RNA-Sequencing And Mass-Spectrometry Proteomic Time-Series Analysis of T-Cell Differentiation Identified Multiple Splice Variants Models That Predicted Validated Protein Biomarkers In Inflammatory Diseases

Rasmus Magnusson
Linköping University
Olof Rundquist
Linköping University
Min Jung Kim
Kyung Hee University
Sandra Hellberg
Linköping University
Chan Hyun Na
Johns Hopkins University School of Medicine
Mikael Benson
Linköping University
David Gomez-Cabrero
Navarrabiomed, Complejo Hospitalario de Navarra, Universidad Pública de Navarra
Ingrid Kockum
Karolinska Institute
Jesper Tegnér
King Abdullah University of Science and Technology (KAUST)
Fredrik Piehl
Karolinska Institute
Maja Jagodic
Karolinska Institute
Johan Mellergård
Linköping University
Claudio Altafini
Linköping University
Jan Ernerudh
Linköping University

Maria C. Jenmalm
Linköping University
Colm E. Nestor
Linköping University
Min-Sik Kim
Daegu Gyeongbuk Institute of Science and Technology
Mika Gustafsson (mika.gustafsson@liu.se)
Linköping University

Research Article

Keywords: Biomarker discovery, disease severity, multiple sclerosis, sCD27, Th1 differentiation, time-series analysis, proteomics, mass-spectronomy, splice-variants, time-delay

Posted Date: September 17th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-880591/v1

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 hypothesized that these correlations could be improved by mathematical models based on measuring splice variants and time delay in protein translation.

Methods We characterized 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, further validating the findings for multiple sclerosis (MS) and asthma.

Results The new models 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 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 (sometimes even impossible with high sensitivity due to ethical constraints) 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 (RNA-seq). Another advantage of using mRNA expression as a core vehicle for biomarker discovery is that mRNA profiling can be performed even in samples of limited amount, like biopsies.

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 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 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 enabling relative quantification of 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 [2], 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-seq, we noted a much better agreement between our mRNA-based measure and proteomics. To test our models, we showed the potential clinical usefulness of our derived models by detecting potential biomarkers in five complex diseases. Throughout the paper we aimed at developing a generally applicable tool to increase the sensitivity in RNA-seq analysis in many different studies, for which some could not identify any genes with FDR < 0.05 due to lack of statistical power. We therefore used the fraction of nominally ($P < 0.05$) differentially expressed genes as a metric (which then will be binomially distributed) of whether our approach did increase enrichment of potential genes which is a necessary requirement for a case control study. 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 multiple sclerosis (MS) supported their biological relevance. Finally, we validated one of the predicted biomarkers, sCD27, using two independent MS 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 different cell types, species and diseases shows its general applicability to increase the power of RNA-seq based studies for biomarker discovery.

Material And Methods

Isolation of $CD4^+$ T helper (T_H) cells and T_H1 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 (1 bead/cell) (Dynal AS, Lillestøm, Norway), 5 ng/µl recombinant human IL-12p70, 10 ng/µl recombinant human IL-2 and 5 µg/µl anti-IL-4 antibodies (clone MAB204) (all three from, Bio-Techne, Minneapolis, USA). All primary CD4-T-cell cultures were cultured and differentiated at 37 ℃, with 5% CO_2 in RPMI 1640 media containing L-glutamine, 10% FBS and 1% Penicillin/Streptomycin mixture (all from Gibco, Paisley, United Kingdom).

RNA-seq and proteomics sample collection

Naive CD45RA$^+$ CD4$^+$ T-cells were isolated, differentiated and sampled as above at baseline, 0.5h, 1h, 2h, 6h and 24h for RNA-seq, and baseline, 1h, 2h, 6h, 24h and 5 days for proteomics. RNA was isolated using a ZR-Duet DNA/RNA kit (Zymo Research, Irvine, USA) and stored at -80°C until transport. During the protein extraction, multiple samples were pooled from twelve different individuals to reach the necessary amount of material for the three biological replicates required for 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). T he false discovery rate (FDR) for peptide level was evaluated to 0.01 for removing false positive data. For highly confident quantifications of protein, protein ratios were calculated from two or more unique quantitative peptides in each replicate. Data was normalized and removed contaminant and razor peptide. To enrich differentially expressed proteins (DEPs), we analysed the quantitative ratios (as the Log2 value). The fold-change ratio cut off was more than 2 or less than 0.5 based on intensity of 0 min. Searched data went through statistical process with Perseus (v 1.5.1.6). Detailed experimental procedures are provided in the Supplementary information.

RNA-seq library preparation and sequencing

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 with pair end sequencing and a read length of 100bp on an Illumina 2500 instrument.

Bioinformatics

All RNA-seq data were processed using the following pipeline. Sample qualities were assessed with fastQC and the mRNA reads were subsequently aligned using STAR[48], with the parameter "–outSAMstrandField intronMotif" and "–outFilterIntronMotifs RemoveNoncanonical", to the "Homo_sapiens.GRCh37.75.dna.primary_assembly.fa" from Ensemble. The resulting read alignment bam files were assembled into transcripts with StringTie, with default parameters, 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 [49]. Correlations were calculated using Pearson correlations across gene expressions, i.e. one coefficient per gene.

