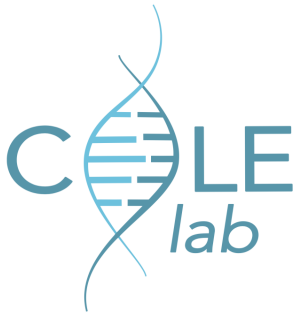
# Using GWAS to Study Genetic Determinants of Fruit Consumption and Aging Gaps

Davyd Sadovskyy

PATH-GDS Rotation 2

12/10/2025



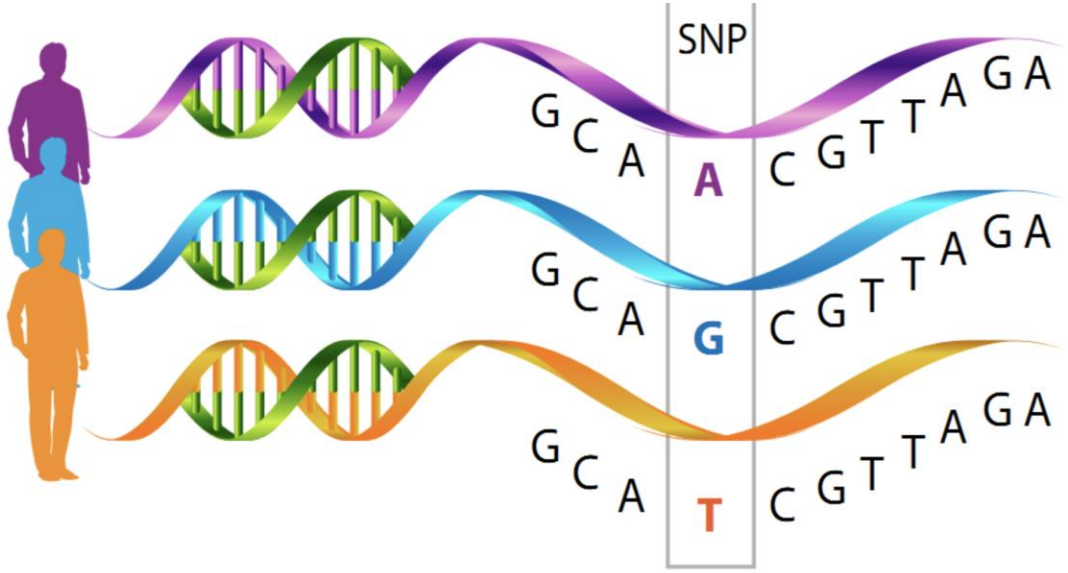

University of Colorado
Anschutz Medical Campus

# GWAS Background

- A genome wide association study (GWAS) is a statistical technique to find associations between a trait and DNA regions

- Many traits can be analyzed using GWAS:
  - memory performance
  - Resilience to sleep deprivation
  - degree of risk taking
  - Fruit intake and aging gaps (how "well" you age)

- Single Nucleotide Polymorphisms (SNPs) are the basis of many GWAS studies.

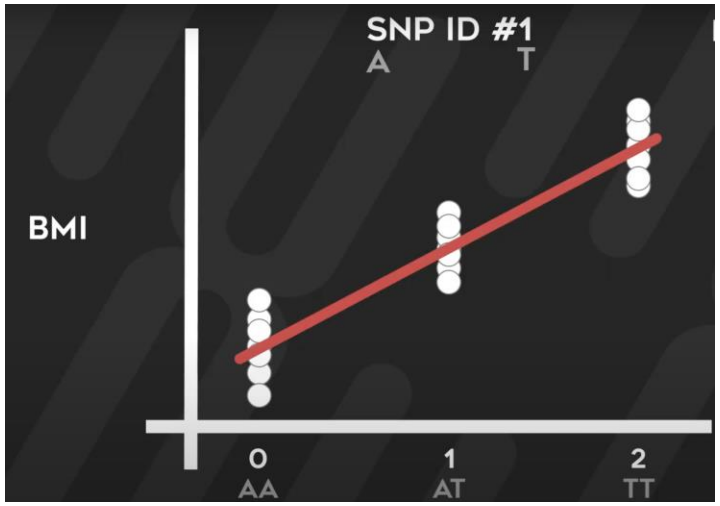


# The Basic Math

- How to find if a SNP is associated with a trait? - use regression

Code SNPs

- `0` : Homozygous alternative ( `A/A` ).
- `1` : Heterozygous ( `G/A` ).
- `2` : Homozygous reference ( `G/G` ).




