Title:
Utilizing Multi-Omic Markers to Predict Age Dysregulation and Identify Underlying Illness

Members: Cimone Jackson, Nathan Weber, Sabrina Su

GitHub Documentation: https://github.com/cjax21/GENE6145

Background:

Aging is a gradual, irreversible natural process that all living organisms experience and various age-related health problems often accompany it (Guo et al., 2022). Aging can lead to various diseases and disabilities, including cardiovascular disease, Alzheimer's disease, and cancer, the leading causes of death in the elderly due to a decline in cellular and tissue functions (Guo et al., 2022). Therefore, there is a need for biomarkers that can indicate aging and its related dysregulation, as they could be used as early indicators that something is wrong in the body. Previous studies looked into routine laboratory biomarkers to identify functionality and age-dependent changes using creatinine, cystatin C, urea, and albumin as renal and liver function markers, glycated hemoglobin (Hba1c) and glucose for diabetes, insulin and c-peptide as general aging markers (Hartmann et al., 2021). Recent advances in genomic technology have made it possible to use multi-omic markers, including gene expression and DNA methylation, to accurately predict a person's age (Hartmann et al., 2021). Known biological markers include the telomere length, which shortens throughout a person's lifespan as the cells in the body undergo division or are affected by stress (Hartmann et al., 2021; Vaiserman & Krasnienkov, 2021). Other biomarkers are dictated through cognitive and physical function, such as bone mass known to deteriorate with age, leading to increased fracture susceptibility (Boskey & Coleman, 2010; Keaveny et al., 2010).

The question is whether biomarkers can prevent or slow disease progression. The field of epigenomics is based on methylation and studies how the environment influences and affects our genes (Tello, 2019). The patterns of methylation or demethylation can impact health, and aging, and lead to the diseases mentioned above. RNA-seq gene expression (FPKM) dataset was generated from human skin fibroblast cell lines (Fleischer et al., 2018). Fibroblasts in skin cells obtained through non-invasive skin biopsies have a low proliferation rate and retain epigenetic changes that occur with age from damage, and phenotypic, epigenomic, and transcriptomic changes (Fleischer et al., 2018). Thirteen epigenetic markers obtained from whole blood samples were utilized to showcase the rate of methylation that occurs at individuals at different "speeds" in one's lifetime (Naue et al., 2017). The biomarkers are the following genes: DDO, ELOVL2, F5, GRM2, HOXC4, KLF14, LDB2, MEIS1-AS3, NKIRAS2, RPA2, SAMD10, TRIM59, ZYG11A, and are selected due to the number of CpG sites correlated to age (Naue et al., 2017). The two datasets can then be used to build a model that takes fibroblast age and whole blood markers to predict discrepancies between biological and chronological age that could indicate an underlying illness or disease.

Justifications:

For General Physicians and Healthcare providers

One of the most significant costs in patient care has always been the lack of early diagnosis. Some patients avoid seeking medical care, leading to undiagnosed major diseases and conditions for months to years (Karczewski & Snyder, 2018). This often culminates with the patient passing due to their undiagnosed condition or being diagnosed in the end stage of illness, which involves subjecting the patient to more aggressive and costly treatment. We provide a solution through our age-dysregulation-based model. However, this model cannot diagnose the illness itself. It can indicate poor health and may result in the subsequent early diagnosis of severe disease. Our model uses a series of multi-omic markers (majority cell cycle markers) and predicts the patients' biological age (C. Jimmy Lin et al., 2021; Cope et al., 2022). By comparing the predicted biological age with the patient's actual age, we can calculate the patient's age dysregulation. The larger the dysregulation, the more likely the patient will experience a health problem. We intend to offer this model as part of routine lab testing with a surface-level skin punch.

