Survey of Dimensionality Reduction Techniques and their Applications for Classifying Disease State in Human RNA-seq Data

NELS BLAIR, ROYA CAMPOS, NOLAN MIDDLETON, PAUL OLA, JEHANZEB SALEEM

WASHINGTON STATE UNIVERSITY, PULLMAN, WA 99164

Introduction

Machine learning can answer biological questions

- TEC-LncMir: identifies RNA-RNA interactions [1]
- Random Forest: can identify TFBS with UV damage [2]
- AlphaFold: predicts protein structure [3]

Introduction

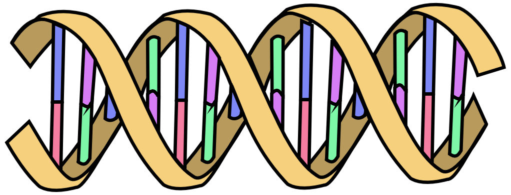
Machine learning can answer biological questions

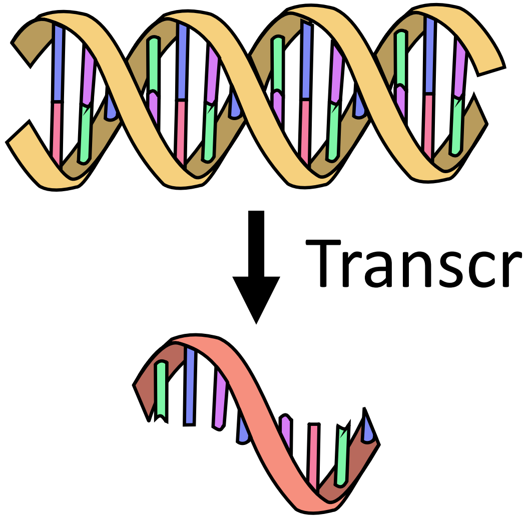
- TEC-LncMir: identifies RNA-RNA interactions [1]
- Random Forest: can identify TFBS with UV damage [2]
- AlphaFold: predicts protein structure [3]

We’re interested in gene expression data

- Informs how the cell is behaving
- Captures disease state, cell type, etc.
- Many diseases (cancers) result from aberrant gene expression

DNA (genes)

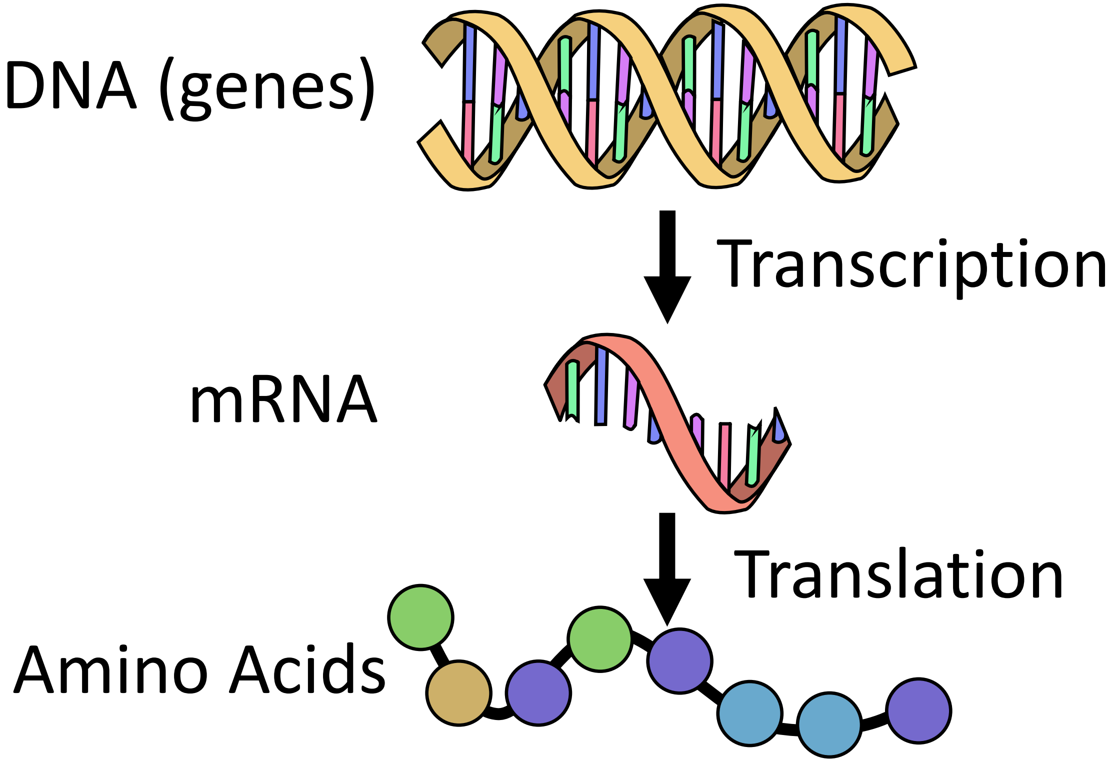




DNA (genes)

Transcription

mRNA



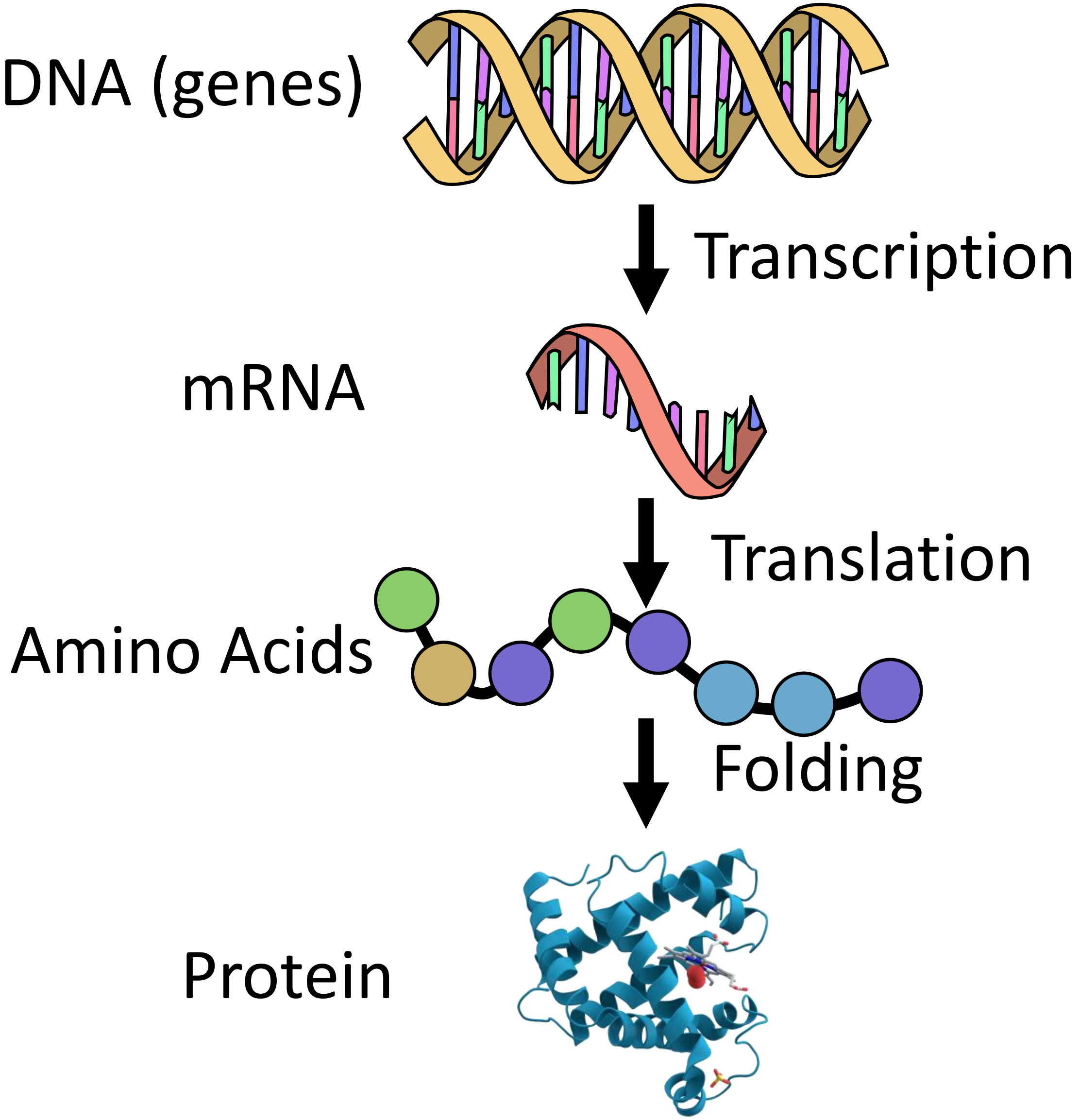
DNA (genes)

Transcription

mRNA

Translation

Amino Acids



DNA (genes)

