Data Science in Human Genomics & Disease Risk Prediction

Natalia Ruzickova*

Fyodor Kondrashov Lab
Institute of Science and Technology Austria

*natalia.ruzickova@ist.ac.at

What is hidden in our genes?

Most human traits, including susceptibility to disease, are shaped both by genes and environment. For example, 60 – 80% of differences in human height and 40 – 60% in BMI are due to genes [1], hypertension and breast cancer are 40 – 60% and 35% heritable [2], [3]. Therefore, **identifying the genetic components contributing to disease susceptibility is key to effective prevention**.

Many of the common disease are highly polygenic – they are not influenced by one gene, but by a **set of genetic variants (typically from hundreds to tens of thousands) scattered all over the genome**. Effects of these variants combine and together determine the genetic susceptibility. For example, for coronary artery disease (CAD), a single mutation present in 0.4% of population elevates the risk in developing the disease 3-fold. However, when taking into account also other variants of smaller effects, the prediction accuracy increases 20 times.

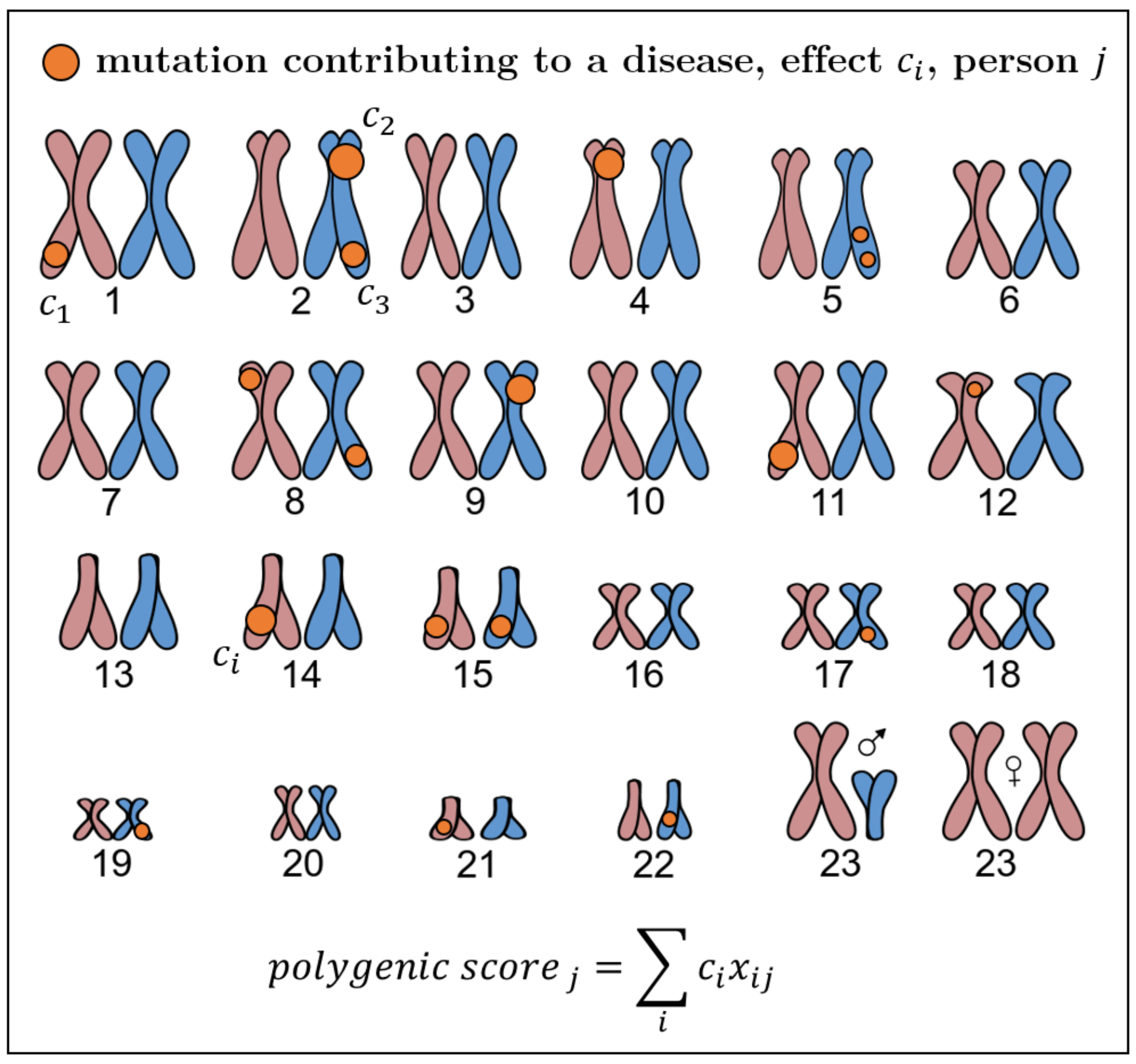
To analyse polygenic disease, the polygenic score (PS, Fig.1) is used. **Effects of mutations are unknown and must be inferred from data on many individuals.** PS is then used as a linear predictor for predicting the risk of each individual. The challenge of to date genomics is to use large sample sizes to construct accurate polygenic scores and thus accurate disease risk prediction.

