EXTENDED REPORT

Multi-OMICS analyses unveil *STAT1* as a potential modifier gene in mevalonate kinase deficiency

Raphael Carapito,[1,2,3,4,5] Christine Carapito,[3,6] Aurore Morlon,[7] Nicodème Paul,[1,2,3,4] Alvaro Sebastian Vaca Jacome,[6] Ghada Alsaleh,[1,3,4] Véronique Rolli,[1,2,3,4,5] Ouria Tahar,[1,2,3,4,5] Ismail Aouadi,[1,2,3,4,5] Magali Rompais,[6] François Delalande,[6] Angélique Pichot,[1,2,3,4] Philippe Georgel,[1,2,3,4] Laurent Messer,[8] Jean Sibilia,[1,3,4,9] Sarah Cianferani,[3,6] Alain Van Dorsselaer,[3,6] Seiamak Bahram[1,2,3,4,5]

ABSTRACT

Objectives The objective of the present study was to explain why two siblings carrying both the same homozygous pathogenic mutation for the autoinflammatory disease hyper IgD syndrome, show opposite phenotypes, that is, the first being asymptomatic, the second presenting all classical characteristics of the disease.

Methods Where single omics (mainly exome) analysis fails to identify culprit genes/mutations in human complex diseases, multiomics analyses may provide solutions, although this has been seldom used in a clinical setting. Here we combine exome, transcriptome and proteome analyses to decipher at a molecular level, the phenotypic differences between the two siblings.

Results This multiomics approach led to the identification of a single gene—*STAT1*—which harboured a rare missense variant and showed a significant overexpression of both mRNA and protein in the symptomatic versus the asymptomatic sister. This variant was shown to be of gain of function nature, involved in an increased activation of the Janus kinase/signal transducer and activator of transcription signalling (JAK/STAT) pathway, known to play a critical role in inflammatory diseases and for which specific biotherapies presently exist. Pathway analyses based on information from differentially expressed transcripts and proteins confirmed the central role of STAT1 in the proposed regulatory network leading to an increased inflammatory phenotype in the symptomatic sibling.

Conclusions This study demonstrates the power of a multiomics approach to uncover potential clinically actionable targets for a personalised therapy. In more general terms, we provide a proteogenomics analysis pipeline that takes advantage of subject-specific genomic and transcriptomic information to improve protein identification and hence advance individualised medicine.

INTRODUCTION

Mevalonate kinase (MVK) deficiency is a recessive autoinflammatory disease caused by mutations in the *MVK* gene. Its clinical picture includes recurring episodes of high fever and a variety of unspecific symptoms, for example, cephalalgia, joint and/or abdominal pain, diarrhoea, skin rash,[1 2] which can be regrouped in several syndromic entities including the severe form known as mevalonic aciduria and the milder phenotype known as hyperimmunoglobulinaemia D and periodic fever syndrome (HIDS), characterised predominantly by recurrent fever with inflammatory symptoms.

We recently reported two sisters (42 and 45 years of age) that despite being both homozygous for the p.V377I mutation in the *MVK* gene (which is found in 90% of patients with HIDS) did present radically opposite phenotypes, that is, one exhibited inflammatory symptoms while the other was disease-free.[3] Briefly, the one sister (subject S2) presented polyarthralgies involving fingers, wrists and ankles, central allodynia and hyperalgesia, small cervical infracentimetric painless lymphadenopathies, optic neuritis, frequent right upper abdominal pains, nausea, vomiting and disturbed night sleep. She presented no evidence for an autoimmune disease. Despite the absence of periodic fever episodes, known pathogenic *MVK* gene mutations were identified. The HIDS diagnosis was confirmed by an abnormal level of polyclonal IgD (690 UI/mL) and IgA (5.82 U/mL), a lowered MVK enzymatic activity in lymphocytes (0.3 μkat/kg prot; cf. 2.5<normal range <7.8 μkat/kg) and urinary mevalonic acid concentration twice above normal values (3.1 and 4.1 mmol/mol creatinine). In contrast, although she carried the same homozygous p.V377I mutation, a similar decrease in MVK enzyme activity (0.3 μkat/kg prot) and elevated serum levels of IgD (140 U/mL) and IgA (2.5 U/mL), the other sister (subject S1) had no history of disease or inflammatory signs.

The obvious hypothesis explaining this differential clinical symptomatology in this family being the existence of a modifier gene, we first turned to family exome sequencing and differential analyses between the two sisters in order to uncover such gene(s). Having failed to achieve this goal, we switched to multiomics analyses (for a review see Hasin *et al*[4]). Over the last few years, there has been indeed a massive expansion of large-scale genomics, transcriptomics and proteomics data that has facilitated the identification of potential markers for disease prognosis, diagnosis and treatment. Notwithstanding these developments, there has been nevertheless a lag in successfully implementing multiomics approaches to a single patient or family, with potential direct benefit to the patient. Here, through a multi(tri)omics analysis, we decipher at a molecular level, the phenotypic differences between patient S2 and her asymptomatic sister S1.

METHODS

Subjects and samples
Subjects reported in this study were members of a single non-consanguineous family of European descent (online supplementary figure S1). Patient S2 was diagnosed as Hyper IgD syndrome, whereas all other members of the family (parents, one sister and one brother) were healthy.[3] All participants gave their informed consent and the study was approved by Strasbourg University Hospitals institutional review board (immune-based diseases project; DC-2013–1911).

Cell culture
Peripheral blood mononuclear cells (PBMC) were isolated by density gradient centrifugation from blood samples of S1 and S2 using Ficoll-Paque. PBMC (10^6 cells) were seeded in 24-well plates in RPMI1640 supplemented with fetal calf serum (FCS), penicillin, streptomycin, amphotericin B (all from Invitrogen, Cergy-Pontoise, France) and stimulated with lipopolysaccharide (LPS) (1 µg/mL; from Salmonella *Abortus equi*; Sigma, Saint-Quentin-Fallavier, France) for 3 or 6 hours at 37°C in a humidified atmosphere containing 5% CO_2. The *STAT1* deficient-U3A cell line (ECACC reference 12021503) was used for recombinant expression of wild-type (wt) and mutated STAT1 proteins. U3A cells were cultivated in Dulbecco's Modified Eagle Medium (DMEM, ThermoFisher Scientific, Waltham, Massachusetts, USA) supplemented with 10% FCS, penicillin and streptomycin.

Exome sequencing
Whole exome sequencing was performed in both parents and the two sisters (S1 asymptomatic and S2 symptomatic). Genomic DNA was extracted and purified from peripheral blood using standard protocols. Exome capture was performed with the SureSelect 50 MB Target Enrichment System V.3 (Agilent Technologies, Santa Clare, California, USA) according to the manufacturer's protocol. Prepared libraries were subjected to 75 bp/35 bp paired-end sequencing on the SOLiD 5500 platform (ThermoFisher Scientific, Waltham, Massachusetts, USA). We obtained an average of 8.6 Gb of mappable sequence data per individual. Colour space reads were mapped to the hg19 reference genome with Lifescope software V.2.5.1 (ThermoFisher Scientific, Waltham, Massachusetts, USA). The mean coverage was 152-fold (median of 148-fold) and 84% of the target region was covered at a minimum of 10×. Single-nucleotide variants and small insertions and deletions were called using the GATK software[5] and algorithms of Lifescope with medium-stringency calling settings. Only variants called by both methods and covered by more than four variant reads were considered. Annotation was performed with the KGGSeq software package.[6]

RNAseq
Total RNA was extracted from $1×10^6$ cells after 3 and 6 hours of culture using TRIzol reagent according to the manufacturer's instruction (Invitrogen, Cergy-Pontoise, France). Experiments were performed in duplicate for each condition. Purified RNA was checked for integrity (RIN>7) and quantified using a Pico Total RNA Bioanalyzer kit and Agilent 2100 BioAnalyzer (Agilent Technologies, Santa Clara, California, USA). Poly(A)+RNA was purified from 1500 ng of total RNA with a Micropoly(A) Purist kit (LifeTechnologies, Carlsbad, California, USA). Barcoded library construction was performed according to the SOLiD Total RNA-Seq Kit (ThermoFisher Scientific, Waltham, Massachusetts, USA) with 30 ng of poly(A)+RNA as input material. After quantification by Qbit (Life Technologies, Carlsbad, California, USA), libraries were pooled in equimolar amounts for template beads preparation using the SOLiD EZ-bead System (ThermoFisher Scientific, Waltham, Massachusetts, USA). Finally, the prepared beads were subjected to sequencing using the SOLiD 5500 sequencer (ThermoFisher Scientific, Waltham, Massachusetts, USA) with the paired-end 75bp/35 pb workflow.

After sequencing, the reads were aligned against the reference genome (hg19) using the Lifescope software V.2.5.1 (ThermoFisher Scientific, Waltham, Massachusetts, USA). The Bioconductor R package DESeq was used to compare the expression levels of the identified genes.[7] After normalisation and hypothesis testing, differentially expressed (DE) genes were selected based using a fold-change cut-off of minimum 1.5 and an adjusted p value of maximum 0.05.

