Figures and figure supplements

Longitudinal proteomic profiling of dialysis patients with COVID-19 reveals markers of severity and predictors of death

Jack Gisby *et al*

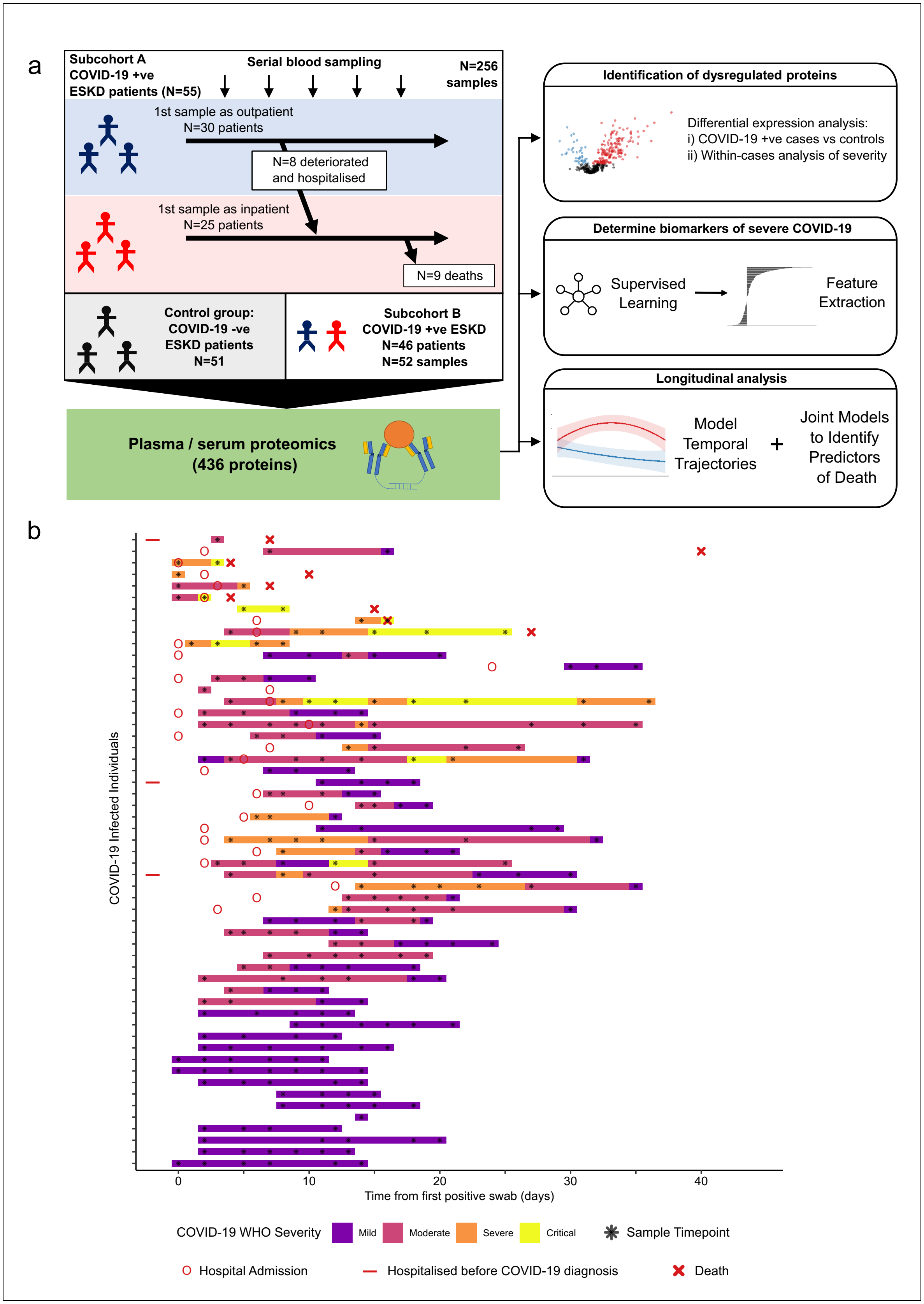
a
Subcohort A
COVID-19 +ve
ESKD patients (N=55)
Serial blood sampling
N=256
samples
1st sample as outpatient
N=30 patients
N=8 deteriorated
and hospitalised
1st sample as inpatient
N=25 patients
N=9 deaths
Control group:
COVID-19 -ve
ESKD patients
N=51
Subcohort B
COVID-19 +ve ESKD
N=46 patients
N=52 samples
Plasma / serum proteomics
(436 proteins)
Identification of dysregulated proteins
Differential expression analysis:
i) COVID-19 +ve cases vs controls
ii) Within-cases analysis of severity
Determine biomarkers of severe COVID-19
Supervised
Learning
Feature
Extraction
Longitudinal analysis
Model
Temporal
Trajectories
+
Joint Models
to Identify
Predictors
of Death
b
COVID-19 Infected Individuals
0
10
20
30
40
Time from first positive swab (days)
COVID-19 WHO Severity
Mild
Moderate
Severe
Critical
Sample Timepoint
Hospital Admission
Hospitalised before COVID-19 diagnosis
Death

Figure 1. Study design. (**a**) Schematic representing a summary of the patient cohorts, sampling, and the major analyses. Blue and red stick figures represent outpatients and hospitalised patients, respectively. (**b**) Timing of serial blood sampling in relation to clinical course of COVID-19 (subcohort A). Black asterisks indicate when samples were obtained. Three patients were already in hospital prior to COVID-19 diagnosis (indicated by red bars).

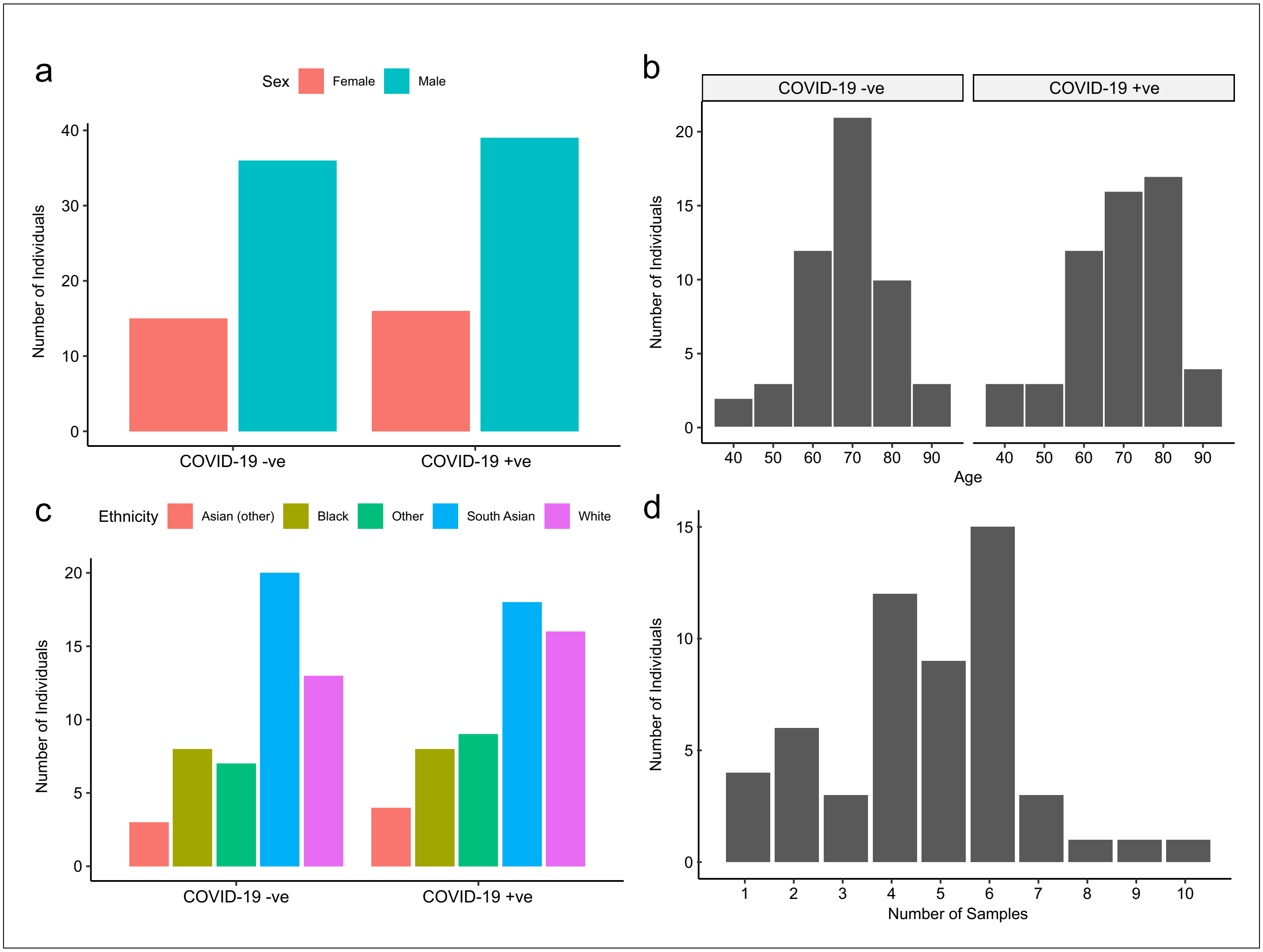

Figure 1—figure supplement 1. Baseline characteristics of subcohort A. The number of COVID-19-positive and -negative patients in subcohort A (plasma), stratified by (**a**) sex, (**b**) age, and (**c**) ethnicity. (**d**) Serial samples obtained for COVID-19 patients.

