Learning the sequence code for mRNA and protein abundance in human immune cells

Benoît P. Nicolet[1,2,*], Anouk P. Jurgens[1], Kaspar Bresser[1], Aurélie Guislain[1], and Monika C. Wolkers[1,*]

[1] Department of Hematopoiesis, Oncode Institute, Sanquin Research, Landsteiner Laboratory, Amsterdam UMC, University of Amsterdam, Plesmanlaan 125, 1066 CX Amsterdam, The Netherlands.

[2] present address: Division of Molecular Oncology & Immunology, Oncode Institute, The Netherlands Cancer Institute; Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands.

[*] Corresponding authors: Email: m.wolkers@sanquin.nl | b.nicolet@nki.nl

ABSTRACT:

Protein abundance is defined by transcriptional, post-transcriptional and post-translational regulatory mechanisms. Understanding the code for gene expression could inform novel therapies. Here, we develop a machine learning pipeline, termed SONAR, to decipher the endogenous sequence code that determines mRNA and protein abundance in human cells. SONAR predicts up to 63% of protein abundance independent of promoter or enhancer information, and reveals a strong—yet dynamic—cell-type specific sequence code. The deep knowledge of SONAR provides a map of biologically active sequence features (SFs), which we leveraged to manipulate protein expression and tailor it to a specific cell-type. Beyond its fundamental findings, our work provides novel means to improve immunotherapies and biotechnology applications.

One Sentence Summary:

SONAR reveal the cell type-specific sequence code for mRNA and protein expression in human immune cells

Main Text:

The development, function and long-term maintenance of immune cells requires an intricate regulation of mRNA and protein abundance, defined by the interplay of transcriptional, post-transcriptional and post-translational events (*1*). Unsurprisingly, dysregulation of gene expression in immune cells is associated with autoimmunity, chronic infection, and cancer progression (*2*, *3*). Understanding the rules for gene expression across all regulatory levels, beyond fundamental interest, is critical to inform novel therapeutic intervention.

All gene expression regulatory levels depend on a crosstalk of SFs contained in DNA (e.g. promoter motifs), RNA (e.g. RNA motifs) and protein (e.g. amino acid residues) together with effectors: e.g. transcription factors, RNA-binding protein (RBP) and protein-protein interaction, respectively (*4*, *5*). Ultimately, all regulation synergizes to affect the cell's capacity to adapt to environmental cues. Immune cells are particularly well adapted to swift alteration phenotype and actively rely on mRNA decay and translation control to sustain preparedness to immunological insults (*6*–*8*). This preparedness requires RBPs and microRNAs (miRs; (*9*, *10*)), mRNA modifications (e.g. m6A) and sequence-intrinsic factors, e.g. GC-content and codon usage (*4*, *11*–*13*). Swift alteration in protein abundance is further tuned by post-translational modifications (PTMs) of amino acids residues (e.g. ubiquitination of lysine (*14*)). Despite SF involvement in all the aforementioned processes, a comprehensive map of their relative biological effect to define protein abundance is lacking.

Integrating the large number of SFs to decipher the sequence code requires advanced mathematical modelling such as machine learning (ML). We aimed to develop a model to learn the endogenous sequence code that defines protein abundance, that: 1) allows one to examine the relative importance of SFs, 2) can be audited, and 3) tolerates the wide range of sequence length found in the human transcriptome.

To achieve this goal, we developed SONAR, an ML pipeline. We generated an SF library, by searching the entire human coding transcriptome (94,348 mRNA isoforms) for SFs, including: (1) sequence-intrinsic SFs such as GC content and codon usage, (2) sequence-specific SFs such as predicted miR seed score and RBP motifs, and (3) predicted mRNA modification sites. These SFs were quantified across the different mRNA regions (i.e., CDS and UTRs, see Methods; Fig. 1A, S1A), resulting in 7,112 SFs. We then trained SONAR on published mRNA abundance data of different cell lines (n=5; (*15*)) and on a variety of

peripheral blood-derived human immune cell subsets (n=82; (*16*, *17*); Table S1), using 80% of the data for training and 20% for testing of the models. SONAR uses XGBoost (*18*), which outperformed other algorithms in modelling mRNA abundance (Fig. S1B). Independently from any knowledge on transcriptional output, the average prediction accuracy of mRNA abundance reached 46.7% for primary immune cells (Fig. 1B, S1C). Models trained on randomly shuffled protein abundance data lacked any predictive capacity (R^2: 0. 0008), confirming the predictive nature of SONAR models. With up to 61.8% prediction accuracy in cell lines (Fig. 1B-C), SONAR even outperformed previous promoter/enhancer-based models in predicting mRNA abundance (*19*).

