A validated strategy to infer protein biomarkers from RNA-Seq by combining multiple mRNA splice variants and time-delay

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 Personalised 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 Institute, 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 and maria.jenmalm@liu.se

Abstract

Background

Profiling of mRNA expression is an important method to identify biomarkers but complicated by limited correlations between mRNA expression and protein abundance. We hypothesised that these correlations could be improved by mathematical models based on measuring splice variants and time delay in protein translation.

Methods

We characterised time-series of primary human naïve $CD4^+$ T cells during early T-helper type 1 differentiation with RNA-sequencing and mass-spectrometry proteomics. We then performed computational time-series analysis in this system and in two other key human and murine immune cell types. Linear mathematical mixed time-delayed splice variant models were used to predict protein abundances, and the models were validated using out-of-sample predictions. Lastly, we re-analysed RNA-Seq datasets to evaluate biomarker discovery in five T-cell associated diseases, validating the findings for multiple sclerosis (MS) and asthma.

Results

The new models demonstrated median correlations of mRNA-to-protein abundance of 0.79-0.94, significantly out-performing models not including the usage of multiple splice variants and time-delays, as shown in cross-validation tests. Our mathematical models provided more differentially expressed proteins between patients and controls in all five diseases. Moreover, analysis of these proteins in asthma and MS supported their relevance. One marker, sCD27, was clinically validated in MS using two independent cohorts, for treatment response and prognosis.

Conclusion

Our splice variant and time-delay models substantially improved the prediction of protein abundance from mRNA data in three immune cell-types. The models provided valuable biomarker candidates, which were validated in clinical studies of MS and asthma. We propose that our strategy is generally applicable for biomarker discovery.

Introduction

A key problem in genome medicine is to find reliable disease biomarkers and therapeutical targets. An important reason is that common diseases involve thousands of proteins across multiple cell types. Proteins are regarded as optimal biomarkers as they are the main drivers of the crucial functions necessary for life, and thus directly connected to patho-physiological processes[1]. Furthermore, many proteins can be readily measured in biological fluids. However, proteome-wide analyses are difficult to perform in clinical studies due to the large quantities of material needed. On the other hand, gene expression profiling can be performed using a range of techniques, such as microarrays or RNA-sequencing. Another advantage of using mRNA expression as a core vehicle for biomarker discovery is that mRNA profiling can be performed even if only samples of limited amount, like biopsies, are available.

Combinations of mRNAs can have high diagnostic efficacy in multiple diseases[2, 3]. An ideal solution could therefore be to perform mRNA profiling to identify protein biomarkers that are needed for diagnosing and subtyping of diseases, as well for the personalisation and monitoring of treatments. However, this approach is complicated by the low correlation between mRNA and protein expression[4-7], which can be tackled with different strategies[8, 9]. The discrepancy between mRNA and protein abundance is due to several factors, including but not limited to differences in the rates of translation and degradation between proteins and cell-types[10]. Moreover, the data resolution of mRNA splice variants and protein isoforms further complicates such analyses, as in the cases of unequal contribution of individual splice variants to the production of a given protein[11], and cell-type specific differences in splice variant use[12].

Thus, the inability to predict protein abundance from mRNA abundance represents a major limitation in biomarker discovery. To this end, we developed a novel method to infer protein levels from mRNA expression data. Our procedure was derived by experimentally analysing early human T helper 1 (T_H1) differentiation and constructing a machine learning modelling approach for time-series RNA-Seq and proteomics data from a dynamical perturbation of the cell-type of interest. T_H differentiation is an optimal model system to dissect the relationship between mRNA and protein as (i) primary human naïve T_H (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[13] and (iii) easy access to large quantities of material allows changes in mRNA and associated protein abundance to be assayed over time[14]. Moreover, T_H cells are important regulators of immunity and thereby

associated with many complex diseases, and T_H1 differentiation itself is pathogenetically relevant in several diseases[15]. The utilised models were based on a time-delayed linear model between mRNA splice-variants of the same gene and protein levels. We generalised the model by applying it onto recent data from human regulatory T (T_{reg}) cell and murine B cell differentiation. By combining the strength of time-series analysis and RNA-sequencing, we were able to increase median mRNA-protein correlations significantly from the initial 0.21 to 0.86. Next, we showed the potential clinical usefulness of our derived models by detecting potential biomarkers in five complex diseases. This application revealed significantly more predicted biomarkers than by using off-the-shelf methods for RNA-Seq data analysis only. Analysis of these predicted proteins in asthma and MS supported their biological relevance. Finally, we validated one of the predicted biomarkers using two independent multiple sclerosis cohorts, which showed a remarkably better stratification between patients and controls than any of our previously reported protein biomarkers. The application of our approach to multiple different cell types, species and diseases shows its general applicability to increase the power of RNA-Seq based studies for biomarker discovery.