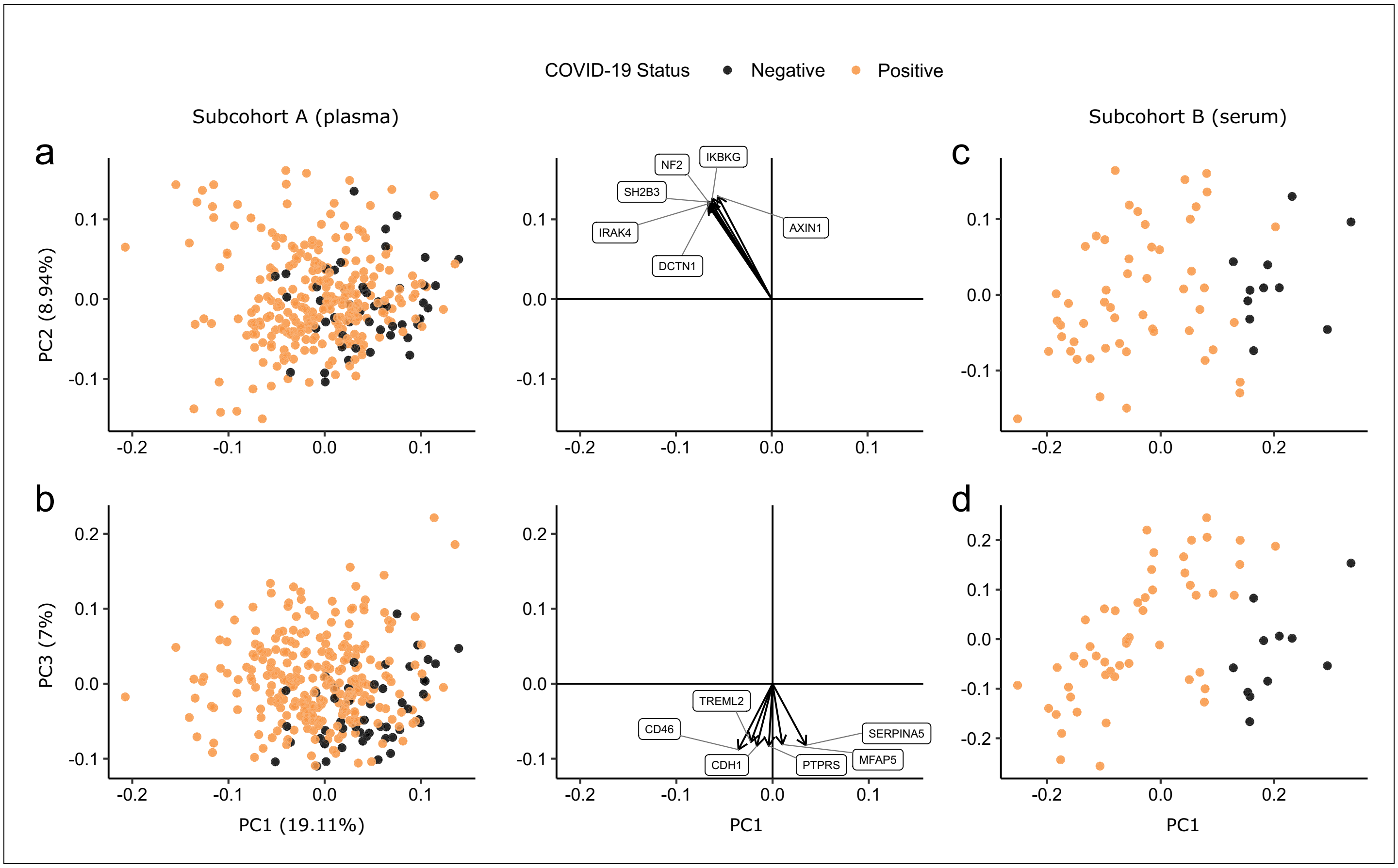

Figure 2. Principal component analysis. PC = principal component. Each point represents a sample. Colouring indicates COVID-19 status. The directions and relative sizes of the six largest PC loadings are plotted as arrows (middle column). (**a**, **b**) Subcohort A. Due to serial sampling, there are multiple samples for most patients. The proportion of variance explained in subcohort A by each PC is shown in parentheses on the axis labels. (**c**, **d**) Subcohort B. Samples are projected into the PCA coordinates from subcohort A.

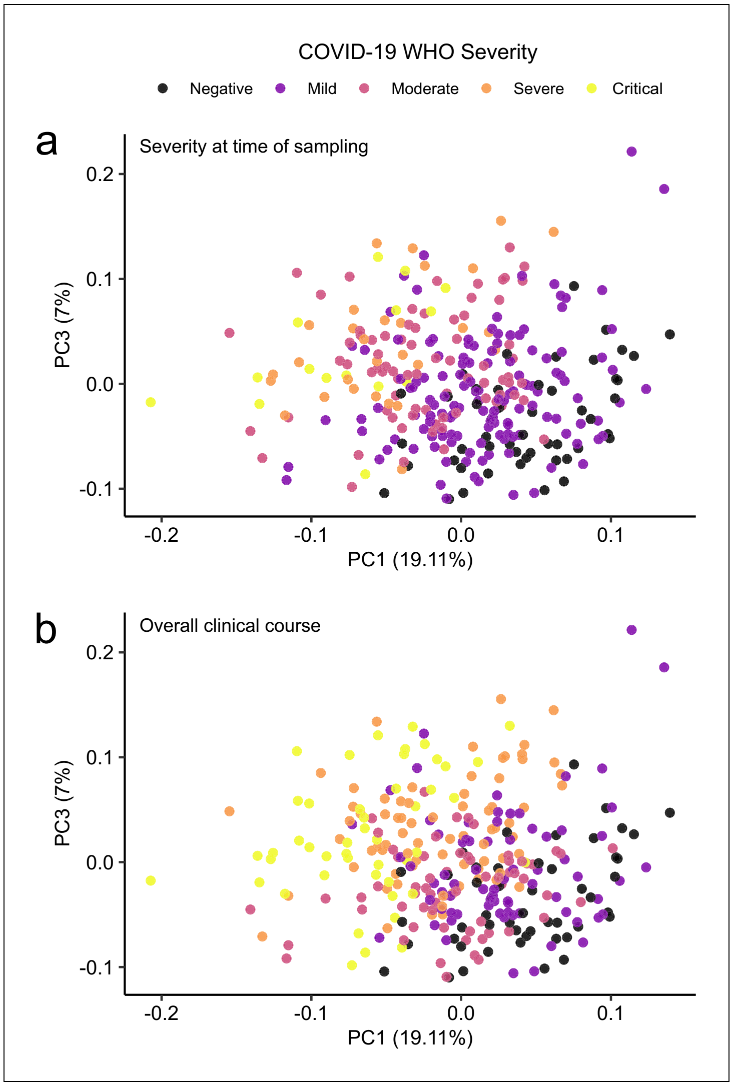

Figure 2—figure supplement 1. Principal component analysis in relation to clinical severity. (**a**) Colouring indicates WHO severity at time of sampling. (**b**) Colouring indicates overall clinical course (indicated by peak WHO severity) for the patient from which that sample was taken.

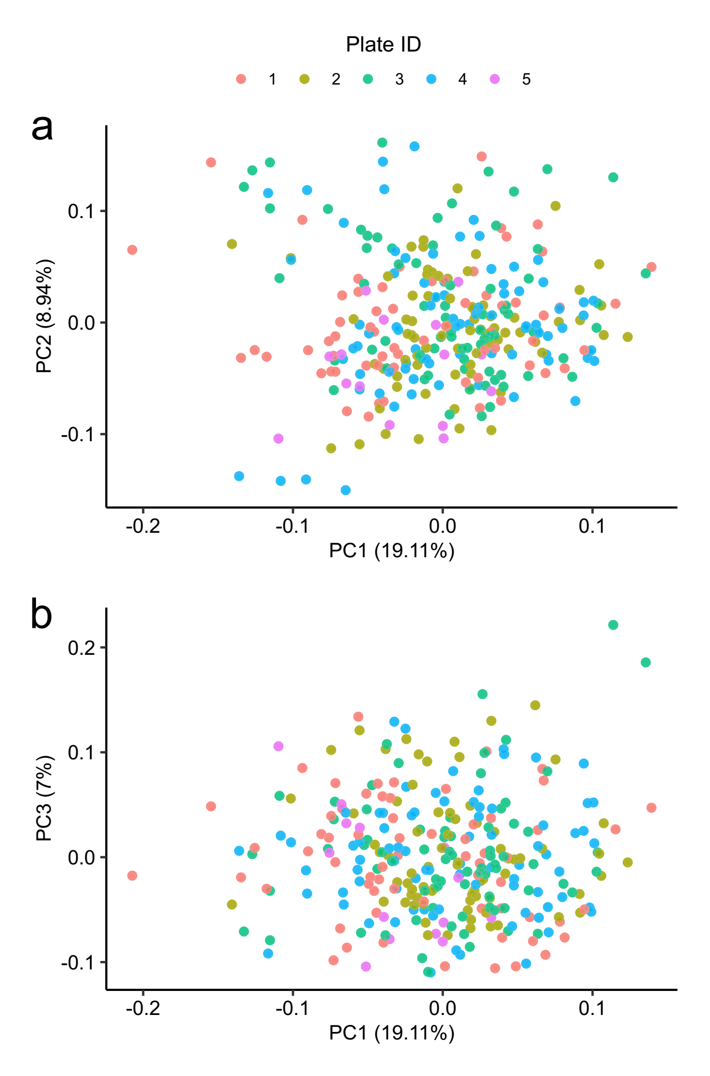

Figure 2—figure supplement 2. Principal component analysis in relation to assay plate. Principal component analysis of the subcohort A coloured by plate.

