. YEO, C. B. BURGE, E. LÉCUYER, D. M. GILBERT, J. DEKKER, J. RINN, E. M. MENDENHALL, J. R. ECKER, M. KELLIS, R. J. KLEIN, W. S. NOBLE, A. KUNDAJE, R. GUIGÓ, P. J. FARNHAM, J. M. CHERRY, R. M. MYERS, B. REN, B. R. GRAVELEY, M. B. GERSTEIN, L. A. PENNACCHIO, M. P. SNYDER, B. E. BERNSTEIN, B. WOLD, R. C. HARDISON, T. R. GINGERAS, J. A. STAMATOYANNOPOULOS, and Z. WENG. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**: Erratum in: *Nature*. 2022 May;605(7909):E3., 699–710, 2020. (see p. 24)
- [230] H. M. AMEMIYA, A. KUNDAJE, and A. P. BOYLE. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific Reports* **9**: Open access, 2019. (see p. 24)
- [231] C. A. LAREAU, S. MA, F. M. DUARTE, and J. D. BUENROSTRO. Inference and effects of barcode multiplets in droplet-based single-cell assays. *Nature Communications* **11**: 866, 2020. (see pp. 25, 101)
- [232] M. D. ROBINSON, D. J. MCCARTHY, and G. K. SMYTH. edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data. *Bioinformatics* **26**: 139–140, 2010. (see p. 25)
- [233] S. ANDERS and W. HUBER. Differential Expression Analysis for Sequence Count Data. *Genome Biology* **11**: R106, 2010. (see p. 25)
- [234] L. J. P. VAN DER MAATEN and G. E. HINTON. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research* **9**: 2579–2605, 2008. (see p. 25)
- [235] E.-A. D. AMIR, K. L. DAVIS, M. D. TADMOR, E. F. SIMONDS, J. H. LEVINE, S. C. BENDALL, D. K. SHENFELD, S. KRISHNASWAMY, G. P. NOLAN, and D. PE’ER. viSNE Enables Visualization of High Dimensional Single-Cell Data and Reveals Phenotypic Heterogeneity of Leukemia. *Nature Biotechnology* **31**: 545–552, 2013. (see p. 25)
- [236] A. SAADATPOUR, G. GUO, S. H. ORKIN, and G.-C. YUAN. Characterizing Heterogeneity in Leukemic Cells Using Single-Cell Gene Expression Analysis. *Genome Biology* **15**: 525, 2014. (see p. 25)
- [237] E. BECHT, L. MCINNES, J. HEALY, C.-A. DUTERTRE, I. W. H. KWOK, L. G. NG, F. GINHOUX, and E. W. NEWELL. Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP. *Nature Biotechnology*, 2018. (see p. 25)
- [238] T. CHARI and L. PACHTER. The Specious Art of Single-Cell Genomics. *PLOS Computational Biology* **19**: e1011288, 2023. (see p. 25)
- [239] T. STUART, A. BUTLER, P. HOFFMAN, C. HAFEMEISTER, E. PAPALEXI, W. M. MAUCK, Y. HAO, M. STOECKIUS, P. SMIBERT, and R. SATIJA. Comprehensive Integration of Single-Cell Data. *Cell* **177**: 1888–1902.e21, 2019. (see pp. 25, 106)
- [240] H. HOTELLING. Relations Between Two Sets of Variates. *Biometrika* **28**: 321–377, 1936. (see p. 25)
- [241] F. A. WOLF, P. ANGERER, and F. J. THEIS. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology* **19**: Open access, 2018. (see p. 25)
- [242] D. A. CUSANOVICH, A. J. HILL, D. AGHAMIRZAI, R. M. DAZA, H. A. PLINER, J. B. BERLETCHE, G. N. FILIPOVA, X. HUANG, L. CHRISTIANSEN, W. S. DEWITT, C. LEE, S. G. REGALADO, D. F. READ, F. J. STEEMERS, C. M. DISTECHE, C. TRAPNELL, and J. SHENDURE. A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174**: 1309–1324.e18, 2018. (see pp. 25, 30)
- [243] C. BRAVO GONZÁLEZ-BLAS, L. MINNOYE, D. PAPASOKRATI, S. AIBAR, G. HULSELMANS, V. CHRISTIAENS, K. DAVIE, J. WOUTERS, and S. AERTS. cisTopic: Cis-Regulatory Topic Modeling on Single-Cell ATAC-Seq Data. *Nature Methods* **16**: 397–400, 2019. (see pp. 26, 27, 106)
- [244] J. K. PRITCHARD, M. STEPHENS, and P. DONNELLY. Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **155**: 945–959, 2000. (see p. 26)

- [245] D. M. BLEI, A. Y. NG, and M. I. JORDAN. Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**: 993–1022, 2003. (see p. 26)
- [246] M. STEYVERS and T. GRIFFITHS. “Probabilistic Topic Models” in: *Handbook of Latent Semantic Analysis* ed. by T. K. LANDAUER, D. S. MCNAMARA, S. DENNIS, and W. KINTSCH. Lawrence Erlbaum Associates Publishers, 2007. 427–448 (see p. 26)
