Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Robust interferon-associated B-cell transcriptional changes in systemic lupus erythematosus: a pseudobulk single-cell analysis with leave-one-batch-out validation

Ali Hamza Bilal

alihamzabilal038@gmail.com

Independent https://orcid.org/0009-0002-2063-2534

Research Article

Keywords: systemic lupus erythematosus, B cells, single-cell RNA-seq, pseudobulk, interferon, robustness, TLR7, USP18

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11816085/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Systemic lupus erythematosus (SLE) is characterised by a prominent type I interferon signature, and B cells are central effectors of the disease. Single-cell RNA sequencing offers cell-type resolution, but differential-expression results from multiplexed cohorts can be sensitive to batch composition and to low-cell-count samples. We re-analysed B cells from the public multiplexed single-cell dataset GSE174188 (162 SLE donors and 99 healthy controls) using a pseudobulk, donor-aware, batch-adjusted limma-voom framework. The primary analysis (284 pseudobulk samples from 220 donors; 7,917 genes tested) identified 1,586 differentially expressed genes at a Benjamini-Hochberg false discovery rate below 0.05 (731 upregulated and 855 downregulated in SLE). Gene Ontology enrichment yielded 66 significant Biological Process terms, including 21 immune-, antiviral-, interferon- or defence-related terms; two interferon-mediated signalling terms were significant, together comprising 19 differentially expressed genes. Six biologically relevant candidates (TLR7, USP18, JAK1, JAK2, IFIH1 and STAT1) were prioritised for focused assessment. Two sensitivity analyses tested robustness: excluding pseudobulk samples with fewer than 100 B cells (259 samples, 202 donors) left all six candidates significant with unchanged direction and a maximum absolute log2 fold-change shift of 0.0446; leave-one-batch-out re-analysis across 19 sequencing batches retained a median of 91.86% of baseline differentially expressed genes (range 84.56–95.19%), with all six candidates significant in every exclusion. An exploratory machine-learning prioritisation is reported with an explicit caveat, because its training labels were derived from a computational score. These results identify a robust interferon-associated B-cell expression programme in SLE and prioritise six candidates for independent and experimental validation.

1. Introduction

In systemic lupus erythematosus (SLE), a breakdown of immune tolerance produces autoantibodies and sustained type I interferon (IFN) signalling that injure multiple organs. B cells are central to SLE pathogenesis as producers of autoantibodies and as antigen-presenting and cytokine-secreting cells, and B-cell-directed therapies are an active area of drug development. Understanding which transcriptional programmes are genuinely altered in SLE B cells, rather than artefacts of cohort composition or technical batch, is therefore of direct therapeutic relevance.

Single-cell RNA sequencing (scRNA-seq) resolves cell-type-specific expression, but differential-expression analysis of multiplexed cohorts faces two practical problems. First, cells from the same donor are not independent observations, so donor-level correlation must be modelled. Second, results can depend on which sequencing batches are included and on samples built from very few cells. Sensitivity analyses that re-fit the model after excluding batches or low-cell-count samples are the standard way to test whether conclusions survive these choices, yet they are not always reported.

Here we present a pseudobulk re-analysis of B cells from GSE174188, a large multiplexed scRNA-seq cohort of SLE patients and healthy controls [1], with two sensitivity analyses: low-cell-count filtering and leave-one-batch-out (LOBO) re-analysis across 19 sequencing batches. We report the full differential-

expression results, pathway enrichment, six prioritised candidate genes, and an exploratory machine-learning prioritisation, with the limitations of each component stated explicitly. The analysis is hypothesis-generating: no experimental validation is claimed.

2. Methods

2.1. Dataset and pseudobulk construction

The public dataset GSE174188 [1] comprises multiplexed scRNA-seq of peripheral blood mononuclear cells from 162 SLE patients and 99 healthy controls (over 1.2 million cells). B cells were aggregated into pseudobulk samples by donor and batch. The saved sample-level pseudobulk count matrix contained 25,716 genes across 347 sample groups. Four sequencing batches, comprising 63 sample groups (47 healthy and 16 SLE managed), were excluded from the primary differential-expression input, leaving 284 pseudobulk samples from 220 independent donors. The analysis records contain no rationale for the exclusion of these four batches, and none could be reconstructed afterwards; this is disclosed as a limitation (Section 4). For reference, differential expression on all 347 sample groups yields 1,706 significant genes; the primary analysis reported here uses the 284-sample input. The upstream scripts that produced the pseudobulk counts from raw single-cell data were not preserved as executable code; a reconstruction script in the repository reproduces the saved 284-sample count matrix exactly from the existing sample-level pseudobulk counts.