Splice variant model construction

We hypothesized that protein abundance could be predicted using a linear combination of the corresponding splice variants. To predict protein abundance, we used the Sklearn[50] implementation of the LASSO[51], an L1-penalized linear regression model.

$$\min_{\beta,\in Re}\{\frac{1}{N}\|Y-\beta X\|_2+\lambda\|\beta\|_1\}$$

Here, the time series of one protein is denoted the vector Y, and the corresponding time series of the splice variants are denoted by the matrix X. The rate constant for each splice variant is contained in the vector β. Furthermore, the λ parameter regulates the influence of the L1 term, and was determined individually for each protein. The λ term was chosen to minimize the prediction error of a leave-one-out cross validation. In the Th1 dataset, the time points differed such that the mRNA abundance also had a measurement at t=30 minutes, while the protein data instead had a measurement of t=120h. For comparison, the protein data for 30 minutes was interpolated, while the 120h time point was omitted. The same procedure was performed using the regulatory T cells from Schmidt *et al.* 2018[14] were Treg induced by either TGF-β, TGF-β and ATRA, or TGF-β and butyrate. Lastly, the same procedure was performed for mice B-cells were B-cell differentiation was induced by the Ikaros transcription factor (GSE75417). Pipe-line and code available from https://gitlab.com/Gustafsson-lab/splice_protein_predictions.

Cross validation

To select the values of λ and τ, a double cross-validation was performed. First, one of the time points of the protein measurements was removed from the set, leaving only 5 data points. Secondly, a leave-one out cross-validation was performed on the remaining 5 time points, giving an estimate of the accuracy of the model approach given a time delay and a lambda value for the penalty term in the Lasso operator. We used the 200 time-delays ranging between 0 and 24h, and a varying set of lambda parameters (increased until all parameters equaled zero). Thirdly, the time delay and penalisation that generated the smallest average squared residuals between the second cross-validation and the data were chosen and used to

predict the 6th data point from splice variants. Fourth, this double cross-validation procedure was repeated for all 6 data points.

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[20]. For each of the studies' datasets we used the T_H1 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 of nominally differential expression from model predictions compared to ingoing splice variants into the model. For eleven different proteins from the two largest biomarker studies of MS we could find protein measurements which was compared with our predicted proteins. One study reported 36 out of 92 proteins as significant [32] and another study [33] reported the expression of four proteins whereof two were significant. We found that all our predicted proteins differential expression agreed with the two studies (9/9 negatively reported from first study and 1/1 negatively and 1/1 positively reported from second study) and the corresponding P-value was calculated as $((92\text{-}36)/92)^9 \times (2/4)^2 = 2.9\text{x}10^{-3}$.

Protein validation

Three of the proteins predicted differentially expressed (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 and a second with 16 patients with relapsing remitting MS before and after one year of treatment with Natalizumab (See Patient cohort above and Supplemental Table S5). Quantification of sCD27 was performed using the Human Instant ELISA™ kit from eBioscience according to the instructions provided by the manufacturer. The optical densities (O.D.) were read at 450 nm with a wavelength correction at 620 nm in a Sunrise™ microplate reader (Tecan, Shanghai, China). Data acquisition was performed using Magellan™ version 7.1 computer software. The lowest detection limit was 0.63 U/ml and values below the detection limit were given half the value of the detection limit. Statistical differences were determined using Mann-Whitney U-test or Wilcoxon matched-pairs signed rank test in Graphpad Prism. Annexin A1, measured by the human Annexin A1 ELISA kit, was undetectable in all analyzed samples (n=32, of whom n=16 were included before and 16 after one year of treatment with Natalizumab). Multiplex Bead Technology was used to measure soluble CD40L according to the manufacturer's description. The samples were analysed on a Luminex®200™ instrument (Invitrogen, Carlsbad, CA, USA) and data was collected using xPONENT 3.1™ and analysed using the MasterPlex® Reader Fit. The lowest detection limit was 1.6 pg/ml and values below the detection limit

were given half the value of the detection limit. sCD40L concentration was below the lowest detection limit in 71 out of 96 samples (74% undetectable) and was therefore considered as undetectable.

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_H1 differentiation. This was done by performing time-series RNA-seq and mass-spectrometry proteomics of primary human NT_H cells (Fig. 1A, S1-2). 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 (Fig. 1A, S1-2), whereof five contained paired omics and could be used for correlation analysis below. This approach detected 6909 expressed proteins and 15699 expressed genes. Out of the 6909 expressed proteins, 5749 could be mapped to genes and out of those, 4920 where found expressed in the RNA-seq data. As expected, a significant fraction of the genes showed a nominally 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, $P < 10^{-11}$) (Fig. 1B, **Table S1**). Notably, the overall median Pearson correlation (rho) between mRNA and protein was only 0.21. Additionally, analysis of the distribution of correlation coefficients revealed significant enrichments of 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). Known T_H cell associated genes, for example *IL7R* and *STX12* [16] contained multiple splice variants, of which several were positively or negatively correlated to their corresponding protein levels (Fig. 1C, S3). 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, Fig. 1D, S4). The rationale for the L1 penalty was to effectively remove splice variants that carry little or no predictive power over protein abundance. In practice this resulted in maximum three splice-variants per protein for the Th1 model which is a method limitation due to the few data points and our regularisation, and the median number of splice variants were two for the human T-cell datasets and

one for the mice B-cells. This simple model resulted in a median gene-protein correlation of rho_{TH1} = 0.86 using cross-validated predictions (Fig. 2A**).** Likewise, to test the generality of the approach 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) (Fig. 2A). Next, 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 using only the most highly correlated 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) (Fig. 2B-C), thus demonstrating that both multiple dynamical splice variants and time delay increase the fit of data needed for optimal performance. In summary, we have identified simple linear models of mRNA splice variants and time delay and we can model the time-courses in T and B-cell differentiation well (see the full models in **Table S1**). We would like to emphasize that this is a minimal requirement for mRNA-protein models to be meaningful, so we proceeded to analyse if the models were useful to translational research by identifying biomarkers in complex diseases.

