AMP 297R @ Harvard IACS

Healthy Aging Signal Research Study

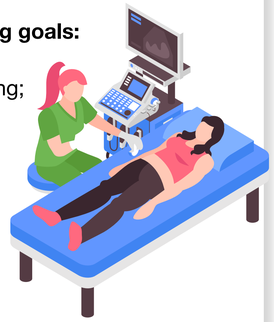
Partner: Merck & Co.
Team members: Daniel Cox, Yaxin Lei, Eleonora Shantsila, Aaron Jacobson

INSTITUTE FOR APPLIED COMPUTATIONAL SCIENCE AT HARVARD UNIVERSITY

1. Introduction and Motivation

Scientist have found that a person's biological age can be different from their chronological age. Our partner Merck has asked us to identify biomarkers and use them to build predictive models of chronological age. Such models could then be used as a baseline against which individuals might be compared whose cellular aging is aberrant due to disease.

2. Our Goals:

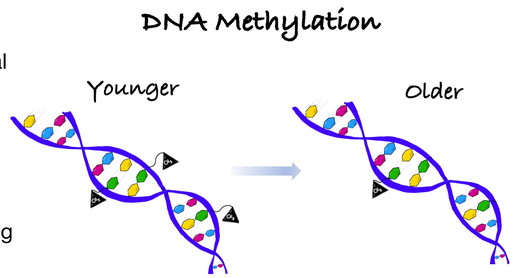


At the outset of this project we defined the following goals:
- identifying **databases** that are relevant to aging;
- Identifying good **biomarkers/features** of healthy aging;
- Building accurate **age-predictive models**;
- **Comparative study** on healthy vs. unhealthy aging, as well as biomarkers from different tissues.

3. Our Data:

We anchored our research on **DNA methylatio**n data, downloaded from the **EWAS** database.

Explanation: DNA methylation is a biological process by which methyl groups are added to the DNA molecule. Methylation can change the activity of a DNA segment without changing the sequence.



A quick peek of our data shows the following, where each sample individual (each row) is associated with a tissue, an age, and ~400,000 columns representing potential methylation site (CpG site).

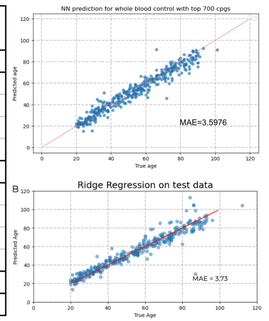
sample_id	tissue	age	cg02494853	cg03706273	cg04023335	cg05213048	cg15295597	cg26520468	cg27539833	cg00008
GSM2334366	whole blood	94	0.078	0.205	0.139	0.904	0.120	0.970	0.912	0.
GSM989863	whole blood	101	0.013	0.008	0.117	0.756	0.033	0.958	0.933	0.

	Healthy Control	Alzheimer's	Parkinson's	Huntington's
Whole Blood	1802	111	222	N/A
Brain	1064	811	N/A	270
Breast	520	N/A	N/A	N/A

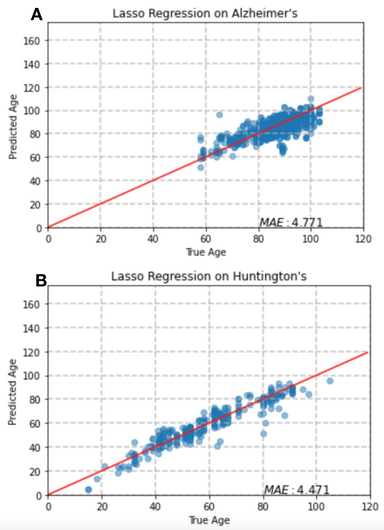
4. Age Predictive Modeling

Among ~400,000 CpG sites (columns), which are important?
To answer this question, we performed feature ranking/selection using feature importances from XGboost regression, and modeled with the top features.
Are features transferrable among tissues?
Using top CpG features ranked by XGboost blood regression, we modeled for age using brain (MAE 5.967), leukocyte (MAE 3.51) and breast (MAE 6.015) tissues, and found considerable feature transferability between tissues.
Are models transferrable among tissues?
No, when we used our blood trained model to predict for brain and breast.
Our results?
Refer to the below for results among different age predictive models. Specifically, Ridge regression and Neural Networks performed the best, both with MAE less than 4 years. Their performance on held out data is shown in the below right graphs.

