SUPPLEMENTARY INFORMATION

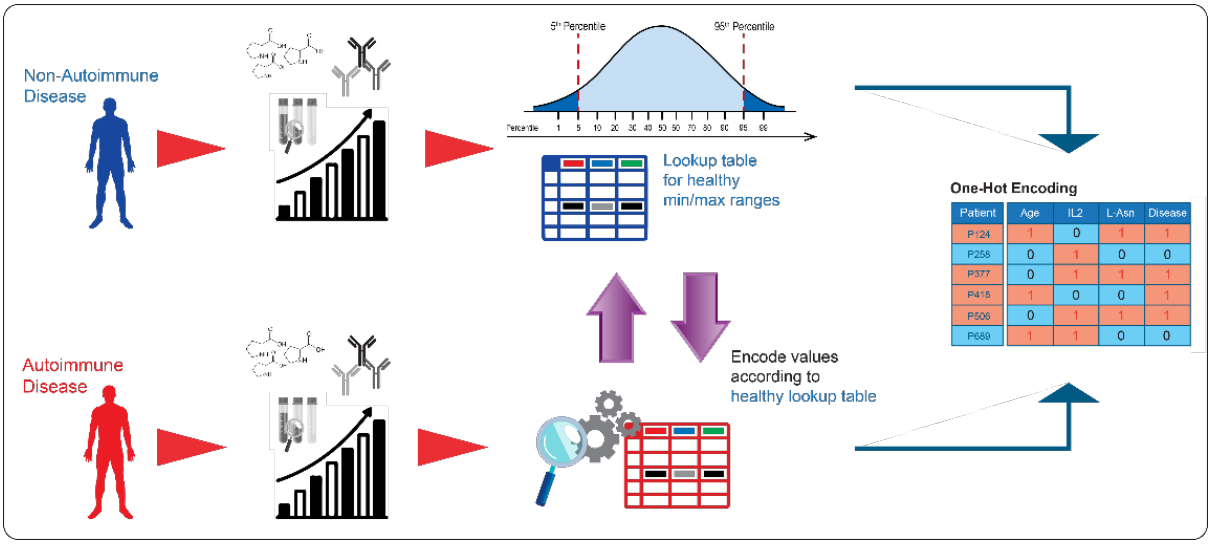

Fig. S1. Illustration of a dataset in which various data types were integrated and one-hot encoded. The red fields indicate the occurrence of one outlier of the specified characteristic per patient.

Measure	**Formula**
Accuracy	(TP + TN) / (TP + TN + FN + FP)
Precision	TP / (TP + FP)
Recall	TP / (TP + FN)
Macro-average (k = number of classes)	$P_{macro} = (P_a + P_b + \ldots + P_n) / k$
Micro-average	$P_{micro} = (TP_a + TP_b + \ldots + TP_n) / (TP_a + TP_b + \ldots + TP_n) + (FP_a + FP_a + FP_b + \ldots + FP_n)$

Table S1. Equations used for determining evaluation of machine learning models to the corresponding data. TP: true positive, TN: true negative, FP: false positive, FN: false negative.