The models showed increased biomarker sensitivity which were further verified 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 datasets for five different diseases [17–20]; asthma, allergic rhinitis, obesity-induced asthma, pro-lymphocytic leukaemia, and 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 $P < 9.8 \times 10^{-4}$) (Fig. 3A). Moreover, we observed a consistently 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 were detected at 20% FDR level 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 (Fig. 4) [21–30]. 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 [21, 22, 24]. 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 in CSF 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 $P = 3.0 \times 10^{-8}$, Fig. 3B, **Table S6**), 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 and CIS patients were able to predict disease activity after four years follow up (AUC = 0.87, $P = 1.2 \times 10^{-3}$, Fig. 3C), which was a stronger prediction than that of all our previously reported 14 biomarkers [31]. Taken together, using the splice variants-to-protein model we were able to *uniquely* identify and validate biomarkers for the prognosis 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 MS we found two large biomarker studies [32, 33] that reported 11 of the analysed genes whereof 10 were non-significant and one were significant, which all agreed with our findings of their significance ($P < 2.9 \times 10^{-3}$; see Methods).

For asthma we found six of the top 20 genes that were differentially expressed by conventional mRNA expression to be previously associated with the disease (**Table S3**). Next, we analysed asthma-associated genes uniquely identified by our model and found seven additional genes. Interestingly, these genes had previously also been reported to be relevant for the disease [34–39], and are currently being evaluated as potential therapeutic targets (Fig. 4; **Table S4).** Examples of those genes include *NDRG1*, which regulates T_H2 differentiation, a key driver in asthmatic disease, downstream of the mTORC2 complex [40, 41], *ADAM17*, a metalloproteinase involved in lung inflammation [36], *PIEZO1*, a mechanosensor regulating T cell activation [42] and pulmonary inflammatory responses [43], and the P-selectin ligand encoding gene *SELPLG*, important for recruitment of lymphocytes to the airways [44, 45]. 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 [35]. Thus, we have validated that our model can identify relevant biomarker candidates 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 improve our ability to predict protein abundance from mRNA abundance. Furthermore, we demonstrated the potential impact of this finding within translational medicine by predicting and validating biomarkers for the inflammatory diseases 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 negative 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 [46]. 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. A necessary requirement for such as model was dynamical data covering a decent number of time-points that allowed for the possibility of including modelling of intermediate time-points and the inference of time delays. From our models we proposed a biomarker discovery strategy which was validated in three steps. First, we found that usage of these models in complex disease enabled identification of more differentially expressed genes, which we therefore predicted as potential biomarkers. Second, we noted that many of the predicted proteins had previously been associated with MS and asthma, confirming that our strategy predicts relevant disease genes. Third, we validated one such protein as a biomarker in MS, namely sCD27. While sCD27 has already been associated with MS, our clinical analysis of two independent cohorts yielded novel findings of remarkably good prognostic capabilities of four years disease activity and of treatment response monitoring, which is important areas for early MS treatment selection.

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. Such models should ideally be functionally validated based on mechanistic principles, described by ordinary differential equations, such as the ones presented by for example Jovanovic et al. [47]. However, given that time-series experiments are expensive, time- and labor intensive, and predictive large-scale models are highly needed for biomarker discoveries, 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 key cell types, namely T_H1 cells, T_{REG} cells and B cells, and we performed wet lab experiments in only one of these cell types. However, we were able to transfer the approach to two other cell types by re-using data of other studies, demonstrating 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 conditions, 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 predict protein expression from mRNA. We showed the general applicability of our approach by developing robust models for datasets from several cell types, and therefore the general principle of the model should be applicable to yet other cell types. For example, we expect this modelling strategy to be applicable to 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 showed that our proposed approach is of clinical relevance in prediction of validated biomarkers.

Declarations

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We would like to thank Jun Hyung Lee for his contribution to the proteomics sample preparation.

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 (NRF-2016K1A3A1A47921601, 2017M3C7A1027472).

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.

Availability of data and materials

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. Pipe-line and code for the mathematical modelling and bioinformatics analysis available from https://gitlab.com/Gustafsson-lab/splice_protein_predictions.

Ethics consent and permissions

The study was performed in accordance with the Declaration of Helsinki and 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 informed and written consent prior to inclusion.

Consent to publish

Not applicable.

