On the prediction of protein abundance from RNA

Rasmus Magnusson[1]*, Olof Rundquist[1]*, Min Jung Kim[2], Sandra Hellberg[3], Chan Hyun Na[4], Mikael Benson[5], David Gomez-Cabrero[6], Ingrid Kockum[7], Jesper Tegnér[8,9,10], Fredrik Piehl[7], Maja Jagodic[7], Johan Mellergård[11], Claudio Altafini[12], Jan Ernerudh[13], Maria C. Jenmalm[3], Colm E. Nestor[3], Min-Sik Kim[14] and Mika Gustafsson[1]

[1]Bioinformatics, Department of Physics, Chemistry and Biology, Linköping University, Linköping, Sweden.

[2]Department of Applied Chemistry, College of Applied Sciences, Kyung Hee University, Yong-in 446-701, Republic of Korea.

[3]Department of Clinical and Experimental Medicine, Linköping University, Linköping, Sweden

[4]Department of Neurology, Institute for Cell Engineering, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[5]Centre for Personalized Medicine, Linköping University, Linköping, Sweden.

[6]Navarrabiomed, Complejo Hospitalario de Navarra, Universidad Pública de Navarra, IdiSNA, 31008 Pamplona, Spain

[7]Department of Clinical Neuroscience, Center for Molecular Medicine, Karolinska Institutet, 171 77, Stockholm, Sweden

[8]Biological and Environmental Sciences and Engineering Division, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955–6900, Saudi Arabia

[9]Unit of Computational Medicine, Department of Medicine, Solna, Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden.

[10]Science for Life Laboratory, Solna, Sweden.

[11]Department of Neurology, Linköping University, Linköping, Sweden

[12]Department of Automatic Control, Linköping University, Linköping, Sweden

[13]Department of Clinical Immunology and Transfusion Medicine and Department of Clinical and Experimental Medicine, Linköping University, Linköping, Sweden

[14]Department of New Biology, Daegu Gyeongbuk Institute of Science and Technology, Daegu 711-873, Republic of Korea

*These authors contributed equally to this work and should be regarded as shared first authors.

Correspondence: mika.gustafsson@liu.se

In eukaryotes, mRNA abundance is often a poor proxy for protein abundance[1-5]. Despite this, the majority of methods used to dissect function in mammalian biology[6] and for biomarker discovery in complex diseases[7] involve manipulation or measurement of mRNA. The discrepancy between mRNA and protein abundance is likely due to several factors, including differences in the rates of translation and degradation between proteins and cell-types[8], unequal contribution of individual splice variants to the production of a given protein[9] and cell-type specific differences in splice variant use[10]. Here we performed experimental and computational time-series analysis of RNA-seq and mass-spectrometry of three key immune cell-types in human and mice and constructed mathematical mixed time-delayed splice variant models to predict protein abundances. These models had median correlations to protein abundance measurements of 0.79-0.94, which is a significant increase from the previously reported 0.21 on human protein atlas data[1], and out-performed less complicated models without the usage of multiple splice variants and time-delay in cross-validation tests. We showed the importance of our models for biomarker discovery by re-analysing RNA-seq data from five different complex diseases, which led to the prediction of new disease proteins that were validated in multiple sclerosis. Our findings suggest that similar protein abundance models may be created for the most critical cell-types in the human body.

To understand the effect of splice-variant selection and translation rate on the relationship between RNA and protein abundance, we performed RNA-sequencing and mass-spectrometry proteomics of primary human naïve CD4$^+$ T helper (NT$_H$) cells at six time points during differentiation into T-helper type 1 (T$_H$1) cells (**Figure 1A, S1, S2**). T$_H$1 differentiation is an optimal model system to dissect the relationship between mRNA and protein as (i) primary human NT$_H$ cells can be isolated in high purity and large quantity from human blood (ii), all NT$_H$ cells are synchronised in the G$_1$ phase of the cell cycle, further reducing inter-cell heterogeneity[11,12] and (iii) changes in mRNA and associated protein abundance can be assayed over time[13]. Moreover, T helper cells are important regulators of immunity and thereby associated with many complex diseases, and T$_H$1 differentiation itself is pathogenetically relevant in several diseases[14,15].