2.2. Differential-expression analysis

Differential expression was tested with a limma-voom framework [2–4] in R version 4.6.1. Lowly expressed genes were filtered with edgeR filterByExpr, library sizes were normalised with the trimmed mean of M-values, and the design matrix adjusted for sequencing batch and disease status (~ Batch + Disease). Repeated samples from the same donor were modelled with duplicateCorrelation using donor as the blocking factor, and empirical Bayes moderation (eBayes) was applied. Significance was assessed at a Benjamini-Hochberg false discovery rate (FDR) below 0.05. The primary analysis tested 7,917 genes.

2.3. Gene Ontology enrichment

Gene Ontology (GO) Biological Process over-representation analysis was performed with GOATOOLS v1.6.9, using the GO ontology release 2026-07-26 and Gene Ontology Annotation human gene associations. The study set comprised the 1,383 differentially expressed genes with GO annotations; the background comprised all 6,973 tested genes with GO annotations. Counts were propagated to parent terms, enrichment was tested with the Fisher exact test, and terms were considered significant at a Benjamini-Hochberg adjusted p-value below 0.05. Immune-related terms were identified by matching term descriptions against the keywords immune, antiviral, interferon and defence/defense. This enrichment analysis was re-run from the differential-expression results for this report with the

documented pipeline above; an earlier exploratory enrichment (94 terms) is retained in the project repository and agrees on the headline terms.

2.4. Candidate gene selection

Six genes (TLR7, USP18, JAK1, JAK2, IFIH1 and STAT1) were selected by the author for focused assessment after reviewing the machine-learning prioritisation outputs, on the basis of significant differential expression in the primary analysis and established biological relevance to interferon signalling or SLE genetics. The six were not selected by rank order from the machine-learning ranking, and selection was hypothesis-driven rather than purely statistical. Differential expression alone was not treated as evidence for a therapeutic direction (inhibition versus activation).

2.5. Sensitivity analyses

Two sensitivity analyses were performed. (i) Low-cell-count filtering: samples assembled from fewer than 100 B cells were dropped and the donor-aware, batch-adjusted model was refitted (259 samples from 202 donors; 7,712 genes tested). (ii) Leave-one-batch-out (LOBO): the model was refitted 19 times, each time excluding one mixed-disease sequencing batch, using as baseline the minimum-100-B-cell dataset derived from the full 347-group matrix (322 samples; 1,768 significant genes). For each exclusion we recorded the number of significant genes, the number shared with the baseline significant set, the percentage of baseline genes retained, the Pearson correlation of log2 fold changes against baseline, and whether each of the six candidates remained significant with unchanged direction. Four further batches, each containing a single disease group, were left out of the one-at-a-time exclusions. Both sensitivity analyses reuse the same cohort and are not independent validation.

2.6. Exploratory machine-learning prioritisation

An exploratory machine-learning component assessed computational target prioritisation from engineered transcriptomic features. Features were summarised per gene from cluster-marker differential-expression results (maximum and mean log2 fold change, maximum and mean expression prevalence, minimum adjusted p-value and number of clusters in which the gene appeared), together with a B-cell specificity score (mean expression in naive B cells divided by mean expression across all other cell types) and composite z-score summaries. A therapeutic-target score was defined as the sum of the z-scores of maximum log2 fold change, mean log2 fold change, cluster count and maximum expression prevalence; genes in the top 10% of this score were labelled positive for classifier training. Repeated LASSO stability selection (100 subsamples of 80% of the training data) recorded feature-selection frequencies. Logistic regression and random forest classifiers were trained on seven specified features (maximum and mean log2 fold change, maximum and mean expression prevalence, minimum adjusted p-value, cluster count and B-cell specificity score) using an 80/20 stratified train/test split. Because the training labels were derived from a computational score built from features related to the predictors, performance figures measure how well the models recover the score's own labels; they do not measure ability to predict experimentally confirmed drug targets. This component is therefore presented strictly as an exploratory exercise in prioritisation.

3. Results

3.1. B-cell differential expression in SLE

The primary analysis identified 1,586 differentially expressed genes at FDR < 0.05, of which 731 were upregulated and 855 downregulated in SLE B cells relative to healthy controls (Fig. 1). The most strongly upregulated genes were dominated by interferon-stimulated genes, including IFI44L (log2 fold change + 2.36), USP18 (+ 2.20), IFIT3 (+ 1.74), IFI27 (+ 1.57), IFI44 (+ 1.49), IFITM1 (+ 1.46) and MX1 (+ 1.42). IFITM1 was the most statistically significant gene (FDR 9.58 x 10^-24), followed by SAMD9L, MX1, EPSTI1 and EIF2AK2. The most strongly downregulated genes included COL4A3 (-1.11), LILRA4 (-1.09), KLK1 (-1.04), ARHGAP24 (-1.03) and IL2RA (-0.99).