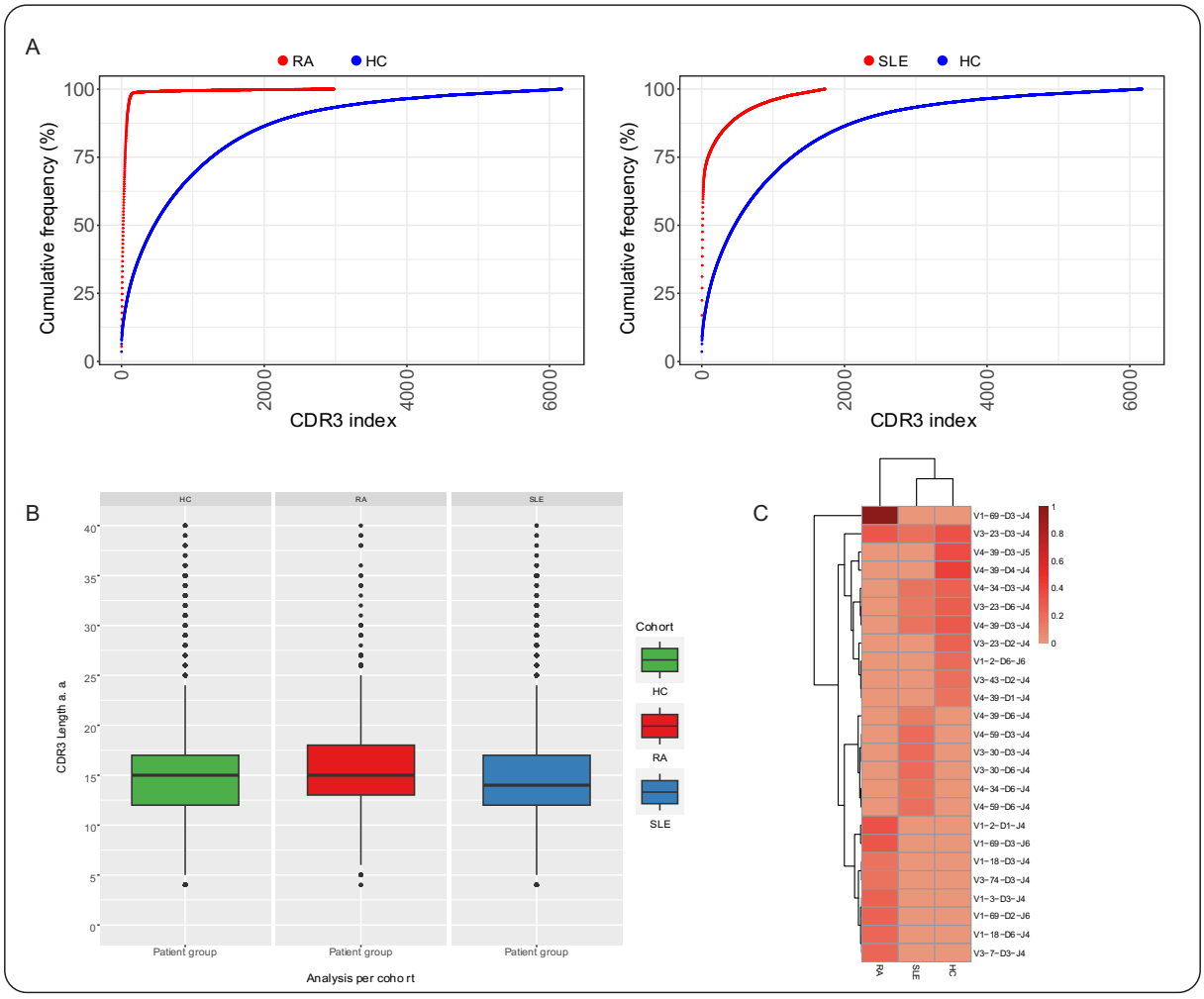

Fig. S2. A) Cumulative frequency plot of one representative RA patient (red line) compared to HC reference (blue), shows the repertoire diversity with CDR3 clones ordered from increasing to decreasing frequency. The most frequent 500 clones account for over 95% of the repertoire, while the cumulative frequency of CDR3 in the healthy cohort shows a diverse repertoire. Furthermore, SLE patient (right side, red line) compared to HC reference reveals that frequent clones account for a significant portion of the repertoire, although to a lower extent than in the RA patient. B) Analysis of the CDR3 lengths of the respective cohorts reveals that specific CDR3 lengths are not a feature that distinguishes in autoimmune diseases, therefore excluding it as a feature represented in data integration. C) VDJ frequency of antibody repertoires in RA, SLE and HC cohorts (pooled samples).