- [247] H. IMRICHOVÁ, G. HULSELMANS, Z. K. ATAK, D. POTIER, and S. AERTS. i-cisTarget 2015 update: generalized cis-regulatory enrichment analysis in human, mouse and fly. *Nucleic Acids Research* **43**: Epub 2015 Apr 29, W57–64, 2015. (see p. 26)
- [248] G. YU, L.-G. WANG, and Q.-Y. HE. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**: 2382–2383, 2015. (see p. 26)
- [249] D. H. HUBEL and T. N. WIESEL. Receptive fields of single neurones in the cat’s striate cortex. *Journal of Physiology* **148**: 574–591, 1959. (see p. 27)
- [250] K. FUKUSHIMA. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics* **36**: 193–202, 1980. (see p. 27)
- [251] Y. LECUN, L. BOTTOU, Y. BENGIO, and P. HAFNER. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**: 2278–2324, 1998. (see p. 27)
- [252] J. ZHOU and O. G. TROYANSKAYA. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods* **12**: Epub 2015 Aug 24, 931–934, 2015. (see p. 27)
- [253] D. R. KELLEY, Y. A. RESHEF, M. BILESCHI, D. BELANGER, C. Y. MCLEAN, and J. SNOEK. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research* **28**: Epub 2018 Mar 27, 739–750, 2018. (see p. 27)
- [254] V. AVSEC, V. AGARWAL, D. VISENTIN, J. R. LEDSAM, A. GRABSKA-BARWIŃSKA, K. R. TAYLOR, Y. ASSAEL, J. JUMPER, P. KOHLI, and D. R. KELLEY. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods* **18**: 1196–1203, 2021. (see p. 28)
- [255] J. LINDER, D. SRIVASTAVA, H. YUAN, V. AGARWAL, and D. R. KELLEY. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. *Nature Genetics* **57**: Epub 2025 Jan 8, 949–961, 2025. (see p. 28)
- [256] A. PAMPARI, A. SHCHERBINA, E. Z. KVON, M. KOSICKI, S. NAIR, S. KUNDU, A. S. KATHIRIA, V. I. RISCA, K. KUNINGAS, K. ALASOO, W. J. GREENLEAF, L. A. PENNACCHIO, and A. KUNDAJE. ChromBPNet: bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants. *bioRxiv Preprint*, 2025. (see p. 28)
- [257] J. WOUTERS, Z. KALENDER-ATAK, L. MINNOYE, K. I. SPANIER, M. DE WAEGENEER, C. BRAVO GONZÁLEZ-BLAS, D. MAUDUIT, K. DAVIE, G. HULSELMANS, A. NAJEM, M. DEWAELE, D. PEDRI, F. RAMBOW, S. MAKHZAMI, V. CHRISTIAENS, F. CEYSENS, G. GHANEM, J. C. MARINE, S. POOVATHINGAL, and S. AERTS. Robust gene expression programs underlie recurrent cell states and phenotype switching in melanoma. *Nature Cell Biology* **22**: Epub 2020 Aug 3, 986–998, 2020. (see p. 28)
- [258] L. MINNOYE, I. I. TASKIRAN, D. MAUDUIT, M. FAZIO, L. VAN AERSCHOT, G. HULSELMANS, V. CHRISTIAENS, S. MAKHZAMI, M. SELTENHAMMER, P. KARRAS, A. PRIMOT, E. CADIEU, E. VAN ROOIJEN, J. C. MARINE, G. EGIDY, G. E. GHANEM, L. ZON, J. WOUTERS, and S. AERTS. Cross-species analysis of enhancer logic using deep learning. *Genome Research* **30**: Epub 2020 Jul 30, 1815–1834, 2020. (see pp. 28, 92)
- [259] J. JANSSENS, S. AIBAR, I. I. TASKIRAN, J. N. ISMAIL, A. E. GOMEZ, G. AUGHEY, K. I. SPANIER, F. V. DE ROP, C. B. GONZÁLEZ-BLAS, M. DIONNE, K. GRIMES, X. J. QUAN, D. PAPASOKRATI, G. HULSELMANS, S. MAKHZAMI, M. DE WAEGENEER, V. CHRISTIAENS, T. SOUTHALL, and S. AERTS. Decoding gene regulation in the fly brain. *Nature* **601**: Epub 2022 Jan 5, 630–636, 2022. (see pp. 28, 103)
- [260] A. VASWANI, N. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES, A. N. GOMEZ, L. KAISER, and I. POLOSUKHIN. Attention Is All You Need 2017 eprint: 1706.03762 (cs.CL) URL: <https://arxiv.org/abs/1706.03762> (see p. 28)

- [261] Y. Ji, Z. ZHOU, H. LIU, and R. V. DAVULURI. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* **37**: 2112–2120, 2021. (see p. 29)
- [262] H. DALLA-TORRE, L. GONZALEZ, J. MENDOZA-REVILLA, N. LOPEZ CARRANZA, A. H. GRZYWACZEWSKI, F. OTERI, C. DALLAGO, E. TROP, B. P. DE ALMEIDA, H. SIRELKHATIM, G. RICHARD, M. SKWARK, K. BEGUIR, M. LOPEZ, and T. PIERROT. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods* **22**: Open access, Published: 28 November 2024, 287–297, 2025. (see p. 29)