Differential proteomic experiments
Detailed analytical procedures are described in online supplementary material. Briefly, proteins were extracted from $1×10^6$ cells after 6 hours of culture with and without LPS activation. Experiments were performed in triplicate for each condition. One hundred µg proteins from each sample were separated on a 1D SDS-PAGE gel. Gel bands were systematically cut and proteins were *in-gel* trypsin digested. Differential analysis using a spectral counting approach was conducted on a NanoAcquity LC-system (Waters, Milford, Massachusetts, USA) coupled to a maXis 4G QToF mass spectrometer (Bruker Daltonics, Bremen, Germany). Liquid chromatography-mass spectrometry/mass spectrometry (LC-MS/MS) data were searched using a local Mascot server (Matrix Science, London, UK) against personalised databases for each subject containing all human entries extracted from the UniProtKB-SwissProt reference database (40 654 target entries with isoforms) and specific exome-derived variants (Personalised Database 1 for S1: 45 834 target entries, Personalised Database 1 for S2: 45 883 target entries). The generation of personalised protein databases is detailed in online supplementary material and in online supplementary figure S3.

Identification validation and spectral counting were performed with Scaffold software (Proteome Software, Portland, USA). Relative protein quantification and testing for differential protein expression were performed using the beta-binomial test implemented in R.[8] The acceptance criteria for statistically different protein abundance changes between two conditions were: a minimum of 4 summed spectral-counts for the three replicates in at least one of the conditions to be compared, a p value lower than 0.05 and a fold change higher than 1.5 (except when the summed spectral-count equals 0 in one of the conditions to be compared).

Proteomic experiments for deep proteome coverage and specific variants identification
Detailed analytical procedures are described in online supplementary material. Briefly, samples were prepared as described above but analysed on a more recent nanoLC-MS/MS coupling in order to increase depth of the proteome coverage. These experiments were conducted on a NanoAcquity LC-system (Waters, Milford, Massachusetts, USA) coupled to a TripleTOF 5600+mass spectrometer (Sciex, Framingham, USA). LC-MS/MS data were searched using Mascot against personalised databases containing specific sequence variants inferred for exome

sequencing on the one hand (personalised databases 1) and containing new splice variants inferred from RNAseq data on the other hand (personalised databases 2). For this purpose, we generated a list of genomic variants and splice junctions from exome and RNAseq datasets, respectively. Based on this list, a modified UniProtKB protein database was constructed containing the identified variant protein sequences in addition to all UniProtKB-SwissProt database entries, including presently known isoforms. Between the two search rounds, peak-list files were processed using an in-house developed software

tool (Recover module of MSDA, https://msda.unistra.fr[9]) and identifications were validated using Proline software[10] (http://proline.profiproteomics.fr/). The overall identification pipeline is outlined in figure 1B. Sixty-nine identified single amino acid variants (SAAVs) were further validated by comparing their chromatographic and MS/MS fragmentation characteristics to heavy labelled synthetic peptides analysed and fragmented in identical conditions. All fragmentation spectra of light and heavy peptide pairs are provided in online supplementary figure S1.
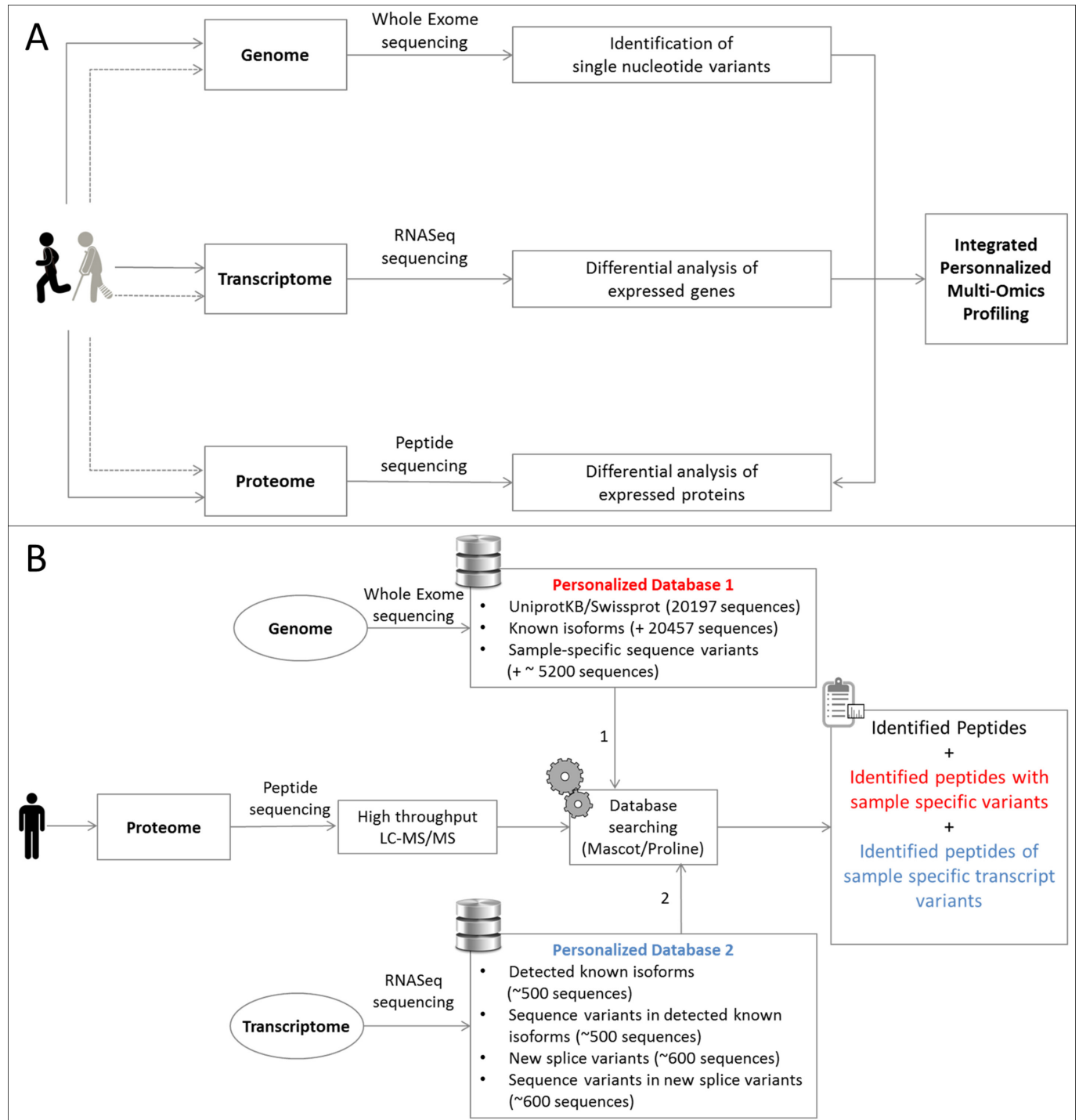

Figure 1 Experimental multiomics strategy. (A) Schematic view of the global strategy combining genomic, transcriptomic and proteomic data. (B) Detailed strategy of the interpretation of proteomic data making use of genomics (personalised database 1) and transcriptomics (personalised database 2) information.

Integrative omics data analyses

The first step of the integrative omics analysis was to assess the correlation of the proteomics and RNAseq datasets. For this purpose, we first applied the voom transformation to the RNAseq read counts in order to convert the counts to log-counts per million with associated precision weight.[11] Then the RNAseq data were analysed using limma method, which is a popular differential expression method developed for microarray data (continuous data), provided by the limma software package.[12] The same approach was used to analyse spectral count data from the proteomics dataset. The resulting datasets (RNAseq: n=24 364; proteins: n=736) were first filtered by removing rows with zero or very low counts. The filtered datasets (RNA-seq n=12 770; proteomic n=727) were then integrated using the multivariate co-inertia analysis (MCIA).[13 14] MCIA projects RNAseq and proteomic datasets, at the same time, into the same dimensional space to identify co-relationships between them. As this method requires identical number of replicates for both datasets, we generated an additional replicate for RNAseq using bootstrap method and validating beforehand that bootstrap did not cause any bias.[15]

The second analysis step consisted in combining S2-specific variants with DE genes in S2 vs S1 at both RNA and protein levels based on fold changes calculated from RNAseq and proteomic datasets, respectively (cf. RNAseq and proteomics sections of materials and methods). By doing so, we aimed at identifying genes that harboured a S2-specific variant and showed a differential expression level at both RNA and protein level as compared with S1.

Finally, identification of molecular network interactions and pathway analysis of DE genes at the RNA and protein levels was completed using the Ingenuity Pathway Analysis (Ingenuity Systems, www.ingenuity.com). The ingenuity knowledge base (genes only) with direct and indirect relationships was used and only molecules and/or relationships that had been experimentally observed in rat, mouse or human were considered.