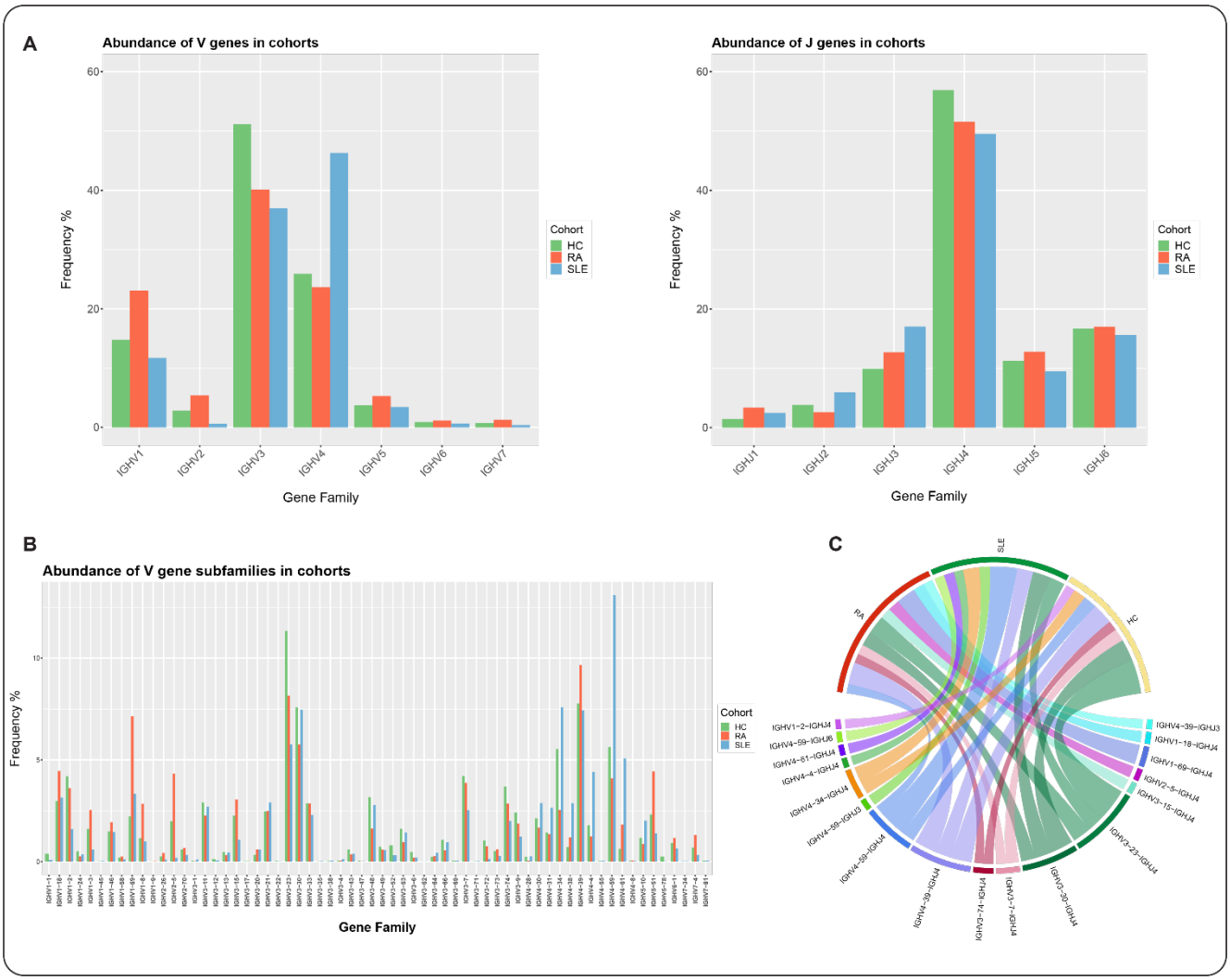

Fig. S3. A) The analysis of V gene frequencies between the cohorts revealed notable variations in IGHV1, IGHV3, and IGHV4, while differences within other V gene families are less pronounced. In the J genes, we observed a decrease in IGHJ4 in both autoimmune groups, an evident increase in IGHJ3 in AID patients, and a increase in IGHJ1 in AID patients. B) Frequency of V gene subfamilies. In SLE patients, there was an increase in IGHV1-69, IGHV4-34, IGHV4-39, IGHV4-59, and IGHV4-61 compared to the control group. The SLE group, on the other hand, presented lower levels of IGHV3-23 than the healthy