Transcription

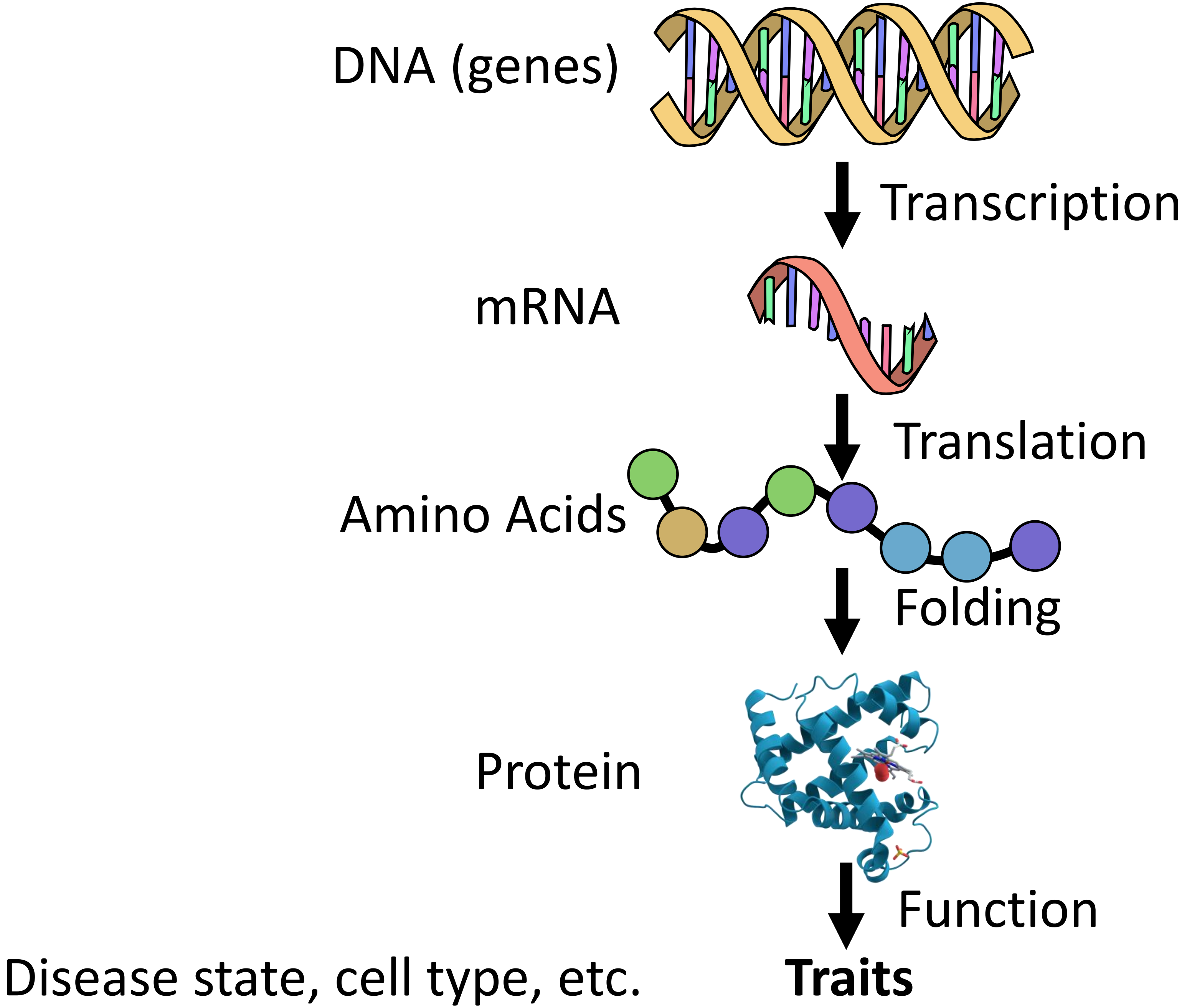
mRNA

Translation

Amino Acids

Folding

Protein



DNA (genes)

Transcription

mRNA

Translation

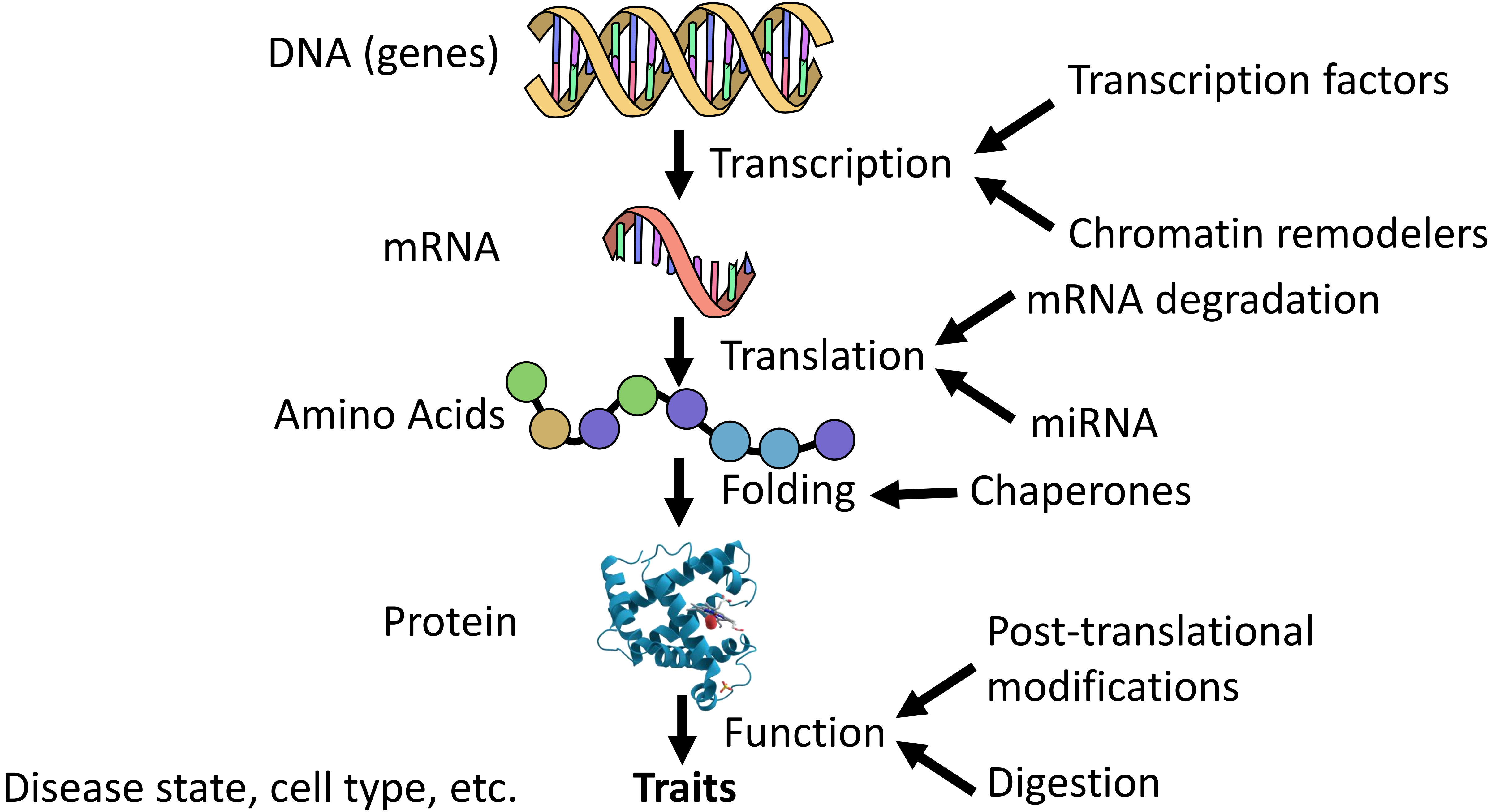
Amino Acids

Folding

Protein

Function

Disease state, cell type, etc. **Traits**



DNA (genes)

Transcription

Transcription factors

Chromatin remodelers

mRNA

Translation

mRNA degradation

miRNA

Amino Acids

Folding

Chaperones

Protein

Function

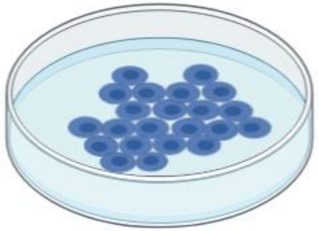
Post-translational modifications

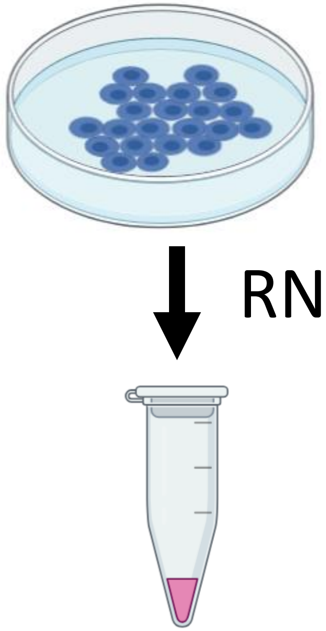
Digestion

Disease state, cell type, etc.

