Studies of mRNA expression and degradation

Thesis by
Ángel Gálvez Merchán

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2023
Defended June 5th, 2023

Ángel Gálvez Merchán
ORCID: 0000-0001-7420-8697

ACKNOWLEDGEMENTS

My PhD has been a long journey, full of learning, efforts, and fun. I feel lucky to have shared these years with many wonderful people that have made this experience so enjoyable, and that have helped me grow so much as a scientist and as a person.

I would like to express my heartfelt gratitude to my two advisors. I was the first person to join Rebecca's lab, and working side-by-side with such an incredible scientist has been an inspiring and transformative experience. I will always remember all that you taught me, your kindness and your patience. Lior welcomed me into his group when I was going through difficult times, and has been an extraordinary mentor ever since. Working with you has been so fun, and I have learnt so much from you. I entered your group as a biochemist, and I am leaving today as a computational biologist. I owe that to your constant support and mentoring. Thank you for all you have given me.

I would also like to thank my committee members, Alexei Aravin and Matt Thomson, for their constant guidance and help in the many projects I embarked on. In addition, I have been very fortunate to build great relationships with many faculty members at Caltech, who helped me in my PhD path when I felt lost, and were always there for me. Thank you Dianne Newman, Lea Goentoro, Ray Deshaies, Bill Dunphy and Carlos Lois. I want to give a special thanks to Kata Fejes Toth, an incredible scientist and one of my biggest supports at Caltech. You always believed in me, and I will always be thankful for all you have done for me. Thank you.

I have been very fortunate to work with an incredible group of people in both of my laboratories. I have had so much fun and learnt so much from everyone at Rebecca and Lior's lab. Besides co-workers, mentors, and collaborators, they have become friends that I will keep in my heart forever. Thank you for everything, and I will greatly miss you.

I made many wonderful friends at Caltech outside my labs, and I am so thankful for all the great times we have spent together. You have all been so important to me. I would also like to thank all my friends from Spain, who have given me constant support in this long process and have helped me soothe my home-sickness.

I would also like to thank my family. Thank you to my parents and my sister for all the love and support throughout my life. I wouldn't have made it to Caltech and through my PhD without you. I want to thank all my cousins, for being such a big

source of happiness and fun throughout my childhood and up until today. I would like to thank my grandparents, most of whom are not here today, but that would feel so proud of what I have achieved.

Finally, I want to thank Carmen. You have been a pillar in my life ever since I met you. Our PhD was not easy, and it would have never made it without you. Thank you for the constant support, for all our moments together, and for being so amazing. I love you. Last but not least, I would like to thank Elvis, for always wagging his tail when I get home and for being the best dog in the world.

PUBLISHED CONTENT AND CONTRIBUTIONS

Gálvez-Merchán, Ángel et al. (2023). "Metadata retrieval from sequence databases with ffq." In: *Bioinformatics* 39.1, btac667. DOI: 10.1093/bioinformatics/btac667. URL: https://academic.oup.com/bioinformatics/article/39/1/btac667/6971839.

This publication is the basis for Chapter 4. A.G.M participated in the development of ffq and in the writing of the paper.

Inglis, Alison J et al. (2023). "Coupled protein quality control during nonsense-mediated mRNA decay". In: *Journal of Cell Science* 136.10. DOI: 10.1242/jcs.261216. URL: https://journals.biologists.com/jcs/article/136/10/jcs261216/310674.

This publication is the basis for Chapter 2. A.G.M participated in the conception of the project, built the fluorescent reporters, participated in the experiments to discover and validate the protein degradation pathway, and performed the whole-genome CRISPRi screen.

Booeshaghi, A. Sina et al. (2022). "Depth normalization for single-cell genomics count data." In: *bioRxiv*, pp. 2022–05. DOI: 10.1101/2022.05.06.490859. URL: https://www.biorxiv.org/content/10.1101/2022.05.06.490859v1.abstract.

This publication is the basis for chapter 3. A.G.M. compiled and pre-processed all datasets. A.G.M participated in the generation of the supplementary material and in the writing of the paper.

TABLE OF CONTENTS

Part I

Protein degradation coupled to Nonsense-mediated mRNA decay

ABSTRACT

Translation of mRNAs containing premature termination codons (PTCs) results in truncated protein products with deleterious effects. Nonsense-mediated decay (NMD) is a surveillance pathway responsible for detecting PTC containing transcripts. While the molecular mechanisms governing mRNA degradation have been extensively studied, the fate of the nascent protein product remains largely uncharacterized. In part 1 of this thesis, we use a fluorescent reporter system in mammalian cells to reveal a selective degradation pathway specifically targeting the protein product of an NMD mRNA. We show that this process is post-translational, and dependent on the ubiquitin proteasome system. To systematically uncover factors involved in NMD-linked protein quality control, we conducted genome-wide flow cytometry-based screens. Our screens recovered known NMD factors, but suggested protein degradation did not depend on the canonical ribosome-quality control (RQC) path-way. A subsequent arrayed screen demonstrated that protein and mRNA branches of NMD rely on a shared recognition event. Our results establish the existence of a targeted pathway for nascent protein degradation from PTC containing mRNAs, and provides a reference for the field to identify and characterize required factors.

Chapter 1

INTRODUCTION TO NONSENSE-MEDIATED MRNA DECAY

1.1 mRNA degradation

Translation lies at the very center of every process in biology. The ribosome synthesizes all proteins in our cells, making its function and accuracy essential prerequisites for the proper operation of all cellular pathways. This importance is highlighted by a plethora of diseases caused by alterations in the translation machinery, with even subtle deviations from normal having great effects in the health of the whole proteome (Lee et al., 2006).

It is hence not surprising that cells have evolved a myriad of mechanisms to regulate translation. These processes ensure that the information encoded in the mRNA is faithfully converted into proteins (Steffen and Dillin, 2016), safeguarding the proteome against the errors inherent to any biological process. But even if such mechanisms were faultless, they are still dependent on a correct template. A perfect translation is of no use if the mRNA contains errors.

This problem is imperative for the cell. Errors in the mRNA can arise from a variety of sources, including genetic mutations, transcriptional errors, RNA processing defects, etc. (Maquat and Carmichael, 2001). Most importantly, its effects are much more severe than translational errors. A single mRNA can undergo translation around 2,000 to 4,000 times on average (Schwanhäusser et al., 2011). In the presence of errors, this can result in the production of thousands of incorrect proteins. This amplification effect, together with the far-reaching consequences of malfunctioning proteins, makes it crucial that mechanisms controlling the quality of mRNA exist in the cell.

Severely damaged mRNAs (such as those lacking a 5' cap or a poly(A) tail) can be easily recognized by the cellular machinery (Shoemaker and Green, 2012). But those with more subtle changes, like mutations in the coding sequence, pose a greater challenge for the cell. Unlike proteins, in which these errors can be identified by their inability to fold, a mutated mRNA offers no biophysical clue for its detection, and must therefore be recognized co-translationally. This principle defines a series of co-translational mRNA surveillance pathways whereby the ribosome not only reads the mRNA, but also scans it in search of errors. (Karamyshev and Karamysheva,

2018). In my PhD, I studied one of those pathways, nonsense-mediated mRNA decay (NMD), and analyzed its tight link with translation.

1.2 Nonsense-mediated mRNA decay (NMD)

NMD is an eukaryotic co-translational mRNA quality control pathway that surveys translation and recognizes, targets and degrades select mRNAs. The discovery of NMD emerged from studies of the genetic disease β-thalassemia, a blood disorder caused by reduced or absent mRNA of the beta chain of hemoglobin (β-globin) (Benz Jr., Swerdlow, and Forget, 1975). In 1979, the β-globin mRNA of a β-thalassemia patient was first sequenced. A single point mutation, leading to the appearance of a premature termination codon (PTC), was identified as the molecular culprit (Chang and Kan, 1979). Subsequent studies showed that a PTC could reduce the half-life of mRNAs across all eukaryotes (Maquat, Kinniburgh, et al., 1981) through a pathway that was coined Nonsense-mediated mRNA decay.

PTCs can arise for a variety of reasons, including genomic insertions, deletions, or even single point mutations along the coding sequence (named nonsense mutations) (He and Jacobson, 2015). mRNAs with PTCs encode for truncated proteins, which can be aggregation-prone, lose or gain activities, or even act as dominant negative factors. This has broad implications in human disease. Out of the nearly 7,000 known rare genetic disorders, approximately 30% arise as a consequence of a premature termination codon (Miller and Pearce, 2014). The list includes diseases such as factor X deficiency (Millar et al., 2000), von Willebrand disease (Schneppenheim et al., 2001), Retinitis pigmentosa (Rosenfeld et al., 1992), etc.

Beyond nonsense mutations, NMD has been shown to modulate around 10% of the human transcriptome (Celik, He, and Jacobson, 2017). This includes mRNAs with upstream open reading frames (uORF), with introns downstream the normal termination codon (Mendell et al., 2004) and mRNAs with defective alternative splicing (Lewis, Green, and Brenner, 2003). Intriguingly, NMD also targets many apparently normal wild-type mRNAs (Lelivelt and Culbertson, 1999; He, Li, et al., 2003; Rehwinkel et al., 2005). While these transcripts have lower codon optimality and a higher rate of out-of-frame translation in average, the exact mechanism and how this process is controlled is yet a mystery of the field (Celik, Baker, et al., 2017). This regulatory function underscores the importance of NMD in the cell, not only as a mechanism against errors, but also as a wide-ranging gene expression control pathway of a significant portion of the genome.

The selection criteria for NMD substrates are based on the effects of a premature termination codon. The position of the start and the stop codons define three fragments in every mRNA: a coding sequence, and a 5' and 3' Untranslated Region (UTR). A nonsense mutation results in the lengthening of the 3'UTR at the expense of the coding sequence. This has two important implications: i) The proteins that interact with the 3'UTR become further from the stop codon. Some of those proteins, such as the Poly(A)-binding protein (PABP), are known to interact with the translation machinery and promote efficient termination (Amrani et al., 2004; Behm-Ansmant et al., 2007). This function is compromised as the distance to the stop codon increases, reducing termination efficiency. ii) proteins that exclusively bind the coding sequence would find themselves bound to the new lengthened 3'UTR. It is believed that a combination of these two factors, as sensed by the ribosome during translation termination, is the molecular cue that triggers degradation of the mRNA by NMD (Maquat, Kinniburgh, et al., 1981).

1.3 EJC-dependent Nonsense-mediated mRNA decay

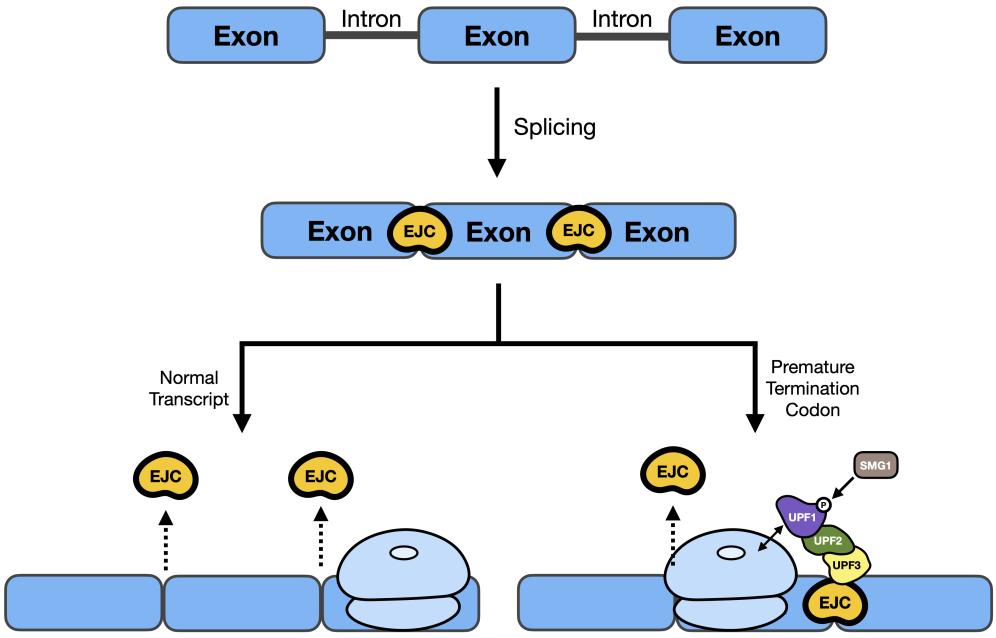


Figure 1.1: EJC-dependent Nonsense-mediated mRNA decay. A nonsense mutation can lead to the deposition of an Exon Junction Complex (EJC) downstream of the premature termination codon. The factors UPF1, UPF2 and UPF3 can bridge the EJC to the terminating ribosome and trigger degradation by NMD.

The most widely studied branch of NMD is the one dependent on the splicing of pre-mRNAs (Shoemaker and Green, 2012). Most mammalian genes contain introns, which are removed by the spliceosome co-transcriptionally (Hoskins and Moore, 2012). In this process, the spliceosome deposits a protein complex called the Exon Junction Complex (EJC) 24 nucleotides upstream of each exon-exon junction. This complex transfers the positional information of the splicing events from the nucleus to the cytoplasm (Woodward et al., 2017). Importantly, the stop codon of the vast majority of mammalian genes is located in the last exon. Therefore, the EJCs of a normal mRNA are always located in the coding sequence (Brogna and Wen, 2009). A PTC in other than the last exon would lead to the appearance of an EJC in the 3'UTR. An EJC located more than 20-24 nucleotides downstream from the termination codon flags the presence of a PTC and triggers mRNA degradation by NMD.

Although the precise molecular mechanisms are not entirely clear, there is a widely recognized model on how NMD occurs. The central NMD factor is UPF1, an RNA-dependent helicase and ATPase that mediates NMD in all tested eukaryotes (Kurosaki, Popp, and Maquat, 2019). UPF2 and UPF3B are also central to NMD. UPF3B associates with the EJC, and is considered one of its peripheral components (Singh et al., 2012). UPF2 associates with UPF3B at the EJCs. In the presence of a PTC, the UPF3B-UPF2 complex recruits UPF1 and stimulate its ATPase and helicase activities (Chamieh et al., 2008). This occurs through a change in UPF1's conformation. Free UPF1 exhibits a closed conformation, driven by the interaction of its terminal domains (Fiorini, Boudvillain, and Le Hir, 2013). UPF2 interacts with one of such terminal domains: the cysteine- and histidine-rich zinc finger domain (CH domain) at UPF1's N-terminus. This induces a relaxed open conformation that activates UPF1 (Kadlec et al., 2006).

The ATPase-dependent helicase activity of UPF1 promotes mRNA degradation by translocating along the substrate and recruiting decay factors. The exact role of UPF1's translocation is unclear. Different studies support that the role of UPF1 helicase activity is to disassemble and displace mRNA-bound proteins that would otherwise hinder RNA degradation (Fiorini, Bagchi, et al., 2015). However, a subsequent study concluded that UPF1 ATPase activity is involved in promoting efficient translation termination and ribosome release (Serdar, Whiteside, and Baker, 2016). This study argues that it is the terminating ribosome that hinders mRNA degradation, and not downstream mRNA-binding proteins as previously thought.

Which of the two models is correct (if any) has not yet been elucidated.

Another key event necessary for mRNA degradation is the recruitment of the kinase SMG1 in the context of translation termination (Yamashita et al., 2001). In a canonical termination event, the release factors eRF1 and eRF3 are recruited to the stop codon to release the nascent peptide from the ribosome. When a ribosome terminates in a PTC, the release factors interact with both UPF1 and SMG1 to form the so-called SMG1-UPF1-eRFs (SURF) complex (Kashima et al., 2006). The interaction of the SURF complex with the downstream EJC activates SMG1, which phosphorylates Upf1 on both its N- and C-terminus. This phosphorylation event is essential for NMD activation and is thought to serve as a commitment step towards mRNA degradation (Kurosaki, Li, et al., 2014).

Upf1 phosphorylation triggers NMD through the recruitment of a number of factors. These include the endonuclease SMG6, as well as the adaptors proteins SMG5-SMG7 and PNRC2, which connect UPF1 to the mRNA degradation machinery (Durand, Franks, and Lykke-Andersen, 2016). PNRC2 interacts with the DCP2 decapping complex, which elicits 5'-to-3' mRNA degradation (Cho, Kim, and Kim, 2009; Lai et al., 2012). SMG6 targets the mRNA via an endonucleolytic cleavage near the PTC (Eberle et al., 2009). The cleavage generates a 5' product that is degraded 3'-to-5' by both the exosome (Schmid and Jensen, 2008) and the exoribonuclease DIS3L2 (Malecki et al., 2013), and a 3' product that is degraded by the exoribonuclease XNR1 after being stripped off its bound factors by UPF1's helicase activity (Franks, Singh, and Lykke-Andersen, 2010)). Finally, the heterodimer SMG5-SMG7 acts as an adaptor for the CCR4-NOT deadenylation complex, which promotes 3'-to-5' exonucleolytic degradation of the target (Loh, Jonas, and Izaurralde, 2013).

The interplay between the SMG5-SMG7 and the SMG6 degradation branches has been the subject of intense study. These branches are considered to be independent from each other, as knock-down of a single one only partially inhibit NMD (Metze et al., 2013). Because the downregulation of SMG6 impairs NMD to a larger extent than the knock-down of SMG7, it is thought that endonucleolytic cleavage is the predominant strategy for mRNA degradation by NMD (Colombo et al., 2017). Nevertheless, the two branches are considered to be redundant, since the knock-down of one appears to be partially compensated by the other (Metze et al., 2013). Intriguingly, a recent paper challenged this notion and provided evidence for dependency between the SMG5-SMG7 and SMG6 branches (Boehm et al., 2020). The

papers shows that knocking-down SMG7 (as made in all previous studies) is insufficient to abolish its function, and that a complete depletion is required to observe a significant effect in NMD. The authors show that complete inactivation of the SMG5-SMG7 pathway also inhibits the SMG6 one, supporting a hierarchy between the two branches. Under this new model, the SMG5-SMG7 heterodimer would first be recruited to phosphorylated UPF1, which would then trigger SMG6 activity and mRNA degradation.

1.4 Contributions of this thesis

While many studies have investigated the mechanisms of mRNA degradation, the link between NMD and translation has been understudied. Since NMD is a co-translational pathway, the degradation of the target mRNA is inextricably associated with the production of a truncated protein. This contradicts the goal of NMD, which is to prevent the production of such truncated products. In the first part of my thesis, I outline our efforts to investigate this conundrum. In Chapter 2, we designed and built a set of NMD fluorescent reporters that deconvolute protein and mRNA degradation, a long-standing challenge in the field. We used the reporter system to discover a novel protein degradation pathway coupled to NMD, which we demonstrated to be dependent on the ubiquitin-proteasome system. Finally, we performed whole-genome CRISPR screens to uncover factors involved in this novel pathway.

References

Amrani, Nadia et al. (2004). “A faux 3’-UTR promotes aberrant termination and triggers nonsense-mediated mRNA decay.” In: *Nature* 432.7013, pp. 112–118.

Behm-Ansmant, Isabelle et al. (2007). “A conserved role for cytoplasmic poly (A)-binding protein 1 (PABPC1) in nonsense-mediated mRNA decay.” In: *The EMBO Journal* 26.6, pp. 1591–1601.

Benz Jr., Edward J., Paul S. Swerdlow, and Bernard G. Forget (1975). “Absence of functional messenger RNA activity for beta globin chain synthesis in ß0-thalassemia.” In: *Blood* 45.1, pp. 1–10.

Boehm, Volker et al. (2020). “Nonsense-mediated mRNA decay relies on "two-factor authentication" by SMG5-SMG7.” In: *bioRxiv*, pp. 2020–07.

Brogna, Saverio and Jikai Wen (2009). “Nonsense-mediated mRNA decay (NMD) mechanisms.” In: *Nature Structural & Molecular Biology* 16.2, pp. 107–113.

Celik, Alper, Richard Baker, et al. (2017). "High-resolution profiling of NMD targets in yeast reveals translational fidelity as a basis for substrate selection." In: *Rna* 23.5, pp. 735–748.

Celik, Alper, Feng He, and Allan Jacobson (2017). "NMD monitors translational fidelity 24/7". In: *Current genetics* 63.6, pp. 1007–1010.

Chamieh, Hala et al. (2008). "NMD factors UPF2 and UPF3 bridge UPF1 to the exon junction complex and stimulate its RNA helicase activity." In: *Nature structural & molecular biology* 15.1, pp. 85–93.

Chang, Judy C and Yuet Wai Kan (1979). "beta 0 thalassemia, a nonsense mutation in man." In: *Proceedings of the National Academy of Sciences* 76.6, pp. 2886–2889.

Cho, Hana, Kyoung Mi Kim, and Yoon Ki Kim (2009). "Human proline-rich nuclear receptor coregulatory protein 2 mediates an interaction between mRNA surveillance machinery and decapping complex." In: *Molecular cell* 33.1, pp. 75–86.

Colombo, Martino et al. (2017). "Transcriptome-wide identification of NMD-targeted human mRNAs reveals extensive redundancy between SMG6-and SMG7-mediated degradation pathways." In: *RNA* 23.2, pp. 189–201.

Durand, Sébastien, Tobias M Franks, and Jens Lykke-Andersen (2016). "Hyperphosphorylation amplifies UPF1 activity to resolve stalls in nonsense-mediated mRNA decay." In: *Nature communications* 7.1, pp. 1–12.

Eberle, Andrea B et al. (2009). "SMG6 promotes endonucleolytic cleavage of nonsense mRNA in human cells." In: *Nature structural & molecular biology* 16.1, pp. 49–55.

Fiorini, Francesca, Debjani Bagchi, et al. (2015). "Human Upf1 is a highly processive RNA helicase and translocase with RNP remodelling activities." In: *Nature communications* 6.1, pp. 1–10.

Fiorini, Francesca, Marc Boudvillain, and Herve Le Hir (2013). "Tight intramolecular regulation of the human Upf1 helicase by its N-and C-terminal domains." In: *Nucleic acids research* 41.4, pp. 2404–2415.

Franks, Tobias M, Guramrit Singh, and Jens Lykke-Andersen (2010). "Upf1 ATPase-dependent mRNP disassembly is required for completion of nonsense-mediated mRNA decay." In: *Cell* 143.6, pp. 938–950.

He, Feng and Allan Jacobson (2015). "Nonsense-mediated mRNA decay: degradation of defective transcripts is only part of the story". In: *Annual review of genetics* 49, p. 339.

He, Feng, Xiangrui Li, et al. (2003). "Genome-wide analysis of mRNAs regulated by the nonsense-mediated and 5' to 3' mRNA decay pathways in yeast." In: *Molecular cell* 12.6, pp. 1439–1452.

Hoskins, Aaron A. and Melissa J. Moore (2012). "The spliceosome: a flexible, reversible macromolecular machine." In: *Trends in Biochemical Sciences* 37.5, pp. 179–188.

Kadlec, Jan et al. (2006). "Crystal structure of the UPF2-interacting domain of nonsense-mediated mRNA decay factor UPF1." In: *RNA* 12.10, pp. 1817–1824.

Karamyshev, Andrey L and Zemfira N Karamysheva (2018). "Lost in translation: ribosome-associated mRNA and protein quality controls". In: *Frontiers in Genetics* 9, p. 431.

Kashima, Isao et al. (2006). "Binding of a novel SMG-1–Upf1–eRF1–eRF3 complex (SURF) to the exon junction complex triggers Upf1 phosphorylation and nonsense-mediated mRNA decay." In: *Genes & Development* 20.3, pp. 355–367.

Kurosaki, Tatsuaki, Wencheng Li, et al. (2014). "A post-translational regulatory switch on UPF1 controls targeted mRNA degradation." In: *Genes & Development* 28.17, pp. 1900–1916.

Kurosaki, Tatsuaki, Maximilian W. Popp, and Lynne E. Maquat (2019). "Quality and quantity control of gene expression by nonsense-mediated mRNA decay." In: *Nature Reviews Molecular Cell Biology* 20.7, pp. 406–420.

Lai, Tingfeng et al. (2012). "Structural basis of the PNRC2-mediated link between mRNA surveillance and decapping." In: *Structure* 20.12, pp. 2025–2037.

Lee, Jeong Woong et al. (2006). "Editing-defective tRNA synthetase causes protein misfolding and neurodegeneration". In: *Nature* 443.7107, pp. 50–55.

Lelivelt, Michael J. and Michael R. Culbertson (1999). "Yeast Upf proteins required for RNA surveillance affect global expression of the yeast transcriptome." In: *Molecular and Cellular Biology* 19.10, pp. 6710–6719.

Lewis, Benjamin P., Richard E. Green, and Steven E. Brenner (2003). "Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans." In: *Proceedings of the National Academy of Sciences* 100.1, pp. 189–192.

Loh, Belinda, Stefanie Jonas, and Elisa Izaurralde (2013). "The SMG5–SMG7 heterodimer directly recruits the CCR4–NOT deadenylase complex to mRNAs containing nonsense codons via interaction with POP2." In: *Genes & Development* 27.19, pp. 2125–2138.

Malecki, Michal et al. (2013). "The exoribonuclease Dis3L2 defines a novel eukaryotic RNA degradation pathway." In: *The EMBO Journal* 32.13, pp. 1842–1854.

Maquat, Lynne E. and Gordon G. Carmichael (2001). "Quality control of mRNA function". In: *Cell* 104.2, pp. 173–176.

Maquat, Lynne E., Alan J. Kinniburgh, et al. (1981). "Unstable β-globin mRNA in mRNA-deficient β0 thalassemia". In: *Cell* 27.3, pp. 543–553.

Mendell, Joshua T. et al. (2004). “Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise”. In: *Nature Genetics* 36.10, pp. 1073–1078.

Metze, Stefanie et al. (2013). “Comparison of EJC-enhanced and EJC-independent NMD in human cells reveals two partially redundant degradation pathways.” In: *RNA* 19.10, pp. 1432–1448.

Millar, D. et al. (2000). “Molecular analysis of the genotype-phenotype relationship in factor VII deficiency”. In: *Human Genetics* 107.4, pp. 327–342.

Miller, Jake N. and David A. Pearce (2014). “Nonsense-mediated decay in genetic disease: friend or foe?” In: *Mutation Research/Reviews in Mutation Research* 762, pp. 52–64.

Rehwinkel, Jan et al. (2005). “Nonsense-mediated mRNA decay factors act in concert to regulate common mRNA targets.” In: *RNA* 11.10, pp. 1530–1544.

Rosenfeld, Philip J. et al. (1992). “A null mutation in the rhodopsin gene causes rod photoreceptor dysfunction and autosomal recessive retinitis pigmentosa”. In: *Nature genetics* 1.3, pp. 209–213.

Schmid, Manfred and Torben Heick Jensen (2008). “The exosome: a multipurpose RNA-decay machine.” In: *Trends in Biochemical Sciences* 33.10, pp. 501–510.

Schneppenheim, Reinhard et al. (2001). “Expression and characterization of von Willebrand factor dimerization defects in different types of von Willebrand disease”. In: *Blood, The Journal of the American Society of Hematology* 97.7, pp. 2059–2066.

Schwanhäusser, Björn et al. (2011). “Global quantification of mammalian gene expression control”. In: *Nature* 473.7347, pp. 337–342.

Serdar, Lucas D., DaJuan L. Whiteside, and Kristian E. Baker (2016). “ATP hydrolysis by UPF1 is required for efficient translation termination at premature stop codons.” In: *Nature Communications* 7.1, pp. 1–8.

Shoemaker, Christopher J. and Rachel Green (2012). “Translation drives mRNA quality control”. In: *Nature Structural & Molecular Biology* 19.6, pp. 594–601.

Singh, Guramrit et al. (2012). “The cellular EJC interactome reveals higher-order mRNP structure and an EJC-SR protein nexus.” In: *Cell* 151.4, pp. 750–764.

Steffen, Kristan K. and Andrew Dillin (2016). “A ribosomal perspective on proteostasis and aging”. In: *Cell Metabolism* 23.6, pp. 1004–1012.

Woodward, Lauren A. et al. (2017). “The exon junction complex: a lifelong guardian of mRNA fate.” In: *Wiley Interdisciplinary Reviews: RNA* 8.3, e1411.

Yamashita, Akio et al. (2001). “Human SMG-1, a novel phosphatidylinositol 3-kinase-related protein kinase, associates with components of the mRNA surveillance complex and is involved in the regulation of nonsense-mediated mRNA decay.” In: *Genes & Development* 15.17, pp. 2215–2228.

Chapter 2

COUPLED PROTEIN QUALITY CONTROL DURING NONSENSE-MEDIATED MRNA DECAY

Inglis, Alison J et al. (2023). "Coupled protein quality control during nonsense-mediated mRNA decay". In: *Journal of Cell Science* 136.10. DOI: 10.1242/jcs.261216. URL: https://journals.biologists.com/jcs/article/136/10/jcs261216/310674.

2.1 Introduction

Like other mRNA surveillance pathways, NMD substrates are recognized and targeted for degradation co-translationally (Belgrader, Cheng, and Maquat, 1993; Wang, Vock, et al., 2002; Zhang and Maquat, 1997), resulting in the synthesis of a potentially aberrant nascent polypeptide chain. Pathways such as no-go and non-stop mRNA decay rely on a coordinated protein quality control pathway, known as ribosome associated quality control (RQC) to both rescue the ribosome and concomitantly target the nascent protein for degradation (Doma and Parker, 2006; Frischmeyer et al., 2002; Juszkiewicz et al., 2018; Van Hoof et al., 2002). In both cases, a terminally stalled ribosome or a collided di-ribosome triggers ribosome splitting (Becker et al., 2011; Pisareva et al., 2011; Shao, Brown, et al., 2015; Shao, Murray, et al., 2016; Shoemaker and Green, 2012) and nascent chain ubiquitination by the E3 ligase LTN1 (facilitated by NEMF, TAE2, and P97) (Brandman, Stewart-Ornstein, et al., 2012; Defenouillère et al., 2013; Lyumkis et al., 2014; Shao, Brown, et al., 2015; Shao, Von der Malsburg, and Hegde, 2013; Verma, Oania, et al., 2013). The ubiquitinated nascent chain is then released from the ribosome by the endonuclease ANKZF1 (Vms1 in yeast) for degradation by the proteasome (Zurita Rendón et al., 2018; Verma, Reichermeier, et al., 2018).

Given the potential dominant negative and proteotoxic effects of even small amounts of a truncated NMD substrate, it has been suggested that a similar protein quality control pathway may exist to recognize and degrade nascent proteins that result from translation of NMD mRNAs. Indeed, proteins produced from PTC-containing mRNAs are less stable than those from normal transcripts (Kuroha, Tatematsu, and Inada, 2009; Kuroha, Ando, et al., 2013; Pradhan et al., 2021; Udy and Bradley,

2022). However, these observations are largely based on comparison of truncated products with longer, potentially more stable polypeptides, making it difficult to distinguish NMD-linked protein degradation from general cellular quality control mechanisms. While recent work has directly tested this using a full-length protein product, there remains no defined mechanism of targeting and degradation, nor direct evidence for the involvement of the ubiquitin-proteasome pathway (Chu et al., 2021; Udy and Bradley, 2022). Furthermore, though it has been postulated that components of the RQC are involved in turnover of nascent NMD substrates (Arribere and Fire, 2018; Chu et al., 2021), the factors required for this process have not been systematically investigated. Because NMD is triggered at a stop codon unlike no-go and non-stop decay, a putative NMD-coupled protein quality control pathway may require a fundamentally different strategy to initiate nascent protein degradation.

Here we describe a reporter system that we have used to identify and interrogate a coupled protein quality control branch of NMD. We demonstrated that in addition to triggering mRNA degradation, NMD concomitantly coordinates degradation of the nascent polypeptide via the ubiquitin-proteasome pathway. Using this reporter system, we systematically identified factors required for NMD-coupled protein degradation, which are distinct from the canonical rescue factors of the RQC. Characterization of a coupled protein-degradation branch of NMD represents a new facet of our understanding of how the cell ensures the integrity and composition of its proteome, and sheds further light on the interplay between mRNA and protein quality control.