SONAR can be audited to reveal the relative contribution of individual SF by quantifying the effect of removing a certain feature from the model, termed 'feature importance' (Data S1). Not all SFs contributed equally to SONAR's predictive capacity, as illustrated by the fact that 2,135 SFs (30% of all SFs) are required to reach 95% of the model's full predictive capacity (Fig. 1D, S1D). Notably, even though the CDS is to date primarily considered to instruct the protein codon sequence, SONAR reveals that the SF importance from the CDS is on average slightly higher than that of UTR-derived features for RNA abundance (Fig. 1E). This finding corroborates with the strong regulatory capacity reported for the CDS of the tumor repressor gene TP53/p53 (*20*), and the high RBP binding to the CDS in cell lines (*21*), further highlighting the regulatory role of CDS features. Together, SONAR is a powerful method to predict and understand the SF code of mRNA abundance.

We next applied SONAR to learn the SF code for protein abundance (Data S2). Again, models trained on randomly shuffled protein data lacked any predictive capacity (R^2: 0.0008), the overall prediction accuracy for protein abundance of primary immune cells averaged 40.7% (n=112; (*22*)), and reached a striking average of 62.7% on in human cell lines (n=9; (*23*); Fig. 2A). We next analyzed the specific contribution of SFs to protein abundance. The most predictive SFs were dominated by sequence motifs, such as GC and TOP-like CT-rich motifs occurring in the 5'UTR and CDS (*24*), CT-rich (CDS, 3'UTR), AU-rich motifs and canonical poly-adenylation signal sites in the 3'UTR (Fig. 2B-C; Fig. S2). GC content, region length, codon usage (CDS), putative m6A and m7G sites, and putative PTM sites (e.g. acetylation and ubiquitination) were also in the top contributing SFs (Fig. S2). SONAR revealed position-dependent SF effects, as exemplified by the effect of m6A site occurrence on different mRNA regions (Fig. S3). SF usage was sufficient to recapitulate

cell types (Fig. 2D), indicating that cell type-specific mechanisms are in place that determine protein abundance. Additional clustering revealed the map of cell type-specific SFs (Fig. S4, Data S3). Together, SONAR can model, understand, and predict the cell type-specific SF landscape needed for protein expression.

We next tested how SONAR models compared to mRNA abundance data to predict protein abundance. The superior performance of SONAR (R^2:0.42) over experimental mRNA measurements (R^2:0.26) to predict protein abundance indicate that SONAR may in part capture mRNA abundance rules (Fig. 2E). To test this, we supplemented mRNA abundance data to SONAR SF models (SF+RNA). mRNA abundance data further improved the prediction accuracy by 11% (R^2:0.53), reaching up to 69% in cell lines (Fig. 2E), and reduced the number of SFs needed for prediction (Fig. S5). This suggests that mRNA abundance rules are partially captured by SONAR protein abundance models. Together, SONAR not only captures the SF landscape of mRNA and protein abundance, but synergized with mRNA measurements to predict protein abundance with up to 69% accuracy.

To further dissect the SF contribution to protein abundance in the context of mRNA abundance and protein translation rate, we used SONAR to model the dynamic mRNA abundance, protein abundance, and translation rate throughout T cell activation ((*7*); Fig. 3A; Data S4). T cells undergo profound transcriptome and proteome remodeling, and are therefore an ideal model system to study the dynamics of transcriptional and post-transcriptional events (*7*, *23*). While all models had similar overall SF usage per mRNA region, the SF usage altered during T cell activation, with those located in the CDS and 3'UTR being the least variable (Fig. S6). Clustering of SF usage throughout T cell activation revealed clear patterns (Fig. S7), suggesting that alterations in SF usage may accompany alterations in gene expression. To test this, we grouped mRNAs and proteins for which abundance altered >2-fold, and we modelled which SFs could explain group differences (Fig. S8). Downregulated mRNAs were enriched for predicted m6A sites, in particular when located within the CDS (Fig. S8, Data S5). Conversely, m7G predicted sites were enriched in mRNAs and proteins that were upregulated upon activation. Similarly, the contribution of codon and amino acid occurrence to SONAR models changed throughout activation (Fig. S9), including the increased use of AAG-codon encoding Lysine shown to be associated with T cells preparedness for rapid activation (*25*, *26*). Codon usage of an mRNA has been associated with regulation of protein expression due to transfer RNA (tRNA) availability and

to ribosome dwell time (*27*). SONAR revealed the strong contribution of codon usage not only to protein models, but also to translation rate and mRNA models (Fig. S9), suggesting that SONAR captures the translation-dependent mRNA decay processes (*28*). SONAR thus reveals the intricate dynamics in SF usage throughout T cell activation, highlighting its value to identify post-transcriptional regulation dynamics governing gene expression.

To uncover the potential biological activity of SFs, we compared the SF contribution to mRNA abundance, translation rates or protein abundance models (Fig. 3B). Translation rate and protein models preferentially used CDS parameters such as length, GC content and TOP-like (CT-rich) motifs (Fig. 3B). In line with previous reports (*11*), m1A, m5C, m6A, m7G putative sites strongly contributed to all 3 models, but most dominantly to mRNA abundance models (Fig. 3B). m7G modifications have been associated with translation efficiency (*29*). Intriguingly, SONAR suggests a novel role for m7G sites in regulating mRNA abundance (Fig. 3C), a finding we confirmed by re-analysis of transcriptome-wide internal m7G mapping data ((*29*); Fig S10; Data S6). Combined, SONAR maps the SF code for protein expression in T cells and allows for the identification of novel regulatory elements of gene expression.