Traits

Cells

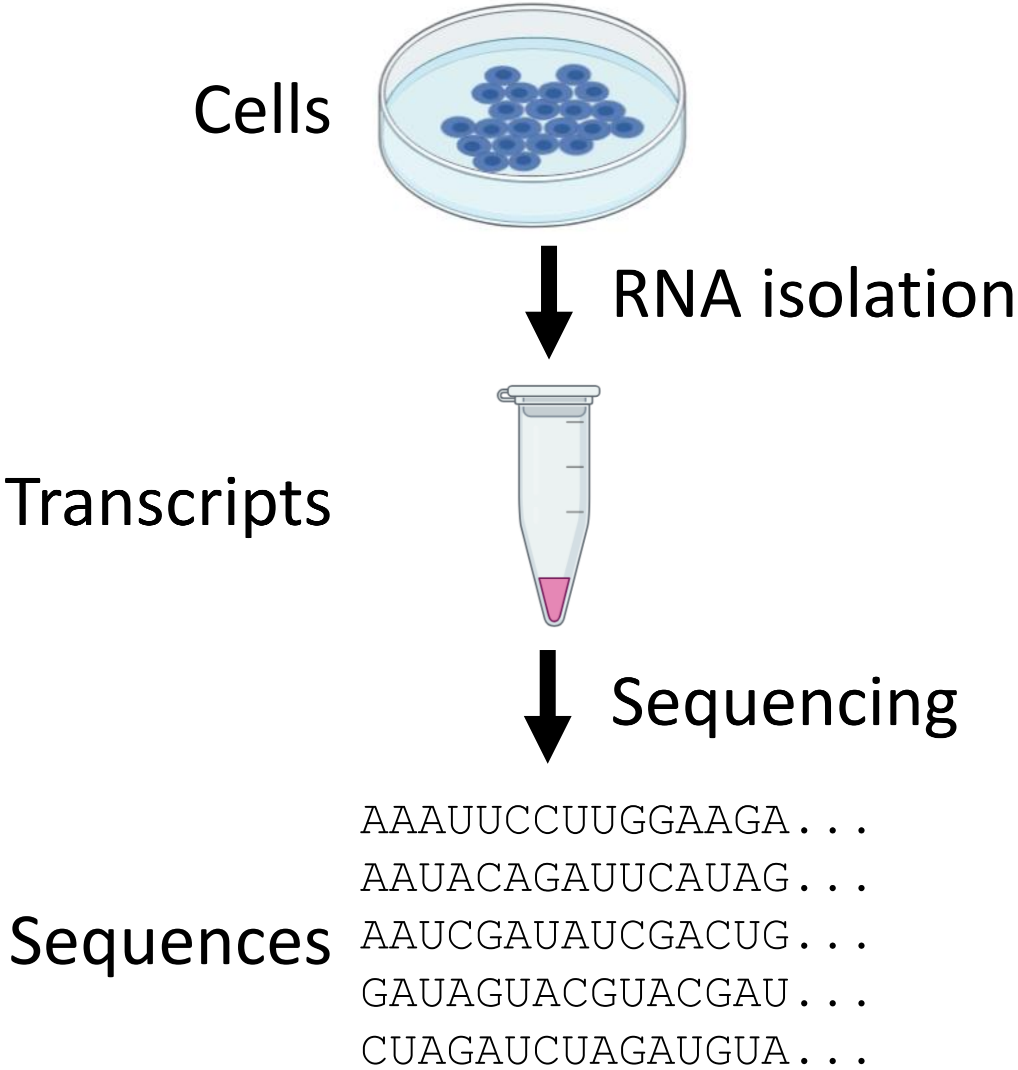




Cells

RNA isolation

Transcripts



Cells

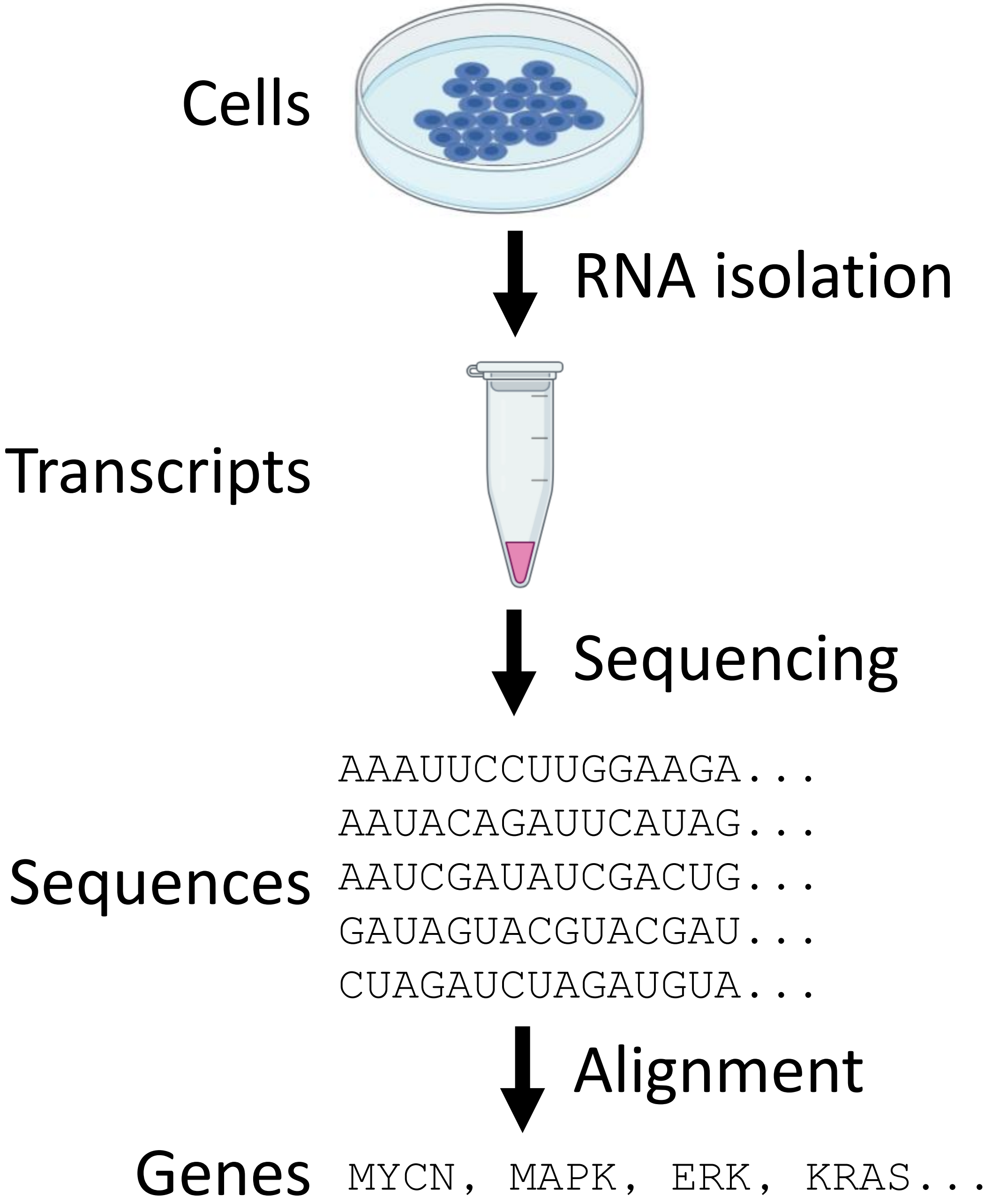
RNA isolation

Transcripts

Sequencing

Sequences

```
AAAUUCCUUGGAAGA...
AAUACAGAUUCAUAG...
AAUCGAUAUCGACUG...
GAUAGUACGUACGAU...
CUAGAUCUAGAUGUA...
```



Cells

RNA isolation

Transcripts

Sequencing

Sequences

AAAUUCCUUGGAAGA...
AAUACAGAUUCAUAG...
AAUCGAUAUCGACUG...
GAUAGUACGUACGAU...
CUAGAUCUAGAUGUA...

Alignment

Genes MYCN, MAPK, ERK, KRAS...

RNA-seq Data

Gene Expression Omnibus

- https://www.ncbi.nlm.nih.gov/sites/GDSbrowser

RNA-seq Data

Gene Expression Omnibus

- https://www.ncbi.nlm.nih.gov/sites/GDSbrowser

Very low sample sizes

- $\approx 10^1, 10^2$
- Especially for humans

RNA-seq Data

Gene Expression Omnibus

- https://www.ncbi.nlm.nih.gov/sites/GDSbrowser

Very low sample sizes

- $\approx 10^1, 10^2$
- Especially for humans

Very high dimensionality

- $\approx 10^4$

RNA-seq Data

Gene Expression Omnibus

- https://www.ncbi.nlm.nih.gov/sites/GDSbrowser

Very low sample sizes

- $\approx 10^1, 10^2$
- Especially for humans

Very high dimensionality

- $\approx 10^4$

Not coordinated

- Inconsistent gene sets
- Variations in methodology
- Different normalization

RNA-seq Data

Gene Expression Omnibus

- https://www.ncbi.nlm.nih.gov/sites/GDSbrowser

Very low sample sizes

- $\approx 10^1, 10^2$
- Especially for humans

Very high dimensionality

- $\approx 10^4$

Not coordinated

- Inconsistent gene sets
- Variations in methodology
- Different normalization

Challenging to learn on

- Lots of data out there
- Untapped potential

Pretrained Models?

No suitable pretrained models exist for RNA-seq data

- Most pretrained models utilize scRNA-seq data
- The few that use RNA-seq data are unsuitable

Pretrained Models?

No suitable pretrained models exist for RNA-seq data

- Most pretrained models utilize scRNA-seq data
- The few that use RNA-seq data are unsuitable

Modern deep learning methods require vast quantities of data

- Go to the single-cell level
- Utilize massive, coordinated data collection effort

Pretrained Models?