2.2 Results

A reporter strategy to decouple mRNA and protein quality control in NMD

To identify a putative NMD-linked protein quality control pathway, we developed a reporter system that uncouples mRNA and protein quality control during NMD. The reporter consists of a single open reading frame expressing GFP and RFP, separated by a viral 2A sequence that causes peptide skipping (Wang, Wang, et al., 2015) (Fig. 2.1A, Fig. S2.1A). A robust example of an endogenous NMD substrate is the beta-globin gene with a nonsense mutation at codon 39, which results in a premature stop codon followed by an intron (Zhang, Sun, et al., 1998). We therefore reasoned that positioning the first intron of the human β-globin gene into the 3' UTR of our reporter after the stop codon would also lead to its recognition as an NMD substrate, as has been previously reported (Chu et al., 2021; Durand and Lykke-Andersen,

2013; Pereverzev et al., 2015). We confirmed that the exogenous β-globin intron is efficiently spliced (Fig. S2.1B), and observed that the mRNA levels of the NMD reporter were 5-fold lower than a matched non-NMD control (Fig. 2.1B). We found that the GFP fluorescence of the NMD reporter and control correlated with their respective mRNA levels, as directly measured by qPCR, suggesting that GFP fluorescence can be used as a proxy for transcript levels (Fig. S2.1D). Further, we saw that knockdown of the core NMD factor UPF1 specifically increased the GFP fluorescence of the NMD reporter (Fig. S2.1E-H), but had no effect on the matched control. We therefore concluded that our fluorescent reporter is recognized and degraded in an NMD-dependent manner. Finally, to ensure that these effects did not result solely from the increase in translation associated with the presence of an EJC (Nott, Le Hir, and Moore, 2004), we also generated a reporter containing an EJC immediately following the stop codon, which is not recognized as an NMD substrate (inert EJC, Fig. 2.1A) (Nagy and Maquat, 1998). Indeed, the mRNA levels of this inert EJC construct were similar to our unspliced control (Fig. S2.1C).

After establishing that our reporters are subject to NMD-dependent mRNA degradation as expected, we sought to exploit them to determine whether there was an additional pathway dedicated to nascent protein degradation. For this, our reporter design has two important physical features. First, it can be used to deconvolute post-transcriptional versus post-translational effects on reporter levels. Upon translation, the GFP is released by the 2A sequence while the RFP remains tethered to the ribosome until the termination codon, where NMD is initiated by interaction between the downstream EJC and the ribosome. We reasoned that if there is an NMD-coupled pathway that triggers degradation of the nascent polypeptide, it would thus act only on the RFP but not the released GFP, resulting in a reduction in the RFP:GFP ratio in comparison to a matched control. In contrast, if NMD functions only in mRNA degradation, we would expect a decrease in both the RFP and GFP levels but would observe no change in the RFP:GFP ratio. Second, these reporters can specifically distinguish nascent protein degradation by a coupled protein quality control pathway from non-specific recognition by general cellular quality control machinery. Canonical NMD substrates contain PTCs that result in translation of a truncated protein, which may be misfolded and thus recognized and degraded by non-specific cytosolic quality control pathways (Popp and Maquat, 2013). By instead using an intact RFP moiety that is recognized as an NMD substrate only because of an intron in its 3' UTR, any destabilization of RFP must result from a coordinated event that occurs prior to its release from the ribosome.

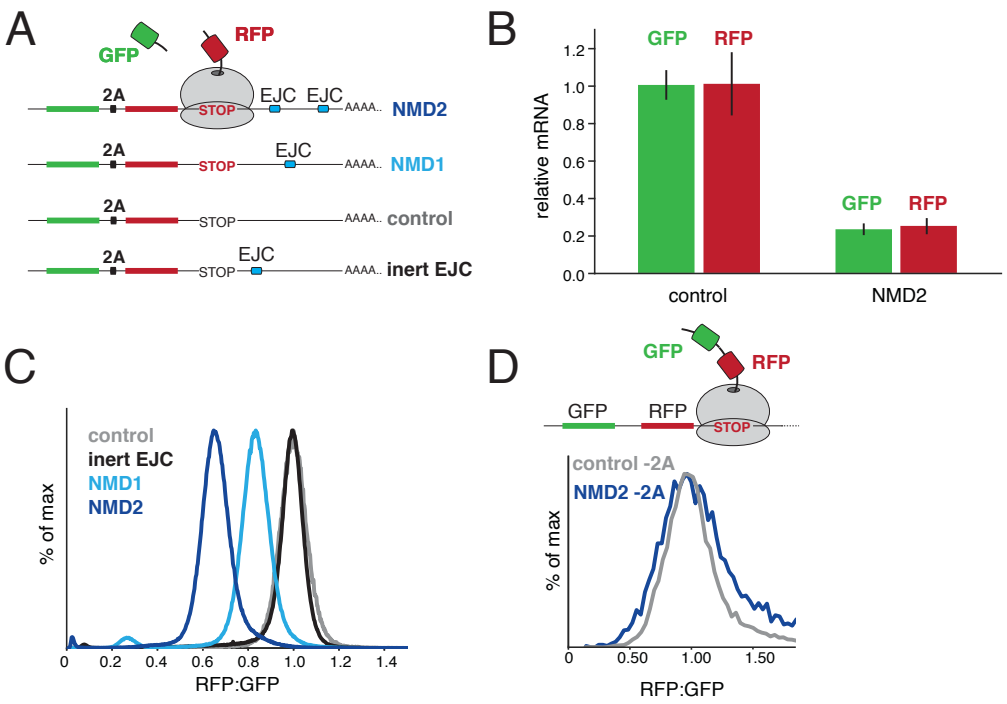


Figure 2.1: Destabilization of nascent proteins from PTC-containing mRNAs. **(A)** Schematic of the reporter strategy used to monitor protein and mRNA degradation in NMD. GFP and RFP are encoded in a single open reading frame separated by a viral 2A sequence. Either one or two introns derived from the β-globin gene are inserted after the stop codon (NMD1 and NMD2, respectively). To control for the documented stimulation in translation that results from the presence of an EJC (Nott, Le Hir, and Moore, 2004), we created a reporter in which the intron was positioned twelve nucleotides after the stop codon, a distance insufficient for recognition as an NMD substrate (inert EJC) (Nagy and Maquat, 1998). **(B)** T-Rex HEK293 cell lines stably expressing either the control or the NMD2 reporter were induced with doxycycline for 24 hours and the total mRNA was then purified. Relative mRNA levels were determined by RT-qPCR using two sets of primers that anneal to the very 5' region of the GFP and 3' region of the RFP open reading frames, respectively. The results were normalized to the control and the standard deviation from three independent experiments is displayed. **(C)** T-Rex HEK293 cell lines stably expressing the indicated reporters were analyzed by flow cytometry. The ratio of RFP:GFP fluorescence, normalized to the control reporter, is depicted as a histogram and quantified in Fig.S2E. **(D)** HEK293T cells were transiently transfected with versions of the control and NMD2 reporters in which the 2A sequence was scrambled, resulting in tethering of both GFP and RFP to the ribosome at the stop codon. Cells were analyzed by flow cytometry after 24 hours and quantified in Fig. S2.2E.

Indeed, using flow cytometry, we observed a decrease in RFP:GFP fluorescence for an NMD substrate compared to a matched control, in two different cell lines (Fig. 2.1C, Fig. S2.2A). Addition of a second β-globin intron to the 3' UTR (Hoek et al., 2019) resulted in a larger decrease in both the mRNA levels and RFP:GFP fluorescence, suggesting the two effects may be tightly coordinated (Hoek et al., 2019). While this decrease in RFP:GFP levels was consistent with NMD-dependent protein quality control, we sought to exclude several alternative models that could

also account for this observation. First, we swapped the order of the RFP and GFP to rule out that differential maturation and/or turnover rates of the fluorophores could explain the decrease in RFP:GFP ratio (Fig. S2.2B, 2.2C) (Amrani et al., 2004; Balleza, Kim, and Cluzel, 2018). A similar effect was observed for this 'reverse' reporter, as previously reported (Chu et al., 2021). Second, we considered whether the decrease in RFP:GFP ratio could be the result of NMD-dependent deadenylation and 3' to 5' exonuclease degradation of the reporter mRNA (Chen and Shyu, 2003; Mitchell and Tollervey, 2003; Takahashi, Araki, Sakuno, et al., 2003). However, we detected no difference in the relative mRNA levels of the RFP and GFP coding regions of the NMD substrate (Fig. 2.1B), confirming that the effect must occur post-transcriptionally. Finally, we addressed two related possibilities: whether slow translational termination, which was shown to occur on NMD substrates in yeast, though potentially not mammals (Amrani et al., 2004; Karousis et al., 2020), or SMG6-dependent endonucleolytic cleavage of the mRNA at the stop codon could explain the RFP:GFP ratio decrease (Eberle et al., 2009). The former could result in increased dwell time of the ribosome at the stop codon when the 30 C-terminal residues of RFP remain occluded in the ribosomal exit tunnel and could potentially affect RFP folding and therefore fluorescence. The latter would lead to production of full-length GFP but truncated RFP, and would be consistent with models proposed for putative NMD-coupled protein quality control in C. elegans (Arribere and Fire, 2018). However, appending a flexible linker to the C-terminus of RFP to ensure it is fully emerged from the ribosome at the stop codon did not affect the RFP:GFP ratio (Fig. S2.2D). This is consistent with the very long maturation time of RFP (mins-hours (Balleza, Kim, and Cluzel, 2018)), which is therefore unlikely to be affected by any putative dwell time (ms-s; (Amrani et al., 2004)) at the stop codon. Conversely, scrambling the 2A sequence, such that both the GFP and RFP are tethered to the ribosome at the stop codon, abolished the ratio difference (Fig. 2.1D, Fig. S2.2E). Together these data exclude that the NMD-dependent decrease in RFP:GFP ratio is due to changes in translation rate, processivity, peptide release, endonucleolytic cleavage, or preferential 3'-5' degradation.

NMD-dependent protein degradation occurs via the ubiquitin proteasome pathway

Having established that an NMD-dependent decrease in RFP fluorescence occurs post-translationally, we tested whether inhibition of the ubiquitin-proteasome pathway could rescue the observed phenotype. We found that both the proteasome

inhibitor MG132 and the E1 ubiquitin-activating enzyme inhibitor MLN7243 specifically increased the RFP:GFP ratio of the NMD reporter (Fig. 2.2A; Fig. S2.3A, C-D). Importantly, this increase was due to an effect on RFP and not GFP (Fig. 2.2B, Fig. S2.3B), consistent with the model that NMD-dependent protein degradation acts post-translationally and selectively toward the polypeptide associated with the ribosome at the PTC. To confirm that the observed changes in fluorescence reflect changes at the protein level, we directly tested for stabilization of RFP upon E1 enzyme inhibition by western blotting (Fig. S2.3D). The apparent absence of truncated RFP would be consistent with a model in which NMD-dependent protein quality control is initiated at the stop codon. Finally, we directly observed a marked increase in ubiquitination of RFP, but not GFP, when expressed from our NMD reporter compared with a matched control, excluding potential indirect effects of ubiquitin-proteasome pathway inhibition (Fig. 2.2C). Therefore, we concluded that in addition to its well-characterized role in mRNA degradation, NMD also triggers degradation of nascent proteins via the ubiquitin proteasome pathway.

Identification of factors required for NMD-coupled protein quality control

Using our characterized NMD2 reporter, we systematically identified factors required for the protein degradation arm of NMD using a fluorescence-activated cell sorting (FACS)-based CRISPR interference (CRISPRi) (Horlbeck et al., 2016) and CRISPR knockout (CRISPR-KO) screen (Fig. 2.3A). We reasoned that the knockdown screen would enable study of essential proteins, including the core NMD factors UPF1 and UPF2 (Hart et al., 2017). Conversely, the knockout screen would identify factors that require near-complete depletion to induce a measurable phenotype, which can lead to false negatives in CRISPRi screens (Rosenbluh et al., 2017). To do this, we engineered two K562 human cell lines that expressed an inducible NMD2 reporter either alone or with the CRISPRi silencing machinery (Gilbert et al., 2014). We transduced the CRISPRi cell line with a single guide RNA (sgRNA) library targeting all known protein-coding open reading frames as previously described (hCRISPRi-v2) (Horlbeck et al., 2016). For the knockout screen, we used a novel 100,000 element library that targets all protein encoding genes (5 sgRNA/gene), which we used to simultaneously deliver both the genome wide sgRNA library and Cas9.

We hypothesized that depletion of factors required for NMD-coupled protein quality control would stabilize RFP, thereby increasing the RFP:GFP ratio. However, depletion of factors that impede NMD-coupled protein quality control would further

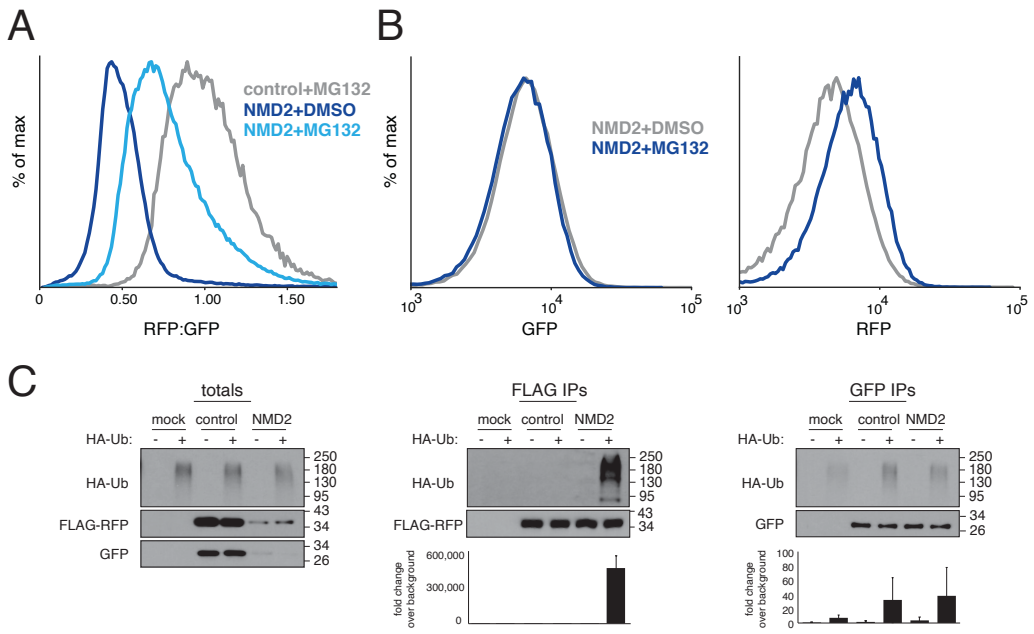


Figure 2.2: NMD-dependent protein degradation occurs via the ubiquitin proteasome pathway. **(A)** Flow cytometry analysis of HEK293T cells transiently transfected with either the con- trol or NMD2 reporter (Fig. 2.1A) and treated with the proteasome inhibitor MG132 or DMSO for 6 hours. See quantification in Fig. S2.4A **(B)** K562 CRISPRi cells stably expressing an inducible NMD2 reporter were treated with either MG132 or DMSO after induction of the reporter and analyzed by flow cytometry. Shown are the GFP (left) and RFP (right) channels for the indicated conditions displayed as a histogram, with fold change quantified in Fig. S2.4B. **(C)** HEK293T cells, stably expressing an HA-tagged ubiquitin (HA-Ub) were transiently transfected with either the control or NMD2 reporter (modified to incorporate a 3xFLAG tag at the N-terminus of RFP). To stabilize ubiquitinated species, cells were treated with MG132 prior to lysis. RFP was immunoprecipitated with anti-FLAG resin and GFP was purified using a GFP nanobody coupled to streptavidin resin (Pleiner et al., 2020). Ubiquitinated species were detected by western blotting for HA-Ub. The quantification of three independent replicates is shown below, with the means and standard deviations plotted.

decrease the RFP:GFP ratio. For the CRISPRi screen, after eight days of knockdown, we sorted cells with high and low RFP:GFP ratios via FACS, and identified sgRNAs enriched in those cells by deep sequencing. For the knockout screen we isolated cells with perturbed RFP:GFP ratios on days eight, ten and twelve post infection of the CRISPR-KO library. We postulated that essential genes would be better represented at the earlier time points before their depletion becomes lethal, while factors that require complete depletion and/or have longer half-lives would be detected at later time points.

In both the knockdown and knockout screens, we find substantial differences between the hits identified here and those from earlier screens designed to identify factors primarily involved in NMD-dependent mRNA degradation (Alexandrov, Shu, and

Steitz, 2017; Baird et al., 2018; Sun et al., 2011; Zinshteyn et al., 2021) suggesting our reporter reflects a distinct aspect of the NMD pathway (Fig. 2.3B-D, Fig. S2.4A). However, we also identified several splicing and core NMD factors as effectors of the RFP:GFP ratio. For example, we found that the core component of the EJC, CASC3 (Gerbracht et al., 2020) is required for NMD-coupled protein degradation (Fig. 2.3B, 2.3D). Furthermore, depletion of several known NMD factors—UPF1, UPF2, UPF3B, SMG6—increased the RFP:GFP ratio of our NMD-reporter. Additionally, we also identified factors that appeared to enhance the degradation of RFP relative to GFP. On day eight of the knockout screen, we found that several essential factors required for 5' to 3' mRNA degradation were enriched in the population of cells with lower RFP:RFP fluorescence (Fig. 2.3C). The phenotype scores for these essential factors decreased from day 8 to day 12, likely due to guide drop out, thereby validating the importance of examining the knock-out screen across multiple time points (Fig. S2.4A). Together, these results suggest a single, shared recognition step for both the mRNA and protein quality control branches of NMD, which requires recognition of an intact EJC downstream of the stop codon via interactions between the canonical NMD factors and the ribosome.

NMD-coupled protein quality control is not mediated by canonical RQC factors

Notably absent in both the knockdown and knockout screen were canonical components of the RQC pathway, suggesting that NMD substrates may rely on an alternative strategy for nascent protein degradation. Because the CRISPRi screen was performed using the same platform and conditions as earlier reporter screens for non-stop decay—including the same cell type, sgRNA library, and sampling time point—the screens are directly comparable (Hickey et al., 2020). While depletion of RQC factors including PELO and the E3 ubiquitin ligase LTN1 were identified in the non-stop reporter screen, neither are significant hits for NMD-dependent protein degradation in our system (Fig. 2.4A, 2.4B). We directly verified that LTN1 knockdown has no effect on our NMD reporter, or the 'reverse' reporter, but did have a marked effect on the fluorescence ratio of an established non-stop decay substrate (Fig. 2.4C-D, Fig. S2.4B-C). We therefore concluded that NMD-coupled protein degradation is mediated by a different set of factors.

Factors required for NMD-coupled protein quality control

Hits from the FACS based reporter screens were validated using an arrayed screen with a matched control. These data confirmed that knockdown of CASC3 increased

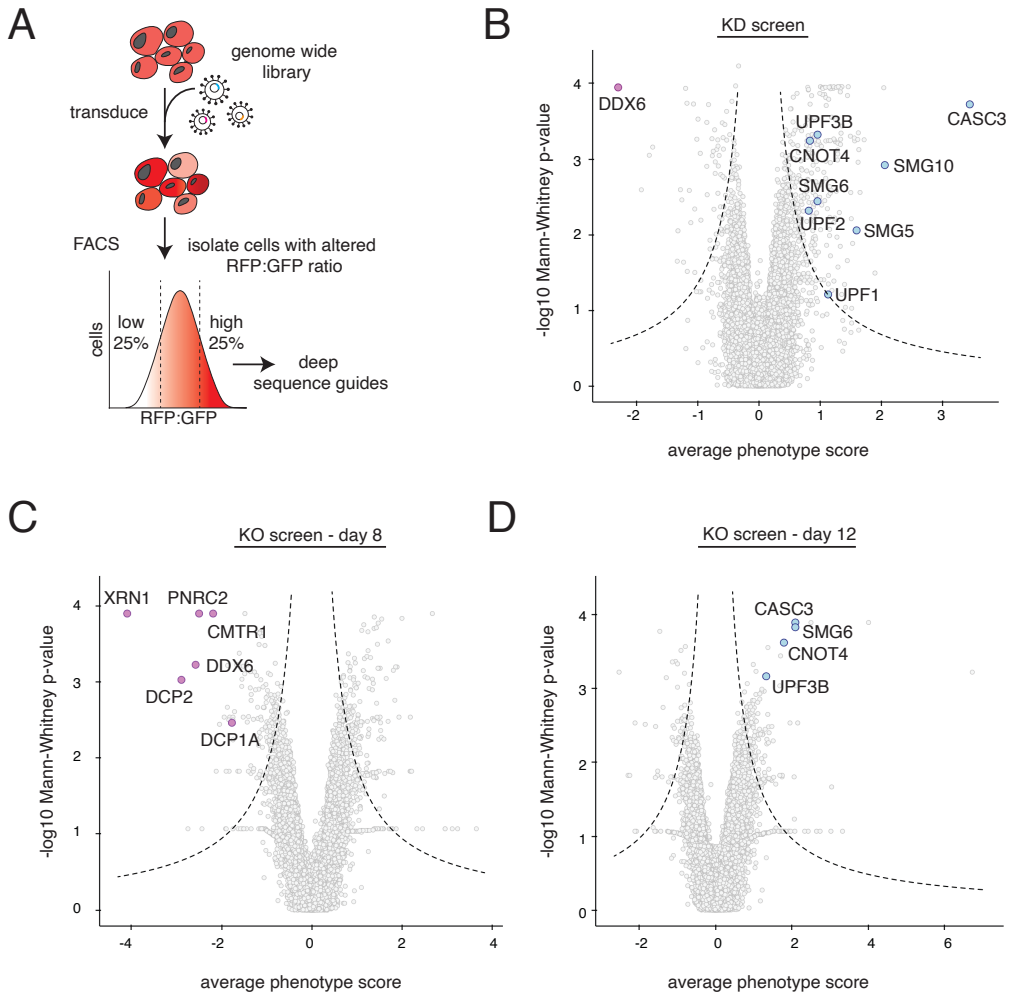


Figure 2.3: Systematic characterization of factors required for NMD-coupled protein quality control. **(A)** Schematic of the workflow. **(B)** Volcano plot of the RFP:GFP stabilization phenotype (log2 for the three strongest sgRNAs per gene) and Mann–Whitney p values from the genome-wide CRISPRi screen, with each point representing one gene. Genes falling outside the dashed lines are statistically significant. Notable hits causing an increase in the RFP to GFP ratio are shown in light blue and include known NMD factors, the splicing factor CASC3, and the E3 ligase CNOT4. DDX6, a known suppressor of NMD, which causes a lower RFP to GFP ratio, is shown in purple. **(C)** Volcano plot as in (B) for the genome-wide CRISPR knock-out screen sorted at the day 8 timepoint. Factors that cause a decrease in RFP relative to GFP include genes involved in mRNA de-capping, DDX6, and the 5'-3' exonuclease XRN1. **(D)** As in (C) but for day 12. In blue are shown known NMD factors and the E3 ligase CNOT4. Highlighted genes can be tracked across the three days of screening in Fig. S2.4A.

both the GFP levels and the RFP:GFP ratio of our NMD reporter (Fig. 2.5, Fig. S2.5A, C, E). The effect of CASC3 (also referred to as MLN51) depletion on our reporter is consistent with its established role as a splicing factor and a critical core component of the EJC (Le Hir, Izaurralde, et al., 2000; Bono et al., 2006). Knockdown of the 5' decapping enzyme DCP1A also increased GFP levels, but

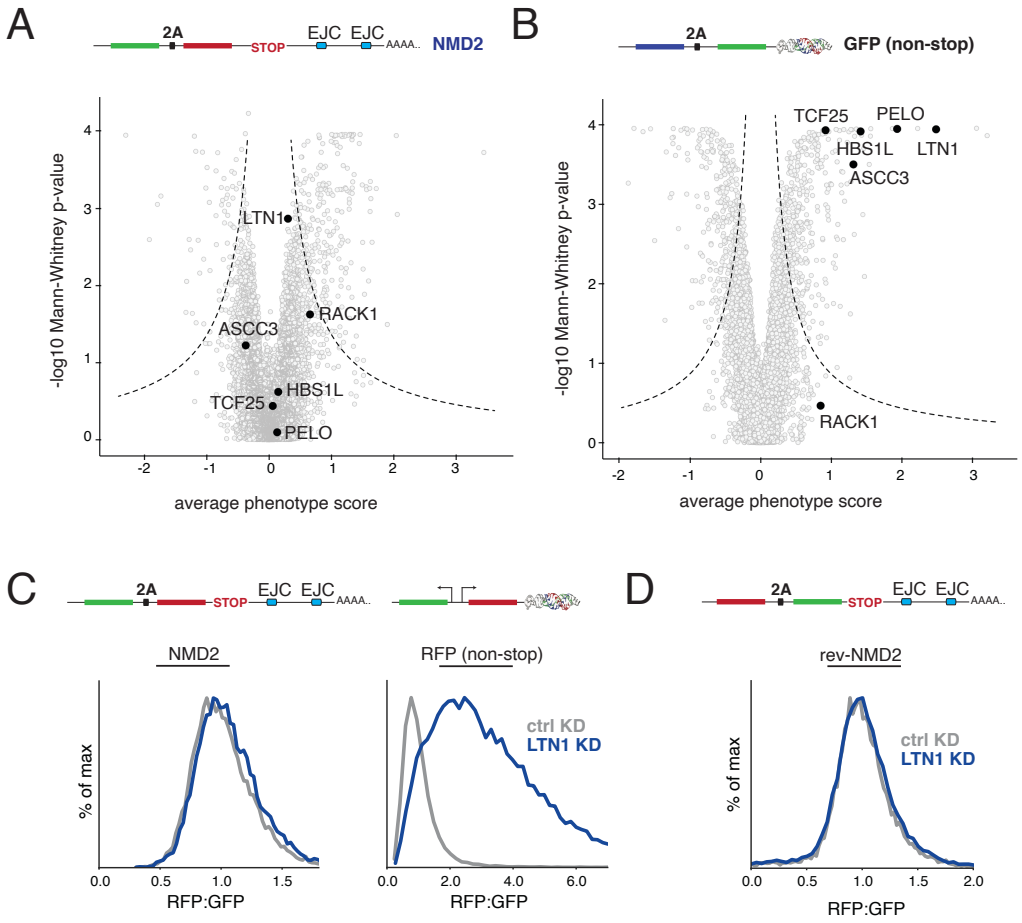


Figure 2.4: NMD-linked protein degradation is not mediated by the canonical RQC pathway. **(A)** Volcano plot of the NMD2 reporter CRISPRi screen as in Fig 3A. Highlighted in black are factors involved in the canonical RQC. **(B)** For comparison, RQC factors are highlighted in black on a volcano plot for an earlier CRISPRi screen using a non-stop reporter (consisting of a BFP, a viral 2A skipping sequence, and a GFP conjugated a triple helix moiety to stabilize the mRNA transcript, which would usually be degraded due to the lack of a stop codon) conducted using identical conditions as in (A) (Hickey et al., 2020). **(C)** K562 CRISPRi cells stably expressing either an inducible NMD2 reporter or a constitutively expressed non-stop reporter with matched GFP and RFP fluorophores (in this case, a functionally equivalent non-stop reporter with two separate promoters, one driving GFP, and the other RFP conjugated to the triple helix moiety; as in Hickey, 2020) were infected with a sgRNA targeting the E3 ligase LTN1. The RFP to GFP ratios for NMD2, and the GFP to RFP ratio for the non-stop reporter as determined by flow cytometry are displayed as a histogram and quantified in Fig. S2.4B. **(D)** K562 CRISPRi cells expressing a reversed version of the NMD2 reporter (rev-NMD2) were infected with an sgRNA against LTN1 and analyzed as in (C) and quantified in Fig. S2.4C.

decreased the RFP:GFP ratio. We confirmed these phenotypes were generalizable using our reverse GFP:RFP reporter (Fig. S2.5B, D, F).

Having observed that the nascent protein is directly ubiquitinated and degraded by the proteasome (Fig. 2.2), we were particularly interested in identifying an E3 ubiq-

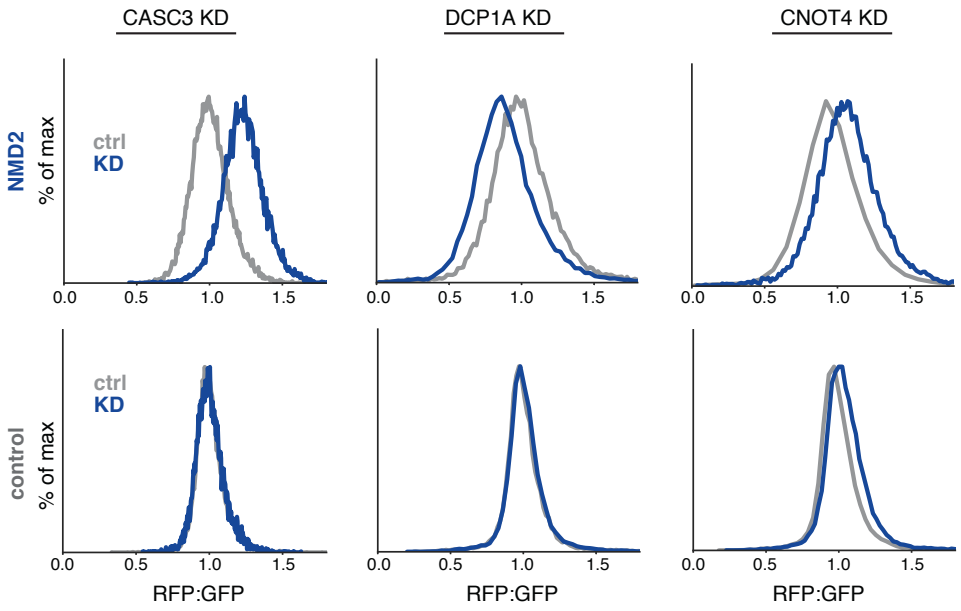


Figure 2.5: Validation of factors involved in NMD-coupled protein quality control. Factors of interest were individually depleted by sgRNA in K562 Zim3 CRISPRi cells expressing the indicated reporter. Displayed are the RFP:GFP ratios for the NMD2 (top) and control (bottom) reporters as determined by flow cytometry, see Fig. S2.5A, C, E for quantification. Similar results were obtained using reverse reporters as in Fig. S2.5B, D, F.

uitin ligase responsible for targeting the NMD-linked nascent chain for degradation. The core NMD factor UPF1 is an E3 ubiquitin RING ligase (Takahashi, Araki, Ohya, et al., 2008) and thus would be well-positioned to mediate nascent chain degradation during NMD. Previous studies have demonstrated that UPF1 stimulates proteasomal degradation of proteins expressed from NMD-targeted mRNA transcripts in yeast, with reporter stability significantly increased in upf1 knockout strains; however, the mechanism underlying this phenotype is unclear and a direct role in nascent chain ubiquitination by UPF1 was not shown (Kuroha, Tatematsu, and Inada, 2009). UPF1 was identified as a weak hit in our CRISPRi screen (Fig. 2.3B), and its depletion resulted in a shift in the RFP:GFP ratio of the NMD reporter (Fig. S2.1F-H). However, rescue of UPF1 knockdown with a RING mutant that disrupts binding with E2 ubiquitin-conjugating enzymes (Feng, Jagannathan, and Bradley, 2017) phenocopied wild-type UPF1 in restoring both the GFP levels and RFP:GFP ratio of our NMD reporter (Fig. S2.6). This result would be inconsistent with a role for the RING domain of UPF1 in ubiquitination of the nascent protein, and suggests that the involvement of UPF1 may instead be upstream of the protein degradation branch. In addition to UPF1, we identified four other E3 ubiquitin ligases in either

the knockdown and knockout screen (KEAP1, MYLIP, CBLL1, and TRIM25). The RING ligase CNOT4 was the only hit to be identified in both screens; however, its effect was not specific to NMD substrates (Fig 2.5, Fig. S2.5E-G), despite efficient depletion (Fig. S2.5G). It therefore is more likely playing a general role in cellular proteostasis, but is unlikely to be specifically involved in NMD-coupled nascent chain degradation.