The deep knowledge obtained with SONAR can be leveraged to design synthetic 3'UTRs and manipulate protein expression without altering its coding sequence. To experimentally validate that a high feature importance assigned by SONAR correctly predicts biological effects, we devised a massively parallel reporter assay (MPRA). To this end, 180nt-long synthetic 3'UTR sequences (synUTR3) containing 6 occurrences of the same 6-mer motif (n=467) were fused to the 3'end of a GFP reporter gene. The resulting synUTR3 library was introduced into human cells, and sequence motif abundances of GFP^{hi} cells were compared to GFP^{lo} cells (Fig. 4A, Fig. S11; Data S7). Analysis of the MPRA revealed distinct biological effects of different synUTR3 motifs on GFP expression (Fig. 4B-C, Fig. S11). Inspection of 6-mer SFs with high importance (SONAR) and 6-mers with strong biological effects (MPRA), independently indicated CT- or AG-content as strong predictors of protein expression (Fig. 4C-E, Fig S11D-E). While SONAR correctly predicted the SFs as strong contributors to protein expression, the MPRA confirmed the directionality of their biological effects: CT-rich motifs were generally enhanced the protein expression, whereas AG-rich motifs were largely repressive (Fig. 4C). These divergent effects of CT and AG motifs were observed across three cell types, indicating a conserved regulatory mechanism (Fig 4F). In

contrast, other motifs displayed some level of cell type specificity (Fig. 4F). To confirm the supportive or repressive activity and cell type specificity, we assessed GFP expression of 4 selected 3'UTR sequence motifs in 3 cell lines, and in primary $CD4^+$ and $CD8^+$ T cells (Fig. 4G, Fig. S12A-B, sup note 2). CTTTCTT-containing 3'UTR reporters consistently boosted the GFP protein production compared to scrambled control, and it did so in a dose-dependent manner (Fig. 4G, H, Fig. S12C). All other 3 motifs displayed a cell-type specific contribution to protein expression (Fig. 4G). These findings confirmed SONAR's capacity to predict protein expression and its potential to help tailor the protein expression for different cell types.

We hypothesized that SF code contained in RNA sequence can inform on the mRNA and protein abundance. We demonstrate that SONAR ML pipeline is a powerful tool to predict gene expression abundance. SONAR not only captured the global SF contribution, uncovered a strong cell-type specific SF landscape, but provides functionally meaningful and exploitable information on SF usage. Together, our findings improve the current transcription-only bias of mRNA and protein abundance ML research. We foresee that the SONAR-guided UTR-mediated manipulation of protein production can be applied to many domains of biology with therapeutic potential such as recombinant protein production or mRNA vaccine technologies.

REFERENCES

1. C. Buccitelli, M. Selbach, mRNAs, proteins and the emerging principles of gene expression control. *Nat Rev Genet*. **21**, 630–644 (2020).

2. E. C. Morris, S. S. Neelapu, T. Giavridis, M. Sadelain, Cytokine release syndrome and associated neurotoxicity in cancer immunotherapy. *Nature Reviews Immunology*. **22**, 85–96 (2022).

3. E.-C. Shin, P. S. Sung, S.-H. Park, Immune responses and immunopathology in acute and chronic viral hepatitis. *Nature Reviews Immunology*. **16**, 509–523 (2016).

4. B. P. Nicolet, N. D. Zandhuis, V. M. Lattanzio, M. C. Wolkers, Sequence determinants as key regulators in gene expression of T cells. *Immunological Reviews*. **304**, 10–29 (2021).

5. C. Vogel, R. De Sousa Abreu, D. Ko, S. Y. Le, B. A. Shapiro, S. C. Burns, D. Sandhu, D. R. Boutz, E. M. Marcotte, L. O. Penalva, Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Molecular Systems Biology*. **6**, 1–9 (2010).

6. F. Salerno, S. Engels, M. van den Biggelaar, F. P. J. van Alphen, A. Guislain, W. Zhao, D. L. Hodge, S. E. Bell, J. P. Medema, M. von Lindern, M. Turner, H. A. Young, M. C. Wolkers, Translational repression of pre-formed cytokine-encoding mRNA prevents chronic activation of memory T cells. *Nature Immunology*. **19**, 828–837 (2018).

7. T. Wolf, W. Jin, G. Zoppi, I. A. Vogel, M. Akhmedov, C. K. E. Bleck, T. Beltraminelli, J. C. Rieckmann, N. J. Ramirez, M. Benevento, S. Notarbartolo, D. Bumann, F. Meissner, B. Grimbacher, M. Mann, A. Lanzavecchia, F. Sallusto, I. Kwee, R. Geiger, Dynamics in protein translation sustaining T cell preparedness. *Nature Immunology*. **21**, 927–937 (2020).

