IT5007: Software Engineering on Application Architecture

Course Project Interim Evaluation

Machine Learning Genomics App

SUPRATIK SEKHAR
BHATTACHARYA_A0228511M

HUANG DAN_A0241594B
28 Feb

1. **Problem Statement:**

Our “ML-based genomics analysis app” aims to utilize state-of-the-art machine learning technology in other research areas.

In terms of research side novelty we are working on developing machine learning algorithm based ancestry breakdowns . This is also informed by domain knowledge of genomics, history and ethnography.

Many researchers and data analysts may have difficulty in running off-the-shelf machine learning models, among other reasons because they need to install the development tools, packages and have to become familiar with the programming languages. There are online tools such as “Visualgo" which implement all kinds of algorithms to help students easily understand algorithms and use the tool. However, there are not such open and free tools in the Genomics research area. Our project is to help them out by implementing all the things together. To help research and analysis with relative ease without having to download any development tools and installing related packages.

We propose an app with 3 major functionalities

1. Genetic Distance Tool:

It is to calculate the genetic distance of the input sample to that with other samples. Will return the nearest neighbors upon pressing a button. The whole distance matrix upon entering n samples

2. Genetic PCA Tool:

This functionality will make the genetic PCA plot
Since PCA is a linear dimensionality reduction method the relative distances are comparable
We might look into some novel things such as making Umap and t-SNE plots for further fine-grained visualization if time permits.

3. Ancestry Breakdown:

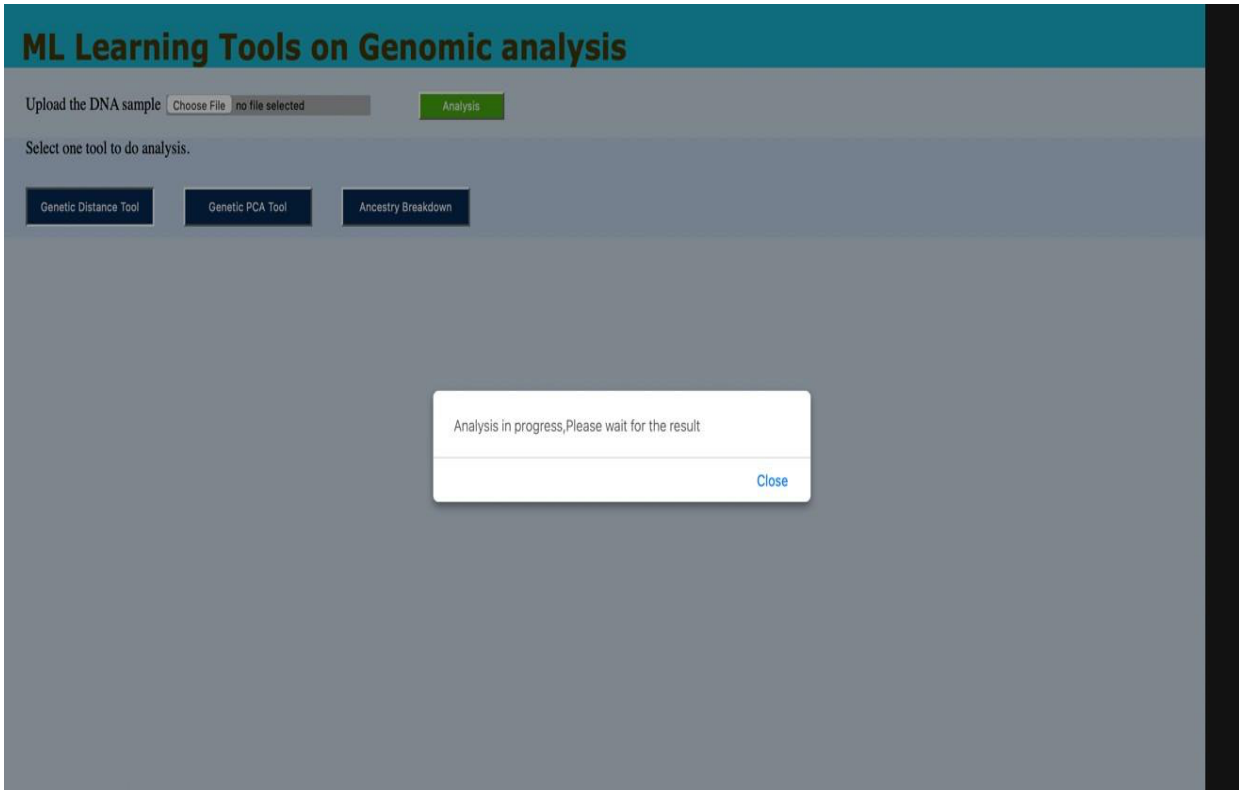
Fuzzification is performed after the machine learning clustering is done. This is after the models have been made and the clusters have been generated. This allows us to come up with the ancestry percentages. Various models and model specifications will be attempted. Domain knowledge comes into play in terms of the results making sense.. The ancestral components are selected such that they correspond to realistic source populations.

