Figures and figure supplements

Epigenetic scores for the circulating proteome as tools for disease prediction

Danni A Gadd, Robert F Hillary and Daniel L McCartney *et al*.

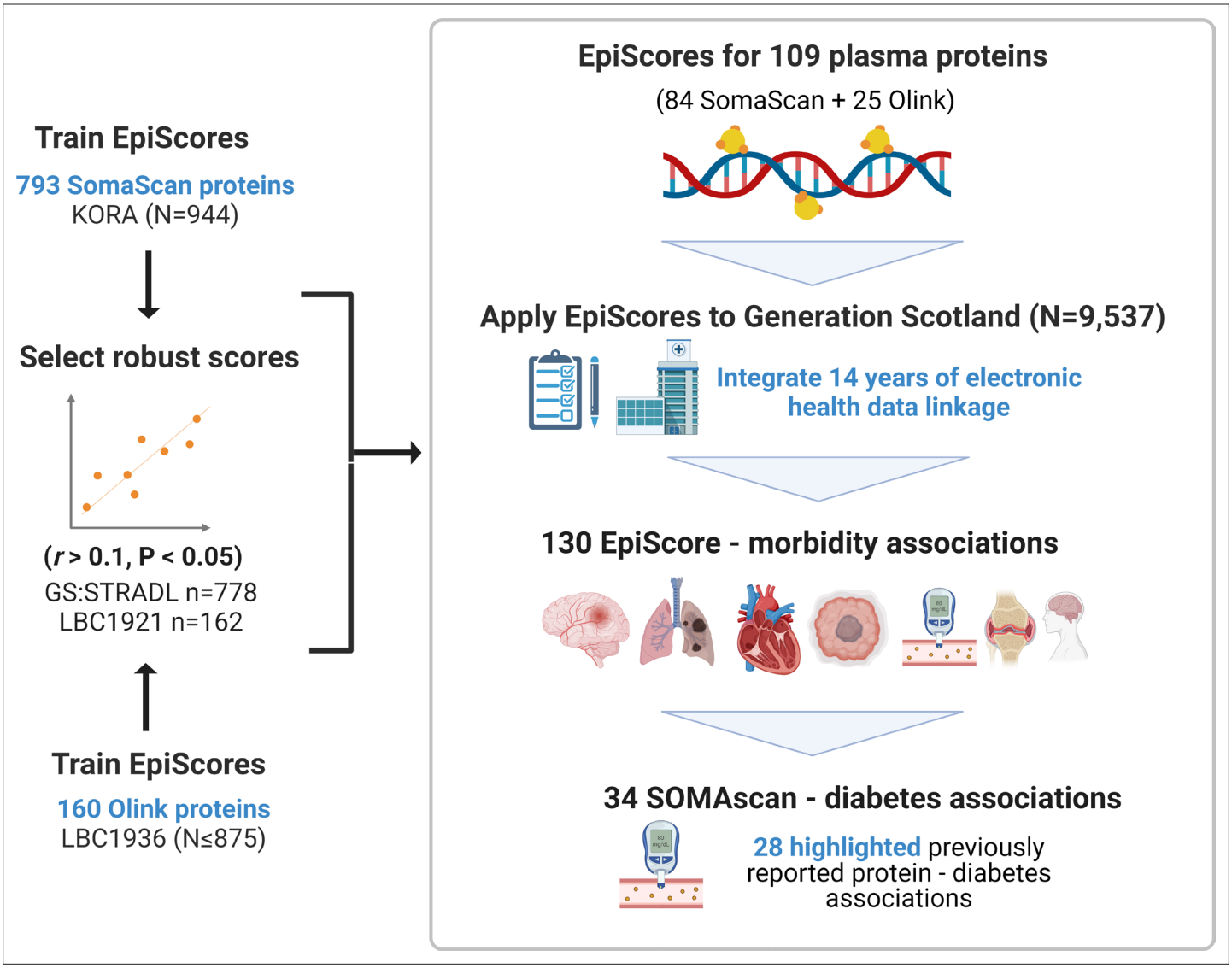

Figure 1. EpiScores for plasma proteins as tools for disease prediction study design. DNA methylation scores were trained on 953 circulating plasma protein levels in the KORA and LBC1936 cohorts. There were 109 EpiScores selected based on performance ($r > 0.1$, $p < 0.05$) in independent test sets. The selected EpiScores were projected into Generation Scotland, a cohort that has extensive data linkage to GP and hospital records. We tested whether levels of each EpiScore at baseline could predict the onset of 12 leading causes of morbidity, over a follow-up period of up to 14 years; 130 EpiScore-disease associations were identified, for 10 morbidities. We then assessed whether EpiScore associations reflected protein associations for diabetes, which is a trait that has been well characterised using SOMAscan protein measurements. Of the 34 SOMAscan-derived EpiScore-diabetes associations, 28 highlighted previously reported protein-diabetes associations.