Abbreviations

MS Multiple sclerosis

TH Help T lymphocytes

IgE Immunoglobulin E

References

1. Clancy, S. and W. Brown, *Translation: DNA to mRNA to Protein*. Nature Education, 2008. **1**: p. 101.
2. Gustafsson, M., et al., *Integrated genomic and prospective clinical studies show the importance of modular pleiotropy for disease susceptibility, diagnosis and treatment*. Genome Med, 2014. **6**(2): p. 17. https://doi.org/10.1186/gm534.
3. Gawel, D.R., et al., *A validated single-cell-based strategy to identify diagnostic and therapeutic targets in complex diseases*. Genome Med, 2019. **11**(1): p. 47. https://doi.org/10.1186/s13073-019-0657-3.
4. Fortelny, N., et al., *Can we predict protein from mRNA levels?* Nature, 2017. **547**(7664): p. E19-E20. https://doi.org/10.1038/nature22293.
5. Maier, T., M. Guell, and L. Serrano, *Correlation of mRNA and protein in complex biological samples*. FEBS Lett, 2009. **583**(24): p. 3966–73. https://doi.org/10.1016/j.febslet.2009.10.036.
6. de Sousa Abreu, R., et al., *Global signatures of protein and mRNA expression levels*. Mol Biosyst, 2009. **5**(12): p. 1512–26. https://doi.org/10.1039/b908315d.
7. Vogel, C. and E.M. Marcotte, *Insights into the regulation of protein abundance from proteomic and transcriptomic analyses*. Nat Rev Genet, 2012. **13**(4): p. 227–32. https://doi.org/10.1038/nrg3185.
8. Liu, Y., A. Beyer, and R. Aebersold, *On the Dependency of Cellular Protein Levels on mRNA Abundance*. Cell, 2016. **165**(3): p. 535–50. https://doi.org/10.1016/j.cell.2016.03.014.
9. Zhao, J., et al., *Translatomics: The Global View of Translation*. Int J Mol Sci, 2019. **20**(1). https://doi.org/10.3390/ijms20010212.

10. Wethmar, K., J.J. Smink, and A. Leutz, *Upstream open reading frames: molecular switches in (patho)physiology.* Bioessays, 2010. **32**(10): p. 885 – 93. https://doi.org/10.1002/bies.201000037.
11. Floor, S.N. and J.A. Doudna, *Tunable protein synthesis by transcript isoforms in human cells*. Elife, 2016. **5**. https://doi.org/10.7554/eLife.10921.
12. Barbosa-Morais, N.L., et al., *The evolutionary landscape of alternative splicing in vertebrate species*. Science, 2012. **338**(6114): p. 1587–93. https://doi.org/10.1126/science.1230612.
13. Sprent, J. and D.F. Tough, *Lymphocyte life-span and memory*. Science, 1994. **265**(5177): p. 1395–400.
14. Schmidt, A., et al., *Time-resolved transcriptome and proteome landscape of human regulatory T cell (Treg) differentiation reveals novel regulators of FOXP3*. BMC Biol, 2018. **16**(1): p. 47. https://doi.org/10.1186/s12915-018-0518-3.
15. Raphael, I., et al., *T cell subsets and their signature cytokines in autoimmune and inflammatory diseases*. Cytokine, 2015. **74**(1): p. 5–17. https://doi.org/10.1016/j.cyto.2014.09.011.
16. Kanduri, K., et al., *Identification of global regulators of T-helper cell lineage specification*. Genome Med, 2015. **7**: p. 122. https://doi.org/10.1186/s13073-015-0237-0.
17. Seumois, G., et al., *Transcriptional Profiling of Th2 Cells Identifies Pathogenic Features Associated with Asthma*. J Immunol, 2016. **197**(2): p. 655 – 64. https://doi.org/10.4049/jimmunol.1600397.
18. Rastogi, D., et al., *CDC42-related genes are upregulated in helper T cells from obese asthmatic children*. J Allergy Clin Immunol, 2018. **141**(2): p. 539–548 e7. https://doi.org/10.1016/j.jaci.2017.04.016.
19. Johansson, P., et al., *SAMHD1 is recurrently mutated in T-cell prolymphocytic leukemia*. Blood Cancer J, 2018. **8**(1): p. 11. https://doi.org/10.1038/s41408-017-0036-5.
20. James, T., et al., *Impact of genetic risk loci for multiple sclerosis on expression of proximal genes in patients*. Hum Mol Genet, 2018. **27**(5): p. 912–928. https://doi.org/10.1093/hmg/ddy001.
21. Colamatteo, A., et al., *Reduced Annexin A1 Expression Associates with Disease Severity and Inflammation in Multiple Sclerosis Patients*. J Immunol, 2019. **203**(7): p. 1753–1765. https://doi.org/10.4049/jimmunol.1801683.
22. van der Vuurst de Vries, R.M., et al., *Soluble CD27 Levels in Cerebrospinal Fluid as a Prognostic Biomarker in Clinically Isolated Syndrome*. JAMA Neurol, 2017. **74**(3): p. 286–292. https://doi.org/10.1001/jamaneurol.2016.4997.
23. Wong, Y.Y.M., et al., *T-cell activation marker sCD27 is associated with clinically definite multiple sclerosis in childhood-acquired demyelinating syndromes*. Mult Scler, 2018. **24**(13): p. 1715–1724. https://doi.org/10.1177/1352458518786655.
24. Masuda, H., et al., *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**: p. 102–107. https://doi.org/10.1016/j.jneuroim.2017.01.024.