8. M. Jovanovic, M. S. Rooney, P. Mertins, D. Przybylski, N. Chevrier, R. Satija, E. H. Rodriguez, A. P. Fields, S. Schwartz, R. Raychowdhury, M. R. Mumbach, T. Eisenhaure, M. Rabani, D. Gennert, D. Lu, T. Delorey, J. S. Weissman, S. A. Carr, N. Hacohen, A. Regev, Dynamic profiling of the protein life cycle in response to pathogens. *Science*. **347**, 1259038–1259038 (2015).

9. P. Kafasla, A. Skliris, D. L. Kontoyiannis, Post-transcriptional coordination of immunological responses by RNA-binding proteins. *Nature Immunology*. **15**, 492–502 (2014).

10. M. Turner, A. Galloway, E. Vigorito, Noncoding RNA and its associated proteins as regulatory elements of the immune system. *Nature Immunology*. **15**, 484–491 (2014).

11. W. V. Gilbert, T. A. Bell, C. Schaening, *Science*, in press, doi:10.1126/science.aad8711.

12. H. Shi, J. Wei, C. He, Where, When, and How: Context-Dependent Functions of RNA Methylation Writers, Readers, and Erasers. *Molecular Cell*. **74**, 640–650 (2019).

13. M. Courel, Y. Clément, C. Bossevain, D. Foretek, O. Vidal Cruchez, Z. Yi, M. Bénard, M.-N. Benassy, M. Kress, C. Vindry, M. Ernoult-Lange, C. Antoniewski, A. Morillon, P. Brest, A. Hubstenberger, H. Roest Crollius, N. Standart, D. Weil, GC content shapes mRNA storage and decay in human cells. *eLife*. **8**, e49708 (2019).

14. V. Akimov, I. Barrio-Hernandez, S. V. F. Hansen, P. Hallenborg, A.-K. Pedersen, D. B. Bekker-Jensen, M. Puglia, S. D. K. Christensen, J. T. Vanselow, M. M. Nielsen, I. Kratchmarova, C. D. Kelstrup, J. V. Olsen, B. Blagoev, UbiSite approach for comprehensive mapping of lysine and N-terminal ubiquitination sites. *Nature Structural & Molecular Biology*. **25**, 631–640 (2018).

15. T. F. Martinez, Q. Chu, C. Donaldson, D. Tan, M. N. Shokhirev, A. Saghatelian, Accurate annotation of human protein-coding small open reading frames. *Nature Chemical Biology*. **16**, 458–468 (2020).

16. G. Monaco, B. Lee, W. Xu, S. Mustafah, Y. Y. Hwang, C. Carré, N. Burdin, L. Visan, M. Ceccarelli, M. Poidinger, A. Zippelius, J. Pedro de Magalhães, A. Larbi, RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. *Cell Reports*. **26**, 1627-1640.e7 (2019).

17. V. Ranzani, G. Rossetti, I. Panzeri, A. Arrigoni, R. J. P. Bonnal, S. Curti, P. Gruarin, E. Provasi, E. Sugliano, M. Marconi, R. De Francesco, J. Geginat, B. Bodega, S. Abrignani, M. Pagani, The long intergenic noncoding RNA landscape of human lymphocytes highlights the regulation of T cell differentiation by linc-MAF-4. *Nature Immunology*. **16**, 318–325 (2015).

18. T. Chen, C. Guestrin, "XGBoost: A Scalable Tree Boosting System" in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, NY, USA, 2016; https://dl.acm.org/doi/10.1145/2939672.2939785), pp. 785–794.

19. V. Agarwal, J. Shendure, Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. *Cell Reports*. **31**, 107663 (2020).

20. S. Mitschka, C. Mayr, Endogenous p53 expression in human and mouse is not regulated by its 3′UTR. *eLife*. **10**, e65700 (2021).

21. E. L. V. Nostrand, P. Freese, G. A. Pratt, X. Wang, X. Wei, R. Xiao, S. M. Blue, J.-Y. Chen, N. A. L. Cody, D. Dominguez, S. Olson, B. Sundararaman, L. Zhan, C. Bazile, L. P. B. Bouvrette, J. Bergalet, M. O. Duff, K. E. Garcia, C. Gelboin-Burkhart, M. Hochman, N. J. Lambert, H. Li, M. P. McGurk, T. B. Nguyen, T. Palden, I. Rabano, S. Sathe, R. Stanton, A. Su, R. Wang, B. A. Yee, B. Zhou, A. L. Louie, S. Aigner, X.-D. Fu, E. Lécuyer, C. B. Burge, B. R. Graveley, G. W. Yeo, A large-scale binding and functional map of human RNA-binding proteins. *Nature*. **583**, 711–719 (2020).