No suitable pretrained models exist for RNA-seq data

- Most pretrained models utilize scRNA-seq data
- The few that use RNA-seq data are unsuitable

Modern deep learning methods require vast quantities of data

- Go to the single-cell level
- Utilize massive, coordinated data collection effort

Older, simpler models are still commonplace

Survey of Dimensionality Reduction

4 dimensionality reductions :

- *a priori* aggregation
- PCA
- Kernelized PCA (kPCA)
- Nonnegative matrix factorization

5 machine learning models:

- Decision Tree (DT)
- k-Nearest Neighbors (k-NN)
- Naïve Bayes (NB)
- Random Forest (RF)
- Support Vector Machine (SVM)

a priori Aggregation

Genes are assigned categories and annotations

- Gene Ontology (GO) categorization [4,5]
- Component, function, and process

a priori Aggregation

Genes are assigned categories and annotations

- Gene Ontology (GO) categorization [4,5]
- Component, function, and process

Dimensionality reduction was done by taking the mean

- Features of reduced dataset correspond to gene groups
- Reduced dimension from $\approx 10^4 \rightarrow 10^3$

a priori Aggregation

Genes are assigned categories and annotations

- Gene Ontology (GO) categorization [4,5]
- Component, function, and process

Dimensionality reduction was done by taking the mean

- Features of reduced dataset correspond to gene groups
- Reduced dimension from $\approx 10^4 \rightarrow 10^3$

Retains biological context

PCA, kPCA, and NMF

Reduce dimension to $d = 2, \dots, 9$

- $d = 2$ allows for visualization
- $d = 9$ is still small compared to the number of data points

PCA and kPCA re-center the data, while NMF does not

All three methods lose biological tractability

- Reduced dimensions have no biological meaning

	GEO					
Dataset	Accession	Title	n	d	k	Ref.
UCC	GDS1615	Ulcerative colitis and Crohn's disease comparison: peripheral blood mononuclear cells	127	22283	3	[6]
SCLC	GDS2373	Squamous cell lung carcinomas	130	22283	6	[7]
SEC	GDS2771	Large airway epithelial cells from cigarette smokers with suspect lung cancer	192	22283	3	[8]
MDS	GDS3795	Myelodysplastic syndrome: CD34+ hematopoietic stem cells	200	54675	2	[9]
ALL	GDS4206	Pediatric acute leukemia patients with early relapse: white blood cells	197	54675	3	[10]
HIV	GDS4228	HIV infection and Antiretroviral Therapy effects on mitochondria in various tissues	166	4825	2	[11]
JIA	GDS4267	Systemic juvenile idiopathic arthritis and non-systemic JIA subtypes: peripheral blood mononuclear cells	154	54675	3	[12]
GBM	GDS5205	Long-term adult survivors of glioblastoma: primary tumors	70	54675	3	[13]
MDG	GDS963	Macular degeneration and dermal fibroblast response to sublethal oxidative stress	36	12625	2	[14]

Methodological Notes

All reported accuracy values are from leave-one-out validation

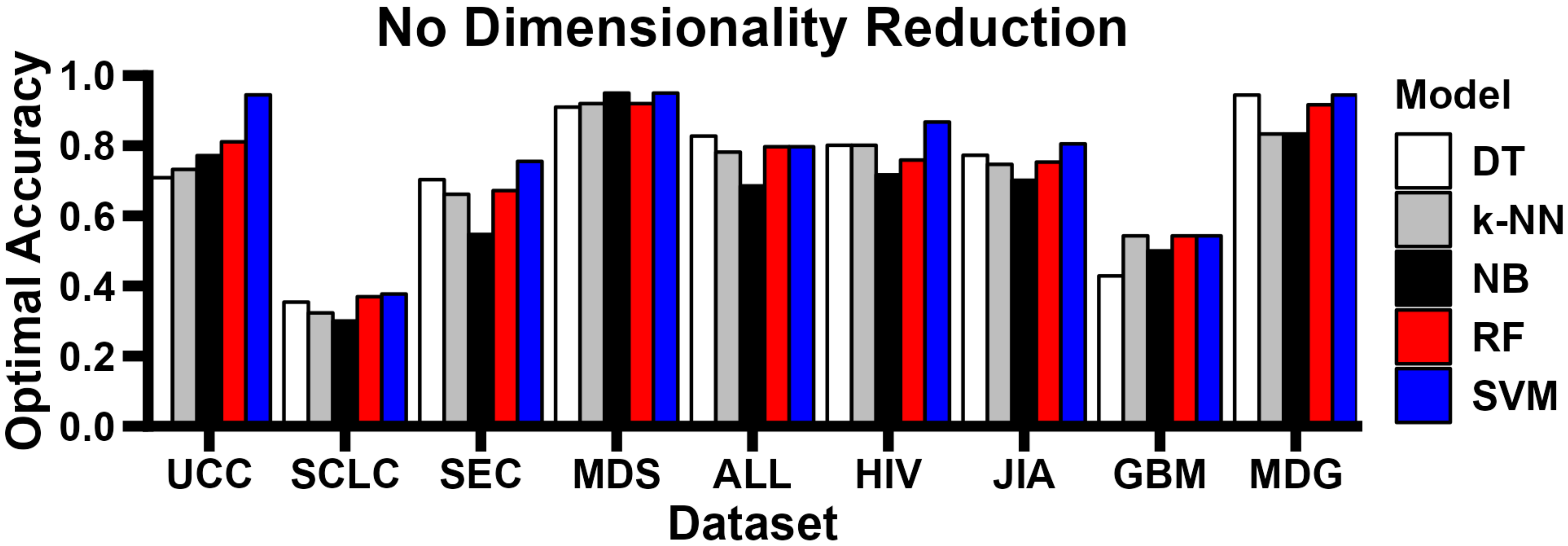
- Small dataset sizes make traditional cross-validation infeasible

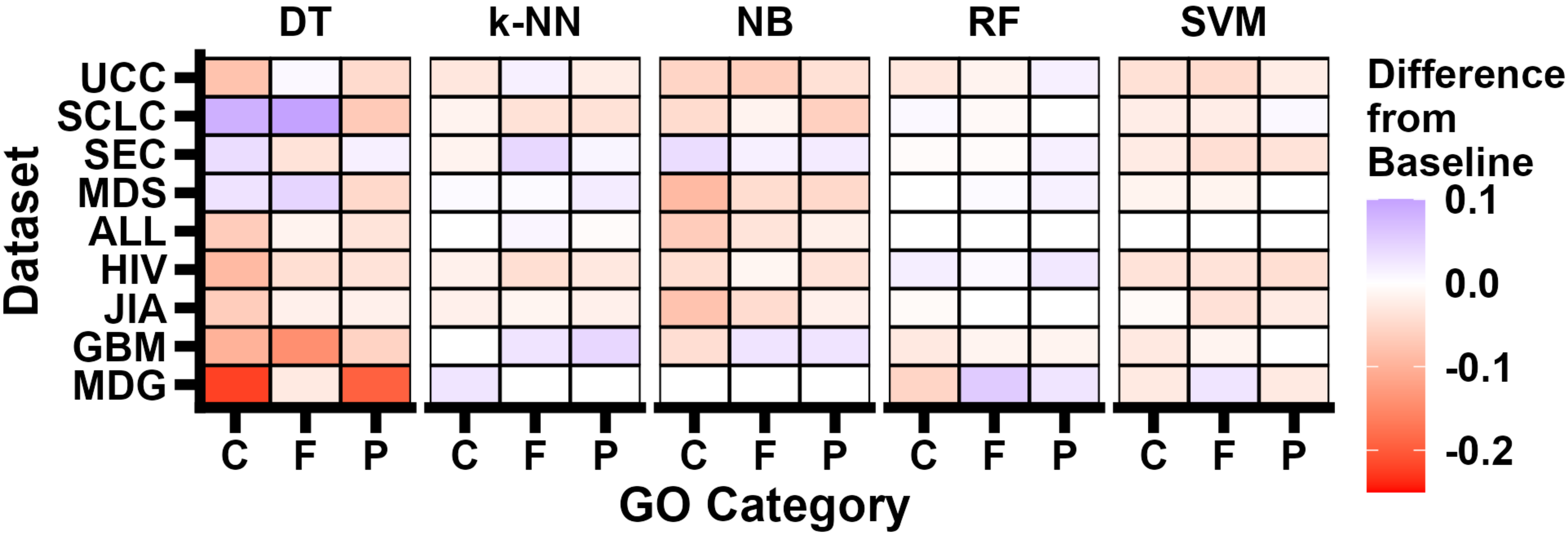
Some dataset features do not correspond to genes

