Supplement

Explanation of the statistical methods used in this article

Software

Data analysis was performed using commercial MS Office tools, Tibco/Spotfire 12.0.0, Partek Flow, Omniviz software version 6.1.13.0 and Textpad version 7.5.1 on a 64 bit MS Windows PC. QIAGEN's Ingenuity® Pathway Analysis (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity) software was used for canonical pathway and upstream regulator analyses.

Summary

1. To compare genes for innate immunity with each other, genes for adaptive immunity with each other, and genes for innate immunity with genes for adaptive immunity.
    a. **Pearson's correlation** between every gene was calculated. An average was taken for the above three groups.
    b. A total of 190 values were calculated, one for each pair of genes in the set of 20 genes.
    c. The results are shown in Supplementary table 2.
        i. Genes co-expressing with IL1B have a high correlation value.
        ii. Genes co-expressing with BLNK have a high correlation value.
        iii. IL1B-related genes are consistently overexpressed compared to BLNK-related genes, as indicated by their slightly negative correlation values (~-0.2).
2. To compare 20 genes (IL1B & BLNK and others) of controls with the SAIDs.
    a. The **Mann-Whitney U test** was used to compare independent samples with non-normal distributions.
    b. A total of 20 values were calculated, one for each gene.
    c. The results are shown in Supplementary table 3.
3. To compare of 20 genes (IL1B & BLNK and others) genes of 52 SAID patients with their follow-up samples.
    a. The **Wilcoxon Signed-Rank test** was used to compare dependent samples with non-normal distributions**.**
    b. A total of 20 values were calculated, one for each gene.
    c. The results are shown in supplementary table 4.
4. To compare 19 ANA-positive patients with 43 true-negative controls.
    a. The **SAM test** was used to find the 30 most differentially expressed genes, 15 for one group and 15 for the other.
5. To compare the proteomic levels of 60 SAID patients before and after treatment.
    a. The **SAM test** was used to find the 30 most differential proteins, 15 for one group and 15 for the other.

Pearson's correlation

Pearson's correlation values (r) and significance values (p) were calculated for each gene pair of the 20 genes examined in Figures 1-4, also shown in Supplementary Table 1. A total of 190 combinations were calculated, of which the first 45 were IL1B-linked gene pairs, the next 45 were BLNK-linked gene pairs and the remaining 100 were pairs between the groups.

For this calculation, we used the normalised RNA expression values of the 20 transcripts described above for all measured samples. Log2 geometric means were calculated for each expression value and mapped to the centre of zero, giving negative values for under-expressed genes for that sample and positive values for over-expressed genes for that sample.

Pearson's correlation coefficient was calculated as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- *x* and *y* are the input lists of expression values for sample x and sample y.
- x_i and y_i are individual elements of that list.
- $\bar{x}$ and $\bar{y}$ are mean values of *x* and *y*, respectively.
- The numerator represents the covariance between *x* and *y.*
- The denominator represents the product of the standard deviations of *x* and *y*.

p-values were then calculated by converting r into a t-statistic, as follows:

$$t = \frac{r \times \sqrt{n-2}}{\sqrt{1-r^2}}$$

- n is the number of samples per gene.
- r is the relevant Pearsons coefficient
- Larger t-statistics corresponded to small p-values, which were then calculated using the survival function.

Mann-Whitney U test

We used the Mann-Whitney U test to compare the RNA expression values of the SAID patients with those of the controls, since we were comparing two independent samples with non-normal distributions.

For each of the 20 genes described in the previous section, we took the normalised RNA expression values. Log2 geometric means were calculated for each expression value and centred on zero, giving negative values for under-expressed genes for that sample and positive values for over-expressed genes for that sample.

Z-scores of the statistical tests were calculated as follows:

1. We performed this test per gene, for example IL1B.
2. We took all the values of the SAID patients in a list, then appended all the values of the controls. In our dataset there were 339 and 68 values, respectively. 407 in total.
    a. We ranked each value, assigning '1' to the lowest and '407' to the highest. We then got a list of size 407 like [304, 49, 294, 105, 5, ... , 95, 392].
3. We then calculated 's', the sum of the ranks within the SAID group.
    a. We then calculated the 's' sum of the first 339 rank values in the list above.
    b. For IL1B this is 70717.
4. We then calculated the 'expected' value:
    a. n1 * (n1+n2+1) / 2
    b. n1 = number of values in SAIDs = 339
    c. n2 = number of values in controls = 68
    d. result = 69156
5. We calculated the standard error of U
    a. $$\sigma_U = \sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}$$
    b. Sqrt(339 * 68 * 408 / 12) = 885,3
6. We calculated the Z-score by (s – expected) / std.errorU: (70717 – 69156) / 885,3 = 1,763
7. The Z-score could then be mapped directly to a corresponding p-value from the standard normal distribution.

Wilcoxon signed-rank test

When comparing SAID patients before and after treatment, we have a paired dependent sample. Therefore, we chose to use a Wilcoxon signed-rank test instead of the Mann-Whitney U test to compare the groups against each other and to test the null hypothesis that the median difference between paired observations is zero.

For each of the 20 genes described above, we took the normalised expression values, calculated their log2 geometric means and then centred the values around zero, so that highly expressed genes were positive and less expressed genes were negative.

Z-scores for the Wilcoxon signed-rank test were calculated as follows:

1. Given two paired samples $X = [x_1, x_2, \ldots, x_n]$ and Y = $[y_1, y_2, \ldots, y_n]$
    a. n is the number of patients with follow-up samples.
    b. x_1 is the first sample's RNA expression of the analyzed gene pre-treatment.
    c. y_1 is the first sample's RNA expression of the analyzed gene post-treatment.
2. We computed the differences: $d_i = x_i - y_i$
3. We then computed the absolute differences: $|d_i|$
4. We ranked the absolute differences; this gave a list of size n with values between 1 and n.
5. We then assigned signs to ranks:
    a. $R_i = rank(|d_i|) * sign(d_i)$
    b. This step ensured that positive and negative differences were in separate categories.
6. We then computed the rank sums:
    a. Positive rank sum:
        i. $W_+ = sum(R_i)$ (for positive d_i)
    b. Negative rank sum:
        i. $W_- = sum(R_i)$ (for negative d_i)
7. We tested the statistics: $W = min(W_+, W_-)$
8. As n was 52 (total transcriptomic samples with follow-up), W was approximately distributed as follows: $Z = (W - (n * (n + 1)) / 4) / sqrt(((n * (n + 1)) * (2n + 1)) / 24)$
9. The Z-score could then be mapped directly to an corresponding p-value from the standard normal distribution.

Significance analysis of microarrays (SAM)

The SAM score is used in differential gene expression analysis to identify significantly differentially expressed genes while accounting for multiple testing.

In our case, we used SAM to find the 30 most differentially expressed genes by comparing ANA-positive SAID patients with a subset of true-negative controls, and by comparing the proteomics of SAID patients before and after treatment.

The calculations for this assay were performed within Omniviz®, a bioinformatics data visualisation tool that supports comparative studies such as those conducted in this study.

In our two SAM analyses, we selected the 15 most significant transcripts on either side of the two-tailed curve, giving us 15 genes that were over-expressed in one group and under-expressed in the other, and vice versa.