22. J. C. Rieckmann, R. Geiger, D. Hornburg, T. Wolf, K. Kveler, D. Jarrossay, F. Sallusto, S. S. Shen-Orr, A. Lanzavecchia, M. Mann, F. Meissner, Social network architecture of human immune cells unveiled by quantitative proteomics. *Nature Immunology*. **18**, 583–593 (2017).

23. T. Geiger, A. Wehner, C. Schaab, J. Cox, M. Mann, *Molecular & Cellular Proteomics*, in press, doi:10.1074/mcp.M111.014050.

24. C. C. Thoreen, L. Chantranupong, H. R. Keys, T. Wang, N. S. Gray, D. M. Sabatini, A unifying model for mTORC1-mediated regulation of mRNA translation. *Nature*. **485**, 109–113 (2012).

25. R. Geiger, J. C. Rieckmann, T. Wolf, C. Basso, Y. Feng, T. Fuhrer, M. Kogadeeva, P. Picotti, F. Meissner, M. Mann, N. Zamboni, F. Sallusto, A. Lanzavecchia, L-Arginine Modulates T Cell Metabolism and Enhances Survival and Anti-tumor Activity. *Cell*. **167**, 829-842.e13 (2016).

26. R. Rak, M. Polonsky, I. Eizenberg-Magar, Y. Mo, Y. Sakaguchi, O. Mizrahi, A. Nachshon, S. Reich-Zeliger, N. Stern-Ginossar, O. Dahan, T. Suzuki, N. Friedman, Y. Pilpel, Dynamic changes in tRNA modifications and abundance during T cell activation. *Proceedings of the National Academy of Sciences*. **118** (2021), doi:10.1073/pnas.2106556118.

27. C. Gobet, B. D. Weger, J. Marquis, E. Martin, N. Neelagandan, F. Gachon, F. Naef, *Proceedings of the National Academy of Sciences*, in press, doi:10.1073/pnas.1918145117.

28. V. Presnyak, N. Alhusaini, Y. H. Chen, S. Martin, N. Morris, N. Kline, S. Olson, D. Weinberg, K. E. Baker, B. R. Graveley, J. Coller, Codon optimality is a major determinant of mRNA stability. *Cell*. **160**, 1111–1124 (2015).

29. L.-S. Zhang, C. Liu, H. Ma, Q. Dai, H.-L. Sun, G. Luo, Z. Zhang, L. Zhang, L. Hu, X. Dong, C. He, Transcriptome-wide Mapping of Internal N7-Methylguanosine Methylome in Mammalian mRNA. *Molecular Cell*. **74**, 1304-1316.e8 (2019).

ACKNOWLEDGMENTS

We thank B. Popovic, M. Messemaker, B. Kwee, T. Schumacher for helpful discussion. We thank B. P. and N.D. Zandhuis for critical reading of this manuscript.

Author contribution: Conceptualization, project administration, writing: BPN, MCW. Software, data curation, analysis, visualization: BPN, KB. Validation, investigation, methodology: BPN, APJ, KB, AG. Funding acquisition, supervision: MCW.

Competing interests: B.P.N. and M.C.W filed a patent application related to this work.

Funding: This research was supported by the Oncode Institute (MCW), and the European research council ERC-Printers 817533 (MCW).

Data and material availability: Code and scripts have been deposited on github (https://github.com/BenNicolet/SONAR-Learning-the-sequence-code-for-mRNA-and-protein-abundance). Models were deposited on Zenodo (https://zenodo.org/record/8263086). All expression data used in this study were previously published: RNA-sequencing data retrieved from (*7*, *15–17*) on the National Center for Biotechnology Information (NCBI)'s gene expression omnibus (GEO) and the European Nucleotide Archive (ENA); for mass spectrometry-based proteomics and protein synthesis rate, datasets were retrieved from (*7*, *22*, *23*). Reporter assay data were deposited on GEO (GSE240919). See details in Table S1. All data are available in the main text or the supplementary materials.

SUPPLEMENTARY MATERIALS:

Materials and Methods

Supplementary Text

Figs. S1 to S12

Tables S1

Data S1 to S7

FIGURE LEGENDS

Fig. 1: Modelling the sequence code of mRNA abundance

(**A**) SONAR was trained to learn the code for mRNA abundance using sequence features (SFs) found in the human transcriptome (see Methods). (**B-C**) mRNA abundance model performance on test sets (R^2, indicated in bar, with S.D.) in (B) human cells (n=88; 2-4 biological replicates; dashed line shows mean R^2), or (C) HEK293T cells (representative of 3; red line shows linear regression line; TPM: transcript per kilobase per million). (**D**) SFs needed to reach indicated prediction ceiling to mRNA models. (**E**) SFs contribution per mRNA region of origin (feature importance (a.u.)). Tn: naïve, Tcm: central-memory, Tem: effector memory, Teff: effector T cells; Treg: regulatory T cells; mem: memory; pDC: plasmacytoid, mDC: myeloid dendritic cells; mono.: monocytes; clas.: classical; inter.: intermediate. NK: natural killer cells.