Expression of recombinant STAT1 and IFN-γ stimulation

The human *STAT1* cDNA was PCR-amplified from HEK293 cells (ATCC CRL-1573) with the following forward and reverse primers respectively 5'-cgcaccggtatgtctcagtggtacgaacttcagca-3' and 5'-cgcttaattaacta tactgtgttcatcatactgtcgaattc-3'. It was then cloned into AgeI and PacI sites of pQCXIP (Clontech, Mountain View, California, USA). The mutated form (p.R241Q) was generated with the QuikChange Site-Directed Mutagenesis kit using the forward 5'-agtggagtggaagcagagacagagcg-3' and reverse 5'-cgctctgctgtctctgcttccactccact-3' primers, following manufacturer's recommendations (Stratagene, San Diego, California, USA). The pQCXIP plasmids harbouring the wt or the mutated *STAT1* were transiently transfected into the STAT1 deficient-U3A cell line with Lipofectamine 3000 according to manufacturer's instructions (Invitrogen, Carlsbad, California, USA). Two days after transfection, cells were stimulated with 2000 IU/mL of interferon-γ (IFN-γ) (RD Systems, Minneapolis, Minnesota, USA) during 30 min. After stimulation, proteins were extracted with NP-40 lysis buffer (150 mM sodium chloride, 50 mM Tris pH8, 1% NP-40) complemented with the cOmplete and phosSTOP protease and phosphatase inhibitor cocktail (Roche, Basel, Switzerland). Proteins were quantified using the Bradford assay according to manufacturer's instructions (Bio-Rad Laboratories, Hercules, California, USA).

Quantitative real-time PCR

The pQCXIP plasmids harbouring the wt or the mutated STAT1 forms were transiently transfected into the STAT1 deficient-U3A cell line with Lipofectamine 3000 according to manufacturer's instructions (Invitrogen, Carlsbad, California, USA). Two days after transfection, cells were stimulated with 1000 IU/mL of IFN-γ during 8 hours. Total RNA was extracted with TRIzol reagent (Invitrogen, Carlsbad, California, USA). Isolated RNA was reverse transcribed with iScript Reverse Transcription supermix (Bio-Rad Laboratories, Hercules, California, USA). PCR was performed with SSO advanced universal SYBRgreen supermix (Bio-Rad Laboratories, Hercules, California, USA) and run on a ABI7000 Real-Time PCR system (ThermoFisher Scientific, Waltham, Massachusetts, USA). PCR conditions were as follows: 10 min at 96°C followed by 45 cycles of 95°C for 10 s and 60°C for 30 s. Primers used were 5'-CTCTAAGTGGCA TTCAAGGAGTACCT-3' (forward) and 5'-TCTCAACACGTG GACAAAATTGGC-3' (reverse). The *GAPDH* gene was used as a normalising gene and was amplified with the forward primer 5'-GGTGAAGGTCGGAGTCAACGGA-3' and the reverse primer 5'-GAGGGATCTCGCTCCTGGAAGA-3'.

Luciferase reporter assay

*GAS (*Gamma interferon activation site) sequences were inserted into the multiple cloning site of the pMCS Tluc16 plasmid (ThermoFisher Scientific, Waltham, Massachusetts, USA). This reporter plasmid was co-transfected with the pQCXIP plasmids harbouring the wt or the mutated STAT1 or the mock vector into STAT1-deficient U3A cells. Transfections were done in 96-well plates using the Lipofectamine 3000 Reagent and following manufacturer's instruction (Invitrogen, Carlsbad, California, USA). Six hours after transfection, growth medium was replaced. One day after transfection, cells were stimulated with 1000 IU/mL of IFNγ during 16 hours and subjected to luciferase assays with TurboLuc one-Step Glow Assay Kit (ThermoFisher Scientific, Waltham, Massachusetts, USA). Experiments were performed in triplicate and data are expressed as fold induction with respect to non-stimulated cells.

Quantification of proinflammatory cytokines

Cytokines were quantified in the supernatant of PBMCs after 6 hours of stimulation with LPS (1 µg/mL, *Salmonella abortus equi* LPS from Sigma). The cytokines IL-6, IL-1β, TNF-α were quantified by ELISA following the manufacturer's instruction (R&D Systems, Minneapolis, Minnesota, USA). The results are presented as mean values±SD in pg/mL. A two-tailed Mann-Whitney test was used to compare two independent groups using the GraphPad software (San Diego, California, USA).

RESULTS

Exomes

The exome of the asymptomatic subject S1, the symptomatic subject S2, their mother and father were sequenced (online supplementary figure S1). On average 49 420 SNPs (median of 49 351) and 1867 indels (median of 1869) were identified in each subject. For SNPs, we focused only on the average 8851 protein altering events (median of 8804), that is, missense, frameshift, splicing, stop loss and stop gain. In the case of indels, only frameshift, non-frameshift, stop loss, stop gain and splicing events were taken into account, which correspond to an average number of 182 (median of 183). To identify the candidate genes, the initial list of 8811 protein altering variants of the symptomatic subject S2 (proband) were filtered by

excluding variants shared with her asymptomatic sister S1. As both rare and common variants associated to the rare homozygous mutation p.V377I present in the two subjects could be associated with the phenotype, no filtering was performed using variant frequency databases. In addition, as it has been reported that homozygous p.V377I carriers could be milder or asymptomatic,[16] we focused only on S2-specific events that represented a total of 1135 variants in 876 different genes.

The complete list of variants is available in the online supplementary table S1.

Proteomes

After extraction of PBMCs from whole blood and culture without (basal state) or with LPS (activated state) during 6 hours, proteins were isolated and analysed using bottom-up nanoLC-MS/MS-based proteomics workflows. In this proteome
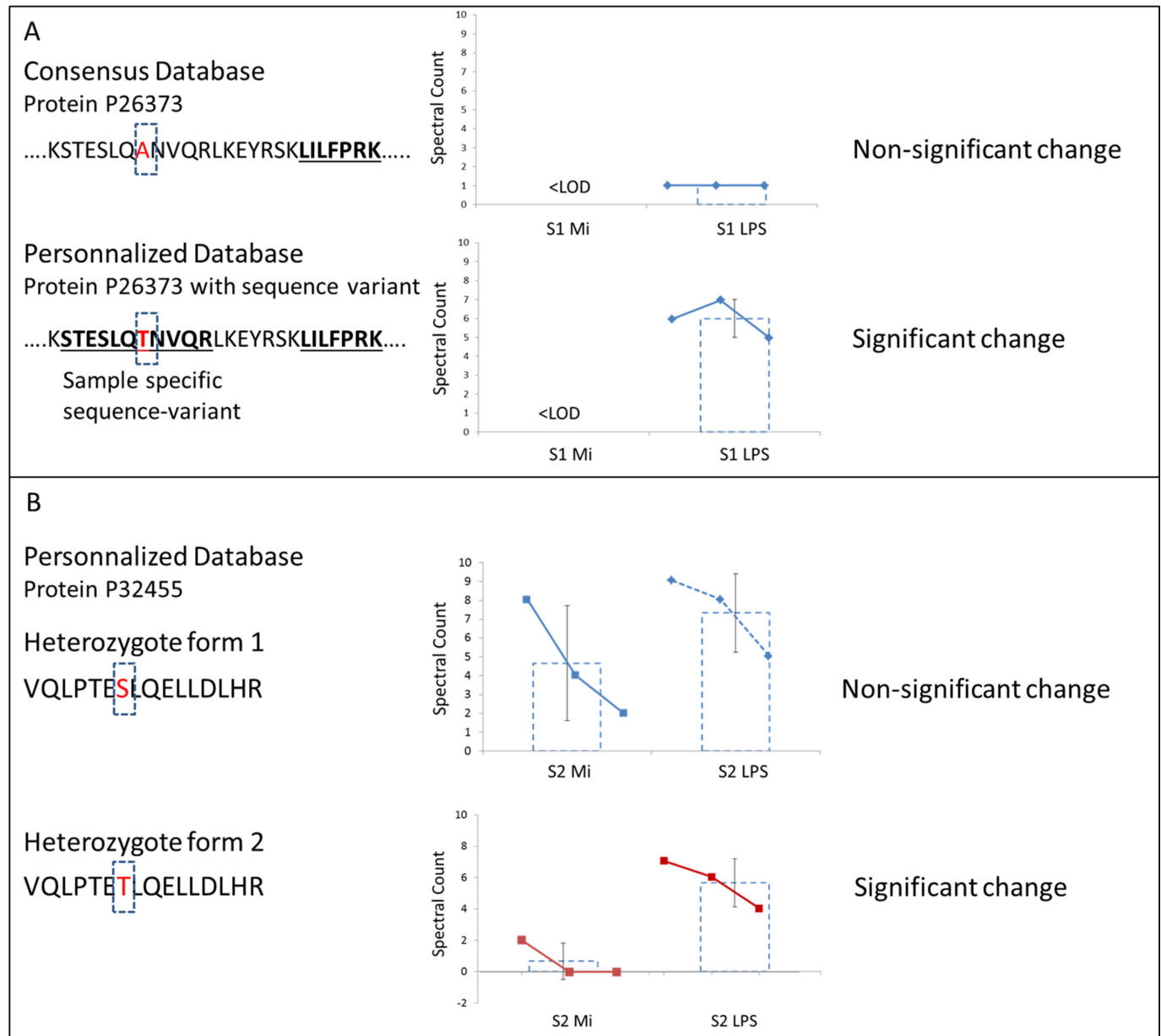