3.2. Pathway enrichment and the interferon programme

GO Biological Process enrichment of the significant genes yielded 66 significant terms. Of these, 21 matched immune, antiviral, interferon or defence terminology, including defence response to virus (FDR 8.5e-10; 62 genes), response to virus (FDR 6.6e-09; 80 genes), type I interferon-mediated signalling pathway (FDR 0.0011; 16 genes), interferon-mediated signalling pathway (FDR 0.0017; 19 genes) and antiviral innate immune response (FDR 0.019; 25 genes) (Fig. 2). The two interferon-mediated signalling terms together comprised 19 differentially expressed genes, including JAK1, JAK2 and STAT1. The most significant terms overall were cytoplasmic translation and regulation of syncytium formation, reflecting the strong translational signature in SLE B cells. The enrichment pattern supports an altered interferon-associated transcriptional programme in SLE B cells rather than a small number of isolated gene changes.

3.3. Six prioritised candidate genes

Six genes were prioritised for focused assessment (Table 1). Five were upregulated in SLE B cells: USP18 showed the largest change (+ 2.20), followed by STAT1 (+ 0.92), TLR7 (+ 0.85), IFIH1 (+ 0.57) and JAK2 (+ 0.33). JAK1 was modestly downregulated (-0.14) but highly significant (FDR 5.16 x 10^-07). All six were significant at FDR < 0.05 in the primary analysis. Increased USP18 expression is of particular note because USP18 is a negative regulator of type I interferon signalling; its upregulation may represent compensatory negative feedback rather than a pathogenic driver, and the appropriate intervention direction is undetermined.

Table 1
Six prioritised candidate genes: differential expression in the primary analysis (284 samples) and after excluding pseudobulk samples with fewer than 100 B cells (259 samples). All six remained significant at FDR < 0.05 with unchanged direction. The largest absolute log2 fold-change shift was 0.0446 (USP18).

Gene	Original log2FC	Original FDR	Filtered log2FC	Filtered FDR	Direction retained
TLR7	+ 0.847	2.55e-08	+ 0.850	8.58e-08	Yes
USP18	+ 2.199	5.21e-16	+ 2.244	2.95e-14	Yes
JAK1	-0.140	5.16e-07	-0.140	5.76e-07	Yes
JAK2	+ 0.329	3.23e-04	+ 0.324	3.66e-04	Yes
IFIH1	+ 0.571	9.58e-11	+ 0.589	1.60e-11	Yes
STAT1	+ 0.920	4.24e-17	+ 0.928	9.15e-16	Yes

3.4. Robustness to low-cell-count filtering

After excluding samples with fewer than 100 B cells, all six candidates remained significant at FDR < 0.05 and retained their original expression directions (Table 1, Fig. 3). Effect sizes were essentially unchanged: the largest absolute shift in estimated log2 fold change was 0.0446 (USP18, from + 2.199 to + 2.244), and the remaining shifts were 0.0175 or smaller. The candidate-level conclusions are therefore not driven by low-cell-count pseudobulk samples.

3.5. Leave-one-batch-out robustness

Across the 19 leave-one-batch-out re-analyses, the median retention of baseline significant genes was 91.86% (range 84.56–95.19%), and the median Pearson correlation of log2 fold changes against baseline was 0.9957 (Fig. 4, Table 2). The lowest retention (84.56%) occurred when batch dmx_AbFlare-4 was excluded and the highest (95.19%) when dmx_YE_8 − 2 was excluded; no single batch drove the results. All six candidate genes remained significant with unchanged direction in all 19 exclusions. The differential-expression signature is therefore robust to the composition of sequencing batches in this cohort, within the limits of a single-cohort sensitivity analysis.

Table 2

Leave-one-batch-out results for all 19 excluded batches. Baseline: 322 samples, 1,768 significant genes. Retention is the percentage of baseline significant genes remaining significant after excluding the batch. FC correlation is the Pearson correlation of log2 fold changes against baseline. All six candidates remained significant with unchanged direction in every exclusion.

Excluded batch	Genes tested	Significant DEGs	Shared DEGs	Retention (%)	FC correlation	Candidates significant
dmx_AbFlare-3	8,439	1,558	1,507	85.24	0.9905	6/6
dmx_AbFlare-4	8,438	1,612	1,495	84.56	0.9856	6/6
dmx_flare1	8,514	1,674	1,638	92.65	0.9970	6/6
dmx_flare2	8,325	1,660	1,629	92.14	0.9975	6/6
dmx_YE_7–13	8,394	1,743	1,607	90.89	0.9924	6/6
dmx_YE_7–19	8,478	1,649	1,600	90.50	0.9952	6/6
dmx_YE_7–20	8,516	1,676	1,641	92.82	0.9979	6/6
dmx_YE_7–26	8,438	1,706	1,614	91.29	0.9956	6/6
dmx_YE_8–16	8,535	1,669	1,615	91.35	0.9959	6/6
dmx_YE_8–17	8,491	1,698	1,630	92.19	0.9972	6/6
dmx_YE_8 – 2	8,546	1,765	1,683	95.19	0.9971	6/6
dmx_YE_8–23	8,522	1,716	1,664	94.12	0.9977	6/6
dmx_YE_8 – 3	8,498	1,694	1,624	91.86	0.9957	6/6
dmx_YS-JY-20_pool3	8,566	1,754	1,637	92.59	0.9954	6/6
dmx_YS-JY-20_pool4	8,477	1,624	1,518	85.86	0.9878	6/6
dmx_YS-JY-21_pool1	8,518	1,791	1,675	94.74	0.9960	6/6
dmx_YS-JY-21_pool2	8,576	1,714	1,646	93.10	0.9968	6/6
dmx_YS-JY-22_pool5	8,553	1,766	1,600	90.50	0.9901	6/6
dmx_YS-JY-22_pool6	8,567	1,704	1,617	91.46	0.9948	6/6