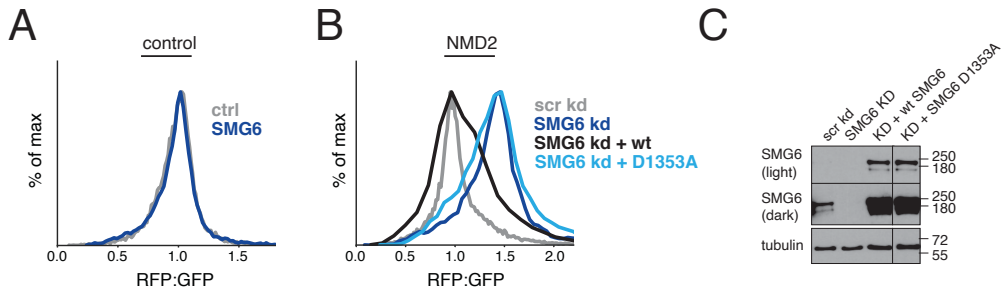


Figure 2.6: NMD-coupled protein quality control is dependent on endonucleolytic cleavage of the mRNA by SMG6. **(A)** HEK293T cells were treated with siRNA against SMG6 for 48 hours, then were transiently transfected with the control reporter. The cells were analyzed by flow cytometry after 24 hours (see quantification in Fig. S2.7A). **(B)** HEK293T cells were treated with siRNA against SMG6 for 48 hours, then were transiently transfected with an siRNA-resistant version of either wild-type SMG6 or a PIN domain mutant (D1353A) along with the NMD2 reporter. The cells were analyzed by flow cytometry after 24 hours (see quantification in Fig. S2.7B). Similar results were obtained with reverse reporters (Fig. S2.7D). **(C)** Levels of SMG6 in the samples from (B) were analyzed by western blotting against SMG6.

Additionally, the endonuclease SMG6 was also identified as a strong hit in both the knock-down and knockout screens (Fig. 2.3). Cleavage by SMG6 is considered a commitment step to degradation of NMD mRNAs, and we sought to determine if the branchpoint of the protein and mRNA degradation pathways was upstream or downstream from this event. To do this we first used small interfering RNA (siRNA) to deplete SMG6, and observed a considerable increase in the RFP:GFP ratio of our NMD reporter compared to its matched control (Fig. 2.6A-B, Fig. S2.7A-B). This phenotype could be rescued by ectopic expression of wild-type, but not a dominant negative inactive mutant Glavan et al., 2006, SMG6 for both our NMD and reverse reporters (Fig. 2.6B-C, Fig. S2.7B-D). Therefore, we concluded that the function SMG6 is required for both mRNA and nascent-chain degradation in NMD, and in both cases depends on its endonuclease activity.

2.3 Discussion

Recognition of an NMD-substrate occurs co-translationally, necessarily resulting in the production of a nascent, potentially cytotoxic polypeptide chain. NMD typically reduces the mRNA level of its substrates 2–50 fold, depending on the transcript and function of the resulting protein product: a reduction that may not be sufficient to maintain proteostasis in the cell. As such, there has been consideration of whether NMD leverages an additional, post-translational pathway to directly target these nascent proteins for degradation (Chu et al., 2021; Kuroha, Tatematsu, and Inada, 2009; Pradhan et al., 2021; Udy and Bradley, 2022).

There are two plausible strategies by which protein degradation of NMD nascent chain may occur. Since many NMD substrates are truncated and thus likely to misfold, they expose hydrophobic degrons that will be recognized by general cytosolic quality control machinery. However, this type of uncoordinated clearance strategy would risk the cell's exposure to transient dominant negative or gain-of-function activity of these truncated or aberrant proteins. In contrast, a coordinated protein quality control pathway that co-translationally initiates protein degradation prior to dissociation from the ribosome would be more consistent with other mRNA surveillance pathways. Indeed, tight coupling of quality control to biogenesis is a strategy used throughout biology to ensure robust and efficient clearance of mRNA and protein products that fail during their maturation (Rodrigo-Brenni and Hegde, 2012).

In the case of NMD, the lack of a robust *in vitro* reconstitution system; the difficulty of deconvoluting post-transcriptional versus post-translational effects on expression of NMD substrates; and the putative contribution of generalized quality control in turnover of the classical truncated NMD substrates has made it difficult to definitively identify this type of coordinated pathway. Using a fluorescent reporter strategy that addresses several of these technical challenges, we demonstrated that in mammals, NMD relies on a coupled protein quality control branch to concomitantly target the nascent protein for degradation via the ubiquitin proteasome pathway.

A coupled protein quality control branch of NMD

We propose the following working model for protein quality control during NMD in mammals (Fig. 2.7). As the ribosome reaches the stop codon during translational elongation, the protein composition of the downstream mRNA serves as the primary cue for initiating NMD. At this point, the nascent polypeptide remains tethered to

the ribosome via the peptidyl tRNA. We postulate that the early recognition steps between the mRNA and protein quality control branches of NMD are shared, and rely on core NMD factors such as UPF1, UPF2, UPF3b, and CASC3. NMD-coupled quality control is thus initiated through the canonical pathway for recognition of PTC-containing mRNAs that involves binding between the ribosome, NMD factors, and the downstream EJC (Gerbracht et al., 2020; Chamieh et al., 2008; Czaplinski et al., 1998; Kim, Kataoka, and Dreyfuss, 2001; Le Hir, Gatfield, et al., 2001). However, because our screens were designed to specifically query factors required for NMD-coupled protein quality control, we find substantial differences between hits identified here and those reported from earlier NMD RNA-degradation screens (Alexandrov, Shu, and Steitz, 2017; Baird et al., 2018; Sun et al., 2011; Zinshteyn et al., 2021). This discrepancy suggests that following recognition of an NMD substrate, the mRNA and protein quality control pathways diverge, relying on distinct sets of factors to target and degrade either the mRNA or nascent protein. However, the pathways are strictly linked, as evidenced by the requirement for endonucleolytic mRNA cleavage by SMG6 for efficient protein degradation.

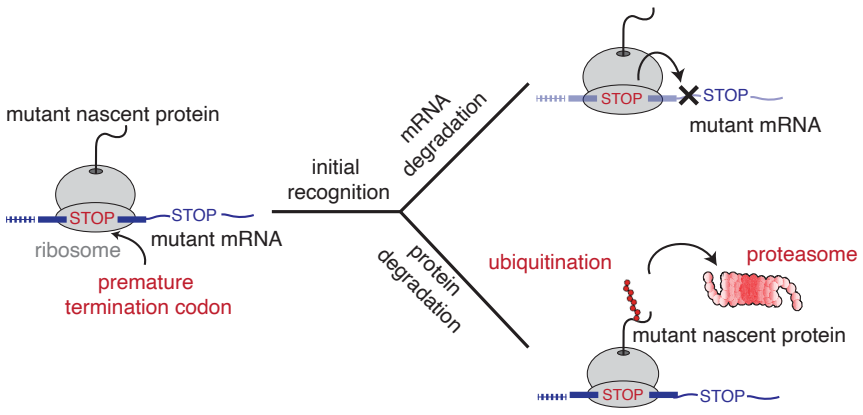


Figure 2.7: Model for NMD-coupled protein quality control. When the ribosome reaches the stop codon, NMD substrates are recognized in a context-dependent manner. These early recognition steps initiate two parallel pathways that rely on distinct suites of factors to concomitantly degrade the mRNA and nascent protein. We postulate that NMD-coupled quality control results in ubiquitination of the nascent protein prior to its release from the ribosome where it subsequently degraded by the proteasome.

We favor a model in which degradation of the nascent polypeptide is initiated prior to its release from the ribosome, as is common to other mRNA surveillance pathways and would minimize potential exposure of an aberrant protein to the cytosol. Consistent with this model we (i) found that only the nascent polypeptide tethered

to the ribosome at the stop codon is subjected to NMD-coupled degradation (Fig. 2.1D, Fig. 2.2B); and (ii) we observe an NMD-specific destabilization of an intact, folded protein compared to a matched control. We therefore concluded that the nascent protein must be somehow 'marked' for degradation prior to its dissociation from the ribosome. However, our data is consistent with earlier studies that suggest that multiple rounds of translation are required before an mRNA is committed to NMD-dependent degradation (Hoek et al., 2019). We similarly observe incomplete degradation of the nascent chain (RFP), in line with only a proportion of ribosomes eliciting NMD-dependent ubiquitination. Following ubiquitination of the nascent protein, it can then be safely released into the cytosol for degradation by the proteasome. In contrast to non-stop and no-go mRNA decay where the primary cue for protein quality control is ribosome stalling (Brandman and Hegde, 2016), NMD is initiated at a stop codon and thus may utilize the typical strategy for nascent protein release and ribosome recycling. The manner by which the nascent protein is recognized as emanating from an NMD substrate is unclear: it has been suggested that at least in yeast, termination at PTCs may occur more slowly than at a canonical stop codon, which could provide a kinetic window for ubiquitination of the nascent protein (Amrani et al., 2004); however, no evidence for this has been found in human cells (Karousis et al., 2020). We therefore cannot differentiate whether nascent protein ubiquitination occurs simultaneously or immediately following translational termination, but we favor a model where ubiquitination is initiated prior to dissociation of the nascent chain from the ribosome.

A potential role for the RQC pathway in NMD-coupled protein quality control

Several non-mutually exclusive models have been proposed for how to coordinate ubiquitination of the nascent protein chain prior to release. Experiments in Drosophila and C. elegans have suggested that at least in some systems, NMD and non-stop decay may be coupled, and levels of some mRNAs and their associated protein products are regulated by both pathways (Arribere and Fire, 2018; Hashimoto et al., 2017). A forward genetic screen in C. elegans further identified the canonical RQC factor PELO (the functional ortholog of dom34/Pelota) as required for repression of an NMD reporter. Based on these and other experiments, the authors proposed a model whereby quality control by NMD in initiated by endonucleolytic cleavage of the mRNA upstream of the stop codon by SMG6. Translation of the resulting truncated mRNA would result in stalling of subsequent ribosomes at its 3' end, triggering further repression at both the mRNA and protein level by the

non-stop decay and RQC pathways (Arribere and Fire, 2018).

If a similar mechanism was occurring in mammalian cells, post-translational degradation of NMD substrates would depend on the canonical RQC factors including the E3 ubiquitin ligase LTN1, and the ribosome rescue factors pelota and HBS1. However, the majority of RQC factors were not significant hits in either of our screens, though were identified in an earlier non-stop decay screen performed using matched conditions (Hickey et al., 2020). Further, depletion of LTN1 directly did not affect our NMD reporter under conditions that robustly stabilized a non-stop decay substrate (Fig. 2.4C). These results suggest that at least for the class of NMD substrates represented by our reporter, NMD-coupled protein degradation does not rely on the canonical RQC pathway. Together these data suggest a functional separation of nonsense and non-stop decay in mammals, as was observed in *S. cerevisiae* Arribere and Fire, 2018 and is consistent with the distinct molecular players identified by NMD versus non-stop mRNA decay screens (Hodgkin et al., 1989; Leeds et al., 1991; Pulak and Anderson, 1993; Wilson, Meaux, and Hoof, 2007).

Direct ubiquitination of the nascent NMD polypeptide

The simplest model for NMD-coupled protein degradation is the direct recruitment of an E3 ligase that ubiquitinates the nascent chain while it remains tethered to the ribosome. Earlier studies have suggested that UPF1, a RING domain E3 ubiquitin ligase and core NMD factor that interacts with both the ribosome and eukaryotic release factors, could carry out this role. UPF1 knockdown has been shown to stabilize protein products produced from NMD substrates mRNAs (Kuroha, Tatematsu, and Inada, 2009; Kuroha, Ando, et al., 2013; Feng, Jagannathan, and Bradley, 2017; Park et al., 2020; Kadlec et al., 2006; Takahashi, Araki, Ohya, et al., 2008). Consistent with these reports, UPF1 was identified in our knockdown screen, and depletion of UPF1 stabilized both the mRNA and protein levels of our NMD reporter. However, we found that point mutations to UPF1 that specifically affect its ability to recruit its E2 ubiquitin-conjugating enzyme while leaving its ribosome-binding and helicase domains intact, did not have any effect on the protein-degradation phenotype of our reporter.

We therefore concluded that UPF1 is required for NMD-coupled protein quality control, but plays a role that does not depend on its E3 ubiquitin ligase activity. To reconcile these results with previous studies, we propose that UPF1 is involved in the early recognition steps of NMD substrates, which affects both the mRNA and

protein degradation branches of NMD. However, our data are inconsistent with a direct role for UPF1 in ubiquitination of the nascent polypeptide. A dedicated E3 ubiquitin ligase that specifically recognizes nascent chains from NMD substrates was not identified through either the knockdown or knockout genome-wide screens. This is either a limitation of the reporter design, or more likely suggests redundancy between E3s in the recognition event.

Implications of nascent protein degradation in proteostasis

The identification of a tightly coupled protein degradation branch of NMD has several immediate implications. Most notably, destabilization at the post-translational level will increase the suppression of NMD substrates. Though we find the effects of NMD-coupled protein degradation on our reporters to be modest (2-fold), in the context of the cell or an organism, this additional level of regulation may be critical to prevent deleterious or off-target effects. Effects on these fluorescent reporters, which are both over-expressed and in which phenotypes require degradation of the remarkably stable RFP moiety, may also underestimate the true effect size on an endogenous substrate. There are numerous physiologically relevant examples where NMD's role in transcriptome regulation, and subsequent production of potentially aberrant proteins, require stringent clearance of the nascent product. During histone production, synthesis must be tightly regulated in a manner coupled to the progression of the cell cycle, and the production of even small amounts of downregulated proteins could be problematic. Our results also have implications for viral infection. Co-translational protein degradation is thought to be a key source of peptides for MHC presentation (Balistreri et al., 2014; Fontaine et al., 2018; Wada et al., 2018; Yewdell and Nicchitta, 2006), with viral messages often targeted by NMD.

Finally, NMD plays an important role in a wide range of genetic diseases: over one third of all human genetic disorders are caused by PTC-creating mutations, including muscular dystrophy and cystic fibrosis. While generally protective, for numerous disease-causing mutations the NMD pathway contributes to pathogenesis by suppressing expression of partially functional mutant proteins (11% of mutations that cause human disease (Mort et al., 2008). The characterization of a second, parallel branch of NMD and the initial identification of potential factors involved in NMD-coupled protein quality control therefore may represent a valuable platform from which to identify potential targets for the new therapeutic strategies.

2.4 Materials and methods

Plasmids and antibodies

Reporter constructs for expression in mammalian cells were generated in either the pcDNA5/FRT/TO (Thermo Scientific) backbone (for expression in HEK293T cells) or the SFFV-tet3G lentiviral backbone with a 3' WPRE element (for inducible expression in K562 cells, from (Jost et al., 2017)). To create the NMD reporters described in Fig. 2.1, a fragment of the beta-globin gene spanning the last 221 nucleotides of exon 2 (the last 35 nucleotides for inert EJC), intron 2 and 129 nucleotides of exon 3 was amplified via PCR from human genomic DNA as described previously (Pereverzev et al., 2015). Either one or two copies were inserted into the 3' UTR of a plasmid encoding GFP-P2A-RFP to generate NMD1 and NMD2 respectively. In the lentiviral constructs, the reporters were inserted in reverse orientation to prevent splicing of the introns during lentiviral production. The presence of functional introns was checked via PCR, using primers that should span the introns (Fig. S2.1B). For this, the RNeasy kit (#74104, Qiagen) was used to purify total RNA from HEK293T transiently expressing the NMD1, 2, or the inert EJC reporter. cDNA was obtained by reverse transcription using the SuperScript III First Strand Synthesis SuperMix (#11752, Invitrogen). PCR amplification from this cDNA with respective primers generated a shorter fragment than that of the reporter plasmids, indicating the introns have been spliced out efficiently.

Modifications of the NMD constructs were created by either replacing the P2A site with a glycine-serine linker of identical length for the linked constructs (Fig. 2.1D), reversing the order of the GFP and RFP for the 'reverse' constructs as in (Chu et al., 2021), or appending the villin headpiece domain (bVHP) downstream of the RFP (Fig. S2.2D). For immunoprecipitation experiments, a FLAG tag was appended to the N-terminus of RFP (Fig. 2.2C). Of note, mCherry and mEGFP versions of the GFP and RFP were used throughout this study, but for simplicity are referred to as GFP and RFP.

cDNA for UPF1 was acquired from Addgene (#99146) and cloned downstream of a BFP-P2A sequence contained in a lentiviral backbone. This was driven by an EF1α promoter from an upstream ubiquitous chromatin opening element (UCOE). The main isoform of UPF1 (isoform 2) was used, as it has been more comprehensively characterized (Nicholson et al., 2014; Fritz et al., 2022). A mutant of UPF1 with mutations in the RING domain (S134A, N148A, T149A) that disrupts binding with E2 ligases was also acquired from Addgene (#99144). Plasmids containing siRNA-

resistant FLAG-tagged SMG6 (wild-type and a D1353A mutant) were a kind gift from Niels Gehring.

To generate knock-downs, single guides against LTN1 (GACTCTGAGCACTCAGACCC), CASC3 (GTGCGTAAGTACCTCGCCGG), and DCP1A (GGCGCTGAGTCGAGCTGGGC) were generated by annealed cloning of top and bottom oligonucleotides (Integrated DNA Technologies, Coralville, IA) into a lentiviral pU6-sgRNA EF1α-Puro-T2A-BFP vector digested with BstXI/BlpI (Addgene, #84832). BFP was removed when the color interfered with the reporter construct. In certain cases, we used a programmed dual sgRNA guide vector (Addgene #140096) to increase the efficiency of knock-down such as for UPF1 (GGCCGCTCGCAGCCTAGAGC and GTTCGAGGGGAGCTGAGGCG) and CNOT4 (GGAGACTCTCAGCTTTCGGT and GGGGCCACCATCTTACATTA).

The following antibodies were used in this study: FLAG (#A2220, Sigma, 1:10,000), HA (#A2095, Sigma, 1:1,000), UPF1 (#A300-038A, Bethyl, 1:1,000), α-tubulin (#T9026, Sigma, 1:5,000), CNOT4 (#12564-1-AP, Proteintech, 1:1,000), SMG6 (#ab87539, Abcam, 1:1,000). Antibodies against GFP and RFP were a kind gift from Ramanujan Hegde. Secondary antibodies used were HRP-conjugated anti-Rabbit (#170-6515, BioRad, 1:5,000) and anti-Mouse (#172-1011, BioRad, 1:5,000), and HRP-conjugated Donkey anti-Goat (ab97110, Abcam, 1:5,000).

siRNAs

Pre-designed Silencer Select siRNAs were ordered from ThermoFisher: control (scrambled 1) and SMG6 (s23489).

Mammalian cell culture

HEK293T cells were grown in Dulbecco's modified eagle medium (DMEM) with 10% FBS (Atlanta Biologicals, #S11550) and 2 mM L-glutamine (Invitrogen, #25030081). siRNA treatments were performed according to manufacturer's instructions in a 6-well plate with 30 pmol of each siRNA, allowing knock-down for a total of 72 hours. siRNA treated cells were transiently transfected with 1 μg of reporter construct DNA 24 hours prior to harvesting.

Stable HEK293 cell lines were generated using Flp-In 293 T-Rex cells purchased from Thermo Fisher Scientific (USA) (RRID: CVCL_U427). Cell lines were grown in DMEM supplemented with 2 mM glutamine, 10% (w/v) FBS, 15 µg/ml Blasticidine S, and 100 µg/ml Zeocin. The open-reading frame to be integrated into the

genomic FRT site was cloned into the pcDNA5/FRT/TO vector backbone and cell lines were generated according to the manufacturer's protocol. Briefly, the reporter construct was transfected together with pOG44 Flp-In recombinase in a 9:1 ratio using Trans-IT 293 transfection reagent (Mirus, USA) according to the manufacturer's instructions. 48 hours after transfection, 100 µg/ml Hygromycin B was used to select for cells that had undergone successful integration.

K562-dCas9-BFP-KRAB Tet-On cells were grown in RPMI-1640 medium with L-Glutamine and HEPES supplemented with 10% Tet System Approved FBS, 100 units/mL penicillin and 100 μg/mL streptomycin (Invitrogen, #15140148). For certain reporter assays, K562 CRISPRi Zim3-hygro Tet-On cells were used to promote better knock-down (Replogle et al., 2022). Cells were maintained at a confluency between 0.5–2 × 106 cells/ml. All cells were tested for contamination regularly.

Lentivirus

Lentivirus was produced by co-transfecting HEK293T cells with two packaging plasmids (pCMV-VSV-G and delta8.9, Addgene #8454) and the desired plasmid using TransIT-293 (Mirus) transfection reagent. 48 hours after transfection, the supernatant was collected, centrifuged and flash frozen. In all instances, virus was rapidly thawed prior to transfection. Virus for the genome-wide CRISPRi screen was generated using this method.

Virus generation for genome wide CRISPR knockout screen HEK-293T cells were seeded at a density of 750,000 cells/ml in 20 ml viral production medium: IMDM (Thermo Fisher Scientific #1244053) supplemented with 20% inactivated fetal serum (GeminiBio #100-106). After 24 hours, media was changed to fresh viral production medium. At 32 hours post-seeding, cells were transfected with a mix containing 76.8 μL Xtremegene-9 transfection reagent (Sigma Aldrich #06365779001), 3.62 μg pCMV-VSV-G (Plasmid #8454, Addgene), 8.28 μg psPAX2 (Plasmid #12260, Addgene), and 20 μg sgRNA plasmid and Opti-MEM (Thermo Fisher Scientific #11058021) to a final volume of 1 ml. Media was changed 16 hours later to fresh viral production medium. At 48 hours after transfection, virus was collected and filtered through a 0.45 μm filter, aliquoted, and stored at -80°C until use.

Generation of K562 reporter cell lines for screening

K562 reporter reporter cell lines were generated by co-transfecting our control or NMD2 viral vectors along with a tet activator element into K562 wild type or K562-dCas9-BFP-KRAB Tet-On cell lines at one copy number per cell. Positive cells

were isolated via FACS on a BD FACSAria2 and grown up to create monoclonal cell lines.

Flow cytometry analysis

HEK293T cells were analyzed by flow cytometry 24 hours after either transient transfection with indicated reporters. T-Rex HEK293 cells stably expressing designated reporters were induced for 24 hours prior to harvesting for flow cytometry. For this, cells were first incubated with trypsin before collection, pelleted, and resuspended in 300 μL of PBS containing 1 μM Sytox Blue Dead Cell Stain (ThermoFisher, #S34857) and analyzed on a Miltenyi Biotech MACSQuant VYB Flow Cytometer. For certain experiments, such as treatment with MG132, K562-dCas9-BFP-KRAB Tet-On NMD2 or control monoclonal cell lines (also used for screening) were induced for 24 hours with 1 μg/ml doxycycline. For transient reporter experiments, K562 Zim3 or KRAB CRISPRi cells were spinfected at a confluency 0.5 × 106 cells/ml. Media was supplemented with 8 μg/ml polybrene (Millipore Sigma, #107689-100G) and the lentivirus of interest was added to the well. The components were mixed by pipetting, and immediately spun down at 1000xg for 2 hours at 30 °C. Expression of the reporter constructs was induced with 1 μg/ml doxycycline and cells were typically analyzed 24 hours later unless otherwise indicated. To generate knock-down, cells were spinfected with both guide and reporter, allowed to grow for 8-10 days and the induced with doxycycline. Guide positive cells were selected with 1 μg/lL puromycin for three days. Flow cytometry data was analyzed either in FlowJo v10.8 Software (BD Life Sciences) or Python using the FlowCytometryTools package.

qPCR analysis

Relative mRNA levels were determined by quantitative PCR. Total cellular RNA was purified from cells using the RNeasy kit (#74104, Qiagen), treated with DNase I (#18068015, Invitrogen) and reverse transcribed using the SuperScript III First Strand Synthesis SuperMix (#11752, Invitrogen), before being subjected to analysis on a StepOnePlus Real-Time PCR system. The relative expression ratios between sample cDNA levels were then analyzed, using primers that amplified either GFP and RFP, and the housekeeping gene HPRT1 (IDT, Hs.PT.58v.45621572). Each set of primers was checked against a standard dilution curve, and the primer efficiencies were between 90 and 110%. The efficiencies were considered in the expression ratio calculation. The primers used were: GFP (fwd: ATTGGACGGAGACGTGAATG,

rev: GTTTCCCGGTAGTGCAGATAA) and RFP (fwd: CCCGCAGACATTCCT-GATTA, rev: AGTCCTGAGTCACTGTAACAAC).

Inhibition of the ubiquitin-proteasome pathway

To look at the effect of MG132 treatment on the NMD2 reporter as shown in Fig. 2.2A, wild-type HEK293T cells were transiently transfected with FLAG-tagged versions of the reporter constructs. 18 hours later, cells were then treated with either 10 μM of the proteasome inhibitor MG132 (Calbiochem, #474790), or a DMSO control for 6 hours. To test the effect of E1 inhibition, this was modified such that cells were treated with either 10 μM of the E1 inhibitor MLN7243 or DMSO for 8 hours. To allow for blotting, cells were then harvested and lysed in 1% SDS. The lysates were normalized to GFP protein levels by serial dilutions and Western-blotting. The normalized lysates were analyzed by SDS-PAGE and Western-blotting using Anti-FLAG and Anti-GFP antibodies. For Fig. 2.2B, our K562 CRISPRi NMD2 monoclonal cell line was induced with 1 μg/ml doxycycline for 10 hours and subsequently treated with 10 μM MG132 or DMSO for 6 hours. Cells were harvested and analyzed by flow cytometry on an Attune NxT Flow Cytometer.

To directly observe ubiquitination of RFP and GFP (Fig. 2.2C), we generated a stable cell line constitutively expressing HA-tagged ubiquitin in HEK293T cells, with a BFP marker. These cells were transiently transfected with reporters where the RFP was FLAG-tagged, and incubated for 42 hours. Cells were then treated with 10 μM MG132 for 6 hours. For blots, cells were harvested by first being resuspended in lysis buffer (50 mM Hepes pH 7.4, 100 mM KOAc, 2 mM MgAc2, 1 × proteasome inhibitor, 1 mM DTT, 50 μM PR-619, 10 μg/ml digitonin) and left on ice for 15 minutes. Mechanical lysis was performed with 10 strokes of a glass dounce and total samples were taken. The amount of RFP and GFP in each sample was determined using a plate reader. Samples for RFP and GFP immunoprecipitations (IPs) were normalized to equivalent RFP and GFP levels, respectively, using HA-Ub-containing cell lysate to maintain the total protein concentration. For the RFP IP: SDS was added to 1% final concentration, and the samples were boiled. They were then diluted with IP buffer (50 mM Hepes pH 7.4, 100 mM KOAc, 2 mM MgAc2, 1% Triton) to a final concentration of 0.1% SDS. Samples were immunoprecipitated with Anti-FLAG M2 affinity resin (Millipore-Sigma) and eluted with SDS. For the GFP IPs: SDS was added to 1% final concentration, then samples were diluted with IP buffer without boiling. Magnetic beads (Pierce) were coupled to a biontinylated

version of a GFP nanobody as described previously (Pleiner et al., 2020), and then were used to immunoprecipitate GFP. Samples were eluted with SDS. The resulting samples were analyzed by Western blotting.

CRISPRi knockdown screen

The genome-scale CRISPRi screen was performed similarly to previously described screens (Gilbert et al., 2014; Horlbeck et al., 2016). The hCRISPRi-v2 compact library (containing 5 sgRNAs per gene, Addgene pooled library #83969) was transduced in duplicate into 330 million K562-dCas9-BFP-KRAB Tet-On-NMD2 cells at MOI < 1 (percentage of transduced cells 48 hours after infection as measured by BFP positive cells: 20%-40%). Cells were grown in 1L of media in 1L spinner flasks (Bellco, SKU: 1965-61010) for the duration of the screen. 48 hours after spinfection, cells were selected with 1 mg/ml puromycin for 3 days. After a 36 hours recovery, cells were induced with 1 μg/ml doxycycline for 24 hours and sorted on a FACS AriaII Fusion Cell Sorter. The cells were maintained at 0.5 × 106 cells/ml for the duration of the screen. This ensured that the culture was maintained at an average coverage of more than 1000 cells per sgRNA for the whole screen.

Cells with high BFP (transduced cells) and with both GFP and RFP signal (successfully induced) were gated. Cells were sorted according to the RFP:GFP ratio of this population.

Around 40 million cells with either the highest (30%) and the lowest (30%) RFP:GFP ratio were collected, pelleted and flash-frozen. Genomic DNA was purified using the Nucleospin Blood XL kit (Takara Bio, #740950.10) and amplified with barcoded primers by index PCR. The library (264 bp) was purified using SPRIbeads (Bulldog Bio, CNGS005), its concentration measured by Qubit fluorometer (Invitrogen) and its integrity checked by Agilent 2100 Bioanalyzer. Samples were analyzed using an Illumina HiSeq2500 high throughput sequencer. Sequencing reads were aligned to the CRISPRi v2 library sequences, counted and quantified (Horlbeck et al., 2016). Generation of negative control genes and calculation of phenotype scores and Mann-Whitney p-values was performed as described previously (Gilbert et al., 2014; Horlbeck et al., 2016). Gene-level phenotypes and counts are available in Supplementary Table 1.

K562 genome-wide CRISPR knockout screen

A genome-wide lentiviral sgRNA library in a Cas9-containing vector (Supplementary Table 3) was used to transduce 500 million a monoclonal K562 cell line con-

taining a tet element and the NMD2 reporter. All other conditions were identical to those used for the CRISPRi KD screen. Cells were induced either at 7, 9, or 11 days with 1 μ g/ml doxycycline for 24 hours and sorted on a FACS Aria II Fusion cell Sorter on days 8, 10, or 12. Data was processed using the pipeline described above and validated by analysis using MAGeCK (Li et al., 2014). Gene-level phenotypes and counts are available in Supplementary Table 2.

For extraction of genomic DNA, QIAamp DNA Blood Maxiprep Kit (Qiagen) was used according to manufacturer's instructions with the following modifications: 500 µL of a 10 mg/ml solution of ProteinaseK in water was used in place of QIAGEN Protease; incubation with ProteinaseK and Buffer AL was performed overnight; centrifugation steps after Buffer AW1 and AW2 were performed for 2 min and 5 min, respectively; gDNA was eluted for 5 min using 1 ml of water preheated to 70 °C, followed by centrifugation for 5 min. gDNA concentration was determined using the Qubit dsDNA HS Assay kit (Thermo Fisher Scientific #Q32851).

References

Alexandrov, Andrei, Mei-Di Shu, and Joan A. Steitz (2017). "Fluorescence amplification method for forward genetic discovery of factors in human mRNA degradation." In: *Molecular Cell* 65.1, pp. 191–201.

Amrani, Nadia et al. (2004). "A faux 3'-UTR promotes aberrant termination and triggers nonsense-mediated mRNA decay." In: *Nature* 432.7013, pp. 112–118.

Arribere, Joshua A. and Andrew Z. Fire (2018). "Nonsense mRNA suppression via nonstop decay." In: *eLife* 7, e33292.

Baird, Thomas D. et al. (2018). "ICE1 promotes the link between splicing and nonsense-mediated mRNA decay." In: *eLife* 7, e33178.

Balistreri, Giuseppe et al. (2014). "The host nonsense-mediated mRNA decay pathway restricts mammalian RNA virus replication." In: *Cell Host & Microbe* 16.3, pp. 403–411.

Balleza, Enrique, J. Mark Kim, and Philippe Cluzel (2018). "Systematic characterization of maturation time of fluorescent proteins in living cells." In: *Nature Methods* 15.1, pp. 47–51.

Becker, Thomas et al. (2011). "Structure of the no-go mRNA decay complex Dom34–Hbs1 bound to a stalled 80S ribosome." In: *Nature Structural & Molecular Biology* 18.6, pp. 715–720.

Belgrader, Phillip, Jiu Cheng, and Lynne E. Maquat (1993). "Evidence to implicate translation by ribosomes in the mechanism by which nonsense codons reduce the

nuclear level of human triosephosphate isomerase mRNA." In: *Proceedings of the National Academy of Sciences* 90.2, pp. 482–486.

Bono, Fulvia et al. (2006). "The crystal structure of the exon junction complex reveals how it maintains a stable grip on mRNA." In: *Cell* 126.4, pp. 713–725.

Brandman, Onn and Ramanujan S. Hegde (2016). "Ribosome-associated protein quality control." In: *Nature Structural & Molecular Biology* 23.1, pp. 7–15.