Fig. 2: Modelling the sequence code of protein abundance

(**A**) SONAR protein abundance model performance (R^2, indicated in bar, with S.D.) on test sets in human cells (n=118; 3-4 biological replicates; dashed line shows mean R^2). (**B**) Prevalence of SF class within the top 1000 most predictive SFs. Percentages (of total) per SF type are indicated on the right. (**C**) Average feature importance of $CD4^+$ and $CD8^+$ naïve T cells protein models, with indicated color-coded SF types. (**D**) UMAP representation of feature importance usage per sample in protein abundance models of (A). (**E**) Protein abundance model performance for all cell types (left panel) trained on mRNA abundance data (RNA only), sequence features (SF only) or both (SF+RNA). Average performance per model type (right panel; n=22; average is indicated in bar; with S.D.). Temra: Tem CD45RA+ T cells; PTM: post-translational modification; Sig.: signal; mod.: modification.

Fig. 3: Sequence features capture the dynamics of mRNA and protein abundance in activated T cells

(**A**) Model performance on test sets of mRNA abundance, translation rate, and protein abundance data of $CD4^+$ T cell during activation (average of 3 donors). (**B**) Comparison of SFs contribution to models in (A). (**C**) Relation between m7G site occurrence with measured mRNA abundance, translation rate, and protein abundance (red line show the linear regression).

Fig. 4: SONAR informs on synthetic 3'UTRs to modulate protein expression.

(**A**) Schematic representation of the GFP reporter assay of 3'UTR motifs. (**B-C**) Log2 fold-change (LFC) enrichment of motifs found in GFP high (GFP^{hi}) or low (GFP^{low}) in (B) $CD4^+$ and $CD8^+$ T cells or (C) HEK293T cells. (**D-E**) SONAR feature importance in relation to (D) doublet or (E) triplet content of 5/6-mers found in 3'UTR derived from protein models (Fig. 2). (**F**) K-means clustering (k=4) of motifs based on LFC(GFP^{hi} – GFP^{lo}). (**G-H**) GFP expression levels normalized to control 3'UTR assessed by flow cytometry in indicated cell types for the (G) 4 indicated 3'UTR reporter, or (G) reporter containing 0 to 6 inclusions of CTTTCTT motifs. Color coding in (C-E) indicate the C+T or A+G nucleotide content of the dimer/trimer within motifs.

Figure 1 Nicolet et al.