3.6. Exploratory machine-learning prioritisation

Repeated LASSO stability selection identified six features selected in all 100 subsamples: maximum and mean log2 fold change, maximum expression prevalence, cluster count, disease score and cell-support

score. The seven features used to train the classifiers differed from this stability-selected set (Section 2.5). Logistic regression and random forest classifiers recorded ROC-AUC values of 1.000 and 0.998 respectively. These figures measure recovery of the derived scoring labels described in the Methods, not prediction of experimentally confirmed targets, and no conclusion of this report depends on them.

4. Discussion

This re-analysis of B cells from GSE174188 finds a broad, batch-robust interferon-associated transcriptional programme in SLE, with 1,586 differentially expressed genes and enrichment of 21 immune- and interferon-related processes, including two interferon-mediated signalling terms. The central contribution of this report is not the gene list itself but the demonstration that it survives two stringent sensitivity checks: low-cell-count filtering and leave-one-batch-out re-analysis across 19 batches. Median DEG retention of 91.86% and fold-change correlations above 0.98 in every exclusion indicate that the signature is a property of the cohort rather than of any single batch or of low-quality samples.

The six prioritised candidates sit at different points of the interferon axis and carry different therapeutic implications. TLR7 gain-of-function genetic variation causes human lupus [5], and dual TLR7/8 inhibition has shown evidence of activity in cutaneous lupus, making TLR7 the most directly tractable candidate. IFIH1 (MDA5) mutations cause SLE with selective IgA deficiency [6], supporting a causal role for nucleic-acid sensing upstream of interferon production. STAT1 expression is elevated in lupus B cells and correlates with disease activity [7], consistent with the upregulation observed here. The Janus kinases present a more nuanced picture: JAK1 was modestly downregulated while JAK2 was upregulated, yet JAK1-selective inhibition (upadacitinib) [8] and JAK1/2 inhibition (baricitinib) [9] have both been investigated in SLE trials, so transcript direction alone does not determine the intervention. USP18 requires the most caution: as a negative regulator of type I interferon signalling whose human deficiency causes severe interferonopathy [10], its strong upregulation (+ 2.20) may be compensatory, and inhibiting it could worsen interferon-driven inflammation. For all six genes, independent SLE validation, B-cell protein validation and functional perturbation remain to be performed, and no therapeutic direction is claimed.

Several limitations must be stated plainly. First, the study is observational and computational, based on a single public cohort; the sensitivity analyses test robustness within that cohort and do not constitute independent validation. Second, the original reasons for excluding four batches from the primary analysis were not documented and could not be established; the exclusion is disclosed here and its effect is bounded by the leave-one-batch-out analysis, which used the fuller sample set, but it remains a weakness of the primary analysis. Third, some upstream inputs are stored locally and the complete workflow has not been independently repeated from the raw sequencing files, although the saved count matrices and reconstruction scripts are provided. Fourth, batch effects, medication exposure and other clinical confounders may influence expression, and flare-versus-managed subgroup comparisons in this dataset are substantially batch-confounded. Fifth, transcript abundance does not measure protein

abundance or signalling activity. Sixth, the machine-learning performance metrics are not validated predictions, for the reasons given above.

Future work should prioritise independent validation in additional SLE cohorts, B-cell-subset-specific analysis, protein-level and signalling measurements, and functional perturbation of the six candidates, together with integration of human genetics, druggability and safety evidence. The full results, scripts and sensitivity outputs are publicly available to support that work.

5. Conclusion

B cells in SLE show a robust interferon-associated transcriptional programme that survives low-cell-count filtering and leave-one-batch-out validation. Six candidates — TLR7, USP18, JAK1, JAK2, IFIH1 and STAT1 — are prioritised for independent and experimental validation. The results are hypothesis-generating and require functional follow-up before any therapeutic conclusion can be drawn.