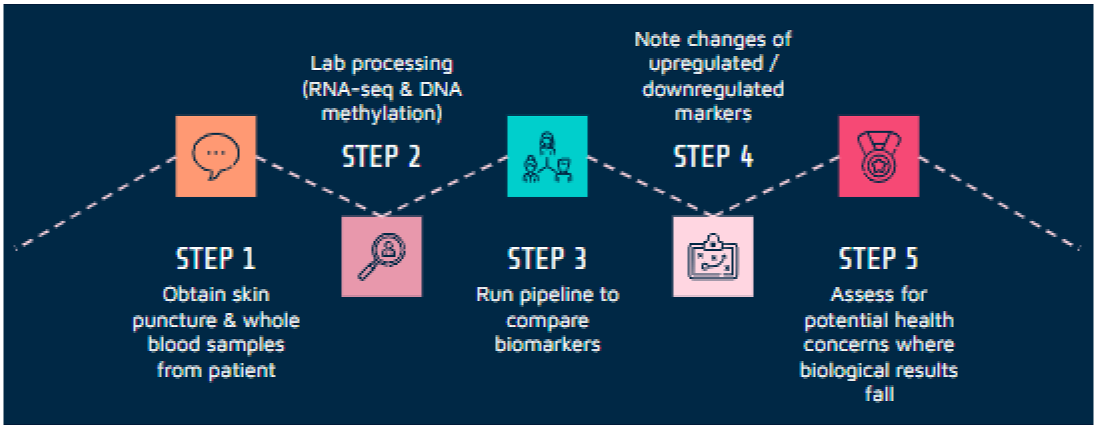
For Patients

Visiting the doctor and staying on top of your health is difficult. We are offering a solution through Age prediction testing. Your doctor can order this test to predict your biological age and compare it to your actual age. In doing so, we calculate something known as age dysregulation. Based on how significant your age dysregulation is, this may indicate a developing illness or poor health, which you can follow up on with your physician to keep you nice and healthy. The testing is quick, cost-effective and minimally invasive, only requiring a blood draw and skin punch which can be collected in the same die. Skin punch only requires local anesthetic and is done in 15 minutes or less.

For Diagnostic testing companies and Clinical Laboratories

We have built a simple-to-integrate multi-omic biological age prediction model which can be used to calculate a patient's age dysregulation through a combination of differential methylation microarray and differential expression microarray of currently 13 markers associated with the cell cycle and aging. We have designed our pipeline in Python to facilitate ease of software integration, and we hope to package our pipeline with common LIMS and bioinformatic package suites. We have identified a potential market opportunity for routine age prediction as age dysregulation is a predictive biomarker for specific illnesses or indicates poor health (Karczewski & Snyder, 2018). We feel that this form of testing is best aimed at adults with a passion for health and health monitoring, and we plan to offer this test initially in general physician clinics as part of an annual checkup, with a plan to expand this test as part of direct-to-consumer blood testing.

Overall Process Flow


STEP 1
Obtain skin puncture & whole blood samples from patient
STEP 2
Lab processing (RNA-seq & DNA methylation)
STEP 3
Run pipeline to compare biomarkers
STEP 4
Note changes of upregulated / downregulated markers
STEP 5
Assess for potential health concerns where biological results fall

The proposed process involves healthcare providers performing a skin puncture to obtain human fibroblast skin cells and a blood draw on the same day. These samples will then be sent to a diagnostic laboratory where RNA-seq analysis will be conducted on the fibroblast cells, and DNA methylation studies will be performed on the whole blood sample. A subset of the biomarkers are associated with increased methylation at a certain age, while others are associated with decreased methylation (refer to the Data Visualization section). Based on these markers, the summarized data will be used to determine the patient's biological age, and healthcare providers will discuss the results with patients to understand their implications in relation to their actual age and potential health conditions. Again the goal is to provide early intervention for those at risk and improve health outcomes.