cohort. Importantly, when compared to healthy participants, RA patients had higher levels of IGHV1-69, IGHV4-39, and IGHV5-51 and lower levels of IGHV3-23 and IGHV4-34. C) the V-J gene recombination indicates a range of variations, no combinations are typical of any specific cohort.

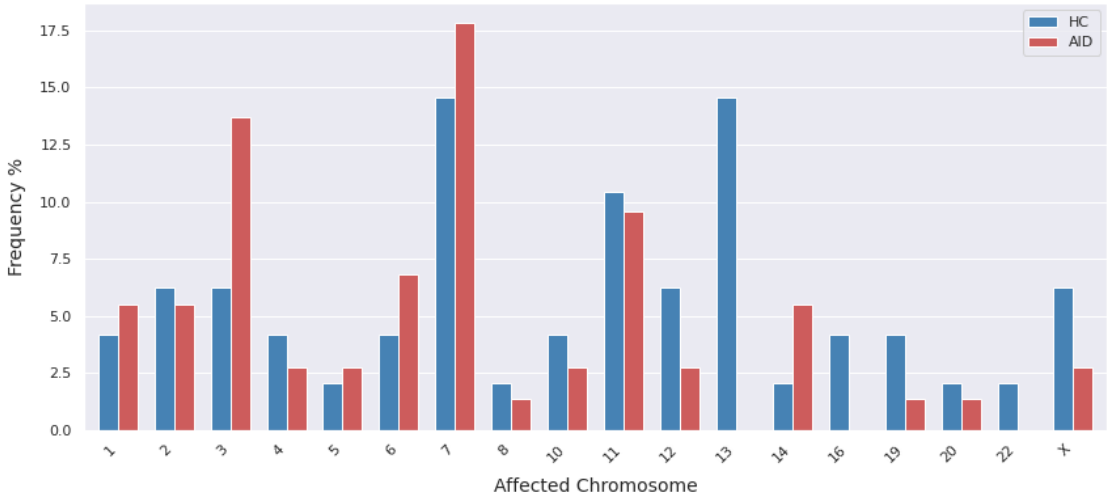

Fig. S4. Frequency of listed mutations per chromosome illustrated as a bar chart using an autoimmune patient and a healthy patient as reference. In this analysis, primarily alterations affecting more than one nucleotide were considered. AID samples show an increased number of variants on chromosomes 3, 6, 7 and 14.