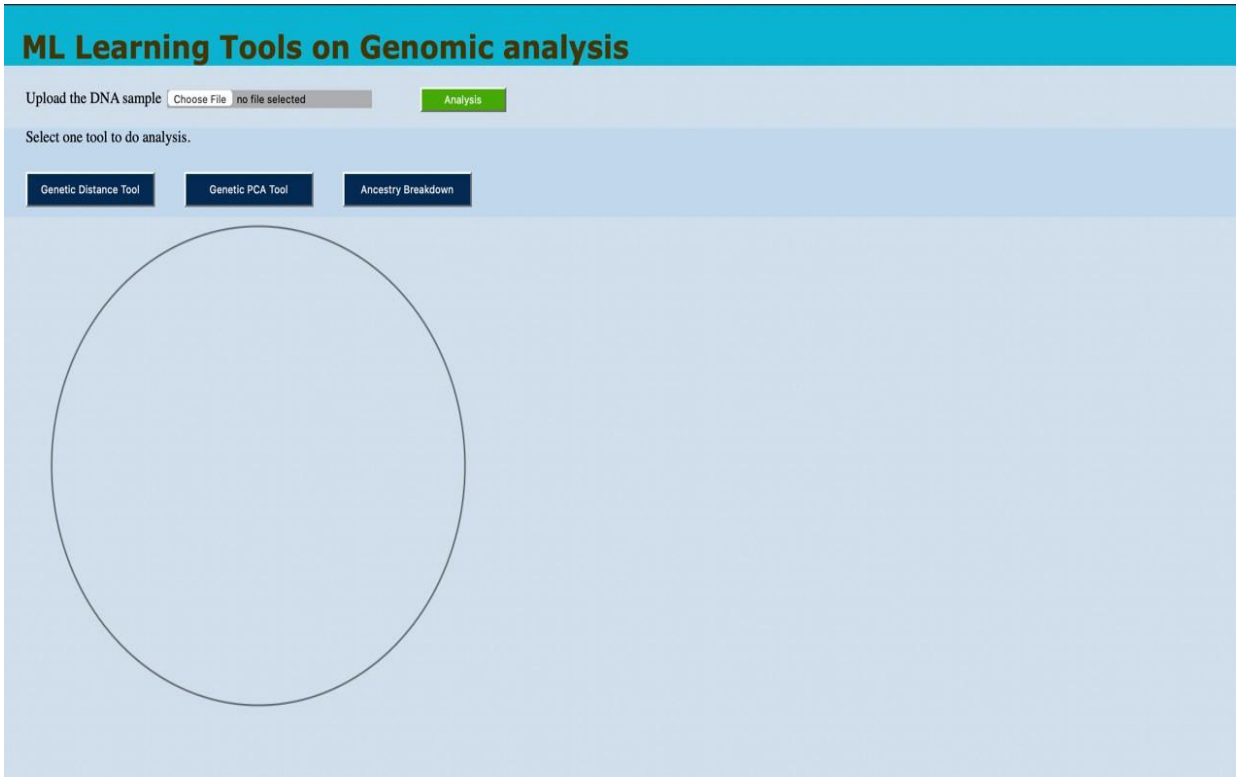
Will look into using various machine learning algorithms to develop the ancestry tool such as:

Fuzzy c means	GaussianMixture
SpectralClustering	AffinityPropagation
AgglomerativeClustering	Birch
DBSCAN	K-means
MiniBatchKMeans	MeanShift
OPTICS	SpectralClustering