Data Management

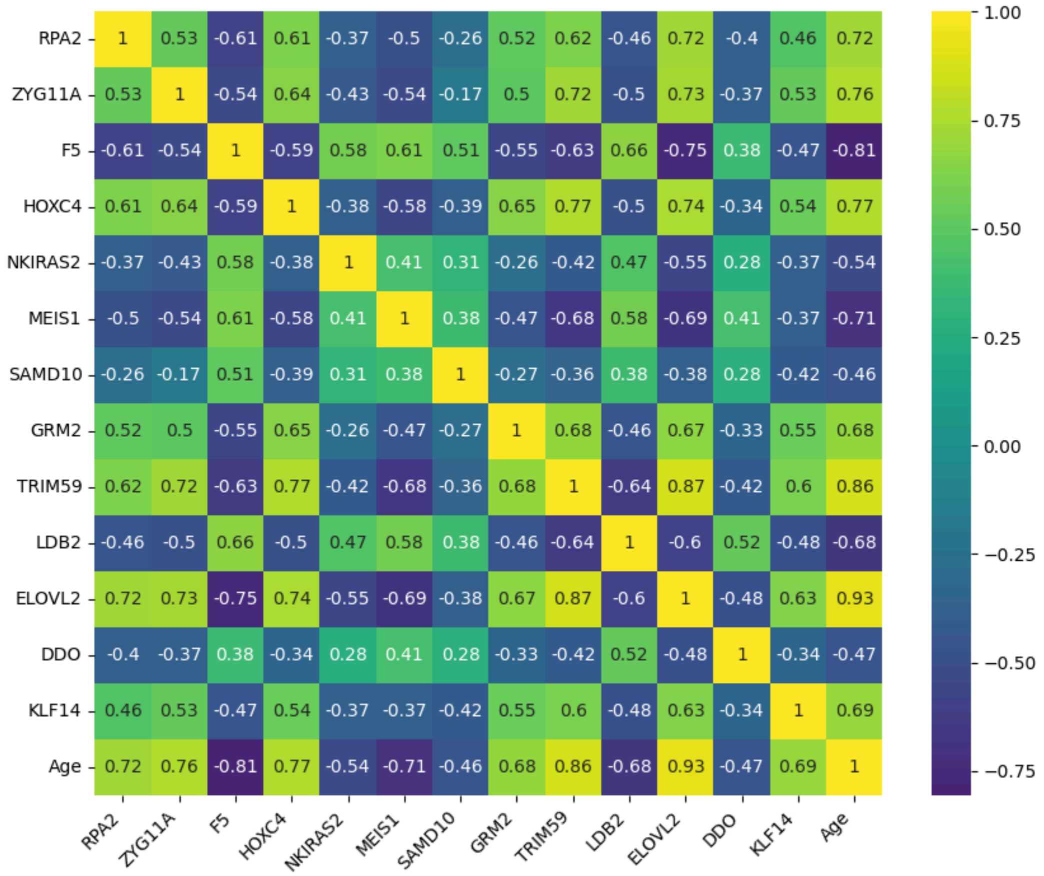
The RNA-seq dataset created by Fleischer et al. contains 133 rows and over 27,000 columns, each column representing expression of a single gene and each row a patient. The patient population is made up of male and females between 1 and 94 years of age. Data was obtained as described above. The RNAseq dataset is especially useful as there are very few datasets that have such a range of ages and an even distribution of women and men.

The Bisulfite-seq (methylation) dataset created by Naue et al., will be used to match a chronological age based on methylation to that of the expression of the same 13 biomarkers named in the introduction. This dataset includes 312 samples split into test and training sets at about a 2:1 ratio. The test set contains 104 rows and the training set has 208 all representing individual patients.

Preferred file formats include CSV, and TXT for their minimal size and ease of integration into electronic medical records. These files can be named according to current lab policy and de-identified to maintain patient security and privacy according to HIPAA, CLIA, and organizational standards. RNA-seq and methylation sequencing are not difficult to integrate as the sequencing methods are similar and the output datasets can easily be converted from CSV to TSV and back. Since the 13 biomarkers have been shown to have correlation with biological age, RNA-seq datasets can be narrowed down to only 13 features (genes) instead of the 27,000 or more genes typically seen in expression matrices. This allows for less data retention as models trained on this data can be saved in the cloud and called upon when needed.