Brandman, Onn, Jacob Stewart-Ornstein, et al. (2012). "A ribosome-bound quality control complex triggers degradation of nascent peptides and signals translation stress." In: *Cell* 151.5, pp. 1042–1054.

Chamieh, Hala et al. (2008). "NMD factors UPF2 and UPF3 bridge UPF1 to the exon junction complex and stimulate its RNA helicase activity." In: *Nature structural & molecular biology* 15.1, pp. 85–93.

Chen, Chyi-Ying A and Ann-Bin Shyu (2003). "Rapid deadenylation triggered by a nonsense codon precedes decay of the RNA body in a mammalian cytoplasmic nonsense-mediated decay pathway." In: *Molecular and Cellular Biology* 23.14, pp. 4805–4813.

Chu, Vincent et al. (2021). "Selective destabilization of polypeptides synthesized from NMD-targeted transcripts." In: *Molecular Biology of the Cell* 32.22, ar38.

Czaplinski, Kevin et al. (1998). "The surveillance complex interacts with the translation release factors to enhance termination and degrade aberrant mRNAs." In: *Genes & Development* 12.11, pp. 1665–1677.

Defenouillère, Quentin et al. (2013). "Cdc48-associated complex bound to 60S particles is required for the clearance of aberrant translation products." In: *Proceedings of the National Academy of Sciences* 110.13, pp. 5046–5051.

Doma, Meenakshi K. and Roy Parker (2006). "Endonucleolytic cleavage of eukaryotic mRNAs with stalls in translation elongation." In: *Nature* 440.7083, pp. 561–564.

Durand, Sébastien and Jens Lykke-Andersen (2013). "Nonsense-mediated mRNA decay occurs during eIF4F-dependent translation in human cells." In: *Nature Structural & Molecular Biology* 20.6, pp. 702–709.

Eberle, Andrea B et al. (2009). "SMG6 promotes endonucleolytic cleavage of nonsense mRNA in human cells." In: *Nature structural & molecular biology* 16.1, pp. 49–55.

Feng, Qing, Sujatha Jagannathan, and Robert K. Bradley (2017). "The RNA surveillance factor UPF1 represses myogenesis via its E3 ubiquitin ligase activity." In: *Molecular Cell* 67.2, pp. 239–251.

Fontaine, Krystal A. et al. (2018). "The cellular NMD pathway restricts Zika virus infection and is targeted by the viral capsid protein." In: *MBio* 9.6, e02126–18.

Frischmeyer, Pamela A. et al. (2002). “An mRNA surveillance mechanism that eliminates transcripts lacking termination codons.” In: *Science* 295.5563, pp. 2258–2261.

Fritz, Sarah E. et al. (2022). “An alternative UPF1 isoform drives conditional remodeling of nonsense-mediated mRNA decay.” In: *The EMBO Journal* 41.10, e108898.

Gerbracht, Jennifer V. et al. (2020). “CASC3 promotes transcriptome-wide activation of nonsense-mediated decay by the exon junction complex.” In: *Nucleic Acids Research* 48.15, pp. 8626–8644.

Gilbert, Luke A. et al. (2014). “Genome-scale CRISPR-mediated control of gene repression and activation.” In: *Cell* 159.3, pp. 647–661.

Glavan, Filip et al. (2006). “Structures of the PIN domains of SMG6 and SMG5 reveal a nuclease within the mRNA surveillance complex.” In: *The EMBO Journal* 25.21, pp. 5117–5125.

Hart, Traver et al. (2017). “Evaluation and design of genome-wide CRISPR/SpCas9 knockout screens.” In: *G3: Genes, Genomes, Genetics* 7.8, pp. 2719–2727.

Hashimoto, Yoshifumi et al. (2017). “Nonstop-mRNA decay machinery is involved in the clearance of mRNA 5’-fragments produced by RNAi and NMD in Drosophila melanogaster cells.” In: *Biochemical and Biophysical Research Communications* 484.1, pp. 1–7.

Hickey, Kelsey L. et al. (2020). “GIGYF2 and 4EHP inhibit translation initiation of defective messenger RNAs to assist ribosome-associated quality control.” In: *Molecular Cell* 79.6, pp. 950–962.

Hodgkin, Jonathan. et al. (1989). “A new kind of informational suppression in the nematode Caenorhabditis elegans.” In: *Genetics* 123.2, pp. 301–313.

Hoek, Tim A. et al. (2019). “Single-molecule imaging uncovers rules governing nonsense-mediated mRNA decay.” In: *Molecular Cell* 75.2, pp. 324–339.

Horlbeck, Max A. et al. (2016). “Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation.” In: *eLife* 5, e19760.

Jost, Marco et al. (2017). “Combined CRISPRi/a-based chemical genetic screens reveal that rigosertib is a microtubule-destabilizing agent.” In: *Molecular Cell* 68.1, pp. 210–223.

Juszkiewicz, Szymon et al. (2018). “ZNF598 is a quality control sensor of collided ribosomes.” In: *Molecular Cell* 72.3, pp. 469–481.

Kadlec, Jan et al. (2006). “Crystal structure of the UPF2-interacting domain of nonsense-mediated mRNA decay factor UPF1.” In: *RNA* 12.10, pp. 1817–1824.

Karousis, Evangelos D. et al. (2020). “Human NMD ensues independently of stable ribosome stalling”. In: *Nature Communications* 11.1, p. 4134.

Kim, V. Narry, Naoyuki Kataoka, and Gideon Dreyfuss (2001). "Role of the nonsense-mediated decay factor hUpf3 in the splicing-dependent exon-exon junction complex." In: *Science* 293.5536, pp. 1832–1836.

Kuroha, Kazushige, Koji Ando, et al. (2013). "The Upf factor complex interacts with aberrant products derived from mRNAs containing a premature termination codon and facilitates their proteasomal degradation." In: *Journal of Biological Chemistry* 288.40, pp. 28630–28640.

Kuroha, Kazushige, Tsuyako Tatematsu, and Toshifumi Inada (2009). "Upf1 stimulates degradation of the product derived from aberrant messenger RNA containing a specific nonsense mutation by the proteasome." In: *EMBO Reports* 10.11, pp. 1265–1271.

Le Hir, Hervé, David Gatfield, et al. (2001). "The exon–exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay." In: *The EMBO Journal* 20.17, pp. 4987–4997.

Le Hir, Hervé, Elisa Izaurralde, et al. (2000). "The spliceosome deposits multiple proteins 20–24 nucleotides upstream of mRNA exon–exon junctions". In: *The EMBO journal* 19.24, pp. 6860–6869.

Leeds, Peltz et al. (1991). "The product of the yeast UPF1 gene is required for rapid turnover of mRNAs containing a premature translational termination codon." In: *Genes & Development* 5.12a, pp. 2303–2314.

Lyumkis, Dmitry et al. (2014). "Structural basis for translational surveillance by the large ribosomal subunit-associated protein quality control complex." In: *Proceedings of the National Academy of Sciences* 111.45, pp. 15981–15986.

Mitchell, Philip and David Tollervey (2003). "An NMD pathway in yeast involving accelerated deadenylation and exosome-mediated 3' -> 5' degradation." In: *Molecular Cell* 11.5, pp. 1405–1413.

Mort, Matthew et al. (2008). "A meta-analysis of nonsense mutations causing human genetic disease". In: *Human mutation* 29.8, pp. 1037–1047.

Nagy, Eszter and Lynne E. Maquat (1998). "When nonsense affects mRNA abundance: a rule for termination codon position within intron-containing genes." In: *Trends in Biochemical Sciences* 23, pp. 198–199.

Nicholson, Pamela et al. (2014). "A novel phosphorylation-independent interaction between SMG6 and UPF1 is essential for human NMD." In: *Nucleic Acids Research* 42.14, pp. 9217–9235.

Nott, Ajit, Hervé Le Hir, and Melissa J. Moore (2004). "Splicing enhances translation in mammalian cells: an additional function of the exon junction complex." In: *Genes & Development* 18.2, pp. 210–222.

Park, Yeonkyoung et al. (2020). "Nonsense-mediated mRNA decay factor UPF1 promotes aggresome formation." In: *Nature Communications* 11.1, p. 3106.

Pereverzev, Anton P. et al. (2015). "Method for quantitative analysis of nonsense-mediated mRNA decay at the single cell level." In: *Scientific Reports* 5.1, p. 7729.

Pisareva, Vera P. et al. (2011). "Dissociation by Pelota, Hbs1 and ABCE1 of mammalian vacant 80S ribosomes and stalled elongation complexes." In: *The EMBO Journal* 30.9, pp. 1804–1817.

Pleiner, Tino et al. (2020). "Structural basis for membrane insertion by the human ER membrane protein complex." In: *Science* 369.6502, pp. 433–436.

Popp, Maximilian Wei-Lin and Lynne E. Maquat (2013). "Organizing principles of mammalian nonsense-mediated mRNA decay." In: *Annual Review of Genetics* 47, pp. 139–165.

Pradhan, Ashis Kumar et al. (2021). "Ribosome-associated quality control mediates degradation of the premature translation termination product Orf1p of ODC antizyme mRNA." In: *FEBS Letters* 595.15, pp. 2015–2033.

Pulak, Rock and Philip Anderson (1993). "mRNA surveillance by the Caenorhabditis elegans smg genes." In: *Genes & Development* 7.10, pp. 1885–1897.

Replogle, Joseph M. et al. (2022). "Maximizing CRISPRi efficacy and accessibility with dual-sgRNA libraries and optimal effectors." In: *eLife* 11, e81856.

Rodrigo-Brenni, Monica C. and Ramanujan S. Hegde (2012). "Design principles of protein biosynthesis-coupled quality control." In: *Developmental Cell* 23.5, pp. 896–907.

Rosenbluh, Joseph et al. (2017). "Complementary information derived from CRISPR Cas9 mediated gene deletion and suppression." In: *Nature Communications* 8.1, p. 15403.

Shao, Sichen, Alan Brown, et al. (2015). "Structure and assembly pathway of the ribosome quality control complex." In: *Molecular Cell* 57.3, pp. 433–444.

Shao, Sichen, Jason Murray, et al. (2016). "Decoding mammalian ribosome-mRNA states by translational GTPase complexes." In: *Cell* 167.5, pp. 1229–1240.

Shao, Sichen, Karina Von der Malsburg, and Ramanujan S. Hegde (2013). "Listerin-dependent nascent protein ubiquitination relies on ribosome subunit dissociation." In: *Molecular Cell* 50.5, pp. 637–648.

Shoemaker, Christopher J. and Rachel Green (2012). "Translation drives mRNA quality control". In: *Nature Structural & Molecular Biology* 19.6, pp. 594–601.

Sun, Yinyan et al. (2011). "A genome-wide RNAi screen identifies genes regulating the formation of P bodies in C. elegans and their functions in NMD and RNAi." In: *Protein & Cell* 2, pp. 918–939.

Takahashi, Shinya, Yasuhiro Araki, Yuriko Ohya, et al. (2008). "Upf1 potentially serves as a RING-related E3 ubiquitin ligase via its association with Upf3 in yeast." In: *RNA* 14.9, pp. 1950–1958.

Takahashi, Shinya, Yasuhiro Araki, Takeshi Sakuno, et al. (2003). "Interaction between Ski7p and Upf1p is required for nonsense-mediated 3'-to-5' mRNA decay in yeast." In: *The EMBO Journal* 22.15, pp. 3951–3959.

Udy, Dylan B. and Robert K. Bradley (2022). "Nonsense-mediated mRNA decay uses complementary mechanisms to suppress mRNA and protein accumulation." In: *Life Science Alliance* 5.3.

Van Hoof, Ambro et al. (2002). "Exosome-mediated recognition and degradation of mRNAs lacking a termination codon". In: *Science* 295.5563, pp. 2262–2264.

Verma, Rati, Robert S. Oania, et al. (2013). "Cdc48/p97 promotes degradation of aberrant nascent polypeptides bound to the ribosome." In: *eLife* 2, e00308.

Verma, Rati, Kurt M. Reichermeier, et al. (2018). "Vms1 and ANKZF1 peptidyl-tRNA hydrolases release nascent chains from stalled ribosomes." In: *Nature* 557.7705, pp. 446–451.

Wada, Masami et al. (2018). "Interplay between coronavirus, a cytoplasmic RNA virus, and nonsense-mediated mRNA decay pathway." In: *Proceedings of the National Academy of Sciences* 115.43, E10157–E10166.

Wang, Jun, Vita M Vock, et al. (2002). "A quality control pathway that down-regulates aberrant T-cell receptor (TCR) transcripts by a mechanism requiring UPF2 and translation." In: *Journal of Biological Chemistry* 277.21, pp. 18489–18493.

Wang, Yuancheng, Feng Wang, et al. (2015). "2A self-cleaving peptide-based multi-gene expression system in the silkworm Bombyx mori." In: *Scientific Reports* 5.1, p. 16273.

Wilson, Marenda A., Stacie Meaux, and Ambro van Hoof (2007). "A genomic screen in yeast reveals novel aspects of nonstop mRNA metabolism." In: *Genetics* 177.2, pp. 773–784.

Yewdell, Jonathan W. and Christopher V. Nicchitta (2006). "The DRiP hypothesis decennial: support, controversy, refinement and extension." In: *Trends in Immunology* 27.8, pp. 368–373.

Zhang, Jing and Lynne E. Maquat (1997). "Evidence that translation reinitiation abrogates nonsense-mediated mRNA decay in mammalian cells." In: *The EMBO journal* 16.4, pp. 826–833.

Zhang, Jing, Xiaolei Sun, et al. (1998). "Intron function in the nonsense-mediated decay of β-globin mRNA: indications that pre-mRNA splicing in the nucleus can influence mRNA translation in the cytoplasm." In: *RNA* 4.7, pp. 801–815.

Zinshteyn, Boris et al. (2021). "Translational repression of NMD targets by GIGYF2 and EIF4E2." In: *PLoS Genetics* 17.10, e1009813.

Zurita Rendón, Olga et al. (2018). "Vms1p is a release factor for the ribosome-associated quality control complex." In: *Nature Communications* 9.1, pp. 1–9.

2.5 Supplementary section

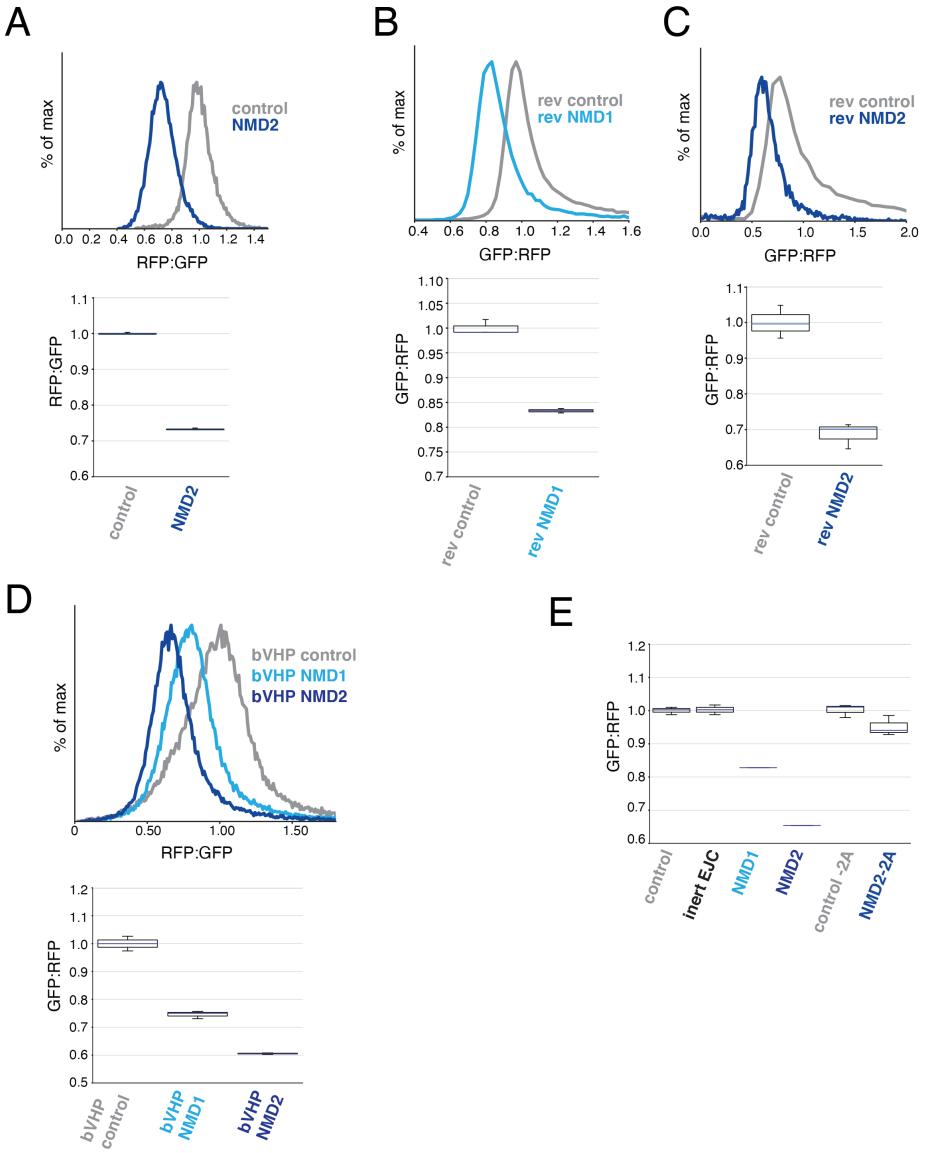


Figure S2.1: NMD-linked protein degradation is independent of cell type, fluorescent protein identity and reporter design. **(A)** K562 CRISPRi cells were virally infected with the control and NMD2 reporters and were then analyzed by flow cytometry after 24 hours of doxycycline induction. Box plots showing the results of three biological replicates are shown below. **(B)** HEK293T cells were transiently transfected with the reversed reporters (in which the GFP and RFP order is reversed), and were analyzed after 24 hours. A box plot showing three biological replicates is below. Note that the NMD1 reverse reporter was used. **(C)** As in A but for the reverse reporters. **(D)** HEK293T cells were transiently transfected with reporters in which a hydrophilic linker domain (bVHP) was inserted between the RFP and the stop codon to ensure the RFP would be fully emerged from the ribosome at the stop codon. The cells were analyzed by flow cytometry after 24 hours, and the results are shown as a histogram. A box plot showing three biological replicates is shown below. **(E)** A box plot showing quantification of three biological replicates for the flow cytometry data shown in Fig. 2.1D.

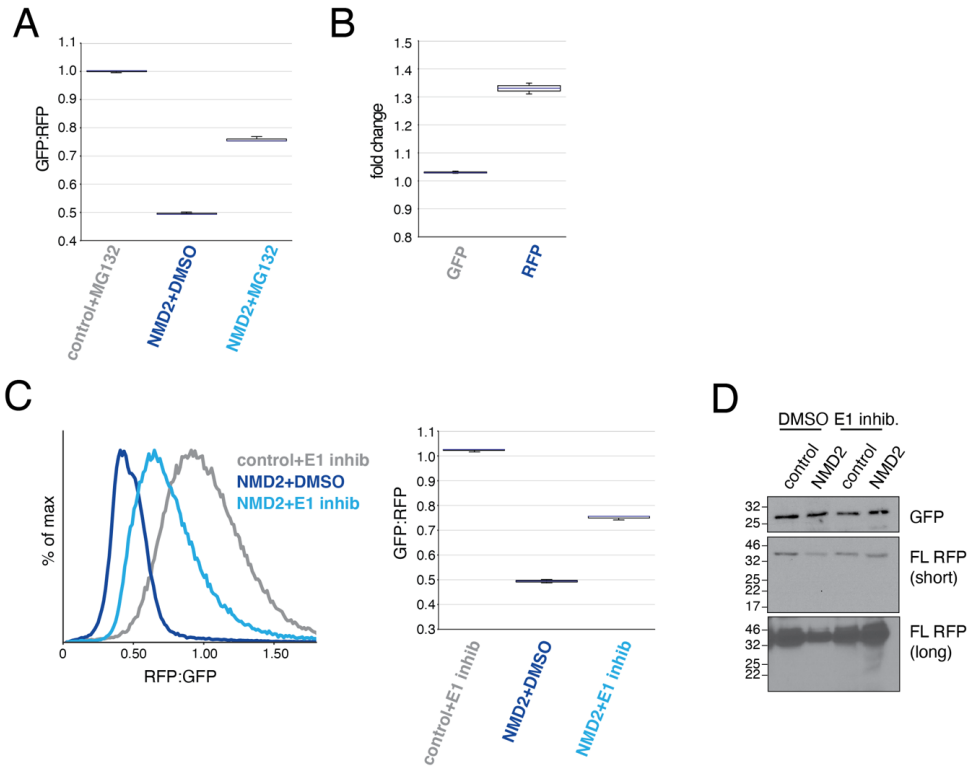


Figure S2.2: The ubiquitin-proteasome system mediates NMD-linked nascent protein degradation. **(A)** A box plot showing three biological replicates for the flow cytometry data presented in Fig. 2A. **(B)** The effect of MG132 treatment on GFP and RFP levels was quantified as fold change for GFP and RFP as shown in Fig. 2B, with three biological replicates plotted. **(C)** HEK293T cells were transiently transfected with either control or NMD2 reporters. After 16 hours, the cells were treated for 8 hours with either 10 μM MLN7243 or a matched DMSO control. Cells were then harvested and analyzed by flow cytometry. The results from three biological replicates are shown on the right. **(D)** Cells were treated as in (C), but were lysed in 1% SDS after MLN7243 treatment. The lysates were boiled and subjected to SDS-PAGE and Western blotting. Samples were normalized to GFP to control for RNA degradation. No RFP degradation products were observed, as seen in the long RFP exposure.

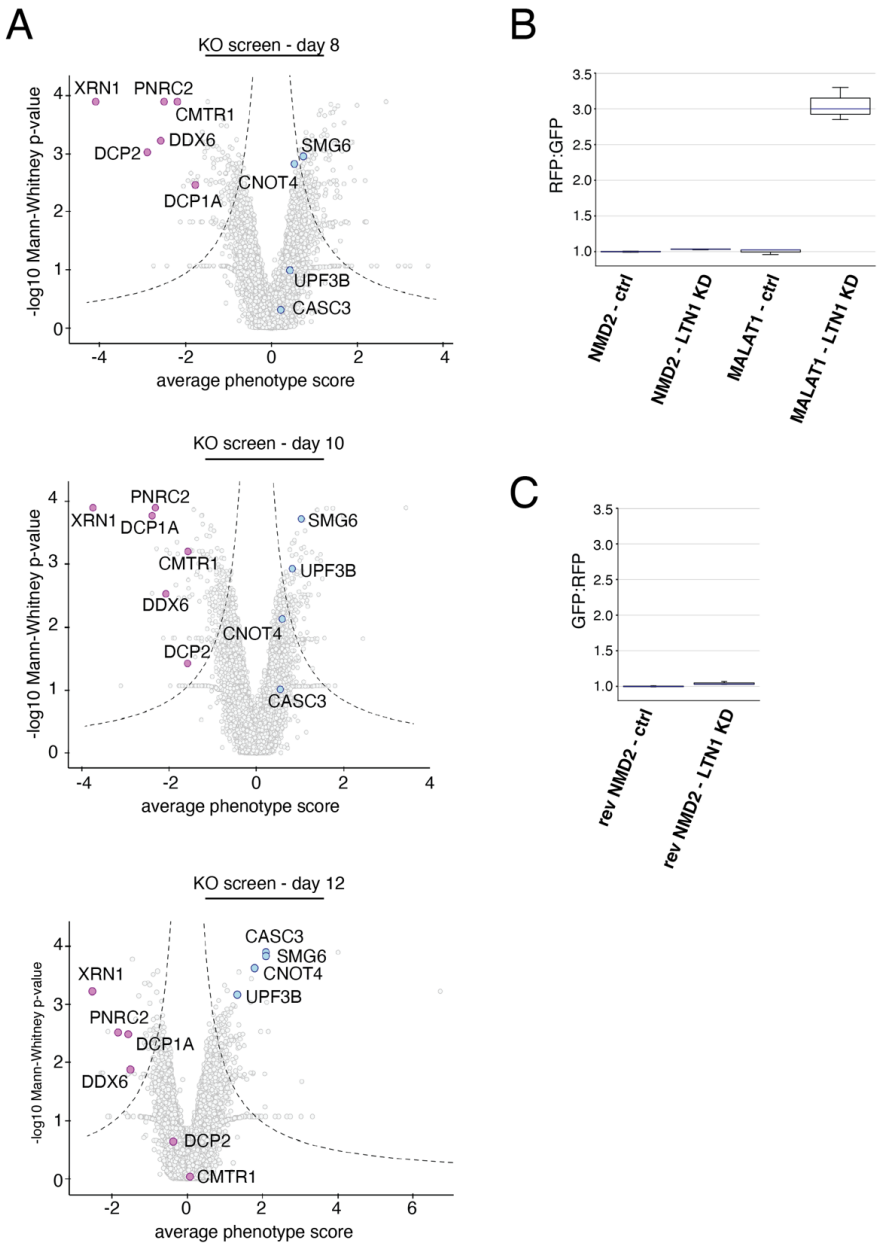


Figure S2.3: CRISPR knock-out screen progression across different time points and quantification of NMD dependence on RQC factors. (A) Shown are volcano plots from days 8, 10, and 12 of the knockout CRISPR screen with factors of interest highlighted. Some genes may drop out over the course of the screen, and so show highest phenotype scores at day 8 (e.g. DCP2). Conversely, other genes require a longer time period to be depleted, and show increased effects on the reporter at later time points (e.g. CASC3). **(B)** A box plot showing the results from three biological replicates of the effect of knocking down the RQC E3 ligase LTN1 on the NMD reporter (NMD2) or the non-stop reporter (MALAT1) as shown in Fig. 4C. **(C)** As in (A) for the effects of LTN1 knock down on the reverse NMD2 reporter as show in Fig. 2.4D.

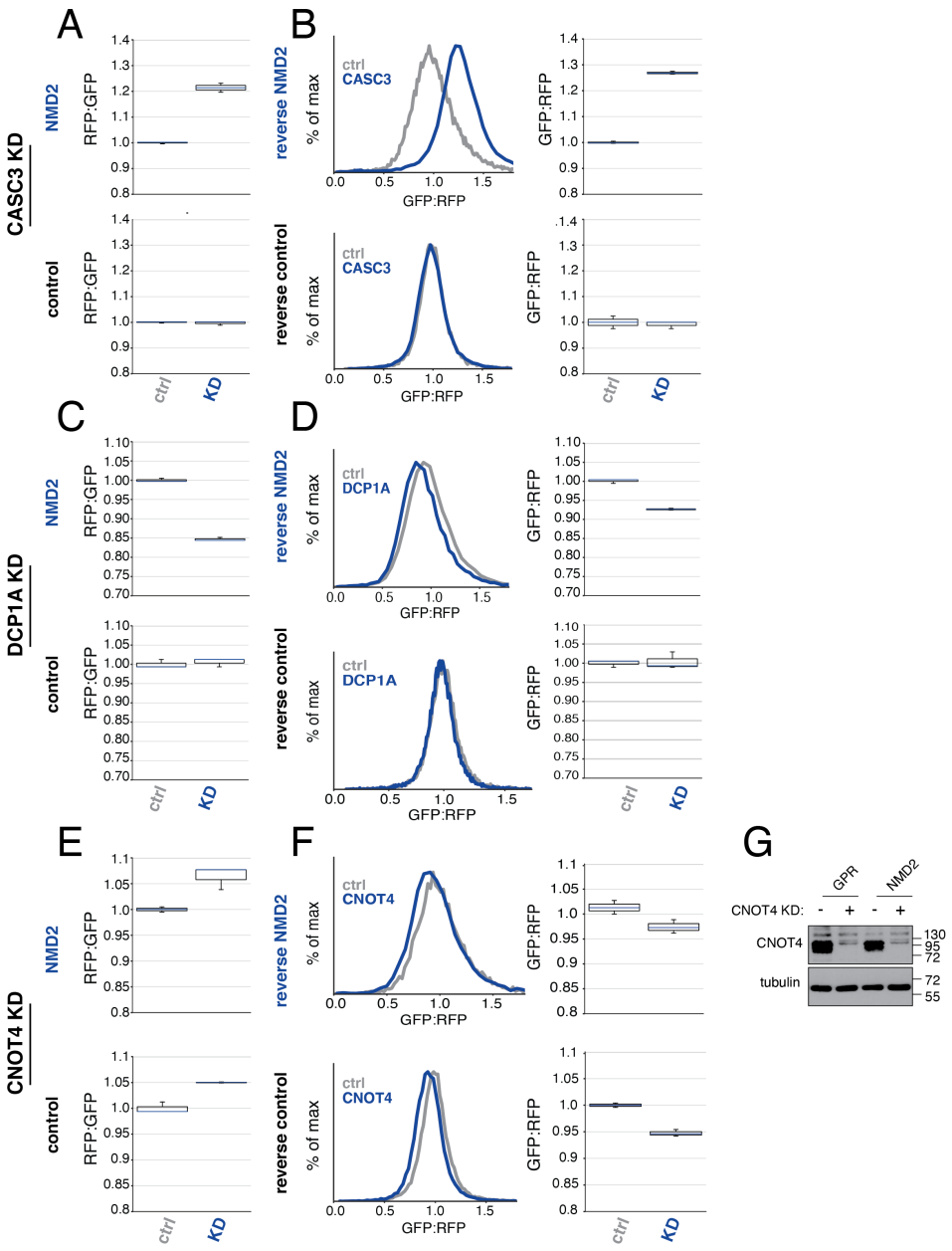


Figure S2.4: Validation of the screen hits. **(A)** Box plots showing the effect of knockdown of CASC3 on the control and NMD2 reporters in K562 Zim3 CRISPRi cells (histograms shown in Fig. 5) across three biological replicates. **(B)** CASC3 was CNOT4 were depleted by sgRNA for 8 days in K562 Zim3 CRISPRi cells expressing either the reverse NMD2 reporter or the reverse control reporter. Displayed are the RFP:GFP ratios for reporters as determined by flow cytometry after 24 hours of induction. Box plots showing three biological replicates are shown next to each histogram. **(C, D)** As above for DCP1A. **(E, F)** As above for CNOT4. (G) CNOT4 depletion was confirmed by Western blot.

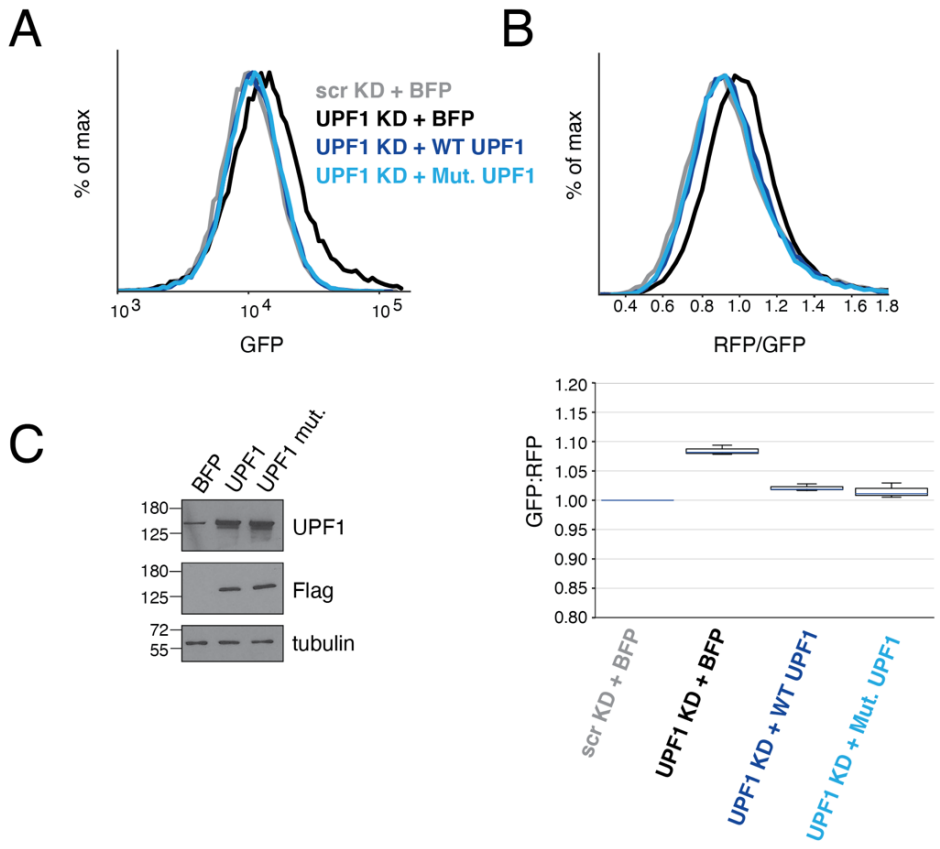


Figure S2.5: The role of UPF1's E3 ligase activity in NMD-linked protein degradation. **(A)** K562 CRISPRi cells stably expressing the inducible NMD2 reporter were constructed to stably express one copy of either BFP, a FLAG-conjugated wild-type UPF1, or a FLAG-conjugated mutant UPF1 (S134A, N148A, T149A) with disruptions that abolish association with E2 conjugating enzymes. WT or mutant UPF1 was separated from BFP by a viral P2A sequence, allowing us to use BFP as a proxy for UPF1 infection. These cells were then infected with dual sgRNA guides targeting UPF1 or a non-targeting control. Note that rescue constructs were resistant to the sgRNA. After 8 days of knockdown, the NMD2 reporter was induced with doxycycline for 24 hours, after which cells were harvested and analyzed by flow cytometry. GFP levels are shown in (A) and the RFP:GFP ratios are shown in **(B)**, with the quantification of three biological replicates below. **(C)** UPF1 wild-type and mutant over-expression levels were confirmed by Western blotting in the K562 line stably expressing NMD2.