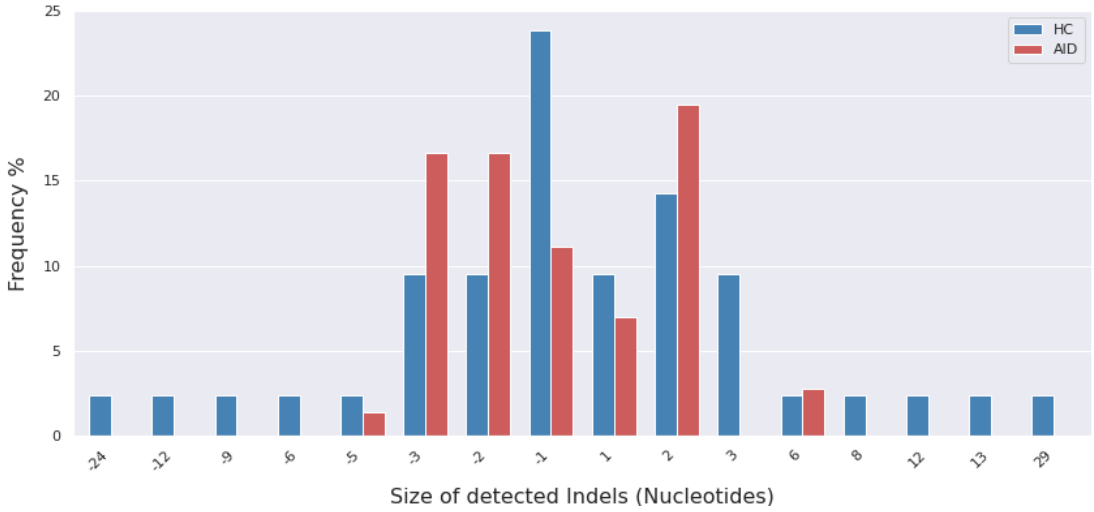

Fig. S5. The number of nucleotides in the insertions (positive values) and deletions (negative values) between the healthy and autoimmune samples. Autoimmune patients had fewer long indels but a higher proportion of shorter mutations such as insertions and deletions.