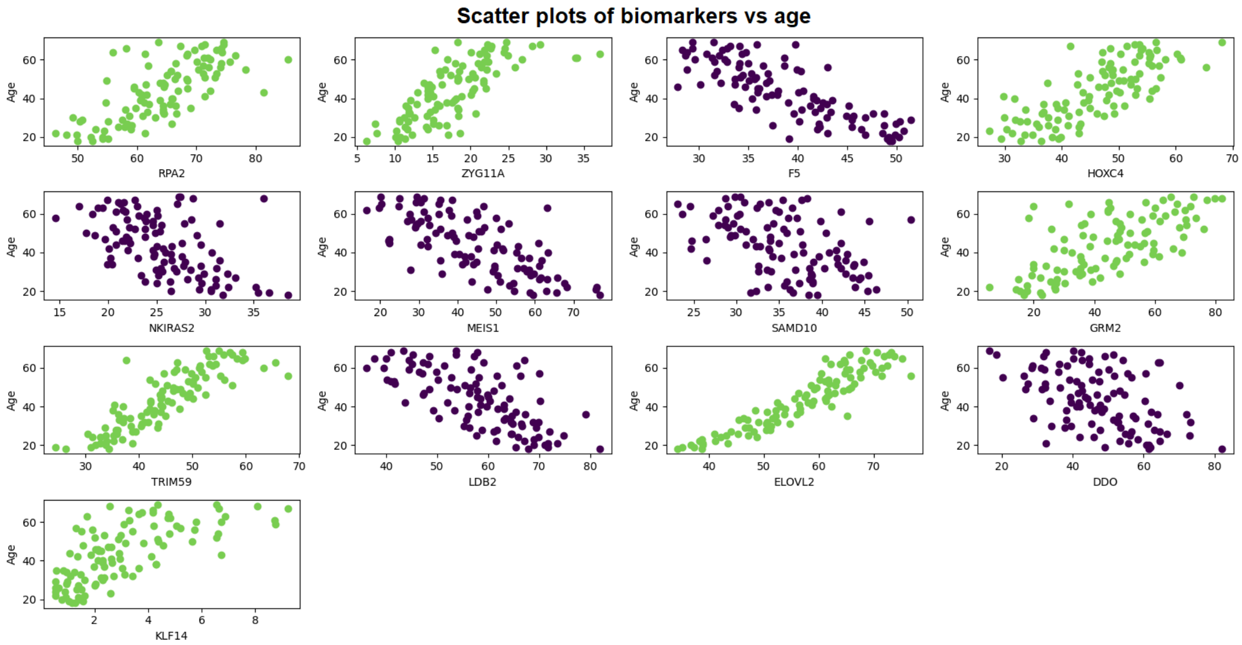
To remain in line with current clinical genomic data practices, preliminary data such as Fastq file formats will need to be housed in a data lake such as AWS S3 object storage. processed data such as BAM, SAM, VCF should not be kept for more than 1 year unless part of a large scale research study. With our data processing and analysis pipeline, original files can be copied and analysis performed as needed. Using open source packages such as Scikit-learn, pandas, and seaborn for data organization and exploratory analysis save time and money as these libraries are widely used. Oftentimes, libraries like Scikit-learn are used as the basis for autoML or other models. This allows for an ease of data integration that is necessary for multi-omic data analysis.

