Functionalizing methylation-driven gene panels in the UK Biobank, TCGA, and HCA datasets

Urvy Mudgal[1] , Michael Pietras[1], Zachary Pitluk[1], Stephen Moore[1], Srikant Sarangi[1]
(1) Paradigm4, Waltham, MA

Background

Conducting population-scale hypotheses-driven validation studies pose challenges such as efficient storage of multi-omics data (e.g., TCGA, UK Biobank, Human Cell Atlas), running complex queries and computations at scale and across datasets, and creation of cohorts based on different metadata fields. The REVEAL FLASH software stack, comprising of a suite of data-specific apps, elastic scaling infrastructure, and a POSIX compliant networked file system, is ideal for performing these kinds of studies.

In this poster, we present a use-case with a methylation-driven eight gene panel (*TCTEX1D4, MAEL, LIME1, KLHL38, HPDL, ESR1, UCP2,* and *COMMD7*) from a published study on breast cancer (*Frontiers in Genetics, 2020*). First, we ran GWAS analysis using the UK Biobank (Application ID: 51518) between variants in the panel from the 200K Whole Exome Sequence dataset and phecodes derived from ICD10 diagnoses of multiple cancer types and other phenotypes of interest. We used PLINK, SAIGE & REGENIE for the GWAS analysis and a custom linkage disequilibrium (LD) plus burden test algorithm for testing pairs of variants.

Next, we did a query across 34 TCGA cancer datasets for single nucleotide variants (SNV) of the gene panel and compared results with the GWAS done on UKBB data. Finally, we looked at the Human Cell Atlas and Tabula Sapiens datasets to assign cell type specificity based on expression levels of the gene panel members.

Methods

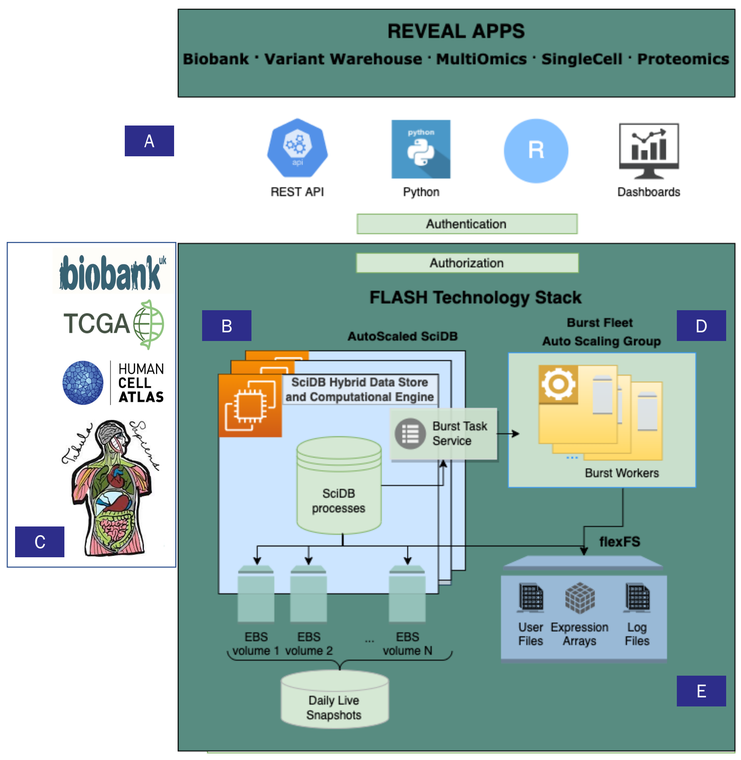

Fig 1 : REVEAL stack architecture – A) REVEAL application layer is a set of tools for user-friendly loading, management, & analysis of large-scale multimodal data. Tools include R, Python, REST API's and extensible dashboards and GUI's; B) SciDB is an array-native database designed to support large linear-algebra operations; C) Population-scale datasets stored in REVEAL can be easily accessed for ad hoc analysis; D) Burst Mode is an elastic task service that automatically distributes large computations across the cloud; E) flexFS is a cloud-native POSIX compliant file system designed to support cost-effective & high I/O throughput across hundreds of machines

- Human Interactome Atlas (HuRI) used to find genes with known PPI's with members of the methylation-driven gene panel
- Custom phecode function used to create phenotype sets based on a high-throughput multimodal automated phenotyping algorithm
- 1662 variants from the 8 genes (maf>1e-05); 42 phecodes generated from 340 ICD10 codes