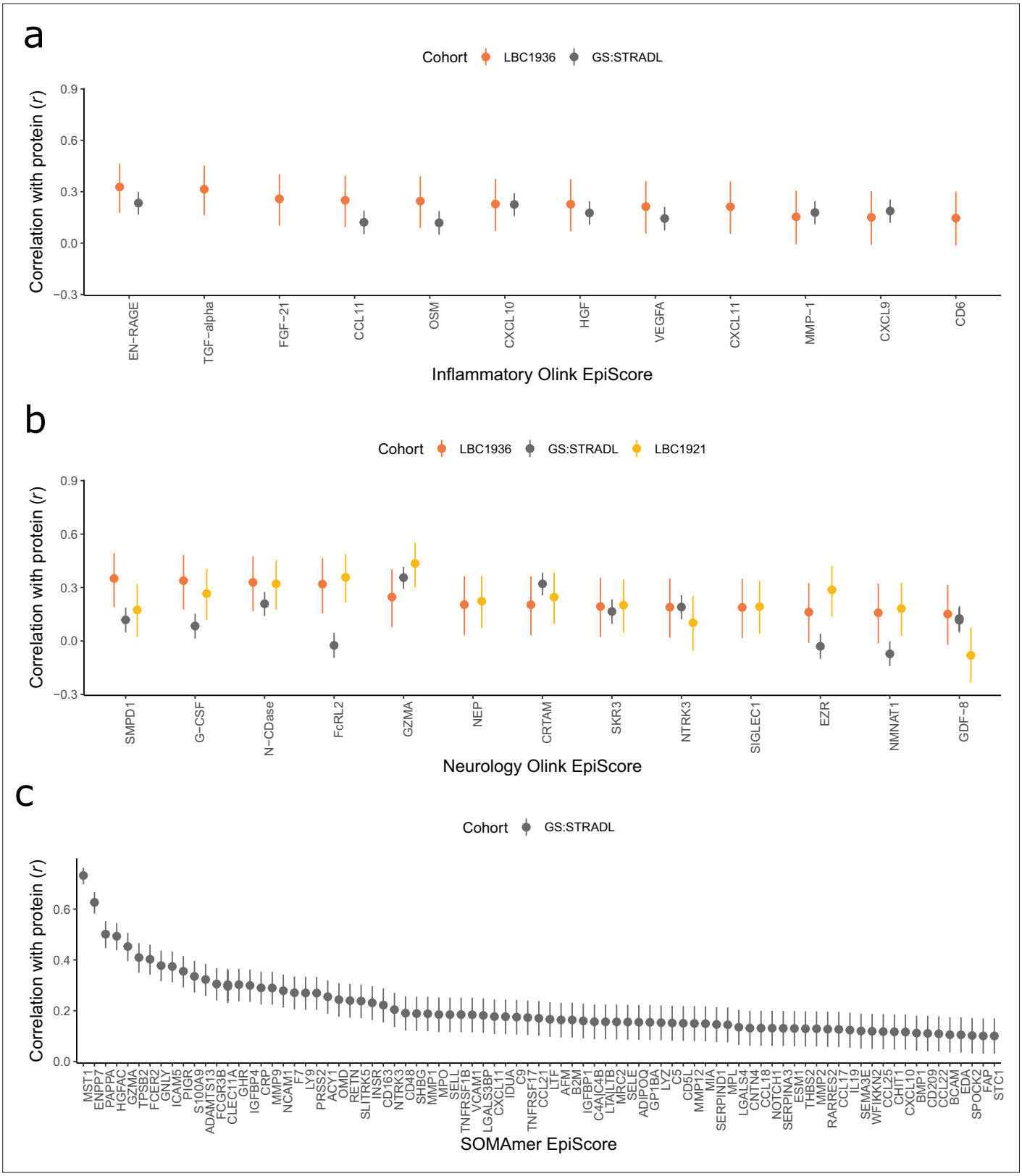

Figure 2. Test performance for the 109 selected protein EpiScores. Test set correlation coefficients for associations between protein EpiScores for (**a**) inflammatory Olink, (**b**) neurology Olink, and (**c**) SOMAmer protein panel EpiScores and measured protein levels are plotted. 95% confidence intervals are shown for each correlation. The 109 protein EpiScores shown had $r > 0.1$ and $p < 0.05$ in either one or both of the GS:STRADL (n = 778) and LBC1921 (n = 162) test sets, wherever protein data was available for comparison. Data shown corresponds to the results included in ***Supplementary file***

Figure 2 continued on next page

Figure 2 continued

1B-C. Correlation heatmaps between the 109 EpiScore measures (***Figure 2—figure supplement 1***) are provided, along with a summary of the most enriched functional pathways for the genes of the 109 proteins used to train EpiScores (***Figure 2—figure supplement 2***).