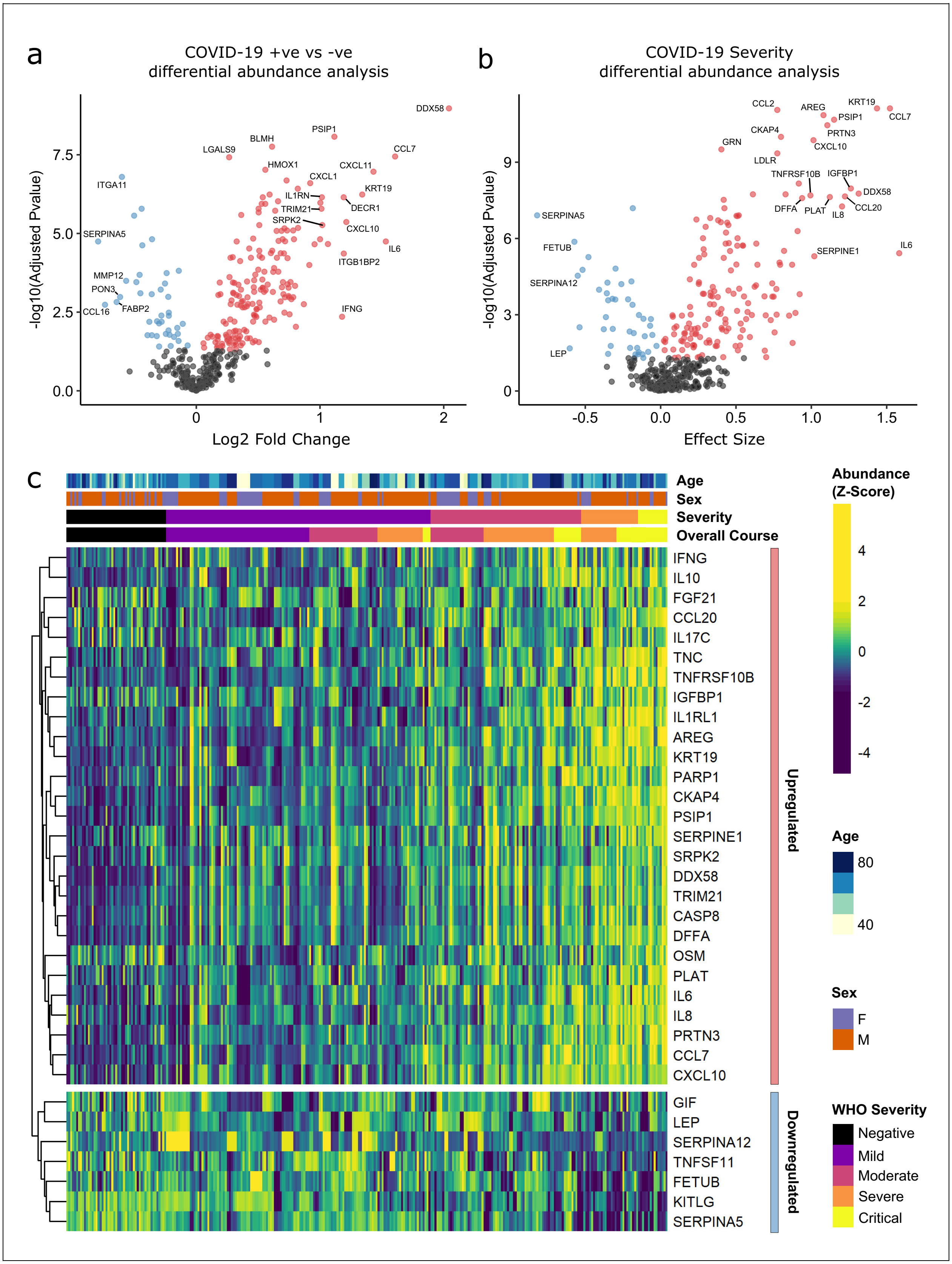

Figure 3. Identification of dysregulated proteins. (a) Proteins upregulated (red) or downregulated (blue) in COVID-19-positive patients versus COVID-19-negative ESKD patients n = 256 plasma samples from 55 COVID-19-positive patients, versus n = 51 ESKD controls (one sample per control patient).

Figure 3 continued on next page

Figure 3 continued

(**b**) Proteins associated with disease severity associations of protein levels against WHO severity score at the time of sampling. Linear gradient indicates the effect size. A positive effect size (red) indicates that an increase in protein level is associated with increasing disease severity and a negative gradient (blue) the opposite. n = 256 plasma samples from 55 COVID-19-positive patients. For (**a, b**), p-values from linear mixed models after Benjamini–Hochberg adjustment; significance threshold = 5% FDR; dark-grey = non-significant. (**c**) Heatmap showing protein levels for selected proteins with strong associations with severity. Each column represents a sample (n = 256 COVID-19 samples and 51 non-infected samples). Each row represents a protein. Proteins are annotated using the symbol of their encoding gene. For the purposes of legibility, not all significantly associated proteins are shown; the heatmap is limited to the 17% most up- or downregulated proteins (by effect size) of those with a significant association. Proteins are ordered by hierarchical clustering. Samples are ordered by WHO severity at the time of blood sample ('Severity'). 'Overall course' indicates the peak WHO severity over the course of the illness.

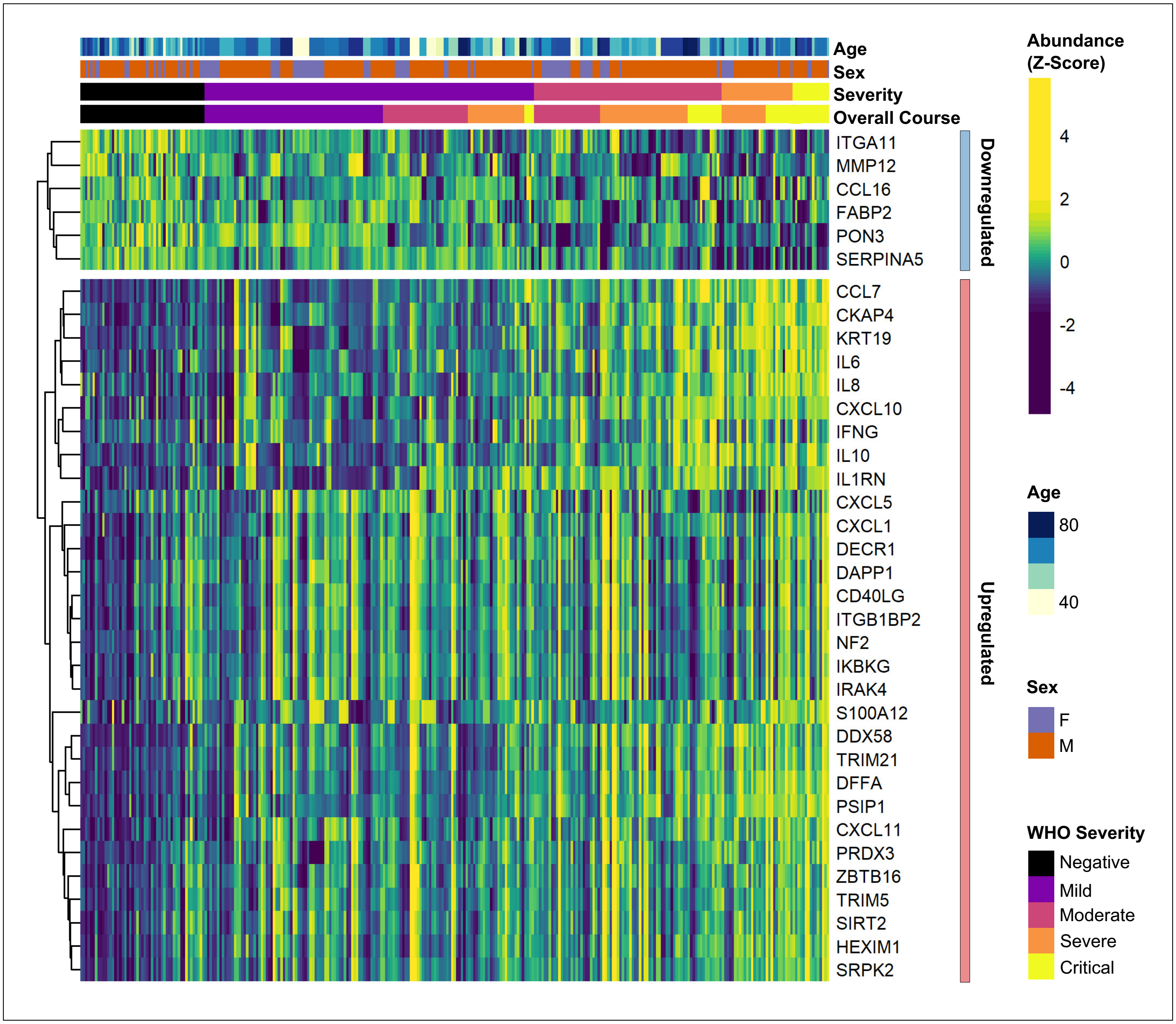

Figure 3—figure supplement 1. Differential abundance analysis between ESKD patients with and without COVID-19. Heatmap showing selected proteins with the largest fold changes in differential abundance analysis (subcohort A). As for ***Figure 3***, the heatmap is limited to the 17% most up- or downregulated proteins (by fold change) of those with a significant association.

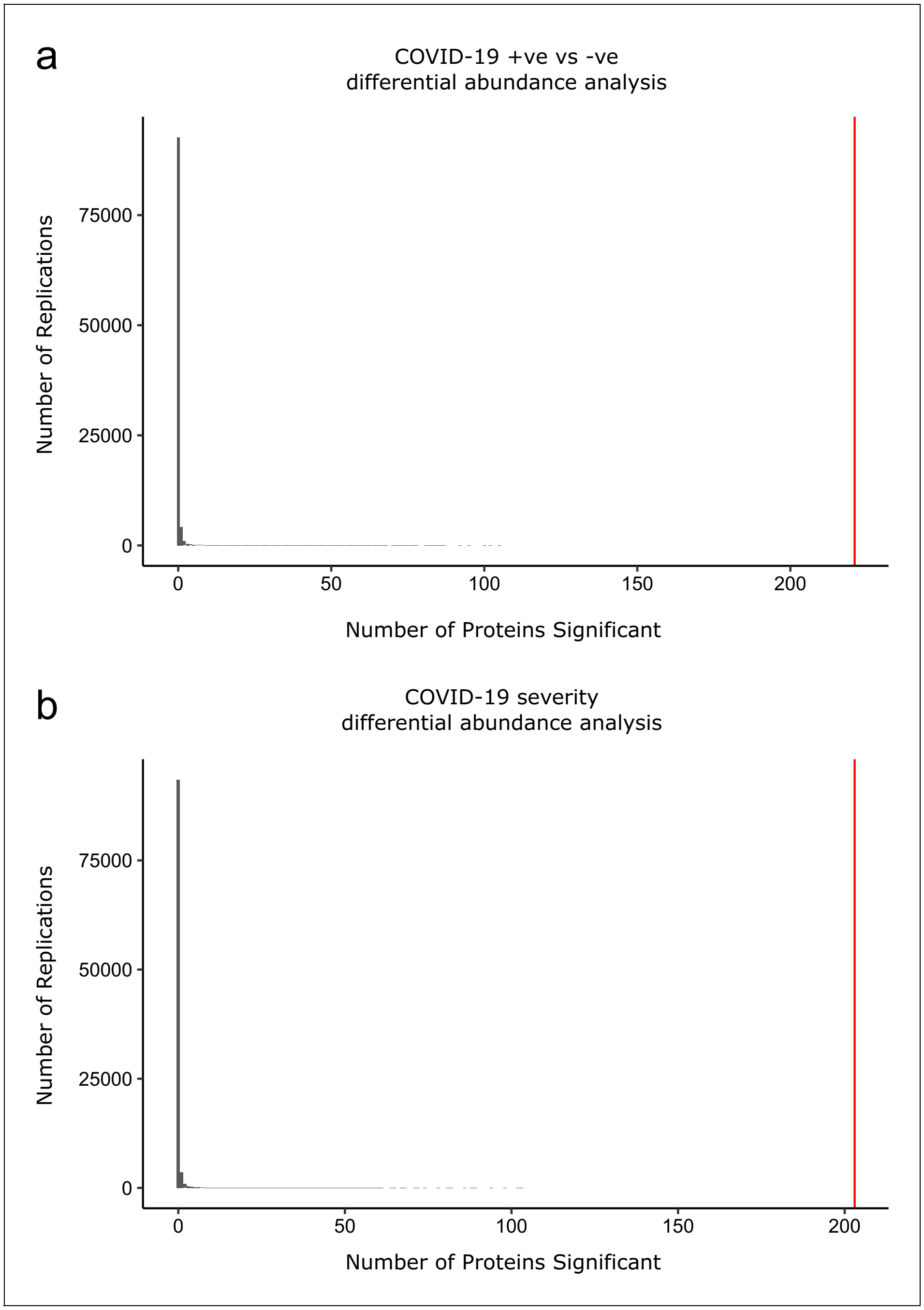

Figure 3—figure supplement 2. Permutation analysis to estimate the null distribution. Histogram showing the distribution of the number of associations declared significant (FDR 5%) after random permutation of class labels (100,000 replications). (**a**) The COVID-19 +ve versus −ve differential abundance analysis. (**b**) The COVID-19 severity differential abundance analysis. The vertical red line denotes the number of proteins we found significant in the analysis with the true sample labels.

