

Figure S8

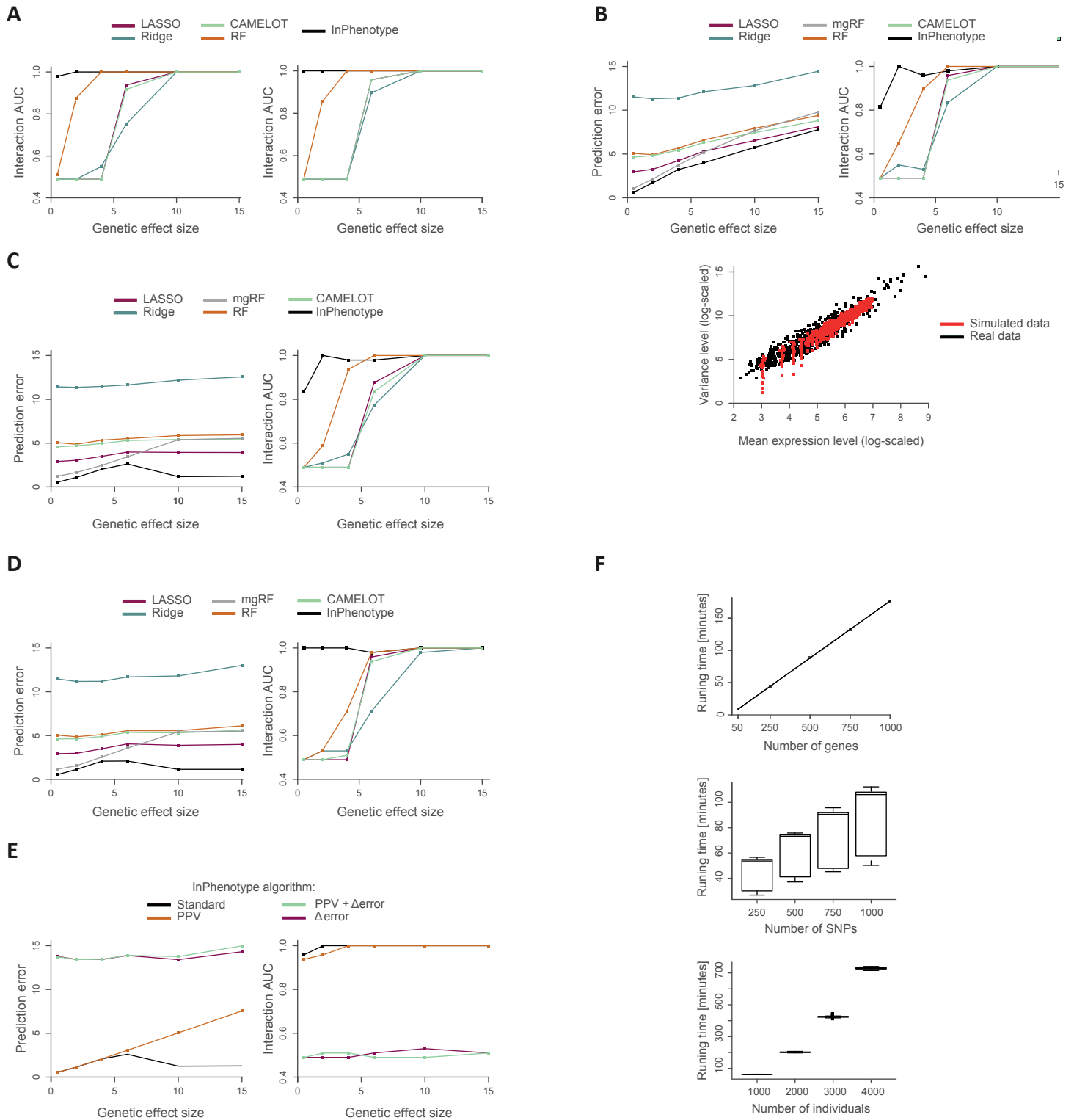


Figure S8. Additional analyses of synthetic datasets. (A) Synthetic datasets of five (left) or ten (right) genes (rather than a single gene) that have an effect on the phenotype. Plots are shown as in **Figure 2**. (B) Synthetic datasets of varying noise levels. **Top:** AUC and prediction error, shown as in **Figure 2**. **Bottom:** the relations between mean expression level and variance in the real influenza dataset (black; a dot for each gene, calculated across all individuals) and in the simulation of varying noise levels (red; a dot for each terminal node, calculated across all individuals reaching that node). (C) Synthetic datasets with biologically relevant allele frequencies, based on the murine genotyping across the collaborative cross mouse lines. Plots are shown as in **Figure 2**. (D) Synthetic datasets with biologically relevant association scores in control genomic loci, based on the murine associations of the influenza dataset. Plots are shown as in **Figure 2**. (E) Synthetic data analysis, where InPhenotype is applied with four different modes: a standard InPhenotype (black), with predictions using the PPV function (orange), with split decision using the Δ error function (maroon), and both (light green). (F) InPhenotype's running time (minutes) for synthetic data collections (25 datasets in each collection), each of which consists of different numbers of genes (**top**), SNPs (**middle**) and individuals (**bottom**). Unless stated otherwise, in all analyses we used the default synthetic data parameters, as specified in **Methods**.