Interestingly, although the majority of genes showed a significant positive correlation between mRNA level and protein level (n=407, expected 123 out of 4920 proteins, binomial test $P<10^{-93}$) during T$_H$1 cell differentiation, a significant fraction of negatively correlated genes was also observed (n=205, expected 123, $P<10^{-11}$) (**Figure 1B, Data S1**). Analysing the correlation of individual splice variants of each gene revealed the presence of both positively and negatively correlating transcript variants for the same genes (binomial test for enrichment of significant negative correlation $P<1.3 \times 10^{-3}$, odds ratio= 1.48). For example, the known T-helper cell associated genes, *IL7R* and *STX12*[16], contained many splice variants, of which several were positively and negatively correlated to their corresponding protein levels (**Figure 1C**). Given the large variation in correlation between splice-variants of a given gene and its corresponding protein, we constructed a simple mRNA-protein model, in which protein expression was defined as a linear combination of the splice variants of a gene, with a time-delay accounting for the dynamical effect induced by post-transcriptional processes and protein synthesis (**Figure 1D**). This simple dynamical model resulted in a gene-protein correlation of $rho_{\mathrm{TH1}} = 0.86$ (**Figure 1E)**, far in excess of previously reported gene-protein prediction models in mammals[1,2,17-19]. A strong correlation between gene and protein levels was also observed when the model was trained on published mRNA-protein datasets from human regulatory T cells ($rho_{\mathrm{TREG}} = 0.79$) and mouse B cells ($rho_{\mathrm{Bcell}} = 0.94$) (**Figure S3**). Importantly, our model out-performed models using only the best correlating splice variant of each gene ($rho_{\mathrm{TH1}} = 0.71$, $rho_{\mathrm{TREG}} = 0.44$, $rho_{\mathrm{Bcell}} = 0.52$), or models using multiple transcripts but without a time delay ($rho_{\mathrm{TH1}} = 0.74$, $rho_{\mathrm{TREG}} = 0.69$, $rho_{\mathrm{Bcell}} = 0.45$) (**Figure 1F**). The median duration of optimal time-delays between splice-variants and proteins was 8h 17 min, 6h 18 min and 8h 49 min for $T_{\mathrm{H}}1$, T_{REG} and murine B cells, respectively. These values are descriptions on the explanatory power of the different models on the same data as it was trained to. Cross-validation confirmed that our models could do out-of-sample prediction significantly better than gene expression based models of protein abundance (binomial test; $P_{\mathrm{TH1}}= 10^{-152}$, $P_{\mathrm{TREG}}= 10^{-247}$, $P_{\mathrm{mice\ B}}= 10^{-59}$), and simpler splice-variant models without time-delays ($P_{\mathrm{TH1}}=10^{-1459}$, $P_{\mathrm{TREG}}= 10^{-8}$, $P_{\mathrm{mice\ B}}= 5\mathrm{x}\ 10^{-4}$, Fig. 1F, Fig. S4). To evaluate mRNA-protein associations in steady state across tissues, we used data from the human protein atlas[20]. We found that this only resulted in marginal improvements in correlation with respect to that previously reported in the literature[1] ($\mathrm{rho}_{\mathrm{ProtAtlas}}= 0.27$), see (**Figure S3). This lack of correlation can be explained by the lack of** dynamic data, and by the presence of different cell types. In further support of cell type specificity, we found only marginal correlations (*rho* = 0.09) when comparing the correlation coefficients of our two T-cell data-sets of T_{H}1and Treg cells. Thus, a common unifying model for many cell-types remains a challenge (**Data S1**). In summary, we have revealed that using a simple linear model of mRNA splice variants and time delay, we could predict protein abundances accurately.