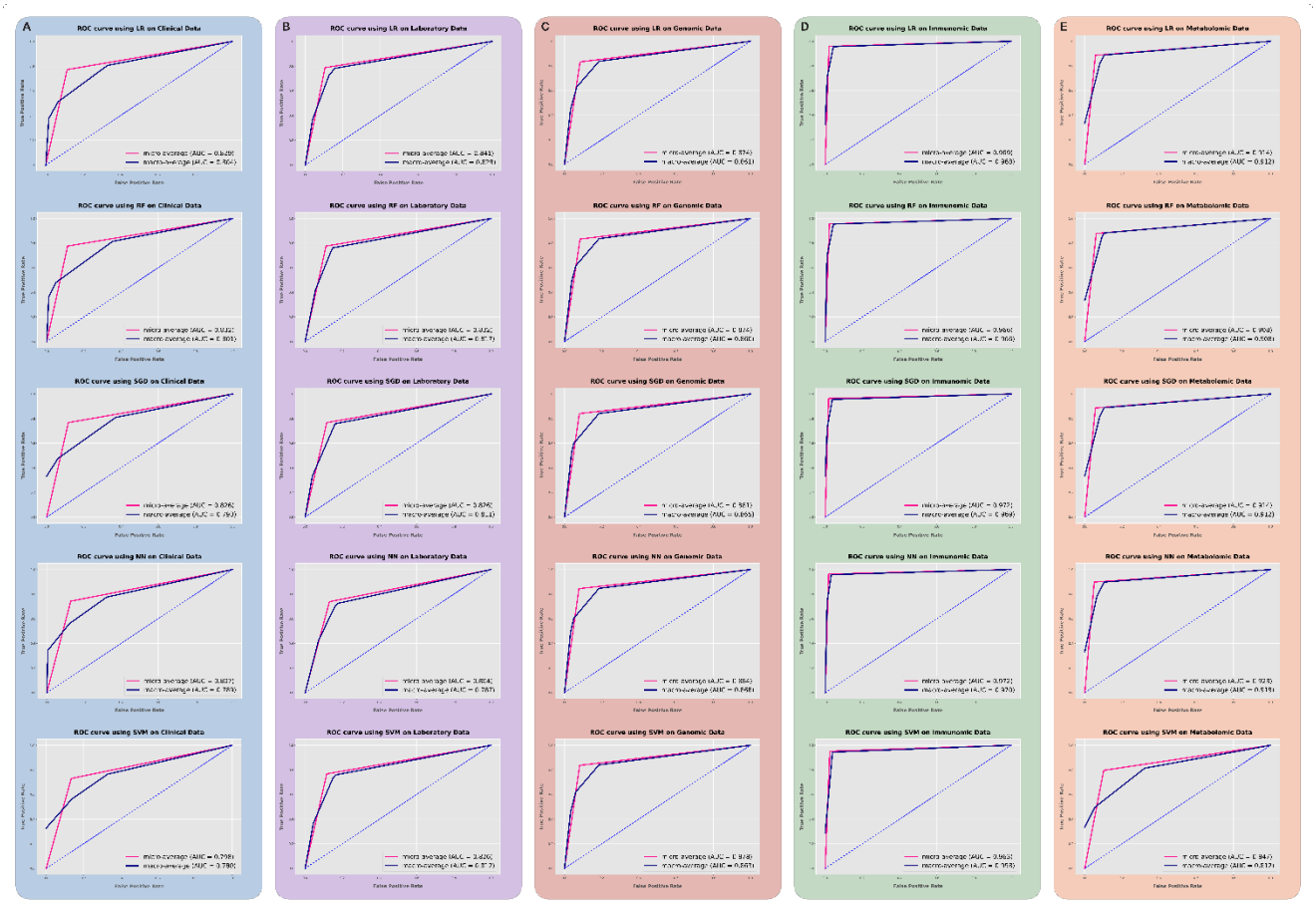

Fig. S6. The performance of support vector machines (SVM), random forest (RF), logistic regression (LR), stochastic gradient descent (SGD), and multilayer perceptron (MLP) classifiers ML models were applied to datasets containing only one datatype each, specifically the performance when incorporating only clinical (blue), laboratory (purple), genomic (red), immunomic (green) or metabolomic (orange) data have been evaluated.