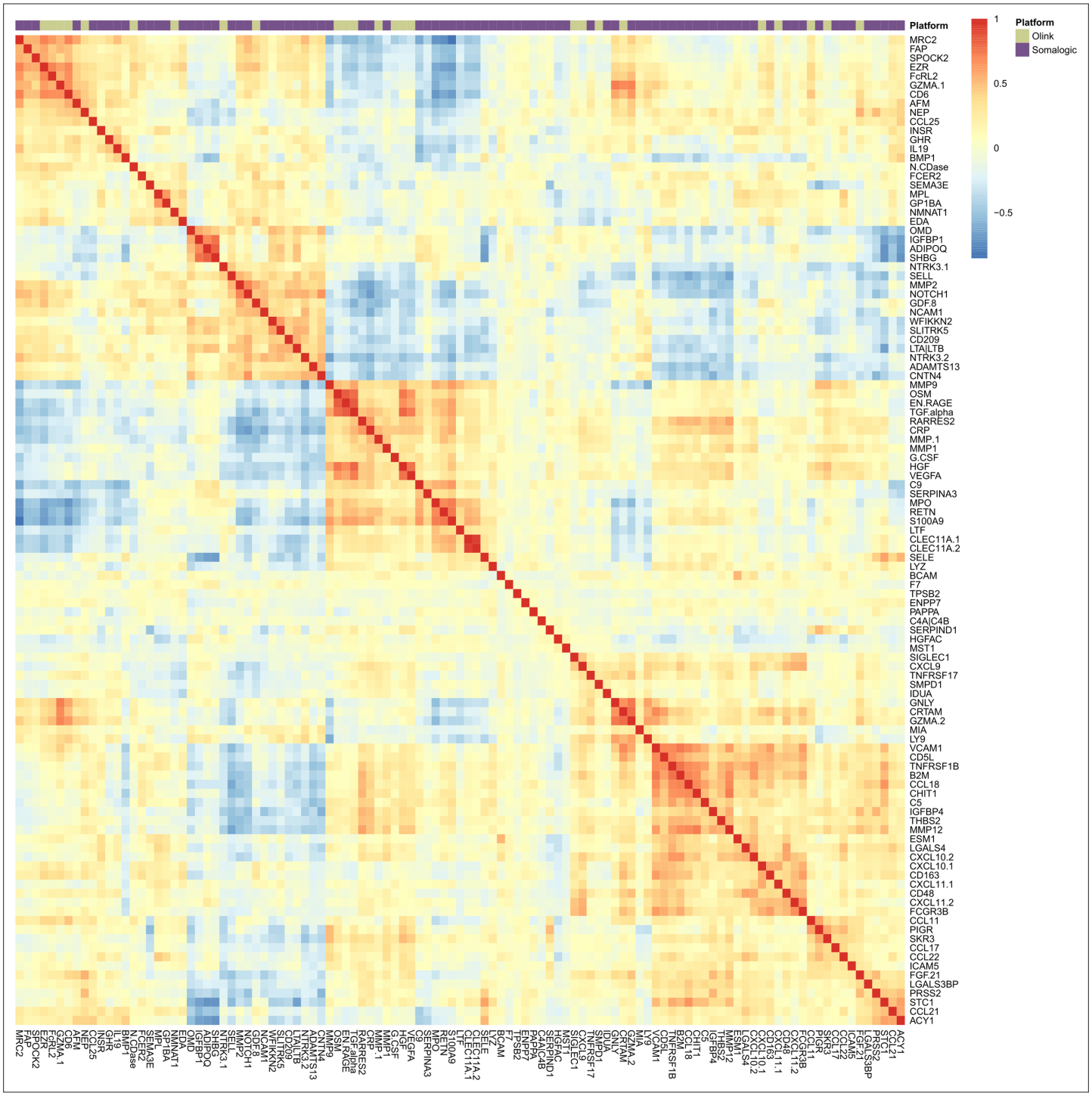

Figure 2—figure supplement 1. Correlation heatmap for protein EpiScore measures in Generation Scotland. Correlation heatmap for EpiScore measures projected into Generation Scotland (N = 9537) for the 109 protein EpiScores selected in the test sample ($r > 0.1$, $p < 0.05$). At the top of the heatmap, an annotation bar is displayed. Olink proteins are shown in pale green and Somalogic proteins are shown in purple.