- [263] G. BRIxi, M. G. DURRANT, J. KU, M. POLI, G. BROCKMAN, D. CHANG, G. A. GONZALEZ, S. H. KING, D. B. LI, A. T. MERCHANT, M. NAGHIPOURFAR, E. NGUYEN, C. RICCI-TAM, D. W. ROMERO, G. SUN, A. TAGHIBAKSHI, A. VORONTSOV, B. YANG, M. DENG, L. GORTON, N. NGUYEN, N. K. WANG, E. ADAMS, S. A. BACCUS, S. DILLMANN, S. ERMON, D. GUO, R. ILANGO, K. JANIK, A. X. LU, R. MEHTA, M. R. MOFRAD, M. Y. NG, J. PANNU, C. RÉ, J. C. SCHMOK, J. ST. JOHN, J. SULLIVAN, K. ZHU, G. ZYNDA, D. BALSAM, P. COLLISON, A. B. COSTA, T. HERNANDEZ-BOUSSARD, E. HO, M.-Y. LIU, T. MCGRATH, K. POWELL, D. P. BURKE, H. GOODARZI, P. D. HSU, and B. L. HIE. Genome modeling and design across all domains of life with Evo 2. *bioRxiv*, 2025. (see pp. 29, 91)
- [264] A. W. SENIOR, R. EVANS, J. JUMPER, J. KIRKPATRICK, L. SIFRE, T. GREEN, C. QIN, A. ŽÍDEK, A. W. R. NELSON, A. BRIDGLAND, H. PENEDONES, S. PETERSEN, K. SIMONYAN, S. CROSSAN, P. KOHLI, D. JONES, D. SILVER, K. KAVUKCUOGLU, and D. HASSABIS. Improved protein structure prediction using potentials from deep learning. *Nature* **577**: Epub 2020 Jan 15, 706–710, 2020. (see p. 29)
- [265] J. JUMPER, R. EVANS, A. PRITZEL, T. GREEN, M. FIGURNOV, O. RONNEBERGER, K. TUNYASUVUNAKOOL, R. BATES, A. ŽÍDEK, A. POTAPENKO, A. BRIDGLAND, C. MEYER, S. A. A. KOHL, A. J. BALLARD, A. COWIE, B. ROMERA-PAREDES, S. NIKOLOV, R. JAIN, J. ADLER, T. BACK, S. PETERSEN, D. REIMAN, E. CLANCY, M. ZIELINSKI, M. STEINEGGER, M. PACHOLSKA, T. BERGHAMMER, S. BODENSTEIN, D. SILVER, O. VINIYALS, A. W. SENIOR, K. KAVUKCUOGLU, P. KOHLI, and D. HASSABIS. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**: Epub 2021 Jul 15, 583–589, 2021. (see p. 29)
- [266] C. V. THEODORIS, L. XIAO, A. CHOPRA, M. D. CHAFFIN, Z. R. AL SAYED, M. C. HILL, H. MANTINEO, E. M. BRYDON, Z. ZENG, X. S. LIU, and P. T. ELLINOR. Transfer learning enables predictions in network biology. *Nature* **618**: Epub 2023 May 31, 616–624, 2023. (see p. 29)
- [267] M. HAO, J. GONG, X. ZENG, C. LIU, Y. GUO, X. CHENG, T. WANG, J. MA, X. ZHANG, and L. SONG. Large-scale foundation model on single-cell transcriptomics. *Nature Methods* **21**: Published: 06 June 2024, 1481–1491, 2024. (see p. 29)
- [268] E. MEREU, A. LAFZI, C. MOUTINHO, C. ZIEGENHAIN, D. J. MCCARTHY, A. ÁLVAREZ VARELA, E. BATLLE, SAGAR, D. GRÜN, J. K. LAU, S. C. BOUTET, C. SANADA, A. OOI, R. C. JONES, K. KAIHARA, C. BRAMPTON, Y. TALAGA, Y. SASAGAWA, K. TANAKA, T. HAYASHI, C. BRAEUNING, C. FISCHER, S. SAUER, T. TREFZER, C. CONRAD, X. ADICONIS, L. T. NGUYEN, A. REGEV, J. Z. LEVIN, S. PAREKH, A. JANJIC, L. E. WANGE, J. W. BAGNOLI, W. ENARD, M. GUT, R. SANDBERG, I. NIKAIDO, I. GUT, O. STEGLE, and H. HEYN. Benchmarking Single-Cell RNA-sequencing Protocols for Cell Atlas Projects. *Nature Biotechnology* **38**: 747–755, 2020. (see pp. 29, 30)
- [269] D. A. CUSANOVICH, R. DAZA, A. ADEY, H. A. PLINER, L. CHRISTIANSEN, K. L. GUNDERSON, F. J. STEEMERS, C. TRAPNELL, and JAY SHENDURE. Multiplex Single-Cell Profiling of Chromatin Accessibility by Combinatorial Cellular Indexing. *Science* **348**: 910–914, 2015. (see p. 29)
- [270] A. R. WU, N. F. NEFF, T. KALISKY, P. DALERBA, B. TREUTLEIN, M. E. ROTHENBERG, F. M. MBURU, G. L. MANTALAS, S. SIM, M. F. CLARKE, and S. R. QUAKE. Quantitative assessment of single-cell RNA-sequencing methods. *Nature Methods* **11**: 41–46, 2013. (see p. 30)
- [271] C. ZIEGENHAIN, B. VIETH, S. PAREKH, B. REINIUS, A. GUILLAUMET-ADKINS, M. SMETS, H. LEONHARDT, H. HEYN, I. HELLMANN, and W. ENARD. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell* **65**: 631–643.e4, 2017. (see p. 30)
- [272] RAN BIOTECHNOLOGIES Gel Beads & NoChipDrop Kit <https://www.ranbiotechnologies.com/products/gel-beads-nochipdrop-kit/> Accessed: 2025-09-23 2025 (see p. 89)

- [273] 10X GENOMICS, INC. Form 10-K: Annual report for the fiscal year ended December 31, 2024 U.S. Securities and Exchange Commission filing 2025 URL: <https://investors.10xgenomics.com/financials/sec-filings/default.aspx> (see p. 90)