Declarations

Data availability

The analysed dataset is the public single-cell RNA-seq dataset GSE174188, available through the NCBI Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE174188). All differential-expression results, enrichment tables, sensitivity outputs and figures are available in the public repository https://github.com/hamzabilal6353-cmd/SLE-RNA-Therapeutic-Discovery. The Gene Ontology re-analysis script and full results table are provided as supplementary files.

Code availability

Analysis scripts are available in the same public repository (https://github.com/hamzabilal6353-cmd/SLE-RNA-Therapeutic-Discovery), including the donor-aware differential-expression reconstruction, the filtering sensitivity script and the leave-one-batch-out function. Some upstream inputs are stored locally and are not included in the repository; independent repetition of the complete workflow from the raw sequencing files has not yet been shown.

Competing interests

The author declares no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Acknowledgements

The author thanks the investigators of the GSE174188 cohort for making their data public. AI assistance (Muse) was used for drafting and language improvement, for verifying reported numbers against the repository files, and for the documented Gene Ontology re-analysis in Section 2.3; the author reviewed and takes responsibility for all content.

References

1. Perez RK et al (2022) Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. Science. ;376(6589):eabf1970. PMID: 35389781
2. Ritchie ME, Phipson B, Wu D et al (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43(7):e47 PMID: 25605792
3. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1):139–140 PMID: 19910308
4. Law CW, Chen Y, Shi W, Smyth GK (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. ;15(2):R29. PMID: 24485249
5. Brown GJ et al (2022) TLR7 gain-of-function genetic variation causes human lupus. Nature 605(7909):349–356 PMID: 35477763
6. Van Eyck L et al (2015) Brief report: IFIH1 mutation causes systemic lupus erythematosus with selective IgA deficiency. Arthritis Rheumatol 67(6):1592–1597 PMID: 25777993
7. Aue A et al (2020) Elevated STAT1 expression but not phosphorylation in lupus B cells correlates with disease activity and increased plasmablast susceptibility. Rheumatology (Oxford) 59(11):3435–3442 PMID: 32357246
8. Merrill JT et al (2024) Efficacy and safety of upadacitinib or elsubrutinib alone or in combination for patients with systemic lupus erythematosus: a phase 2 randomized controlled trial. Arthritis Rheumatol 76(10):1518–1529 PMID: 38923871
9. Morand EF et al (2023) Baricitinib for systemic lupus erythematosus: a double-blind, randomised, placebo-controlled, phase 3 trial (SLE-BRAVE-I). Lancet 401(10381):1001–1010 PMID: 36848918
10. Meuwissen MEC et al (2016) Human USP18 deficiency underlies type 1 interferonopathy leading to severe pseudo-TORCH syndrome. J Exp Med 213(7):1163–1174 PMID: 27325888