2. Solution Architecture

The frontend has been designed to be simple and easy to use. Users need only to upload their DNA sample followed by selecting a machine learning tool(“Genetic Distance Tool”, “Genetic PCA Tool”, “Ancestry Breakdown”) and then click on the “Analysis” button, our machine learning backend will output the results



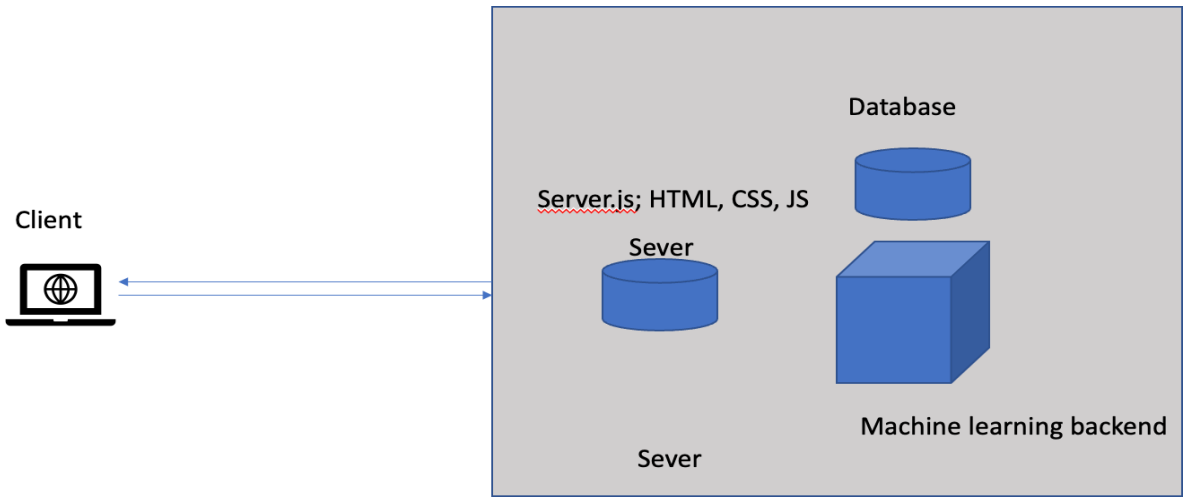


We plan to implement the machine learning model as the backend and deploy it

The generation of ancestries involves fuzzification of results from unsupervised learning algorithms

This step is done ex-post the clustering for ease of computation

The architecture will be as given below:



3. **Legal/Other Aspects**:

Open source libraries to be used:

Numpy

Pandas

Matplotlib

Scikit-Learn

Umap

etc.

Various data processing, genomics and machine learning packages will be used.

4. Competition Analysis

Vahaduo has a separate custom PCA tool and a Admixture tool which has the genetic distance functionality

The research originality is terms of the ancestry breakdown.
There is no competition for the machine learning based ancestry breakdowns to the best of our knowledge.
(https://vahaduo.github.io/vahaduo/)

5. Feature List

Project Schedule			
Documentation	**PIC**	**Due date**	**Status**
Project discussion report	HUANG DAN and SUPRATIK BHATTACHARYA	23-Feb	Done
Interim Evaluation report	HUANG DAN and SUPRATIK BHATTACHARYA	01-Mar	Done
Final Evaluation report	HUANG DAN and SUPRATIK BHATTACHARYA	01-Apr	Onging
FrontEnd			
Mainpage	HUANG DAN	31-Mar	Onging
Genetic Distance Tool	HUANG DAN	31-Mar	Onging
Genetic PCA Tool	HUANG DAN	31-Mar	Onging
Ancestry Breakdown	HUANG DAN	31-Mar	Onging
BackEnd			
Backend related work	HUANG DAN and SUPRATIK BHATTACHARYA	23-Apr	Onging
Genetic Distance Tool	SUPRATIK BHATTACHARYA	23-Apr	Onging
Genetic PCA Tool	SUPRATIK BHATTACHARYA	23-Apr	Onging
Ancestry Breakdown	SUPRATIK BHATTACHARYA	23-Apr	Onging

6. Git Repository :

The Repository on Github is given below:
https://github.com/dandelion793/IT5007CourseProject

Class Participation (2 points): Participation in the class discussions on MS Teams will be evaluated. Cases where the student writes an explanatory post for the peers will be considered:
We have attended the discussion meeting with the professor at 3:15pm on 23 Feb.