Results

Phenotype	chr:pos:ref:alt	Gene1_Gene2	p-value	r^2	Cases	Ctrls
Cancer of Tongue	6:152098883:G:A__12:110722709:A:G	ESR1_PPP1CC	1.1e-06	3.3e-07	128	137070
Cervical Cancer	6:152122621:C:T__12:110730429:C:T	ESR1_PPP1CC	7.8e-06	4.4e-07	164	126703
Cancer of Mouth	8:123645884:T:C__6:41649742:G:A	KLHL38_MDFI	2.4e-08	2.8e-08	291	137070
Cancer of Mouth	8:123652558:G:T__6:41649742:G:A	KLHL38_MDFI	4.1e-07	4.3e-08	291	137070
Cancer of Brain & Nervous System	20:63735726:CTG:C__7:30923432:C:T	LIME1_AQP1	3.9e-07	5.6e-08	306	137119
Cancer of Brain & Nervous System	20:63735726:CTG:C__7:30923982:T:C	LIME1_AQP1	3.5e-06	3.6e-07	306	137119
Cancer of other Female Genital organs	1:44806775:C:A__6:41646256:C:T	TCTEX1D4_MDFI	1.5e-06	2.9e-05	204	129358
Cancer of Stomach	1:44808883:C:A__6:41639553:T:G	TCTEX1D4_MDFI	8.7e-06	1.1e-07	292	131041
Cancer of Tongue	11:73975391:C:CTG__21:44600709:G:A	UCP2_KRTAP10	1.6e-09	1.1e-07	128	137070
Cancer of Tongue	11:73977099:C:T__21:44600709:G:A	UCP2_KRTAP10	1.9e-09	1.2e-07	128	137070

Table 1: Top significant associations between pairs of variants and phenotypes using a custom LD + Burden test

Phenotype	Gene	Chr	Pos	Ref	Alt	p-value	Consequence	Cases	Ctrls
Cancer of other female genital organs excluding Uterus and Ovary	COMMD7	20	32704469	A	G	6.9e-06	synonymous var	137	129358
Cancer of Prostate	KLHL38	8	123651965	C	T	5.2e-06	missense var	3272	129866
Cancer of Prostate	UCP2	11	73978011	C	T	8.3e-06	missense var	3272	129866
Cancer of Eye	MAEL	1	166989443	C	T	2.9e-07	missense var	74	137119
Cancer of the Gums	KLHL38	8	123651810	G	C	7.9e-06	missense var	19	137070
Cancer of the Gums	KLHL38	8	123652375	G	A	4.1e-06	synonymous var	19	137070
Cancer of other Endocrine glands	KLHL38	8	123651970	C	T	6.8e-06	synonymous var	49	137264
Cancer suspected or other	KLHL38	8	123645887	G	A	1.9e-06	missense var	381	123266
Cancer of the Mouth floor	MAEL	1	166989782	G	A	2.6e-06	missense var	22	137070
Cancer of Hypopharynx	MAEL	1	167021733	C	T	3.6e-06	missense var	16	137070
Cancer of Hypopharynx	ESR1	6	152098959	C	T	2.7e-06	missense var	16	137070
Cancer of Nasal cavities	KLHL38	8	123645899	C	T	1.1e-08	missense var	22	137070
Cancer of Connective tissue	ESR1	6	152098959	C	T	2.4e-06	missense var	166	137691
Breast Cancer female	KLHL38	8	123651810	G	C	7.8e-06	missense var	1003	129103
Breast Cancer female	COMMD7	20	32704469	A	G	8.6e-06	synonymous var	1003	129103

Table 2: Top PLINK GWAS associations for variants also present across TCGA datasets

Phenotype	Gene	Chr	Pos	Ref	Alt	p-value	beta	Consequence	Cases	Ctrls
Breast Cancer Female	LIME1	20	63737733	G	A	4.5e-06	44.9	Intron variant	1003	129103

Table 3: Only association with pval<1e-05 found between variants and phenotypes tested when using SAIGE & REGENIE

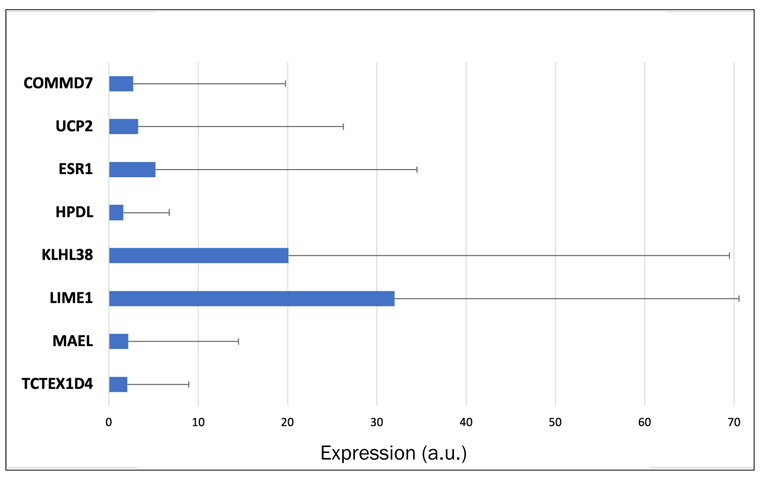

Fig 2: Mean (+/- SD) expression values for the 8 member gene panel across 13 HCA datasets (~4.5 million cells)