Figures

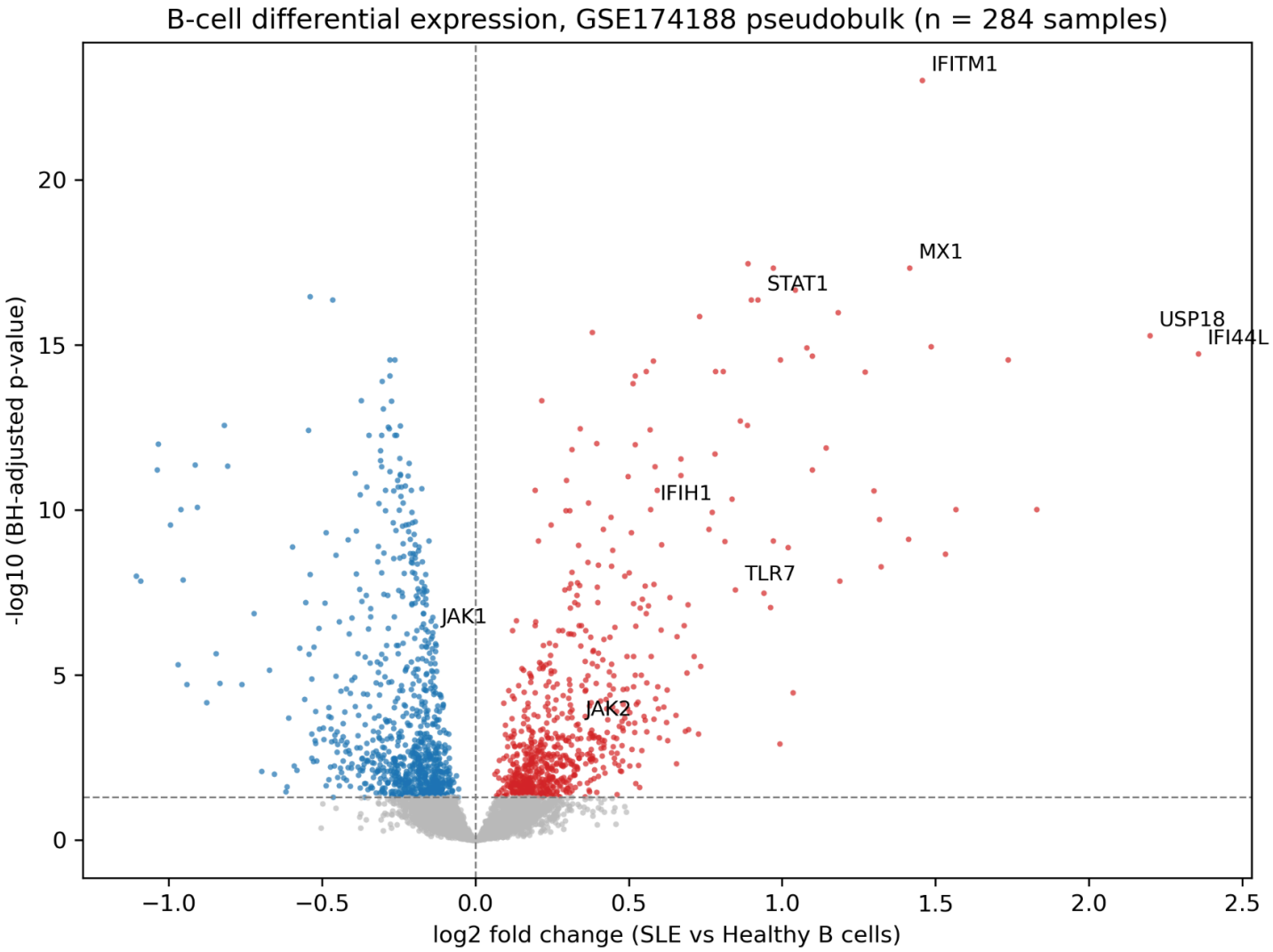

Figure 1

Volcano plot of B-cell differential expression (SLE versus healthy; 284 pseudobulk samples; 7,917 genes tested). Red points are significantly upregulated and blue points significantly downregulated at FDR < 0.05; grey points are not significant. The six prioritised candidates and the top interferon-stimulated genes IFI44L, IFITM1 and MX1 are labelled. The dashed horizontal line marks FDR = 0.05.