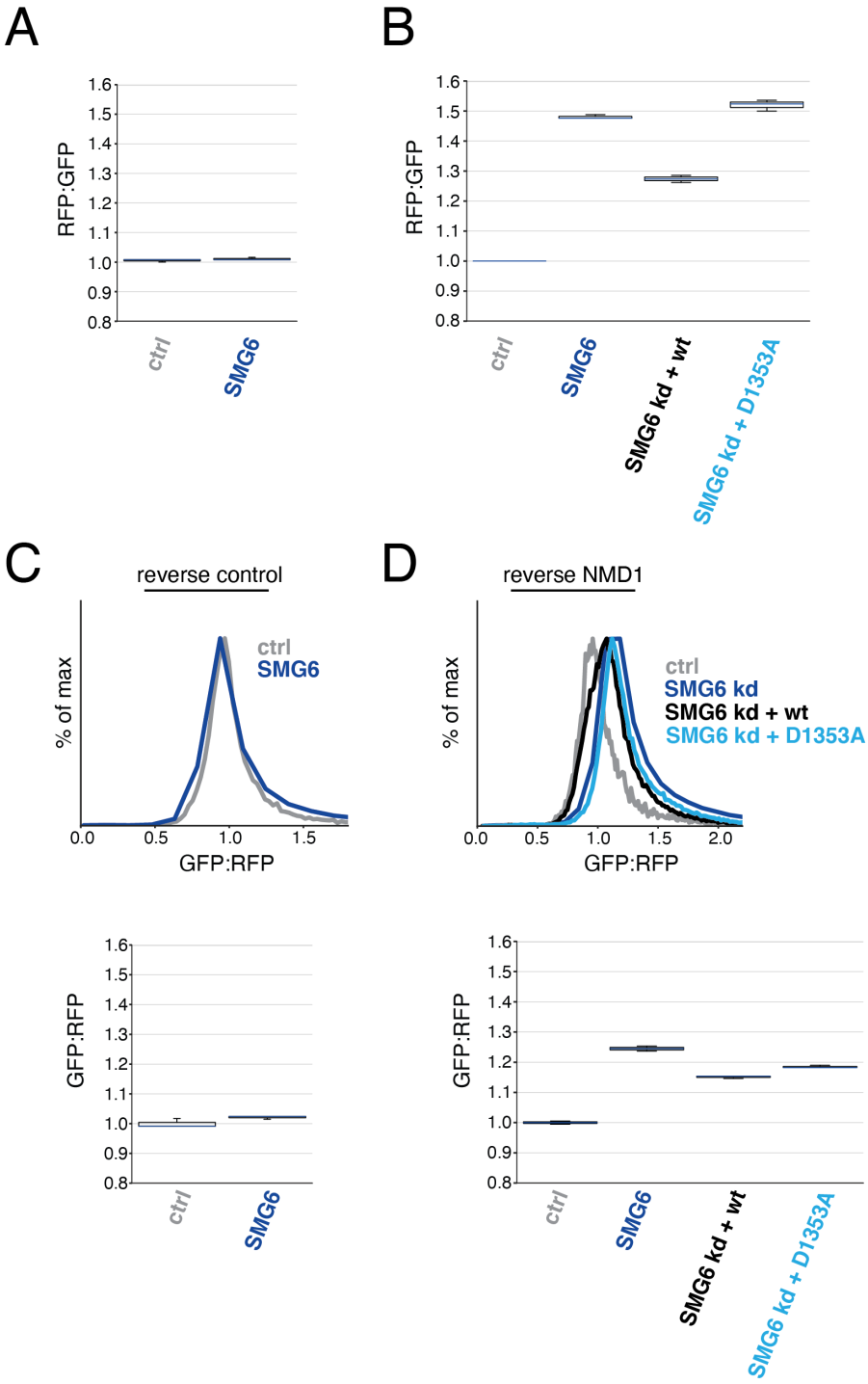


Figure S2.6: The effect of SMG6 is independent of fluorescent protein order. **(A)** Effect of SMG6 depletion on the control reporter (Fig. 2.6A) across three biological replicates. **(B)** As in (A) for NMD2 with rescue by the wild type and mutant SMG6. **(C)** HEK293T cells were treated with siRNA against SMG6 for 48 hours, then were transiently transfected with the reversed control reporter. Cells were analyzed in triplicate by flow cytometry after 24 hours. (results plotted below). **(D)** HEK293T cells were treated with an siRNA against SMG6, transfected with an siRNA-resistant version of either wild-type SMG6 or a PIN domain mutant version and the reversed NMD reporter, and analyzed by flow cytometry after 24 hours. A box plot showing the results from three biological replicates is shown.

Part II

The Commons Cell Atlas

ABSTRACT

Current cell atlas projects aim to curate representative datasets, cell-types, and marker genes for tissues across an organism. Despite their ubiquity, atlas projects rely on duplicated and manual effort to curate marker genes and annotate cell-types. Importantly, the lack of data-compatible tools and a fixed representation of the atlas make their reanalysis near-impossible. To overcome these challenges, we present a collection of data, algorithms, and tools to automate cataloging and analyzing cell-types across all tissues in an organism. We leveraged this work to build a Human Commons Cell Atlas comprising 2.9 million cells across 27 tissues that can be easily updated and that is structured to facilitate custom analyses. To showcase the flexibility of the atlas, we demonstrate that it can be used for isoform analyses. In particular, we study cell-type specificity of isoforms of OAS1, which has recently been shown to offer SARS-CoV-2 protection in certain individuals that display higher expression of the p46 isoform. Using our Commons Cell Atlas, we localize the OAS1 p44b isoform to the testis, and find that it is specific to germ line cells. By virtue of enabling customized analyses via a modular and dynamic atlas structure, the Commons Cell Atlas should be useful for exploratory analyses that are intractable within the rigid framework of current gene-centric static atlases.

Chapter 3

INTRODUCTION

3.1 Classification of cells

"Among all the marvels I have discovered in nature, these are the most marvelous of all." –Antonie van Leeuwenhoek

Back in the 17th century, two scientists peered through the lenses of their microscopes and uncovered the mystery of the basic unit of living systems. Robert Hooke and Antonie van Leeuwenhoek described the *cell* as the structural unit of life for the first time, laying the foundation for the field of cell biology (Hooke, 1665; Leeuwenhoek, 1977). Ever since then, scientists have sought to classify and characterize cells in order to understand the living organisms of which they are a part.

For centuries, researchers have relied on the visual appearance of cells as the basis for their classification into different cell-types (Arendt et al., 2016). Physical traits, like cell shape (e.g., rods and cones in the retina) or internal structures (e.g., mast cells in the immune system) have been used to distinguish between cell-types. With the advent of new biological techniques, the criteria for cell classification expanded to include additional features, such as surface protein expression (Delmonte and Fleisher, 2019), and the secretion of certain molecules (Romer and Sussel, 2015). However, the development of novel technologies based on Next-Generation Sequencing (NGS) brought about a paradigm shift: to no longer rely on just one or a few defining features, but rather to classify cells according to their complete gene expression profile (Trapnell, 2015) (Fig 3.1).

3.2 Single Cell RNA-seq

Single-cell RNA sequencing (scRNA-seq) enables the dissection of gene expression at the single-cell level (Tang et al., 2009; Chen, Ning, and Shi, 2019). There are several different scRNA-seq technologies available, all of them following the same basic steps (Jovic et al., 2022). First, individual cells are isolated, which can be achieved by a number of methods, including fluorescence-activated cell sorting (FACS) (Soumillon et al., 2014), and microfluidic systems (Macosko et al., 2015; Klein et al., 2015). Second, the transcripts of each cell are converted to cDNA and tagged with a Unique Molecular Identifier (UMI). Last, the cDNA is amplified, and

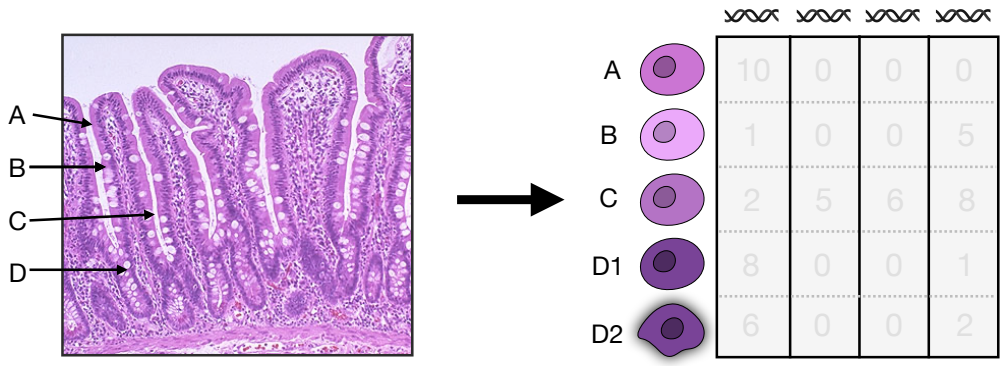


Figure 3.1: A new level of resolution for cell-type classification. scRNA-seq allows to quantify the expression of all genes for each individual cell of a given sample, enabling the classification of cell-types based on the full transcriptome rather than a limited set of features. In addition, new cell-types that were indistinguishable through other methods can be defined and characterized.

the resulting library is sequenced by NGS to identify and quantify the transcripts that were present in each original cell. During this process, transcripts coming from each cell are uniquely barcoded, which allows subsequent computational demultiplexing (Jovic et al., 2022).

The outcome of DNA sequencing is saved in a FASTQ file, where each entry (referred to as a *read*) represents the original sequence of a molecule in the library (Robinson, Piro, and Jäger, 2017). The reads are then aligned to a reference to count the number of occurrences of each gene or transcript in the FASTQ reads. This reference can contain the entire sequence of the genome, as used by alignment methods like STAR (Dobin et al., 2013), or just the transcriptome, as used by methods like kallisto (Bray et al., 2016). Next, transcripts originating from the same cell are grouped together using their corresponding barcode sequences; and reads originating from the same molecule are collapsed using their UMIs. The result of this process is a gene count matrix, in which each entry displays the number of molecules that mapped to the respective gene in the corresponding cell (Chen, Ning, and Shi, 2019).

Filtering and normalization of the gene count matrix are essential steps that must be taken before any statistical analysis can be performed on the data (Hu et al., 2022). While a unique barcode is assigned to the transcripts of each cell, not all barcodes present in the data correspond to real cells. The use of a *knee plot* to identify barcodes with low UMI counts is a popular technique for filtering out low-quality cells. Once filtered, the matrix needs to be normalized to achieve two goals: i) depth-normalize the data and ii) stabilize the variance. Depth-normalization refers to adjusting the gene expression values in each cell to account for differences

in sequencing depth, while variance stabilization aims to remove the gene mean-variance relationship, which is considered technical. Once the matrix is normalized, cells can be classified into subpopulations according to their gene expression profiles, and statistical analysis can be performed to gain a better understanding of the biological processes at play.

3.3 Single cell atlases

The advent of scRNA-seq has paved the way for addressing once-intractable problems in biology. scRNA-seq has been used to discover and define new cell-types, examine the temporal progression of developmental processes (Behjati et al., 2014), explore gene regulatory networks (Akers and Murali, 2021) and study random allelic gene expression (Deng et al., 2014). Many of these goals have been achieved through the generation of so-called *single cell atlases*, which constitute a prime example of how scRNA-seq has enabled us to gain a deeper understanding of cellular diversity and function.

A single cell atlas can be defined as a comprehensive catalog of gene expression profiles of individual cells within a given tissue or organism. The first single cell atlases were created in the early 2010s (Kolodziejczyk et al., 2015), and since then, several large-scale projects have been established, including the Human Cell Atlas (HCA) (Lindeboom, Regev, and Teichmann, 2021), the Tabula Muris project (Consortium et al., 2018), and the BRAIN Initiative Cell Census Network (Ecker et al., 2017). These atlases have profiled millions of individual cells from various tissues and organs, providing unprecedented insights into the diversity and function of cells within complex biological systems.

3.4 Contribution of this thesis

Single cell atlases hold enormous potential, but in their current form have two significant limitations, namely that raw data is difficult to access and atlases are "static", i.e., it is challenging to reprocess results in response to new data or annotations. In the second part of my thesis, I outline my efforts to address these challenges. In Chapter 4, I describe the development of a novel tool to extract metadata from genomic databases, an essential step needed to efficiently build atlases that wasn't available. In Chapter 5, we performed the first ever comprehensive benchmark of normalization methods across an entire cell atlas, providing quantitative results on the performance and suitability of each method and proposing a novel statistical method that outperforms other existing ones. Moreover, we have developed a suite

of tools that solve fundamental challenges in scRNA-seq processing, such as a fully automated method to filter low quality cells and a novel cell-type assignment algorithm that outperforms the current gold standard (Chapter 6). Using these tools, we have created an open-access framework to build single cell atlases for any organism. By virtue of its modular nature, the framework can be customized to build atlases that allow researchers to tackle novel research questions that are inaccessible to current atlases, such as the study of isoform expression (Chapter 6). Using this framework, we built the first completely reproducible Human Cell Atlas, encompassing 27 organs from 525 datasets (Chapter 7). All the elements of the atlas can be both downloaded and updated, from the raw data to the cell-type marker genes. Finally, we used this atlas to study the cell-type specificity of OAS1 isoforms, a previously unexplored question with high clinical relevance (Chapter 7).

References

Akers, Kyle and T.M. Murali (2021). "Gene regulatory network inference in single-cell biology." In: *Current Opinion in Systems Biology* 26, pp. 87–97.

Arendt, Detlev et al. (2016). "The origin and evolution of cell types." In: *Nature Reviews Genetics* 17.12, pp. 744–757.

Behjati, Sam et al. (2014). "Genome sequencing of normal cells reveals developmental lineages and mutational processes." In: *Nature* 513.7518, pp. 422–425.

Bray, Nicolas L. et al. (2016). "Near-optimal probabilistic RNA-seq quantification." In: *Nature Biotechnology* 34.5, pp. 525–527.

Chen, Geng, Baitang Ning, and Tieliu Shi (2019). "Single-cell RNA-seq technologies and related computational data analysis." In: *Frontiers in genetics*, p. 317.

Consortium, Tabula Muris et al. (2018). "Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris". In: *Nature* 562.7727, pp. 367–372.

Delmonte, Ottavia M. and Thomas A. Fleisher (2019). "Flow cytometry: Surface markers and beyond." In: *Journal of Allergy and Clinical Immunology* 143.2, pp. 528–537.

Deng, Qiaolin et al. (2014). "Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells." In: *Science* 343.6167, pp. 193–196.

Dobin, Alexander et al. (2013). "STAR: ultrafast universal RNA-seq aligner." In: *Bioinformatics* 29.1, pp. 15–21.

Ecker, Joseph R et al. (2017). "The BRAIN initiative cell census consortium: lessons learned toward generating a comprehensive brain cell atlas". In: *Neuron* 96.3, pp. 542–557.

Hooke, Robert (1665). *Micrographia: or some physiological descriptions of minute bodies made by magnifying glasses.* Printed by Jo. Martyn and Ja. Allestry, printers to the Royal Society.

Hu, Jialu et al. (2022). "Pre-processing, dimension reduction, and clustering for single-cell RNA-seq data." In: *Handbook of Statistical Bioinformatics*. Springer, pp. 37–51.

Jovic, Dragomirka et al. (2022). "Single-cell RNA sequencing technologies and applications: A brief overview." In: *Clinical and Translational Medicine* 12.3, e694.

Klein, Allon M. et al. (2015). "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells." In: *Cell* 161.5, pp. 1187–1201.

Kolodziejczyk, Aleksandra A. et al. (2015). "The technology and biology of single-cell RNA sequencing." In: *Molecular Cell* 58.4, pp. 610–620.

Leeuwenhoek, Antony van (1977). *The Selected Works of Antony van Leeuwenhoek Containing His Microscopical Discoveries in Many of the Works of Nature.* Arno Press.

Lindeboom, Rik G.H., Aviv Regev, and Sarah A Teichmann (2021). "Towards a human cell atlas: taking notes from the past". In: *Trends in Genetics* 37.7, pp. 625–630.

Macosko, Evan Z. et al. (2015). "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets." In: *Cell* 161.5, pp. 1202–1214.

Robinson, Peter N., Rosario M. Piro, and Marten Jäger (2017). "FASTQ Format." In: *Computational Exome and Genome Analysis*. Chapman and Hall/CRC, pp. 57–65.

Romer, Anthony I. and Lori Sussel (2015). "Pancreatic islet cell development and regeneration." In: *Current Opinion in Endocrinology, Diabetes, and Obesity* 22.4, p. 255.

Soumillon, Magali et al. (2014). "Characterization of directed differentiation by high-throughput single-cell RNA-Seq." In: *BioRxiv*, p. 003236.

Tang, Fuchou et al. (2009). "mRNA-Seq whole-transcriptome analysis of a single cell." In: *Nature methods* 6.5, pp. 377–382.

Trapnell, Cole (2015). "Defining cell types and states with single-cell genomics." In: *Genome research* 25.10, pp. 1491–1498.

Chapter 4

METADATA RETRIEVAL FROM GENOMICS DATABASES WITH FFQ

Gálvez-Merchán, Ángel et al. (2023). "Metadata retrieval from sequence databases with ffq." In: *Bioinformatics* 39.1, btac667. DOI: 10.1093/bioinformatics/btac667. URL: https://academic.oup.com/bioinformatics/article/39/1/btac667/6971839.

4.1 Introduction

The extraordinarily large volume of user-generated sequencing data available in public databases is increasingly being utilized in research projects alongside novel experiments (Simon et al., 2018; Razmara et al., 2019; Lung et al., 2020; Rajesh et al., 2021; Hippen and Greene, 2021; Wartmann et al., 2021; Kasmanas et al., 2021; Huang et al., 2021; Klie et al., 2021; Booeshaghi et al., 2022). Collation of metadata is crucial for effective use of publicly available data. Accurate metadata can provide information about the samples assayed and can facilitate the acquisition of raw data. For example, sra-tools enables users to query and download data from the National Center for Biotechnology Information Sequence Read Archive (NCBI SRA), which currently hosts 13.67 PB of data. An alternative to sra-tools is the pysradb tool (Choudhary, 2019). pysradb was developed to access metadata from the Sequence Read Archive (SRA), using metadata obtained from the regularly updated SRAdb SQLite database (Zhu et al., 2013). MetaSRA adds additional standardized metadata on top of the SRAdb SQLite database (Bernstein, Doan, and Dewey, 2017) and also provides an API for accessing them. While these and other tools (Mahi et al., 2019; Li, Li, and Yu, 2018; Eaton, 2020; Bernstein, Gladstein, et al., 2020) have proven to be very useful, they provide access to a limited scope of databases. We developed *FetchFastQ* (*ffq*) to facilitate metadata retrieval from a diverse set of databases, including

1. National Center for Biotechnology Information Sequence Read Archive (SRA) and Gene Expression Omnibus (GEO),

2. European Molecular Biology Lab-European Bioinformatics Institute European Nucleotide Archive (EMBL-EBI ENA),

3. DNA Data Bank of Japan Gene Expression Archive (DDBJ GEA), and

4. Encyclopedia of DNA Elements (ENCODE) database (Davis et al., 2018).

In order to facilitate a modular architecture for *ffq*, we first studied the structure of these databases in detail to identify commonalities and relationships between them (Fig. 4.1). The SRA, ENA, and DDBJ databases all follow a similar hierarchical structure where studies are grouped into samples, experiments, and runs, a shared architecture that is useful and likely the result of the longstanding International Nucleotide Sequence Database Collaboration (INSDC) between the ENA, NCBI, and DDBJ. We note that the Genome Sequence Archive (GSA) (Chen et al., 2021) is not a member of the INSDC. However it also uses a similar hierarchical structure for its database, and regularly ingests data from the SRA, but does not expose its publicly available data for programmatic access.

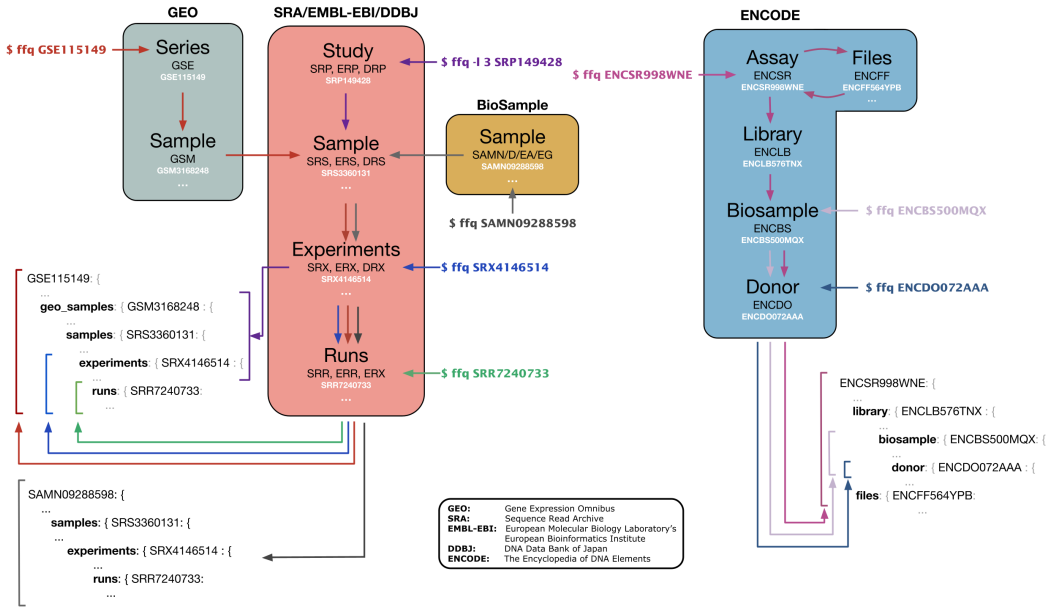


Figure 4.1: Metadata retrieval. *ffq* fetches and returns metadata as a JSON object by traversing the database hierarchy. Subsets of the database hierarchy can be returned by specifying *-l [level]*.

The consistent database schemas used by members of the INSDC greatly simplifies metadata retrieval for *ffq*. For example, GEO accession codes are grouped hierarchically through Series and Samples and have external relations to SRA accession codes for raw sequencing data submitted to the SRA. This enables *ffq* to fetch metadata and processed data from GEO that submitters have associated with raw sequencing data stored in the SRA.

4.2 Description

Based on the database architectures, we created *ffq* to fetch metadata using database accessions or paper DOIs as input. Importantly, *ffq* only fetches metadata and links to data files and does not offer data downloading. This deliberate design decision was motivated by the UNIX philosophy "Make each program do one thing well" (McIlroy, Pinson, and Tague, 1978).

The *ffq* options are summarized below:

ffq [accession(s)] where [accession] can be any of the following: SR(R/X/S/P), ER(R/X/S/P), DR(R/X/S/P), GS(E/M), ENC(SR/BS/DO), CXR, SAM(N/D/EA/EG), DOI.

ffq [-l level] [accession(s)] where [level] defines the hierarchy in the database to which data is subset data.

ffq [–ftp] [–aws] [–gcp] [–ncbi] [accession(s)] where the flags correspond to the types of data-storage links for the raw data.

ffq [-o out] [–split] [accession(s)] where [out] corresponds to a path on disk to save the JSON file and [–split] splits the metadata from multiple accessions into their own file.

The *ffq* codebase consists of 58 functions and 2,198 lines of code across six files and relies on only four software dependencies. Users supply an accession or DOI and the tool returns metadata for the sequencing data associated with that accession or DOI. Accession-based *ffq* metadata retrieval uses the NCBI Entrez programming utilities, ENA API, GEO FTP, and ENCODE API to programmatically access metadata with HTTP requests. DOI-based metadata retrieval first converts the DOI to the manuscript title via the CrossRef API (Hendricks et al., 2020) and then retrieves all study accessions associated with the manuscript title with the ENA search API. The reliance on these external dependencies can make it challenging to track API updates that may break *ffq* functionality. To provide resilience to such changes, we have implemented extensive quality control via an automated testing framework that validates behavior against all external APIs and five Python versions (3.6, 3.7, 3.8, 3.9, and 3.10) that cover 78% of the code. This makes it easy to detect and address API updates within *ffq*. Once fetched, metadata is returned as a Javascript Object Notation (JSON) object. Run times for metadata retrieval vary depending on database up-time, server connection speed, and database rate-limiting, but generally we find that *ffq* can download metadata at a rate of 10s per sample. This rate

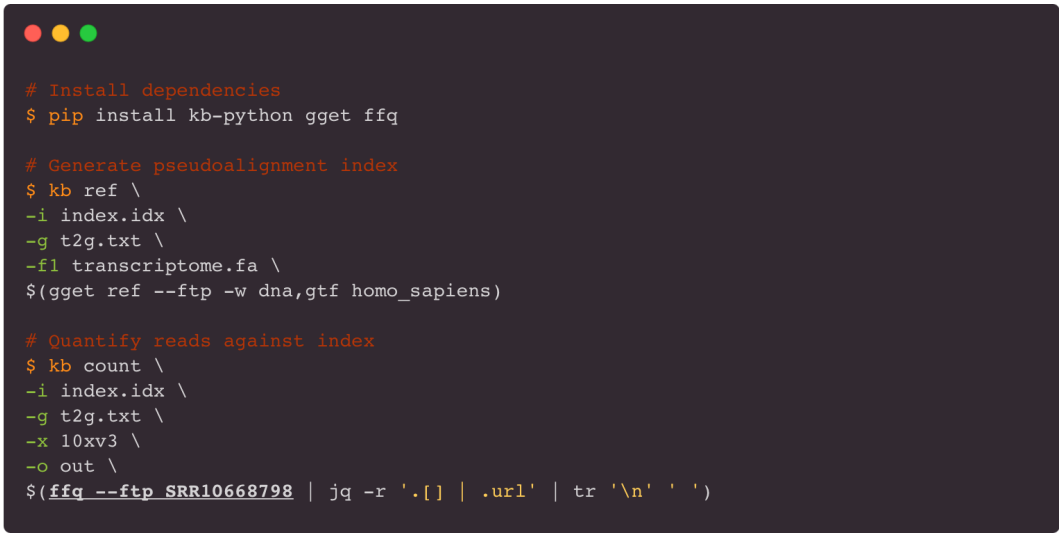
includes short and deliberate delays we have added between HTTP requests to prevent a perceived Denial-of-Service. External factors may impact *ffq*'s ability to fetch metadata that are independent of the tool. Internet connection, improperly formatted accessions, missing or incomplete metadata are some of the failure modes that users may face. To aid users in debugging missing or incomplete metadata, custom exceptions have been implemented and possible failure modes and caveats have been listed in the documentation.

4.3 Usage and documentation

The *ffq* tool is written in Python and can be installed with pip and conda. Users supply an accession or DOI and the tool returns metadata for the associated sequencing data. The JSON-return objects make *ffq* interoperable with other tools such as jq for easy command-line parsing. Additionally, *ffq*'s modularity and simplicity make it extensible to other genomic databases. By leveraging existing APIs, *ffq* offers a lightweight solution for querying data that is guaranteed to be more up-to-date than tools that rely on regular database builds. These features enable researchers to use *ffq* to refine research questions. For example, *ffq* can be used to fetch publicly available scRNAseq data, which can be preprocessed with existing tools (Melsted et al., 2019) and compared against newly generated data (Fig. 4.2). Alternatively, *ffq* can be used for sequencing quality control; sequencing reads can be fetched with *ffq* and piped into common command-line tools to count the number of reads or assess the per-base quality scores. These and other use cases are explained in the *ffq* documentation. The modularity of *ffq* makes possible streamwise processing of publicly available FASTQ files for any number of applications.

4.4 Discussion

While *ffq* facilitates downloading of data from numerous genomic databases, the results retrieved are only useful to the extent that the metadata uploaded is meaningful and complete. Meaningful and complete user-generated data underlies the curation of genomic references essential for comparative genomic data analysis (Luebbert and Pachter, 2022). Unfortunately, there is little to no standardization of user-uploaded sequencing metadata (Wang, Lachmann, and Ma'ayan, 2019; Rajesh et al., 2021), and metadata descriptions can become exceedingly complex for current multiplexed experiments, where different assays with distinct data types are combined. Improvement of metadata uploading in machine-readable standard formats is essential if publicly available genomic data are to be usable by scientists in the



```
# Install dependencies
$ pip install kb-python gget ffq

# Generate pseudoalignment index
$ kb ref \
-i index.idx \
-g t2g.txt \
-f1 transcriptome.fa \
$(gget ref --ftp -w dna,gtf homo_sapiens)

# Quantify reads against index
$ kb count \
-i index.idx \
-g t2g.txt \
-x 10xv3 \
-o out \
$(ffq --ftp SRR10668798 | jq -r '.[] | .url' | tr '\n' ' ')
```

Figure 4.2: Example use case. Publicly available scRNAseq data is fetched with *ffq* and quantified with kb-python to generate a gene count matrix. The *ffq* command is underlined.

future. Users who wish to refine research questions with complete and accurate publicly available data will benefit from *ffq*. By providing direct links to sequencing data and metadata, *ffq* allows any number of downstream procedures that operate on sequencing reads. Importantly, the modularity of *ffq* enables streamwise processing of data and metadata that obviates the need for large amounts of storage and lessens the cost of computing.

References

Bernstein, Matthew N, AnHai Doan, and Colin N Dewey (2017). "MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive". In: *Bioinformatics*.

Bernstein, Matthew N, Ariella Gladstein, et al. (2020). "Jupyter notebook-based tools for building structured datasets from the Sequence Read Archive". In: *F1000Research* 9.

Booeshaghi, A. Sina et al. (2022). "Depth normalization for single-cell genomics count data." In: *bioRxiv*, pp. 2022–05. DOI: 10.1101/2022.05.06.490859. URL: https://www.biorxiv.org/content/10.1101/2022.05.06.490859v1.abstract.

Chen, Tingting et al. (2021). "The genome sequence archive family: toward explosive data growth and diverse data types". In: *Genomics, Proteomics & Bioinformatics* 19.4, pp. 578–583.

Choudhary, Saket (2019). "pysradb: A Python package to query next-generation sequencing metadata and data from NCBI Sequence Read Archive". In: *F1000Research* 8.

Davis, Carrie A et al. (2018). ".(2018) The encyclopedia of DNA elements (ENCODE): data portal update". In: *Nucleic Acids Res* 46, pp. D794–D801.

Eaton, Katherine (2020). "NCBImeta: efficient and comprehensive metadata retrieval from NCBI databases". In: *Journal of Open Source Software* 5.46, p. 1990.