Data Visualization

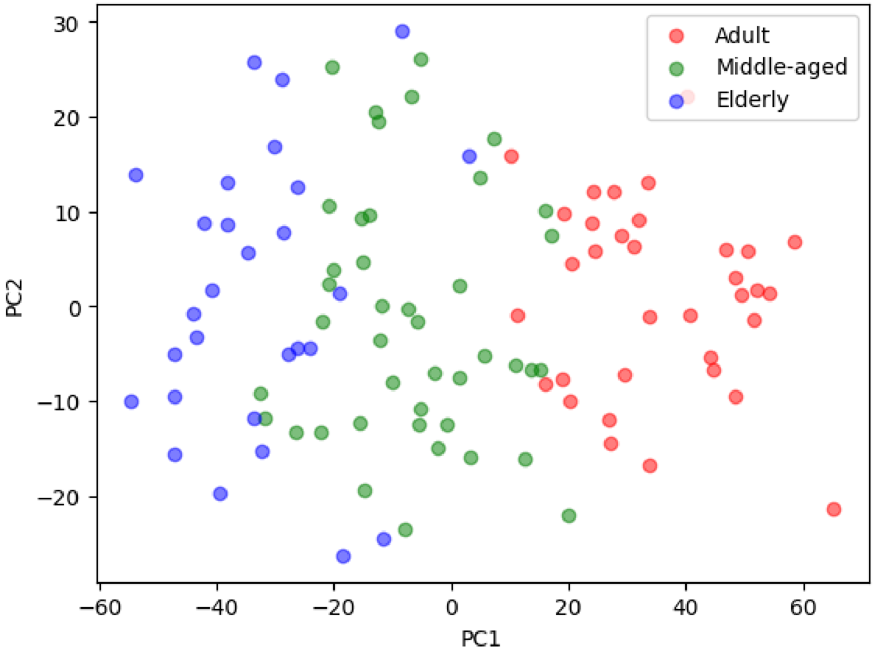


	RPA2	ZYG11A	F5	HOXC4	NKIRAS2	MEIS1	SAMD10	GRM2	TRIM59	LDB2	ELOVL2	DDO	KLF14	Age
RPA2	1	0.53	-0.61	0.61	-0.37	-0.5	-0.26	0.52	0.62	-0.46	0.72	-0.4	0.46	0.72
ZYG11A	0.53	1	-0.54	0.64	-0.43	-0.54	-0.17	0.5	0.72	-0.5	0.73	-0.37	0.53	0.76
F5	-0.61	-0.54	1	-0.59	0.58	0.61	0.51	-0.55	-0.63	0.66	-0.75	0.38	-0.47	-0.81
HOXC4	0.61	0.64	-0.59	1	-0.38	-0.58	-0.39	0.65	0.77	-0.5	0.74	-0.34	0.54	0.77
NKIRAS2	-0.37	-0.43	0.58	-0.38	1	0.41	0.31	-0.26	-0.42	0.47	-0.55	0.28	-0.37	-0.54
MEIS1	-0.5	-0.54	0.61	-0.58	0.41	1	0.38	-0.47	-0.68	0.58	-0.69	0.41	-0.37	-0.71
SAMD10	-0.26	-0.17	0.51	-0.39	0.31	0.38	1	-0.27	-0.36	0.38	-0.38	0.28	-0.42	-0.46
GRM2	0.52	0.5	-0.55	0.65	-0.26	-0.47	-0.27	1	0.68	-0.46	0.67	-0.33	0.55	0.68
TRIM59	0.62	0.72	-0.63	0.77	-0.42	-0.68	-0.36	0.68	1	-0.64	0.87	-0.42	0.6	0.86
LDB2	-0.46	-0.5	0.66	-0.5	0.47	0.58	0.38	-0.46	-0.64	1	-0.6	0.52	-0.48	-0.68
ELOVL2	0.72	0.73	-0.75	0.74	-0.55	-0.69	-0.38	0.67	0.87	-0.6	1	-0.48	0.63	0.93
DDO	-0.4	-0.37	0.38	-0.34	0.28	0.41	0.28	-0.33	-0.42	0.52	-0.48	1	-0.34	-0.47
KLF14	0.46	0.53	-0.47	0.54	-0.37	-0.37	-0.42	0.55	0.6	-0.48	0.63	-0.34	1	0.69
Age	0.72	0.76	-0.81	0.77	-0.54	-0.71	-0.46	0.68	0.86	-0.68	0.93	-0.47	0.69	1