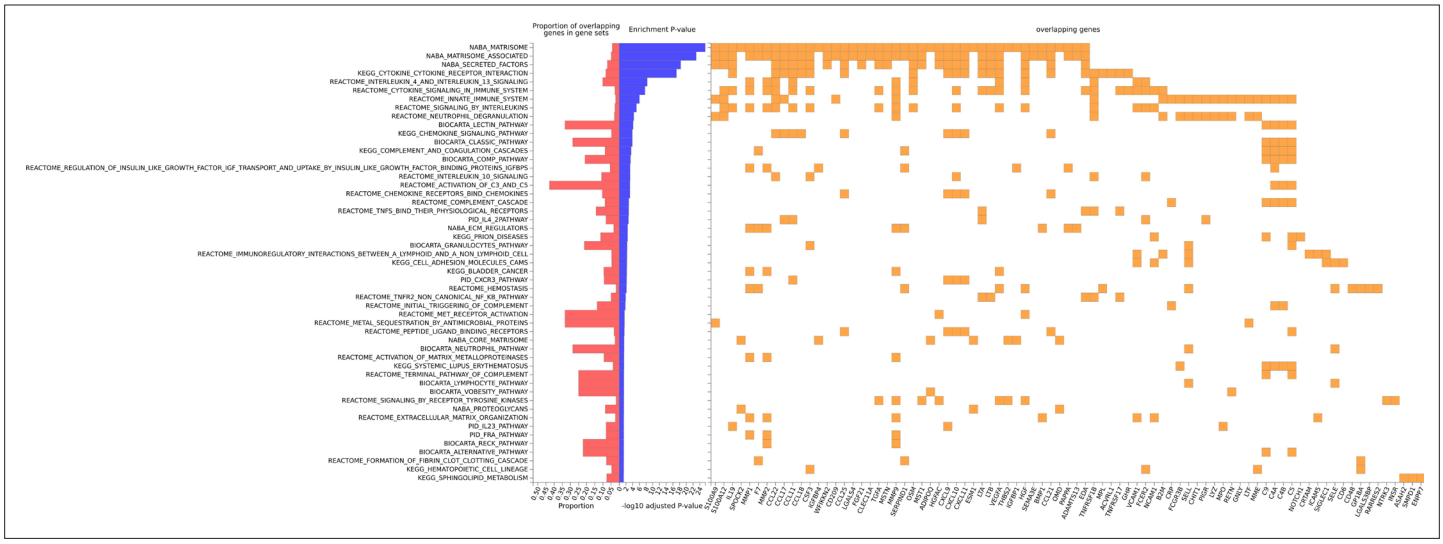

Figure 2—figure supplement 2. GeneSet enrichment of canonical pathways common to the genes encoding proteins that were used to train the 109 selected EpiScores. Genes selected for pathway enrichment (false discovery rate [FDR]-adjusted p < 0.05) are summarised, with the proportion of overlapping genes enriched in the gene-set also shown. The corresponding data for this figure can be accessed in full in ***Supplementary file 1H***.