Source of large datasets: UK Biobank

UK Biobank [4] is a biorepository of large amounts of human genetic and other biological (phenotypic) data. It provides the data to researchers and allows for extensive studies on human genetics and effects of environment on human health.

UK Biobank in numbers:

- 500k participants and their genotypes (50k full genome sequences, rest genotyped and imputed)
- >3k sets of phenotypic data (blood and urine samples, health records, lifestyle and family history questionnaire, cognitive testing)
- 93 million genetic variants (polymorphisms)



$$\textit{polygenic score}_j = \sum_i c_i x_{ij}$$

Fig.1: Polygenic disease. Mutations affecting the disease are spread all over the genome. Their effects combine and determine one's genetic risk of developing the disease. Polygenic score is the sum of all effects (c) present in an individual j ($x_{ij} \in 0, 1, 2$ depending on how many copies an individual carries).

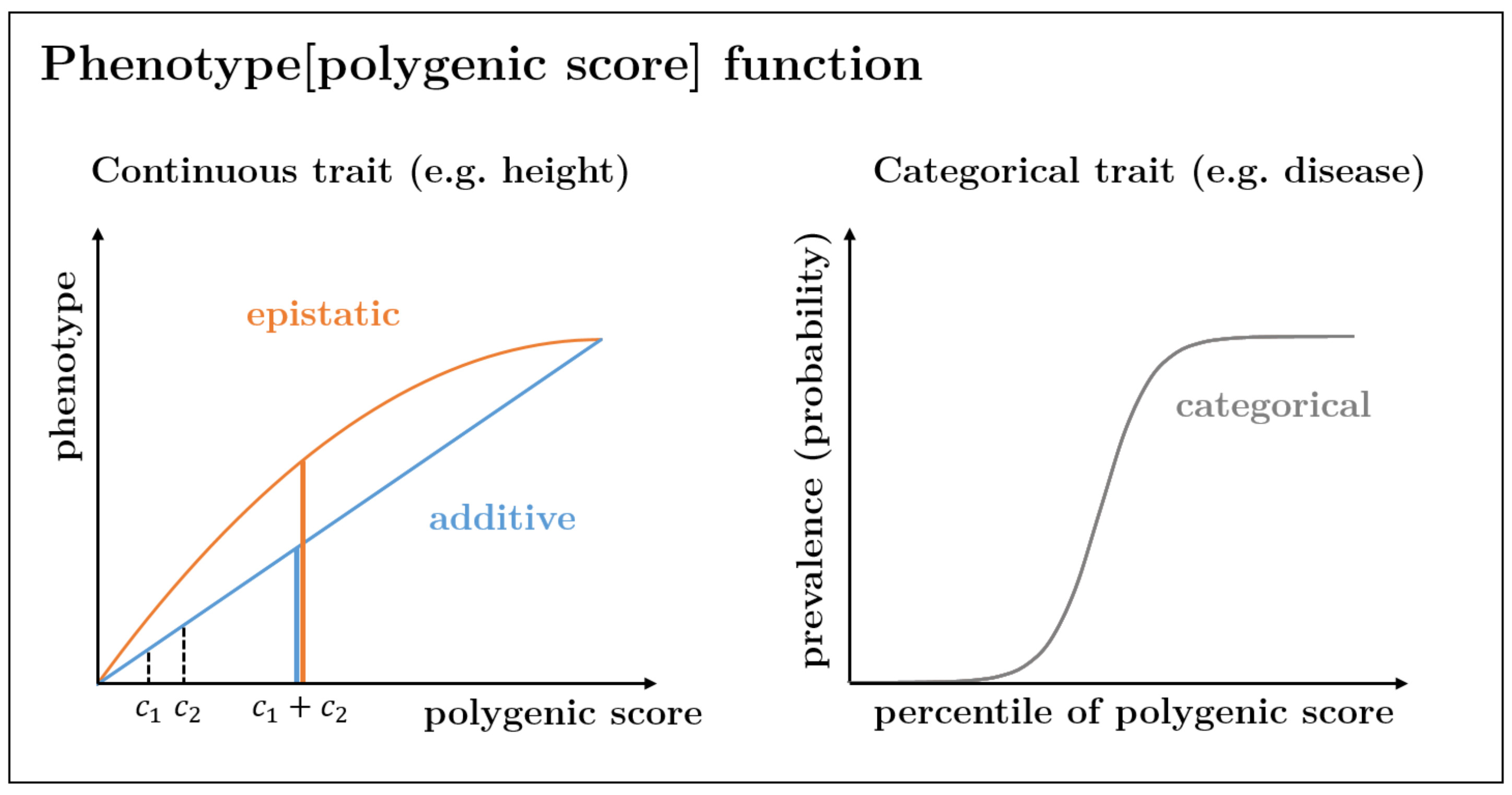


Fig.2: Dependence of phenotype on polygenic score. Phenotype (response variable) is generally proportional to the polygenic score (linear predictor). The particular form of dependence may reveal nonlinear gene-gene interactions and the underlying biological mechanisms. For disease, when mutations accumulate over a certain treshold, the disease is much more likely to develop.

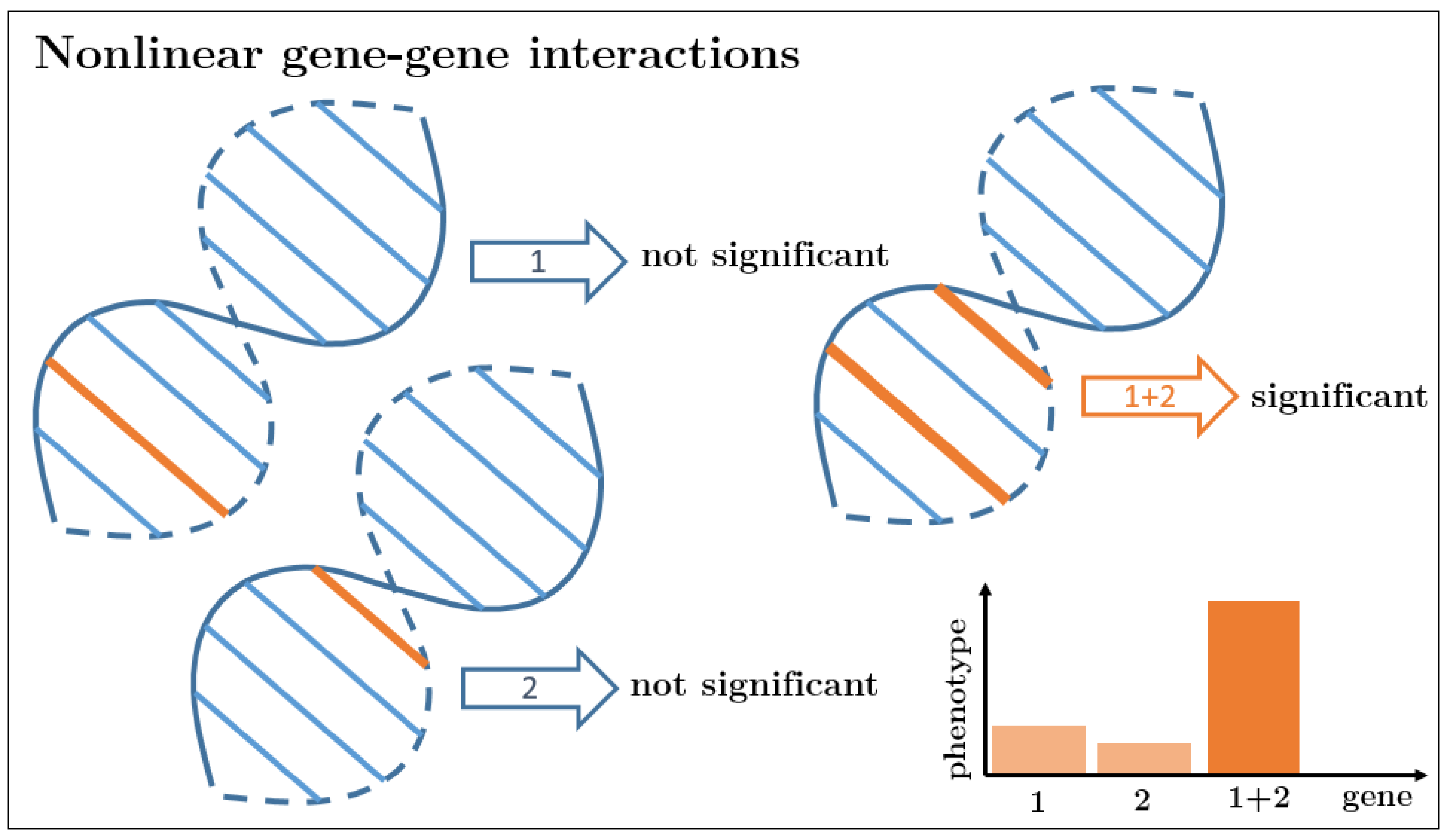


Fig.3: Epistasis on a microscopic level. Mutations may interact nonlinearly, e.g. enhance each others' effects. Such mutations may not be discovered by linear models, even though they do affect the disease.