Appendix A:

Data sources-Ancient DNA samples:

- Lazaridis, I., Mittnik, A., Patterson, N., Mallick, S., Rohland, N., Pfrengle, S., ... & Stamatoyannopoulos, G. (2017). Genetic origins of the Minoans and Mycenaeans. *Nature*, *548*(7666), 214-218.
- Clemente, F., Unterländer, M., Dolgova, O., Amorim, C. E. G., Coroado-Santos, F., Neuenschwander, S., ... & Papageorgopoulou, C. (2021). The genomic history of the Aegean palatial civilizations. *Cell*, *184*(10), 2565-2586.
- Seguin-Orlando, A., Donat, R., Der Sarkissian, C., Southon, J., Thèves, C., Manen, C., ... & Orlando, L. (2021). Heterogeneous hunter-gatherer and steppe-related ancestries in Late Neolithic and Bell Beaker genomes from present-day France. *Current Biology*, *31*(5), 1072-1083.
- Olalde, I., Brace, S., Allentoft, M. E., Armit, I., Kristiansen, K., Booth, T., ... & Reich, D. (2018). The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature*, *555*(7695), 190-196.
- Patterson, N., Isakov, M., Booth, T., Büster, L., Fischer, C. E., Olalde, I., ... & Mihovilić, K. (2022). Large-scale migration into Britain during the Middle to Late Bronze Age. *Nature*, *601*(7894), 588-594.
- Wang, W., Ding, M., Gardner, J. D., Wang, Y., Miao, B., Guo, W., ... & Fu, Q. (2021). Ancient Xinjiang mitogenomes reveal intense admixture with high genetic diversity. *Science Advances*, *7*(14), eabd6690.
- Llorente, M. G., Jones, E. R., Eriksson, A., Siska, V., Arthur, K. W., Arthur, J. W., ... & Manica, A. (2015). Ancient Ethiopian genome reveals extensive Eurasian admixture in Eastern Africa. *Science*, *350*(6262), 820-822.
- Papac, L., Ernée, M., Dobeš, M., Langová, M., Rohrlach, A. B., Aron, F., ... & Haak, W. (2021). Dynamic changes in genomic and social structures in third millennium BCE central Europe. *Science Advances*, *7*(35), eabi6941.
- Lipson, M., Cheronet, O., Mallick, S., Rohland, N., Oxenham, M., Pietrusewsky, M., ... & Reich, D. (2018). Ancient genomes document multiple waves of migration in Southeast Asian prehistory. *Science*, *361*(6397), 92-95.
- Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S. M., Bondarev, A. A., ... & Pääbo, S. (2014). Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*, *514*(7523), 445-449.
- Yang, M. A., Fan, X., Sun, B., Chen, C., Lang, J., Ko, Y. C., ... & Fu, Q. (2020). Ancient DNA indicates human population shifts and admixture in northern and southern China. *Science*, *369*(6501), 282-288.
- Harney, É., Nayak, A., Patterson, N., Joglekar, P., Mushrif-Tripathy, V., Mallick, S., ... & Rai, N. (2019). Ancient DNA from the skeletons of Roopkund Lake reveals Mediterranean migrants in India. *Nature communications*, *10*(1), 1-10.
- Narasimhan, V. M., Patterson, N., Moorjani, P., Rohland, N., Bernardos, R., Mallick, S., ... & Reich, D. (2019). The formation of human populations in South and Central Asia. *Science*, *365*(6457), eaat7487.
- Yao, H., Wang, M., Zou, X., Li, Y., Yang, X., Li, A., ... & He, G. (2021). New insights into the fine-scale history of western–eastern admixture of the northwestern Chinese population in the Hexi Corridor via genome-wide genetic legacy. *Molecular Genetics and Genomics*, *296*(3), 631-651.
- Wang, C. C., Yeh, H. Y., Popov, A. N., Zhang, H. Q., Matsumura, H., Sirak, K., ... & Reich, D. (2021). Genomic insights into the formation of human populations in East Asia. *Nature*, *591*(7850), 413-419.