Hendricks, Ginny et al. (2020). "Crossref: The sustainable source of community-owned scholarly metadata". In: *Quantitative Science Studies* 1.1, pp. 414–427.

Hippen, Ariel A and Casey S Greene (2021). "Expanding and remixing the metadata landscape". In: *Trends in cancer* 7.4, pp. 276–278.

Huang, Yu-Ning et al. (2021). "The systematic assessment of completeness of public metadata accompanying omics studies". In: *bioRxiv*, pp. 2021–11.

Kasmanas, Jonas Coelho et al. (2021). "HumanMetagenomeDB: a public repository of curated and standardized metadata for human metagenomes". In: *Nucleic acids research* 49.D1, pp. D743–D750.

Klie, Adam et al. (2021). "Increasing metadata coverage of SRA BioSample entries using deep learning–based named entity recognition". In: *Database* 2021.

Li, Zhao, Jin Li, and Peng Yu (2018). "GEOMetaCuration: a web-based application for accurate manual curation of Gene Expression Omnibus metadata". In: *Database* 2018.

Luebbert, Laura and Lior Pachter (2022). "Efficient querying of genomic databases for single-cell RNA-seq with gget". In: *bioRxiv*.

Lung, Pei-Yau et al. (2020). "Maximizing the reusability of gene expression data by predicting missing metadata". In: *PLoS computational biology* 16.11, e1007450.

Mahi, Naim Al et al. (2019). "GREIN: An interactive web platform for re-analyzing GEO RNA-seq data". In: *Scientific reports* 9.1, pp. 1–9.

McIlroy, M. Douglas, Elliot N. Pinson, and Berkley A. Tague (1978). "UNIX time-sharing system". In: *The Bell system technical journal* 57.6, pp. 1899–1904.

Melsted, Páll et al. (2019). "Modular and efficient pre-processing of single-cell RNA-seq". In: *BioRxiv*, p. 673285.

Rajesh, Anushka et al. (2021). "Improving the completeness of public metadata accompanying omics studies". In: *Genome Biology* 22.1, pp. 1–5.

Razmara, Ashkaun et al. (2019). "recount-brain: a curated repository of human brain RNA-seq datasets metadata". In: *bioRxiv*, p. 618025.

Simon, Lukas M. et al. (2018). "MetaMap, an interactive webtool for the exploration of metatranscriptomic reads in human disease-related RNA-seq data". In: *bioRxiv*, p. 425439.

Wang, Zichen, Alexander Lachmann, and Avi Ma'ayan (2019). "Mining data and metadata from the gene expression omnibus". In: *Biophysical reviews* 11.1, pp. 103–110.

Wartmann, Hannes et al. (2021). “Bias-invariant RNA-sequencing metadata annotation”. In: *GigaScience* 10.9, giab064.

Zhu, Yuelin et al. (2013). “SRAdb: query and use public next-generation sequencing data from within R”. In: *BMC bioinformatics* 14.1, pp. 1–4.

Chapter 5

DEPTH NORMALIZATION FOR SINGLE-CELL GENOMICS COUNT DATA

Booeshaghi, A. Sina et al. (2022). "Depth normalization for single-cell genomics count data." In: *bioRxiv*, pp. 2022–05. DOI: 10.1101/2022.05.06.490859. URL: https://www.biorxiv.org/content/10.1101/2022.05.06.490859v1.abstract.

5.1 Introduction

A central theme in single-cell RNA-seq "count normalization" is the importance of achieving depth normalization alongside variance stabilization (Vallejos et al., 2017; Evans, Hardin, and Stoebel, 2018; Robinson and Oshlack, 2010). While variance stabilization has been studied for over 85 years (Bartlett, 1936), the question of how to achieve both variance stabilization and depth normalization is unsolved. An important condition that is often overlooked when evaluating normalization and variance-stabilization methods is that structure must be preserved in the data, which is why classic variance stabilizing transformations are monotonic by design (Doob, 1935). This is why the constant transformation, which sets all counts equal to each other and results in a fully variance-stabilized matrix with all cell depths equal, is not a good normalization.

While many methods have been proposed for single-cell RNA-seq normalization (Cole et al., 2019; Tian et al., 2019; You et al., 2021; Lytal, Ran, and An, 2020; Borella et al., 2022; Ahlmann-Eltze and Huber, 2021; Breda, Zavolan, and Nimwegen, 2021), the approach of equalizing depth for all cells, often to a "size factor" such as ten thousand (CP10k) or one million (CPM), followed by the application of a variance stabilizing transform like log plus one (log1p) is most popular. These methods are implemented in the widely used Seurat and Scanpy (Wolf, Angerer, and Theis, 2018) programs, but they do not explicitly model cell depth as a covariate. The recently published sctransform method (Hafemeister and Satija, 2019), which has quickly become the most widely used normalization method for single-cell RNA-seq, aims to address the challenge of variance stabilization and depth normalization by transforming data to Pearson residuals derived from a regularized

negative binomial regression. This regression-based method includesincorporates sequencing depth as a covariate in a model, rather than usingutilizing a size factor (Anders and Huber, 2010). However, despite the claims in (Hafemeister and Satija, 2019), benchmarking of sctransform in (Crowell et al., 2020) shows that the method fails to completely remove the effects of variable depth. The authors of the benchmarking study show that as a result, sctransform produces "unacceptably high false discovery rates [when used for differential expression]." Similarly, (Urban et al., 2009) find that sctransform performs poorly (see their Figure 3) and the veracity of the claim in (Hafemeister and Satija, 2019) that sctransform "can successfully remove the influence of technical characteristic from downstream analyses" is brought into question by the authors' own results. Figure 6 of their paper shows a UMAP plot of 33,148 PBMCs that the authors claim displays "a gradient that is correlated with sequencing depth" for log-normalized data, but not for data normalized with sctransform (Fig. 5.1a).

The figure belies this claim. Contrary to the authors' assertions, an examination of the plots shows that the Monocytes have a depth gradient with both methods. While this may be due to challenges in interpreting the UMAP embeddings (Chari, Banerjee, and Pachter, 2021), it could also be an indication that both methods fail to depth normalize the data. Furthermore, a differential expression benchmark of sctransform in (Hafemeister and Satija, 2019) shows that it produces almost no false positives, whereas a similar benchmark in a later paper (Choudhary and Satija, 2022) shows the opposite (Fig. 5.1b, c, d). Aside from questions about depth normalization, it is also unclear whether sctransform is effective at variance stabilization (Ahlmann-Eltze and Huber, 2021). These issues raise the question of how effective sctransform, or any other currently used method, is at achieving both depth normalization and variance stabilization.

Furthermore, an analysis of how normalization is used in practice, shows that normalization methods are applied in a task-specific manner, resulting in numerous normalizations sometimes being mixed together in a single analysis. For example, sctransform is not, in practice, a single method for computing Pearson residuals from raw counts, but rather a program that implements multiple normalization methods, where each method is used for a different task in the standard Seurat workflow. This highlights the importance of benchmarking the fundamental properties of each normalization technique in a way that is motivated by, and cognizant of, the downstream analysis tasks it may be applied to. In this paper we evaluate several

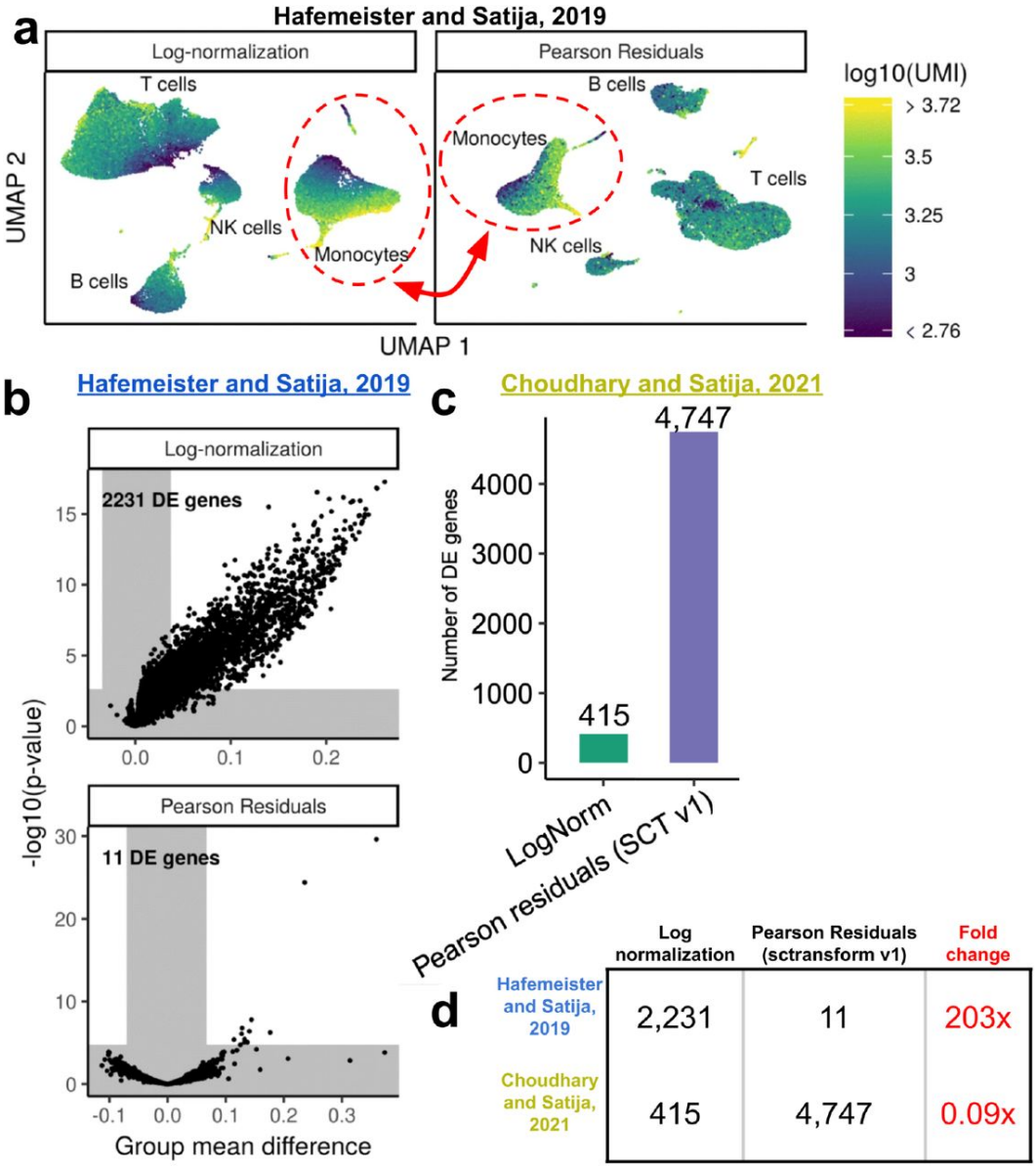


Figure 5.1: Questions about the efficacy of the sctransform depth normalization. **(a)** A reproduction of Figure 6 from (Hafemeister and Satija, 2019) shows a UMAP generated from the 10x Genomics "33k PBMCs from a Healthy Donor, v1 Chemistry" dataset, where the data has been normalized with the log1 pCP10k transform. The figure on the right shows a UMAP generated from the raw data normalized with sctransform. The authors state that "..correlations [between locations of embedded cells and sequencing depth] are strikingly reduced for Pearson residuals [in comparison to log-normalized data" but the difference for Monocytes (circled in red) does not look striking. **(b)** A differential expression control experiment from (Hafemeister and Satija, 2019) showing sctransform greatly reduces false positive genes in comparison to the log transform whereas **(c)** the opposite is shown in a similar control experiment in (Choudhary and Satija, 2022). The figures are all licensed under CC BY 4.0, and have been reproduced from the papers they were published in with only minor modifications (cropping, the addition of arrows and circles, and addition of number in the plot shown in (c)).

commonly used normalization methods based on how they perform with respect to three criteria that are crucial for common analysis methods: variance stabilization, normalization, and monotonicity of the transformations.

5.2 Results

Evaluation criteria

In considering how to evaluate normalization methods, we focused on downstream applications and their respective assumptions. Dimensionality reduction with PCA is an initial step in many analyses that relies on equal gene variances. If variance is not stabilized, genes with a high variance may have an outsized impact on the singular values solely due to having a high mean (Nguyen and Holmes, 2019). Similarly, without depth-normalization, the key step of identifying genes that are differentially expressed between cell types, may yield false-positive genes simply due to certain groups of cells being sampled more deeply than others (Robinson and Oshlack, 2010). An additional property of normalization techniques that is important for tasks such as marker gene selection is monotonicity of the transformations, especially for constructing heatmaps or similar visualizations.

To assess effectiveness of variance stabilization, we plotted the mean of each gene vs. its variance across cells, and measured the coefficient of variation of the gene variance (CV) after transformation as a scale-independent measure of the effectiveness of variance stabilization. Depth normalization was assessed by plotting, for each cell, the total raw cell counts vs. the total transformed cell counts. Since the total abundance of a gene per cell may not be measured with respect to an absolute scale, we computed the r2 correlation with raw cell depth as a proxy for the extent to which raw cell counts were reflected in the transformed data. Finally, for each cell, we computed the Spearman rank correlation between cells prior to, and after, transformation to measure deviations from a monotonic transformation. These three metrics allow for quantifying the trends observed in the three plots and offer a measure of the effectiveness of each normalization technique.

To verify that these metrics are reasonable for benchmarking normalization methods, we first examined cells from a NIH/3T3 mouse cell line dataset published in (Svensson, 2020) and studied in (Ahlmann-Eltze and Huber, 2021). We found that these metrics, which we computed for each normalization technique, were concordant with the analysis performed in (Ahlmann-Eltze and Huber, 2021), and provided useful summaries of the performance of different normalization techniques.

Benchmarks of 526 datasets

In recognition of the fact that the patterns we observed in 10xv3_NIH_3T3 were not necessarily representative of other datasets, we analyzed a further 525 datasets of which 437 passed quality control. We evaluated eight normalization techniques; in addition to sctransform, we selected seven other methods based on their use in popular single-cell RNA-seq analysis packages, as well as a novel method we decided to investigate after examining initial results (see Methods). The most widely used approach for depth normalization and variance stabilization is depth normalization of cell counts to ten thousand counts (CP10k), followed by variance stabilization of the gene counts with the log(x+1) transform (denoted by log1p, with the combined procedures denoted log1pCP10k). This is the default in the Seurat and Scanpy packages. Seurat and Scanpy also recommend an additional scaling step (scalelog1pCP10k) for some analyses. Scaling consists of two steps: centering gene expression values by subtracting the mean expression of each gene, and equalizing gene variances by dividing the counts for each gene by the standard deviation (computed across cells). We also benchmarked a method that has been adapted for single-cell RNA-seq from bulk RNA-seq, namely cell depth normalization to the mean cell depth, followed by log1p (log1pPF). This "proportional fitting" approach, our name for the method because the first step constitutes one step of iterative proportional fitting (Deming and Stephan, 1940), is similar to log1pCP10k (Love, Huber, and Anders, 2014), and is the method underlying the Monocle single-cell analysis package (Cao et al., 2019). We also tested the square root transformation that forms a part of the scprep package default transformation, as well as a log1pCPM, which is a popular option in Seurat, and is similar to log1pCP10k but with a scaling factor of one million rather than ten thousand. Finally, we included a benchmark of PF for completeness.

Our benchmarks revealed high variability in the extent of variance stabilization for any given method (Fig. 5.2a); to the extent that even though one method might be better at stabilizing variance than another, on one dataset it may produce worse results than the inferior method on another. For example, the sqrt transformation results in a CV of 1.61 for GSM3738540 whereas the log1p transformation yields a higher CV of 6.78 for GSM3396184. Some datasets are also particularly sensitive to the method used. The sqrt transformation gives a CV of 9.45 for GSM3178783 and 46.53 for GSM3396177, whereas the log1p transformation gives consistent results for these datasets with 5.77 GSM3178783 and 5.8 for GSM3396177; interestingly there is even a slight reversal in behavior. This highlights the importance of large-

scale benchmarking for evaluating normalization methods.

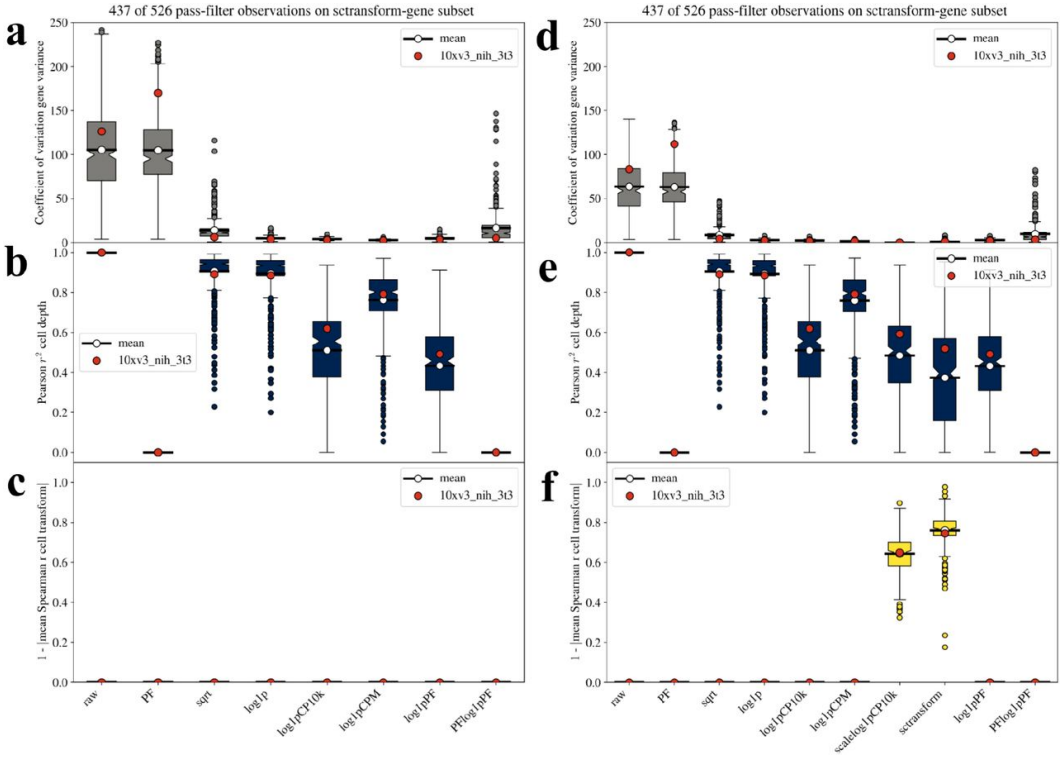


Figure 5.2: Benchmarking normalization techniques on 437 of 526 datasets passing filter. **(a)-(c)** demonstrate metrics computed on all genes, a task which is computationally intractable to compute on sctransform and scalelog1pCP10k due to their size. **(d)-(f)** demonstrate metrics computed on a subset of genes as identified by sctransform's default gene filtering. (Methods). **(a)** and **(d)** show the coefficient of variation on the gene variances for each dataset. **(b)** and **(e)** show the Pearson r2 between the raw cell depth and the transformed cell depth. **(c)** and **(f)** show one minus the absolute value of the mean Spearman r on the raw vs transformed cell. A bar is plotted to the mean of each distribution (also marked with a red circle). The 10xv3_nih_3t3 dataset is marked with a blue circle.

The sctransform method subsets the genes analyzed (see Methods), so to compare sctransform to other methods we redid the analysis of each method with respect to the sctransform selected genes (Fig. 5.2d); we found the results to be qualitatively consistent with the full analysis using all genes. In terms of depth normalization, we found that even methods that claim to normalize for depth, e.g., sctransform, do not succeed in completely removing depth effects and retain information about depth in the normalized data (Fig. 5.2e). Popular normalization methods such as log1pCP10k are similar in terms of removing the effects of depth on downstream analysis (Fig. 5.2b, 2e). The sctransform normalized cells, for example, exhibit similar cell-depth correlation (r2 = 0.37) as log1pPF cells (r2 = 0.43) on average, with some sctransform normalized datasets exhibiting very high depth correlation. Finally, while most transformations are monotonic, we find that sctransform scram-

bles the rank order of genes in individual cells (Fig. 5.2c, f), a straightforward result of the normalization procedure that can negatively influence downstream analyses if not taken into consideration. Variance stabilization Our analysis of current normalization methods shows that they exhibit a stark tradeoff between variance stabilization and depth normalization. To understand the implications of each normalization technique we analyzed data from (Angelidis et al., 2019), as studied in (Ahlmann-Eltze and Huber, 2021).

Interestingly, there has been much focus on variance stabilization, perhaps because variance stabilization has a long history dating back to (Bartlett, 1936). A relationship between expression levels of a gene and its variance can mask biological variation and affect data analysis methods such as PCA as a result of technical artifacts (e.g., sampling). Highly expressed genes may dominate PCA components, regardless of biologically meaningful variation. When analyzing angelidis_2019 we found that PF, like the raw counts, was not variance stabilized resulting in non-uniform PC loadings corresponding to low entropy for genes (Fig. 5.2a), with PC loadings increasing with increasing gene mean. sctransform had the highest entropy, a finding that can be explained by the heuristic clipping procedure performed on the gene variances (Choudhary, 2019).

To address this problem, approximations to variance stabilizing transforms, such as log1p or sqrt are used, often in conjunction with a depth normalization step such as PF, CP10k, and CPM. Variance stabilizing transforms like log1p and sqrt reduce the CV of the genes from angelidis_2019 by a factor of 29.1 and 7.9, respectively (from 98.8 to 3.4 and 12.5). The addition of depth normalization step does not greatly affect the CV for log1pPF (3.0). Therefore normalization techniques that include a variance stabilization step will greatly reduce the effects that highly expressed, and thus highly variable, genes have on PC components.

The log transformation is often used with a pseudocount, and the size of the pseudocount can be seen to reflect assumptions about the extent of overdispersion (Ahlmann-Eltze and Huber, 2021; Booeshaghi and Pachter, 2021). For negative binomial data, the overdispersion is the constant α in a quadratic mean (μ) - variance ($\sigma 2$) relationship of $\sigma 2 = \mu + \alpha\ \mu 2$. Depth normalizations prior to logarithmic transformation with a pseudocount of 1 therefore reflect assumptions about the overdispersion as reflected in the size factor. As pointed out in (Ahlmann-Eltze and Huber, 2021), a large size factor represents an assumption of high overdispersion. For example, (Ahlmann-Eltze and Huber, 2021) show that scaling counts with a size

factor of one million by computing CPM in a scRNAseq dataset with an average of 5,000 counts per cell, is equivalent to using a pseudo-count of 0.005. This amounts to assuming an overdispersion of $\alpha = 50$. This calculation is based on a variance stabilizing approximation derived by the Delta method that yields a pseudocount of $1/4\alpha$; interestingly, there is some disagreement over the denominator of the pseudocount as $1/2\alpha$ (Anscombe, 1948) is frequently preferred. In a simulation study (see Methods), we found that in the range of relevant overdispersion parameters, $1/4\alpha$ provides a slightly better variance stabilizing transform than $1/2\alpha$.

Our results (Fig. 5.3) reflect the different assumptions about overdispersion underlying the use of log1pPF, log1pCP10k, or log1pCPM depth-normalization, however they also show that a smaller CV is not necessarily indicative of better variance stabilization. For example, the CPM assumption of overdispersion, that is at least two orders of magnitude larger than present in biological datasets, results in overcorrection and the removal of biological variation (Ahlmann-Eltze and Huber, 2021) and results in the smallest CV in angelidis_2019 of 1.7. The sqrt transformation did not perform as well at stabilizing the variance as log1p, which is not surprising given the overdispersion (relative to the Poisson distribution) of single-cell RNA-seq data. As noted previously, many methods display a linear relationship between gene mean and gene variance for cells with very low counts. This phenomenon is well known and is a consequence of Theorem 1 of (Warton, 2018). The sctransform method is an exception, because when the program computes the Pearson residuals, the standard deviation for each gene is artificially required to be at least nzmedian/5 where nzmedian is the median number of counts for each gene computed over non-zero cells (Choudhary, 2019).

The log1pPF method, which stabilizes the variance after depth normalization, performs well in all metrics, a result which is consistent with the findings of (Ahlmann-Eltze and Huber, 2021). Overall, while some methods achieve better variance stabilization than others since they better match the overdispersion characteristics of biological data (e.g., log1p vs. sqrt), even sqrt is effective at achieving an absolute reduction in the coefficient of variation of variance, which explains its adequacy in scprep. Similarly, while log1pCP10k is preferable to log1pCPM, the use of log1pCPM does not preclude obtaining some meaningful results in analysis (Chen et al., 2021). Indeed, all current variance stabilization procedures are heuristics that ignore the fact that Poisson and negative binomial distributions may arise due to biophysical stochasticity in bursty transcription and RNA degradation (Amrhein,

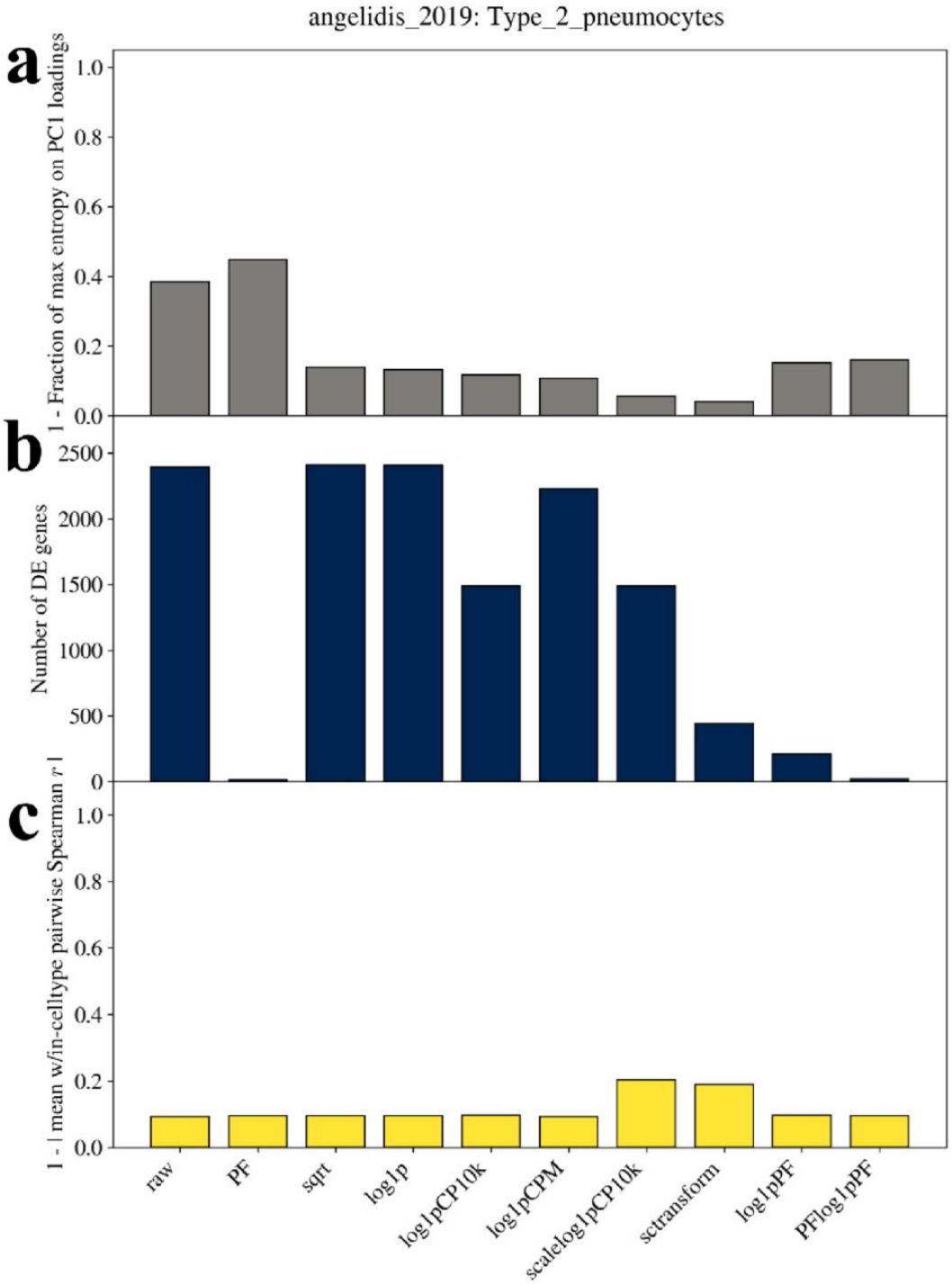


Figure 5.3: Cell-type level metrics. Three metrics are computed for cells within the Type 2 pneumocytes from angelidis_2019 for all normalization methods. **(a)** The fraction entropy of the PC1 loadings for all genes as a fraction of the max entropy. **(b)** The number of false-positive DE genes. **(c)** The absolute value of the mean within-cell-type-pairwise Spearman r.

Harsha, and Fuchs, 2019; Jahnke and Huisinga, 2007). The development of "mechanistically justified normalization" is a pressing challenge for single-cell RNA-seq

analysis. Implicit use of cell depth Current depth normalization procedures that are applied alongside variance stabilization procedures implicitly assume that differences in cell-count depth is a technical artifact due to sampling differences between cells, rather than the result of different numbers of RNA molecules in different cells. While this assumption may be flawed, in the absence of effective data and procedures for assessing variation in the amount of RNA between cells, normalization for cell sequencing depth is essential. This is because standard statistical tests that are employed in Seurat and Scanpy, such as the t-test and wilcoxon rank-sum, do not explicitly model cell-depth as a technical covariate.

To investigate the effect that poor depth-normalization can have on analysis, we selected two subsets of cells from Type 2 pneumocytes with high and low depth, respectively (see Methods). An analysis of the number of differential expressed genes detected after transformation with different methods shows that poor depth normalization can lead to many false positives (Fig. 5.3b). Interestingly, the default normalization used for differential expression analysis in Seurat and Scanpy (CP10k) finds 1,490 false-positive DE genes, about seven times more than log1pPF (Love, Huber, and Anders, 2014), which is used in (Cao et al., 2019), and has been recently recommended again (Ahlmann-Eltze and Huber, 2021). In comparison, sctransform finds 442 DE genes, about twice as many as log1pPF and about three times fewer than log1pCP10k.

Depth normalization is also important for identifying clusters with biologically meaningful gene-expression patterns. Standard clustering techniques first construct a cell-cell distance matrix based on a distance metric. Next, a k-nearest neighbor graph is constructed from the distance matrix with tools such as annoy (Bernhardsson, 2018). Finally, graph-partition methods, e.g., Louvain (Blondel et al., 2008) or Leiden (Traag, Waltman, and Van Eck, 2019), identify "communities" of cells in this graph that exhibit similar expression patterns. In the absence of proper depth normalization, cell-cell distances, computed with metrics like the l1 distance can be correlated with cell depth. For Type 2 Pneumocytes in angelidis_2019, the cell-cell distances were correlated with cell depth when normalized with sctransform (0.71) and scalelog1pCP10k (0.54) but not with PFlog1pPF (0.06). Proper depth normalization ensures that the k-nearest neighbor graph can be built with a distance metric that is cell-depth independent and results in cell communities that exhibit similar gene expression patterns.

The interpretation of PCA requires confidence that explained variation is biological

rather than technical (Lun 2018) and in the absence of depth normalization, PCA components can correlate strongly with cell depth. Normalization techniques that do not include a final depth-normalization step, like sqrt, log1p, and log1pCP10k, demonstrate a high correlation with PC1. Techniques that end with a depth normalization step like PF and PFlog1pPF, and techniques that model cell depth as a covariate, exhibit lower correlation with PC1 with the former exhibiting almost no relation to PC1. In the absence of depth normalization, subsequent analyses that rely on PCA, such as clustering or UMAP, may produce results that are affected by technical artifacts, rather than reflecting biological structure (Fig. 5.1a).

Finding markers versus differential expression

Classic variance stabilizing transformations such as the logarithm or square root functions are monotonic, a property that is rarely highlighted, but is of crucial importance. For instance, monotonicity of the transformation applied to single-cell RNA-seq counts is important for the task of finding marker genes.

