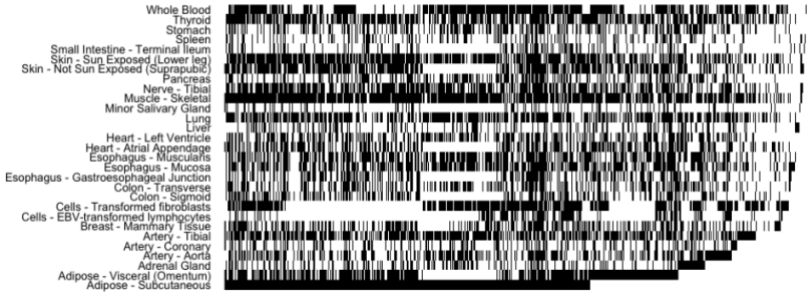
# Multi-Tissue Transcriptomics Learning for Chronic Diseases

**Project Description.** As part of multi-omics data analytics for several diseases, transcriptome-wide association studies based on genetically predicted gene expression have the potential to identify novel regions associated with various complex traits. It has been shown that incorporating expression quantitative information corresponding to multiple tissue types can improve power for association studies involving complex etiology.




A heatmap of missingness in GTEx gene expression data for the 29 tissues. Rows correspond to the 29 tissues and columns correspond to the 613 subjects with expression measured in at least one of the corresponding tissues. White spaces denote missing measured expression, whereas blacks denote observed measured expression.

The GTEx Consortium reported that hierarchical clustering of RNA profiles from 25 unique tissue types among 1,641 individuals accurately distinguished the tissue types, but a multidimensional scaling or principal component analysis and others failed to generate a 2D projection of the data that separates tissue subtypes.

We propose an artificial intelligence framework for learning a multivariate response model -jointly- with the error precision matrix, i.e., the tissue-tissue expression correlation, for predicting gene expression in multiple tissues simultaneously. Unlike existing methods for multi-tissue, our approach incorporates tissue-tissue expression correlation, which allows us to handle missing expression measurements more efficiently and more accurately predict gene expression using a weighted summation of genotypes. We use a Bayesian approach to estimate the missing information related to several tissues, and this allows us to handle missing measurements more efficiently and more accurately. We will be using 29 tissues collected by Genotype-Tissue Expression (GTEx) database and compare our method with the existing algorithms.

**Project Type.** Research with a focus on producing a serious publication.

**Internship Batch**: Batch 1 from May 7 to June 29

**Duties/Activities.** Some code exists, but there is still some work to be done to make it usable, and to produce results. Participate in the understanding and analysis of the GTEx data, and extensive experimentation.

**Required Skills:** Python, R

**Preferred Intern Academic Level:** B.Sc. (3$^{rd}$ year or 4$^{th}$ year) or MS student + serious commitment to the project.

**Learning Opportunities.** Students will be exposed to the exciting research in "multiomics analytics for chronic diseases" while they need not have prior knowledge about the concepts mentioned above. They will acquire new knowledge in machine learning and enhance their programming skills in Python and R. They will experiment and learn dimension reduction techniques such as PCA, tSNE and UMAP; multivariate machine learning models; ML techniques for imputing missing measurements; EM algorithm; Bayesian methodology.

**Expected Team Size:** 2 students.

**Mentors:** Dr. Abdelkader Baggag, Dr. Aisha Al-Qahtani, Dr. Halima Bensmail
**Emails:** abaggag@hbku.edu.qa, aialqahtani@hbku.edu.qa, hbensmail@hbku.edu.qa