- HCA datasets queried include: Human Tissue TCell Activation, Adult Retina, HPSI Human Cerebral Organoids, 1M Immune Cells, Single Cell Liver Landscape, Tissue Stability, Human Hematopoietic Profiling, Fetal Maternal Interface, Reprogrammed Dendritic Cells, Single Cell Transcriptome Analysis of Human Pancreas, Human Inhibitory Interneuron Diff, Kidney Single Cell Atlas, Human Colonic Mesenchyme IBD

Results

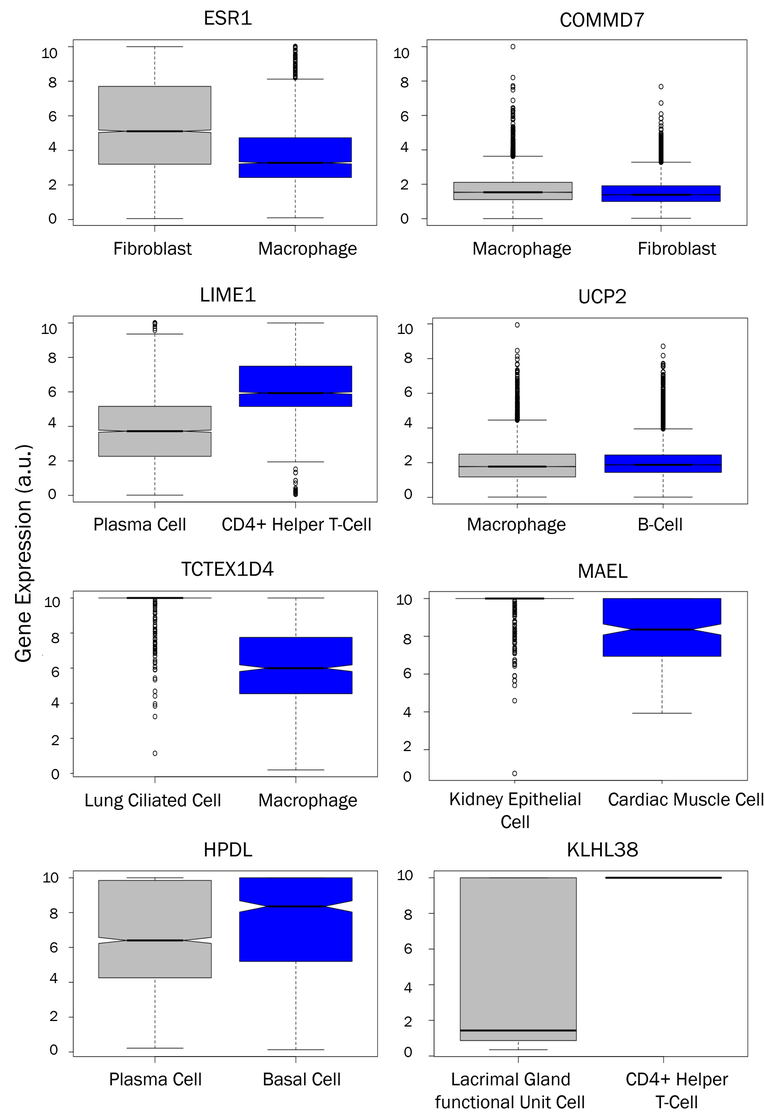

Fig 3 : Gene expression values in top 2 cell types, identified by the number of cells expressing the corresponding gene, sampled across 29 Tabula Sapiens datasets

Gene	Cell Type	Num Cells	Mean Exp.	Std Dev.
TCTEX1D4	luminal epithelial cell of mammary gland	48	6.7	2.5
MAEL	fibroblast of breast	10	7.1	4.7
LIME1	t cell	10	4.5	3.9
KLHL38	luminal epithelial cell of mammary gland	7	9.4	1.7
HPDL	fibroblast of breast	42	9.0	1.4
ESR1	luminal epithelial cell of mammary gland	549	4.0	1.8
UCP2	luminal epithelial cell of mammary gland	1042	1.4	0.7
COMMD7	luminal epithelial cell of mammary gland	2149	1.6	0.7

Table 4: Gene expression in the top cell type, identified by the number of cells expressing the gene in the Tabula Sapiens Mammary Gland dataset

Conclusions

- A frequently encountered situation in biological research is finding a set of "biomarkers" that are based on a specific dataset
- In this poster, we examined whether evidence for a TCGA derived set of biomarkers could be validated in orthogonal datasets
- For expression data, we used single cell data from the publicly available Human Cell Atlas and Tabula Sapiens
- For genetic evidence we conducted GWAS analyses to determine if there are any associations between the gene panel variants in the 200K WES data from the UK Biobank (project 51518) and phecodes derived from individual cancer ICD10 codes
- We used PLINK, SAIGE, REGENIE, and a custom LD + Burden test function for the analyses at the cost of $30
- Performing validation experiments across population-scale orthogonal datasets cost-effectively and efficiently is made possible using REVEAL
- Thank you to the UK Biobank, TCGA, HCA, Tabula Sapiens, and their participants for making such research projects possible!