To test the clinical usefulness of our results, we applied our model to available RNA-sequencing data-sets involving T_H-associated diseases. Using data derived from human total $CD4^+$ T cells in asthma, allergic rhinitis, obesity-induced asthma, pro-lymphocytic leukaemia, and multiple sclerosis (MS), we found for each disease a higher fraction of nominally differentially expressed proteins than standard differential expression analysis (**Figure 2A**). For MS, the T_H1 model resulted in the highest fraction and 21 proteins were predicted as differentially expressed at $FDR<0.05$, whereof three (Annexin A1, sCD40L and sCD27) were annotated as extracellular according to gene ontology. To validate these predictions, we analysed if cerebrospinal fluid (CSF) levels of these proteins related to clinical outcome and immunomodulatory treatment in two independent cohorts, namely newly diagnosed MS (clinically isolated syndrome (CIS) and relapsing/remitting MS, n=41) *vs* healthy controls (HC, n=23), and response to Natalizumab treatment in relapsing remitting MS patients (see Methods, n=16). In both cohorts, only sCD27 was present at a detectable level. Analysis of all patients (n=57) *vs* HC (n=23) showed high separation (AUC=0.88, non-parametric $P=3.0 \times 10^{-8}$, **Figure 2B**), and treatment with Natalizumab reduced the sCD27 levels by 34% ($P = 4.9 \times 10^{-4}$). Lastly, we tested the prognostic value of sCD27 and found that the baseline levels in the newly diagnosed MS patients were able to predict disease activity after four years follow up (AUC= 0.87, $P=1.2 \times 10^{-3}$, Figure 2B), which was stronger than that of all our previously reported 14 biomarkers[21]. Taken together, the high correlation between predicted and measured protein levels from mass-spectrometry, the increased fraction of differentially expressed predicted proteins and successful biomarker validation, show the relevance of our $CD4^+$ T cell models for discovery of protein biomarkers in T_H-cell mediated diseases from RNA-seq data alone.

In conclusion, we have shown that simple mRNA-protein models, in which the protein expression is defined as a linear combination of the splice variants of a gene, with a time-delay accounting for the dynamical effect induced by post-transcriptional processes and protein synthesis, can profoundly improve our ability to predict protein abundance from mRNA abundance. We expect this modelling strategy to be generally applicable to other cellular differentiation systems, such as embryonic stem cell differentiation, and to be increasingly useful for understanding basic biology and identification of new biomarkers as more RNA-seq and proteomic data sets become publicly available.