25. Wanke, F., et al., *EBI2 Is Highly Expressed in Multiple Sclerosis Lesions and Promotes Early CNS Migration of Encephalitogenic CD4 T Cells*. Cell Rep, 2017. **18**(5): p. 1270–1284. https://doi.org/10.1016/j.celrep.2017.01.020.
26. Bomprezzi, R., et al., *Gene expression profile in multiple sclerosis patients and healthy controls: identifying pathways relevant to disease*. Hum Mol Genet, 2003. **12**(17): p. 2191–9. https://doi.org/10.1093/hmg/ddg221.
27. Aquino, D.A., et al., *Multiple sclerosis: altered expression of 70- and 27-kDa heat shock proteins in lesions and myelin*. J Neuropathol Exp Neurol, 1997. **56**(6): p. 664–72.
28. Bonetti, B., et al., *Activation of NF-kappaB and c-jun transcription factors in multiple sclerosis lesions. Implications for oligodendrocyte pathology*. Am J Pathol, 1999. **155**(5): p. 1433–8. https://doi.org/10.1016/s0002-9440(10)65456-9.
29. Achiron, A., et al., *Impaired expression of peripheral blood apoptotic-related gene transcripts in acute multiple sclerosis relapse*. Ann N Y Acad Sci, 2007. **1107**: p. 155–67. https://doi.org/10.1196/annals.1381.017.
30. de, J.G.-G.J., et al., *Decreased serum levels of sCD40L and IL-31 correlate in treated patients with Relapsing-Remitting Multiple Sclerosis*. Immunobiology, 2018. **223**(1): p. 135–141. https://doi.org/10.1016/j.imbio.2017.10.001.
31. Håkansson, I., et al., *Neurofilament levels, disease activity and brain volume during follow-up in multiple sclerosis*. J Neuroinflammation, 2018. **15**(1): p. 209. https://doi.org/10.1186/s12974-018-1249-7.
32. Huang, J., et al., *Inflammation-related plasma and CSF biomarkers for multiple sclerosis*. Proc Natl Acad Sci U S A, 2020. **117**(23): p. 12952–12960. https://doi.org/10.1073/pnas.1912839117.
33. Mahler, M.R., et al., *Multiplex assessment of cerebrospinal fluid biomarkers in multiple sclerosis*. Mult Scler Relat Disord, 2020. **45**: p. 102391. https://doi.org/10.1016/j.msard.2020.102391.
34. Nestor, C.E., et al., *DNA methylation changes separate allergic patients from healthy controls and may reflect altered CD4 + T-cell population structure*. PLoS Genet, 2014. **10**(1): p. e1004059. https://doi.org/10.1371/journal.pgen.1004059.
35. Ferreira, M.A., et al., *Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology*. Nat Genet, 2017. **49**(12): p. 1752–1757. https://doi.org/10.1038/ng.3985.
36. Dreymueller, D., S. Uhlig, and A. Ludwig, *ADAM-family metalloproteinases in lung inflammation: potential therapeutic targets*. Am J Physiol Lung Cell Mol Physiol, 2015. **308**(4): p. L325-43. https://doi.org/10.1152/ajplung.00294.2014.
37. Poole, A., et al., *Dissecting childhood asthma with nasal transcriptomics distinguishes subphenotypes of disease*. J Allergy Clin Immunol, 2014. **133**(3): p. 670-8 e12. https://doi.org/10.1016/j.jaci.2013.11.025.
38. Persson, H., et al., *Transcriptome analysis of controlled and therapy-resistant childhood asthma reveals distinct gene expression profiles*. J Allergy Clin Immunol, 2015. **136**(3): p. 638–48. https://doi.org/10.1016/j.jaci.2015.02.026.

39. Enomoto, Y., et al., *Tissue remodeling induced by hypersecreted epidermal growth factor and amphiregulin in the airway after an acute asthma attack*. J Allergy Clin Immunol, 2009. **124**(5): p. 913–20 e1-7. https://doi.org/10.1016/j.jaci.2009.08.044.
40. Heikamp, E.B., et al., *The AGC kinase SGK1 regulates TH1 and TH2 differentiation downstream of the mTORC2 complex*. Nat Immunol, 2014. **15**(5): p. 457–64. https://doi.org/10.1038/ni.2867.
41. Murray, J.T., et al., *Exploitation of KESTREL to identify NDRG family members as physiological substrates for SGK1 and GSK3.* Biochem J, 2004. **384**(Pt 3): p. 477 – 88. https://doi.org/10.1042/BJ20041057.
42. Liu, C.S.C., et al., *Cutting Edge: Piezo1 Mechanosensors Optimize Human T Cell Activation*. J Immunol, 2018. **200**(4): p. 1255–1260. https://doi.org/10.4049/jimmunol.1701118.
43. Solis, A.G., et al., *Mechanosensation of cyclical force by PIEZO1 is essential for innate immunity*. Nature, 2019. **573**(7772): p. 69–74. https://doi.org/10.1038/s41586-019-1485-8.
44. Purwar, R., et al., *Resident memory T cells (T(RM)) are abundant in human lung: diversity, function, and antigen specificity*. PLoS One, 2011. **6**(1): p. e16245. https://doi.org/10.1371/journal.pone.0016245.
45. Leath, T.M., M. Singla, and S.P. Peters, *Novel and emerging therapies for asthma.* Drug Discov Today, 2005. **10**(23–24): p. 1647-55. https://doi.org/10.1016/S1359-6446(05)03646-9.
46. Eraslan, B., et al., *Quantification and discovery of sequence determinants of protein-per-mRNA amount in 29 human tissues*. Mol Syst Biol, 2019. **15**(2): p. e8513. https://doi.org/10.15252/msb.20188513.
47. Jovanovic, M., et al., *Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens*. Science, 2015. **347**(6226): p. 1259038. https://doi.org/10.1126/science.1259038.
48. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner*. Bioinformatics, 2013. **29**(1): p. 15–21. https://doi.org/10.1093/bioinformatics/bts635.
49. BC., T., *Homo.sapiens: Annotation package for the Homo.sapiens object. R package version 1.3.1..* 2015.
50. Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python*. JMLR, 2011. **12**: p. 2825–2830.
51. Tibshirani, R., *Regression Shrinkage and Selection via the Lasso*. Journal of the Royal Statistical Society. Series B (Methodological), 1996. **58**(1): p. 267–288.