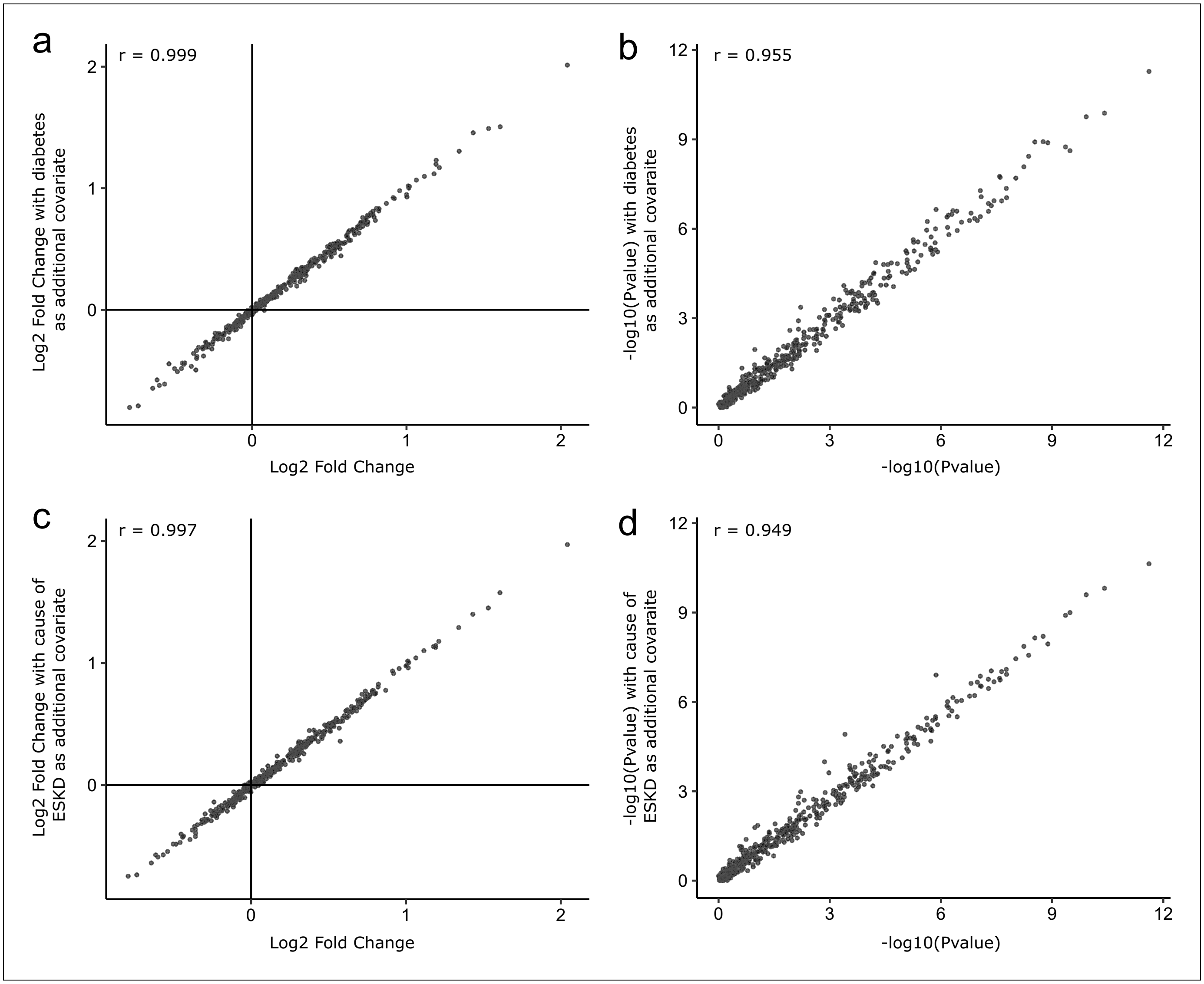

Figure 3—figure supplement 3. Sensitivity analyses adjusting for diabetes status and cause of ESKD. As sensitivity analyses, the COVID-19-positive versus -negative differential abundance regressions were repeated adding diabetes status (**a**, **b**) and cause of ESKD (**c**, **d**) as additional covariates. The basic model included age, sex, and ethnicity as covariates. Each point represents a protein. A comparison of −log10 p-values and effect sizes is shown for all 436 proteins. r indicates Pearson's correlation coefficient.

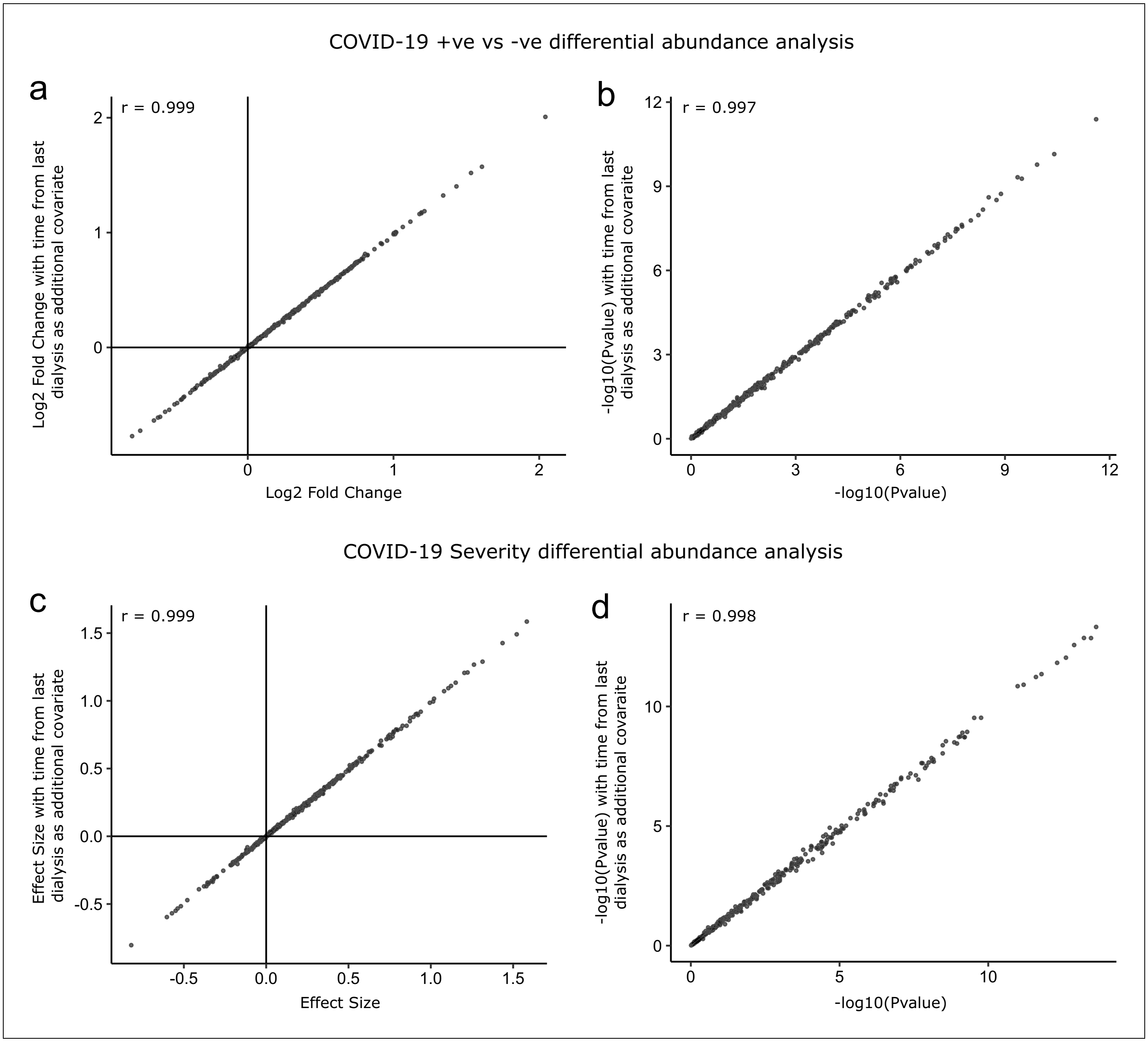

Figure 3—figure supplement 4. Sensitivity analysis adjusting for time since last haemodialysis. Comparison of results obtained with and without adding time since last haemodialysis as an additional covariate to the regression models. (**a**, **b**) COVID-19 positive versus negative differential expression analysis. (**c**, **d**) Severity analysis. Each point represents a protein. r indicates Pearson's correlation coefficient.