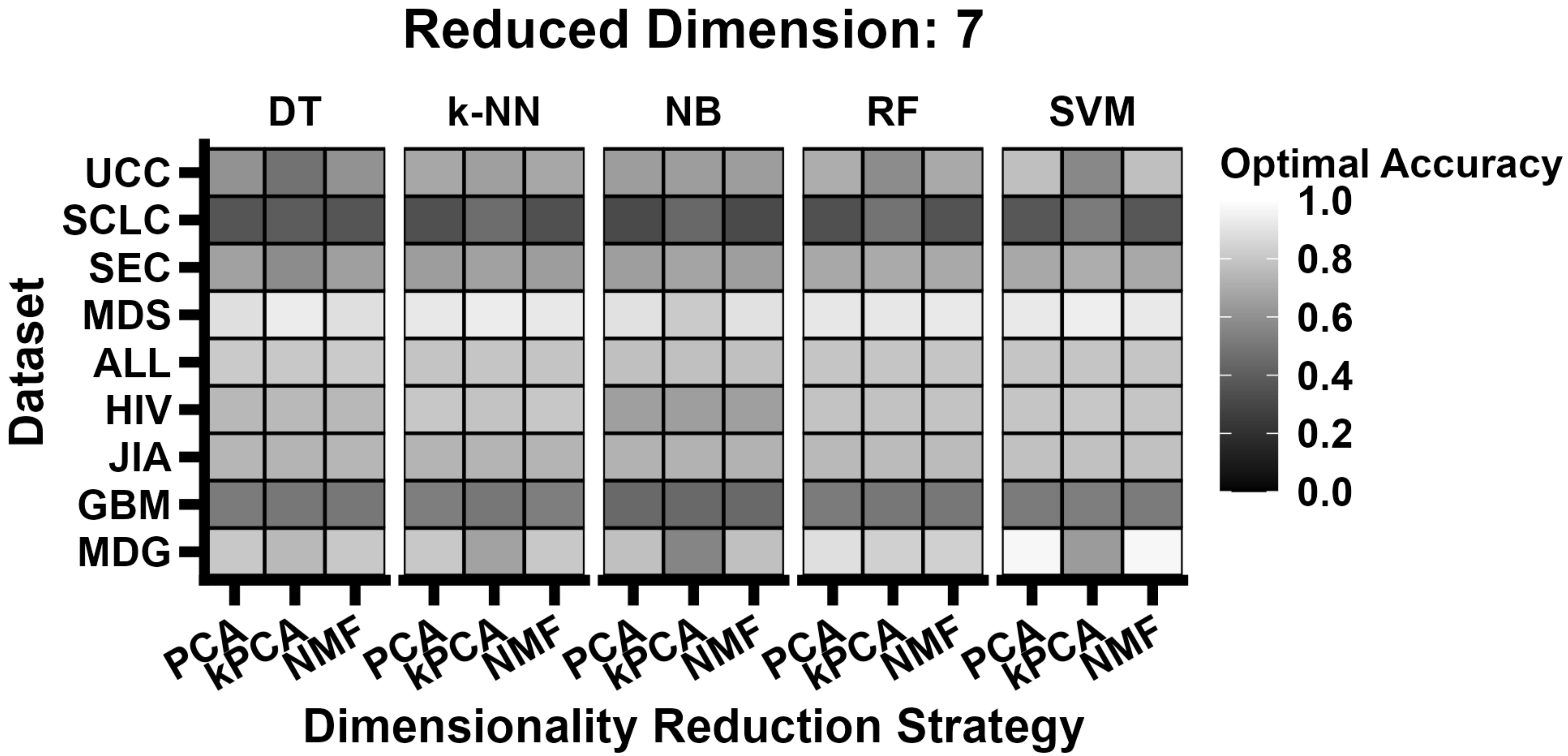
- Sequencing platform controls, etc.
- These were removed

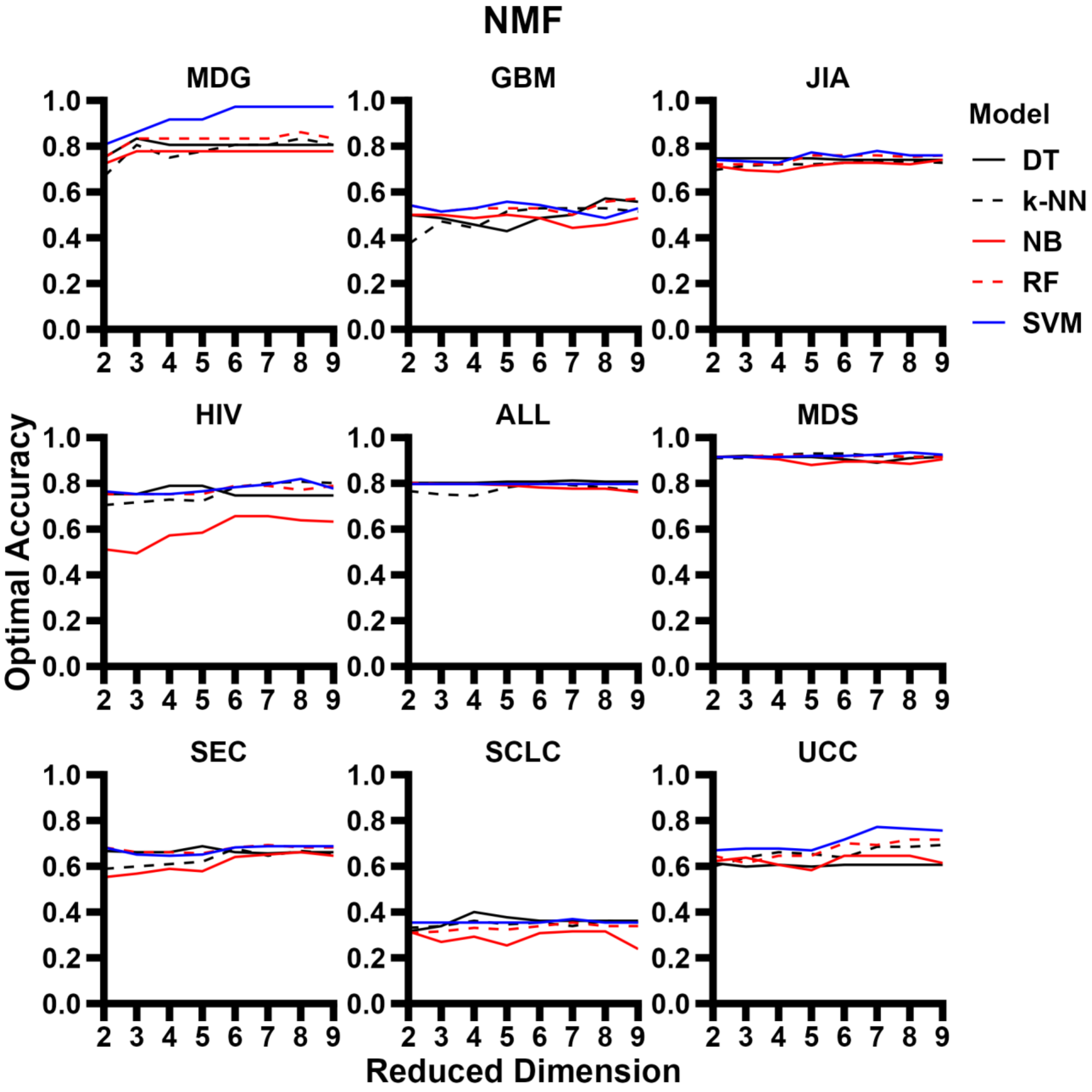
All data/code is available on GitHub:

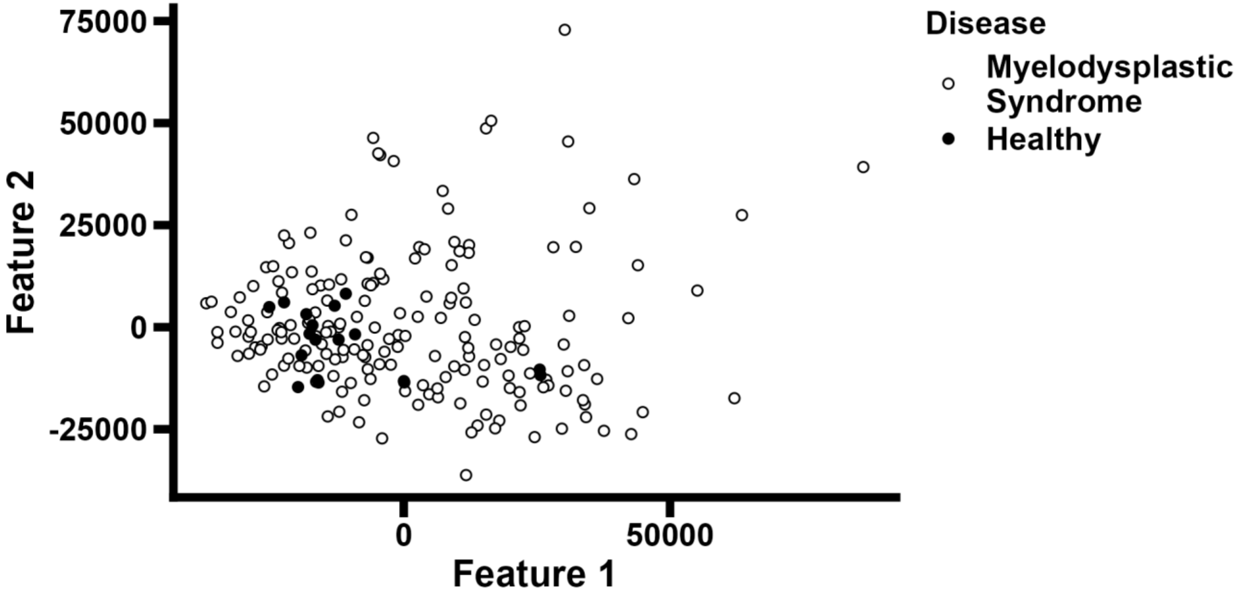
- https://github.com/nolan-middleton/CPT_S-Group-Project-Fall-2025