Figures

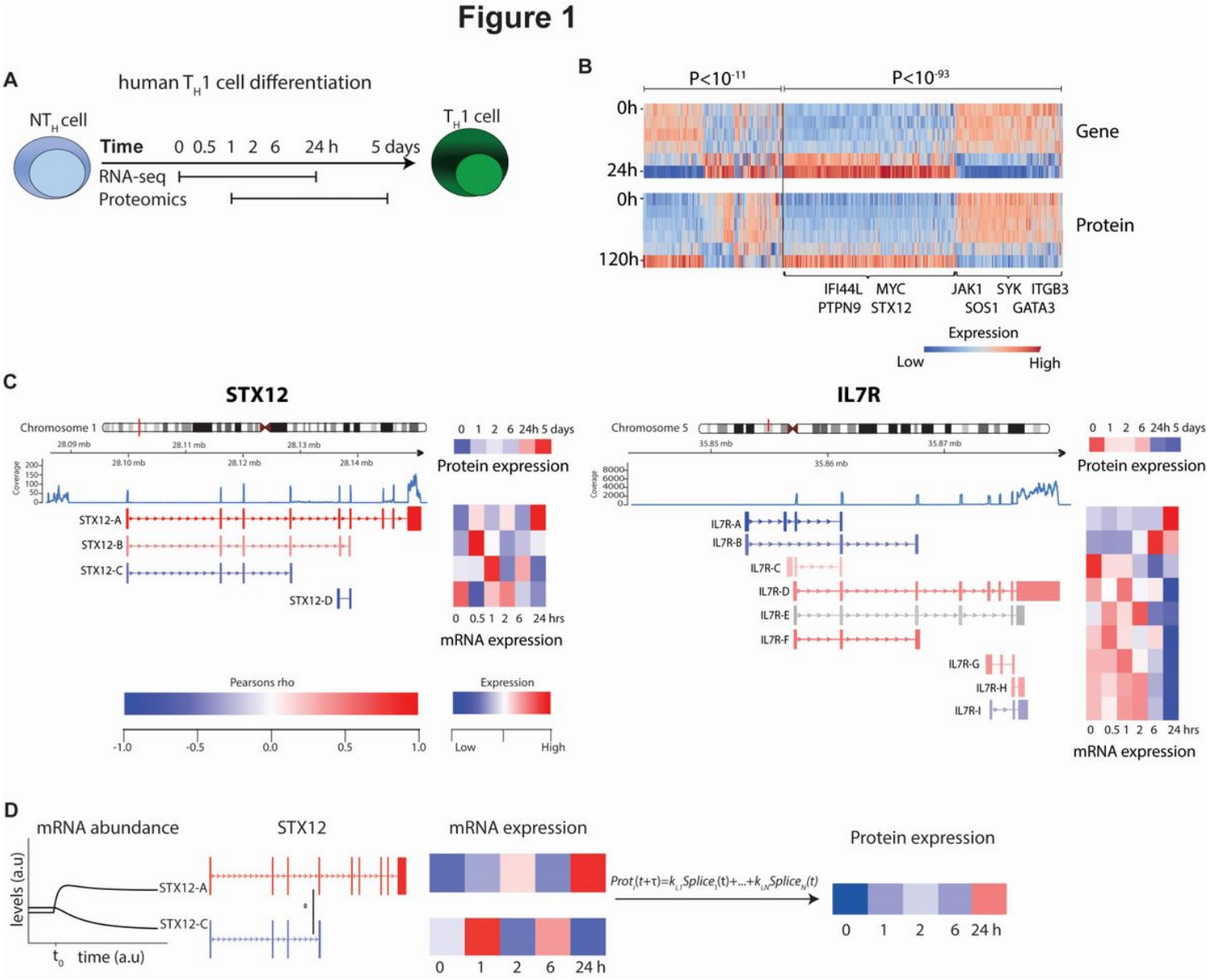

Figure 1

RNA-Seq and mass-spectrometry analysis of TH1 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