Results

A significant portion of T-cell genes showed diverse correlations between RNA splice variants and proteins

In order to generate accurate mRNA and protein models, taking into account the major factors of time-delay and splice variant usage, we first developed a model analysing early T-helper type 1 (T_H1) differentiation. This was done by performing time-series RNA-sequencing and mass-spectrometry proteomics of primary human NT_H cells (**Figure 1A, S1, S2**). RNA-seq (> 40x 10^6 reads per sample) and proteome profiling was performed to detect differentially expressed mRNA splice variants and proteins at six time points from 30 minutes to five days of T_H1 differentiation (**Figure 1A, S1, S2**). This approach detected 6909 proteins, of which 4920 could be mapped to genes expressed in the RNA-Seq data. As expected, a significant fraction of the genes showed a significant positive correlation between mRNA and protein levels (n=407, expected 123 out of 4920 proteins, binomial test $P<10^{-93}$) during T_H1 cell differentiation. Interestingly, a significant fraction of negatively correlated genes was also observed (n=205, expected 123, $P<10^{-11}$) (**Figure 1B, Table S1**). Remarkably, the overall median Pearson correlation (rho) between mRNA and protein was only 0.21. We hypothesised that this could depend on variable correlations between mRNA splice variants of each gene and the protein it encoded. Indeed, we found both positive and negative correlations between splice variants and their corresponding proteins (binomial test for enrichment of significant negative correlation $P<1.3 \times 10^{-3}$, odds ratio= 1.48). For example, the known T_H cell associated genes, *IL7R* and *STX12*[16] contained multiple splice variants, of which several were positively or negatively correlated to their corresponding protein levels (**Figure 1C**). Given the large variation in correlation between different splice-variants of a given gene and its corresponding protein, we proceeded to construct predictive splice-variant models of protein abundance.

A linear model combining the expressions of multiple splice variant transcripts showed substantially stronger correlations with protein abundance than individual transcripts

In order to construct generally applicable and predictive mRNA-to-protein models, we applied a simple linear relation between the protein abundance of a gene and its associated mRNA splice-variants. Furthermore, we allowed for different translation times for each gene. Firstly, we used a cross-validated L1 penalised linear regression model to favour simple models using single splices without any time-delays (Methods, **Figure 1D**). The rationale for

the L1 penalty was to effectively remove splice variants that carry little or no predictive power over protein abundance. This simple model resulted in a median gene-protein correlation of rho_{TH1} = 0.86 (**Figure 2A)**, far in excess of previously reported gene-protein prediction models in mammals[5, 7, 10, 11]. Likewise, we also trained similar models for two existing mRNA-protein time-series datasets with similar results, that is from human T_{REG} cells[14](rho_{TREG} = 0.79) and mouse B cells (GSE75417) (rho_{Bcell} = 0.94) (**Figure 2A**). In order to test whether the increase in correlation was due to the incorporation of negatively correlating splice variants, multiple transcripts, or time-delay we also constructed such models without each of these effects. Importantly, our model out-performed models with one splice variant for each gene (rho_{TH1} = 0.71, rho_{TREG} = 0.44, rho_{Bcell} = 0.52), and models using multiple transcripts but without a time delay (rho_{TH1} = 0.74, rho_{TREG} = 0.69, rho_{Bcell} = 0.45) (**Figure 2B-C**), thus demonstrating that both multiple dynamical splice variants and time delay are needed for optimal performance. In order to define the optimal time-delays between splice-variants and proteins, we analysed the time delay distributions and found it to have a mean of 8h 17 min, 6h 18 min and 8h 49 min for T_H1, T_{REG} and murine B cells, respectively. The detailed parameters of our models are fully displayed in Table S1.Next, by using cross-validation we confirmed that our models could do out-of-sample prediction significantly better than gene expression-based models of protein abundance (binomial test; $P_{TH1}= 10^{-152}$, $P_{TREG}= 10^{-247}$, $P_{mice\ B}= 10^{-59}$), and better than static splice-variant models which did not include time-delays ($P_{TH1}=10^{-1459}$, $P_{TREG}= 10^{-8}$, $P_{mice\ B}= 5x\ 10^{-4}$, Fig. 2B). To evaluate mRNA-protein associations in steady state across tissues, we used mRNA expression data from the human protein atlas[17]. We found only marginal improvements by using splicing information in the multi-tissue models with respect to what had previously reported in the literature[5]($rho_{ProtAtlas}$= 0.27, see **Figure S3)**. This lack of correlation may be explained by the lack of dynamic data, and by the presence of different cell types, and we speculate that differences in splice variant specificity between tissues effectively hinders this type of models. 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 datasets of $T_H1$and T_{REG} cells. Thus, a common unifying model for many cell-types remains a challenge (**Table S1**). In summary, we have revealed that by using a simple linear model of mRNA splice variants and time delay, we could predict protein abundances accurately.

Applying the model to clinical datasets revealed potential biomarkers which were validated in multiple sclerosis and asthma

Lastly, we aimed to test the potential usefulness of our derived models for the identification of protein biomarkers by applying them on available RNA-Seq datasets from human total $CD4^+$ T cells. We found data-sets for five different diseases[18-21]; asthma, allergic rhinitis, obesity-induced asthma, pro-lymphocytic leukaemia, and multiple sclerosis (MS), as well as corresponding controls. Because our models correlated well to protein abundances, we hypothesised that differential expression tests using the predicted proteins between patients and controls to be more sensitive than testing directly on the mRNA expression for all splice variants individually. Indeed, we observed that the fraction of nominally differentially expressed genes was higher than using an individual differential expression analysis for all ten comparisons (binomial $P< 9.8 \times 10^{-4}$) (**Figure 3A**). Moreover, we consistently observed a higher enrichment for the T_H1 model compared to the T_{REG} model ($P<0.03$), with the highest enrichments in MS and asthma. We therefore proceeded to use our T_H1 model on MS and asthma.

For MS, we found 20 genes with FDR<0.05, of which none could be found by testing for differential expression on the mRNA expression data directly (**Table S2**). Interestingly, eight of the 20 proteins had previously been associated with MS (**Figure 4**)[22-31]. In order to further justify the relevance of the added proteins as potential biomarkers, we proceeded to study three secreted proteins that our model predicted to be differentially expressed in the MS dataset (Annexin A1, sCD40L and sCD27). Notably, these proteins have been associated with MS previously[22, 23, 25]. 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 patients (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 supplementary notes, n=16). In both cohorts, only sCD27 was present at a detectable level, while Annexin A1 and sCD40L were not. 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 3B**), and treatment with Natalizumab reduced the sCD27 levels by 34% ($P=4.9 \times 10^{-4}$). Notably, sCD27 levels at baseline of 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 3B**), which was a stronger prediction than that of all our previously reported 14 biomarkers[32]. Taken together, using the splice variants-to-protein model we were able to *uniquely* identify and validate biomarkers of MS in an independent patient cohort, while these

genes could not be discovered using previous state-of-the-art test for differential gene expression.