Figure 2 Benefits of using personalised databases for protein identification and quantification. (A) A sequence stretch of 60S ribosomal protein L13 (P26373) is given. Peptides identified in searches using the consensus database (UniProtKB-SwissProt) and the personalised databases (including subject-specific sequence variants) are underlined. When using the consensus database only a single peptide was identified and the spectral count values for protein P26373 show a non-significant change in abundance when comparing the basal to the activated state. However, when using the personalised database, an additional tryptic peptide including a subject-specific variant was identified (amino acid change from Alanine to Threonine at position 112). The relative quantification using spectral count is improved as the sequence coverage is greater and a significant overexpression of the protein could be detected. (B) In this example, two sequence variants of the Guanylate-binding protein 1 (P32455) are present in the personalised database 1 of S2. Both peptides, one with Threonine and one with Serine at position 349, were identified. This is the unambiguous proof that two heterozygote variants of the same gene are expressed. The spectral count-based relative quantification shows that one of the heterozygote forms is overexpressed when comparing the basal to the activated state. This example demonstrates that allele-specific quantification is possible at protein level.
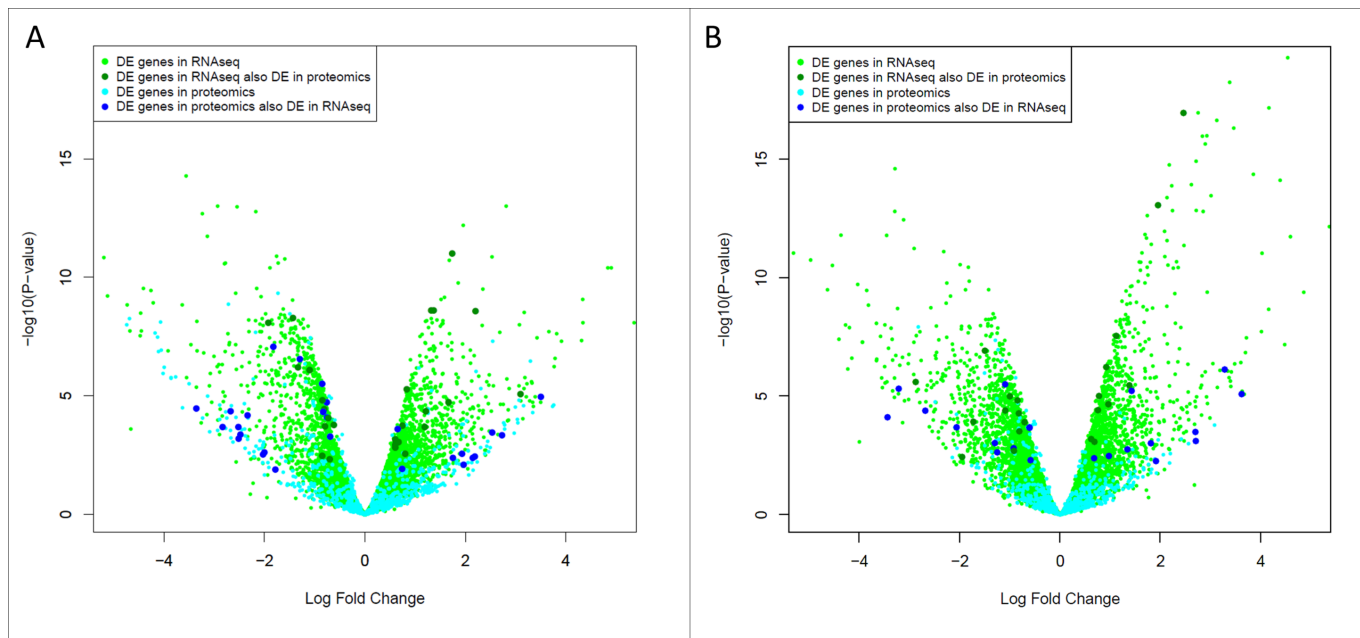
Figure 3 Volcano plots of DE genes in RNAseq and proteomics analyses. Statistical significance (−log10 p value) is plotted against log2 fold change for either basal (without LPS, Panel A) and activated (with LPS, Panel B) states. Data were selected at the cut-off values adjusted p<0.05 and fold change >1.5. For this analysis, RNAseq data at 3 and 6 hours were pooled. DE, differentially expressed; LPS, lipopolysaccharide.

analysis, we identified a total of 8335 distinct peptides among all samples, corresponding to 244 450 assigned MS/MS spectra in an assembly of 1428 unique protein groups with a false discovery rate below 1% at both peptide and protein levels (see online supplementary table S2). Protein abundances were estimated using a spectral count approach where the total number of MS/MS spectra acquired for the peptides of a given protein were used as a proxy for the protein quantities. Based on a minimal spectral count requirement of 4 (for the replicates in at least one of the conditions to be compared), 736 unique protein groups could be confidently quantified across samples. These quantitative results were aligned with those obtained at the transcriptome level using an RNAseq approach as described below (figure 1A, online supplementary table S2). Notably, in order to take advantage of the unique availability of multilevel omics data acquired on identical individuals/samples, the proteomics data interpretation strategy was adapted to use personalised protein databases including subject-specific information inferred from exome and transcriptome data. Therefore, for each subject, the human reference UniProtKB-SwissProt database was modified to take into account, on one hand, non-synonymous SNPs derived from exome sequencing (personalised databases 1) and on the other, the alternative transcripts uncovered from RNASeq experiments (personalised databases 2) (figure 1B, online supplementary material). Then, in order to further increase the depth of proteome coverage, improve overall numbers of identifications and leverage benefits of using personalised databases, all four samples were reinjected on a later generation nanoLC-MS/MS coupling, namely a TripleTOF 5600+Q TOF instrument (Sciex). At the end, the complete dataset searched against the personalised databases enabled the unambiguous identification of 100 and 108 SAAVs for S1 and S2, respectively and two new splice variants common to S1 and S2 (online supplementary table S3). In order to further validate these SAAVs identifications, we synthesised a random selection of 69 SAAVs-containing peptides in heavy labelled form. These peptides were mixed and analysed using the same LC-MS/MS conditions as the samples. Both

chromatographic and MS/MS fragmentation characteristics were compared and allowed ultimate validation of all SAAVs. Fragmentation spectra of light and heavy peptide pairs are provided in online supplementary figure S2. Finally, additional advantages of using personalised databases are illustrated in figure 2 where we show that using an incomplete consensus reference database can lead to errors in quantification (this has far ranging implications beyond the present manuscript). Indeed, in the example of figure 2A, reference database searches led to a non-significant change in abundance while the personalised database search enabled the identification of an additional SAAV peptide that allowed concluding for a significant overexpression. In figure 2B, we demonstrate that heterozygote variants of a given gene can be unequivocally identified and even individually quantified for allele-specific expression analysis.

Differentially expressed genes and proteins

DE genes were analysed at the RNA and protein levels. Figure 3 shows the statistical significance of differential transcript and protein levels by volcano plots along with respective fold changes. A total of 1661 transcripts and 190 proteins were DE in S2 vs S1 at the basal state (without LPS) using a cut-off value for adjusted p value of 0.05 and a minimum fold change of 1.5. Twenty-seven genes were shared by the protein and RNAseq datasets. In the activated state (with LPS), we identified 2502 and 69 genes that were DE in S2 vs S1 at the transcripts and protein levels, respectively. Twenty-one genes were common to both datasets. The complete lists of genes with significantly different expression levels at the RNA or protein levels are presented in online supplementary table S2.

Integrative analysis of RNAseq and proteome data

MCIA was applied to RNAseq and proteomic data to assess the correlation of the two datasets. Figure 4 shows the projection of samples onto the first two principal components (PCs) of MCIA. The coordinates of RNAseq and proteomic measurements for
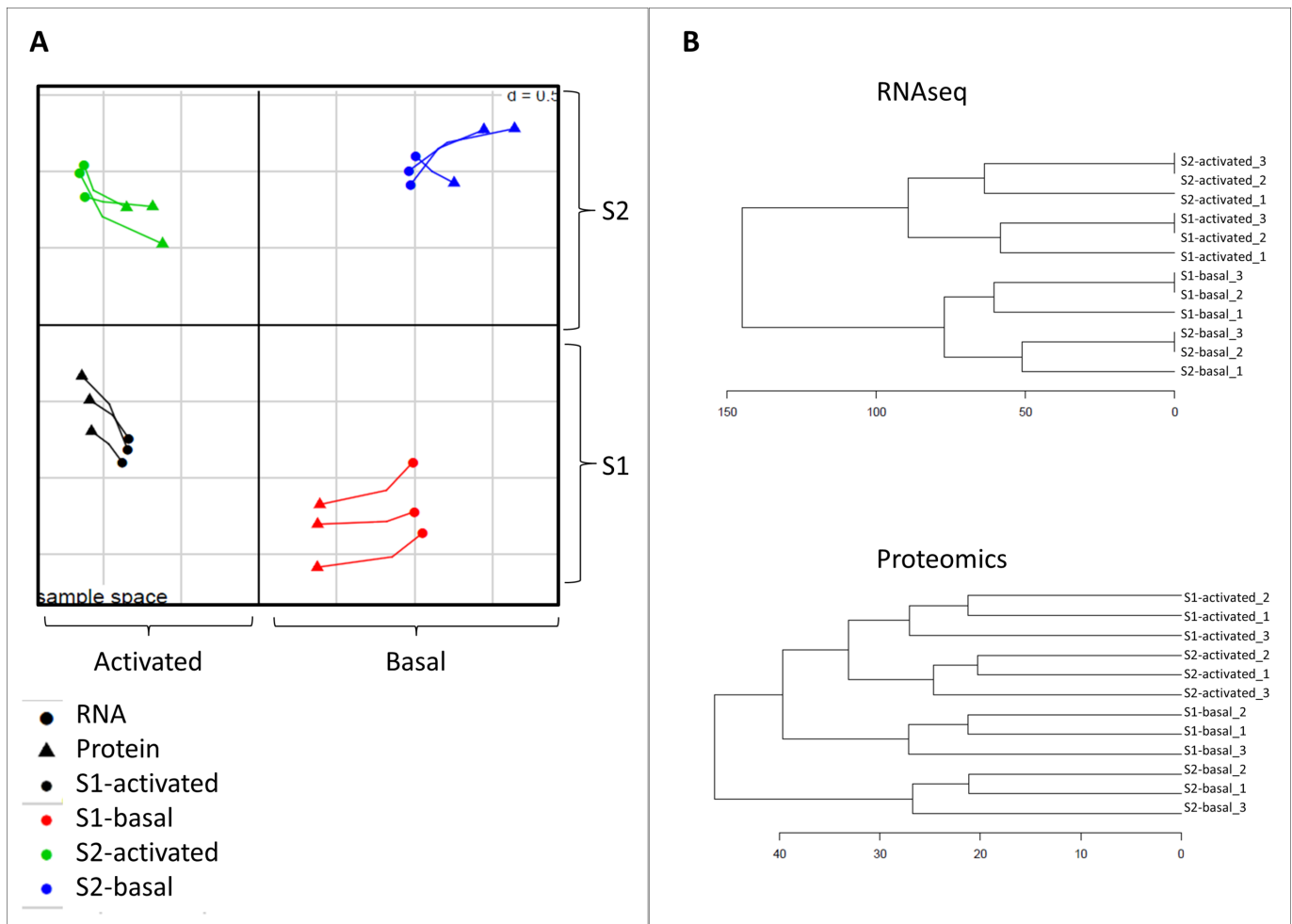
Figure 4 Integrative analysis of RNAseq and proteomics data. (A) MCIA of RNAseq and proteomic data. The projection of the samples on the first two PCs of MCIA is depicted. (B) Hierarchical clustering of RNA-seq and protein expression data. Dendrograms show the hierarchical clustering using Euclidean distance. MCIA, multiple co-inertia analysis; PCs, principal components.