- [274] I. C. CLARK, K. M. FONTANEZ, R. H. MELTZER, Y. XUE, C. HAYFORD, A. MAY-ZHANG, C. D'AMATO, A. OSMAN, J. Q. ZHANG, P. HETTIGE, J. S. A. ISHIBASHI, C. L. DELLEY, D. W. WEISGERBER, J. M. REPLOGLE, M. JOST, K. T. PHONG, V. E. KENNEDY, C. A. C. PERETZ, E. A. KIM, S. SONG, W. KARLON, J. S. WEISSMAN, C. C. SMITH, Z. J. GARTNER, and A. R. ABATE. Microfluidics-free single-cell genomics with templated emulsification. *Nature Biotechnology* **41**: Epub 2023 Mar 6, 1557–1566, 2023. (see p. 90)
- [275] ILLUMINA, INC. Illumina acquires Fluent BioSciences to accelerate single-cell analysis and discovery to a broader customer base <https://investor.illumina.com/news/press-release-details/2024/Illumina-acquires-Fluent-BioSciences-to-accelerate-single-cell-analysis-and-discovery-to-a-broader-customer-base/default.aspx> Press release July 2024 URL: <https://investor.illumina.com/news/press-release-details/2024/Illumina-acquires-Fluent-BioSciences-to-accelerate-single-cell-analysis-and-discovery-to-a-broader-customer-base/default.aspx> (see p. 90)
- [276] H. DICKMÄNKEN, M. WOJNO, K. THEUNIS, E. C. EKŞİ, L. MAHIEU, V. CHRISTIAENS, N. KEMPYNCK, F. V. DE ROP, N. ROELS, K. I. SPANIER, R. VANDEPOEL, G. HULSELMANS, S. POOVATHINGAL, and S. AERTS. HyDrop v2: Scalable atlas construction for training sequence-to-function models. *bioRxiv*, 2025. (see p. 90)
- [277] R. SUTTON The Bitter Lesson Mar. 2019 URL: <http://www.incompleteideas.net/IncIdeas/BitterLesson.html> (see p. 91)
- [278] WAYMO New Insights for Scaling Laws in Autonomous Driving Waypoint: The official Waymo blog June 2025 URL: <https://blog.waymo.com/2025/06/new-insights-for-scaling-laws-in.html> (see p. 91)
- [279] T. B. BROWN, B. MANN, N. RYDER, M. SUBBIAH, J. KAPLAN, P. DHARIWAL, A. NEELAKANTAN, P. SHYAM, G. SASTRY, A. ASKELL, S. AGARWAL, A. HERBERT-VOSS, G. KRUEGER, T. HENIGHAN, R. CHILD, A. RAMESH, D. M. ZIEGLER, J. WU, C. WINTER, C. HESSE, M. CHEN, E. SIGLER, M. LITWIN, S. GRAY, B. CHESS, J. CLARK, C. BERNER, S. MCCANDLISH, A. RADFORD, I. SUTSKEVER, and D. AMODEI. Language Models are Few-Shot Learners. 2020. (see p. 91)
- [280] J. DEVLIN, M.-W. CHANG, K. LEE, and K. TOUTANOVA. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019. (see p. 91)
- [281] Y. LIU, M. OTT, N. GOYAL, J. DU, M. JOSHI, D. CHEN, O. LEVY, M. LEWIS, L. ZETTLEMOYER, and V. STOYANOV. RoBERTa: A Robustly Optimized BERT Pretraining Approach 2019 DOI: 10.48550/arXiv.1907.11692 arXiv: 1907.11692 [cs.CL] (see p. 91)
- [282] K. PARVEEN, T. Q. B. PHUC, A. A. ALGHAMDI, F. HAJJEJ, W. J. OBIDALLAH, Y. A. ALDURAYWISH, and M. SHAFIQ. Unraveling the dynamics of ChatGPT adoption and utilization through Structural Equation Modeling. *Scientific Reports* **14**: Open access, 2024. (see p. 91)
- [283] J. KAPLAN, S. MCCANDLISH, T. HENIGHAN, T. B. BROWN, B. CHESS, R. CHILD, S. GRAY, A. RADFORD, J. WU, and D. AMODEI Scaling Laws for Neural Language Models 2020 arXiv: 2001.08361 [cs.LG] URL: <https://arxiv.org/abs/2001.08361> (see p. 91)
- [284] J. HOFFMANN, S. BORGEAUD, A. MENSCH, E. BUCHATSKAYA, T. CAI, E. RUTHERFORD, D. DE LAS CASAS, L. A. HENDRICKS, J. WELBL, A. CLARK, T. HENNIGAN, E. NOLAND, K. MILLICAN, G. VAN DEN DRIESSCHE, B. DAMOC, A. GUY, S. OSINDERO, K. SIMONYAN, E. ELSER, J. W. RAE, O. VINYALS, and L. SIFRE Training Compute-Optimal Large Language Models 2022 arXiv: 2203.15556 [cs.CL] URL: <https://arxiv.org/abs/2203.15556> (see p. 91)
- [285] K. A. WETTERSTRAND DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) <https://www.genome.gov/sequencingcostsdata> Accessed: 2025-06-19 2023 (see p. 91)

- [286] A. K. ADDURI, D. GAUTAM, B. BEVILACQUA, A. IMRAN, R. SHAH, M. NAGHIPOURFAR, N. TEYSSIER, R. ILANGO, S. NAGARAJ, M. DONG, C. RICCI-TAM, C. CARPENTER, V. SUBRAMANYAM, A. WINTERS, S. TIRUKKOVULAR, J. SULLIVAN, B. S. PLOSKY, B. ERASLAN, N. D. YOUNGBLUT, J. LESKOVEC, L. A. GILBERT, S. KONERMANN, P. D. HSU, A. DOBIN, D. P. BURKE, H. GOODARZI, and Y. H. ROOHANI. Predicting cellular responses to perturbation across diverse contexts with State. 2025. (see p. 91)