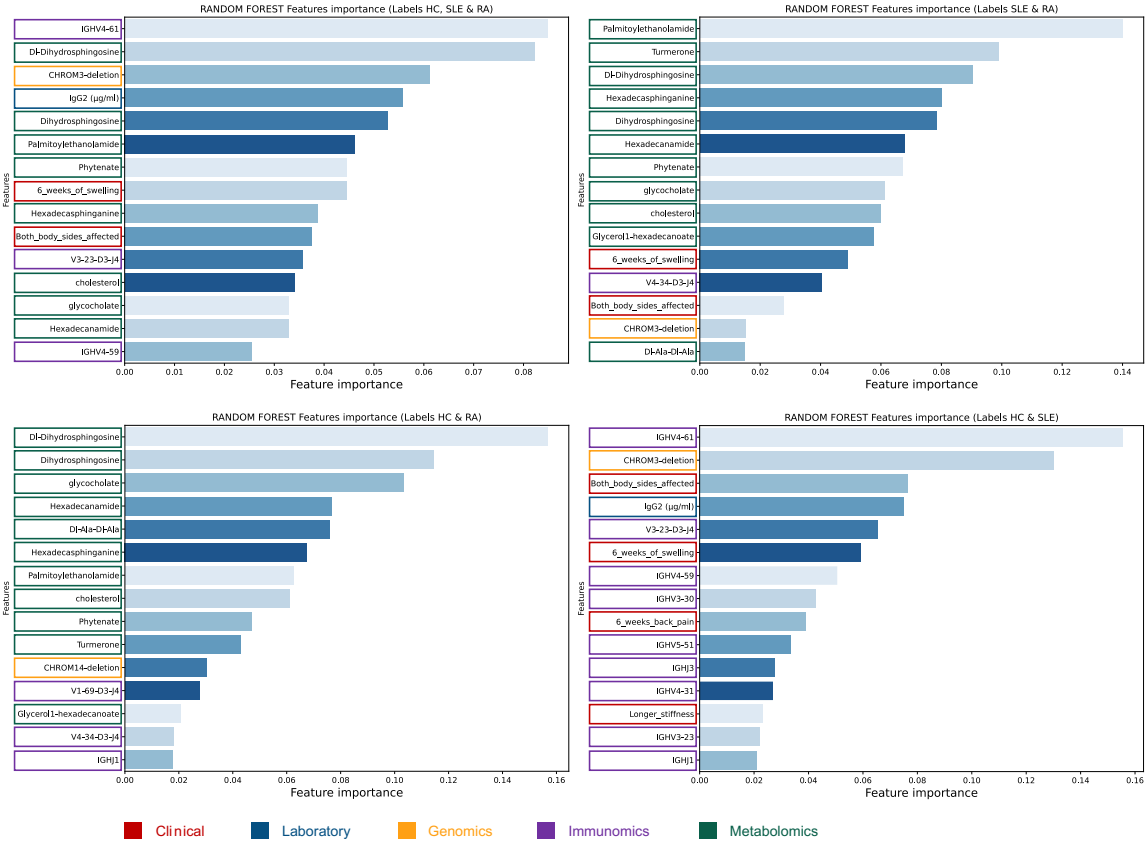

Fig. S8. Weighting of feature importance generated by grid search cross-validated Random Forest classifier applied to integrated dataset containing clinical data (red), laboratory data (blue) and multi-omics data, namely genomics (orange), immunomics (purple) and metabolomics (green).

Determined soluble mediators
IL-4 (pg/ml), BAFF (pg/ml), IgM (µg/ml), IgG2 (µg/ml), IL-23 (pg/ml), IL-6 (pg/ml), IL-12p70 (pg/ml), PTX3 (pg/ml), APRIL (pg/ml), IgG1 (µg/ml), IL-27 (pg/ml), CXCL12 (pg/ml), sST2 (pg/ml)

Table S2. Soluble mediators, obtained using laboratory data analysis procedures, which were included as features to be used for data integration.

Variant Effect Predictor (VEP) characteristics	Applied Filter
Consequence	● missense_variant ● splice_acceptor_variant&non_coding_transcript_variant ● splice_region_variant&intron_variant ● missense_variant&NMD_transcript_variant ● stop_lost ● stop_gained ● splice_acceptor_variant ● frameshift_variant ● splice_region_variant&intron_variant&NMD_transcript_variant ● splice_donor_variant ● start_lost ● inframe_deletion ● splice_region_variant&synonymous_variant ● inframe_insertion ● regulatory_region_variant
VARIANT_CLASS	● deletion ● insertion ● sequence_alteration ● indel
MAX_AF	<= 0.01

Table S3. A list of Ensembl VEP filtering properties used to determine which mutations were likely to have a greater impact on health status. Consequence refers to the predicted effect of a genetic variant on the corresponding gene or transcript, therefore provides information regarding the predicted functional consequences of genetic variants.

Determined genomic features
CHROM1-deletion, CHROM3-deletion, CHROM5-deletion, CHROM6-insertion, CHROM6-deletion, CHROM7-insertion, CHROM7-deletion, CHROM14-deletion, CHROMX-insertion

Table S4. Whole Exome Sequence (WES) data was used to define features to integrate as genomic variants based on the analysis WES.

Extracted V genes and J genes	Extracted CDR3s
IGHV1-69, IGHV3-23	IGHV1-69-IGHD3-IGHJ4
IGHV3-30, IGHV4-30	IGHV3-23-IGHD3-IGHJ4
IGHV4-31, IGHV4-34	IGHV3-23-IGHD6-IGHJ4
IGHV4-39, IGHV4-59	IGHV3-30-IGHD6-IGHJ4
IGHV4-61, IGHV5-51	IGHV3-43-IGHD2-IGHJ4
IGHJ1	IGHV4-34-IGHD3-IGHJ4
IGHJ3	IGHV4-39-IGHD3-IGHJ4
IGHJ4	IGHV4-59-IGHD3-IGHJ4

Table S5. Selected immunomic features for data integration based on the analysis of NGS data.

Extracted metabolomics data	
Hexadecanamide	Glycerol 1-hexadecanoate
Glycocholate	Dl-Dihydrosphingosine
Palmitoylethanolamide	Hexadecasphinganine
Cholesterol	Palmitic acid
Turmerone	Epsilon caprolactam
Phytenate	Dihydrosphingosine
Dl-Ala-Dl-Ala	Kynurenic acid
Oxoproline	Lysophosphatidylethanolamine
Deoxyadenosine	

Table S6. Metabolomics characteristics that were determined to distinguish the corresponding cohorts through data analysis, and used for data integration.