each sample are linked by a line. The shorter the line, the higher is the level of consensus between the transcript and protein expression levels. The transcriptome and proteome profiles of the individual samples were projected close in space showing that the expression profiles of both techniques shared a high degree of concordance. The first axis of the MCIA (PC1; horizontal axis) separated the activated (negative on PC1) from the basal state (positive on PC1), suggesting that the transcriptome and proteome profiles are different in the two conditions. These results are consistent with independently performed hierarchical clustering analysis[17] (figure 4B).

Combined analysis of exome, RNAseq and proteome
Combining the lists of S2-specific non-synonymous variants, DE transcripts and proteins in S2 vs S1 led to the identification of a single gene—*STAT1*—harbouring a rare coding variant (minor allele frequency of 0.029% and 0% at heterozygous and homozygous states, respectively in GnomAD) and showing mRNA and protein abundances significantly more important in S2 than in S1 (figure 5). At basal state, mRNA and protein fold changes were 1.78 ($p=1.74\times10^{-4}$) and 11.32 ($p=2.22\times10^{-4}$), respectively. In the activated state, these values were 1.71 ($p=2.34\times10^{-4}$) for *STAT1* mRNA and 12.27 ($p=5.27\times10^{-4}$) for the protein. The variant (c.722G>A, rs146273341, NM_007315.3) is inherited from the father and causes an Arg to Gln modification

(p.R241Q, NP_009330.1) in the coiled coil domain of the protein without predicted deleterious impacts on protein function (online supplementary table S4).

Activation of the interferon-γ/STAT1 pathway
Functional pathway analyses were performed on lists of DE genes in S2 vs S1 after activation with LPS. The top canonical pathways, upstream regulators and disease/biological functions are presented in table 1. Based on RNAseq data, the top canonical pathway identified was the interferon signalling pathway with 13 and 14 out of 36 genes of the pathway that were DE at the mRNA level, 3 and 6 hours post-LPS-activation, respectively. This pathway was associated with the activation of IFN-γ as the top upstream regulator. Protein level analyses that were restricted to intracellular protein revealed activation of pathways related to immune responses including granzyme A signalling or the antigen presenting pathway. Immune-mediated pathways were further confirmed by the identification of known disease pathways of immunological, inflammatory, antimicrobial, dermatological or connective/skeletal/muscular origin (table 1).

Figure 6A,B represents the molecules of the IFN-γ network and their interaction with STAT1, which is central in this pathway. Based on RNAseq data at 3 hour post-LPS-activation, 7 proteins of the IFN-γ pathway were found to be upregulated. To further validate the activation of the IFN-γ/STAT1 pathway,
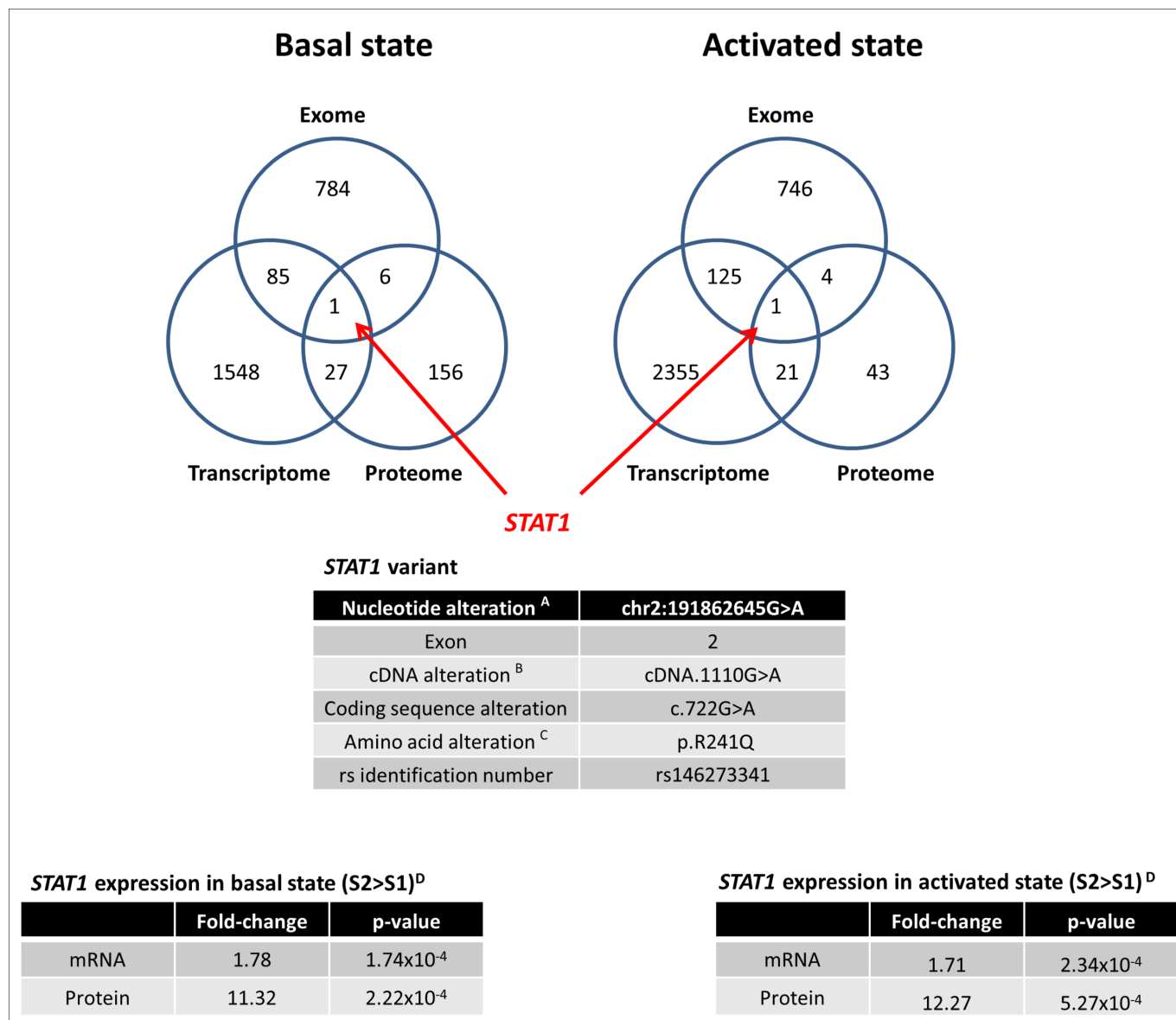
Basal state

Exome

784

85   6

1

1548   27   156

Transcriptome   **Proteome**

Activated state

Exome

746

125   4

1

2355   21   43

Transcriptome   **Proteome**

STAT1

STAT1 variant

Nucleotide alteration [A]	chr2:191862645G>A
Exon	2
cDNA alteration [B]	cDNA.1110G>A
Coding sequence alteration	c.722G>A
Amino acid alteration [C]	p.R241Q
rs identification number	rs146273341

STAT1 expression in basal state (S2>S1) [D]

	Fold-change	p-value
mRNA	1.78	1.74×10^{-4}
Protein	11.32	2.22×10^{-4}

STAT1 expression in activated state (S2>S1) [D]

	Fold-change	p-value
mRNA	1.71	2.34×10^{-4}
Protein	12.27	5.27×10^{-4}

Figure 5 Combined analysis of exome, transcriptome and proteome datasets. The figure shows Venn diagrams of genes with S2-specific variants and genes that were upregulated in S2 at the mRNA (transcriptome) and protein (proteome) levels. The tables in the lower panel indicate the details of the *STAT1* variant and the differential expression values in RNAseq and proteomics datasets in both basal and activated states. [A]Positions refer to Hg19 (GRCh37). [B]Positions refer to GenBank transcript NM_007315.3. [C]Position refers to protein accession number NP_009330.1. [D]Increased expression in S2 versus S1.

we stimulated the STAT1 deficient U3A cell line transfected with the wild-type or mutated *STAT1* (c.722G>A, p.R241Q) gene with IFN-γ. On stimulation with IFN-γ, cells transfected with the R241Q allele responded two times more strongly than those transfected with the wild-type allele, as shown by measurement of the induction of GAS (γ-activated sequence)–dependent reporter gene transcription activity (figure 7A). Accordingly, the transcription of the CXCL10 target gene was enhanced (figure 7B).