Figure 1

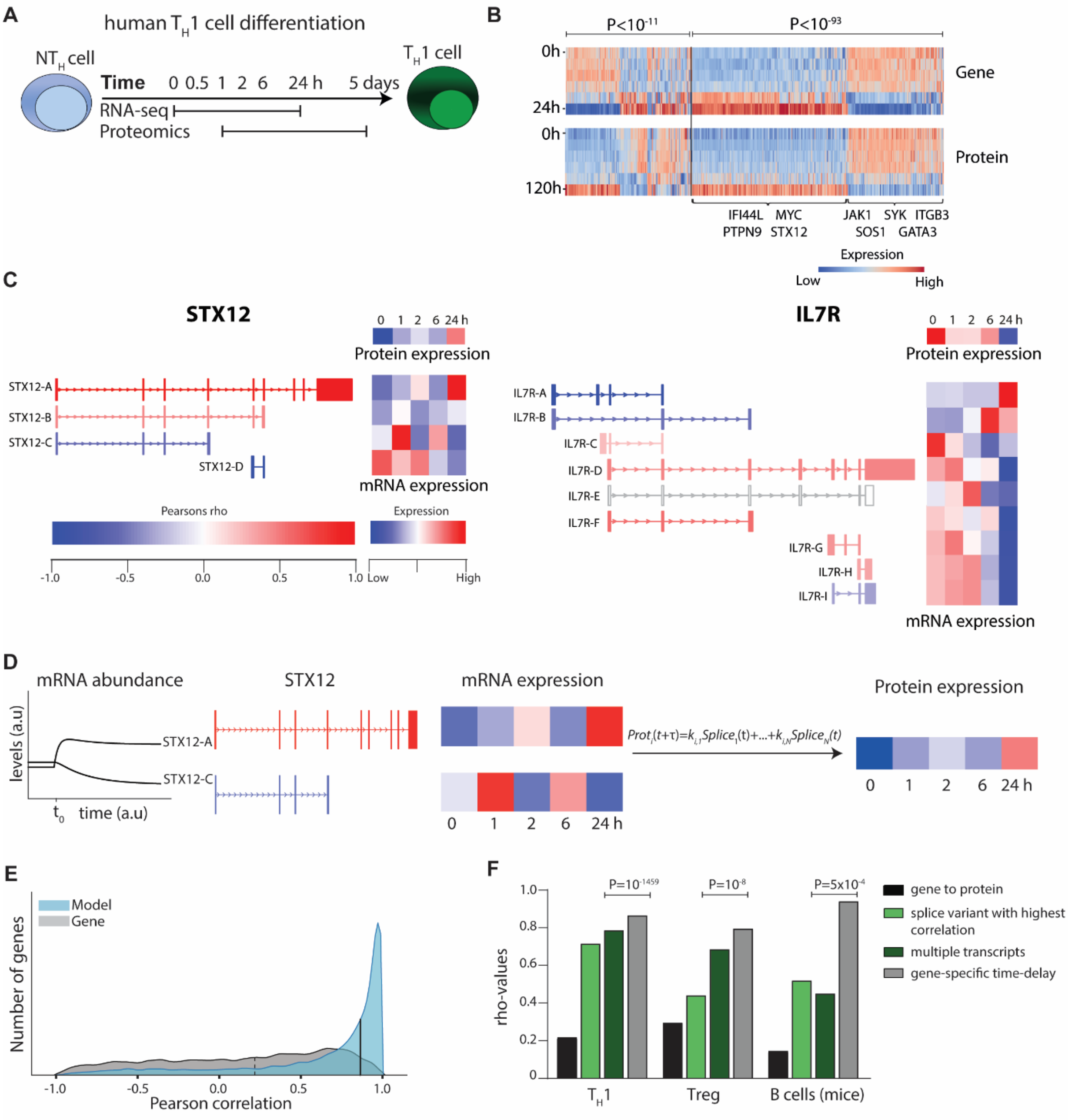

Fig 1. Predictive protein abundance models derived from splice variants showed high correlation to experiments in multiple cell types. (A) Experimental design. (B) Heat map of transcript and protein abundance dynamics in genes that show significant negative (left) and

positive (right) correlations. (C) Transcript splice variants of *STX12* (left) and *IL7R* (right) were both significantly negatively and positively correlated with protein levels. (D) Illustration of the modelling procedure for resolving the poor correlation, using STX12 as an example. (E) Gene/protein correlations in Th1 differentiation. In the histogram, in grey a single value lumping all splice variants of a transcript (as in *e.g.* Fortelny *et al.*[1]) is used to quantify mRNA abundance (median: dashed line at 0.21), while in the blue histogram our time-delayed multiple splice variant based model is used (median at 0.86). Only cross-validated protein predictions (PPs) are shown for the 3410 out of 4920 proteins for which the null-model could be rejected. (F) Median correlation coefficients (rho) for different mathematical protein prediction models derived from RNA with increasing protein abundance correlations. P-values are derived from predictions using leave-one-out cross-validation.