- [287] A. RIVES, J. MEIER, T. SERCU, S. GOYAL, Z. LIN, J. LIU, D. GUO, M. OTT, C. L. ZITNICK, J. MA, and R. FERGUS. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences of the United States of America* **118**: e2016239118, 2021. (see p. 91)
- [288] Z. LIN, H. AKIN, R. RAO, B. HIE, Z. ZHU, W. LU, N. SMETANIN, R. VERKUIL, O. KABELI, Y. SHMUELI, A. DOS SANTOS COSTA, M. FAZEL-ZARANDI, T. SERCU, S. CANDIDO, and A. RIVES. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**: 1123–1130, 2023. (see p. 91)
- [289] T. HAYES, R. RAO, H. AKIN, N. J. SOFRONIEW, D. OKTAY, Z. LIN, R. VERKUIL, V. Q. TRAN, J. DEATON, M. WIGGERT, R. BADKUNDRI, I. SHAFKAT, J. GONG, A. DERRY, R. S. MOLINA, N. THOMAS, Y. A. KHAN, C. MISHRA, C. KIM, L. J. BARTIE, M. NEMETH, P. D. HSU, T. SERCU, S. CANDIDO, and A. RIVES. Simulating 500 million years of evolution with a language model. *Science* **387**: 850–858, 2025. (see p. 91)
- [290] J. DURAIRAJ, Y. ADESHINA, Z. CAO, X. ZHANG, V. OLEINIKOVAS, T. DUGNAN, Z. MCCLURE, X. ROBIN, D. KOVTUN, E. ROSSI, G. ZHOU, S. VECCHAM, C. ISERT, Y. PENG, P. SUNDARESON, M. AKDEL, G. CORSO, H. STÄRK, Z. CARPENTER, M. BRONSTEIN, E. KUCUKBENLI, T. SCHWEDE, and L. NAEF. PLINDER: The protein-ligand interactions dataset and evaluation resource. 2024. (see p. 91)
- [291] P. ŠKRINJAR, J. EBERHARDT, J. DURAIRAJ, and T. SCHWEDE Have protein-ligand co-folding methods moved beyond memorisation? *bioRxiv* 2025.02.03.636309 2025 DOI: 10.1101/2025.02.03.636309 (see p. 91)
- [292] E. W. SAYERS, M. CAVANAUGH, L. FRISSE, K. D. PRUITT, V. A. SCHNEIDER, B. A. UNDERWOOD, L. YANKIE, and I. KARSCH-MIZRACHI. GenBank 2025 update. *Nucleic Acids Research* **53**: D56–D61, 2025. (see p. 92)
- [293] E. CLOUGH, T. BARRETT, S. E. WILHITE, P. LEDOUX, C. EVANGELISTA, I. F. KIM, M. TOMASHEVSKY, K. A. MARSHALL, K. H. PHILLIPPY, P. M. SHERMAN, H. LEE, N. ZHANG, N. SEROVA, L. WAGNER, V. ZALUNIN, A. KOCHERGIN, and A. SOBOLEVA. NCBI GEO: archive for gene expression and epigenomics data sets: 23-year update. *Nucleic Acids Research* **52**: D138–D144, 2024. (see p. 92)
- [294] F. ZUO, N. SU, X. XIE, L. JIANG, M. FANG, Y. ZHAO, L. ZHU, X. CHEN, and Y. YANG. Live-cell imaging of RNA dynamics using bright and stable fluorescent RNAs. *Nature Protocols* **21**: 3765–3794, 2026. (see p. 92)
- [295] M. J. CARROZZA, B. LI, L. FLORENS, T. SUGANUMA, S. K. SWANSON, K. K. LEE, W.-J. SHIA, S. ANDERSON, J. YATES, M. P. WASHBURN, and J. L. WORKMAN. Histone H3 Methylation by Set2 Directs Deacetylation of Coding Regions by Rpd3S to Suppress Spurious Intragenic Transcription. *Cell* **123**: 581592, 2005. (see p. 92)
- [296] F. CAPUANO, M. MÜLLEDER, R. KOK, H. J. BLOM, and M. RALSER. Cytosine DNA Methylation Is Found in *Drosophila melanogaster* but Absent in *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and Other Yeast Species. *Analytical Chemistry* **86**: 36973702, 2014. (see p. 92)
- [297] A. ANDERSSON-ROLF and H. CLEVERS. New developments and applications of human organoids. *Nature Reviews Molecular Cell Biology* **27**: 543558, 2026. (see p. 93)
- [298] TAHOE THERAPEUTICS, ARC INSTITUTE, and BIOHUB Tahoe Therapeutics, Arc Institute, and Biohub Partner to Generate the Largest Perturbation Dataset for Virtual Cell Models Announced dataset comprising more than 120 million single-cell measurements across 225,000 drug–patient perturbation interactions Jan. 2026 URL: <https://arcinstitute.org/news/tahoe-arc-biohub> (see pp. 93–95)
- [299] A. C. HUANG, T.-H. S. HSIEH, J. ZHU, J. MICHUDA, A. TENG, S. KIM, X-Atlas/Orion: Genome-wide Perturb-seq Datasets via a Scalable Fix-Cryopreserve Platform for Training Dose-Dependent Biological Foundation Models. *bioRxiv* Preprint, 2025. (see pp. 93–95)

- [300] R. ZHU, E. DANN, J. YAN, J. REYES RETANA, R. GOTO, R. C. GUITCHE, L. BRIxi, M. OTA, A. HARTMAN, T. L. ROTH, A. T. SATPATHY, J. K. PRITCHARD, and A. MARSON. Genome-scale Perturb-seq in primary human CD4+ T cells maps context-specific regulators of T cell programs and human immune traits. *Cell* Genome-scale CRISPRi Perturb-seq across approximately 22 million primary human CD4+ T cells, 2026. (see p. 93)