This is a correlation matrix displaying the relationships between age and the 13 biomarkers: DDO, ELOVL2, F5, GRM2, HOXC4, KLF14, LDB2, MEIS1-AS3, NKIRAS2, RPA2, SAMD10, TRIM59, ZYG11A. The positive and negative bars on the right side of the matrix indicate the strength and direction of the correlations. Some genes are highly associated with each other, while others are not. The primary focus is on how each biomarker is linked to age. By examining the bottom row of the matrix, we observe that ELOVL2, TRIM59, HOXC4, ZYG11A, and RPA2 are the top five genes that are most strongly associated with age. ELOVL2 is involved in fatty acid metabolism, TRIM59 plays a role in the immune system, HOXC4 is connected with nervous system development, ZYG11A regulates the cell cycle, and RPA2 is involved in DNA repair, replication, transcription, cell cycle, and cellular response to external stimuli (Jana Naue et al., 2017). These top genes are all involved in pathways that are susceptible to external factors and can affect DNA methylation, a well-known aspect of epigenetics. Therefore, it is reasonable that these genes are the top biomarkers.

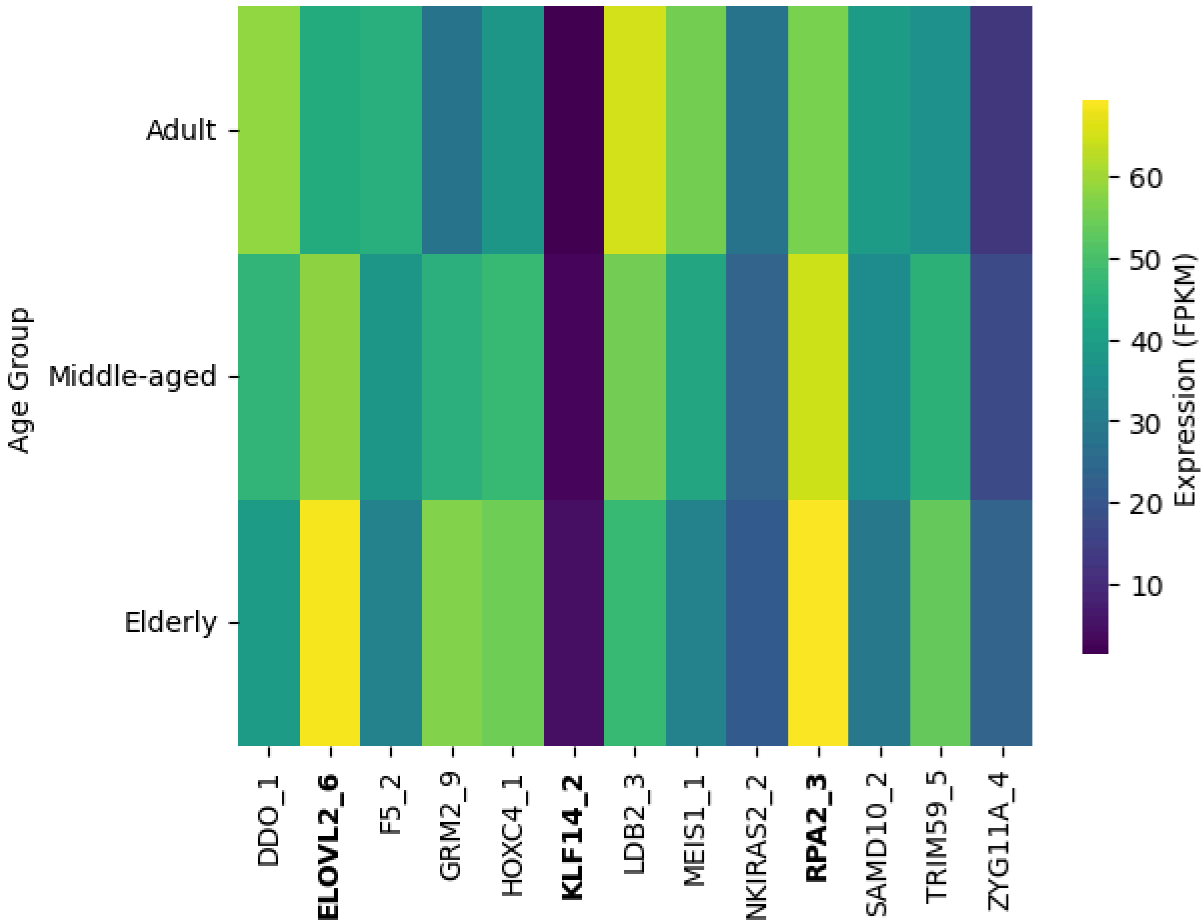


The scatter plot shown here displays the relationship between methylation levels and age for each individual biomarker. The color of each plot represents whether the gene is upregulated (green) or downregulated (dark blue-purple) with increasing age. This corresponds to the top 5 genes identified in the correlation matrix. The main aim of this scatter plot is to identify a patient's methylation value for each biomarker and use it to determine their biological age, which can be presented as feedback to the patient.



Principal component analysis (PCA) is a statistical method that helps to reduce the complexity of a dataset by transforming a large number of variables into a smaller set of uncorrelated variables, which are referred to as principal components (Jaadi, Z., 2022). These principal components are essentially linear combinations of the original variables that help to capture variations in the dataset (Jaadi, Z., 2022). The primary purpose of PCA is to identify patterns in data and to identify factors that contribute to variation in the dataset, which in our case, would be age groups (Jaadi, Z., 2022).

The results of the PCA analysis on the dataset of 13 biomarkers across more than 100 patients revealed the presence of three distinct age groups: adult, middle-aged, and elderly. According to the study's definition, adults are individuals below the age of 34, middle-aged are between 35 and 54, and elderly are above the age of 55. The 2D PCA plot shows a clear separation between these three age groups, indicating that these biomarkers can be used to distinguish between different age groups. The variance ratios for PC1 and PC2 were 64% and 11%, respectively, indicating that most of the variability in the dataset is explained by PC1. However, the study suggests that the dataset could be improved by including more data for individuals over the age of 69 and those under 18 to better understand the differences in the biomarker profiles between these age groups.