Calculate slope (Beta value)

Calculate p – value and see if it's a significant association

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$$

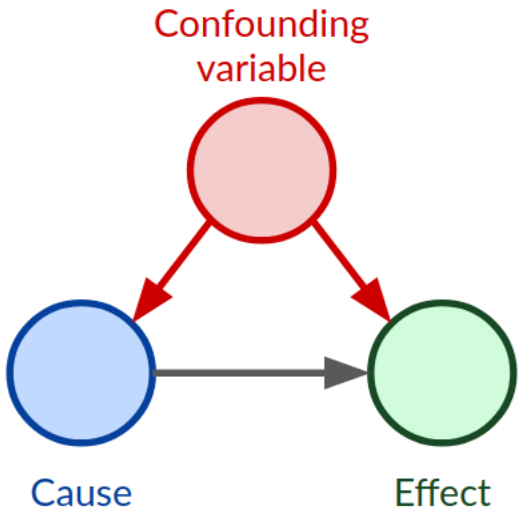
Trait of interest

Slope

SNP value

# Statistical Modeling Considerations

- Confounding variables like age, sex, and ancestry must be controlled for.




  - Ex. Chopstick use – Chinese people have more common chopstick use. When differentiating what SNPs influence chopstick use, GWAS will say it's the "Chinese ethnicity" SNPs. But these aren't the genuine genetic contributors to chopstick use we are looking for.

- If you use standard significance threshold of 0.05, just by random chance, 50k SNPs will show association (1million x 0.05 = 50000)

$$\alpha_{\text{corrected}} = \frac{\alpha}{\text{Number of SNPs}} = \frac{0.05}{1,000,000} = 5 \times 10^{-8}$$

# The GWAS Pipeline

- Genetic data is sensitive and can't be downloaded. All data is stored  DNA Nexus cloud. Scripts can be sent to the computing platform to perform analysis.

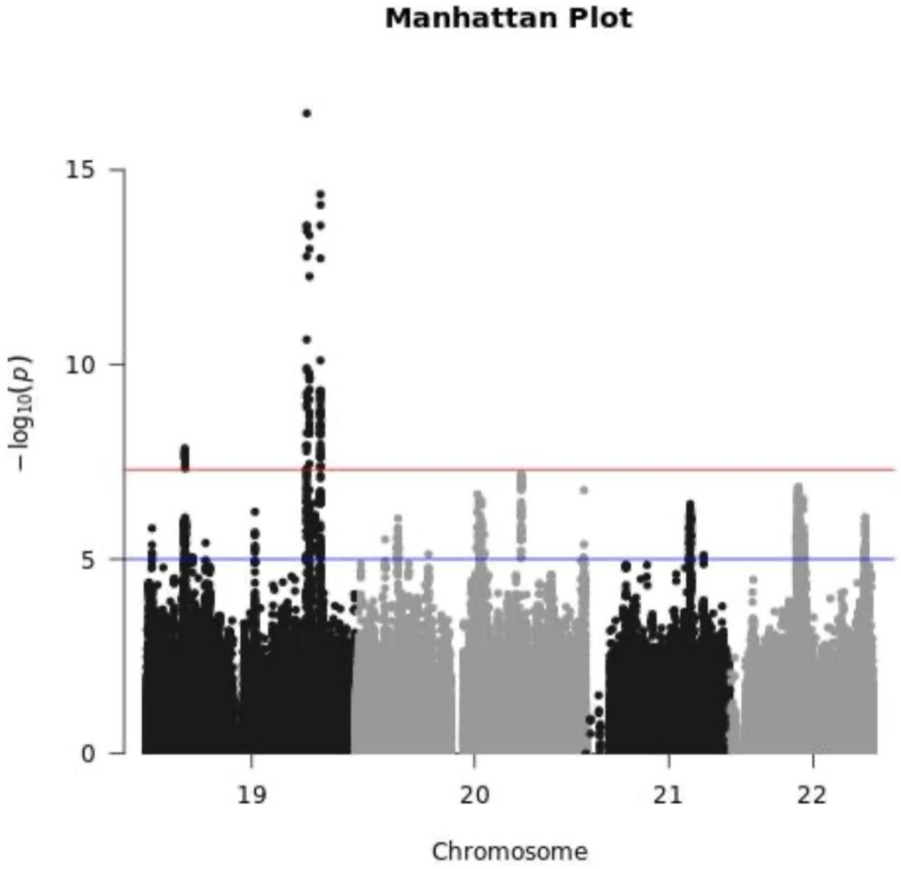
```
(base) davydsadovskyy@davyds-mbp scripts % dx 01-example_gwas_merge.sh
```