Figure 2

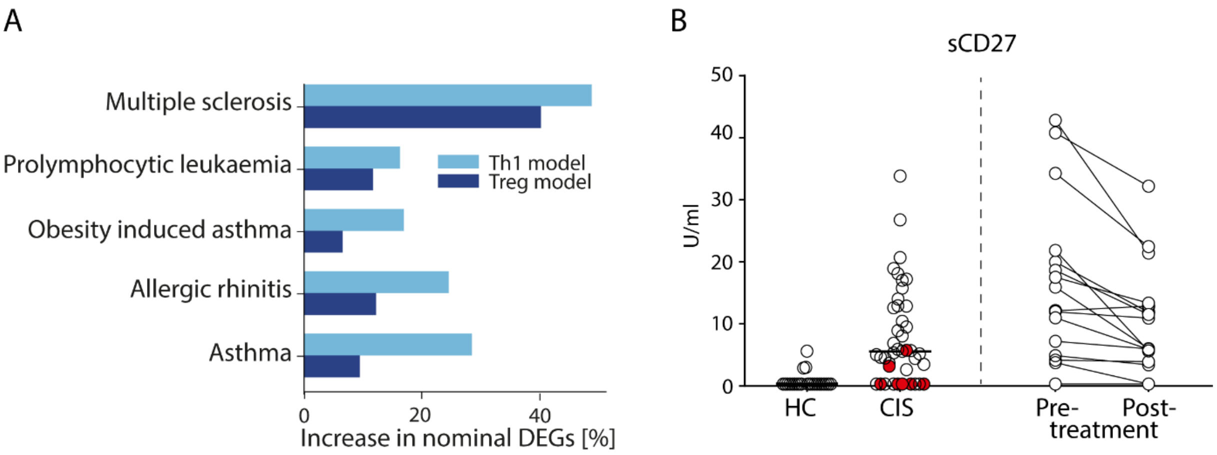

Fig 2. Proteins models led to the discovery of new potential biomarkers of complex diseases that were validated in multiple sclerosis (MS). (A) Differential predicted protein (PP) analysis of five diseases using the Th1 (light blue) and Treg (dark blue) models showed higher fraction of nominally significant genes than normal gene tests. (B) Validation of multiple sclerosis PP in cerebrospinal fluid (CSF) from early MS (clinically isolated syndrome (CIS)) *vs* healthy controls (HC) and pre *vs* post one-year treatment with Natalizumab. Red filled circles represent patients with no evidence of disease activity) at four years follow up and the remaining are coded by unfilled circles.

Material and methods

Experimental protocol

Isolation of CD4$^+$ T helper (TH) cells and TH1 polarization: Peripheral blood mononuclear cells (PBMC) were isolated from blood donor derived buffy coats through gradient centrifugation (Lymphoprep, Axis shields diagnostics, Dundee, Scotland). Naive CD45RA$^+$ CD4$^+$ T cells were subsequently isolated with magnetic bead separation using the "Naive CD4$^+$ T Cell Isolation Kit II, human" (Miltenyi Biotec, Bergisch Gladbach, Germany). The cells were then activated and polarized towards T_H1 using Dynabeads™ Human T-Activator CD3/CD28 (Dynal AS, Lillestøm, Norway), recombinant human IL-12p70, recombinant human IL-2 and anti-IL-4 antibodies (clone MAB204) (all three from, Bio-Techne, Minneapolis, USA), in RPMI 1640 media (Gibco, Paisley, United Kingdom). A portion of T-cells used for RNA and protein isolation was obtained at baseline and after 0.5 h, 1 h, 2 h, 6 h, 24 h and 5 days. The cells were washed twice in PBS, snap frozen in a dry ice ethanol bath and stored at -80°C until use. During the protein and RNA extractions, multiple samples were pooled from twelve different individuals to reach the necessary amount of material for the subsequent analysis steps.

Mass-spec proteomics: The cells were lysed by sonication. Proteins were digested with trypsin through an in-solution digestion protocol and desalted peptides were labelled with 6-plex TMT reagents (Thermofisher Scientific, Massachusetts, USA). Then, the labelled peptides were mixed and separated using high-pH reverse-phase liquid chromatography, each fraction of which was analysed on an Orbitrap Fusion Lumos Tribrid mass spectrometer (Thermofisher Scientific, *Massachusetts*, USA). The tandem mass spectrometry data were analysed using MaxQuant (v 1.6.0.1). Detailed experimental procedures are provided in the Supplementary information.

RNA-seq: RNA was isolated using a ZR-Duet DNA/RNA kit (Zymo Research, Irvine, USA) and stored at -80°C. RNA library preparation and the subsequent RNA-sequencing were carried out by the Beijing Genomics Institute (https://www.bgi.com/global/). Library preparation was performed using the TruSeq RNA Library Prep Kit v2 (Illumina, San Diego, USA). Each sample was sequenced to the depth of 40 million reads per samples (Fig 1A) with pair end sequencing and a read length of 100bp on an Illumina 2500 instrument.

Bioinformatics: All RNA-seq data were processed similarly using the following pipeline. Sample qualities were assessed with fastQC and the mRNA reads were subsequently aligned using STAR[22] to the "Homo_sapiens.GRCh37.75.dna.primary_assembly.fa" from Ensemble. The resulting read alignment bam files were assembled into transcripts with StringTie[23] using the GRCh37.75 gtf annotation from Ensemble. To evaluate mRNA to protein relationship, mRNA reads were mapped to the mass spectrometry signal of protein abundance using the Homo.sapiens and Mus.musculus package in R[24]. Correlations were calculated using Pearson correlations across gene expressions, i.e. one coefficient per gene.

