# **VisGen:** Visualising your own genetic data and its influence on traits and diseases

## Motivation

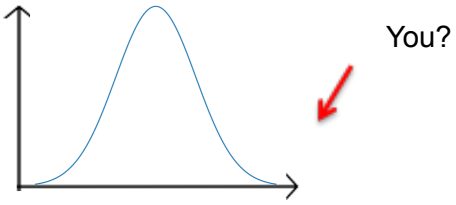
Genetic information can offer health benefits for predicting, diagnosing and preventing diseases. Recent research has shown that 1000s of common genetic variants with small effects on specific diseases can be combined as polygenic scores (PGSs). For common disorders, some studies find a 3- to 5-fold increased disease risk for patients with high disease-specific PGSs [1]. Thus, PGS are growing in importance in clinical genetic diagnosis and as overall health information for interested citizens. The KardioKompassi App is a great example of using PGS to improve cardiovascular health behavior in a clinical setting for Finnish citizens [3]. Direct-to-consumer companies, however, only offer information on single genetic markers and tools that calculate individual PGS in a user-friendly and secure environment are rare and do not often use all genetic markers [2] limiting predictive performance. PGS are typically computed from genotype data that can be generated at relatively low cost (ca. 25-100 EUR, depending on commercial/nonprofit provider e.g. 23andMe, ancestryDNA).

## Goal

The goal of this Master project is to build an application that allows users to upload their raw genetic data from genotype arrays (in the form of vcf files) to a web portal which will then assess the user's ancestry and compute PGS for a number of traits and diseases. The user's PGS will be visualized by comparing it to PGSs of up to 500,000 individuals from the UK biobank. Scientific questions will include assessing the predictive value of PGS for diseases such as cardiovascular disease (e.g. heart attack) risk, together with commonly used clinical risk factors (e.g. smoking, LDL cholesterol, BMI).

Few genetic risk markers
→ Low genetic risk for disease

Many genetic risk markers
→ High genetic risk for disease




**PGS of UK biobank participants**

Counting genetic risk markers gives you a single number, the PGS. User-specific PGS will be computed for a number of traits and diseases and compared with the PGS of UK biobank participants.

## Tasks

1. Write a pipeline that infers genetic ancestry of user's vcf file (projecting it with individuals with known genetic ancestry from the 1000 genomes project)
2. Write a pipeline that computes user PGS for different traits and diseases (using existing PGS weights files or generate PGS weights files)
3. Visualize polygenic risk scores of user genotype data and UK biobank data, in conjunction with clinical risk factors
4. Establish web portal where data can be securely uploaded and analysed together with UKB data on the DHC server

## Project Summary

- Develop an application where an individual user can upload their raw genetic data to receive their personal PGS for several traits and diseases, visualized in comparison with PGSs of 500,000 individuals from the UK biobank
- Assess the predictive value of PGS for disease; for cardiovascular disease we will additionally include commonly used clinical risk factors.

## What you will learn

- Basic principles on how genetics influence traits
- Basic principles on polygenic scores including analysis tools
- Basic analysis of personal genotyping data, and relevant data formats
- Domain-specific tools and software, such as plink or PRS-CS
- Building a user-specific application for highly sensitive personal data
- Reproducible pipeline development

## About You

Depending on the tasks you are most interested in, having a software engineering background can be helpful for pipeline development or building the application. Having prior experience with genetics/genomics data e.g. a bioinformatics background, helps with processing genetic data. Having experience with visualization and design is a great plus for design-related tasks. So you should be able to work with a multidisciplinary team and of course, skills can also be learned on the job. This is what we expect from the group:

- Excited about designing applications for user-facing genomic analysis
- Experience with back-end programming and in setting up analysis pipelines
- Interest in working with data visualization and potentially basic front-end experience

## Team


**Dr. Henrike Heyne**

Senior Research Scientist Digital Health and Personalized Medicine
Hasso Plattner Institute Digital Health Center
Campus III, Rudolf-Breitscheid-Str. 187, 14482 Potsdam
Phone: +49 331 5509 4845
Email: henrike.heyne@hpi.de



**Julian Wanner, M. Sc.**

PhD candidate
Hasso Plattner Institute Digital Health Center
Campus III, Rudolf-Breitscheid-Str. 187, 14482 Potsdam
Phone: +49 331 5509 4817
Email: julian.wanner@hpi.de



**Prof. Dr. Erwin Böttinger**

Head of the Digital Health Center
Professor for Digital Health - Personalized Medicine
Campus III, Rudolf-Breitscheid-Str. 187, 14482 Potsdam
Phone: +49-(0)331 5509-164
Email: erwin.boettinger@hpi.de


## References:


1. Khera AV, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet 2018;50:1219-24.
2. Folkersen L, Pain O, Ingason A, et al. Impute.me: An Open-Source, Non-profit Tool for Using Data From Direct-to-Consumer Genetic Testing to Calculate and Interpret Polygenic Risk Scores. Front Genet. 2020 Jun 30;11:578.
3. Widen E, Junna N, Ruotsalainen S, et al. Communicating polygenic and non-genetic risk for atherosclerotic cardiovascular disease - An observational follow-up study. medRxiv 2020:2020.09.18.20197137.