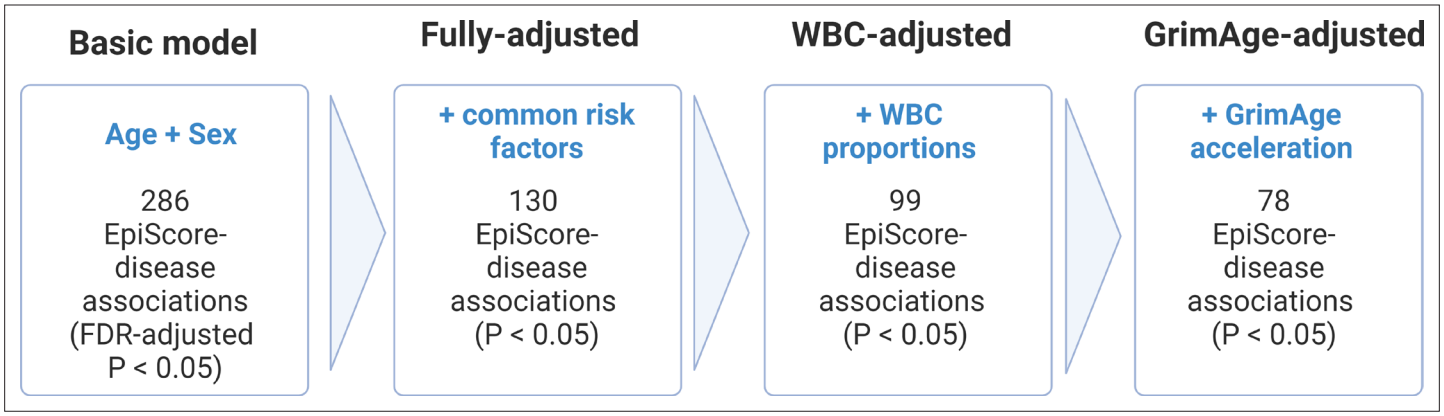

Figure 3. Nested Cox proportional hazards assessment of protein EpiScore-disease prediction. Mixed effects Cox proportional hazards analyses in Generation Scotland (n = 9537) tested the relationships between each of the 109 selected EpiScores and the incidence of 12 leading causes of morbidity (***Supplementary file 1I-J***). The basic model was adjusted for age and sex and yielded 286 associations between EpiScores and disease diagnoses, with false discovery rate (FDR)-adjusted p < 0.05. In the fully adjusted model, which included common risk factors as additional covariates (smoking, deprivation, educational attainment, body mass index (BMI), and alcohol consumption), 130 of the basic model associations remained significant with p < 0.05. In a sensitivity analysis, the addition of estimated white blood cells (WBCs) to the fully adjusted models led to the attenuation of 31 of the 130 associations. In a further sensitivity analysis, 78 associations remained after adjustment for both immune cell proportions and GrimAge acceleration.