Model	MSE	rMSE	MAE	r^2	Corr
1000 cpgs					
Linear	88.680	9.417	6.669	0.803	0.912
Ridge	36.610	6.051	3.733	0.918	0.959
Lasso	36.830	6.609	3.866	0.918	0.958
XGboost	34.580	5.880	4.109	0.923	0.961
100 cpgs					
Linear	41.635	6.453	4.224	0.907	0.953
Ridge	37.580	6.130	3.950	0.916	0.957
Lasso	37.510	6.125	3.881	0.916	0.957
Xgboost	35.380	5.948	4.126	0.921	0.960
700 cpgs					
Neural Net	23.470	4.841	3.597	—	—



5. Healthy vs Unhealthy



Are models transferable between healthy and unhealthy cohorts?
Models for the same tissues are transferable between healthy and unhealthy cohorts. For example for brain tissue the lasso regression model trained on healthy data achieved a MAE of 5.431 for the healthy cohort, 4.771 for Alzheimer's and 4.471 for Huntington's
Can the features be used to classify healthy and unhealthy cohorts?
Our best classification model using logistic regression achieved a class accuracy of 0.76 suggesting the features most associated with aging cannot be used effectively for healthy vs unhealthy cohort classification.

6. Biological Significance

Table below focuses on the top 20 highest ranked CpG sites and their associated genes:
1. KLF14, ELOVL2 and ZNF423 (blue) are ssociated with fat cells or fat metabolism. Thus, it may be that **processes involving fat metabolism and storage have an important influence on aging**.

2. Four genes are associated with the ubiquitin-proteasome pathway (red), OTUD7A, TRIM59, RNF180, and NHLRC1, an important pathway for **protein degradation**. In fact three of these genes are E3 ubiquitin ligases, which are responsible for marking proteins for degradation.

3. Third, many of the genes in the below table contain a structure known as a **zinc finger**, a structure that binds Zn2+ and is often involved in **DNA-protein and protein-protein interactions**.

4. Nearly all of the methylation sites in the table have been identified in other studies as being related to aging. Thus, our results are in accord with the growing literature on DNA methylation and aging.

Rank	cpg	Gene	Function	Zinc finger	Refs related to aging
1	cg14361627	KLF14	Krüppel-Like Factor 14 (KLF14), transcription factor, master regulator of gene expression in the adipose tissue	x	16, 8, 5, 7
2	cg16867657	ELOVL2	Fatty Acid Elongase 2, involved in the synthesis of very long polyunsaturated fatty acids		21, 15, 14, 17,5, 7
3	cg24724428	ELOVL2			15, 14, 17,5, 7
4	cg11649376	ACSS3	Acyl-CoA Synthetase Short Chain Family Member 3, Ligates acetate and CoA6		1
5	cg24079702	FHL2	Four And A Half LIM Domains 2, Assembly of extracellular membranes, double zinc finger, LIM protein	x	5,17,2,
6	cg04875128	OTUD7A	OTU Deubiquitinase 7A, deubiquitinizing enzyme and possible tumor suppressor, zinc finger	x	21, 17,7
7	cg08097417	KLF14		x	21, 16, 8, 5, 7
8	cg00292135	C7orf13	Not much known		
9	cg02046143	IGSF9B	Immunoglobulin Superfamily Member 9B, cell adhesion, localized to inhibitory synapses		21, 7
10	cg07553761	TRIM59	Tripartite Motif Containing 59, E3 ubiquitin ligase, zinc finger, RING finger protein	x	15, 7
11	cg21572722	ELOVL2			15, 14, 17,5, 7
12	cg04208403	ZNF423	Zinc Finger Protein 423, Krüppel-Like Factor, zinc finger transcription factor, KO affects adipogenesis	x	16
13	cg23500537				21
14	cg08262002	LDB2	LIM Domain Binding 2, adapter molecule, binds LIM		14,15
15	cg04955333	IQCE	IQ Motif Containing E, signaling by GPCR and Hedgehog		21
16	cg09809672	EDARADD	EDAR Associated Death Domain, Ectodysplasin-A receptor-associated adapter protein		21, 16, 4, 9
17	cg06639320	FHL2		x	21, 5,17,2,
18	cg17621438	RNF180	E3 Ubiquitin-Protein Ligase RNF180, promotes protein degradation by the proteasome pathway	x	21
19	cg22736354	NHLRC1	E3 Ubiquitin-Protein Ligase NHLRC1 , promotes protein degradation by the proteasome pathway	x	9
20	cg22454769	FHL2		x	21, 5,17,2,
21	cg19344626	NWD1	NACHT And WD Repeat Domain Containing 1, modulator of androgen receptor activity		
22	cg23744638				21
23	cg07850154	RNF180		x	

7. Conclusion

- By studying the relationship between chronological age and DNA methylation in blood, we have been able to build models to **predict age** with a **mean absolute error of 3.6 years**.
- From the ~ 400,000 DNA methylation sites (CpG sites) we started with, we have identified **~700 that are optimal for age predictive modeling**.
  - These same CpGs can also be used in building models with DNA methylation data from **other tissues** and in **both healthy and unhealthy cohorts.**
- Our top ranked CpGs are often associated with genes that regulate **adipose tissue gene expression** and the **ubiquitin-proteosome protein degradation pathway**.

Special thanks to Doctor Chris Tanner, TF Phoebe Wong and the amazing team from Merck & Co for their help! For full technical report, code base and reference, refer to https://github.com/AC297r-Merck/Healthy-Aging