Proinflammatory cytokine response of lipopolysaccharide (LPS)-stimulated peripheral blood mononuclear cells (PBMCs)
In order to assess the impact of the *STAT1* variant in presence or absence of the *MVK* variant at the homozygous state, the cytokine response of PBMCs isolated from different members of the family was evaluated (table 2). As expected, LPS-stimulated PBMCs from the symptomatic patient S2 released large

amounts of TNF-α (10437±2804 pg/mL), IL-1β (5790±743 pg/mL) and IL-6 (24649±1408 pg/mL), while those of her asymptomatic sister S1 released lower quantities of those cytokines (4804±1251 pg/mL, 3868±709 pg/mL and 18316±1518 pg/mL for TNF-α, IL-1β and IL-6, respectively), at about the same level as those of the mother's (4804±1251 pg/mL, 3995±794 pg/mL and 14642±700 pg/mL for TNF-α, IL-1β and IL-6, respectively). Interestingly, the concentrations of cytokines produced by the PBMCs of the father, who also carries the *STAT1* variant but does not show any disease manifestation, were equivalent to those of the symptomatic patient S2 (8580±1073 pg/mL, 7388±831 pg/mL and 24727±479 pg/mL for TNF-α, IL-1β and IL-6, respectively; table 2).

DISCUSSION
Here we report a multiomics study in two siblings harbouring the same homozygous p.V377I pathogenic mutation in the *MVK*

Table 1 Pathway analysis of differentially expressed genes identified in the LPS-activated state using RNAseq and proteomics datasets

Top canonical pathways

RNAseq data 3-hour post-LPS stimulation	P value	Overlap	RNAseq data 6-hour post-LPS stimulation	P value	Overlap	Proteomics data 6-hour post-LPS stimulation	P value	Overlap
Interferon signalling	9.47×10^{-14}	36.1%—13/36	Granulocyte adhesion and diapedesis	3.18×10^{-17}	18.2%—30/165	Granzyme A signalling	3.53×10^{-06}	23.5%—4/17
Crosstalk between dendritic cells and natural killer cells	2.29×10^{-10}	16.9%—15/89	Agranulocyte adhesion and diapedesis	1.11×10^{-14}	16.0%—28/175	Antigen presentation pathway	3.62×10^{-06}	13.5%—5/37
Granulocyte adhesion and diapedesis	8.07×10^{-10}	11.5%—19/165	Interferon signalling	1.15×10^{-13}	38.9%—14/36	Caveolar-mediated endocytosis signalling	6.12×10^{-06}	8.5%—6/71
T helper cell differentiation	8.75×10^{-10}	18.8%—13/69	Communication between innate and adaptive immune cells	2.58×10^{-12}	22.0%—18/82	Integrin signalling	8.89×10^{-06}	4.2%—9/212
Activation of IRF by cytosolic pattern recognition receptors	1.90×10^{-09}	20.0%—12/60	Altered T cell and B cell signalling in rheumatoid arthritis	3.42×10^{-11}	20.5%—17/83	Cdc42 signalling	1.91×10^{-05}	5.4%—7/129

Top upstream regulators

	P value of overlap	Predicted activation		P value of overlap	Predicted activation		P value of overlap	Predicted activation
IFNG	3.46×10^{-31}	Activated	IFNG	5.09×10^{-62}	Activated	EBI3	6.48×10^{-06}	Activated
TLR3	7.38×10^{-28}	Activated	IFNL1	1.33×10^{-60}	Activated	mir-122	9.91×10^{-06}	Activated
IFNA2	1.09×10^{-27}	Activated	PRL	1.51×10^{-58}	Activated	TGM2	2.75×10^{-05}	Activated
IL1RN	4.90×10^{-27}	Inhibited	MAPK1	1.86×10^{-45}	Inhibited	RFX5	4.28×10^{-05}	Inhibited
PRL	9.67×10^{-27}	Activated	IFNA2	2.07×10^{-45}	Activated	CEBPA	8.03×10^{-05}	Activated

Top diseases and bio functions

	P value	Number of molecules		P value	Number of molecules		P value	Number of molecules
Antimicrobial response	1.33×10^{-16} –1.07×10^{-16}	35	Dermatological diseases and conditions	1.46×10^{-05} –1.75×10^{-32}	269	Dermatological diseases and conditions	1.88×10^{-02} –4.98×10^{-11}	37
Inflammatory response	2.88×10^{-04} –1.07×10^{-16}	131	Organismal injury and abnormalities	1.49×10^{-05} –1.75×10^{-32}	476	Immunological disease	1.37×10^{-02} –4.98×10^{-11}	43
Connective tissue disorders	1.49×10^{-04} –1.19×10^{-16}	79	Connective tissue disorders	2.68×10^{-06} –3.09×10^{-25}	110	Inflammatory disease	1.42×10^{-02} –4.98×10^{-11}	40
Inflammatory disease	1.99×10^{-04} –1.19×10^{-16}	113	Inflammatory disease	1.46×10^{-05} –3.09×10^{-25}	147	Organismal injury and abnormalities	1.90×10^{-02} –4.98×10^{-11}	121
Organismal injury and abnormalities	2.96×10^{-04} –1.19×10^{-16}	343	Skeletal and muscular disorders	1.44×10^{-05} –3.09×10^{-25}	143	Inflammatory response	1.57×10^{-02} –2.58×10^{-10}	50

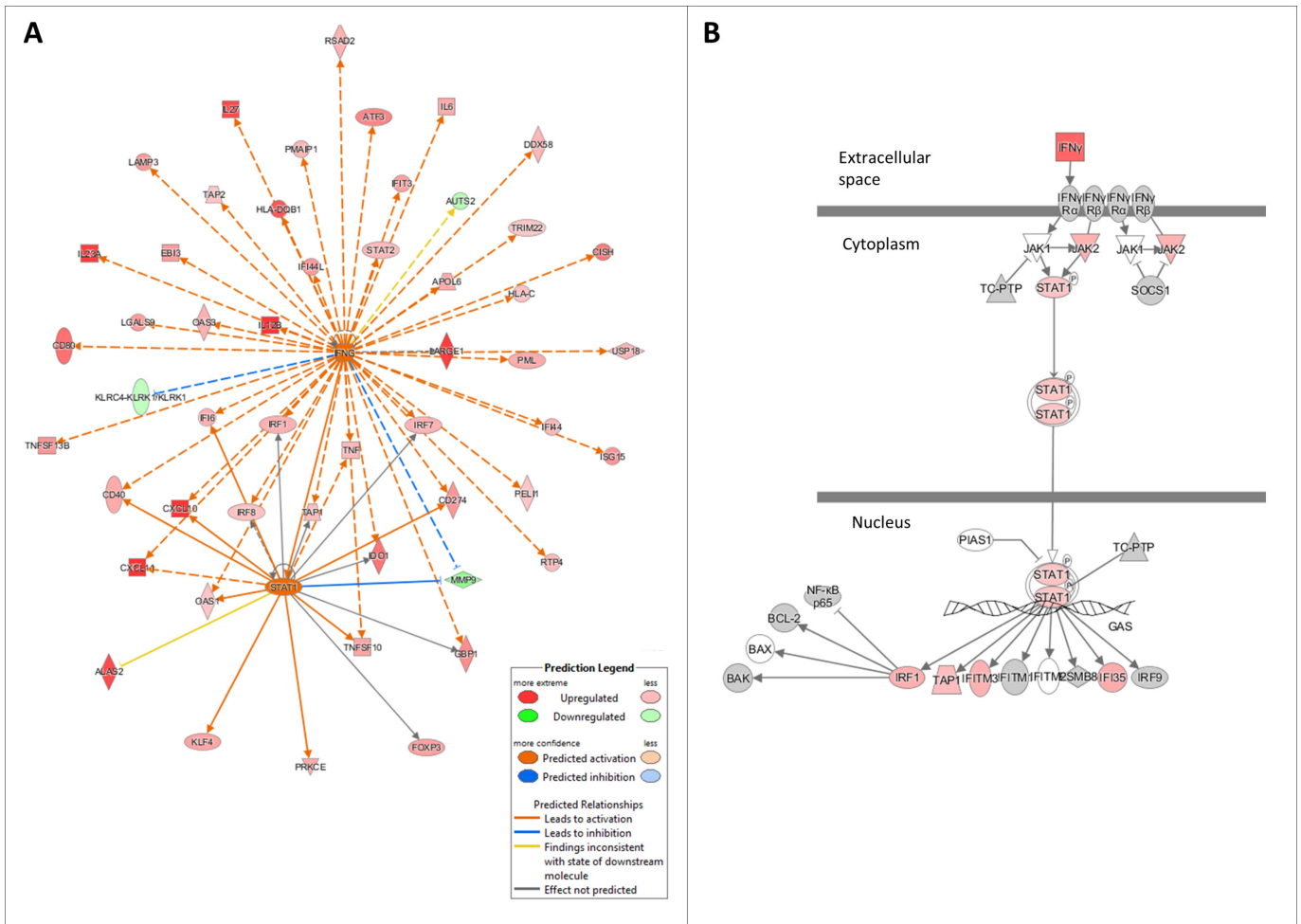IRF, interferon regulatory factor; LPS, lipopolysaccharide.