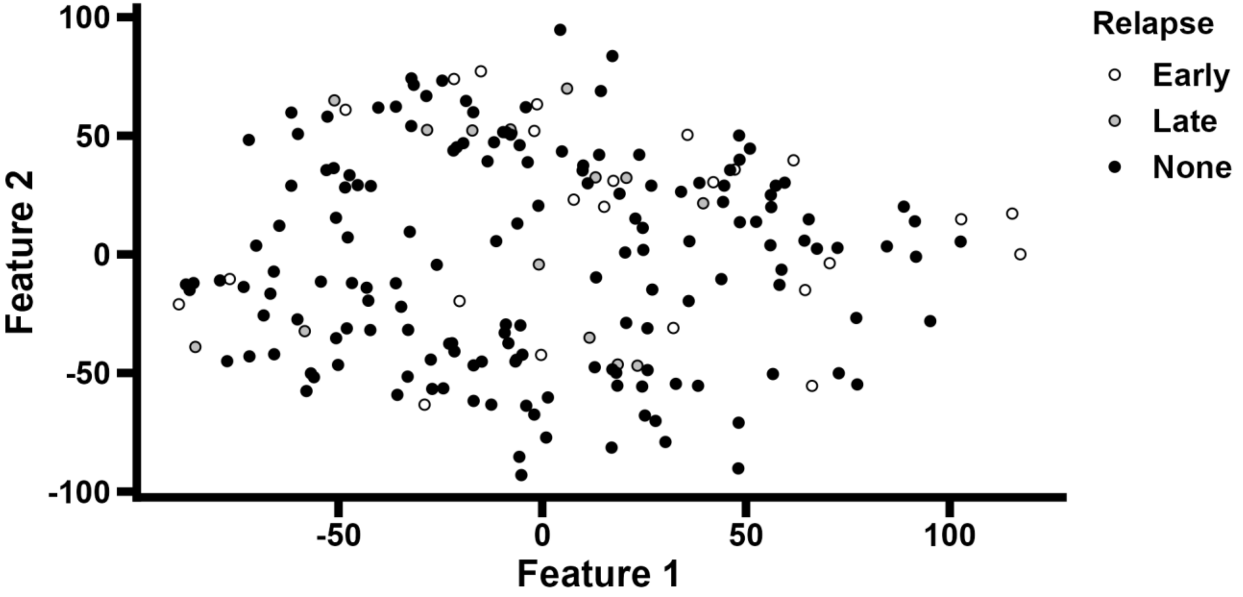

No Dimensionality Reduction
Optimal Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
UCC
SCLC
SEC
MDS
ALL
HIV
JIA
GBM
MDG
Dataset
Model
DT
k-NN
NB
RF
SVM

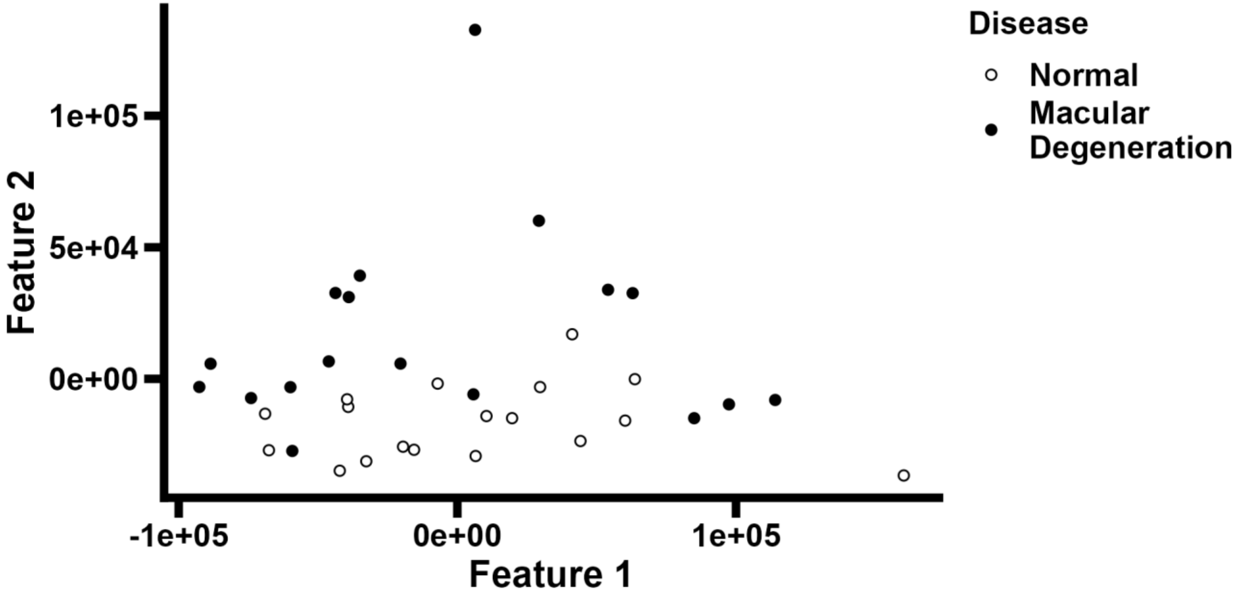

DT
k-NN
NB
RF
SVM
Dataset
UCC
SCLC
SEC
MDS
ALL
HIV
JIA
GBM
MDG
C
F
P
GO Category
Difference from Baseline
0.1
0.0
-0.1
-0.2

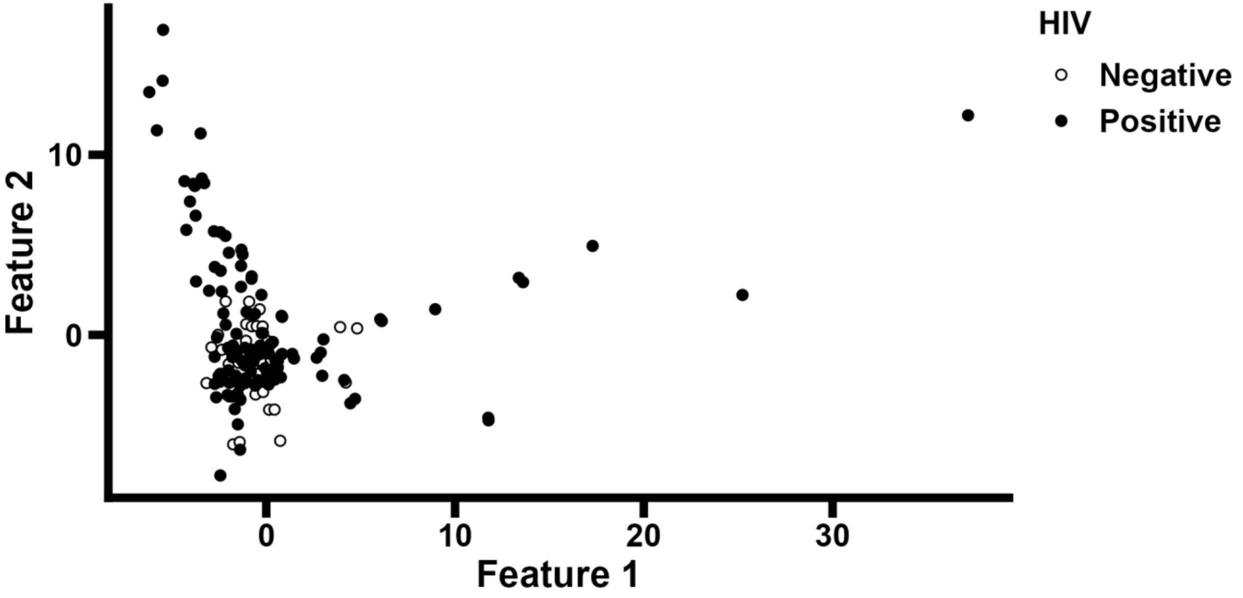

Reduced Dimension: 7
DT
k-NN
NB
RF
SVM
Optimal Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
Dataset
UCC
SCLC
SEC
MDS
ALL
HIV
JIA
GBM
MDG
PCA
KPCA
NMF
Dimensionality Reduction Strategy

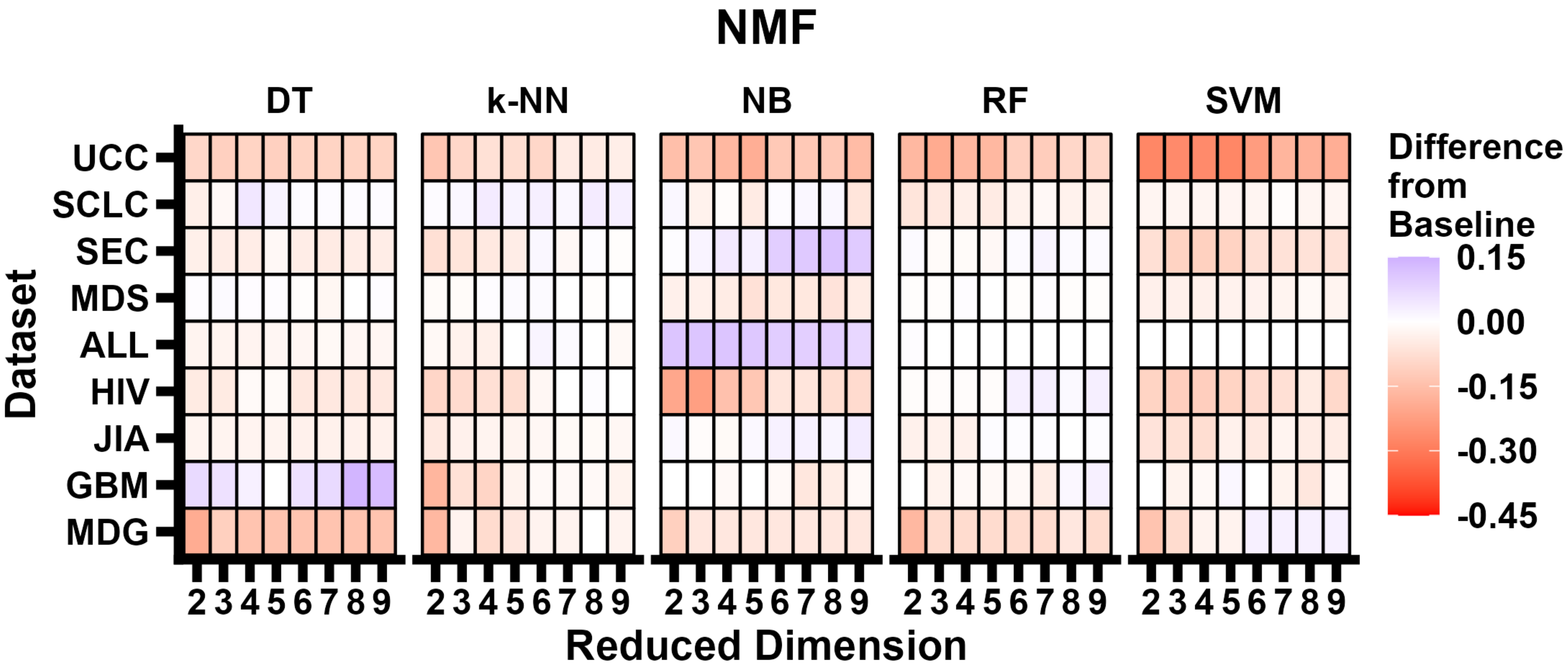

NMF
MDG
GBM
JIA
HIV
ALL
MDS
SEC
SCLC
UCC
Optimal Accuracy
Reduced Dimension
0.0
0.2
0.4
0.6
0.8
1.0
2 3 4 5 6 7 8 9
Model
DT
k-NN
NB
RF
SVM