For asthma we found six of the top 20 genes that were differentially expressed to previously be reported for the disease (**Table S3**). Next, we analysed asthma genes uniquely identified by our model and found seven genes that had previously also been reported to be associated with disease[33-38] and are currently being evaluated as potential therapeutic targets (**Figure 4**; **Table S4).** Examples of those genes include *NDRG1*, which regulates Th2 differentiation, a key driver in asthmatic disease, downstream of the mTORC2 complex[39, 40], *ADAM17*, a metalloproteinase involved in lung inflammation[35], *PIEZO1*, a mechanosensor regulating T cell activation[41] and pulmonary inflammatory responses[42], and the P-selectin ligand encoding gene *SELPLG*, important for recruitment of lymphocytes to the airways[43, 44]. Furthermore, the immunomodulatory genes *TNFAIP8* and *ARHGAP15* were identified in GWAS studies as shared risk variants for several IgE-mediated diseases including asthma, allergic rhinitis and atopic eczema[34]. Thus, we have validated that our model can identify important biomarkers and therapeutical targets also in the context of another immune-mediated disease, *i.e.* asthma.

Discussion

In the present study 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. Furthermore, we demonstrated the impact that this finding can have within genome medicine by predicting and validating biomarkers for MS and asthma.

Despite being part of the central dogma and of uttermost importance in biology and medicine, the prediction of protein levels from mRNA levels has long been associated with low precision, which has been a matter of debate[4]. Due to the complex process of mRNA-to-protein translation, there are several aspects that need to be considered[8]. In this paper we thoroughly addressed two presumed main aspects; (1) how to incorporate splice variants into the prediction protein expression, and (2) how to deal with the time-delay of the translation between mRNA and protein expression. Interestingly, both aspects were found to impact prediction of protein abundance, as shown in our combined model, although the incorporation of splice variants influenced the protein abundance prediction the most. Herein, we report

splice variants to have a wider correlation profile, both positive and negative, than what would be expected, and our novel approach takes advantage of this anti-correlation between splice variants and proteins. In previous work, the impact of incorporating splice variants into protein predictions has been analysed. These studies have focused on mechanistic cell-type independent factors such as splice variant-specific degradation rates[45]. Instead, we found that the correlations were cell-type specific and we constructed data-driven predictive models. In order to construct those models, we performed activation of NT_H cells followed by time-series analysis, which enabled us to infer the system based on its dynamics. These models were simplistic linear and time-delayed and validated through low out-of-sample prediction error. We found that usage of these models in complex disease enabled identification of more differentially expressed genes, which we therefore predicted as potential biomarkers. One such protein was validated as a biomarker for the MS disease prognosis. Thus, a main biological message is that intra-gene splice variant expressions influence translation, but the multifaceted nature of this mechanism remains too complex to capture with linear regression models.

Although incorporating splice variant information into the model was the main influential factor on the correlation, time delay also had an impact. The kinetics in translation of mRNA to protein is of general interest given its crucial importance in the design of experiments, for example in verifying relevance of mRNA expression to protein expression. Given that time-series experiments are time- and labor intensive, as well as expensive, a database that provides the relevant time delay between mRNA expression and the expression of its corresponding protein would be immensely valuable. Here, we present such an atlas, comprising almost 5000 gene expression-to-protein translation kinetics (**Table S1**).

A limitation with the paper is that we investigated few cell types, namely T_H1 cells, T_{REG} cells and B cells. We also only performed wet lab experiments in one of these cell types, but were able to transfer the approach to two other cell types *in silico*, showing the robustness of the model assumptions. Furthermore, the chosen cell types are central in regulation of immune responses, and the T_H cells indeed are involved in many complex and common illnesses, like infectious, allergic, autoimmune and cardiovascular diseases and cancer.

In conclusion, we have constructed data-driven linear models incorporating splice variant information and time delay to with high accuracy predict protein expression from mRNA expression. We have shown the general applicability of our approach by developing models for datasets from several cell types and shown the robustness of our approach. In addition, the

general principle of the model should be applicable to other cell types and can be used when that data becomes available. However, our data show that the model should be applied in a cell-specific manner given the low correlation in mixed tissue samples. 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. Finally, we have shown that approach is of clinical relevance through applying it to predict validated biomarkers.

Data availability statement

The raw and processed RNA-seq data was submitted to the EMBL-EBI sequencing archive arrayexpress and is available under the accession number E-MTAB-7775. The proteomics data was submitted to the EMBL-EBI proteomics repository PRIDE under the accession PXD013361.

Ethics consent and permissions

The study was approved by the Regional Ethics Committee in Linköping, Sweden (Dnr M180-07 and M2-09). All patients were recruited at the Department of Neurology, Linköping, University Hospital Sweden and both patients and controls gave written consent prior to inclusion.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

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 the Swedish foundation for strategic research.

Author contributions

MG initiated and supervised the study. RM and OR performed bioinformatics analyses, and RM performed the modelling. These analyses were led by MG, CA, JT, and DGC. OR

performed experimental work on T-cell differentiation, which were supervised by CEN, MCJ, JE and MB. MJK and CHN performed the proteomics analysis, which was supervised by MSK. FP and JM recruited patients and collected clinical material, and SH performed and analysed the biomarker validation assays, which were led by IK, MCJ, and JE. All authors contributed to and approved the final draft for publication.