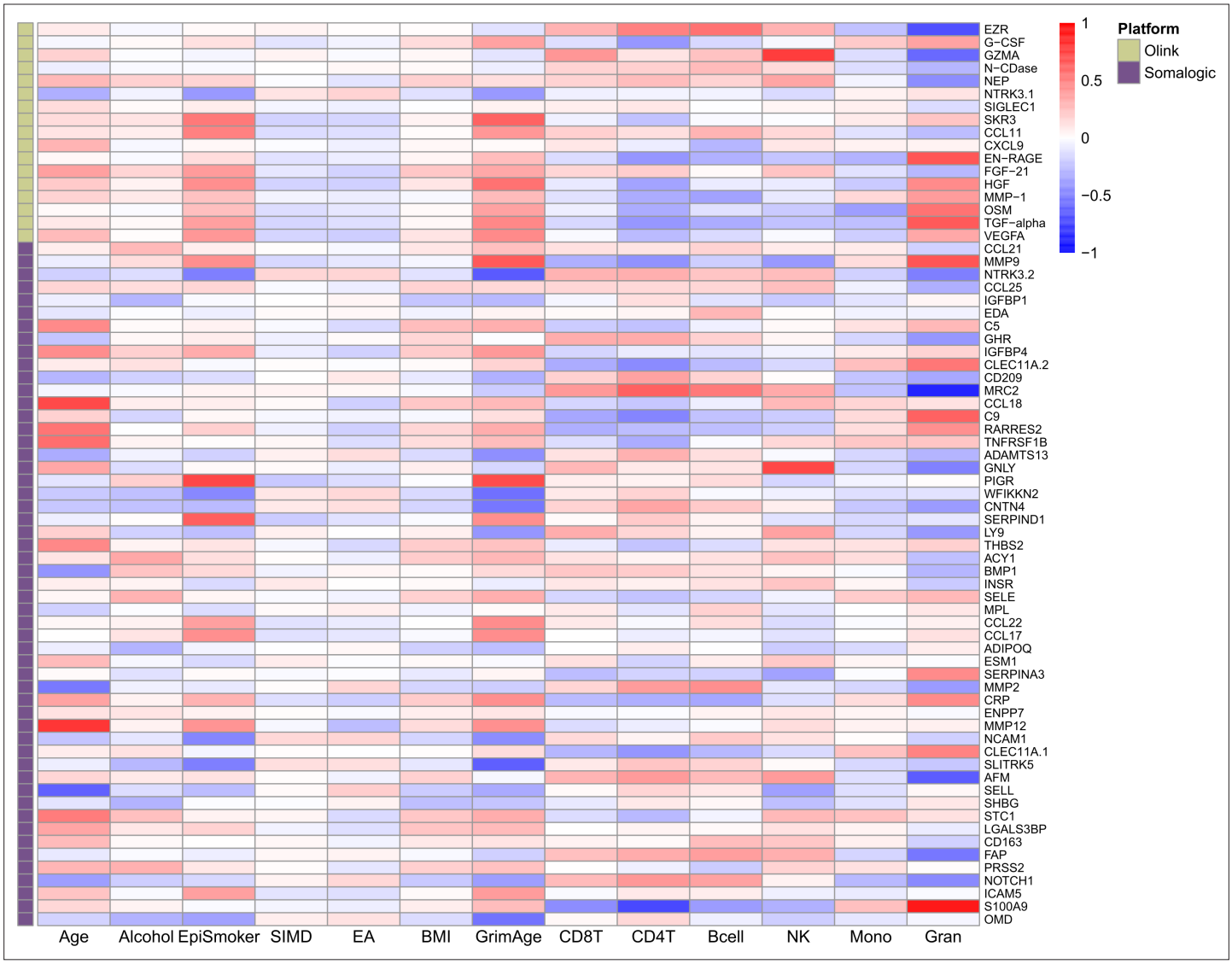

Figure 3—figure supplement 1. Phenotypic trait and estimated white blood cell proportion correlations with EpiScores. Heatmap of Pearson's correlations (*r*) between the 70 protein EpiScore measures that were associated with incident disease (with $p < 0.05$ in the fully adjusted Cox mixed effects proportional hazards models) and continuous phenotypic/lifestyle trait variables and Houseman-estimated white blood cell proportions in Generation Scotland (total N = 9537). Protein measurements used to train the predictors were adjusted for age and sex. The maximum sample size available was used for each correlation. GrimAge: GrimAge acceleration. Units: weekly units of alcohol. EpiSmoker: DNAm-derived score for smoking. SIMD: Scottish Index of Multiple Deprivation. EA: educational attainment. Mono: monocytes. Gran: granulocytes. NK: natural killer cells.

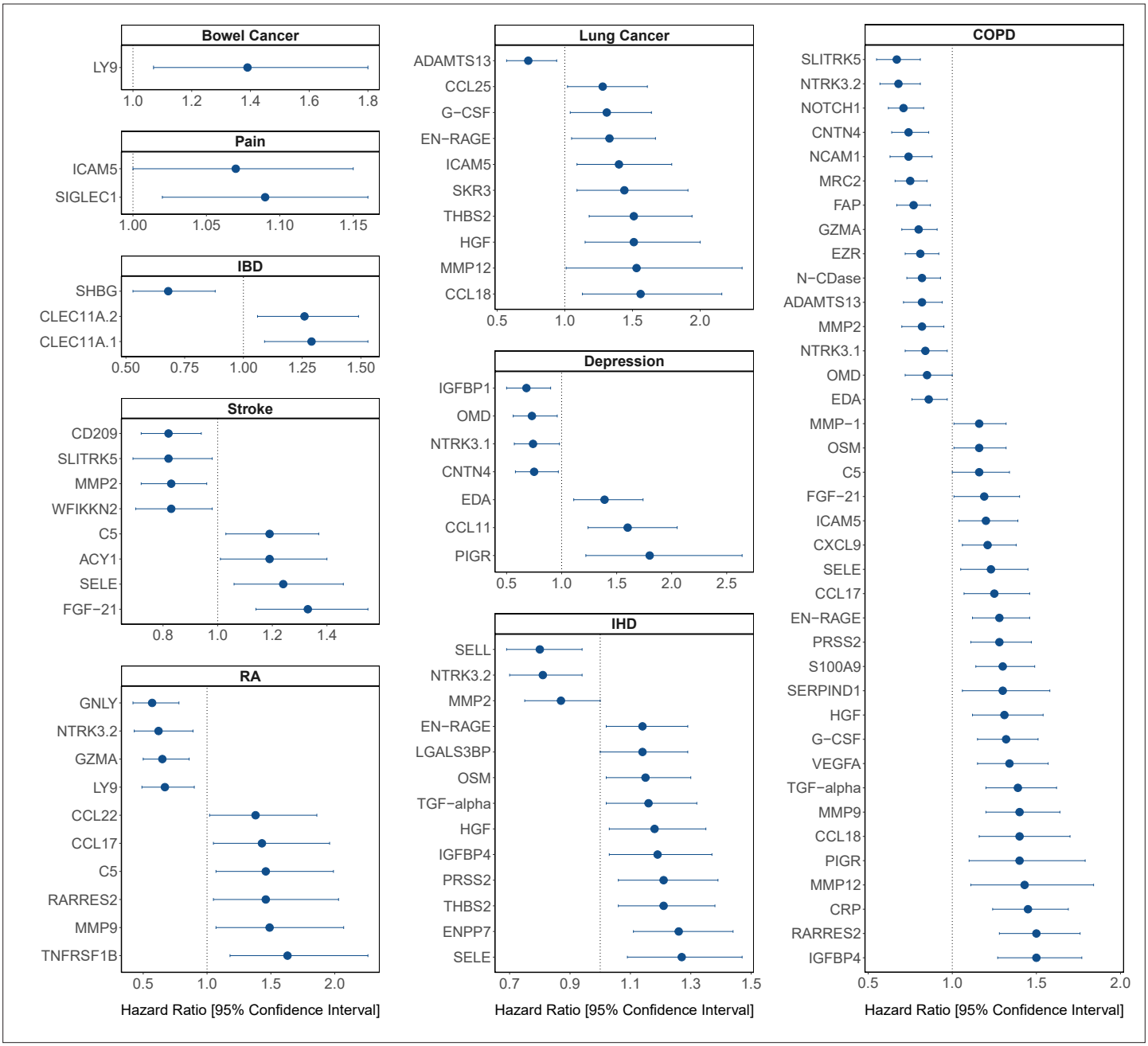

Figure 4. Protein EpiScore associations with incident disease. EpiScore-disease associations for 9 of the 11 morbidities with associations where $p < 0.05$ in the fully adjusted mixed effects Cox proportional hazards models in Generation Scotland (n = 9537). Hazard ratios are presented with confidence intervals for 92 of the 130 EpiScore-incident disease associations reported. Models were adjusted for age, sex, and common risk factors (smoking, body mass index (BMI), alcohol consumption, deprivation, and educational attainment). IBD: inflammatory bowel disease. IHD: ischaemic heart disease. COPD: chronic obstructive pulmonary disease. For EpiScore-diabetes associations, see ***Figure 6***. Data shown corresponds to the results included in ***Supplementary file 1J***.