The heatmap shown above, highlights the different gene expression profiles across the three age groups. Where lighter areas indicate high expression of the corresponding gene which in turn implies low methylation of CpGs associated with said gene. Darker tiles indicate the inverse relationship showing very low expression of the associated gene and so those genes experience high methylation. We can also see that certain genes such as *ELOVL2* and *RPA2* increase in expression as the patient increases in biological age. One thing to note however, is that heatmap colors represent the most common expression level within a group for a gene and therefore can be subject to bias. In order for a heatmap to be most effective the data has to be normalized to reduce the impact of outlier bias.

Conclusion

In conclusion, our project aims to provide an easily accessible biomarker to identify age dysregulation and its association with underlying illnesses. By utilizing multi-omic markers to predict a patient's age, we aim to improve health outcomes and provide early intervention for patients who need it the most.

For general physicians/clinicians, we know that patients can struggle with coming in for check ups or health monitoring reasons. Our model offers a easy and convenient solution for accurate health monitoring. This age dysregulation testing model can be easily integrated into an annual routine check-up and is currently minimally invasive but we have plans to move towards reduced invasiveness. Although this model cannot serve as a diagnostic test, age dysregulation can be used as a general marker of poor health which can help accurately flag patients for follow-up testing.

For our patients, we know that it can be difficult and inconvenient to go see the doctor whenever you are feeling a bodily pain and oftentimes we do not know that something is wrong until it becomes too late. Our age dysregulation testing model allows for quick and cost-effective health monitoring that can be easily integrated into your annual checkups.

For genetic testing labs/diagnostic companies, out age dysregulation testing analysis pipeline is developed in python and thus can be easily added to a LIMS or a bioinformatic package suite. Our model features fast and easily accessible DNA which can be sourced from a routine blood draw and skin punch. We have identified an available patient market in health monitoring screens which may be potentially lucrative and have plans to become less invasive—removing the skin punch requirement—and to move towards more accessible direct-to-consumer testing.