Reference list

1. Clancy S, Brown W: **Translation: DNA to mRNA to Protein.** *Nature Education* 2008, **1**: 101.

2. Gustafsson M, Edstrom M, Gawel D, Nestor CE, Wang H, Zhang H, Barrenas F, Tojo J, Kockum I, Olsson T, Serra-Musach J, Bonifaci N, Pujana MA, Ernerudh J, Benson M: **Integrated genomic and prospective clinical studies show the importance of modular pleiotropy for disease susceptibility, diagnosis and treatment.** *Genome Med* 2014, **6**: 17.

3. Gawel DR, Serra-Musach J, Lilja S, Aagesen J, Arenas A, Asking B, Bengner M, Bjorkander J, Biggs S, Ernerudh J, Hjortswang H, Karlsson JE, Kopsen M, Lee EJ, Lentini A, Li X, Magnusson M, Martinez-Enguita D, Matussek A, Nestor CE, Schafer S, Seifert O, Sonmez C, Stjernman H, Tjarnberg A, Wu S, Akesson K, Shalek AK, Stenmarker M, Zhang H, Gustafsson M, Benson M: **A validated single-cell-based strategy to identify diagnostic and therapeutic targets in complex diseases.** *Genome Med* 2019, **11**: 47.

4. Fortelny N, Overall CM, Pavlidis P, Freue GVC: **Can we predict protein from mRNA levels?** *Nature* 2017, **547**: E19-E20.

5. Maier T, Guell M, Serrano L: **Correlation of mRNA and protein in complex biological samples.** *FEBS Lett* 2009, **583**: 3966-73.

6. de Sousa Abreu R, Penalva LO, Marcotte EM, Vogel C: **Global signatures of protein and mRNA expression levels.** *Mol Biosyst* 2009, **5**: 1512-26.

7. Vogel C, Marcotte EM: **Insights into the regulation of protein abundance from proteomic and transcriptomic analyses.** *Nat Rev Genet* 2012, **13**: 227-32.

8. Liu Y, Beyer A, Aebersold R: **On the Dependency of Cellular Protein Levels on mRNA Abundance.** *Cell* 2016, **165**: 535-50.

9. Zhao J, Qin B, Nikolay R, Spahn CMT, Zhang G: **Translatomics: The Global View of Translation.** *Int J Mol Sci* 2019, **20**.

10. Wethmar K, Smink JJ, Leutz A: **Upstream open reading frames: molecular switches in (patho)physiology.** *Bioessays* 2010, **32**: 885-93.

11. Floor SN, Doudna JA: **Tunable protein synthesis by transcript isoforms in human cells.** *Elife* 2016, **5**.

12. Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Colak R, Kim T, Misquitta-Ali CM, Wilson MD, Kim PM, Odom DT, Frey BJ, Blencowe BJ: **The evolutionary landscape of alternative splicing in vertebrate species.** *Science* 2012, **338**: 1587-93.

13. Sprent J, Tough DF: **Lymphocyte life-span and memory.** *Science* 1994, **265**: 1395-400.

14. Schmidt A, Marabita F, Kiani NA, Gross CC, Johansson HJ, Elias S, Rautio S, Eriksson M, Fernandes SJ, Silberberg G, Ullah U, Bhatia U, Lahdesmaki H, Lehtio J, Gomez-Cabrero D, Wiendl H, Lahesmaa R, Tegner J: **Time-resolved transcriptome and proteome landscape of human regulatory T cell (Treg) differentiation reveals novel regulators of FOXP3.** *BMC Biol* 2018, **16**: 47.

15. Raphael I, Nalawade S, Eagar TN, Forsthuber TG: **T cell subsets and their signature cytokines in autoimmune and inflammatory diseases.** *Cytokine* 2015, **74**: 5-17.

16. Kanduri K, Tripathi S, Larjo A, Mannerstrom H, Ullah U, Lund R, Hawkins RD, Ren B, Lahdesmaki H, Lahesmaa R: **Identification of global regulators of T-helper cell lineage specification.** *Genome Med* 2015, **7**: 122.

17. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigyarto CA, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, Edqvist PH, Berling H, Tegel H, Mulder J, Rockberg J, Nilsson P, Schwenk JM, Hamsten M, von Feilitzen K, Forsberg M, Persson L, Johansson F, Zwahlen M, von Heijne G, Nielsen J, Ponten F: **Proteomics. Tissue-based map of the human proteome.** *Science* 2015, **347**: 1260419.

18. Seumois G, Zapardiel-Gonzalo J, White B, Singh D, Schulten V, Dillon M, Hinz D, Broide DH, Sette A, Peters B, Vijayanand P: **Transcriptional Profiling of Th2 Cells Identifies Pathogenic Features Associated with Asthma.** *J Immunol* 2016, **197**: 655-64.

19. Rastogi D, Nico J, Johnston AD, Tobias TAM, Jorge Y, Macian F, Greally JM: **CDC42-related genes are upregulated in helper T cells from obese asthmatic children.** *J Allergy Clin Immunol* 2018, **141**: 539-548 e7.

20. Johansson P, Klein-Hitpass L, Choidas A, Habenberger P, Mahboubi B, Kim B, Bergmann A, Scholtysik R, Brauser M, Lollies A, Siebert R, Zenz T, Duhrsen U, Kuppers R, Durig J: **SAMHD1 is recurrently mutated in T-cell prolymphocytic leukemia.** *Blood Cancer J* 2018, **8**: 11.