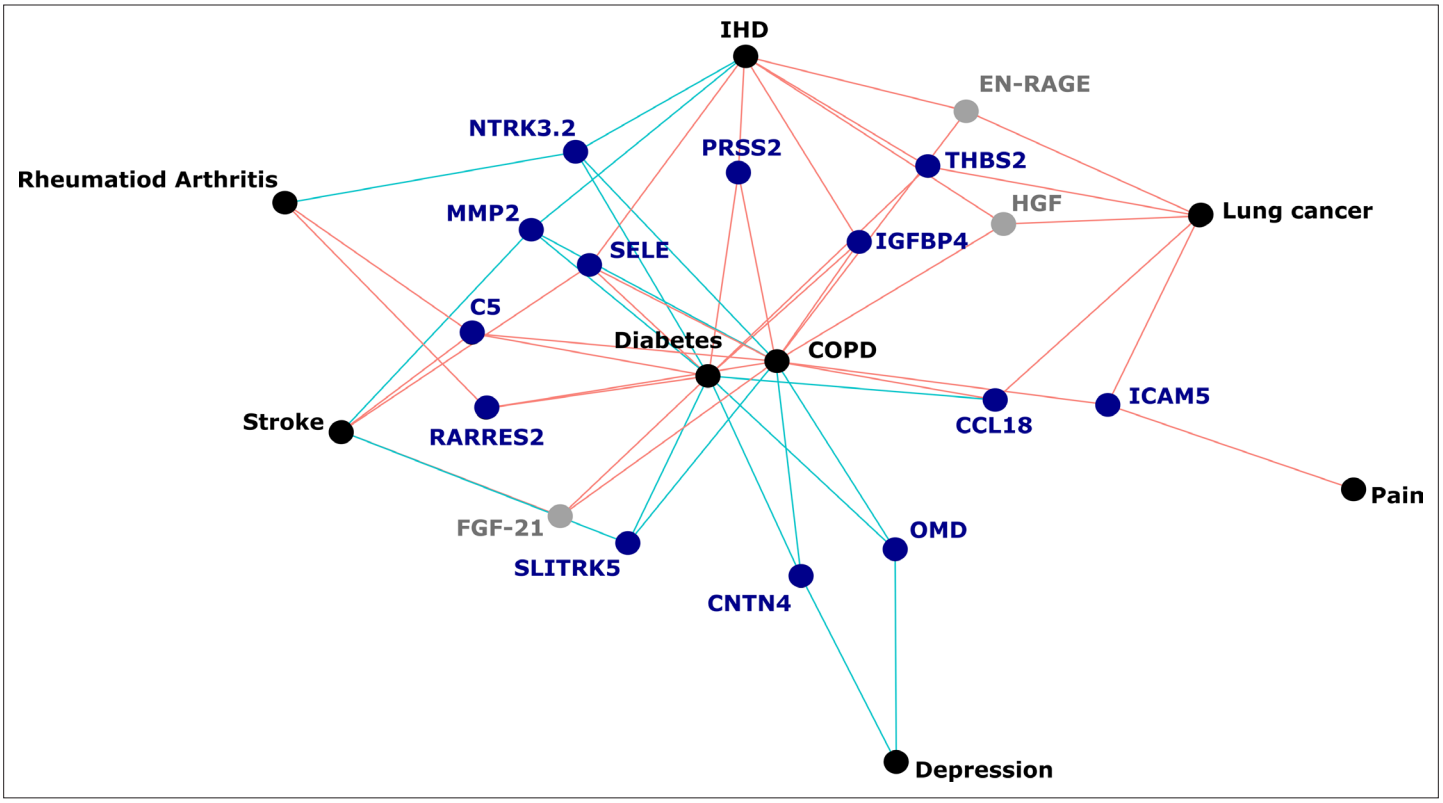

Figure 5. Protein EpiScores that associated with the greatest number of morbidities. EpiScores with a minimum of three relationships with incident morbidities in the fully adjusted Cox models. The network includes 16 EpiScores as dark blue (SOMAscan) and grey (Olink) nodes, with disease outcomes in black. EpiScore-disease associations with hazard ratios < 1 are shown as blue connections, whereas hazard ratios > 1 are shown in red. COPD: chronic obstructive pulmonary disease. IHD: ischaemic heart disease. Data shown corresponds to the results included in ***Supplementary file 1J***.