The term "marker gene identification" is frequently used interchangeably with "differential expression" as noted in (Dumitrascu et al., 2021), but the two tasks are not the same. Differential expression analysis, which is the task of identifying genes exhibiting significantly different expression between groups of cells yields what can be considered a set of computational marker genes. However such genes may not constitute useful experimental markers, i.e., genes that mark cell types in a way that is experimentally actionable. Experimental marker genes for a group of cells are not only statistically differential with respect to other cells, but also specifically expressed (i.e., not present in high abundance in other cells). Thus, while experimental markers will be included in computational markers, not all computational markers are experimental markers.

One popular approach for identifying experimental marker genes is manual inspection of heatmaps, because in principle these can allow for identifying genes that not only distinguish among cell types, but that are also exceptionally highly expressed within cell types (Bonnycastle et al., 2019). The accurate depiction of gene expression in heatmaps is challenging due to the wide range of gene expression in typical experiments. To address this problem, programs such as Seurat and Scanpy scale the gene expression values across cells, by normalizing them to have mean zero and variance 1, and then clip extreme values for each gene. These values are visualized using a continuous color scale.

However, the use of heatmaps to identify marker genes requires some care. First, the additional scaling step introduces a non-monotonic transformation on cells that can scramble the relative expression of two marker genes within a cell type, possibly even reversing the ranking between them. For example in the angelidis_2019 dataset, Syce2 is a DE gene for Red Blood Cells that switches ranking within the Eosinophils cell type, from rank 76 to rank 41 out of 96 top DE genes, after the heatmap scaling procedure (Methods). Secondly, by only selecting genes that are highly expressed between cell types, one may successfully identify computational markers that may not be appropriate experimental markers due to the low expression of those markers within that cell type relative to other genes.

Use of monotonic transformations for normalizations results in cells within a cell type exhibiting a pairwise Spearman r correlation of 1 whereas for transformations such as sctransform the Spearman correlation can be much lower (Fig. 5.3c). By avoiding an initial scrambling of genes within cells, further heatmap scaling procedures can then be applied to create two heatmaps that more faithfully represent gene expression ranking within and across cells.

Scalable normalization

Compute resource constraints imposes practical limits on matrix operations. One issue that arises in the context of normalization is that some methods transform sparse matrices into dense matrices that can surpass standard RAM availability. For example, we found that the scalelog1pCP10k matrix ERX2756720 was 219 times larger than the log1p sparse matrix. Memory and speed requirements can inhibit scalable computation on increasingly large scRNA-seq datasets and drive higher cloud-computing costs (Supplementary Table 1 of (Melsted et al., 2021)). In contrast, sparse matrices have been used for high-performance computing for a long time (Orchard-Eays, 1956; Markowitz, 1957), and can drastically reduce the memory overhead required to perform memory-intensive computations. While recently developed "sketching" procedures (Hao et al., 2022) that subsample matrix operations for scalable computation may provide workarounds for dense matrices, we believe that sparsity will remain an important consideration for normalization transformations for the foreseeable future.

The PFlog1pPF heuristic

The Seurat and Scanpy workflows offer users the ability to choose different matrix types for different analysis tasks. This is a good design decision, in principle,

because different tasks make different assumptions on the count matrix. However, without clear guidelines or appropriate defaults, matrix managers like the Seurat and AnnData objects can confuse users and make analysis error-prone. A single normalization technique resulting in a single (sparse) matrix can make data sharing and reproducibility more straightforward.

While depth normalization is achieved perfectly with proportional fitting (PF), the addition of a log1p transform in log1pPF does reintroduce some depth heterogeneity (Fig. 5.2b). The importance of depth normalization therefore motivated us to explore adding an additional proportional fitting step to log1pPF. We hypothesized that an additional round of proportional fitting might achieve depth equalization without drastically affecting variance stabilization. We tested this method (PFlog1pPF) and found that to be the case on 10xv3_nih_3T3, angelidis_2019, and the other benchmark datasets.

We observed that PFlog1pPF can be seen to only slightly decrease variance stabilization (Fig. 5.2a) while ensuring depth normalization and monotonicity. With the addition of a PF step, gene variance CV suffers only slightly making PFlog1pPF comparable to sqrt with the additional benefit of full depth normalization of PF resulting in almost no false-positive differentially expressed genes (Fig. 5.3b). PFlog1pPF also recapitulates cell-type marker gene expression for angelidis_2019 and is consistent with other normalization techniques tested in (Ahlmann-Eltze and Huber, 2021). Additionally, PCA components computed on PFlog1pPF have similar loadings to log1pPF (Fig 5.3a) and within-celltype pairwise gene expression rankings are better preserved than sctransform and scalelog1pCP10k (Fig. 5.3c) both of which exhibit high concordance.

5.3 Discussion

Count normalization is a crucial first step in all scRNAseq analysis that, in principle, comprises a single step in a standard workflow. However in practice normalization is a collection of techniques, data representations, analysis types, and visualizations that interact with each other in non-obvious and frequently undocumented ways. In Seurat and Scanpy, the analysis software used for the majority of scRNAseq analysis, some normalization implementations can also limit users by requiring large amounts of memory. Thus, while users frequently think of normalization as a single data transformation step in analyses, it is often not; the software engineering choices made by developers of the tools used can affect analyses in unpredictable,

and sometimes unintended ways.

Despite the complexity of normalization in practice, much work on scRNAseq has focused on statistical details that, while important, are not necessarily the primary determinants of results. For example, the debate over whether gene-specific over-dispersion parameters should be used when computing Pearson residuals (Hafemeister and Satija, 2019; Lause, Berens, and Kobak, 2021; Choudhary and Satija, 2022) ignores the fact that Pearson residuals are not the result of a monotonic transformation, and they create dense matrices that can lead to significant analysis limitations (Borella et al., 2022). These problems have significant implications for common tasks such as finding marker genes, as discussed above. Newer methods that explicitly couple statistical methods with software engineering considerations are needed; we examined several recent publications proposing new ideas but restricted the paper to widely used methods common in existing workflows (Brown et al., 2021; Breda, Zavolan, and Nimwegen, 2021; Borella et al., 2022; Bacher et al., 2017). A detailed analysis and review of these methods is an important next step. Furthermore, normalization should ideally include modeling of transcriptional dynamics so as to be able to evaluate the contribution of technical noise to count data (Gorin and Pachter, 2023).

We have argued that a single, sparse, variance-stabilized and depth-normalized matrix on which all analysis and visualizations are performed can simplify current workflows. The PFlog1PF heuristic we have proposed is a monotonic transform on the raw counts that results in a fully depth normalized matrix and offers variance stability similar to sqrt. Importantly, we have shown that for downstream analysis, PFlog1pPF effectively stabilizes variance for PCA, produces low false-positive DE genes, and has the same within cell-type Spearman correlation as unnormalized matrices. Having said that, we believe it is an interesting challenge to develop more principled approaches that achieve depth normalization and variance stabilization while preserving sparsity and respecting monotonicity

Regardless of the normalization transformation that is applied, our work shows that assessment of data quality and normalization effectiveness is crucial in practice. Measures such as the overdispersion, coefficient of variation of the transformed-gene variances, and raw to transformed cell-depth Pearson correlation ought to be collected as part of standard quality control of experiments. It's also crucial that practitioners understand the assumptions implicit in the normalizations applied, and the implications for interpretation of results, such as whether variation is technical

or biological.

5.4 Methods

Preprocessing

Raw matrices were filtered by removing cells beneath a selected knee-plot threshold. The knee plot and threshold used for each dataset are reported in the dataset folders. Datasets for which the average count per cell was less than 818.46 (the average count per cell in angelidis_2019) were not used in Fig. 5.1.

Collecting metadata

Dataset metadata was collected with the ffq program version 0.2.1 available at https://github.com/pachterlab/ffq by running ‘ffq -l 2 -o DATASETID_metadata.json DATASETID‘. 18 out of the 526 datasets processed did not have metadata associated with their dataset ID.

Normalizing matrices

We applied seven normalization methods to the cell-filtered matrix: PF, sqrt, log1p, log1pCP10k, log1pPF, log1pCPM, PFlog1pPF. The normalization transformations were computed by running the ‘norm_sparse.sh‘ script.

We then ran ‘norm_sctransform.sh‘ on the original cell-filtered matrix to generate the sctransform matrix. The sctransform function was called with:

var_features_n = number_of_genes_in_dataset, vst_flavor = "v2",

and default parameters. In order to perform a uniform analysis, we filtered the original cell-filtered matrix to the set of genes returned by sctransform- since sctransform has a built-in gene filtering step.

We then ran *‘norm_sparse.sh‘* to create the seven normalized matrices mentioned above, and finally ran *‘norm_cp10k_log_scale.sh‘* to create the scalelog1pCP10k matrix.

Running sctransform

We performed all of our benchmarks of sctransform with v2. The sctransform v1 regression model has been shown to be overspecified (Lause, Berens, and Kobak 2021) and has been superseded by v2. We opted to benchmark sctransform v2 over analytical Pearson residuals as the latter’s validation consisted of comparing two dimensional t-SNE embeddings computed on the principle components, to compare

and contrast methods on simulated and ground truth data.

In order to run sctransform v2, a one-line modification was made to pysctransform.py (in the develop branch), namely casting *'params["order"]'* as a numpy array with *'npy.asarray(params["order"])'* in line 333. This modification fixed an issue described in https://github.com/saketkc/pySCTransform/issues/4#issue-912930103 which was causing the pip-installed version of pySCTransform not to work. An additional modification to the pySCTransform code allowed for the corrected counts matrix to be returned- line 759 *'return (vst_out["residuals"], vst_out["corrected_counts"])'*.

Computing dataset metrics

For each normalization method we computed three metrics: the coefficient of variation on the transformed-gene variances (CV), the Pearson r2 correlation between the transformed-cell counts and the raw cell counts, and the average Spearman r between the transformed-cell counts and the raw cell counts. The CV was computed by calculating the variance for each gene, across all cells, and then calculating the variance and mean across all genes, and dividing the two. The Pearson r2 was computed by summing the transformed cell counts and running *'sklearn.linear_model.LinearRegression().fit()'* followed by score() on the transformed cell counts and the raw cell counts. The average Spearman r was computed by first performing paired stats.spearmanr on all transformed-raw cell pairs and then taking the mean.

Computing cell-type metrics

Cell-type metrics were computed on cells from the Type 2 pneumocytes in the angelidis_2019 dataset. For each normalization method, *'sklearn.decomposition.PCA()'* was run with n_components=1 and svd_solver="full" and the absolute value of the loadings were l1-normalized. The entropy was computed with *'scipy.stats.entropy()'* and the max entropy was computed with *'np.log(ngenes)'*. Additionally, the Pearson r2 was computed on PC1, derived from PCA on the normalized matrix, and raw-cell depth.

To compute the number of false-positive DE gene genes, we performed differential expression on two groups of cells: 500 cells with the highest raw-cell count and 500 cells with the lowest raw-cell count. Then, for each normalization method, we performed differential expression as previously described (Booeshaghi et al. 2021). The number of differentially expressed genes with a corrected p-value less than 0.01 were recorded.

To compute the average pairwise-Spearman gene-rank correlation, we first found the smallest non-zero difference in counts between entries in each normalization matrix. We added a random number between zero and one-fourth of this minimum to each gene vector to break ties. After adjusting the matrix counts, pairwise-Spearman correlations were calculated on all cells and the average was computed.

To compute the correlation between pairwise-difference in cell depth and pairwise l1 distance, for each matrix we subsampled to 1,000 cells and then computed all pairwise differences in cell depth by running *'sklearn.metrics.pairwise_distances'* with metric="l1" on the cell sums. Then we computed the pairwise l1 distances in the same manner but on with the entire gene vectors. Lastly, *'sklearn.linear_model.LinearRegression.fit()'* and *'score()'* were used to compute the Pearson correlation.

Computing matrix metrics

The following matrix-level metrics were computed for each matrix, on both all genes and those subset by sctransform: the number of cells (ncells), the number of genes (ngenes), the number of non-zero entries in the matrix (nvals), the fraction of non-zero entries (density), the average depth per cell (avg_per_cell), the average depth per gene (avg_per_gene), the minimum depth per cell (min_cell), the maximum depth per cell (max_cell), the total number of counts in the matrix (total_count), the empirical overdispersion (overdispersion). These metrics were computed with 'metrics_matrix.sh'.

Creating multi-panel normalization figure

For each dataset and normalization, the following three plots were made: 1. a scatterplot of the transformed gene variance vs raw gene mean, 2. a scatterplot of the transformed cell depth vs raw cell depth, and 3. a histogram of the distribution of transformed-to-raw cell Spearman rank correlations. To make visualization easier, a min-max procedure was performed to scale the x and y axes of plot 2 where the min cell depth was subtracted from each cell and the result was divided by the max cell depth. These figures were made on all genes, for normalizations that were computationally tractable, and on the gene subset by sctransform for all normalizations.

Plotting styles for the gene mean-variance relationship In order to consistently visualize variance stabilization of normalization procedures against each other, we plotted all transformations on a log-log axis with the x and y-axis limits set equal.

We also plotted the identity line y equals x to illustrate the asymptotic behavior of the mean-variance relationship for genes with small mean.

Pseudocount simulation

We simulated negative binomial count data for 8,000 genes, g1, g2, . . . , g8000, as follows: we first drew the mean expression for each gene from an exponential distribution with mean 3, obtaining $\mu 1, \mu 2, \ldots, \mu 8000$. We considered the overdispersion parameters γ = 0.3, 0.5, 1, 1.5, 2, 3, 4, 5. This spans a larger range than is evident in typical single-cell RNA-seq experiments, but is informative. For each gene we generated gene counts for 10,000 cells from a negative binomial distribution with mean μi and overdispersion γk to form a 10,000 cells x 8,000 genes count matrix. We then filtered this data to remove genes with average count less than 4. Two hundred simulations were performed for each parameter setting.

Generating heatmaps

The top 100 expressed genes were found for each cell type in angelidis_2019. Then a cell type x gene matrix was made by averaging the expression of all cells within a cell type on the set of top 100 genes for that cell type. Lastly, the genes within each cell type were ranked from lowest to highest expressed using ‘scipy.stats.rankdata()‘ and the matrix of ranks was plotted on a heatmap.

To create the cell and gene-scaled cell x gene heatmaps, the top 96 DE genes for all cell types were selected and the cell x gene matrix on those 96 genes was scaled to unit variance and zero mean using ‘sklearn.preprocessing.scale()‘ across the cells to create the gene-scaled heatmap, and across the genes to create the cell-scaled heatmap. To find genes that switch rank, we first rank the raw gene expression within a cell type for the top marker genes, and then compare gene ranks to scaled (mean zero and variance one) gene expression ranks.

Data and code availability

All data and code to reproduce the figures and results in the paper are available at https://github.com/pachterlab/BHGP_2022.

References

Ahlmann-Eltze, Constantin and Wolfgang Huber (2021). “Transformation and preprocessing of single-cell RNA-seq data”. In: *bioRxiv*, pp. 2021–06.

Amrhein, Lisa, Kumar Harsha, and Christiane Fuchs (2019). "A mechanistic model for the negative binomial distribution of single-cell mRNA counts". In: *bioRxiv*, p. 657619.

Anders, Simon and Wolfgang Huber (2010). "Differential expression analysis for sequence count data". In: *Nature Precedings*, pp. 1–1.

Angelidis, Ilias et al. (2019). "An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics". In: *Nature communications* 10.1, p. 963.

Anscombe, Francis J (1948). "The transformation of Poisson, binomial and negative-binomial data". In: *Biometrika* 35.3/4, pp. 246–254.

Bacher, Rhonda et al. (2017). "SCnorm: robust normalization of single-cell RNA-seq data". In: *Nature methods* 14.6, pp. 584–586.

Bartlett, Maurice S (1936). "The square root transformation in analysis of variance". In: *Supplement to the Journal of the Royal Statistical Society* 3.1, pp. 68–78.

Bernhardsson, Erik (2018). "Annoy: Approximate nearest neighbors in C++/Python". In: *Python package version* 1.0.

Blondel, Vincent D et al. (2008). "Fast unfolding of communities in large networks". In: *Journal of statistical mechanics: theory and experiment* 2008.10, P10008.

Bonnycastle, Lori L et al. (2019). "Single-cell transcriptomics from human pancreatic islets: sample preparation matters". In: *Biology Methods and Protocols* 4.1, bpz019.

Booeshaghi, A Sina and Lior Pachter (2021). "Normalization of single-cell RNA-seq counts by log (x+ 1) or log (1+ x)". In: *Bioinformatics* 37.15, pp. 2223–2224.

Borella, Matteo et al. (2022). "PsiNorm: a scalable normalization for single-cell RNA-seq data". In: *Bioinformatics* 38.1, pp. 164–172.

Breda, Jérémie, Mihaela Zavolan, and Erik van Nimwegen (2021). "Bayesian inference of gene expression states from single-cell RNA-seq data". In: *Nature Biotechnology* 39.8, pp. 1008–1016.

Brown, Jared et al. (2021). "Normalization by distributional resampling of high throughput single-cell RNA-sequencing data". In: *Bioinformatics* 37.22, pp. 4123–4128.

Cao, Junyue et al. (2019). "The single-cell transcriptional landscape of mammalian organogenesis". In: *Nature* 566.7745, pp. 496–502.

Chari, Tara, Joeyta Banerjee, and Lior Pachter (2021). "The specious art of single-cell genomics". In: *BioRxiv*, pp. 2021–08.

Chen, Wanqiu et al. (2021). "A multicenter study benchmarking single-cell RNA sequencing technologies using reference samples". In: *Nature Biotechnology* 39.9, pp. 1103–1114.

Choudhary, Saket (2019). “pysradb: A Python package to query next-generation sequencing metadata and data from NCBI Sequence Read Archive”. In: *F1000Research* 8.

Choudhary, Saket and Rahul Satija (2022). “Comparison and evaluation of statistical error models for scRNA-seq”. In: *Genome biology* 23.1, p. 27.

Cole, Michael B et al. (2019). “Performance assessment and selection of normalization procedures for single-cell RNA-seq”. In: *Cell systems* 8.4, pp. 315–328.

Crowell, Helena L et al. (2020). “Muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data”. In: *Nature communications* 11.1, p. 6077.

Deming, W Edwards and Frederick F Stephan (1940). “On a least squares adjustment of a sampled frequency table when the expected marginal totals are known”. In: *The Annals of Mathematical Statistics* 11.4, pp. 427–444.

Doob, Joseph L. (1935). “The limiting distributions of certain statistics.” In: *Annals of Mathematical Statistics* 6.3, pp. 160–69.

Dumitrascu, Bianca et al. (2021). “Optimal marker gene selection for cell type discrimination in single cell analyses”. In: *Nature communications* 12.1, p. 1186.

Evans, Ciaran, Johanna Hardin, and Daniel M Stoebel (2018). “Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions”. In: *Briefings in bioinformatics* 19.5, pp. 776–792.

Gorin, Gennady and Lior Pachter (2023). “Length biases in single-cell RNA sequencing of pre-mRNA”. In: *Biophysical Reports* 3.1.

Hafemeister, Christoph and Rahul Satija (2019). “Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression”. In: *Genome biology* 20.1, p. 296.

Hao, Yuhan et al. (2022). “Dictionary learning for integrative, multimodal, and scalable single-cell analysis”. In: *bioRxiv*, pp. 2022–02.

Jahnke, Tobias and Wilhelm Huisinga (2007). “Solving the chemical master equation for monomolecular reaction systems analytically”. In: *Journal of mathematical biology* 54, pp. 1–26.

Lause, Jan, Philipp Berens, and Dmitry Kobak (2021). “Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data”. In: *Genome biology* 22.1, pp. 1–20.

Love, Michael I, Wolfgang Huber, and Simon Anders (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome biology* 15.12, pp. 1–21.

Lytal, Nicholas, Di Ran, and Lingling An (2020). “Normalization methods on single-cell RNA-seq data: an empirical survey”. In: *Frontiers in genetics* 11, p. 41.

Markowitz, Harry M (1957). “The elimination form of the inverse and its application to linear programming”. In: *Management Science* 3.3, pp. 255–269.

Melsted, Páll et al. (2021). “Modular, efficient and constant-memory single-cell RNA-seq preprocessing”. In: *Nature biotechnology* 39.7, pp. 813–818.

Nguyen, Lan Huong and Susan Holmes (2019). “Ten quick tips for effective dimensionality reduction”. In: *PLoS computational biology* 15.6, e1006907.

Orchard-Eays, Wm (1956). “An efficient form of inverse for sparse matrices”. In: *Proceedings of the 1956 11th ACM national meeting*, pp. 154–157.

Robinson, Mark D and Alicia Oshlack (2010). “A scaling normalization method for differential expression analysis of RNA-seq data”. In: *Genome biology* 11.3, pp. 1–9.

Svensson, Valentine (2020). “Droplet scRNA-seq is not zero-inflated”. In: *Nature Biotechnology* 38.2, pp. 147–150.

Tian, Luyi et al. (2019). “Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments”. In: *Nature methods* 16.6, pp. 479–487.

Traag, Vincent A, Ludo Waltman, and Nees Jan Van Eck (2019). “From Louvain to Leiden: guaranteeing well-connected communities”. In: *Scientific reports* 9.1, p. 5233.

Urban, Jan et al. (2009). “Expertomica metabolite profiling: getting more information from LC-MS using the stochastic systems approach”. In: *Bioinformatics* 25.20, pp. 2764–2767.

Vallejos, Catalina A et al. (2017). “Normalizing single-cell RNA sequencing data: challenges and opportunities”. In: *Nature methods* 14.6, pp. 565–571.

Warton, David I (2018). “Why you cannot transform your way out of trouble for small counts”. In: *Biometrics* 74.1, pp. 362–368.

Wolf, F Alexander, Philipp Angerer, and Fabian J Theis (2018). “SCANPY: large-scale single-cell gene expression data analysis”. In: *Genome biology* 19, pp. 1–5.

You, Yue et al. (2021). “Benchmarking UMI-based single-cell RNA-seq preprocessing workflows”. In: *Genome biology* 22.1, p. 339.

Chapter 6

ALGORITHMS FOR A COMMONS CELL ATLAS

6.1 Introduction

Cell atlas projects like the Human Cell Atlas (HCA) aim to produce "reference" maps for all cells in the human body (Regev et al., 2017). Specifically, the stated goal of the HCA is *"To create [...] reference maps of all human cells [...] as a basis for both understanding human health and diagnosing, monitoring, and treating disease"* (Lindeboom, Regev, and Teichmann, 2021). This aim, shared by various atlas projects like Azimuth (Hao et al., 2021) and Tabula Sapiens (Consortium* et al., 2022), entails generating a catalogue of cell types, states, locations, transitions, and lineages in all cells in the human body using a variety of data sources.

However, these current atlas projects, have multiple drawbacks that limit the scale of data preprocessing and the generation of new reference maps. First, quantifications are often limited to the gene-level and do not distinguish between spliced and unspliced forms (Rozenblatt-Rosen et al., 2017). Second, compatible tools for reanalyzing processed data are lacking. Third, the data is not necessarily preprocessed uniformly, introducing computational variability (Delorey et al., 2021). Fourth, the infrastructure of current atlases is static, limiting the ease with which new cell-type, markers and reference transcriptomes can be used to update quantifications. Finally, the task of annotating cell-types from marker gene lists is a manual and time-intensive process (Clarke et al., 2021). Addressing these challenges will be essential in facilitating data reprocessing and creating reference maps in light of the increasing volume of single-cell data being generated.

In order to overcome these drawbacks in current atlas design, we developed a set of algorithms and tools that enable the creation of what we called the Commons Cell Atlas (CCA) infrastructure. This infrastructure allows uniform single cell data preprocessing, and generates atlases that can easily incorporate new data, as well as updated information such as marker genes or cell-type definitions.

6.2 Results

To facilitate the generation of single cell atlases, we developed a collection of tools, *'mx'* and *'ec'*, that operate on cell by feature (gene/isoform/protein/peak) matrices

and marker equivalence class files, respectively. These tools solve key algorithmic and infrastructure problems in scRNA-seq preprocessing, namely barcode filtering, automated cell type assignment, marker gene selection, and iterative data reprocessing (You et al., 2021).

The first step in single-cell data analysis is filtering out low quality cells (You et al., 2021). This is usually achieved by using a *knee plot*, where the user has to visually find the inflection point that separates good from bad cells (Macosko et al., 2015). However, this method is manual and subjective, and therefore unsuitable for automated and reproducible data analysis. We solved this problem by implementing *'mx filter'*, which runs a Gaussian Mixture Model (Reynolds et al., 2009) on the 1D histogram of the UMI counts. The tool uses the fact that inflection points in a knee plot correspond to points between peaks in the 1D histogram (Fig. 6.1). Therefore, by finding the points of maximum entropy (i.e. maximum uncertainty for the GMM model (Benavent, Ruiz, and Sáez, 2009)), *'mx assign'* can identify the corresponding *knee* and use it to filter out cells in an automated and efficient way.

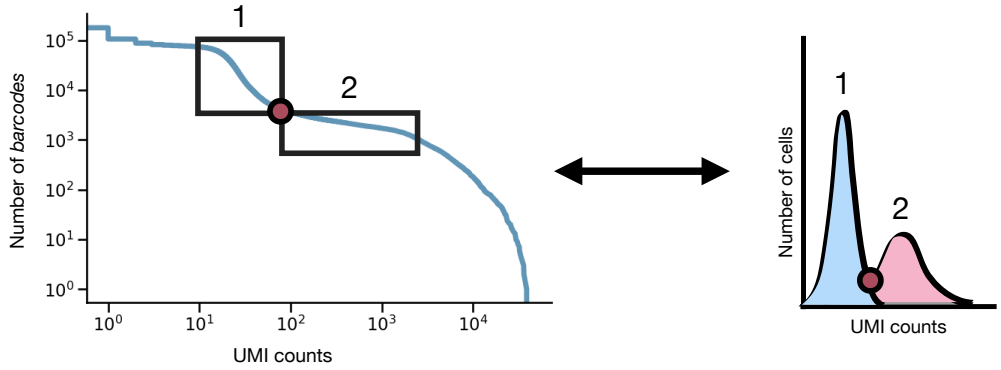


Figure 6.1: Filtering low quality cells with *'mx filter'*. Finding the point of maximum entropy between two Gaussians is equivalent to find the knee in a knee plot.

Next, we aimed to address cell-type annotation, a time-consuming task that relies on manual effort to sift through literature and lists of differentially expressed genes (Hay et al., 2018). This procedure also suffers from the *double-dipping problem*, where statistically significant genes are found by testing on groups of cells that were defined based on the differential expression of that same set of genes. Other tools, such as CellAssign (Zhang et al., 2019), address this issue by using a predefined set of marker genes for assignment. However, they still suffer from long runtimes and high memory usage (Zhang et al., 2019). We developed *'mx assign'*, which takes in a single-cell matrix and a marker gene file and performs cell-type assignment using

a modified Gaussian Mixture Model (GMM). The *'mx assign'* algorithm operates on the submatrix of marker genes, like standard algorithms such as CellAssign, but is different in two ways. First, instead of joining multiple batches together to perform assignment (Zhang et al., 2019), we perform assignments on a per matrix basis. Second, assignments are not performed on gene expression, but on the rank of the gene across cells. Because the Euclidean distance between ranked data is a constant factor multiple of the Spearman correlation, our method effectively assigns cells by Spearman correlation rather than Euclidean distance (Hotelling and Pabst, 1936).

We benchmarked *'mx assign'* against CellAssign on simulated data generated using the Splatter package (Zappia, Phipson, and Oshlack, 2017). We first assessed the accuracy of CellAssign and *'mx assign'* on a varying number of marker genes across different number of cell types. We found that *'mx assign'* accurately assigns cells to celltypes with as few as three marker genes per celltype while CellAssign performs less accurately on fewer genes (Fig. 6.2A). Next, we tested the robustness of each algorithm to the mis-specification of marker genes. We simulated the selection of a bad marker gene by setting its counts to 0 for all cells in the sample. We observed that CellAssign loses the ability to correctly assign the cell-type after the misidentification of 10 marker genes, while *'mx assign'* remains highly accurate even with a single correct marker and 14 incorrect ones (Fig. 6.2B). Finally, we benchmarked the runtime of each algorithm as a function of the number of cells to be assigned. *'mx assign'* was 350 times faster than Cellassign at assigning 8,000 cells, demonstrating the efficiency of our algorithm (Fig. 6.2C).

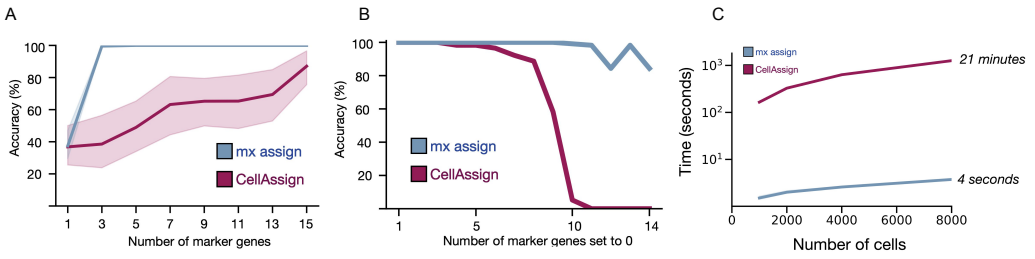


Figure 6.2: Benchmarking *'mx assign'*. **(A)** *'mx assign'* is more accurate than CellAssign with fewer marker genes. **(B)** Results from *'mx assign'* remain robust in the presence of misidentified marker genes. **(C)** *'mx assign'* runs 350 times faster than CellAssign.

Our suite of tools include other key steps in scRNA-seq processing, such as normalization *('mx norm')* and quality control *('mx inspect')*, as well as other more general matrix manipulation procedures whose utility extends beyond single cell matrices. With all these tools in hand, we set to design a workflow and associated

infrastructure for the reproducible generation of single cell atlas, which we named the Commons Cell Atlas (CCA). The workflow, illustrated in Fig. 6.3, consists on a series of *'mx'* and *'ec'* commands that can efficiently and modularly process count matrices. The final output of the workflow; the filtered, normalized and assigned matrix, constitutes the unit of the Commons Cell Atlas, which can be constantly modified by the automated and continuous running of the tools. The infrastructure is hosted on Github, making any Commons Cell Atlas easily accessible, as well as highly collaborative. We envision the creation of Commons Cell Atlas for many organisms, as well as for many purposes, each of which will grow and evolve as we learn more about the system of study.

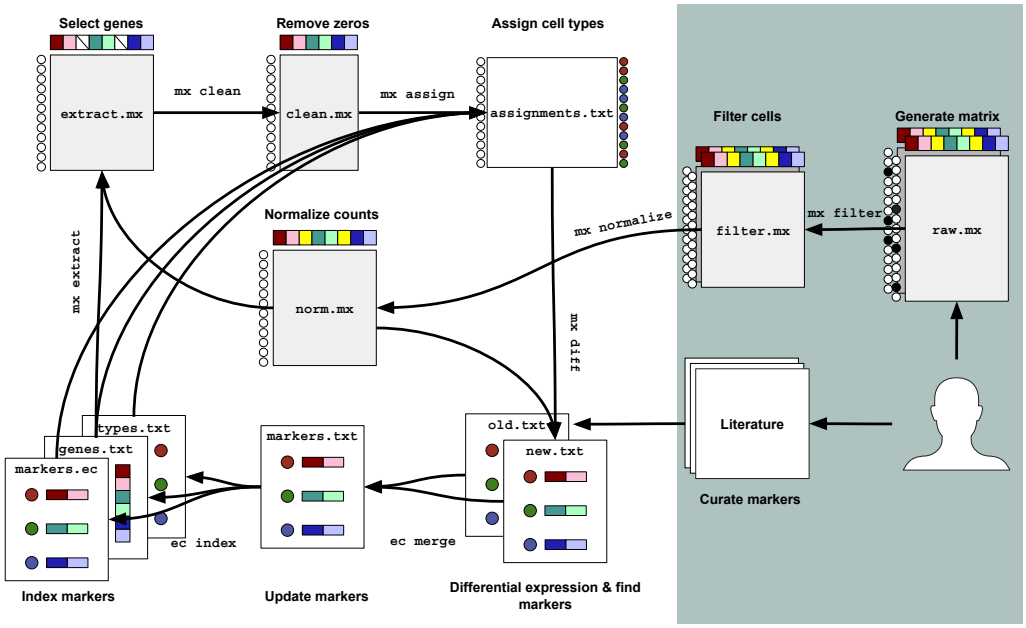


Figure 6.3: The Commons Cell Atlas infrastructure. The workflow, composed of *'mx'* and *'ec'* commands, can process raw single cell data in an efficient and reproducible manner. All elements in the figure are updatable and dynamic.