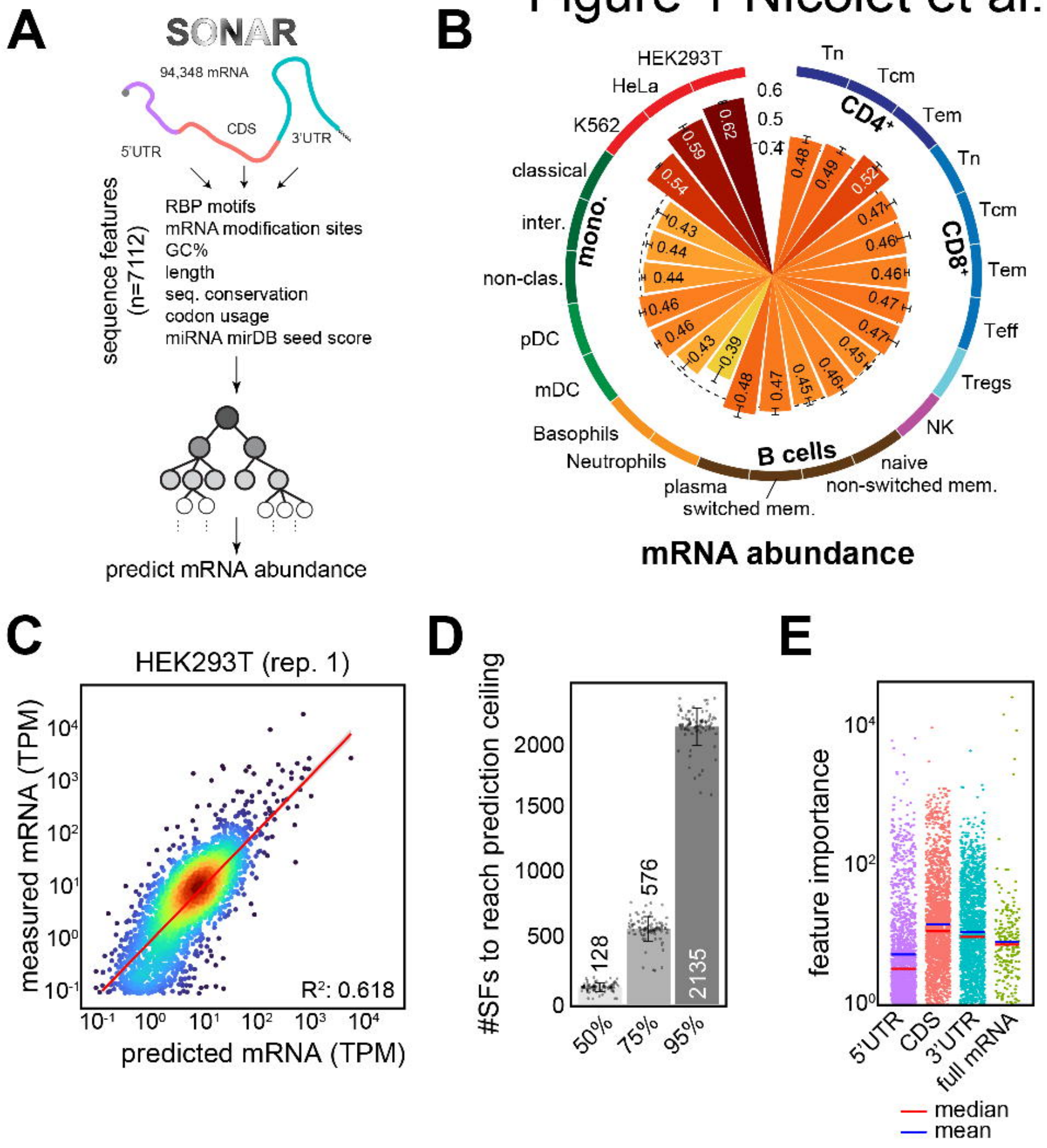

Fig. 2 Nicolet et al.

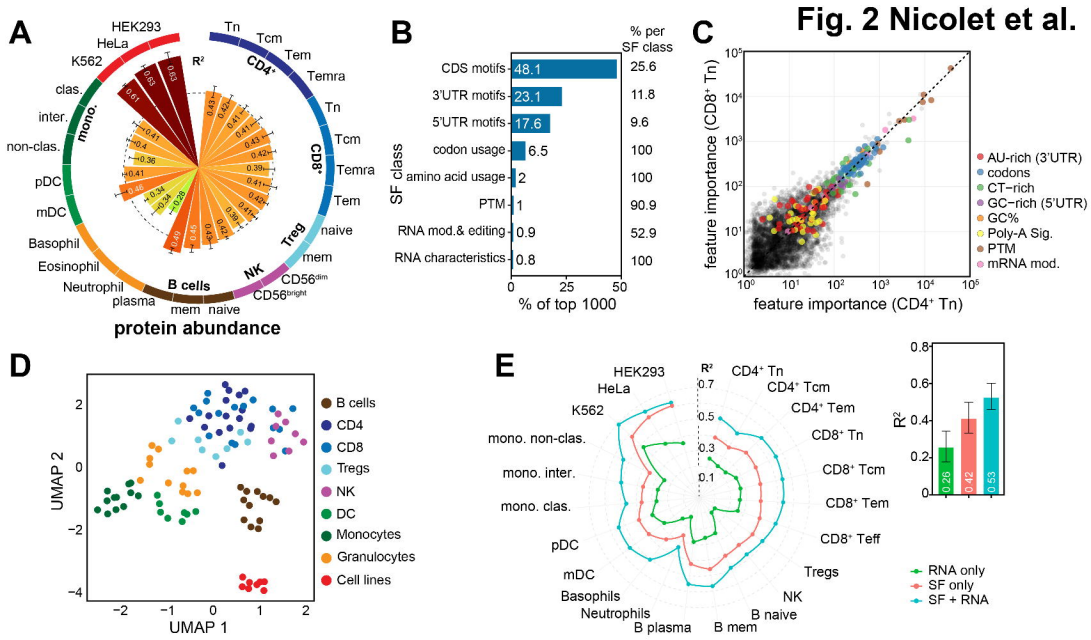

Figure 3 Nicolet et al.