Data Analysis techniques used for disease risk prediction

The usual procedure of predicting phenotype and calculating polygenic scores is as follows:

1. **identify *individual* mutations** associated with the disease (GWAS: chi-square test for difference in case and control mutation frequency)
2. **construct the polygenic score** (linear predictor): fit the additive effects of individual mutations using a training dataset (linear or logistic regression + regularisation [1], [2])
3. choose the polygenic score with best predictive power using a validation dataset
4. **predict disease prevalence** in a test population using the generalised linear model

Because of the large number of variants and genomes, these methods can only identify variants that are individually significant and assume the effects act independently, thus **neglecting nonlinear gene-gene interactions - epistasis.**

Possible advantages of a Deep Learning approach

Nonlinear gene-gene interactions (epistasis) and their importance is an open problem in biology. Conventional methods neglect epistasis. Moreover, accounting for epistasis would improve predictions of susceptibility to a disease in patients. A deep learning approach can reveal epistasis for the following reasons:

1. fits not only the polygenic score but also the phenotype[PS] function (Fig.2) - may predict more accurately than linear or logistic model
2. determines the polygenic score and the phenotype[PS] function *simultaneously*
3. reveals mutations that enhance each other and are significant only together (Fig.3)

This approach is, however, computationally very demanding (92 million variants x 500k people). Methods need to be found to make it tractable. I am happy to hear your ideas!

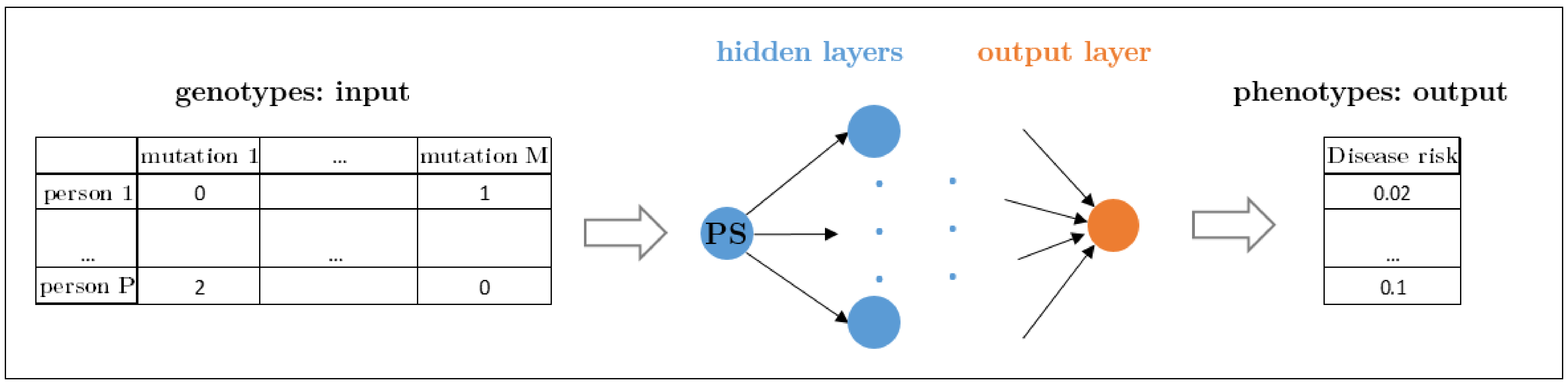
References

[1] L. Lello et al.:. Accurate genomic prediction of human height. *Genetics*, 2018.

[2] A. V. Khera et al.:. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics Letters*, 2018.

[3] Tian Ge et al.:. Phenome-wide heritability analysis of the uk biobank. *PLOS Genetics*, 2017.

[4] The uk biobank, www.ukbiobank.ac.uk.



genotypes: input

	mutation 1	...	mutation M
person 1	0		1
...		...	
person P	2		0

phenotypes: output

Disease risk
0.02
...
0.1

Fig.4: Genomic data and NN architecture. The architecture used to predict disease risk of a person from their genotype. The input is the list of mutations for each person - their genotype. The first hidden layer is a single neuron, its output is the polygenic score (linear predictor) which carries information about effects of individual mutations. The output is the genetic disease risk.