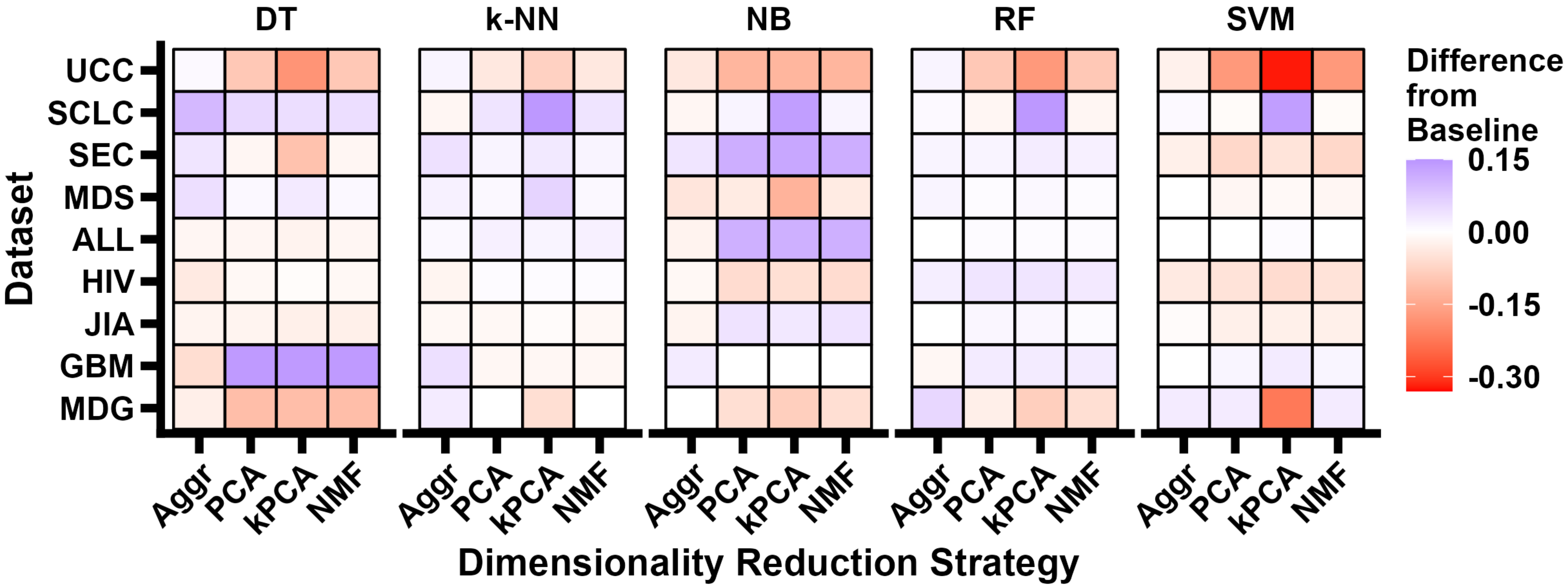

Feature 2
75000
50000
25000
0
-25000
0
50000
Feature 1
Disease
○ Myelodysplastic Syndrome
● Healthy


Feature 2
100
50
0
-50
-100
-50
0
50
100
Feature 1
Relapse
○ Early
● Late
● None


Feature 2
1e+05
5e+04
0e+00
-1e+05
0e+00
1e+05
Feature 1
Disease
○ Normal
● Macular Degeneration


Feature 2
10
0
0
10
20
30
Feature 1
HIV
○ Negative
● Positive


NMF
DT
k-NN
NB
RF
SVM
Dataset
UCC
SCLC
SEC
MDS
ALL
HIV
JIA
GBM
MDG
2 3 4 5 6 7 8 9
Reduced Dimension
Difference from Baseline
0.15
0.00
-0.15
-0.30
-0.45


DT
k-NN
NB
RF
SVM
Dataset
UCC
SCLC
SEC
MDS
ALL
HIV
JIA
GBM
MDG
Aggr
PCA
kPCA
NMF
Dimensionality Reduction Strategy
Difference from Baseline
0.15
0.00
-0.15
-0.30

Conclusions

The impacts of dimensionality reduction are varied

- Can help and harm classification performance
- Highly variable across datasets, dimension, and classification model

Conclusions

The impacts of dimensionality reduction are varied

- Can help and harm classification performance
- Highly variable across datasets, dimension, and classification model

Likely depends on:

- Underlying biological context
- Spatial features of dataset
- Dataset parameters (number of points, number of classes, etc.)