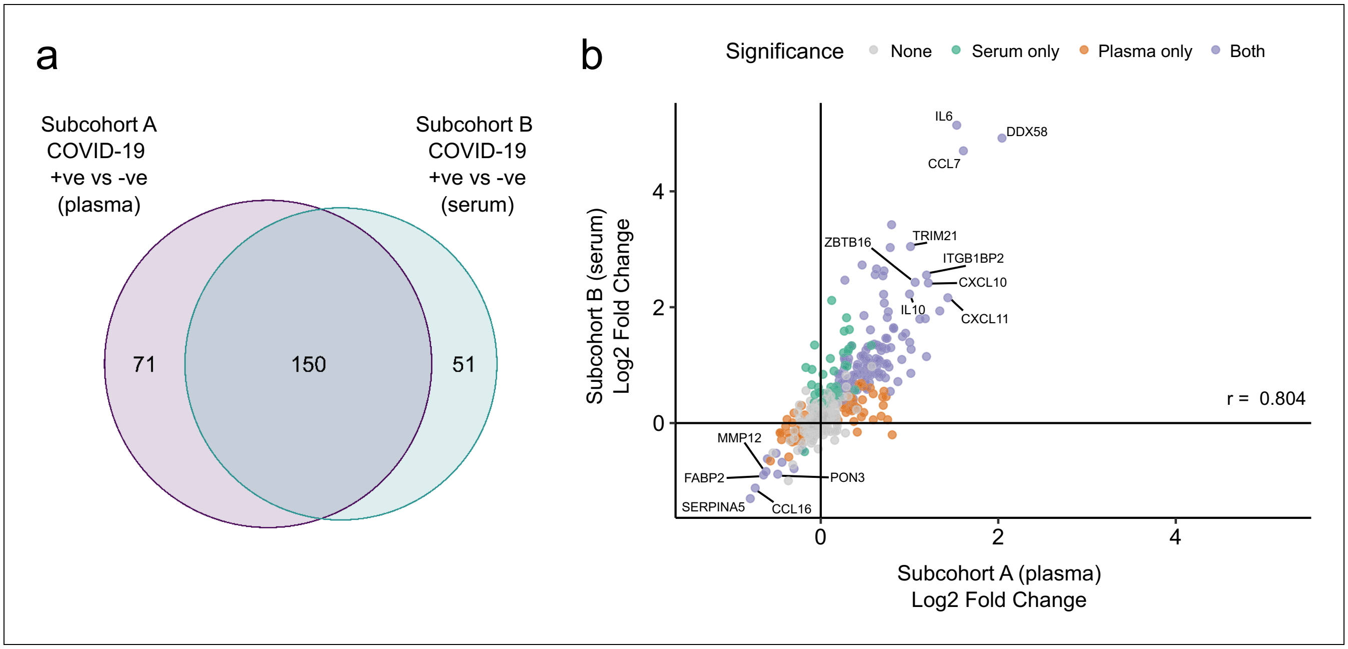

Figure 4. Validation. (**a**) Overlap between the significant associations in the differential abundance analysis between ESKD patients with and without COVID-19 in subcohorts A and B. 5% FDR was used as the significance threshold in both analyses. (**b**) Comparison of estimated effect sizes for all 436 proteins in the differential abundance analyses (COVID-19 positive versus negative) in subcohort A and B. Each point represents a protein. Pearson's r is shown. Differential abundance analyses were performed using linear mixed models. Subcohort A analysis (plasma samples): 256 samples from 55 COVID-19 patients versus 51 non-infected patient samples (single time-point). Subcohort B (serum samples): 52 samples from 55 COVID-19 patients and 11 non-infected patient samples (single timepoint).

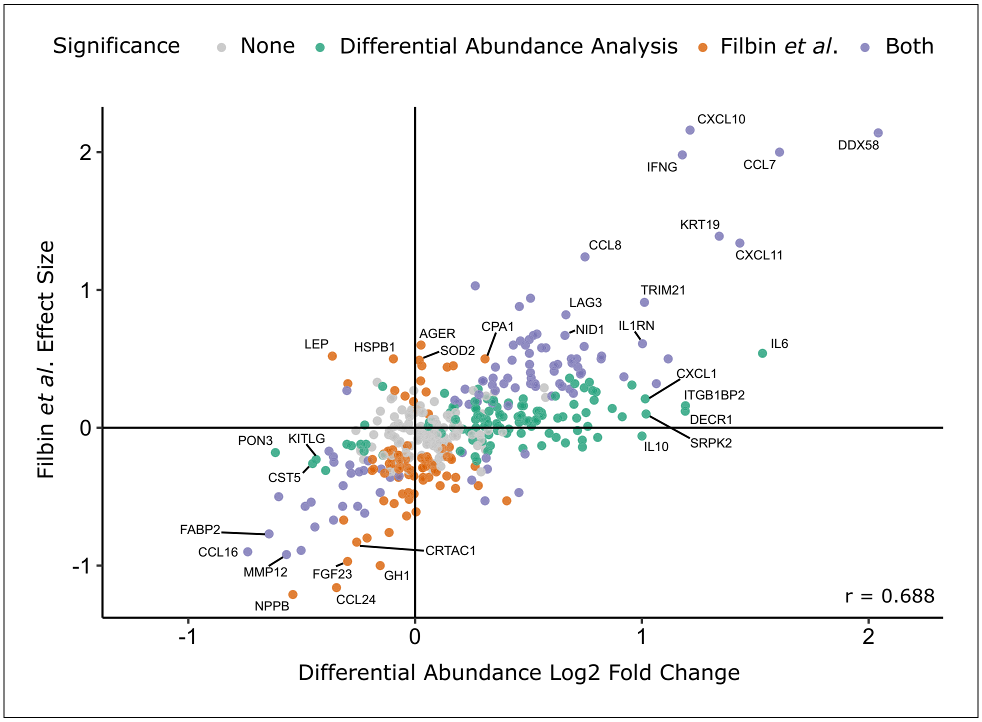

Figure 4—figure supplement 1. Comparison with the report of ***Filbin et al., 2020***. Comparison of log2 fold change for COVID-19-positive versus -negative ESKD patients in our study versus COVID-19-positive versus -negative respiratory distress patients in the report by ***Filbin et al., 2020***. Colours indicate whether a protein was significantly differentially abundant in each study. Pearson's r is shown.

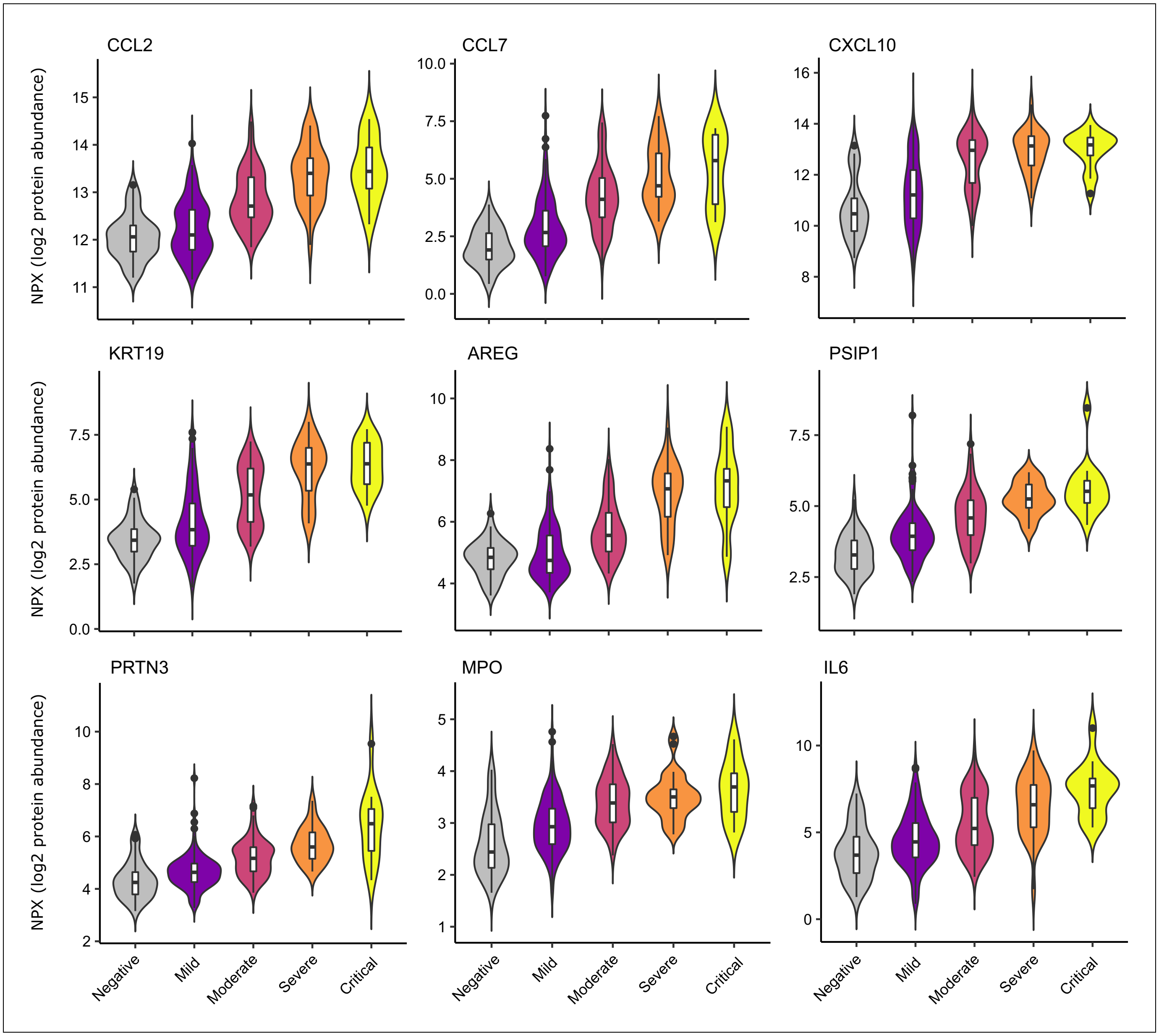

Figure 5. Selected proteins strongly associated with COVID-19 severity. Violin plots showing distribution of plasma protein levels according to COVID-19 status at the time of blood draw. Boxplots indicate median and inter-quartile range. n = 256 samples from 55 COVID-19 patients and 51 samples from non-infected patients. WHO severity indicates the clinical severity score of the patient at the time the sample was taken. Mild n = 135 samples; moderate n = 77 samples; severe n = 29 samples; critical n = 15 samples. Upper: monocyte chemokines. Middle: markers of epithelial injury. Lower: two neutrophil proteases and IL6.