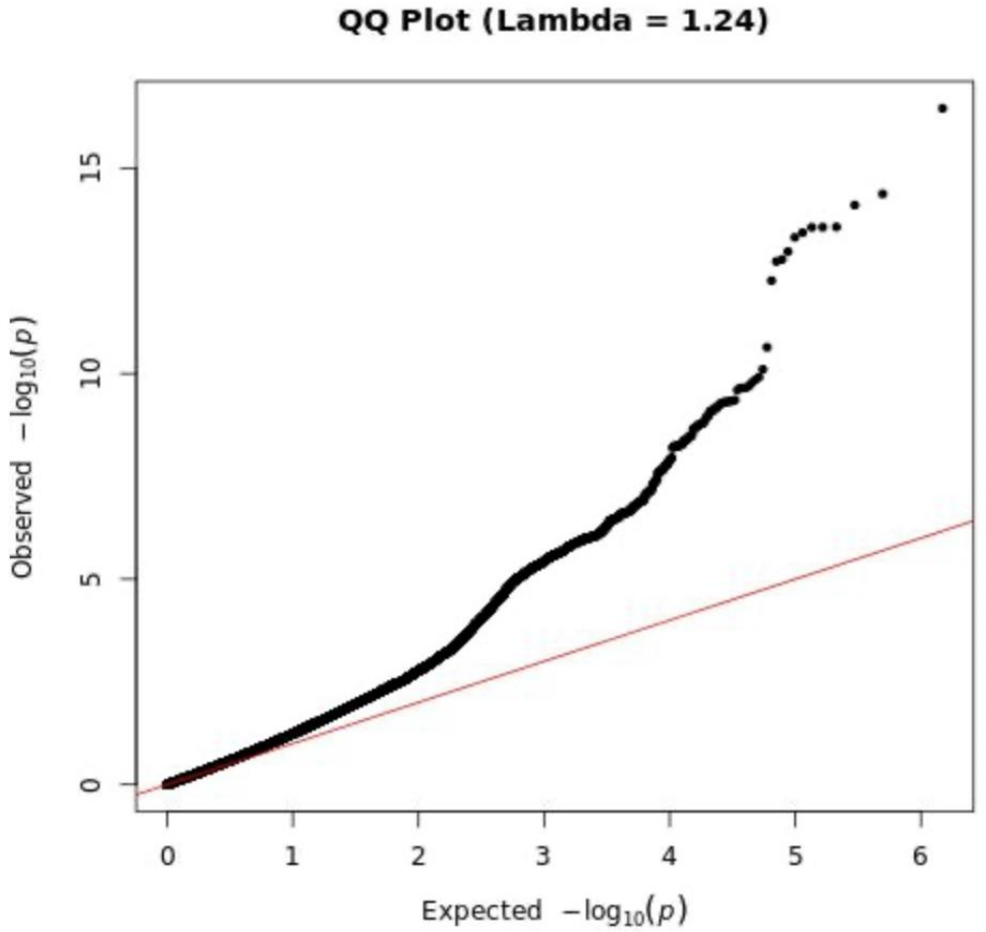
- 01-example_gwas_merge.sh
  - Merge individual chromosome data into 1 file
- 02-example_gwas_filter.sh
  - Create .txt file for high quality individuals and high quality SNPs
- 03-example_gwas_regenie_step1.sh
  - Step 1 of regenie
- 04-example_gwas_regenie_step2.sh
  - Step 2 of regenie
- 05-example_gwas_merge_filter_results.sh
- 06-example_gwas_plot_results.sh
- 07-example_gwas_clump_results.sh