6.3 Discussion

The "single-cell atlas" has, over the past decade, gained a specific meaning within the field of genomics (Quake, 2022). It is widely understood that a single-cell atlas constitutes a collection of single-cell genomics datasets, along with defined clusters of cells that are labeled as cell types according to marker genes. We posit that a single-cell atlas includes the tools used for building, querying and analyzing the constituent data. With the CCA methods framework, we have developed several of the key tools for single-cell atlases. Our tools, such as *'mx assign'*, have been developed according to design criteria that facilitate scaling to large numbers of

samples, and regenerating / updating of atlases. In this way, the tools we have developed make a Commons Cell Atlas a "living" atlas, where data can easily be added, annotations and metadata improved, and cell type annotations regenerated with the addition of new information.

The tools we have developed are all transparent, facilitate reproducible research, are well-documented making them usable, and have been evaluated in rigorous benchmarking. We expect that this standard will be of value in other atlas projects, and the modular nature of the CCA infrastructure makes possible easy adoption of our tools for other projects. The CCA methods are also all licensed under the BSD-2 license making them freely usable both in academia and industry.

6.4 Methods

Data simulation

We simulated scRNA-seq data using the R package Splatter (Zappia, Phipson, and Oshlack, 2017) with default parameters, 10,000 genes, a DE probability of 0.2 and an even probability for each of the simulated groups. We performed simulations varying the number of cells (1000, 2000, 4000 and 8000) and the number of groups (2, 4, 6, and 8).

Assignment benchmarking

We benchmarked CellAssign and mx assign using the simulated data described above. We followed the same approach as (Zhang et al., 2019) to select marker genes. Both CellAssign and *'mx assign'* were run with default parameters for all the simulated datasets varying the number of marker genes used for the assignment. We calculated the accuracy by comparing the assignment labels of each method to the group ground truth from the Splatter output. For the robustness benchmark, we chose one group and set the expression of each of its markers to 0 one by one in random order, and we measured the accuracy of the assignment as above. For the runtime benchmark, we ran mx assign and CellAssign on the simulated dataset containing 8 groups and 1,000, 2,000, 4,000 or 8,000 cells using 15 markers per group. The runtime was calculated using the results for 3 independent runs for each condition.

Code availability

The code, as well as the complete documentation for each tool can be found in the following repositories:

'mx': https://github.com/pachterlab/mx

'ec': https://github.com/pachterlab/ec

References

Benavent, Antonio Peñalver, Francisco Escolano Ruiz, and Juan Manuel Sáez (2009). "Learning Gaussian mixture models with entropy-based criteria". In: *IEEE transactions on neural networks* 20.11, pp. 1756–1771.

Clarke, Zoe A et al. (2021). "Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods". In: *Nature protocols* 16.6, pp. 2749–2764.

Consortium*, Tabula Sapiens et al. (2022). "The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans". In: *Science* 376.6594, eabl4896.

Delorey, Toni M et al. (2021). "COVID-19 tissue atlases reveal SARS-CoV-2 pathology and cellular targets". In: *Nature* 595.7865, pp. 107–113.

Hao, Yuhan et al. (2021). "Integrated analysis of multimodal single-cell data". In: *Cell* 184.13, pp. 3573–3587.

Hay, Stuart B et al. (2018). "The Human Cell Atlas bone marrow single-cell interactive web portal". In: *Experimental hematology* 68, pp. 51–61.

Hotelling, Harold and Margaret Richards Pabst (1936). "Rank correlation and tests of significance involving no assumption of normality". In: *The Annals of Mathematical Statistics* 7.1, pp. 29–43.

Lindeboom, Rik G.H., Aviv Regev, and Sarah A Teichmann (2021). "Towards a human cell atlas: taking notes from the past". In: *Trends in Genetics* 37.7, pp. 625–630.

Macosko, Evan Z. et al. (2015). "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets." In: *Cell* 161.5, pp. 1202–1214.

Quake, Stephen R (2022). "A decade of molecular cell atlases". In: *Trends in Genetics*.

Regev, Aviv et al. (2017). "The human cell atlas". In: *elife* 6, e27041.

Reynolds, Douglas A et al. (2009). "Gaussian mixture models." In: *Encyclopedia of biometrics* 741.659-663.

Rozenblatt-Rosen, Orit et al. (2017). "The Human Cell Atlas: from vision to reality". In: *Nature* 550.7677, pp. 451–453.

You, Yue et al. (2021). "Benchmarking UMI-based single-cell RNA-seq preprocessing workflows". In: *Genome biology* 22.1, p. 339.

Zappia, Luke, Belinda Phipson, and Alicia Oshlack (2017). "Splatter: simulation of single-cell RNA sequencing data". In: *Genome biology* 18.1, p. 174.

Zhang, Allen W et al. (2019). “Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling”. In: *Nature methods* 16.10, pp. 1007–1015.

Chapter 7

A HUMAN COMMONS CELL ATLAS REVEALS CELL TYPE SPECIFICITY FOR OAS1 ISOFORMS

7.1 Introduction

The innate immune system plays a crucial role in defending the body against viruses (Takeuchi and Akira, 2009; Koyama et al., 2008; Carty, Guy, and Bowie, 2021). One of the key components of this system is Oligoadenylate synthetase 1 (OAS1), a protein that gets activated during viral infection through its binding to double-stranded RNA (Melchjorsen et al., 2009). Activated OAS1 produces 2',5'-oligoadenylates, promoting the activity of RNase L and triggering the degradation of cellular and viral RNAs to halt viral replication (Hovanessian and Justesen, 2007). Importantly, the last exon of the human OAS1 gene undergoes alternative splicing, leading to the production of isoforms with unique C-terminal sequences (Di, Elbahesh, and Brinton, 2020). Because of their distinct antiviral activities (Soveg et al., 2021), the differential expression of OAS1 isoforms correlates with susceptibility to certain viruses (Li et al., 2017). For example, the expression of OAS1-p46 has been shown to provide protection against viral infections such as West Nile virus (Lim et al., 2009), Dengue virus (Lin et al., 2009), Hepatitis C virus (El Awady et al., 2011), and most recently, SARS-CoV-2 (Zhou et al., 2021). Given the clinical relevance, several studies have investigated the regulation of the expression of OAS1 isoforms, with a particular focus on the impact of various SNPS on the relative isoform abundance (Li et al., 2017). However, no study has explored whether this regulation is tissue or cell-type specific. This question is significant: an OAS1 isoform can only protect against a virus if it is expressed in the cell-type the virus infects. The tropism of the virus could therefore render a protective OAS1's SNP useless, even if the protective isoform is overall overexpressed across the body.

A single cell atlas is the ideal tool to study cell-type specific OAS1 isoform regulation. There are a number of available cell atlases that together contain data from most human organs, such as the Human Cell Atlas (Rozenblatt-Rosen et al., 2017), the adult human cell atlas (He et al., 2020), Tabula Sapiens (Consortium* et al., 2022), the Human Cell Landscape (Han et al., 2020), Descartes (Cao et al., 2020), and Azimuth (Hao et al., 2021). However, our attempts to use these atlases for

our research question uncovered three important limitations. First, current atlases exclusively provide gene-level expression data, and lack information on isoforms. This deficiency is not minor, as evidenced by the growing body of literature supporting the critical role of isoform expression in major biological processes (Wang et al., 2008; Chaponnier and Gabbiani, 2004; Warren et al., 2003), as well as in the identification and definition of cell-types (Booeshaghi et al., 2021). Second, current cell atlas projects depend on assay-specific preprocessing tools that can introduce computationally-induced batch effects that can be challenging to identify and correct. Finally, cell atlas are static objects that cannot be easily updated or re-processed to facilitate interpretation of data according to the continuous stream of new findings emerging from single cell studies. To address these limitations and study the cell-type specificity of OAS1 isoforms in humans, we created a Human Commons Cell Atlas using the Commons Cell Atlas (CCA) infrastructure (See Chapter 6).

7.2 Results

Building the Human Commons Cell Atlas

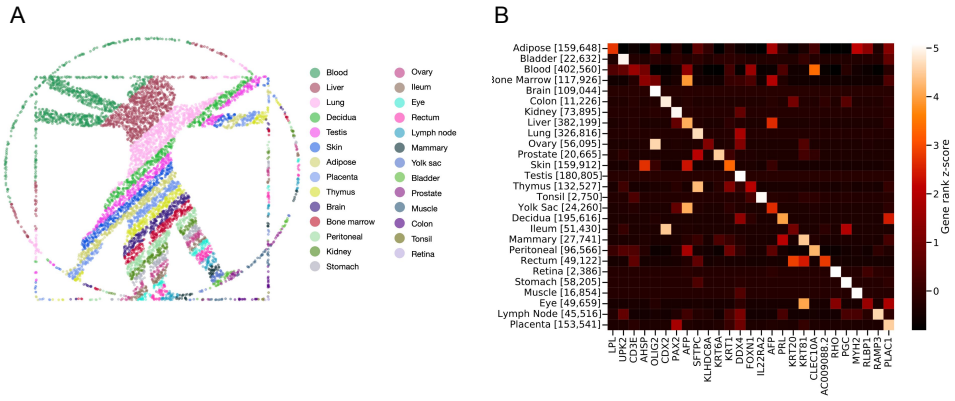


Figure 7.1: The Human Commons Cell Atlas. **(A)** 2D representation of the atlas. Cells were downsampled to match the number of coordinates of the Vitruvian man maintaining the original proportions by tissue. **(B)** Heatmap displaying the z-score of the average rank of select tissue marker genes calculated across tissues

The Human CCA comprises over 2.9 million cells from 525 publicly available scRNA-seq datasets across 27 tissues (Fig. 7.1A). These datasets were compiled from publicly available single-cell RNA-seq datasets deposited across GEO, SRA, ENA and DDBJ (Barrett et al., 2010; Leinonen, Sugawara, et al., 2010; Leinonen, Akhtar, et al., 2010; Ogasawara et al., 2020). This collection of data consisted of 147 billion sequencing reads. We chose to start with raw FASTQs, instead of gene count matrices, which is crucial for i) ensuring a uniform read alignment strategy

and barcode error correction, and ii) enabling isoform quantification, which is lost when counts are aggregated at the gene level. Data and metadata were downloaded and organized by "observation". This term maps to the GEO sample accession (GSM), and refers to a group of FASTQ files coming from a single sample and experiment.

Atlas building requires uniform processing to minimize computational variability. To that end, we leveraged the recently developed kallisto | bustools (*'kb-python'*) program (Melsted et al., 2019) to generate all 525 gene count matrices from raw sequencing data. The Human Commons Cell atlas was built in about two weeks (305 hours), with less than 8GB of memory usage. Reads were pseudoaligned to the human transcriptome, cell barcodes were corrected within hamming-1 distance of a barcode "onlist", and naÏve UMI collapsing was performed to generate gene counts (see Methods).

Gene-level Human CCA

In order to study OAS1 expression, we first sought to assess the suitability and robustness of the Human CCA for computationally identifying tissue-level marker genes (Fig. 7.1B. We first performed differential expression between the 27 tissues on the rank of all genes and identified tissue-level markers from the list of DE genes. We identified genes that are highly and specifically expressed in tissues, with most of them representing bona fide markers for each tissue. The list of genes include the LPL gene (encoding Lipoprotein Lipase) in adipose tissue (Zechner et al., 2000), the RLBP1 gene (encoding Retinaldehyde Binding Protein 1) in eye (Morimura, Berson, and Dryja, 1999), and the SFPTC gene (encoding pulmonary-associated surfactant protein C) in lung (Tredano et al., 2004) (Fig 7.1B). These findings serve as a positive control on the atlas's ability to join tissue-level groupings and prior known marker genes.

We then sought to identify if the OAS1 gene exhibits tissue-level specificity. As expected given its crucial function, OAS1 was detected in most tissues, without significant enrichment in any particular one (Fig. 7.2A). Notably, OAS1 was up-regulated in lung samples from COVID-19 infected individuals, which is consistent with OAS1 being a type I interferon (IFN)-induced gene (Melchjorsen et al., 2009) (Fig 7.2B).

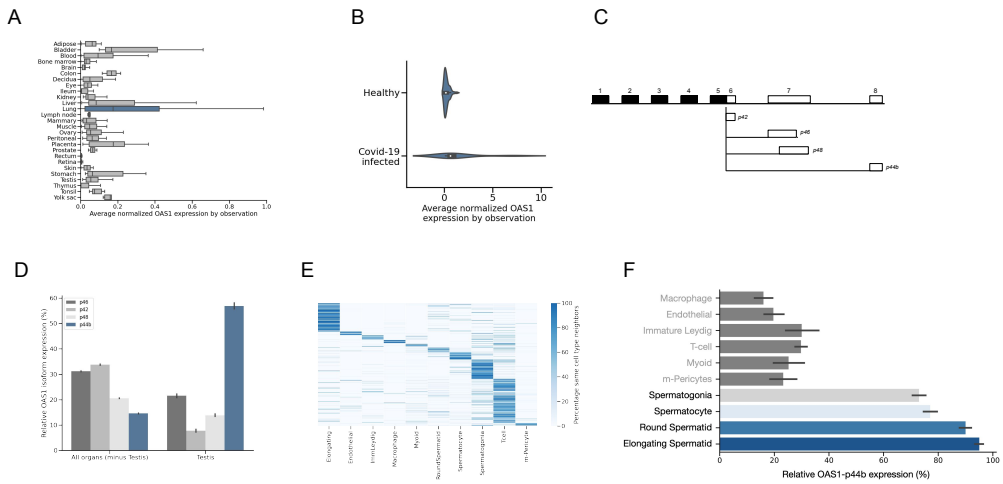


Figure 7.2: Cell-type specificity of OAS1 isoforms. **(A)** Average normalized OAS1 gene expression by dataset. Results for lung, the tissue with highest expression, are highlighted in blue. **(B)** Normalized OAS1 expression in lung datasets by health status of the individual. **(C)** Diagram of the 4 main OAS1 isoforms we found in our data. **(D)** Relative OAS1 isoform expression in all tissues except testis, and testis. OAS1-p44b is highly and specifically expressed in testis. **(E)** Validation of testis cell-type assignments. The clusters indicate that cells from the same cell-type are frequently neighbors in the K-Nearest Neighbor Graph. **(F)** OAS1-p44b is the main OAS1 isoform in germ cells undergoing spermatogenesis.

Isoform-level Human CCA

The expression of isoforms within genes can vary greatly, even when the overall gene expression remains unchanged (Booeshaghi et al., 2021). This information is lost in currently published atlases, which fail to quantify transcript isoforms. We hypothesized that OAS1 isoforms exhibited tissue-level specificity. To test this hypothesis, we rebuilt our atlas at the isoform level leveraging transcript compatibility counts (See Methods) (Ntranos et al., 2019). However, since most of the publicly available single-cell data derives from 3' technologies, our quantification was limited to isoforms with distinct 3' ends.

OAS1 isoform cell-type specificity

We leveraged the unique 3'UTR of OAS1 isoforms to study their differential expression. We found that, out of the 11 OAS1 isoforms annotated in the human transcriptome, only 4 of them were significantly expressed across the atlas: p46, p42, p48 and p44-b (Fig. 7.2C). We failed at detecting cell-type specificity of the protective isoform p46 (Zhou et al., 2021), which exhibited broad expression across all tissues and was among the most highly expressed isoforms alongside p42 (Fig. 7.2D). The least expressed isoform, p44-b, was responsible for only around 10% of the total OAS1 expression. This is in line with studies consistently showing that

amplifying the p44b isoform by qPCR is difficult due to its very low or undetectable expression level (Noguchi et al., 2013; Iida et al., 2021). Interestingly, we found that this trend was reversed in testis, with p44b accounting for almost 60% of total OAS1 expression and being the predominant isoform (Fig. 7.2D).

To investigate if OAS1-p44b expression was cell-type specific, we assigned cell-types using *'mx assign'*. We validated the output of *'mx assign'* by calculating, for each cell, the percentage of cells belonging to the same cell-type within their 20-nearest neighbors (Zhang, 2016). Cells from the same cell-type clustered together, validating our assignment results (Fig. 7.2E). We observed that the high OAS1-p44b expression was specific to germ cells undergoing spermatogenesis, where p44b represented over 80% of total OAS1 expression (Fig. 7.2F). To discard any possible artifacts caused by pseudoalignment, we visualized the alignments of one of the testis samples (GSM3302525) and observed high density of reads mapping to OAS1's exon 8, which is unique to the isoform p44b. Moreover, this result was not sample or paper-dependent, with Round Spermatids across all testis samples expressing high levels of OAS1-p44b.

Screen of cell-type specific isoform switching

Given the ability of the human CCA to quantify isoforms with distinct 3' UTRs, we decided to screen for cell-type specific isoform switching within testis (Fig. 7.3). We found genes with differential isoform usage in all cell-types except for T-cells. Some of the hits, such as IFT27 in Spermatocytes, have been shown to have essential roles within the testis (Zhang et al., 2017). Our results suggest that some of these roles may be carried out by specific isoforms, and that isoform switching may have be an important regulatory mechanism in a number biological processes.

7.3 Discussion

The question of how to organize a single-cell atlas is complex, and there is little agreement on even the most basic questions, such as what constitutes a "cell-type". In addition to conceptual problems, engineering challenges abound. Single-cell genomics datasets are growing in both number and size, and it is non-trivial to engineer atlases so that they can be updated when new datasets or biology are discovered. As we have demonstrated, the Commons Cell Atlas concept offers a solution to many of these problems by virtue of reframing single-cell atlases as a dynamic collection of data and, crucially, tools for processing, querying and interpreting the data. Our isoform analysis was possible thanks to this design principle; in order to obtain iso-

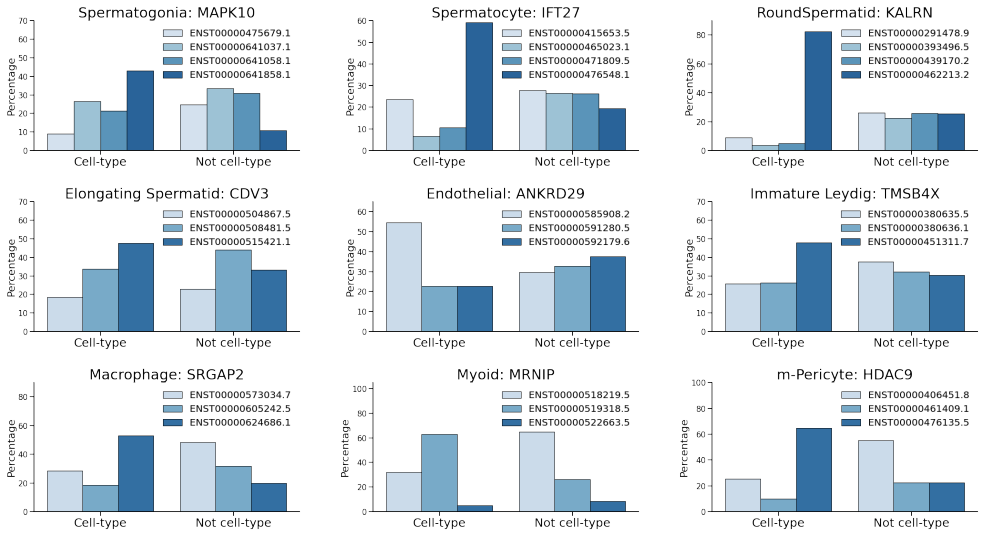


Figure 7.3: Quantifying isoforms with distinct 3' UTRs. Examples of cell-type specific isoform switches in testis.

form quantifications we needed to reprocess the data several times. Other questions may demand alternative atlas processing that, with the Commons Cell Atlas architecture, should be tractable to implement. Most importantly, the Commons Cell Atlas principle dictates that there is no definitive Commons Cell Atlas, but rather numerous Commons Cell Atlases that are customized and specific to the questions being explored.

One of the main aims of cell atlases is to provide a comprehensive characterization and classification of cells (Regev et al., 2017), which relies heavily on identifying their cell-type. The cell-type assignments within the Commons Cell Atlas can be constantly updated, thereby enabling continual refinement of the derived results as new information emerges. To show that this is a feasible strategy, we obtained all marker genes from single cell publications whose data was not used to build the atlas. Despite data and marker genes coming from different sources, we were able to successfully assign cells from most tissues in the atlas, as measured by the percentage of same-cell-type neighbors (as shown in Fig. 7.2E for testis). We expect that these assignments will change as we learn more about each tissue, enabled by the dynamism and efficiency of the Common Cell Atlas infrastructure.

The pivotal function of OAS1 in innate immunity has been well-established, and recent studies have demonstrated that differential expression of its isoforms can affect susceptibility to viruses, including SARS-CoV-2 (Zhou et al., 2021). The work

described here enables the study of this differential expression across organs and cell-types, providing a valuable resource for our understanding of viral immunity. We have used our Human Commons Cell Atlas to discover previously unknown cell-type specificity for an OAS1 isoform. Interestingly, this isoform was deemed undetectable across many studies, and our atlas provides an explanation for this results. Our finding is preliminary, and a comprehensive assessment of p44b function and activity in the testis is beyond the scope of our paper. But our discovery highlights the utility of a Commons Cell Atlas, and more generally, points towards the importance of carefully assessing isoform cell-type specificity.

The Human Cell Atlas is composed of 27 tissues, 526 cell-types and 3,554 marker genes. The whole atlas is hosted on Github, and it is therefore readily available to download, inspect, modify and use. We envision that the Human Commons Cell Atlas will constantly evolve as we increase our knowledge on tissues and cell-types, moving away from the idea of achieving a "final" or "complete" atlas. For as long as we continue discovering new cell-types, developing new single cell technologies, and gathering new single cell data, the Human Commons Cell Atlas and all its derived results will continue to grow and improve.

7.4 Methods

Downloading data

Datasets accession were obtained from the following database (Svensson, Veiga Beltrame, and Pachter, 2020). Metadata and links to raw data were collected using the ffq program version 0.2.1 (available at https://github.com/pachterlab/ffq) by running 'ffq DATASETID'.

Preprocessing data

Reads from each dataset were pseudoaligned to the human transcriptome, which was obtained by running textit'kb ref -i index.idx -g t2g.txt -d human'. Reads were uniformly processed using the textit'kallisto | bustools' python wrapper textit'kb-python', running the command *'kb count -i index.idx -g t2g.txt -x [technology]'*.

Filtering matrices

To filter the gene count matrices, barcodes with low UMIs were filtered with *'mx filter'*, which uses a derivation of the knee plot approach (See Chapter 6). Barcodes with more than 40% mitochondrial genes were discarded.

Normalizing matrices

Gene count matrices were normalized using *'mx norm'*, which uses the log1pPF method (See Chapter 6).

Marker genes curation

A list of marker genes for each tissue was generated from supplementary tables of single cell publications containing differential expression information. Markers were selected applying filters to the corrected p-value, log fold change, and percentage of cells expressing the gene. A Google Colab notebook that downloads the supplementary table, filters it, and generates a markers file is available at the Human Cell Atlas repository for each tissue.

Cell-type assignment and validation

Cell-types were assigned running 'mx assign' on each individual dataset using the marker gene file generated above. Assignments were validated by calculating, for each cell, the percentage of cells that belong to the same cell-type in the k-nearest neighbors (KNN) graph, with k=20, within each dataset. The KNN graph was calculated using the union of marker genes of the corresponding tissue.

2D representation of the atlas

To create a 2D latent space in which cells from the same tissue are neighbors, we used MCML (Chari, Banerjee, and Pachter, 2021) with the fracNCA parameter set to 1 (this is, optimizing only the Neighborhood Component Analysis (Goldberger et al., 2004) (NCA) loss). We then calculated the pairwise distances of each cell's 2D coordinates to the Vitruvian man 2D coordinates, using the L_1 norm or manhattan distance. The distances were used as input to the scipy 'linear_sum_assignment' function to map the 2D latent space to the 2D shape coordinates, assigning each cell coordinate to a shape coordinate while minimizing the total cost or distance (as per the distance matrix).

Tissue-level markers

For each dataset, we calculated the average expression of each gene across all cells, and used that value to rank each gene in the dataset. The gene ranks of datasets from the same tissue were averaged, resulting in a tissues x genes matrix. Genes with high value in one tissue and low in the others were selected, and the Z-scores across tissues were plotted using a heatmap.

OAS1 gene expression by tissue

The average normalized OAS1 expression for each dataset was calculated across all cells. Each data point in Fig. 7.2A and B corresponds to a different dataset.

Isoform quantification

A transcript Compatibility Counts (TCC) matrix for each sample was obtained by running *'bustools count'* without the *'–genecount'* option on the bus files generated after pseudoaligning the raw reads. Transcript abundances were quantified using the EM algorithm by running *'kallisto quant-tcc'* on the TCC matrices. The transcript abundance matrix of each sample was normalized within each cell-type using log1pPF. The normalized matrix was then subsetted to isoforms that i) derived from genes with more than one isoform, ii) had reads in the samples that mapped uniquely to it and iii) had a minimum average normalized expression of 0.002 per cell.

Data and code availability

The Human Commons Cell Atlas can be accessed here: https://github.com/cellatlas/human

References

Barrett, Tanya et al. (2010). "NCBI GEO: archive for functional genomics data sets—10 years on". In: *Nucleic acids research* 39.suppl_1, pp. D1005–D1010.

Booeshaghi, A Sina et al. (2021). "Isoform cell-type specificity in the mouse primary motor cortex". In: *Nature* 598.7879, pp. 195–199.

Cao, Junyue et al. (2020). "A human cell atlas of fetal gene expression". In: *Science* 370.6518, eaba7721.

Carty, Michael, Coralie Guy, and Andrew G Bowie (2021). "Detection of viral infections by innate immunity". In: *Biochemical pharmacology* 183, p. 114316.

Chaponnier, Christine and Giulio Gabbiani (2004). "Pathological situations characterized by altered actin isoform expression". In: *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland* 204.4, pp. 386–395.

Chari, Tara, Joeyta Banerjee, and Lior Pachter (2021). "The specious art of single-cell genomics". In: *BioRxiv*, pp. 2021–08.

Consortium*, Tabula Sapiens et al. (2022). "The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans". In: *Science* 376.6594, eabl4896.

Di, Han, Husni Elbahesh, and Margo A Brinton (2020). "Characteristics of human OAS1 isoform proteins". In: *Viruses* 12.2, p. 152.

El Awady, Mostafa K et al. (2011). “Single nucleotide polymorphism at exon 7 splice acceptor site of OAS1 gene determines response of hepatitis C virus patients to interferon therapy”. In: *Journal of gastroenterology and hepatology* 26.5, pp. 843–850.

Goldberger, Jacob et al. (2004). “Neighbourhood components analysis”. In: *Advances in neural information processing systems* 17.

Han, Xiaoping et al. (2020). “Construction of a human cell landscape at single-cell level”. In: *Nature* 581.7808, pp. 303–309.

Hao, Yuhan et al. (2021). “Integrated analysis of multimodal single-cell data”. In: *Cell* 184.13, pp. 3573–3587.

He, Shuai et al. (2020). “Single-cell transcriptome profiling of an adult human cell atlas of 15 major organs”. In: *Genome biology* 21, pp. 1–34.

Hovanessian, Ara G and Just Justesen (2007). “The human 2’-5’ oligoadenylate synthetase family: unique interferon-inducible enzymes catalyzing 2’-5’ instead of 3’-5’ phosphodiester bond formation”. In: *Biochimie* 89.6-7, pp. 779–788.

Iida, Kei et al. (2021). “Switching of OAS1 splicing isoforms mitigates SARS-CoV-2 infection”. In: *BioRxiv*, pp. 2021–08.

Koyama, Shohei et al. (2008). “Innate immune response to viral infection”. In: *Cytokine* 43.3, pp. 336–341.

Leinonen, Rasko, Ruth Akhtar, et al. (2010). “The European nucleotide archive”. In: *Nucleic acids research* 39.suppl_1, pp. D28–D31.

Leinonen, Rasko, Hideaki Sugawara, et al. (2010). “The sequence read archive”. In: *Nucleic acids research* 39.suppl_1, pp. D19–D21.

Li, He et al. (2017). “Identification of a Sjögren’s syndrome susceptibility locus at OAS1 that influences isoform switching, protein expression, and responsiveness to type I interferons”. In: *PLoS genetics* 13.6, e1006820.

Lim, Jean K et al. (2009). “Genetic variation in OAS1 is a risk factor for initial infection with West Nile virus in man”. In: *PLoS pathogens* 5.2, e1000321.

Lin, Ren-Jye et al. (2009). “Distinct antiviral roles for human 2’, 5’-oligoadenylate synthetase family members against dengue virus infection”. In: *The Journal of Immunology* 183.12, pp. 8035–8043.

Melchjorsen, Jesper et al. (2009). “Differential regulation of the OASL and OAS1 genes in response to viral infections”. In: *Journal of interferon and cytokine research* 29.4, pp. 199–208.

Melsted, Páll et al. (2019). “Modular and efficient pre-processing of single-cell RNA-seq”. In: *BioRxiv*, p. 673285.

Morimura, Hiroyuki, Eliot L Berson, and Thaddeus P Dryja (1999). “Recessive mutations in the RLBP1 gene encoding cellular retinaldehyde-binding protein in a form of retinitis punctata albescens.” In: *Investigative ophthalmology & visual science* 40.5, pp. 1000–1004.

Noguchi, Satoshi et al. (2013). “Differential effects of a common splice site polymorphism on the generation of OAS1 variants in human bronchial epithelial cells”. In: *Human Immunology* 74.3, pp. 395–401.

Ntranos, Vasilis et al. (2019). “A discriminative learning approach to differential expression analysis for single-cell RNA-seq”. In: *Nature methods* 16.2, pp. 163–166.

Ogasawara, Osamu et al. (2020). “DDBJ Database updates and computational infrastructure enhancement”. In: *Nucleic acids research* 48.D1, pp. D45–D50.

Regev, Aviv et al. (2017). “The human cell atlas”. In: *elife* 6, e27041.

Rozenblatt-Rosen, Orit et al. (2017). “The Human Cell Atlas: from vision to reality”. In: *Nature* 550.7677, pp. 451–453.

Soveg, Frank W et al. (2021). “Endomembrane targeting of human OAS1 p46 augments antiviral activity”. In: *Elife* 10, e71047.

Svensson, Valentine, Eduardo da Veiga Beltrame, and Lior Pachter (2020). “A curated database reveals trends in single-cell transcriptomics”. In: *Database* 2020.

Takeuchi, Osamu and Shizuo Akira (2009). “Innate immunity to virus infection”. In: *Immunological reviews* 227.1, pp. 75–86.

Tredano, Mohammed et al. (2004). “Mutation of SFTPC in infantile pulmonary alveolar proteinosis with or without fibrosing lung disease”. In: *American journal of medical genetics Part A* 126.1, pp. 18–26.

Wang, Eric T et al. (2008). “Alternative isoform regulation in human tissue transcriptomes”. In: *Nature* 456.7221, pp. 470–476.

Warren, Chad M et al. (2003). “Titin isoform expression in normal and hypertensive myocardium”. In: *Cardiovascular research* 59.1, pp. 86–94.

Zechner, Rudolf et al. (2000). “The role of lipoprotein lipase in adipose tissue development and metabolism”. In: *International Journal of Obesity* 24.4, S53–S56.

Zhang, Yong et al. (2017). “Intraflagellar transporter protein (IFT27), an IFT25 binding partner, is essential for male fertility and spermiogenesis in mice”. In: *Developmental biology* 432.1, pp. 125–139.

Zhang, Zhongheng (2016). “Introduction to machine learning: k-nearest neighbors”. In: *Annals of translational medicine* 4.11.

Zhou, Sirui et al. (2021). “A Neanderthal OAS1 isoform protects individuals of European ancestry against COVID-19 susceptibility and severity”. In: *Nature medicine* 27.4, pp. 659–667.