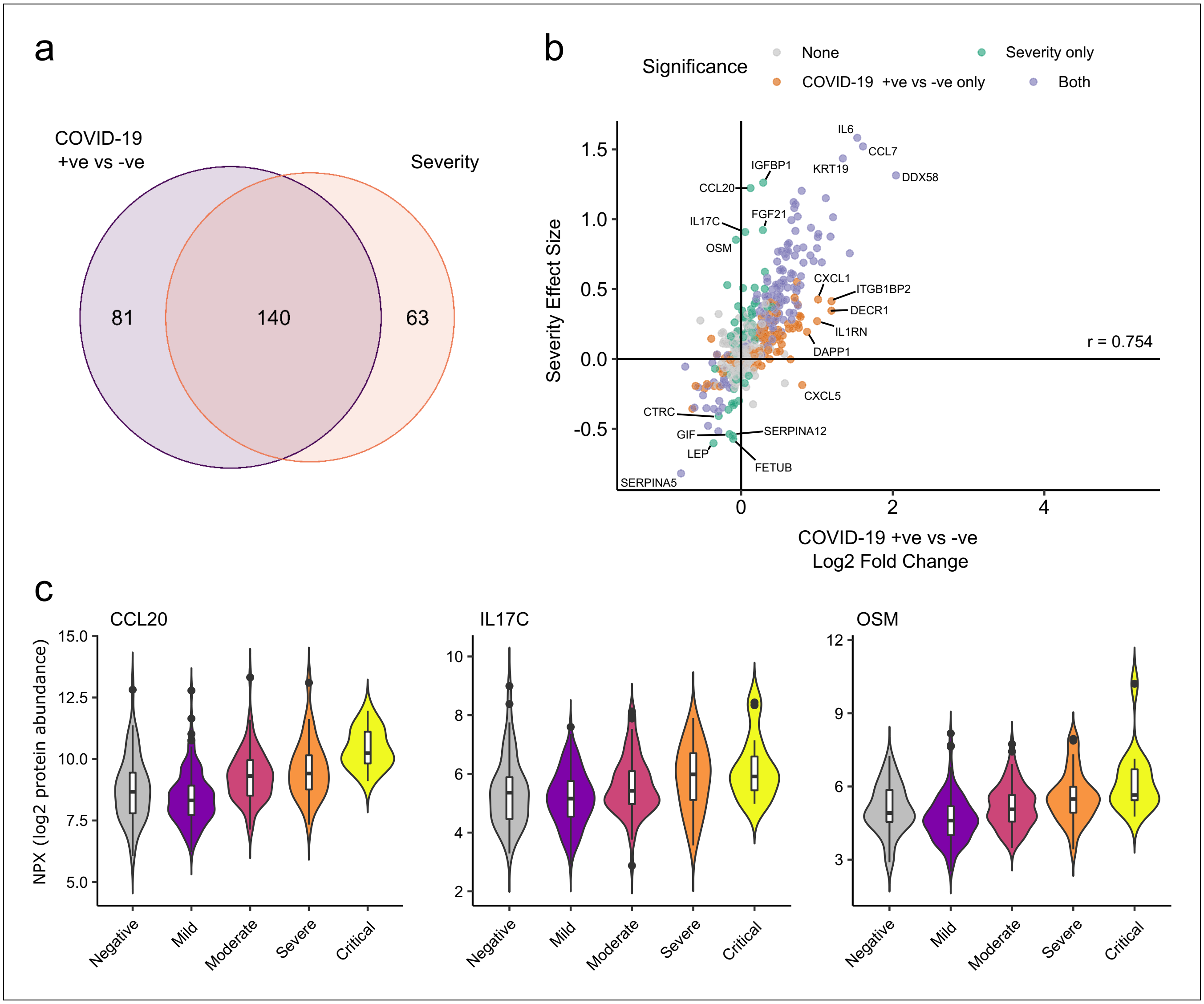

Figure 6. Comparison of proteins differentially expressed in COVID-19 with those associated with clinical severity. (**a**) Overlap between the proteins significantly differentially expressed in COVID-19 (n = 256 COVID-19 samples and 51 non-infected samples) versus those associated with severity (within-case analysis, n = 256 samples) (subcohort A). 5% FDR was used as the significant cut-off in both analyses. (**b**) Comparison of effect sizes for each protein in the COVID-19-positive versus -negative analysis (x-axis) and severity analysis (y-axis). Each point represents a protein. Pearson's r is shown. (**c**) Examples of proteins specifically associated with severity, but not significantly differentially abundant in the comparison of all cases versus controls. Violin plots showing distribution of plasma protein levels according to COVID-19 status at the time of blood draw. Boxplots indicate median and inter-quartile range. n = 256 samples from 55 COVID-19 patients and 51 samples from non-infected patients. WHO severity indicates the clinical severity score of the patient at the time the sample was taken. Mild n = 135 samples; moderate n = 77 samples; severe n = 29 samples; critical n = 15 samples.

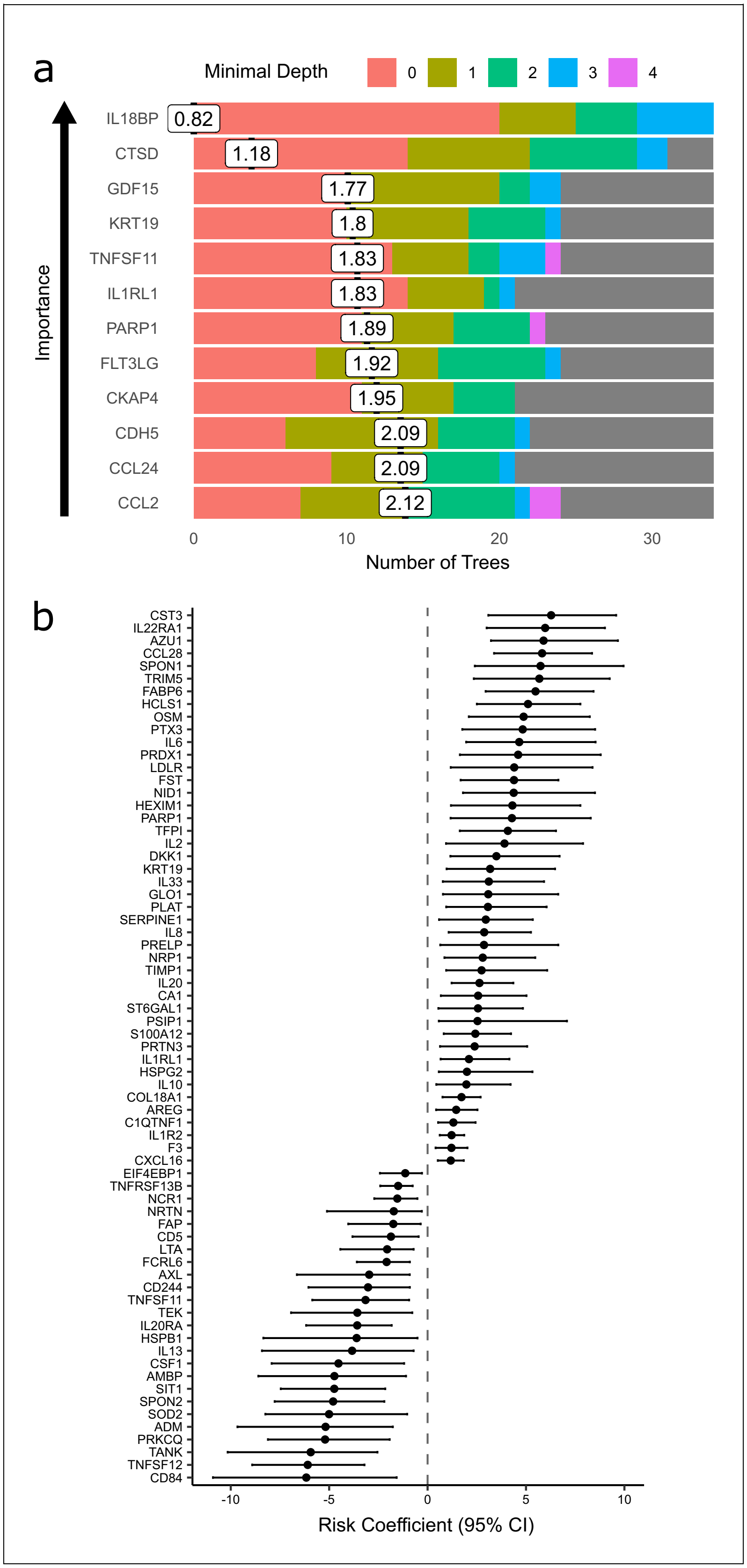

Figure 7. Prediction of severe COVID-19 and death. (a) The 12 most important proteins for predicting overall clinical course (defined by peak COVID-19 WHO severity) using Random Forests supervised learning. If a variable

Figure 7 continued on next page

Figure 7 continued

is important for prediction, it is likely to appear in many decision trees (number of trees) and be close to the root node (i.e. have a low minimal depth). The mean minimal depth across all trees (white box) was used as the primary feature selection metric. (**b**) Proteins that are significant predictors of death (Benjamini–Hochberg adjusted p<0.05). n = 256 samples from 55 COVID-19-positive patients, of whom nine died. Risk coefficient estimates are from a joint model. Bars indicate 95% confidence intervals. For proteins with a positive risk coefficient, a higher concentration corresponds to a high risk of death, and vice versa for proteins with negative coefficients.