Splice variant model construction:

We trained a linear regression model with a LASSO constraint model using the python package Sklearn[25]. We implemented a time-delay shift which was defined using a grid of 195 values, exponentially increasing from 0 to 24h. Linear interpolation was used to estimate the corresponding protein values. Specifically, for different τ we solve the following:

$$\min\{\frac{1}{N}\|Y(t+\tau)-\beta X(t)\|_2+\lambda\|\beta\|_1\}$$

Here, the time series of one protein is denoted by the vector Y, and the corresponding time series of the splice variants are denoted by the matrix X. The λ term was chosen to minimize the prediction error of a leave-one-out cross validation. For each τ, we did a leave-one-out cross validation on top of the one used for determining the λ parameter. Next, the τ yielding the lowest error on an outer cross validation error was selected. Testing for τ=0 was used when studying the information gain from combining splice variants. Pipe-line and code available from https://gitlab.com/Gustafsson-lab/IMUNA-an-integrated-multilevel-Th1-analysis .

Disease prediction

Disease relevance of the splice variant models was tested by re-analysis of deep RNA-sequenced case control material of samples containing total CD4$^+$ T-cells, *i.e.* CD4$^+$ T-cells with all its sub-types. We found T-cell prolymphocytic leukemia (T-PLL, GSE100882), asthma in obese children (GSE86430), and allergic rhinitis/asthma (GSE75011) studies through a Gene Expression Omnibus (GEO) repository search and multiple sclerosis (MS) through collaboration[26]. For each of the studies' datasets, we used the T$_H$1 and Treg derived models on how to combine mRNA splice variants to predict protein abundance. The resulting sets of predicted protein levels were tested for differential expression between patients and controls using a non-parametric Kruskal-Wallis test. We also applied Kruskal-Wallis tests to the individual splice variants that were used by the models. We assessed model effects by measuring the increase in nominally differential expression from model predictions compared to ingoing splice variants into the model.

Protein validation

Three of the DEPPs (Annexin A1, sCD40L and sCD27) were measured in cerebrospinal fluid (CSF) from two different cohorts, one with of 41 patients with newly diagnosed MS and 23 healthy matched controls (Supplemental table) and a second with 16 patients with relapsing remitting MS before and after one year of treatment with Natalizumab (Supplemental table). Quantification of Annexin A1 was performed using Human Annexin A1 ELISA kit (Abcam, Cambridge, United Kingdom) and sCD27 was measured using the Human Instant ELISA^{TM} kit (Thermo Fischer Scientific, Waltham, MA, USA), according to the instructions provided by the manufactures. Multiplex Bead Technology (MILLIPLEX® MAP Kit, Cat. #: HCYTOMAG-60K-01, Merck Millipore, Burlington, MA, USA) was used to measure soluble CD40L, according to the manufacturer's description. A more detailed description is available in the Supplementary information.

Ethics statement

The study was approved by the Regional Ethics Committee in Linköping, Sweden (Dnr M180-07 and M2-09).

Acknowledgments

Funding: This work was supported by the Swedish Cancer Society grants (CAN 2017/625), East Gothia Regional Funding, Åke Wiberg foundation, Neuro Sweden, the Swedish Research Council grants 2015-02575, 2015-03495, 2015-03807, 2016-07108, 2018-02776, National Research foundation of Korea, and Swedish foundation for strategic research.

Author contribution

MG initiated and supervised the study. RM, and OR performed bioinformatics analyses, which were led by MG, CA, JT, and DGC. OR, MJK, CHN, SH performed experimental work on T-cell differentiation, which were supervised by CEN, MCJ, JE, MB, and MSK. SH, FP, JM performed sample collection and clinical lab work, which were led by IK, MJ, and JE. All authors contributed to and approved the final draft for publication.

Reference list

1 Fortelny, N., Overall, C. M., Pavlidis, P. & Freue, G. V. C. Can we predict protein from mRNA levels? *Nature* **547**, E19-E20, doi:10.1038/nature22293 (2017).

2 Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet* **13**, 227-232, doi:10.1038/nrg3185 (2012).