21. James T, Linden M, Morikawa H, Fernandes SJ, Ruhrmann S, Huss M, Brandi M, Piehl F, Jagodic M, Tegner J, Khademi M, Olsson T, Gomez-Cabrero D, Kockum I: **Impact of genetic risk loci for multiple sclerosis on expression of proximal genes in patients.** *Hum Mol Genet* 2018, **27**: 912-928.

22. Colamatteo A, Maggioli E, Azevedo Loiola R, Hamid Sheikh M, Cali G, Bruzzese D, Maniscalco GT, Centonze D, Buttari F, Lanzillo R, Perna F, Zuccarelli B, Mottola M, Cassano S, Galgani M, Solito E, De Rosa V: **Reduced Annexin A1 Expression Associates with Disease Severity and Inflammation in Multiple Sclerosis Patients.** *J Immunol* 2019, **203**: 1753-1765.

23. van der Vuurst de Vries RM, Mescheriakova JY, Runia TF, Jafari N, Siepman TA, Hintzen RQ: **Soluble CD27 Levels in Cerebrospinal Fluid as a Prognostic Biomarker in Clinically Isolated Syndrome.** *JAMA Neurol* 2017, **74**: 286-292.

24. Wong YYM, van der Vuurst de Vries RM, van Pelt ED, Ketelslegers IA, Melief MJ, Wierenga AF, Catsman-Berrevoets CE, Neuteboom RF, Hintzen RQ: **T-cell activation marker sCD27 is associated with clinically definite multiple sclerosis in childhood-acquired demyelinating syndromes.** *Mult Scler* 2018, **24**: 1715-1724.

25. Masuda H, Mori M, Uchida T, Uzawa A, Ohtani R, Kuwabara S: **Soluble CD40 ligand contributes to blood-brain barrier breakdown and central nervous system inflammation in multiple sclerosis and neuromyelitis optica spectrum disorder.** *J Neuroimmunol* 2017, **305**: 102-107.

26. Wanke F, Moos S, Croxford AL, Heinen AP, Graf S, Kalt B, Tischner D, Zhang J, Christen I, Bruttger J, Yogev N, Tang Y, Zayoud M, Israel N, Karram K, Reissig S, Lacher SM, Reichhold C, Mufazalov IA, Ben-Nun A, Kuhlmann T, Wettschureck N, Sailer AW, Rajewsky K, Casola S, Waisman A, Kurschus FC: **EBI2 Is Highly Expressed in Multiple**

Sclerosis Lesions and Promotes Early CNS Migration of Encephalitogenic CD4 T Cells. *Cell Rep* 2017, **18**: 1270-1284.

27. Bomprezzi R, Ringner M, Kim S, Bittner ML, Khan J, Chen Y, Elkahloun A, Yu A, Bielekova B, Meltzer PS, Martin R, McFarland HF, Trent JM: **Gene expression profile in multiple sclerosis patients and healthy controls: identifying pathways relevant to disease.** *Hum Mol Genet* 2003, **12**: 2191-9.

28. Aquino DA, Capello E, Weisstein J, Sanders V, Lopez C, Tourtellotte WW, Brosnan CF, Raine CS, Norton WT: **Multiple sclerosis: altered expression of 70- and 27-kDa heat shock proteins in lesions and myelin.** *J Neuropathol Exp Neurol* 1997, **56**: 664-72.

29. Bonetti B, Stegagno C, Cannella B, Rizzuto N, Moretto G, Raine CS: **Activation of NF-kappaB and c-jun transcription factors in multiple sclerosis lesions. Implications for oligodendrocyte pathology.** *Am J Pathol* 1999, **155**: 1433-8.

30. Achiron A, Feldman A, Mandel M, Gurevich M: **Impaired expression of peripheral blood apoptotic-related gene transcripts in acute multiple sclerosis relapse.** *Ann N Y Acad Sci* 2007, **1107**: 155-67.

31. de JG-GJ, Rojas-Mayorquin AE, Valle Y, Padilla-Gutierrez JR, Castaneda-Moreno VA, Mireles-Ramirez MA, Munoz-Valle JF, Ortuno-Sahagun D: **Decreased serum levels of sCD40L and IL-31 correlate in treated patients with Relapsing-Remitting Multiple Sclerosis.** *Immunobiology* 2018, **223**: 135-141.

32. Håkansson I, Tisell A, Cassel P, Blennow K, Zetterberg H, Lundberg P, Dahle C, Vrethem M, Ernerudh J: **Neurofilament levels, disease activity and brain volume during follow-up in multiple sclerosis.** *J Neuroinflammation* 2018, **15**: 209.

33. Nestor CE, Barrenas F, Wang H, Lentini A, Zhang H, Bruhn S, Jornsten R, Langston MA, Rogers G, Gustafsson M, Benson M: **DNA methylation changes separate allergic patients from healthy controls and may reflect altered CD4+ T-cell population structure.** *PLoS Genet* 2014, **10**: e1004059.

34. Ferreira MA, Vonk JM, Baurecht H, Marenholz I, Tian C, Hoffman JD, Helmer Q, Tillander A, Ullemar V, van Dongen J, Lu Y, Ruschendorf F, Esparza-Gordillo J, Medway CW, Mountjoy E, Burrows K, Hummel O, Grosche S, Brumpton BM, Witte JS, Hottenga JJ, Willemsen G, Zheng J, Rodriguez E, Hotze M, Franke A, Revez JA, Beesley J, Matheson MC, Dharmage SC, Bain LM, Fritsche LG, Gabrielsen ME, Balliu B, andMe Research T, collaborators A, consortium B, LifeLines Cohort S, Nielsen JB, Zhou W, Hveem K,

Langhammer A, Holmen OL, Loset M, Abecasis GR, Willer CJ, Arnold A, Homuth G, Schmidt CO, Thompson PJ, Martin NG, Duffy DL, Novak N, Schulz H, Karrasch S, Gieger C, Strauch K, Melles RB, Hinds DA, Hubner N, Weidinger S, Magnusson PKE, Jansen R, Jorgenson E, Lee YA, Boomsma DI, Almqvist C, Karlsson R, Koppelman GH, Paternoster L: **Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology.** *Nat Genet* 2017, **49**: 1752-1757.