Figure 6 Activation of the interferon-γ/STAT1 pathway. (A) Functional network analysis by IPA using RNAseq data at 3 hour in the activated state. The genes involved in the top scoring pathway (ie, INFγ signalling) and their interaction with the STAT1 pathway are represented. (B) Canonical INFγ signalling pathway. The genes upregulated in the dataset are represented in red. IFN-γ, interferon-γ; IPA, ingenuity pathway analysis.

gene but presenting with opposite phenotypes (symptomatic in S2 vs asymptomatic in S1). By combining exome, transcriptome and proteome data, we uncovered the S2-specific variant p.R241Q in the *STAT1* gene which is central in the regulatory pathway of inflammation mediated by IFN-γ.[18] Differential expression and molecular pathway analyses at both mRNA and protein levels confirmed the key role of *STAT1* in the development of a higher level of IFN-γ mediated inflammation in S2. Finally, the *STAT1* pathway was experimentally shown to be more activated in the presence of the S2-specific *STAT1* variant p.R241Q.

The *STAT1* variant p.R241Q of patient S2 is inherited from her healthy father whose PBMCs' inflammatory response to LPS stimulation is equivalent to that of S2 (table 2). This result suggests that the presence of both the homozygous *MVK* p.V377I mutation and the heterozygous *STAT1* p.R241Q mutation is required to elicit clinical symptoms. The clinical symptoms of patient S2 could therefore be explained by a synergistic effect of *MVK* and *STAT1*.

STAT1 is a key component of the Janus kinase/signal transducer and activator of transcription signalling pathway (JAK/STAT). This pathway is a major signalling route for a wide range of cytokines and growth factors. Its activation stimulates broad cellular functions including cell proliferation, differentiation, migration and apoptosis, which are critical events in key processes such as haematopoiesis or immune response.[19 20] The JAK/STAT signalling is initiated on binding of cytokines

to their cognate receptors. As a consequence, the receptors dimerise and phosphorylate JAKs, which in turn phosphorylate the STAT family members. These phosphorylated STATs then form parallel dimers and translocate to the nucleus where they act as transcription factors.[21] Dysregulation of JAK/STAT signalling is involved in a variety of conditions including hyper-IgE syndrome, haematological malignancies, diabetes or yet obesity.[22–24] When considering the transcription factor *STAT1* in particular, the activation of this gene has been observed in a number of inflammatory disorders such as liver inflammation, rheumatoid arthritis or yet cerebral ischaemia.[25–27] Based on these information, the involvement of STAT1 in the pathophysiology of the inflammatory phenotype observed in S2 is possible.

At the genetic level, inborn errors of this gene can be classified into four types[1]: (1) autosomal recessive complete *STAT1* deficiency, (2) autosomal recessive partial *STAT1* deficiency, (3) autosomal dominant STAT1 deficiency and (4) autosomal-dominant gain of *STAT1* activity.[28] The latter category is characterised by an increased activation of STAT1, which is known to be associated with a broad clinical spectrum including inherited infectious diseases (ie, the most common genetic cause of chronic mucocutaneous candidiasis), autoimmune/inflammatory features, carcinomas or yet aneurysms.[29–32] The functional translation of gain of function STAT1 variants is gain of phosphorylation by impairment of
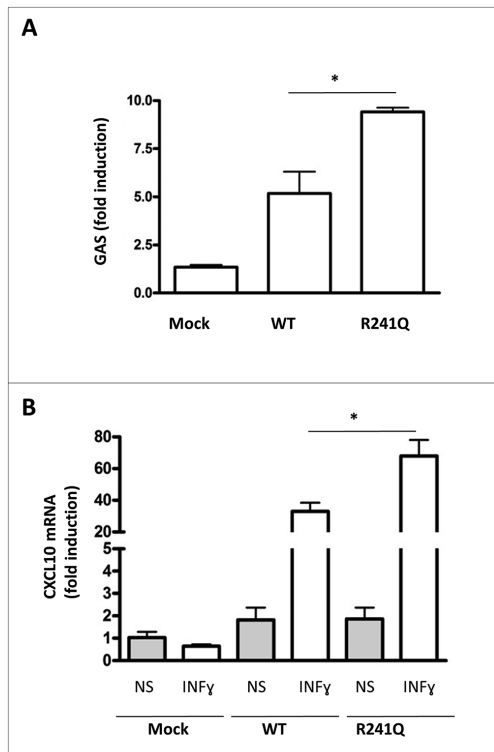
Figure 7 Functional characterisation of the gain-of-function variant R241Q U3A cells were transfected with a mock vector, a WT, or the mutant allele R241Q of STAT1. All experiments were performed at least three times independently. (A) Response to IFN-γ evaluated by luciferase activity of a reporter gene under the control of the GAS (Interferon-Gamma Activated Sequence) promoter. (B) Induction of CXCL10 after stimulation with IFN-γ measured by quantitative RT-PCR. IFN-γ, interferon-γ; WT, wild type.

nuclear dephosphorylation.[29] The heterozygous variant identified in S2 has several attributes to be a new gain of function variant[1]: it is located in the coiled coil domain of the protein, where the vast majority of variants described so far are dominant gain of function mutations[2 28] bioinformatics prediction tools classify the variant as non-deleterious for protein function (online supplementary table S4),[3] an important number of proteins of the STAT1 signalling pathway are upregulated in S2 vs S1,[4] in a cellular model, STAT1 with the p.R241Q variant showed a higher transcriptional activity compared with the wild type protein.

Given the important amount of data linking JAK/STAT signalling to health and disease, this pathway has become an important drug target. Inhibitors of JAKs have notably proven effective in the treatment of myeloproliferative diseases and rheumatoid arthritis.[33 34] STATs are also being investigated as targets using either oligonucleotide-based inhibitors[35] or inhibitory peptides and small interfering RNAs.[36] The results of the present study suggest that targeting the JAK/STAT pathway with approved JAK inhibitors such as tofacitinib or ruxolitinib could be a good therapeutic option to treat patient S2 who is currently under anti-IL1 (Anakinra 100 mg/day) without substantial amelioration.[37]

From a more general point of view, the unique availability of exome, transcriptome and proteome data for the two studied siblings allowed us to refine the common data interpretation strategy. Indeed, database search-driven algorithms used for proteomics data processing rely on exact matching between measured and theoretical sequences. Therefore, the use of personalised protein databases including subject-specific information derived from exome and RNASeq data allows improving proteomics data interpretation both qualitatively and quantitatively.[38] We could even perform allele-specific gene expression analysis up to the protein level. Finally, we were able to identify a significant number of SAAVs with regard to the achieved proteome coverage. Reverse benefits are also noteworthy as, for instance, the ability of RNASeq strategies to identify new splice variants and confirm their expression with proteomics thanks to personalised databases including them, will overall enable improving genome annotation.[39]

Our study has been focused on STAT1 as a unique potential trigger for increased inflammation in patient S2. Although we believe that it is a good candidate based on the results presented here, we cannot exclude other genes and mechanisms such as regulatory elements that could contribute to the observed phenotypic difference between the two siblings. Technically, quantitative data of the proteomics approach would have been more precise with the latest generation mass spectrometers that use an extracted ion chromatogram quantification strategy rather than a spectral count approach that is limited in dynamics and quantitative accuracy.[40] Globally, integration of multilevel datasets improves with the increase in depth of coverage of proteomics data.[41] As observed in other studies, we confirmed that mRNA and protein levels quite poorly correlate.[42–44] In this regard, performing a multilevel study provides a more precise view of the global genome expression profile. Finally, 20 years of technical progress has allowed access to multiomics characterisation at the level of a single patient and its potential to become an

Table 2 Clinical, mutational and inflammatory statuses of the family members

Subject	Clinical status (symptomatic)	*MVK* Homozygous p.V377I variant	*STAT1* Heterozygous p.R241Q variant	TNF-α (pg/mL)*	IL-1β (pg/mL)*	IL-6 (pg/mL)*
Mother	No	No	No	+ (4135±1838)	+ (3995±794)	+ (14642±700)
Subject S1†	No	Yes	No	+ (4804±1251)	+ (3868±709)	+ (18316±1518)
Father	No	No	Yes	++ (8580±1073)	++ (7388±831)	++ (24727±479)
Patient S2†	Yes	Yes	Yes	++ (10437±2804)	++ (5790±743)	++ (24649±1408)

*Peripheral blood mononuclear cells were stimulated with LPS (1 µg/mL) for 6 hours. Cytokine levels in the cell culture supernatant were quantified by ELISA. + and ++ represent the global estimate of concentrations. Values are the mean of at least three experiments±SD. + vs ++ conditions were statistically different for all three cytokines (Mann-Whitney U test p<0.05).
†Results from Messer *et al*.[3]
LPS, lipopolysaccharide.

efficient tool for personalised medicine is demonstrated here. However, to transform the proof-of-concept study to a routine application in clinics, a few challenging aspects still need to be improved, including delays in multilevel data generation and processing and the availability of more sophisticated data integration software solutions.