3 de Sousa Abreu, R., Penalva, L. O., Marcotte, E. M. & Vogel, C. Global signatures of protein and mRNA expression levels. *Mol. Biosyst.* **5**, 1512-1526, doi:10.1039/b908315d (2009).

4 Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**, 535-550, doi:10.1016/j.cell.2016.03.014 (2016).

5 Li, J. J. & Biggin, M. D. Gene expression. Statistics requantitates the central dogma. *Science* **347**, 1066-1067, doi:10.1126/science.aaa8332 (2015).

6 Maier, T., Guell, M. & Serrano, L. Correlation of mRNA and protein in complex biological samples. *FEBS Lett.* **583**, 3966-3973, doi:10.1016/j.febslet.2009.10.036 (2009).

7 Hellberg, S. *et al.* Dynamic Response Genes in CD4+ T Cells Reveal a Network of Interactive Proteins that Classifies Disease Activity in Multiple Sclerosis. *Cell reports* **16**, 2928-2939, doi:10.1016/j.celrep.2016.08.036 (2016).

8 Wethmar, K., Smink, J. J. & Leutz, A. Upstream open reading frames: molecular switches in (patho)physiology. *Bioessays* **32**, 885-893, doi:10.1002/bies.201000037 (2010).

9 Floor, S. N. & Doudna, J. A. Tunable protein synthesis by transcript isoforms in human cells. *Elife* **5**, doi:10.7554/eLife.10921 (2016).

10 Barbosa-Morais, N. L. *et al.* The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**, 1587-1593, doi:10.1126/science.1230612 (2012).

11 Sprent, J., Cho, J. H., Boyman, O. & Surh, C. D. T cell homeostasis. *Immunol. Cell Biol.* **86**, 312-319, doi:10.1038/icb.2008.12 (2008).

12 Sprent, J. & Tough, D. F. Lymphocyte life-span and memory. *Science* **265**, 1395-1400 (1994).

13 Schmidt, A. *et al.* Time-resolved transcriptome and proteome landscape of human regulatory T cell (Treg) differentiation reveals novel regulators of FOXP3. *BMC Biol.* **16**, 47, doi:10.1186/s12915-018-0518-3 (2018).

14 DuPage, M. & Bluestone, J. A. Harnessing the plasticity of CD4(+) T cells to treat immune-mediated disease. *Nature reviews* **16**, 149-163, doi:10.1038/nri.2015.18 (2016).

15 Nestor, C. E. *et al.* 5-Hydroxymethylcytosine Remodeling Precedes Lineage Specification during Differentiation of Human CD4(+) T Cells. *Cell reports* **16**, 559-570, doi:10.1016/j.celrep.2016.05.091 (2016).

16 Kanduri, K. *et al.* Identification of global regulators of T-helper cell lineage specification. *Genome medicine* **7**, 122, doi:10.1186/s13073-015-0237-0 (2015).

17 Gry, M. *et al.* Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC Genomics* **10**, 365, doi:10.1186/1471-2164-10-365 (2009).

18 Lundberg, E. *et al.* Defining the transcriptome and proteome in three functionally different human cell lines. *Mol. Syst. Biol.* **6**, 450, doi:10.1038/msb.2010.106 (2010).

19 Sharma, K. *et al.* Cell type- and brain region-resolved mouse brain proteome. *Nat Neurosci* **18**, 1819-1831, doi:10.1038/nn.4160 (2015).

20 Uhlen, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419, doi:10.1126/science.1260419 (2015).

21 Håkansson, I. *et al.* Neurofilament levels, disease activity and brain volume during follow-up in multiple sclerosis. *J Neuroinflammation* **15**, 209, doi:10.1186/s12974-018-1249-7 (2018).

22 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

23 Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290-295, doi:10.1038/nbt.3122 (2015).

24 Team BC. Homo.sapiens: Annotation package for the Homo.sapiens object. R package version 1.3.1. (2015).

25 Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *JMLR* **12**, 2825−2830 (2011).

26 James, T. *et al.* Impact of genetic risk loci for multiple sclerosis on expression of proximal genes in patients. *Human molecular genetics* **27**, 912-928, doi:10.1093/hmg/ddy001 (2018).