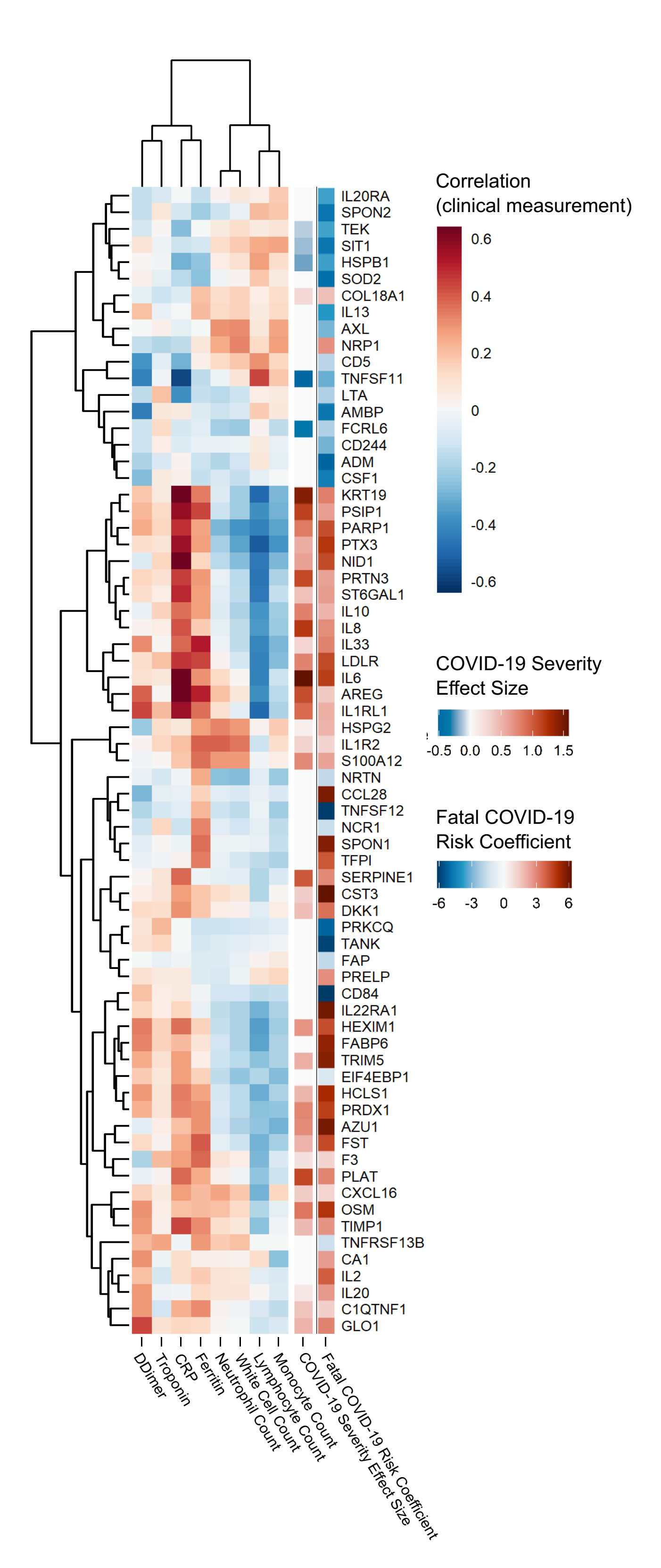

Figure 7—figure supplement 1. Proteins associated with risk of death: correlation to clinical severity and clinical laboratory measurements. Proteins significantly associated with risk of death (5% FDR) are shown. The estimated

Figure 7—figure supplement 1 continued on next page

Figure 7—figure supplement 1 continued

effect size from the linear mixed model testing association with severity is also shown. Correlations between protein levels and contemporaneous clinical laboratory marker values were calculated using rmcorr (***Bakdash and Marusich, 2017***) for each of the proteins significant (5% FDR) in the joint model. The rows and columns of the clinical marker correlation matrix are ordered by hierarchical clustering.

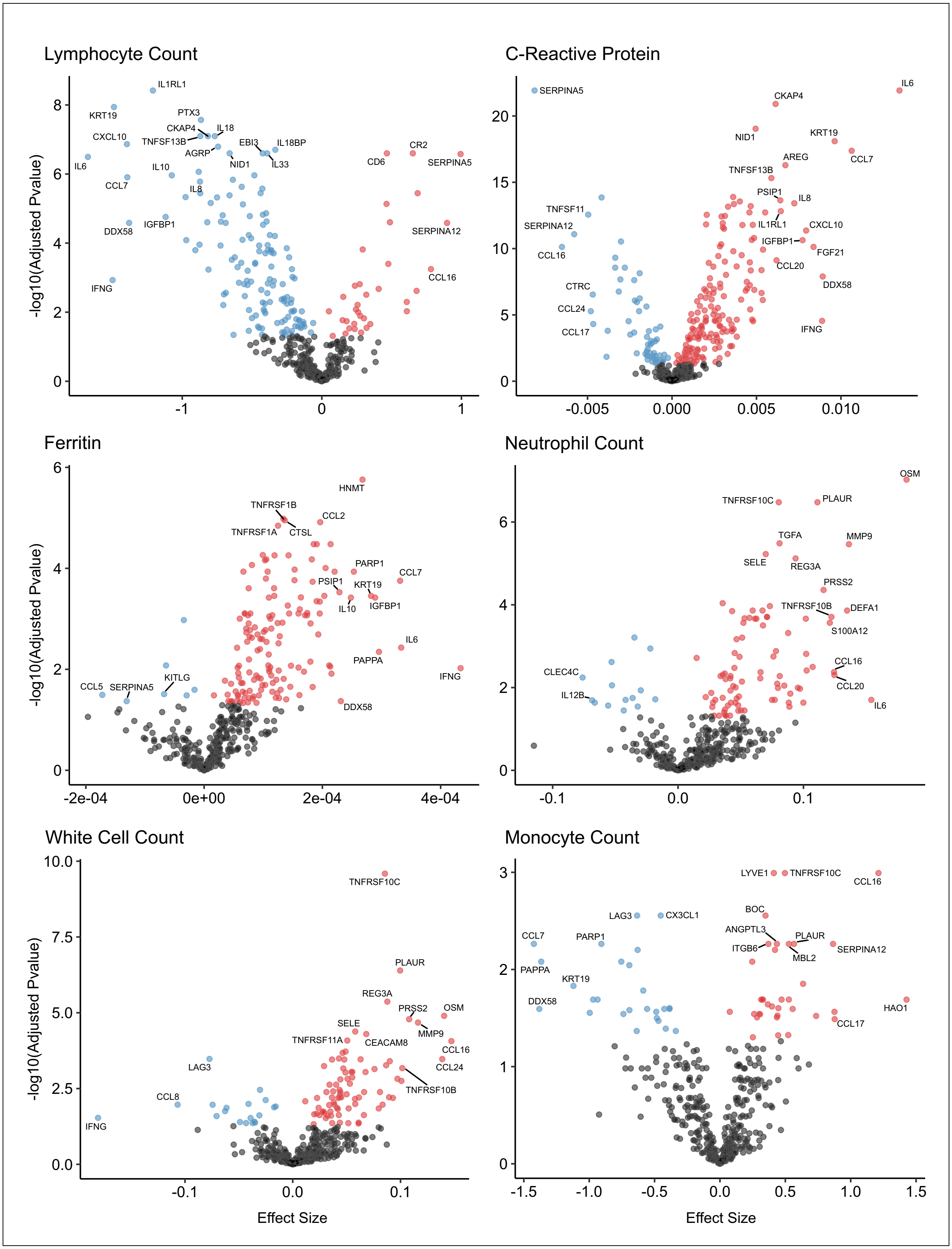
Lymphocyte Count
-log10(Adjusted Pvalue)
0
2
4
6
8
-1
0
1
IL1RL1
KRT19
PTX3
CKAP4
IL18
CXCL10
TNFSF13B
AGRP
EBI3
IL18BP
CR2
IL6
NID1
IL33
CD6
SERPINA5
IL10
CCL7
IL8
DDX58
IGFBP1
SERPINA12
CCL16
IFNG
C-Reactive Protein
0
5
10
15
20
-0.005
0.000
0.005
0.010
SERPINA5
IL6
CKAP4
NID1
KRT19
CCL7
AREG
TNFSF13B
PSIP1
IL8
TNFSF11
SERPINA12
IL1RL1
CXCL10
IGFBP1
FGF21
CCL16
CCL20
DDX58
CTRC
CCL24
CCL17
IFNG
Ferritin
-log10(Adjusted Pvalue)
0
2
4
6
-2e-04
0e+00
2e-04
4e-04
HNMT
TNFRSF1B
CCL2
TNFRSF1A
CTSL
PARP1
CCL7
PSIP1
KRT19
IL10
IGFBP1
IL6
PAPPA
IFNG
KITLG
CCL5
SERPINA5
DDX58
Neutrophil Count
0
2
4
6
-0.1
0.0
0.1
OSM
TNFRSF10C
PLAUR
TGFA
MMP9
SELE
REG3A
PRSS2
TNFRSF10B
DEFA1
S100A12
CCL16
CLEC4C
CCL20
IL12B
IL6
White Cell Count
-log10(Adjusted Pvalue)
0.0
2.5
5.0
7.5
10.0
-0.1
0.0
0.1
Effect Size
TNFRSF10C
PLAUR
REG3A
PRSS2
OSM
SELE
MMP9
TNFRSF11A
CEACAM8
CCL16
LAG3
CCL24
TNFRSF10B
CCL8
IFNG
Monocyte Count
0
1
2
3
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Effect Size
LYVE1
TNFRSF10C
CCL16
BOC
LAG3
CX3CL1
ANGPTL3
CCL7
PARP1
PLAUR
ITGB6
SERPINA12
MBL2
PAPPA
KRT19
DDX58
HAO1
CCL17

Figure 8. Associations of clinical laboratory markers with plasma proteins. Proteins that are positively (red) or negatively (blue) associated with clinical laboratory parameters (5% FDR). p-values from differential abundance analysis using linear mixed models after Benjamini–Hochberg adjustment. Dark-grey = non-significant. Two associations were found for d-dimer (not shown – see ***Supplementary file 1g***).