35. Dreymueller D, Uhlig S, Ludwig A: **ADAM-family metalloproteinases in lung inflammation: potential therapeutic targets.** *Am J Physiol Lung Cell Mol Physiol* 2015, **308**: L325-43.

36. Poole A, Urbanek C, Eng C, Schageman J, Jacobson S, O'Connor BP, Galanter JM, Gignoux CR, Roth LA, Kumar R, Lutz S, Liu AH, Fingerlin TE, Setterquist RA, Burchard EG, Rodriguez-Santana J, Seibold MA: **Dissecting childhood asthma with nasal transcriptomics distinguishes subphenotypes of disease.** *J Allergy Clin Immunol* 2014, **133**: 670-8 e12.

37. Persson H, Kwon AT, Ramilowski JA, Silberberg G, Soderhall C, Orsmark-Pietras C, Nordlund B, Konradsen JR, de Hoon MJ, Melen E, Hayashizaki Y, Hedlin G, Kere J, Daub CO: **Transcriptome analysis of controlled and therapy-resistant childhood asthma reveals distinct gene expression profiles.** *J Allergy Clin Immunol* 2015, **136**: 638-48.

38. Enomoto Y, Orihara K, Takamasu T, Matsuda A, Gon Y, Saito H, Ra C, Okayama Y: **Tissue remodeling induced by hypersecreted epidermal growth factor and amphiregulin in the airway after an acute asthma attack.** *J Allergy Clin Immunol* 2009, **124**: 913-20 e1-7.

39. Heikamp EB, Patel CH, Collins S, Waickman A, Oh MH, Sun IH, Illei P, Sharma A, Naray-Fejes-Toth A, Fejes-Toth G, Misra-Sen J, Horton MR, Powell JD: **The AGC kinase SGK1 regulates TH1 and TH2 differentiation downstream of the mTORC2 complex.** *Nat Immunol* 2014, **15**: 457-64.

40. Murray JT, Campbell DG, Morrice N, Auld GC, Shpiro N, Marquez R, Peggie M, Bain J, Bloomberg GB, Grahammer F, Lang F, Wulff P, Kuhl D, Cohen P: **Exploitation of KESTREL to identify NDRG family members as physiological substrates for SGK1 and GSK3.** *Biochem J* 2004, **384**: 477-88.

41. Liu CSC, Raychaudhuri D, Paul B, Chakrabarty Y, Ghosh AR, Rahaman O, Talukdar A, Ganguly D: **Cutting Edge: Piezo1 Mechanosensors Optimize Human T Cell Activation.** *J Immunol* 2018, **200**: 1255-1260.

42. Solis AG, Bielecki P, Steach HR, Sharma L, Harman CCD, Yun S, de Zoete MR, Warnock JN, To SDF, York AG, Mack M, Schwartz MA, Dela Cruz CS, Palm NW, Jackson R, Flavell RA: **Mechanosensation of cyclical force by PIEZO1 is essential for innate immunity.** *Nature* 2019, **573**: 69-74.

43. Purwar R, Campbell J, Murphy G, Richards WG, Clark RA, Kupper TS: **Resident memory T cells (T(RM)) are abundant in human lung: diversity, function, and antigen specificity.** *PLoS One* 2011, **6**: e16245.

44. Leath TM, Singla M, Peters SP: **Novel and emerging therapies for asthma.** *Drug Discov Today* 2005, **10**: 1647-55.

45. Eraslan B, Wang D, Gusic M, Prokisch H, Hallstrom BM, Uhlen M, Asplund A, Ponten F, Wieland T, Hopf T, Hahne H, Kuster B, Gagneur J: **Quantification and discovery of sequence determinants of protein-per-mRNA amount in 29 human tissues.** *Mol Syst Biol* 2019, **15**: e8513.

Figure legends

Fig 1. RNA-Seq and mass-spectrometry analysis of T_H1 differentiation revealed highly variable correlations. (A) Experimental design. (B) Heat map of transcript and protein abundance dynamics in genes that show significant negative (left) and positive (right) correlations. (C) Examples of transcript splice variants showing that both *STX12* (left) and *IL7R* (right) were significantly negatively and positively correlated with protein levels. (D) Illustration of the modelling procedure for resolving the poor correlation, using STX12 as an example.

Fig 2. Multiple transcripts and time-delays increased mRNA and protein correlations significantly in multiple cell-types. (A) Gene/protein Pearson correlations in T_H1 (left), T_{reg} (middle left), and murine B-cell (middle right) differentiation. In the histogram, the grey curve shows the correlation distribution when the sum of all splice variant expressions of a transcript [4] is used to quantify mRNA abundance (median: dashed line), while in the blue histogram our time-delayed multiple splice variant based models are used (medians: solid lines at 0.86, 0.79, and 0.94 for T_H1, T_{reg} and murine B-cells, respectively). Only cross-validated protein predictions are shown for the proteins for which the null-model could be rejected. (B) Out-of-sample cross validation prediction of the three models. Aiming to quantify the predictive power of each added input to the model, we observed that a linear model with gene-specific time-delays was the model that generated predictions with the smallest sum of squared residuals. (C) Median correlation coefficients (rho) for different mathematical protein prediction models derived from RNA with increasing protein abundance correlations. P-values were derived from predictions using leave-one-out cross-validation.