**The actual GWAS**

DNAnexus®

# Fruit Consumption GWAS Results

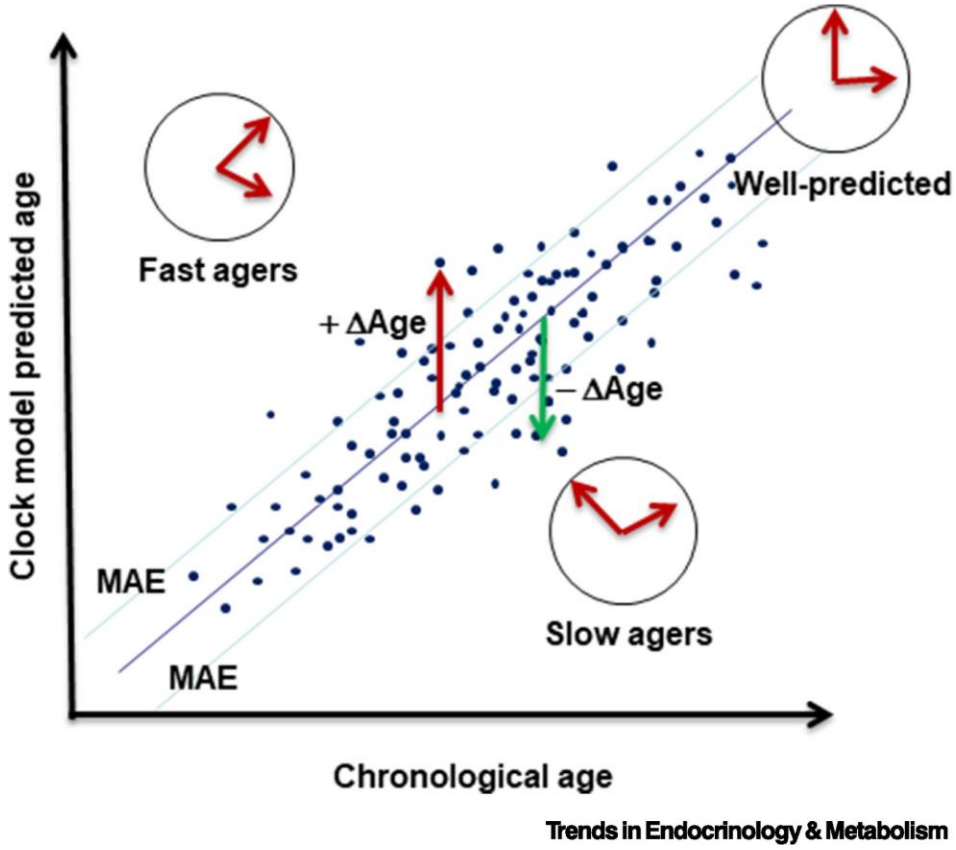



1. **Straight Line at the Start:**
   - The lower range of the plot (non-significant p-values) should closely follow the diagonal.
   - This indicates that the majority of SNPs behave as expected under the null hypothesis.
2. **Upward Deviation at the End:**
   - The upper-right part of the plot may show points deviating above the diagonal.
   - These represent SNPs with observed p-values much smaller (more significant) than expected, indicating potential true associations.

# Aging Gaps

- An aging clock is a model that predicts age as a function of biomarkers.
  - The biomarkers can be transcriptomics, metabolomics, epigenetic, proteomics, brainwaves, etc
- Aging gap is the difference between someone's chronological age and their predicted age.
  - It has biological significance.
- **Are there genetic determinants to one's "aging gap"?**



*(Argentieri et al., 2024)*

# Aging Gap GWAS


- UKB database has OLINK protein data that can be used to build aging clocks.

| eid | ins_index | birth_date | age_at_blood_draw | date_of_blood_draw | a1bg | aamd | aarsd | ... |
|---|---|---|---|---|---|---|---|---|
| 1000001 | 2 | xx/xxxx | xx.xx | xx/xxxx | 0.1234 | 0.2512 | -0.323 | ... |
| 1000002 | 0 | xx/xxxx | xx.xx | xx/xxxx | null | 0.3561 | 0.754 | ... |
| 1000003 | 0 | xx/xxxx | xx.xx | xx/xxxx | null | null | null | ... |
| 1000004 | 0 | xx/xxxx | xx.xx | xx/xxxx | 0.0464 | 0.4999 | -0.170 | ... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |

**53057 rows**
**2923 columns**

- Created a clean dataset ready for building aging clock but ran into computing issues.