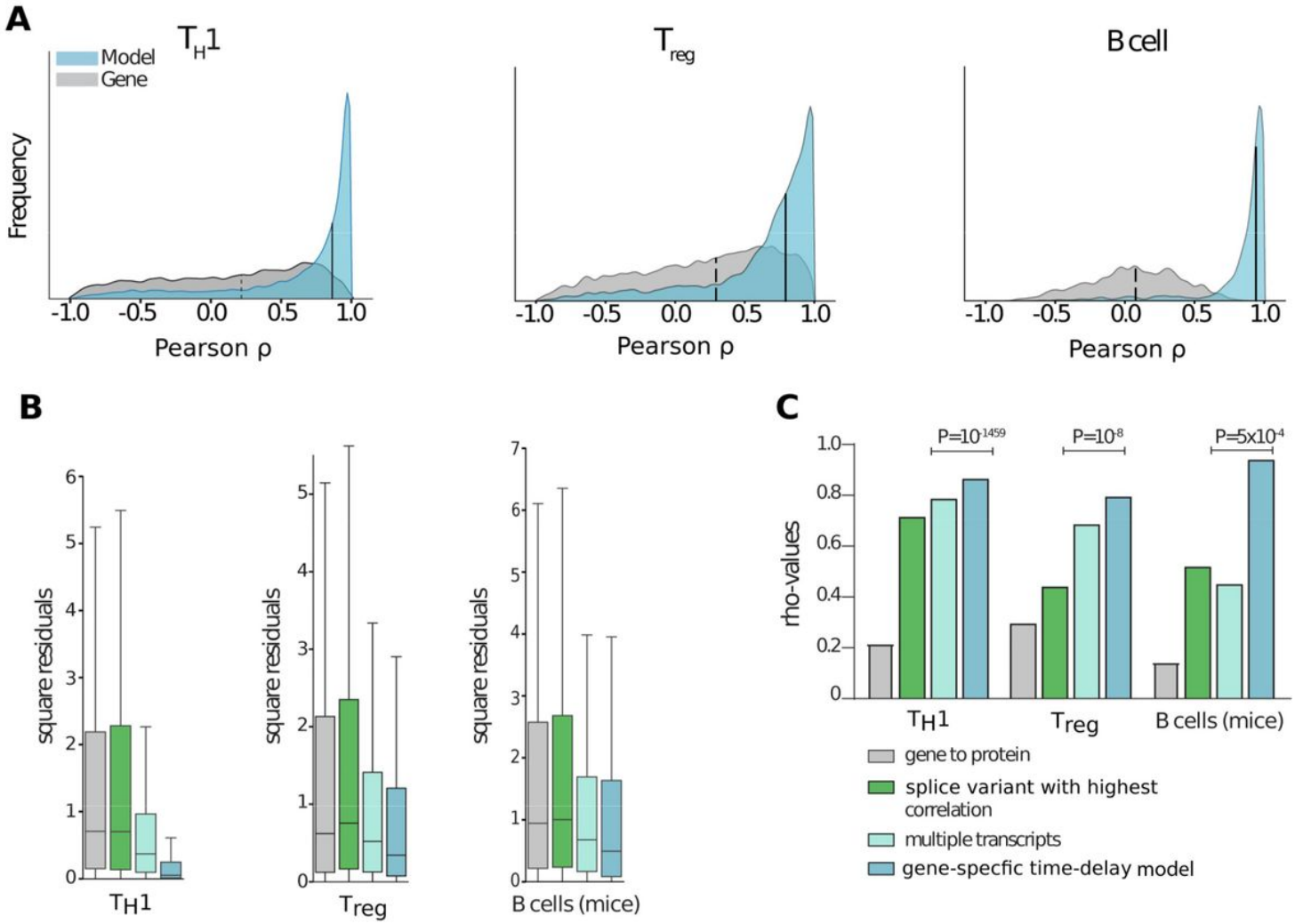

Figure 2

Multiple transcripts and time-delays increased mRNA and protein correlations significantly in multiple cell-types. (A) Gene/protein Pearson correlations in TH1 (left), Treg (middle), and murine B-cell (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 TH1, Treg and murine B-cells, respectively). Only cross-validated protein predictions are shown for the protein model 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 mRNA with increasing protein abundance correlations. P-values were derived from predictions using leave-one-out cross-validation.

