Background

Diabetes affects ~12–14% of adult Americans and imposes >$400 billion per year in U.S costs.[2] Rare non-coding variants exert large, cell-type-specific regulatory effects, but molecular QTL studies remain underpowered across pancreas cell types. Deep sequence-to-function (S2F) models infer regulatory effects directly from DNA sequence, but most are trained on bulk tissues and miss single-cell specificity. We will train a pancreas foundation model, scooby-PanK, to predict gene expression and chromatin accessibility at single-cell resolution to quantify variant effects (ΔRNA, ΔATAC) across pancreatic and immune cell types relevant to T1D(Fig. 1).[3]

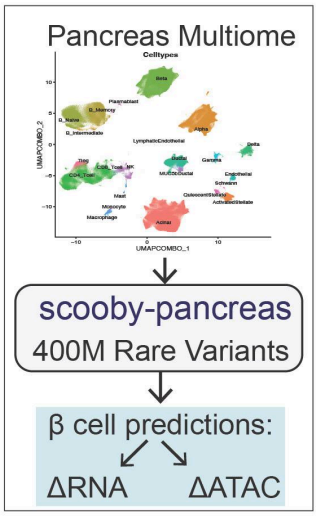

Fig 1. Schematic of Approach

Methods

We integrate single-cell RNA+ATAC data from PanKbase and other pancreas/islet cohorts, into a harmonized pancreas atlas (~2.2M cells from 327 donors across endocrine, exocrine, stromal and immune populations)[4-9,14].[4] Cell embeddings are computed with scVI/MultiVI. Following the scooby framework,[3] we adapt a bulk S2F foundation model to single cells by conditioning on these embeddings and fine-tuning LoRA adapters and a cell-conditioned decoder. Variant effects are scored by predicting reference and alternate 524-kb windows and computing ΔRNA and ΔATAC per gene/peak and cell type. We will predict and record variant effects for all ~400M variants in TOPMed across modeled cell types.[18] We evaluate ΔRNA with bulk human islet eQTL effect sizes (n=420 donors)[10] and correlate ΔATAC with caQTLs and regulatory programs from islet multiome and chromatin studies.[13],[14] To assess rare variants (MAF <0.1%), we will test whether donors carrying alleles with large predicted ΔRNA/ΔATAC exhibit allele-specific expression/accessibility and outlier gene expression, using established methods for rare-variant effects, ASE, and QTL fine-mapping[15=1].

Results (scooby-PBMC preliminary data)

Using a scooby-PBMC model trained on multiome PBMC data from 982 donors[3,11] , we generated ΔRNA predictions and evaluated them against eQTLs in an independent cohort of 1,925 individuals (5.4M PBMCs, 28 cell types).[12] After filtering ($p<1\times10^{-50}$, $|\beta|>0.5$) to 3,377 likely causal variants, ΔRNA predictions correlated with eQTL effect sizes (Spearman $\rho=0.32$), reaching Pearson $R=0.65$ for 860 high-confidence predictions, demonstrating accurate prediction of regulatory effects from sequence. We further used scooby-PBMC to predict effects for 391,605 rare variants across 9 immune genes and observed a spectrum of cell-type-specific effects, including variants with opposite effects in different cell types (e.g., decreased CD79B expression in B cells but increased expression in monocytes), highlighting the model's ability to resolve context-specific regulation.

Conclusions

Scooby-PanK will deliver cell-type-resolved ΔRNA/ΔATAC predictions for rare and common variants in pancreas and immune cell types. By combining a multi-donor pancreas atlas, scooby-style S2F modeling, and genome-wide predictions for ~400M TOPMed variants,[18] we will create a resource to prioritize putative causal variants and enable enrichment and carrier-level analyses in large biobanks.