- Agranat-Tamir, L., Waldman, S., Martin, M. A., Gokhman, D., Mishol, N., Eshel, T., ... & Reich, D. (2020). The genomic history of the Bronze Age southern Levant. *Cell*, *181*(5), 1146-1157.
- Wang, C. C., Reinhold, S., Kalmykov, A., Wissgott, A., Brandt, G., Jeong, C., ... & Haak, W. (2019). Ancient human genome-wide data from a 3000-year interval in the Caucasus corresponds with eco-geographic regions. *Nature communications*, *10*(1), 1-13.
- Zhao, Y. B., Zhang, Y., Zhang, Q. C., Li, H. J., Cui, Y. Q., Xu, Z., ... & Zhu, H. (2015). Ancient DNA reveals that the genetic structure of the northern Han Chinese was shaped prior to 3,000 years ago. *PLoS One*, *10*(5), e0125676.
- Wang, C. C., Yan, S., Yao, C., Huang, X. Y., Ao, X., Wang, Z., ... & Li, H. (2013). Ancient DNA of Emperor CAO Cao's granduncle matches those of his present descendants: a commentary on present Y chromosomes reveal the ancestry of Emperor CAO Cao of 1800 years ago. *Journal of human genetics*, *58*(4), 238-239.
- Jeong, C., Wang, K., Wilkin, S., Taylor, W. T. T., Miller, B. K., Bemmann, J. H., ... & Warinner, C. (2020). A dynamic 6,000-year genetic history of Eurasia's Eastern Steppe. *Cell*, *183*(4), 890-904.
- Omrak A., Günther T., Valdiosera C., Svensson E. M., Malmström H., Kiesewetter H., Aylward W., Storå J., Jakobsson M., Götherström A., Genomic evidence establishes Anatolia as the source of the European Neolithic gene pool. *Curr. Biol.* ***26***, 270–275 (2016).
- Gallego-Llorente, M., Connell, S., Jones, E. R., Merrett, D. C., Jeon, Y., Eriksson, A., ... & Pinhasi, R. (2016). The genetics of an early Neolithic pastoralist from the Zagros, Iran. *Scientific reports*, *6*(1), 1-7.
- Broushaki, F., Thomas, M. G., Link, V., López, S., van Dorp, L., Kirsanow, K., ... & Burger, J. (2016). Early Neolithic genomes from the eastern Fertile Crescent. *Science*, *353*(6298), 499-503.

References

- Pardo-Seco, J., Gómez-Carballa, A., Amigo, J., Martinón-Torres, F., & Salas, A. (2014). A genome-wide study of modern-day Tuscans: revisiting Herodotus's theory on the origin of the Etruscans. *PLoS One*, *9*(9),