References

1. Tingpeng Yang, Yonghong He, and Yu Wang. Introducing tec-lncmir for prediction of lncrna-mirna interactions through deep learning of rna sequences. Brief Bioinform, 26(1):bbaf046, 02 2025. ISSN 1477-4054. doi: 10.1093/bib/bbaf046. URL https://doi.org/10.1093/bib/bbaf046.
2. Hannah E Wilson, Scott Stevison, Levi Lamprey, and John J Wyrick. Mapping transcription factor binding sites by learning uv damage fingerprints. Nucleic Acids Res, 53(19):gkaf1014, 10 2025. ISSN 1362-4962. doi: 10.1093/nar/gkaf1014. URL https://doi.org/10.1093/nar/gkaf1014.
3. John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873):583–589, August 2021. doi: 10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.
4. M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. Nat Genet, 25(1):25–29, May 2000.
5. The Gene Ontology Consortium, Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, David P Hill, Raymond Lee, Huaiyu Mi, Sierra Moxon, Christopher J Mungall, Anushya Muruganugan, Tremayne Mushayahama, Paul W Sternberg, Paul D Thomas, Kimberly Van Auken, Jolene Ramsey, Deborah A Siegele, Rex L Chisholm, Petra Fey, Maria Cristina Aspromonte, Maria Victoria Nugnes, Federica Quaglia, Silvio Tosatto, Michelle Giglio, Suvarna Nadendla, Giulia Antonazzo, Helen Attrill, Gil dos Santos, Steven Marygold, Victor Strelets, Christopher J Tabone, Jim Thurmond, Pinglei Zhou, Saadullah H Ahmed, Praoparn Asanitthong, Diana Luna Buitrago, Meltem N Erdol, Matthew C Gage, Mohamed Ali Kadhum, Kan Yan Chloe Li, Miao Long, Aleksandra Michalak, Angeline Pesala, Armalya Pritazahra, Shirin C C Saverimuttu, Renzhi Su, Kate E Thurlow, Ruth C Lovering, Colin Logie, Snezhana Oliferenko, Judith Blake, Karen Christie, Lori Corbani, Mary E Dolan, Harold J Drabkin, David P Hill, Li Ni, Dmitry Sitnikov, Cynthia Smith, Alayne Cuzick, James Seager, Laurel Cooper, Justin Elser, Pankaj Jaiswal, Parul Gupta, Pankaj Jaiswal, Sushma Naithani, Manuel Lera-Ramirez, Kim Rutherford, Valerie Wood, Jeffrey L De Pons, Melinda R Dwinell, G Thomas Hayman, Mary L Kaldunski, Anne E Kwitek, Stanley J F Laulederkind, Marek A Tutaj, Mahima Vedi, Shur-Jen Wang, Peter D'Eustachio, Lucila Aimo, Kristian Axelsen, Alan Bridge, Nevila Hyka-Nouspikel, Anne Morgat, Suzi A Aleksander, J Michael Cherry, Stacia R Engel, Kalpana Karra, Stuart R Miyasato, Robert S Nash, Marek S Skrzypek, Shuai Weng, Edith D Wong, Erika Bakker, Tanya Z Berardini, Leonore Reiser, Andrea Auchincloss, Kristian Axelsen, Ghislaine Argoud-Puy, Marie-Claude Blatter, Emmanuel Boutet, Lionel Breuza, Alan Bridge, Cristina Casals-Casas, Elisabeth Coudert, Anne Estreicher, Maria Livia Famiglietti, Marc Feuermann, Arnaud Gos, Nadine Gruaz-Gumowski, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Ivo Pedruzzi, Lucille Pourcel, Sylvain Poux, Catherine Rivoire, Shyamala Sundaram, Alex Bateman, Emily Bowler-Barnett, Hema Bye-A-Jee, Paul Denny, Alexandr Ignatchenko, Rizwan Ishtiaq, Antonia Lock, Yvonne Lussi, Michele Magrane, Maria J Martin, Sandra Orchard, Pedro Raposo, Elena Speretta, Nidhi Tyagi, Kate Warner, Rossana Zaru, Alexander D Diehl, Raymond Lee, Juancarlos Chan, Stavros Diamantakis, Daniela Raciti, Magdalena Zarowiecki, Malcolm Fisher, Christina James-Zorn, Virgilio Ponferrada, Aaron Zorn, Sridhar Ramachandran, Leyla Ruzicka, and Monte Westerfield. The gene ontology knowledgebase in 2023. Genetics, 224(1):iyad031, 03 2023. ISSN 1943-2631. doi: 10.1093/genetics/iyad031. URL https://doi.org/10.1093/genetics/iyad031.
6. Michael E Burczynski, Ron L Peterson, Natalie C Twine, Krystyna A Zuberek, Brendan J Brodeur, Lori Casciotti, Vasu Maganti, Padma S Reddy, Andrew Strahs, Fred Immermann, Walter Spinelli, Ulrich Schwertschlag, Anna M Slager, Monette M Cotreau, and Andrew J Dorner. Molecular classification of crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. J Mol Diagn, 8(1):51–61, February 2006.
7. Mitch Raponi, Yi Zhang, Jack Yu, Guoan Chen, Grace Lee, Jeremy M G Taylor, James Macdonald, Dafydd Thomas, Christopher Moskaluk, Yixin Wang, and David G Beer. Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. Cancer Res, 66(15):7466–7472, August 2006.
8. Adam M Gustafson, Raffaella Soldi, Christina Anderlind, Mary Beth Scholand, Jun Qian, Xiaohui Zhang, Kendal Cooper, Darren Walker, Annette McWilliams, Gang Liu, Eva Szabo, Jerome Brody, Pierre P Massion, Marc E Lenburg, Stephen Lam, Andrea H Bild, and Avrum Spira. Airway PI3K pathway activation is an early and reversible event in lung cancer development. Sci Transl Med, 2(26):26ra25, April 2010.
9. Petra Gorombei, Fabien Guidez, Saravanan Ganesan, Mathieu Chiquet, Andrea Pellagatti, Laure Goursaud, Nilgun Tekin, Stephanie Beurlet, Satyananda Patel, Laura Guerenne, Carole Le Pogam, Niclas Setterblad, Pierre de la Grange, Christophe LeBoeuf, Anne Janin, Maria-Elena Noguera, Laure Sarda-Mantel, Pascale Merlet, Jacqueline Boultwood, Marina Konopleva, Michael Andreeff, Robert West, Marika Pla, Lionel Adès, Pierre Fenaux, Patricia Krief, Christine Chomienne, Nader Omidvar, and Rose Ann Padua. BCL-2 inhibitor ABT-737 effectively targets leukemia-initiating cells with differential regulation of relevant genes leading to extended survival in a NRAS/BCL-2 mouse model of high risk-myelodysplastic syndrome. Int J Mol Sci, 22(19):10658, September 2021.
10. Lüder Hinrich Meyer, Sarah Mirjam Eckhoff, Manon Queudeville, Johann Michael Kraus, Marco Giordan, Jana Stursberg, Andrea Zangrando, Elena Vendramini, Anja Möricke, Martin Zimmermann, Andre Schrauder, Georgia Lahr, Karlheinz Holzmann, Martin Schrappe, Giuseppe Basso, Karsten Stahnke, Hans Armin Kestler, Geertruy Te Kronnie, and Klaus- Michael Debatin. Early relapse in ALL is identified by time to leukemia in NOD/SCID mice and is characterized by a gene signature involving survival pathways. Cancer Cell, 19(2): 206–217, February 2011.
11. Caryn G Morse, Joachim G Voss, Goran Rakocevic, Mary McLaughlin, Carol L Vinton, Charles Huber, Xiaojun Hu, Jun Yang, Da Wei Huang, Carolea Logun, Robert L Danner, Zoila G Rangel, Peter J Munson, Jan M Orenstein, Elisabeth J Rushing, Richard A Lempicki, Marinos C Dalakas, and Joseph A Kovacs. HIV infection and antiretroviral therapy have divergent effects on mitochondria in adipose tissue. J Infect Dis, 205(12):1778–1787, June 2012.
12. Claas H Hinze, Ndate Fall, Sherry Thornton, Jun Q Mo, Bruce J Aronow, Gerlinde Layh-Schmitt, Thomas A Griffin, Susan D Thompson, Robert A Colbert, David N Glass, Michael G Barnes, and Alexei A Grom. Immature cell populations and an erythropoiesis gene-expression signature in systemic juvenile idiopathic arthritis: implications for pathogenesis. Arthritis Res Ther, 12(3):R123, June 2010.

References

13. Guido Reifenberger, Ruthild G Weber, Vera Riehmer, Kerstin Kaulich, Edith Willscher, Henry Wirth, Jens Gietzelt, Bettina Hentschel, Manfred Westphal, Matthias Simon, Gabriele Schackert, Johannes Schramm, Jakob Matschke, Michael C Sabel, Dorothee Gramatzki, Jörg Felsberg, Christian Hartmann, Joachim P Steinbach, Uwe Schlegel, Wolfgang Wick, Bernhard Radlwimmer, Torsten Pietsch, Jörg C Tonn, Andreas von Deimling, Hans Binder, Michael Weller, Markus Loeffler, and German Glioma Network. Molecular characterization of long-term survivors of glioblastoma using genome- and transcriptome-wide profiling. Int J Cancer, 135(8):1822–1831, October 2014.
14. Nataly Strunnikova, Sara Hilmer, Jessica Flippin, Michael Robinson, Eric Hoffman, and Karl G Csaky. Differences in gene expression profiles in dermal fibroblasts from control and patients with age-related macular degeneration elicited by oxidative injury. Free Radic Biol Med, 39(6): 781–796, September 2005.
15. Images taken from BioRender. https://biorender.com.
16. Jocelyn Krebs, Elliott Goldstein, and Stephen Kilpatrick. Lewin’s Genes XII. Jones & Bartlett Learning, 5 Wall Street, Burlington, MA 01803, 12 edition, 2018. ISBN 9781284104493.
17. Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet, 10(1):57–63, January 2009. doi: 10.1038/nrg2484. URL https://doi.org/10.1038/nrg2484.
18. Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, Kaiqin Lao, and M Azim Surani. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods, 6(5):377–382, May 2009.
19. Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods, 21(8):1470–1480, August 2024. doi: 10.1038/s41592-024-02201-0. URL https://doi.org/10.1038/s41592-024-02201-0.
20. Shreshth Gandhi, Farnoosh Javadi, Valentine Svensson, Umair Khan, Matthew G. Jones, Johnny Yu, Daniele Merico, Hani Goodarzi, and Nima Alidoust. Tahoe-x1: Scaling perturbation-trained single-cell foundation models to 3 billion parameters. bioRxiv, 2025. doi: 10.1101/2025.10.23. 683759. URL https://www.biorxiv.org/content/10.1101/2025.10.23.683759v1.
21. Zhuoyi Wei Yun Su Ningyuan Shangguan Shuangyu Yang Chengyang Zhang Wenbing Li Jinbo Zhang Nan Fang Hongyu Zhang Huiying Zhao Yutong Lu Jue Fan Weijiang Yu Yuansong Zeng, Jiancong Xie and Yuedong Yang. Cellfm: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells. Nat Commun, 16(4679), May 2025. doi: 10.1038/s41467-025-59926-5. URL https://doi.org/10.1038/s41467-025-59926-5.
22. Tracy Boakye Serebour, Adam P. Cribbs, Mathew J. Baldwin, Collen Masimirembwa, Zedias Chikwambi, Angeliki Kerasidou, and Sarah J. B. Snelling. Overcoming barriers to single-cell rna sequencing adoption in low- and middle-income countries. Eur J Hum Genet, 32(10):1206–1213, October 2024. doi: 10.1038/s41431-024-01564-4. URL https://doi.org/10.1038/s41431-024-01564-4.
23. Maxence Gélard, Guillaume Richard, Thomas Pierrot, and Paul-Henry Cournède. Bulkrnabert: Cancer prognosis from bulk rna-seq based language models. In Stefan Hegselmann, Helen Zhou, Elizabeth Healey, Trenton Chang, Caleb Ellington, Vishwali Mhasawade, Sana Tonekaboni, Peniel Argaw, and Haoran Zhang, editors, Proceedings of the 4th Machine Learning for Health Symposium, volume 259 of Proceedings of Machine Learning Research, pages 384–400. PMLR, 15–16 Dec 2025. URL https://proceedings.mlr.press/v259/gelard25a.html.
24. Liu D Wang N He D Wu Z Zhu X Wen X Li X Li J Wang Z Wang X, Wang H. Deep learning using bulk rna-seq data expands cell landscape identification in tumor microenvironment. Oncoimmunology, 25(11), February 2022. doi: 10.1080/2162402X.2022.2043662.
25. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.
26. Derek A. Pisner and David M. Schnyer. Chapter 6 - support vector machine. In Andrea Mechelli and Sandra Vieira, editors, Machine Learning, pages 101–121. Academic Press, 2020. ISBN 978-0-12-815739-8. doi: https://doi.org/10.1016/B978-0-12-815739-8.00006-7. URL https://www.sciencedirect.com/science/article/pii/B9780128157398000067.
27. Musalula Sinkala. Mutational landscape of cancer-driver genes across human cancers. Sci Rep, 13(1), August 2023. doi: 10.1038/s41598-023-39608-2. URL https://doi.org/10.1038/s41598-023-39608-2.
28. Ales Varabyou, Steven L Salzberg, and Mihaela Pertea. Effects of transcriptional noise on estimates of gene and transcript expression in RNA sequencing experiments. Genome Res, 31 (2):301–308, February 2021.