Figure S2. Distribution of T2D gene p-values for small, large and all genes before and after correction for confounders. (A) The distribution of the mean $p^{BestSNP}$ (best SNP association p-value per gene $g$) calculated across 1,000 phenotype permutations of the Diabetes Genetics Initiative (DGI) GWA study is shown for all genes in genome (blue line), only large genes ($\geq$100 kilobase (kb); red line), and only small genes ($\leq$10 kb; green line). Large genes tended to receive on average a more significant gene score (lower $p$-values) than all genes in the permuted datasets, and small genes tended to receive on average a less significant gene score (higher $p$-values) than all genes. (B-D)
The distribution of gene association $p$-values is shown for the actual DGI study for all gene sizes (blue line), large genes (red line) and small genes (green line) (B) before correcting for confounders ($P_{g}^{\text{BestSNP}}$), and after correcting for confounders on $P_{g}^{\text{BestSNP}}$, such as gene size, using either (C) phenotype permutation analysis ($P_{g}^{\text{Gene}}$) or (D) step-wise multivariate linear regression analysis ($P_{g}^{\text{Gene'}}$). The regression-based correction transforms the gene $p$-values to a distribution that is close to uniform and removes the confounding effect of gene size, similar to the permutation-based correction, which corrects for all confounding effects without $a$ priori knowledge of them. The regression correction seems to slightly over-correct the gene $p$-values of large genes (red line in D) in the high $p$-value end of the distribution ($p>0.8$). A bin of 0.01 was used for all four plots.