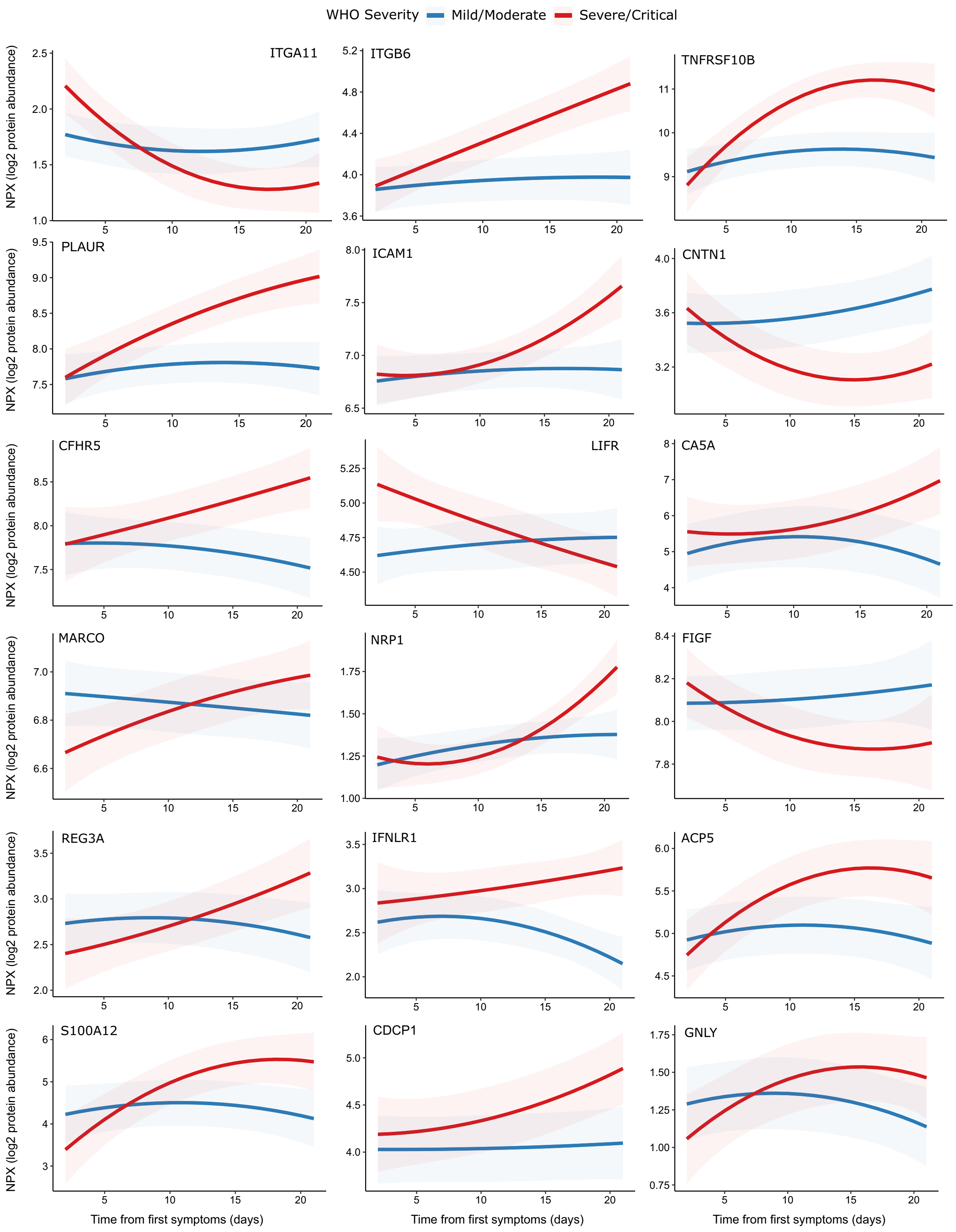

Figure 9. Modelling of temporal protein trajectories. The top 18 proteins displaying the most significantly (5% FDR) different longitudinal trajectories between patients with a mild or moderate (n = 28) versus severe or critical (n = 27) overall clinical course (defined by peak WHO severity). Means and

Figure 9 continued on next page

Figure 9 continued

95% confidence intervals for each group, predicted using linear mixed models (see Materials and methods), are plotted. The remainder of significant proteins are shown in ***Figure 9—figure supplement 1***. Individual data points are shown in ***Figure 9—figure supplement 2***.

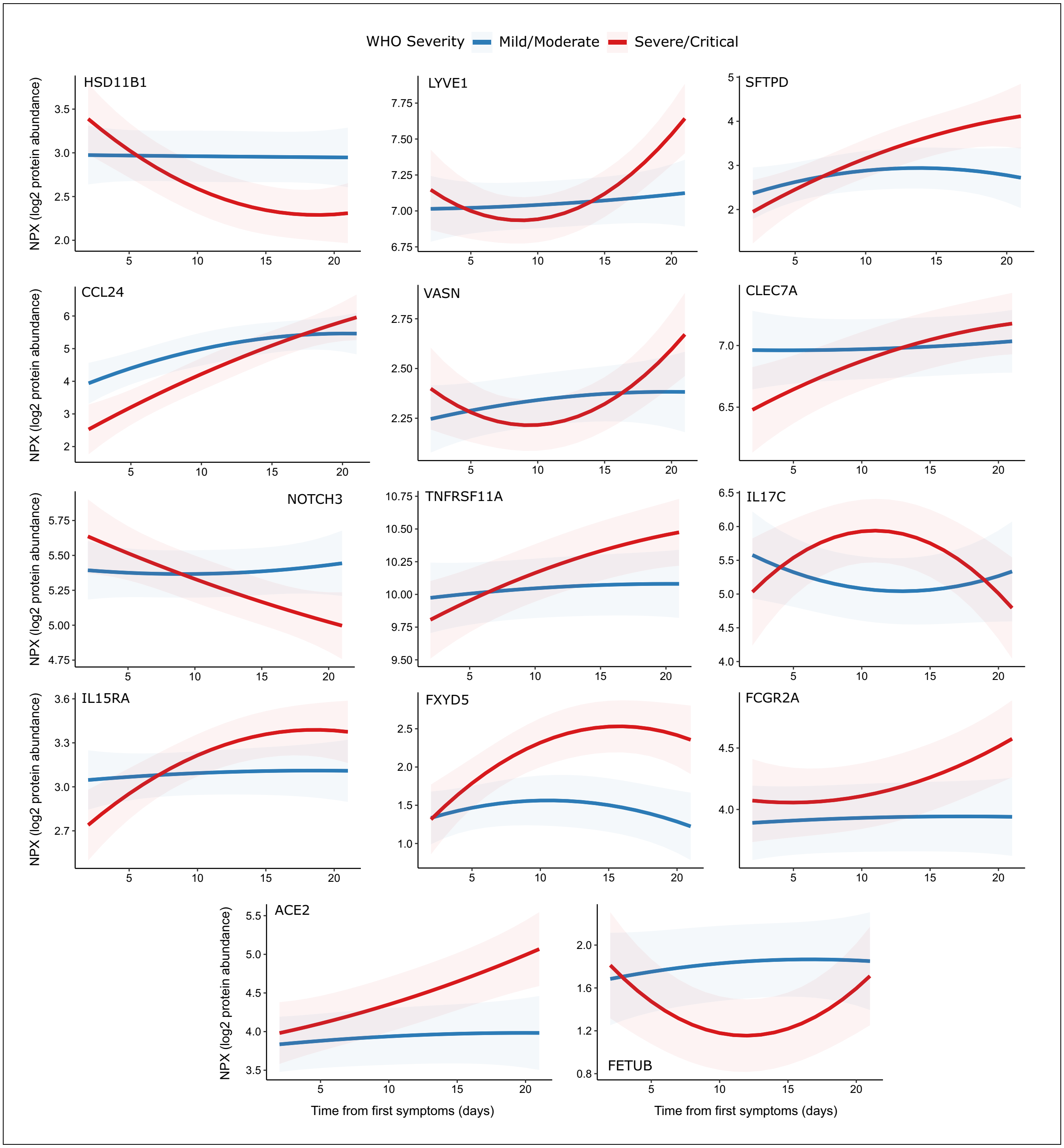

Figure 9—figure supplement 1. Display of modelled temporal trajectories for other proteins with a significant time × severity interaction. Proteins significant at 5% FDR but not shown in ***Figure 9*** is displayed here.

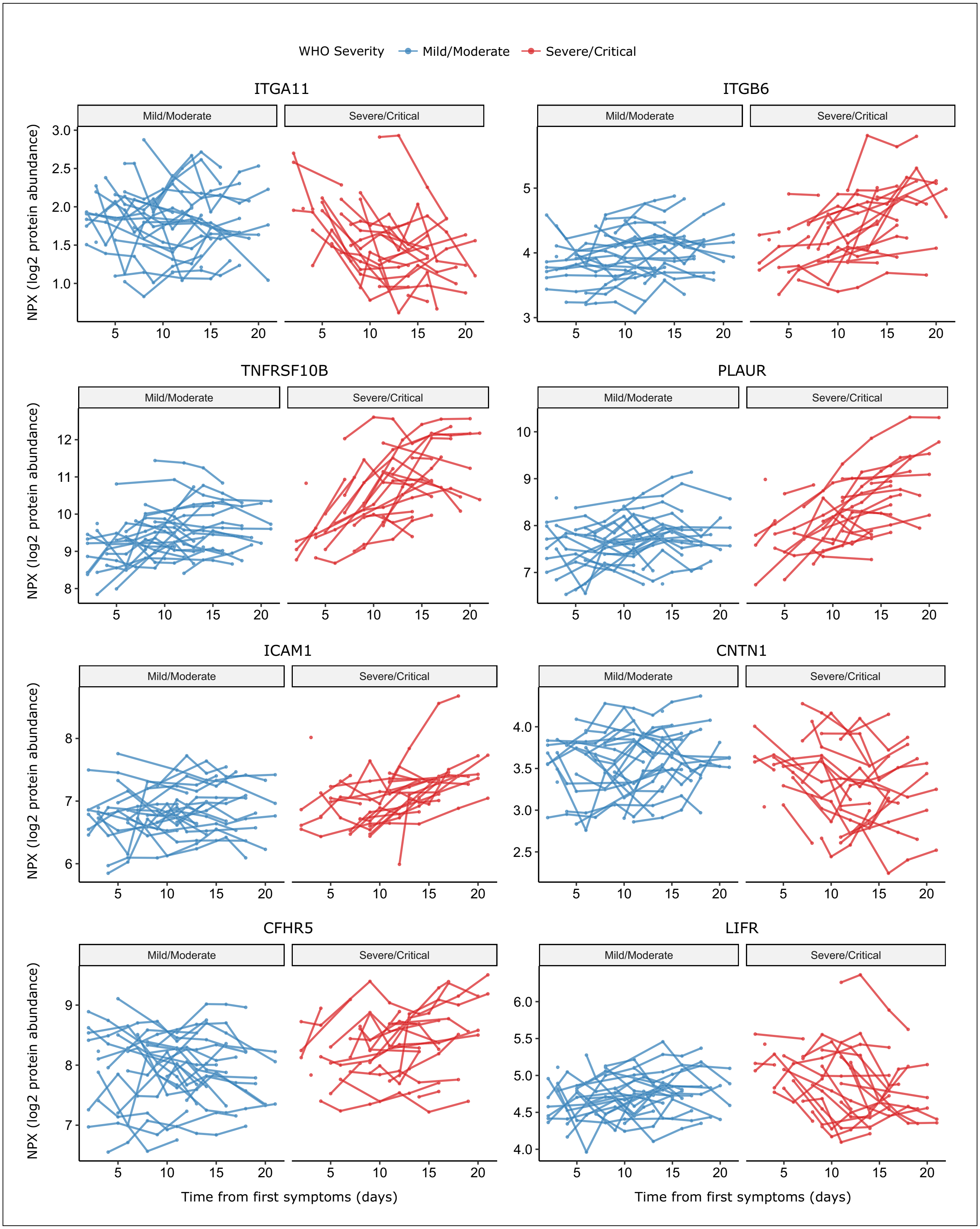

Figure 9—figure supplement 2. Raw data points for modelling of temporal protein trajectories. The eight most significant proteins from ***Figure 9*** are displayed.