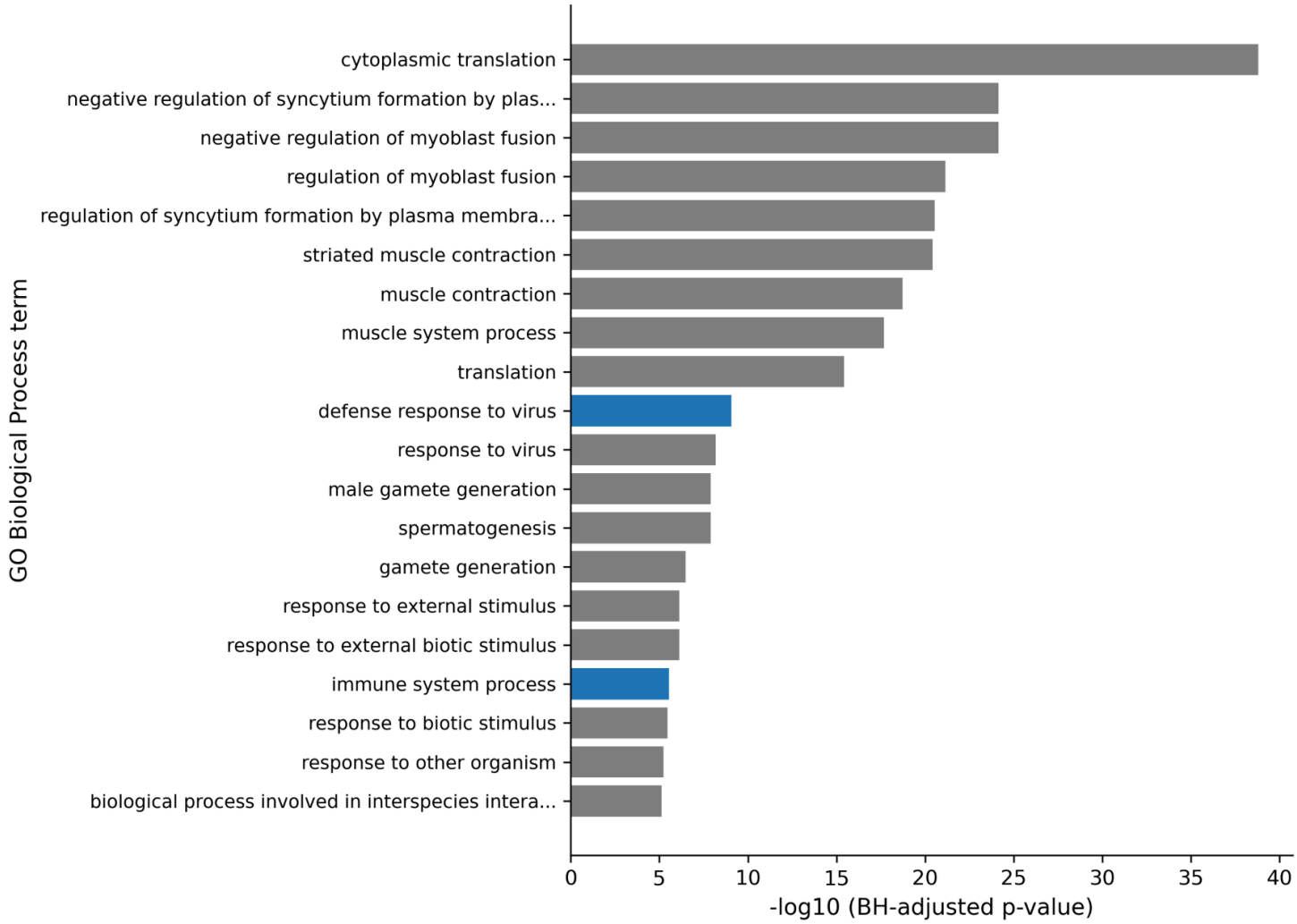

Figure 2

Gene Ontology Biological Process enrichment among significant B-cell differentially expressed genes (re-run; Section 2.3). Bars show -log10 Benjamini-Hochberg adjusted p-values for the 20 most significant terms; immune-, antiviral-, interferon- or defence-related terms are shown in blue. 66 terms were significant in total, including 21 immune-related terms.