- Future Directions:
  - A recent study used the same OLINK dataset to build an aging clock and found 204 of the 2923 proteins to be most predictive of age (Argentieri et al). **How would GWAS of the aging gap compare, and what genes might overlap?**

# Regenie Details

*(Mbatchou et al., 2020)*

**Step 1**

- First, remove effects of covariates

$$\tilde{y} = P_X y = (I_N - X(X^T X)^{-1} X^T) y \qquad \tilde{G} = P_X G_S$$

- Condense G (huge matrix with SNP data) into W.

This greatly reduces memory usage, but retains information.

$$\widetilde{\mathbf{y}} = \widetilde{\mathbf{G}}_i \boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

$$\mathbf{W}_i = (\widetilde{\mathbf{G}}_i \widehat{\gamma}_{\lambda_1}, \ldots, \widetilde{\mathbf{G}}_i \widehat{\gamma}_{\lambda_R}), \qquad i = 1, \ldots, B$$

$$\mathbf{W} = (\mathbf{W}_1, \ldots, \mathbf{W}_B)$$

- One more regression

$$\widetilde{\mathbf{y}} = \mathbf{W} \boldsymbol{\eta} + \boldsymbol{\epsilon}$$

**Step 2**

- Association testing for individual SNPs

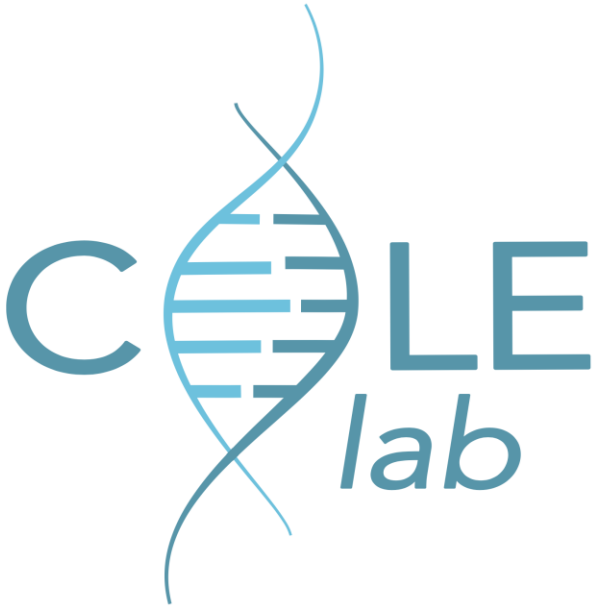
$$T_{\log} = \frac{\widetilde{\mathbf{g}}^T (\mathbf{y} - \widehat{\mathbf{p}})}{[\widetilde{\mathbf{g}}^T \Gamma \widetilde{\mathbf{g}}]^{1/2}}$$

↓

**P-values**

# Acknowledgments


- Joanne Cole, PhD
- Kristen J. Sutton, PhD
- Maizy Brasher, MA

# References

Argentieri, M. A., Xiao, S., Bennett, D., Winchester, L., Nevado-Holgado, A. J., Ghose, U., Albukhari, A., Yao, P., Mazidi, M., Lv, J., Millwood, I., Fry, H., Rodosthenous, R. S., Partanen, J., Zheng, Z., Kurki, M., Daly, M. J., Palotie, A., Adams, C. J., … van Duijn, C. M. (2024). Proteomic aging clock predicts mortality and risk of common age-related diseases in diverse populations. *Nature Medicine, 30*(9), 2450–2460. https://doi.org/10.1038/s41591-024-03164-7

Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., Habegger, L., Ferreira, M., Baras, A., Reid, J., Abecasis, G., Maxwell, E., & Marchini, J. (2020). Computationally efficient whole genome regression for quantitative and binary traits. *BioRxiv*. https://doi.org/10.1101/2020.06.19.162354