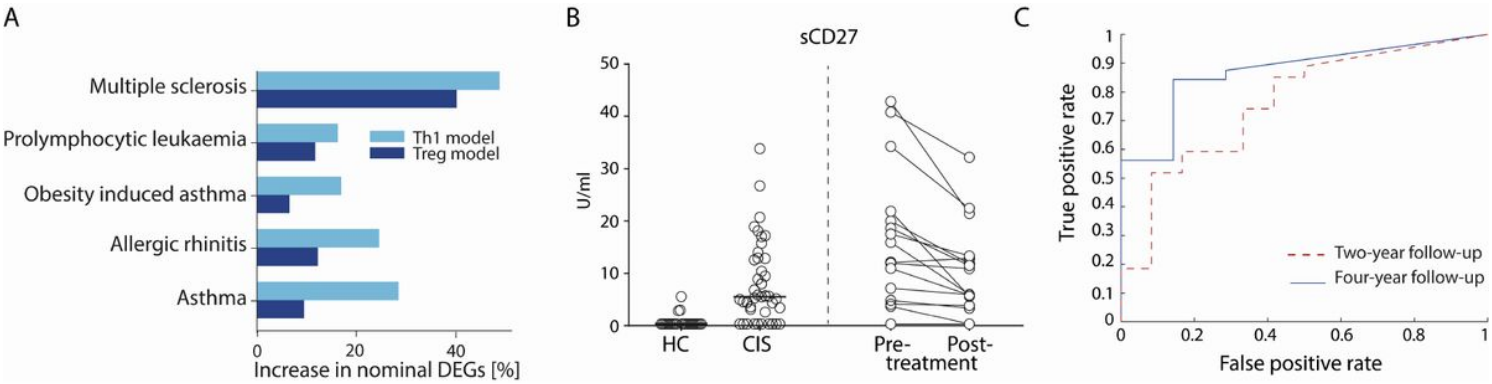

Figure 3

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 TH1 (light blue) and Treg (dark blue) models showed higher fraction of nominally significant genes than that of normal differential gene expression tests. (B) Measurement of actual protein levels of the predicted proteins in patients with early MS (clinically isolated syndrome (CIS)) vs healthy controls (HC) and pre vs post one-year treatment with Natalizumab. sCD27 was measured in cerebrospinal fluid (CSF) using ELISA. Left plots show healthy controls vs CIS where patients with no evidence of disease activity (NEDA) at four years are shown as 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.

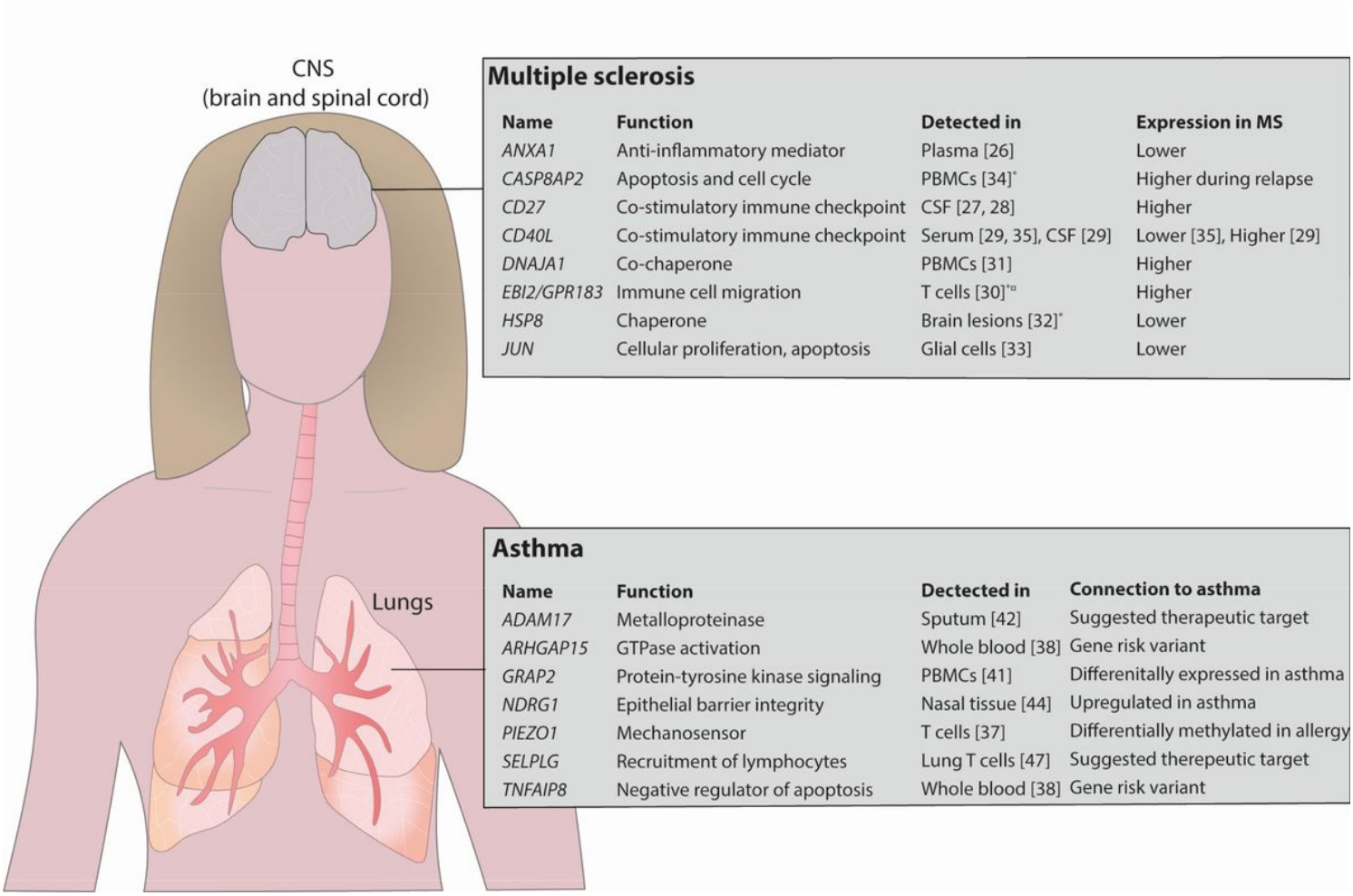

Figure 4

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.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementinformation.docx
- TableS1.xlsx
- TableS2.xlsx
- TableS3.xlsx
- TableS4.xlsx
- TableS5.xlsx
- TableS6.xlsx