- [301] K. HUANG, R. LOPEZ, J.-C. HÜTTER, T. KUDO, A. RIOS, and A. REGEV. Sequential Optimal Experimental Design of Perturbation Screens Guided by Multi-modal Priors. *bioRxiv* Preprint, 2023. (see p. 93)
- [302] D. GRÜN, L. KESTER, and A. VAN OUDENAARDEN. Validation of noise models for single-cell transcriptomics. *Nature Methods* **11**: 637–640, 2014. (see p. 93)
- [303] V. SVENSSON, K. N. NATARAJAN, L.-H. LY, R. J. MIRAGAIA, C. LABALETTE, I. C. MACAULAY, A. CVEJIC, and S. A. TEICHMANN. Power analysis of single-cell RNA-sequencing experiments. *Nature Methods* **14**: 381–387, 2017. (see p. 93)
- [304] HUMAN CELL ATLAS Human Cell Atlas achieves leap in understanding of the human body 2024 URL: https://www.humancellatlas.org/press_release/human-cell-atlas-achieves-leap-in-understanding-of-the-human-body/ (visited on 09/12/2026) (see pp. 93, 94)
- [305] CHAN ZUCKERBERG INITIATIVE Go Behind the Scenes of CZ CELLxGENE URL: <https://chanzuckerberg.com/blog/behind-the-scenes-of-chan-zuckerberg-cellxgene/> (visited on 09/12/2026) (see pp. 93, 94)
- [306] J. KAISER NIH’s BRAIN Initiative puts \$500 million into creating most detailed ever human brain atlas Science (news) Sept. 2022 URL: <https://www.science.org/content/article/nihs-brain-initiative-puts-dollar500-million-creating-detailed-ever-human-brain-atlas> (see pp. 93, 94)
- [307] J. R. ECKER, M. HAWRYLYCZ, E. LEIN, B. REN, C. THOMPSON, H. ZENG, O. WHITE, and G.-Q. ZHANG. A Community Standard Multispecies Cell Atlas of the Basal Ganglia. *bioRxiv*, 2026. (see pp. 93, 94)
- [308] HUBMAP CONSORTIUM. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature* **574**: 187–192, 2019. (see pp. 93, 95)
- [309] ALIGNING SCIENCE ACROSS PARKINSON’S Team Voet URL: <https://www.asapcrn.org/research-community/teams/voet/> (visited on 09/12/2026) (see pp. 93, 94)
- [310] O. M. SIGALOVA, A. PANÍKOVÁ, J. DE MAN, K. THEUNIS, G. J. HULSELMANS, V. KONSTANTAKOS, B. STUYVEN, A. DE BRABANDERE, J. GEURTS, A. MIKORSKA, S. MUKHERJEE, S. ABOUELASRAR SALAMA, K. VANDEREYKEN, K. DAVIE, L. MAHIEU, C. H. ADLER, T. G. BEACH, G. E. SERRANO, T. VOET, J. DEMEULEMEESTER, and S. AERTS. Modeling cis-regulatory variation in human brain enhancers across a large Parkinson’s Disease cohort. *bioRxiv* Preprint, 2026. (see pp. 93, 94)
- [311] BGI GROUP Launch of the 10 Billion Cells Alliance 2025 URL: <https://en.genomics.cn/scientificdiscoveries-7179.html> (visited on 09/12/2026) (see pp. 93, 95)
- [312] J. YIN, Y. ZHENG, Z. HUANG, W. ZHOU, Y. YUAN, P. CAI, Y. BAI, S. YANG, Y. GAO, S. DUAN, Y. WANG, Z. XU, W. ZHANG, X. ZHANG, Y. WEI, Y. HUANG, Y. LIU, W. WANG, T. YANG, Z. ZHANG, X. CHEN, X. ZHANG, J. LV, F. LI, Y. ZHANG, G. ZENG, X. WANG, W. MA, G. HOU, S. HAO, C. LIU, Y. LAI, P. LIU, B. WANG, Y. LI, W. ZHANG, P. GAO, J. XIE, M. A. ESTEBAN, Y. GU, X. LIU, J. JI, T. QI, B. LIU, H. WANG, Y. ZHAO, X. YANG, X. WANG, R. CHEN, J. YANG, Y. YIN, J. WANG, Y. CAO, X. XU, L. LIU, X. JIN, and C. LIU. Chinese Immune Multi-Omics Atlas. *Science* **391**: eadt3130, 2026. (see pp. 93, 94)
- [313] S. YAZAR, J. ALQUICIRA-HERNANDEZ, K. WING, A. SENABOUTH, M. G. GORDON, S. ANDERSEN, Q. LU, A. ROWSON, T. R. P. TAYLOR, L. CLARKE, K. MACCORA, C. CHEN, A. L. COOK, C. J. YE, K. A. FAIRFAX, A. W. HEWITT, and J. E. POWELL. Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* **376**: eabf3041, 2022. (see pp. 93, 95)
- [314] GARVAN INSTITUTE OF MEDICAL RESEARCH \$27 million to map 50 million human cells and uncover genetic fingerprints of disease Press release May 2024 URL: <https://www.scienceinpublic.com.au/tenk10k/> (see pp. 93, 95)

- [315] NOETIK Generating Multimodal Human Data Company report; accessed 10 September 2026 URL: <https://www.noetik.ai/human-data> (visited on 09/10/2026) (see pp. 94, 95)