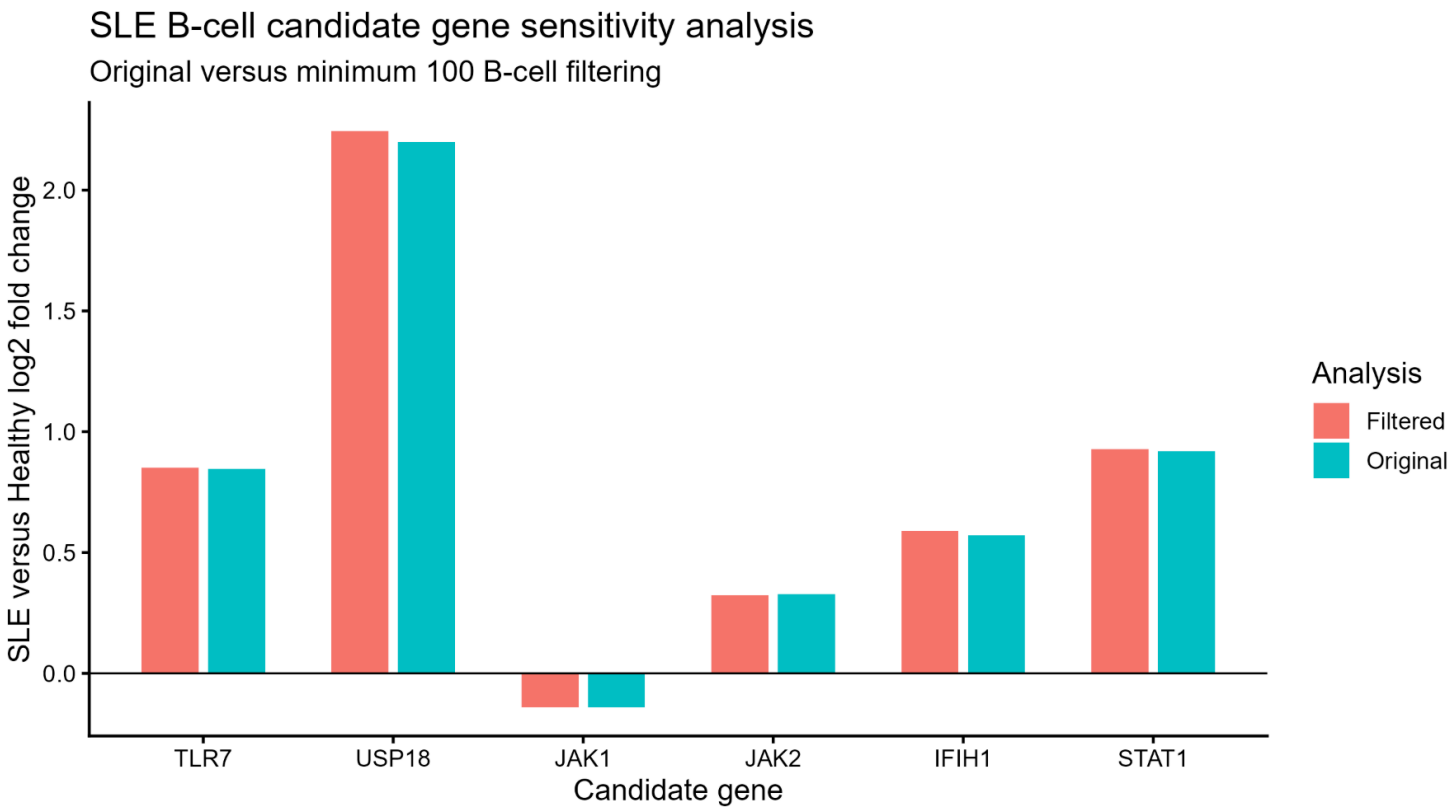

Figure 3

Six-candidate sensitivity analysis: log2 fold changes in the original analysis (284 samples) versus after minimum-100-B-cell filtering (259 samples). All six genes retained their direction and significance.

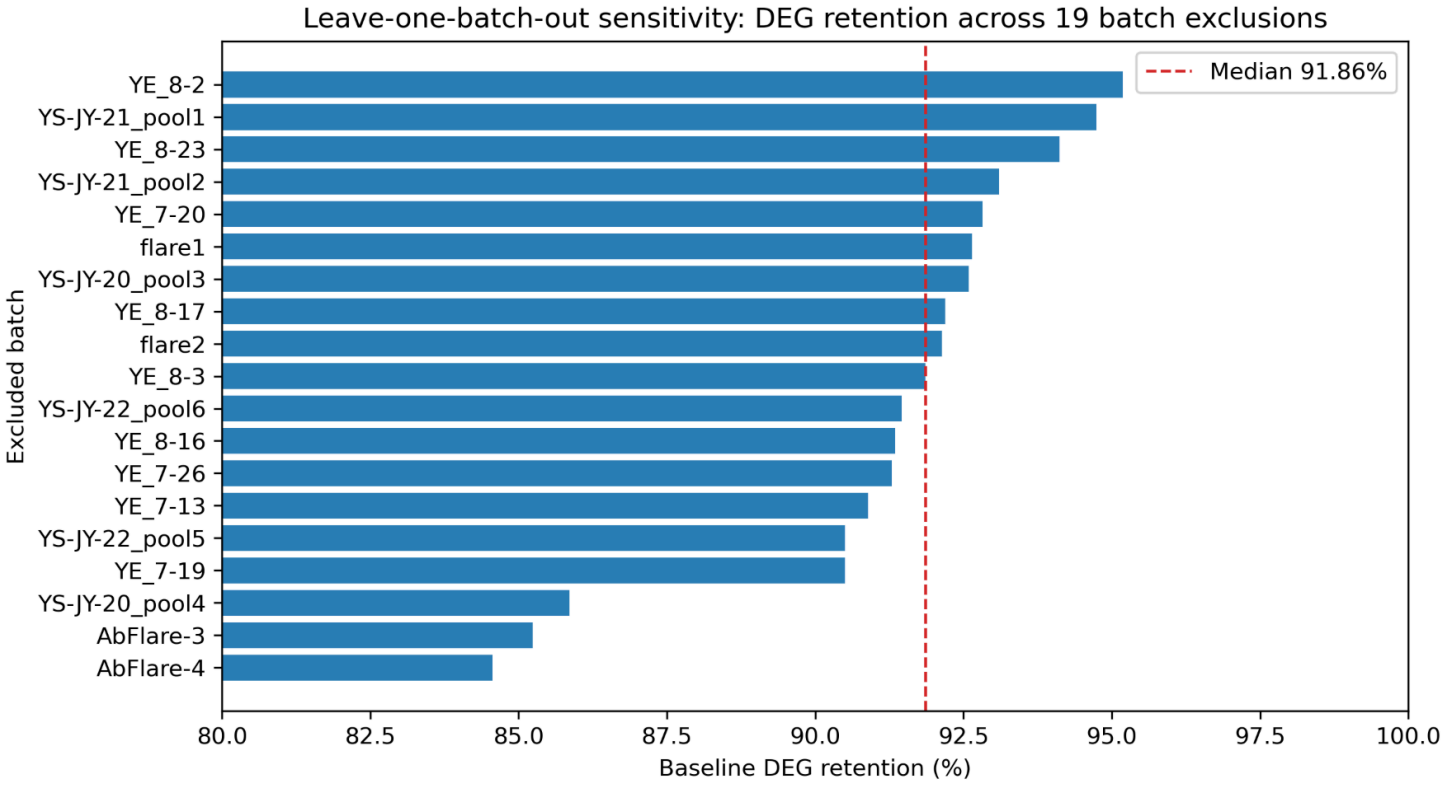

Figure 4

Leave-one-batch-out sensitivity: percentage of baseline significant genes retained after excluding each of 19 sequencing batches. The dashed line marks the median retention of 91.86%.

Supplementary Files

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

- SLEBcellGOBPrerunSupplementary.csv
- SLEBcellGOrerunscript.py