Fig 3. 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 T_H1 (light blue) and T_{reg} (dark blue) models showed higher fraction of nominally significant genes than that of normal differential gene expression tests. (B) Validation of PP from early MS (clinically isolated syndrome (CIS)) *vs* healthy controls (HC) and pre *vs* post one-year treatment with Natalizumab. Validation is performed measuring sCD40 in cerebrospinal fluid (CSF) and

stratifying on phenotyping. Left plots show healthy controls vs CIS showing patients with no evidence of disease activity (NEDA) at four years treatment with filled circles. (C) Receiver operating curve using sCD27 concentration as a single prognostic marker of NEDA at four (solid line) and two years (dashed line) after CIS.

Fig 4. Overview of detected potential biomarkers in asthma and MS. The model identified several proteins that have previously been identified in MS and asthma. The upper panel shows the potential biomarkers identified in MS and the lower panel shows the same in asthma. *mRNA expression, ¤ identified in mice. PBMCs, peripheral blood mononuclear cells. References are given in the figure.

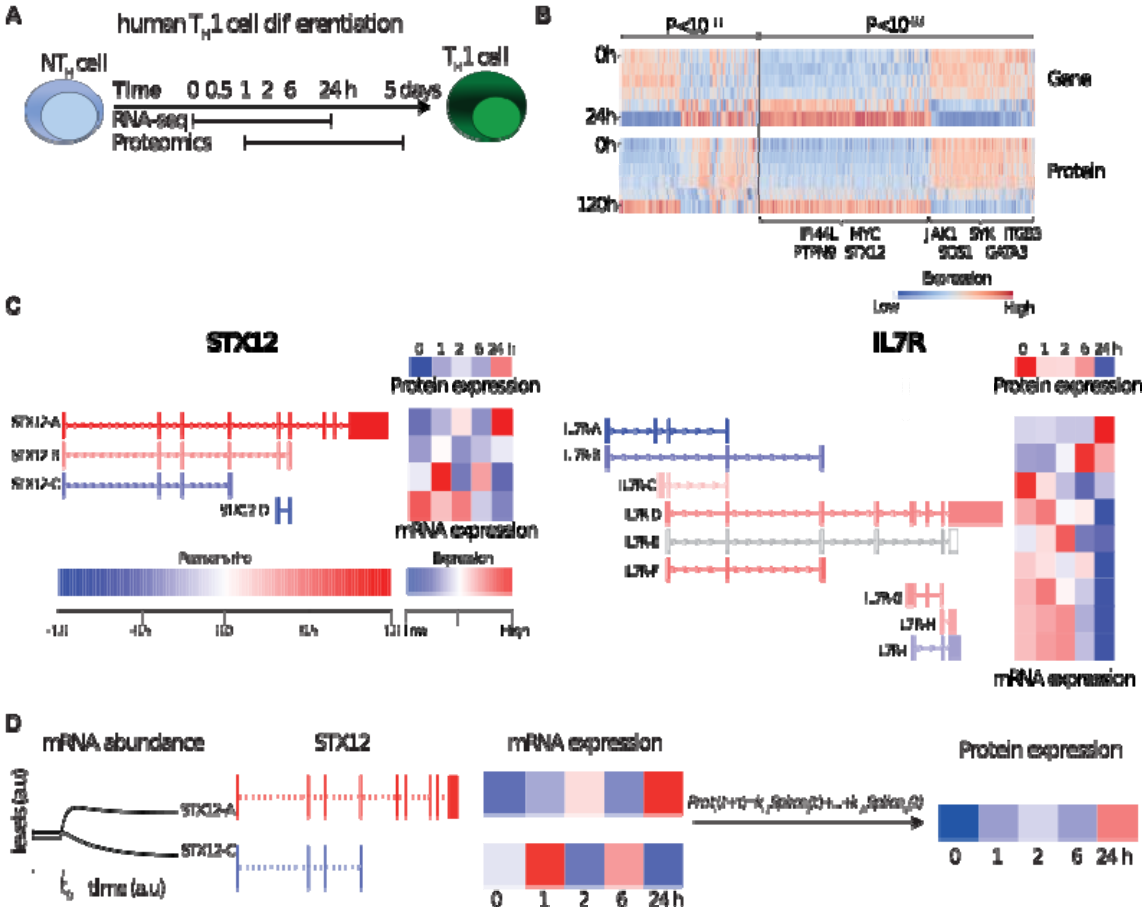


Figure 1