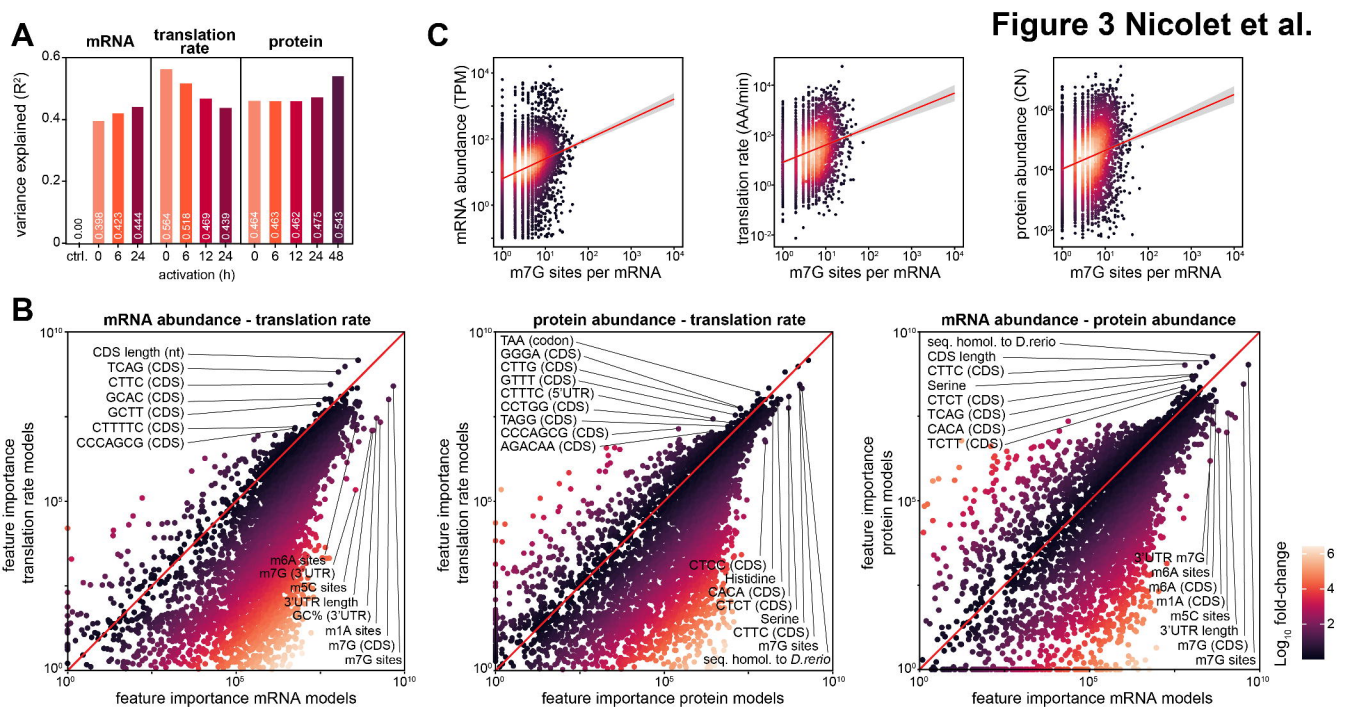
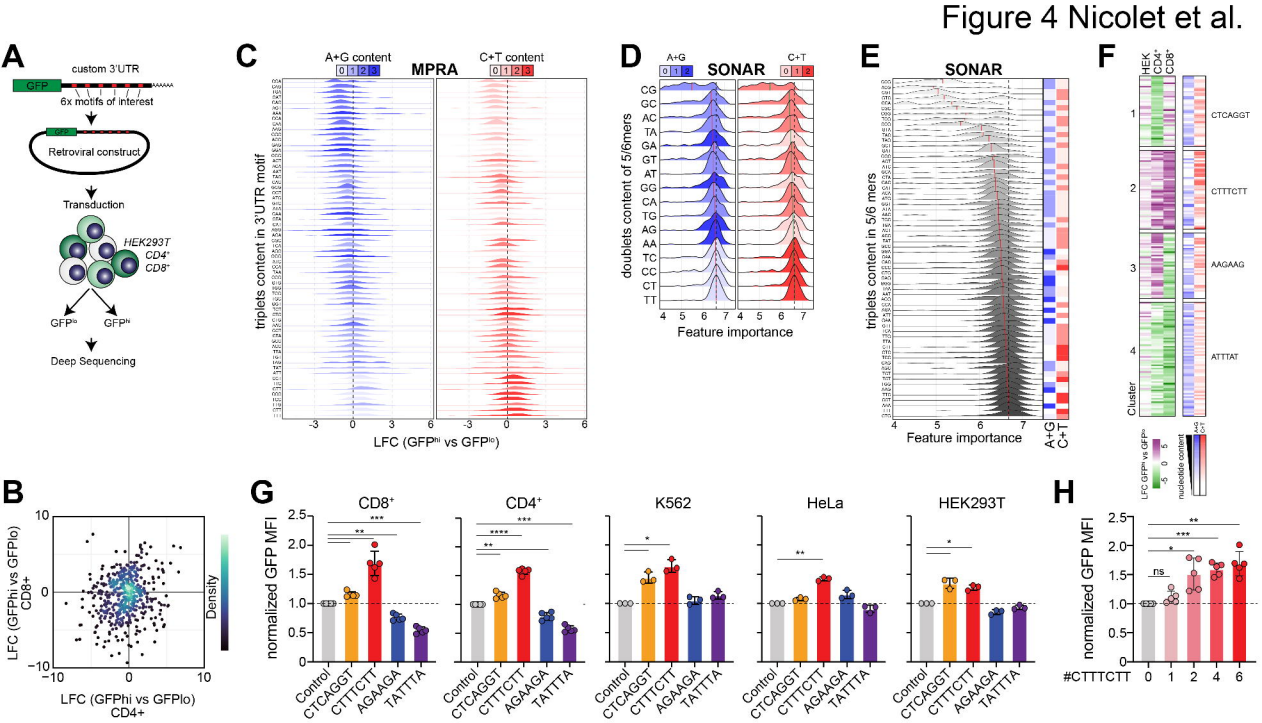
Figure 4 Nicolet et al.