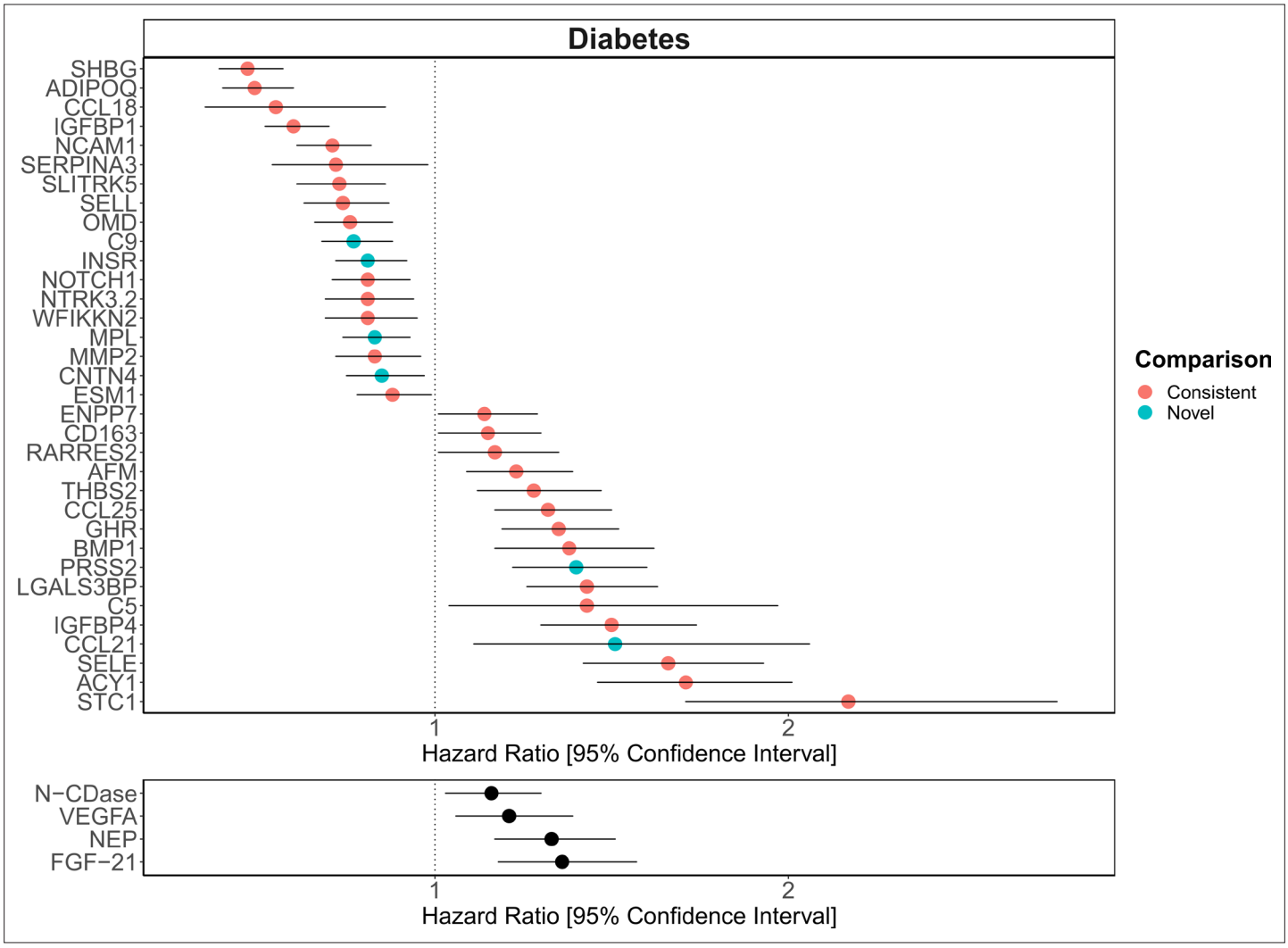

Figure 6. Replication of known protein-diabetes associations with protein EpiScores. EpiScore-incident diabetes associations in Generation Scotland (n = 9537). The 34 SOMAscan (top panel) and four Olink (bottom panel) associations shown with $p < 0.05$ in fully adjusted mixed effects Cox proportional hazards models. Of the 34 SOMAscan-derived EpiScores, 28 associations were consistent with protein-diabetes associations (pink) in one or more of the comparison studies that used SOMAscan protein levels. Six associations were novel (blue). Data shown corresponds to the results included in ***Supplementary files 1J and M***.