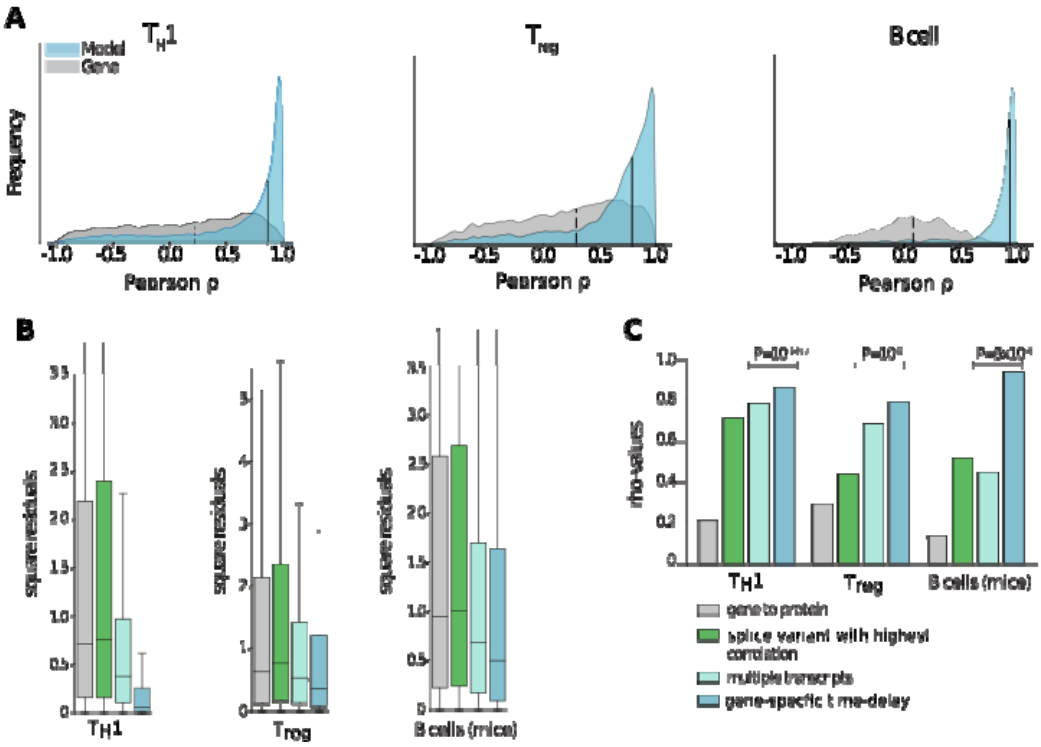


Figure 2

Figure 3

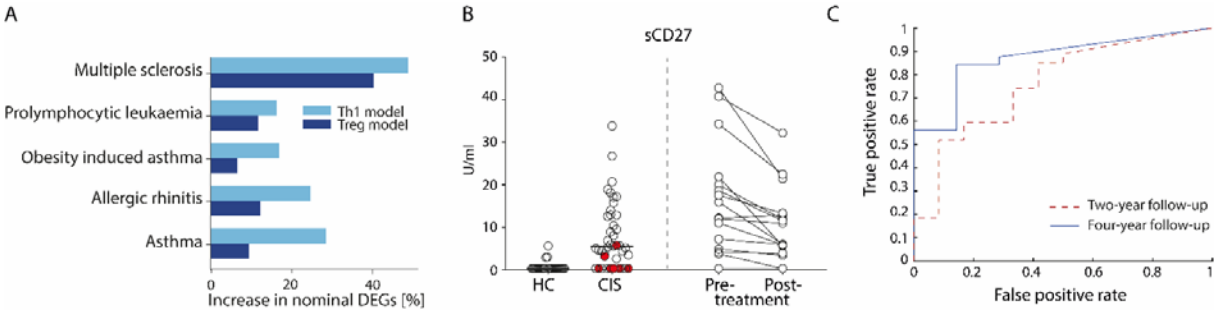
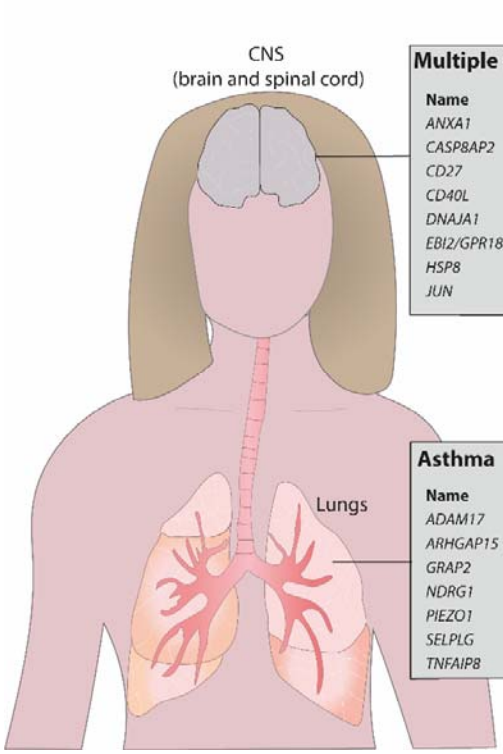


Figure 4



Multiple sclerosis

Name	Function	Detected in	Expression in MS
ANXA1	Anti-inflammatory mediator	Plasma [22]	Lower
CASP8AP2	Apoptosis and cell cycle	PBMCs[30]*	Higher during relapse
CD27	Co-stimulatory immune checkpoint	CSF [23,24]	Higher
CD40L	Co-stimulatory immune checkpoint	Serum[25,31], CSF[25]	Lower [31], Higher[25]
DNAJA1	Co-chaperone	PBMCs[27]	Higher
EBI2/GPR183	Immune cell migration	T cells[26]*¤	Higher
HSP8	Chaperone	Brain lesions[28]*	Lower
JUN	Cellular proliferation, apoptosis	Glial cells[29]	Lower

Asthma

Name	Function	Dectected in	Connection to asthma
ADAM17	Metalloproteinase	Sputum[38]	Suggested therapeutic target
ARHGAP15	GTPase activation	Whole blood[34]	Gene risk variant
GRAP2	Protein-tyrosine kinase signaling	PBMCs[37]	Differenitally expressed in asthma
NDRG1	Epithelial barrier integrity	Nasal tissue[40]	Upregulated in asthma
PIEZO1	Mechanosensor	T cells[33]	Differentially methylated in allergy
SELPLG	Recruitment of lymphocytes	Lung T cells [43]	Suggested therepeutic target
TNFAIP8	Negative regulator of apoptosis	Whole blood[34]	Gene risk variant