References

1. Boskey, A. L., & Coleman, R. (2010). Aging and Bone. *Journal of Dental Research*, *89*(12), 1333–1348. https://doi.org/10.1177/0022034510377791
2. C. Jimmy Lin et al. (2021, January 15). *Freenome's Multiomics Blood Test Shows Promising Results in Detecting Colorectal Advanced Adenomas in a Prospective, Multi-Center Clinical Study*. Freenome. https://www.freenome.com/blood-based-detection-of-advanced-adenomas
3. Cope, H., Willis, C. R. G., MacKay, M. J., Rutter, L. A., Toh, L. S., Williams, P. M., Herranz, R., Borg, J., Bezdan, D., Giacomello, S., Muratani, M., Mason, C. E., Etheridge, T., & Szewczyk, N. J. (2022). Routine omics collection is a golden opportunity for European human research in space and analog environments. *Patterns*, *3*(10), 100550. https://doi.org/10.1016/j.patter.2022.100550
4. Fleischer, J. G., Schulte, R., Tsai, H. H., Tyagi, S., Ibarra, A., Shokhirev, M. N., Huang, L., Hetzer, M. W., & Navlakha, S. (2018). Predicting age from the transcriptome of human dermal fibroblasts. *Genome Biology*, *19*(1), 221. https://doi.org/10.1186/s13059-018-1599-6
5. Guo, J., Huang, X., Dou, L., Yan, M., Shen, T., Tang, W., & Li, J. (2022). Aging and aging-related diseases: From molecular mechanisms to interventions and treatments. *Signal Transduction and Targeted Therapy*, *7*(1), Article 1. https://doi.org/10.1038/s41392-022-01251-0
6. Hartmann, A., Hartmann, C., Secci, R., Hermann, A., Fuellen, G., & Walter, M. (2021). Ranking Biomarkers of Aging by Citation Profiling and Effort Scoring. *Frontiers in Genetics*, *12*. https://www.frontiersin.org/articles/10.3389/fgene.2021.686320
7. Karczewski, K. J., & Snyder, M. P. (2018). Integrative omics for health and disease. *Nature Reviews. Genetics*, *19*(5), 299–310. https://doi.org/10.1038/nrg.2018.4
8. Keaveny, T. M., Kopperdahl, D. L., Melton, L. J., Hoffmann, P. F., Amin, S., Riggs, B. L., & Khosla, S. (2010). Age-Dependence of Femoral Strength in White Women and Men. *Journal of Bone and Mineral Research*, *25*(5), 994–1001. https://doi.org/10.1359/jbmr.091033
9. Naue, J., Hoefsloot, H. C. J., Mook, O. R. F., Rijlaarsdam-Hoekstra, L., van der Zwalm, M. C. H., Henneman, P., Kloosterman, A. D., & Verschure, P. J. (2017). Chronological age prediction based on DNA methylation: Massive parallel sequencing and random forest regression. *Forensic Science International: Genetics*, *31*, 19–28. https://doi.org/10.1016/j.fsigen.2017.07.015
10. Tello, C. (2019, December 11). *Methylation—How Does It Affect Our Health & Aging?* SelfDecode Health. https://health.selfdecode.com/blog/what-is-methylation-and-how-does-it-affect-our-health/
11. Vaiserman, A., & Krasnienkov, D. (2021). Telomere Length as a Marker of Biological Age: State-of-the-Art, Open Issues, and Future Perspectives. *Frontiers in Genetics*, *11*, 630186. https://doi.org/10.3389/fgene.2020.630186
12. Erickson, Nick, et al. "AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data." arXiv preprint arXiv:2003.06505 (2020).
13. Jaadi, Z. (2022.). Principal component analysis (Pca) explained | built in. Retrieved May

6, 2023, from https://builtin.com/data-science/step-step-explanation-principal-component-analysis