- [316] NOETIK TARIO-2: A Whole-Transcriptome Foundation Model from H&E Alone Company report; accessed 10 September 2026 URL: <https://www.noetik.blog/p/tario-2-a-whole-transcriptome-foundation> (visited on 09/10/2026) (see pp. 94, 95)
- [317] RECURSION PHARMACEUTICALS, INC. RxRx: Open-Science Data and Tools to Accelerate Machine Learning Research Company report; accessed 12 September 2026 URL: <https://www.rxr.x.ai/> (visited on 09/12/2026) (see pp. 94, 95)
- [318] XAIRA THERAPEUTICS Xaira Therapeutics launches X-Cell, its first virtual cell model, trained on the largest-ever genome-wide perturbation dataset, X-Atlas/Pisces Business Wire Mar. 2026 URL: <https://www.businesswire.com/news/home/20260317710096/en/Xaira-Therapeutics-Launches-X-Cell-Its-First-Virtual-Cell-Model-Trained-on-the-Largest-Ever-Genome-Wide-Perturbation-Dataset-X-AtlasPisces> (see pp. 94, 95)
- [319] OWKIN Owkin powers a new era in oncology research with MOSAIC – an unprecedented \$50 million spatial atlas of cancer cells Press release June 2023 URL: <https://www.owkin.com/newsfeed/owkin-powers-a-new-era-in-oncology-research-with-mosaic-an-unprecedented-50-million-spatial-atlas-of-cancer-cells> (see pp. 94, 95)
- [320] IGVF CONSORTIUM. Deciphering the impact of genomic variation on function. *Nature* **633**: 47–57, 2024. (see pp. 94, 95)
- [321] D. LI, S. LIU, P. R. ASSIS, M. LI, S. DONG, I. WHALING, O. JOLANKI, M. KAGDA, W. ZHANG, J. F. MACIAS-VELASCO, T. LIU, S. CODY, L. ANTONACCI-FULTON, Y. HUANG, J. LIU, M. T. MONTGOMERY, D. ZEIBERG, S. JAIN, V. PEJAVER, T. BERGQUIST, Y. CHEN, P. RADIVOJAC, C. A. GERSBACH, R. N. SHERPA, C. P. CASTRO, A. P. BOYLE, L. M. STARITA, D. M. FOWLER, N. AHITUV, K. K. DEY, W. H. MAJOROS, T. E. REDDY, M. CRAVEN, R. SINHA, Y. SVERCHKOV, X. CAI, M. Z. NZIMA, M. A. CALDERWOOD, J. ROZOWSKY, M. GERSTEIN, J. MA, F. YUE, J. M. CHERRY, M. I. LOVE, J. M. ENGREITZ, B. C. HITZ, and T. WANG. The IGVF catalog—from genetic variation to function. *Nucleic Acids Research* **54**: D1437–D1445, 2026. (see pp. 94, 95)
- [322] CHAN ZUCKERBERG INITIATIVE Billion Cells Project with 10x Genomics and Ultima Genomics 2025 URL: <https://chanzuckerberg.com/newsroom/billion-cells-project-launches-advance-ai-biology/> (visited on 09/12/2026) (see pp. 94, 95)
- [323] ILLUMINA Illumina introduces Billion Cell Atlas to accelerate AI and drug discovery Company report; accessed 12 September 2026 2026 URL: <https://www.illumina.com/company/news-center/press-releases/2026/fda84c92-b4b3-4691-a402-35555abe8605.html> (visited on 09/12/2026) (see pp. 94, 96)
- [324] ILLUMINA Illumina expands Billion Cell Atlas program with new AI drug developers Company report; accessed 12 September 2026 2026 URL: <https://investor.illumina.com/node/27191/pdf> (visited on 09/12/2026) (see p. 94)
- [325] O. ROZENBLATT-ROSEN. The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. *Cell* **181**: 236–249, 2020. (see p. 95)
- [326] RECURSION PHARMACEUTICALS, INC. Quarterly Report for the Period Ended March 31, 2026 Company report; accessed 12 September 2026 2026 URL: <https://www.sec.gov/Archives/edgar/data/1601830/000160183026000078/rxr.x-20260331.htm> (visited on 09/12/2026) (see p. 95)
- [327] RECURSION PHARMACEUTICALS, INC. Recursion Reports Second Quarter Financial Results; Genentech Options First Neuroscience Target into Early Discovery Program Company report; accessed 12 September 2026 2026 URL: <https://ir.recursion.com/news-releases/news-release-details/recursion-reports-second-quarter-financial-results-genentech/> (visited on 09/12/2026) (see p. 95)

- [328] A. J. RUBIN, K. R. PARKER, A. T. SATPATHY, Y. QI, B. WU, A. J. ONG, M. R. MUMBACH, A. L. JI, D. S. KIM, S. W. CHO, B. J. ZARNEGAR, W. J. GREENLEAF, H. Y. CHANG, and P. A. KHAVARI. Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks. *Cell* **176**: 361–376.e17, 2019. (see p. 96)