e105920.
- Trivedi, R., Sahoo, S., Singh, A., Bindu, G. H., Banerjee, J., Tandon, M., ... & Kashyap, V. K. (2008). Genetic imprints of pleistocene origin of indian populations: a comprehensive Phylogeographic sketch of Indian Y-chromosomes. *International Journal of Human Genetics*, *8*(1-2), 97-118.
- Metspalu, M., Kivisild, T., Metspalu, E., Parik, J., Hudjashov, G., Kaldma, K., ... & Villems, R. (2004). Most of the extant mtDNA boundaries in south and southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. *BMC genetics*, *5*(1), 1-25.
- Pathak, A. K., Kadian, A., Kushniarevich, A., Montinaro, F., Mondal, M., Ongaro, L., ... & Villems, R. (2018). The genetic ancestry of modern Indus valley populations from northwest India. *The American Journal of Human Genetics*, *103*(6), 918-929.
- Mahal, D. G., & Matsoukas, I. G. (2017). Y-STR haplogroup diversity in the Jat population reveals several different ancient origins. *Frontiers in genetics*, *8*, 121.
- Sharma, S., Rai, E., Sharma, P., Jena, M., Singh, S., Darvishi, K., ... & Bamezai, R. N. (2009). The Indian origin of paternal haplogroup R1a1* substantiates the autochthonous origin of Brahmins and the caste system. *Journal of human genetics*, *54*(1), 47-55.
- Sengupta, S., Zhivotovsky, L. A., King, R., Mehdi, S. Q., Edmonds, C. A., Chow, C. E. T., ... & Underhill, P. A. (2006). Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. *The American Journal of Human Genetics*, *78*(2), 202-221.
- Watson, J. D. (1990). The human genome project: past, present, and future. *Science*, *248*(4951), 44-49.
- Collins, F. S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project: lessons from large-scale biology. *Science*, *300*(5617), 286-290.
- Gibbs, R. A. (2020). The human genome project changed everything. *Nature Reviews Genetics*, *21*(10), 575-576.
- Diaz-Papkovich, A., Anderson-Trocmé, L., Ben-Eghan, C., & Gravel, S. (2019). UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS genetics*, *15*(11), e1008432.
- Dorrity, M. W., Saunders, L. M., Queitsch, C., Fields, S., & Trapnell, C. (2020). Dimensionality reduction by UMAP to visualize physical and genetic interactions. *Nature communications*, *11*(1), 1-6.
- Daly, K. G., Mattiangeli, V., Hare, A. J., Davoudi, H., Fathi, H., Doost, S. B., ... & Bradley, D. G. (2021). Herded and hunted goat genomes from the dawn of domestication in the Zagros Mountains. *Proceedings of the National Academy of Sciences*, *118*(25).
- Miller, W., Schuster, S. C., Welch, A. J., Ratan, A., Bedoya-Reina, O. C., Zhao, F., ... & Lindqvist, C. (2012). Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. *Proceedings of the National Academy of Sciences*, *109*(36), E2382-E2390.
- Irion, D. N., Schaffer, A. L., Famula, T. R., Eggleston, M. L., Hughes, S. S., & Pedersen, N. C. (2003). Analysis of genetic variation in 28 dog breed populations with 100 microsatellite markers. *Journal of Heredity*, *94*(1), 81-87.
- Rimbault, M., & Ostrander, E. A. (2012). So many doggone traits: mapping genetics of multiple phenotypes in the domestic dog. *Human molecular genetics*, *21*(R1), R52-R57.
- Turcsán, B., Kubinyi, E., & Miklósi, Á. (2011). Trainability and boldness traits differ between dog breed clusters based on conventional breed categories and genetic relatedness. *Applied Animal Behaviour Science*, *132*(1-2), 61-70.

- Larson, G., Karlsson, E. K., Perri, A., Webster, M. T., Ho, S. Y., Peters, J., ... & Lindblad-Toh, K. (2012). Rethinking dog domestication by integrating genetics, archeology, and biogeography. *Proceedings of the National Academy of Sciences*, *109*(23), 8878-8883.
- Trut, L., & Dugatkin, L. A. (2017). How to build a dog. *Scientific American*, *316*(5), 68-73.
- Udell, M. A., Dorey, N. R., & Wynne, C. D. (2010). What did domestication do to dogs? A new account of dogs' sensitivity to human actions. *Biological reviews*, *85*(2), 327-345.
- Verdugo, M. P., Mullin, V. E., Scheu, A., Mattiangeli, V., Daly, K. G., Maisano Delser, P., ... & Bradley, D. G. (2019). Ancient cattle genomics, origins, and rapid turnover in the Fertile Crescent. *Science*, *365*(6449), 173-176.
- Clark, G. (2014). *The son also rises*. Princeton University Press.
- Clark, G. (2008). *A farewell to alms*. Princeton University Press.
- Clark, G., Leigh, A., & Pottenger, M. (2020). Frontiers of mobility: Was Australia 1870–2017 a more socially mobile society than England?. *Explorations in Economic History*, *76*, 101327.
- Clark, G. (2016). Microbes and markets: was the Black Death an economic revolution?. *Journal of Demographic Economics*, *82*(2), 139-165.
- Hao, Y. (2013). *Social Mobility under Three Regimes, China, 1645–2012*. University of California, Davis.
- Clark, G. (2012). What is the true rate of social mobility in Sweden? A surname analysis, 1700-2012. *Manuscript, Univ. California, Davis*.
- Plomin, R. (2019). *Blueprint, with a new afterword: How DNA makes us who we are*. Mit Press.
- Plomin, R., & Von Stumm, S. (2018). The new genetics of intelligence. *Nature Reviews Genetics*, *19*(3), 148-159.
- Plomin, R., & Bergeman, C. S. (1991). The nature of nurture: Genetic influence on “environmental” measures. *Behavioral and brain sciences*, *14*(3), 373-386.
- Smith-Woolley, E., Ayorech, Z., Dale, P. S., von Stumm, S., & Plomin, R. (2018). The genetics of university success. *Scientific Reports*, *8*(1), 1-9.