Author affiliations
[1]Laboratoire d'ImmunoRhumatologie Moléculaire, INSERM UMR_S1109, Plateforme GENOMAX, LabEx TRANSPLANTEX, Faculté de Médecine, Université de Strasbourg, Strasbourg, France
[2]Franco-Japanese Nextgen HLA laboratory, Laboratoire International Associé (LIA) INSERM, Nagano, Japan
[3]Fédération Hospitalo-Universitaire OMICARE, Université de Strasbourg, Strasbourg, France
[4]Fédération de Médecine Translationnelle de Strasbourg (FMTS), Université de Strasbourg, Strasbourg, France
[5]Laboratoire d'Immunologie, Plateau Technique de Biologie, Pôle de Biologie, Nouvel Hôpital Civil, Strasbourg, France
[6]Laboratoire de Spectrométrie de Masse BioOrganique, Université de Strasbourg, CNRS, IPHC, UMR 7178, Strasbourg, France
[7]Molecular Immunology Unit, BIOMICA SAS, Strasbourg, France
[8]Service de Rhumatologie, Hôpitaux Civils de Colmar, Colmar, France
[9]Service de Rhumatologie, Centre National de Référence pour les Maladies Autoimmunes Systémiques Rares, Hôpitaux Universitaires de Strasbourg, Strasbourg, France

REFERENCES

1 Haas D, Hoffmann GF. Mevalonate kinase deficiency and autoinflammatory disorders. *N Engl J Med* 2007;356:2671–3.

2 Haas D, Hoffmann GF. Mevalonate kinase deficiencies: from mevalonic aciduria to hyperimmunoglobulinemia D syndrome. *Orphanet J Rare Dis* 2006;1:13.

3 Messer L, Alsaleh G, Georgel P, *et al*. Homozygosity for the V377I mutation in mevalonate kinase causes distinct clinical phenotypes in two sibs with hyperimmunoglobulinaemia D and periodic fever syndrome (HIDS). *RMD Open* 2016;2:e000196.

4 Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biol* 2017;18:83.

5 McKenna A, Hanna M, Banks E, *et al*. The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–303.

6 Li MX, Gui HS, Kwan JS, *et al*. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic Acids Res* 2012;40:e53.

7 Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010;11:R106.

8 Pham TV, Piersma SR, Warmoes M, *et al*. On the beta-binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics. *Bioinformatics* 2010;26:363–9.

9 Carapito C, Burel A, Guterl P, *et al*. MSDA, a proteomics software suite for in-depth mass spectrometry data analysis using grid computing. *Proteomics* 2014;14:1014–9.

10 Carapito C, Lane L, Benama M, *et al*. Computational and mass-spectrometry-based workflow for the discovery and validation of missing human proteins: application to chromosomes 2 and 14. *J Proteome Res* 2015;14:3621–34.

11 Law CW, Chen Y, Shi W, *et al*. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 2014;15:R29.

12 Smyth G. In: Dudoit S, Irizarry R, Huber W, eds. *Bioinformatics and computational biology solutions using R and bioconductor*. New York: Springer, 2005:397–420.

13 Meng C, Kuster B, Culhane AC, *et al*. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics* 2014;15:162.

14 Bady P, Dolédec S, Dumont B, *et al*. Multiple co-inertia analysis: a tool for assessing synchrony in the temporal variability of aquatic communities. *C R Biol* 2004;327:29–36.

15 Chernick MR, F RH. *Introduction biostatistics for the health sciences: modern applications including bootstrap. Wiley series in probability and statistics*. New York, United States: John Wiley and Sons Ltd, 2003:424.

16 Houten SM, van Woerden CS, Wijburg FA, *et al*. Carrier frequency of the V377I (1129G>A) MVK mutation, associated with Hyper-IgD and periodic fever syndrome, in the Netherlands. *Eur J Hum Genet* 2003;11:196–200.

17 Eisen MB, Spellman PT, Brown PO, *et al*. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 1998;95:14863–8.

18 Rauch I, Müller M, Decker T. The regulation of inflammation by interferons and their STATs. *JAKSTAT* 2013;2:e23820.

19 Darnell JE, Kerr IM, Stark GR. Jak-STAT pathways and transcriptional activation in response to IFNs and other extracellular signaling proteins. *Science* 1994;264:1415–21.

20 Rawlings JS, Rosler KM, Harrison DA. The JAK/STAT signaling pathway. *J Cell Sci* 2004;117:1281–3.

21 Kiu H, Nicholson SE. Biology and significance of the JAK/STAT signalling pathways. *Growth Factors* 2012;30:88–106.

22 Minegishi Y, Karasuyama H. Defects in Jak-STAT-mediated cytokine signals cause hyper-IgE syndrome: lessons from a primary immunodeficiency. *Int Immunol* 2009;21:105–12.

23 Vainchenker W, Constantinescu SN. JAK/STAT signaling in hematological malignancies. *Oncogene* 2013;32:2601–13.

24 Gurzov EN, Stanley WJ, Pappas EG, *et al*. The JAK/STAT pathway in obesity and diabetes. *Febs J* 2016;283:3002–15.

25 Gao B. Cytokines, STATs and liver disease. *Cell Mol Immunol* 2005;2:92–100.

26 Kasperkovitz PV, Verbeet NL, Smeets TJ, *et al*. Activation of the STAT1 pathway in rheumatoid arthritis. *Ann Rheum Dis* 2004;63:233–9.

27 Takagi Y, Harada J, Chiarugi A, *et al*. STAT1 is activated in neurons after ischemia and contributes to ischemic brain injury. *J Cereb Blood Flow Metab* 2002;22:1311–8.

28 Boisson-Dupuis S, Kong XF, Okada S, *et al*. Inborn errors of human STAT1: allelic heterogeneity governs the diversity of immunological and infectious phenotypes. *Curr Opin Immunol* 2012;24:364–78.

29 Liu L, Okada S, Kong XF, *et al*. Gain-of-function human STAT1 mutations impair IL-17 immunity and underlie chronic mucocutaneous candidiasis. *J Exp Med* 2011;208:1635–48.

30 van de Veerdonk FL, Plantinga TS, Hoischen A, *et al*. STAT1 mutations in autosomal dominant chronic mucocutaneous candidiasis. *N Engl J Med* 2011;365:54–61.

31 Toubiana J, Okada S, Hiller J, *et al*. Heterozygous STAT1 gain-of-function mutations underlie an unexpectedly broad clinical phenotype. *Blood* 2016;127:3154–64.

32 Uzel G, Sampaio EP, Lawrence MG, *et al*. Dominant gain-of-function STAT1 mutations in FOXP3 wild-type immune dysregulation-polyendocrinopathy-enteropathy-X-linked-like syndrome. *J Allergy Clin Immunol* 2013;131:1611–23.

33 Hodge JA, Kawabata TT, Krishnaswami S, *et al*. The mechanism of action of tofacitinib - an oral Janus kinase inhibitor for the treatment of rheumatoid arthritis. *Clin Exp Rheumatol* 2016;34:318–28.

34 Sonbol MB, Firwana B, Zarzour A, *et al*. Comprehensive review of JAK inhibitors in myeloproliferative neoplasms. *Ther Adv Hematol* 2013;4:15–35.

35 Sen M, Grandis JR. Nucleic acid-based approaches to STAT inhibition. *JAKSTAT* 2012;1:285–91.

36 Miklossy G, Hilliard TS, Turkson J. Therapeutic modulators of STAT signalling for human diseases. *Nat Rev Drug Discov* 2013;12:611–29.

37 Aittomäki S, Pesu M. Therapeutic targeting of the Jak/STAT pathway. *Basic Clin Pharmacol Toxicol* 2014;114:18–23.

38 Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. *Nat Methods* 2014;11:1114–25.

39 Marx H, Hahne H, Ulbrich SE, *et al*. Annotation of the Domestic Pig Genome by Quantitative Proteogenomics. *J Proteome Res* 2017;16:2887–98.

40 Ramus C, Hovasse A, Marcellin M, *et al*. Benchmarking quantitative label-free LC-MS data processing workflows using a complex spiked proteomic standard dataset. *J Proteomics* 2016;132:51–62.

41 Mitchell CJ, Getnet D, Kim MS, *et al*. A multi-omic analysis of human naïve CD4+ T cells. *BMC Syst Biol* 2015;9:75.

42 Zhang B, Wang J, Wang X, *et al*. Proteogenomic characterization of human colon and rectal cancer. *Nature* 2014;513:382–7.

43 Gry M, Rimini R, Strömberg S, *et al*. Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC Genomics* 2009;10:365.

44 Edfors F, Danielsson F, Hallström BM, *et al*. Gene-specific correlation of RNA and protein levels in human cells and tissues. *Mol Syst Biol* 2016;12:883.