- [329] N. LISCOVITCH-BRAUER, A. MONTALBANO, J. DENG, A. MÉNDEZ-MANCILLA, H.-H. WESSELS, N. G. MOSS, C.-Y. KUNG, A. SOOKDEO, X. GUO, E. GELLER, S. JAINI, P. SMIBERT, and N. E. SANJANA. Profiling the genetic determinants of chromatin accessibility with scalable single-cell CRISPR screens. *Nature Biotechnology* **39**: 1270–1277, 2021. (see p. 96)
- [330] S. E. PIERCE, J. M. GRANJA, and W. J. GREENLEAF. High-throughput single-cell chromatin accessibility CRISPR screens enable unbiased identification of regulatory networks in cancer. *Nature Communications* **12**: 2969, 2021. (see p. 96)
- [331] R. P. WIERENGA, S. GOLAS, W. HO, C. COLEY, and K. M. ESVELT. PyLabRobot: An Open-Source, Hardware Agnostic Interface for Liquid-Handling Robots and Accessories. *bioRxiv* Preprint, 2023. (see p. 96)
- [332] F. V. DE ROP, J. N. ISMAIL, C. BRAVO GONZÁLEZ-BLAS, G. J. HULSELMANS, C. C. FLERIN, J. JANSSENS, K. THEUNIS, V. M. CHRISTIAENS, J. WOUTERS, G. MARCASSA, J. DE WIT, S. POOVATHINGAL, and S. AERTS. Hydrophobic Enables Droplet-Based Single-Cell ATAC-seq and Single-Cell RNA-seq Using Dissolvable Hydrogel Beads. *eLife* **11**: e73971, 2022. (see pp. 99, 102, 105)
- [333] F. V. DE ROP. Optimisation of the inDrop scRNA-seq Platform and Development of a scATAC-seq Protocol. *KU Leuven*, 2019. (see pp. 99, 100)
- [334] Y. XIA and G. M. WHITESIDES. Soft Lithography. *Angewandte Chemie International Edition* **37**: 550–575, 1998. (see p. 100)
- [335] R. M. MULQUEEN, B. A. DEROSA, C. A. THORNTON, Z. SAYAR, K. A. TORKENCZY, A. J. FIELDS, K. M. WRIGHT, X. NAN, R. RAMJI, F. J. STEEMERS, B. J. O’ROAK, and A. C. ADEY. Improved single-cell ATAC-seq reveals chromatin dynamics of in vitro corticogenesis. *bioRxiv*, 2020. (see p. 103)
- [336] R. R. STICKELS, E. MURRAY, P. KUMAR, J. LI, J. L. MARSHALL, D. J. DI BELLA, P. ARLOTTA, E. Z. MACOSKO, and F. CHEN. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nature Biotechnology*, 2020. (see p. 103)
- [337] F. V. DE ROP, G. HULSELMANS, C. FLERIN, P. SOLER-VILA, A. RAFELS, V. CHRISTIAENS, C. B. GONZÁLEZ-BLAS, D. MARCHESE, G. CARATÙ, S. POOVATHINGAL, O. ROZENBLATT-ROSEN, M. SLYPER, W. LUO, C. MUUS, F. DUARTE, R. SHRESTHA, S. T. BAGDATLI, M. R. CORCES, L. MAMANOVA, A. KNIGHTS, K. B. MEYER, R. MULQUEEN, A. TAHERINASAB, P. MASCHMEYER, J. PEZOLDT, C. L. G. LAMBERT, M. IGLESIAS, S. R. NAJLE, Z. Y. DOSSANI, L. G. MARTELOTTO, Z. BURKETT, R. LEBOSKY, J. I. MARTIN-SUBERO, S. PILLAI, A. SEBÉ-Ó, B. DEPLANCKE, S. A. TEICHMANN, L. S. LUDWIG, T. P. BRAUN, A. C. ADEY, W. J. GREENLEAF, J. D. BUEN-ROSTRO, A. REGEV, S. AERTS, and H. HEYN. Systematic Benchmarking of Single-Cell ATAC-sequencing Protocols. *Nature Biotechnology*, 1–11, 2023. (see p. 105)
- [338] P. DI TOMMASO, M. CHATZOU, E. W. FLODEN, P. P. BARJA, E. PALUMBO, and C. NOTREDAME. Nextflow enables reproducible computational workflows. *Nature Biotechnology* **35**: 316–319, 2017. (see p. 106)
- [339] N. OTSU. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics* **9**: 62–66, 1979. (see p. 106)
- [340] S. L. WOLOCK, R. LOPEZ, and A. M. KLEIN. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Systems* **8**: 281–291.e9, 2019. (see p. 106)
- [341] F. ZHANG, H. M. KANG, and Y. YUN. popsicle. *GitHub*, 2019. (see p. 106)
- [342] Y. ZHANG, T. LIU, C. A. MEYER, J. EECKHOUTE, D. S. JOHNSON, B. E. BERNSTEIN, C. NUSBAUM, R. M. MYERS, M. BROWN, W. LI, and X. S. LIU. Model-based analysis of ChIP-Seq (MACS). *Genome Biology* **9**: R137, 2008. (see p. 106)

-
- [343] S. AIBAR, C. B. GONZÁLEZ-BLAS, T. MOERMAN, V. A. HUYNH-THU, H. IMRICHOVA, G. HULSELMANS, F. RAMBOW, J.-C. MARINE, P. GEURTS, J. AERTS, J. VAN DEN OORD, Z. K. ATAK, J. WOUTERS, and S. AERTS. SCENIC: Single-Cell Regulatory Network Inference and Clustering. *Nature Methods* **14**: 1083–1086, 2017. (see p. 106)
- [344] I. KORSUNSKY, N. MILLARD, J. FAN, K. SLOWIKOWSKI, F. ZHANG, K. WEI, Y. BAGLAENKO, M. BRENNER, P. R. LOH, and S. RAYCHAUDHURI. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods* **16**: Epub 2019 Nov 18, 1289–1296, 2019. (see p. 107)