Module 3: Polygenic Risk Scoring

A polygenic risk score (PRS) is a comprehensive method used to estimate the relative genetic risks of being affected by a certain condition or inheriting a certain trait. In a series of articles, we will cover the basics of polygenic risk scoring, as a core methodology we, at BioCertica, use to generate results for our users. Therefore, we will talk about the following:

- 3.1. What is PRS and how does it work?
- 3.2. What is a genome-wide association study (GWAS) & how does it work?
- 3.3. How do we obtain PRS results?
- 3.4 What do PRS results mean?
- 3.5 Factors that affect PRS estimation (Summary)
- 3.6. Application of PRS in clinical medicine (Summary)
- 3.7 Key challenges for PRS in the future (Summary)

Topic 3.1: What is Polygenic Risk Scoring (PRS) and how does it work?

Genetic testing at BioCertica

DNA testing is a vital and versatile tool nowadays, as its discovery has revolutionized our lives in many ways. Not only is DNA a blueprint of life, but it also provides a broad spectrum of answers essential to improving various aspects of our lives. Therefore, by performing DNA tests, we aim to better understand our genetic makeup and how it may affect our lives. We at BioCertica use scientifically proven procedures and protocols for performing genotyping, and the quality and accuracy of obtained genotype data are at the highest level. All the traits we include in our

reports adhere to the strict quality protocols they have to pass, ensuring that the information we provide is accurate and up to date according to the latest scientific research.

We have implemented the process of developing polygenic risk scores for all our traits and conditions. Calculation of polygenic risk scores is based on the latest methodology for estimating genetic risk for specific diseases or disorders, taking into account hundreds and in some cases thousands of SNPs. Therefore, this enables us to provide even more accurate results to our users and makes us unique in the local market and globally. To our knowledge, only a few international genetic testing companies have implemented this methodology so far.

What is PRS?

As you'll remember from module 2, certain conditions and traits are controlled by a single gene mutation, and they are called **monogenic**. For example, a single mutation in the beta-globin chain of hemoglobin causes sickle cell anemia or a mutation on the *CFTR* gene that causes cystic fibrosis.

However, the whole story is not that simple when it comes to complex conditions like diabetes, heart disease, cancer, or traits like height, ability to match musical pitch, etc. Here we talk about conditions and traits that are determined by the interactions of thousands of genes - therefore being called **polygenic**. For a better explanation of the difference between single-gene diseases and complex diseases, the illustrations below show the human chromosomes.

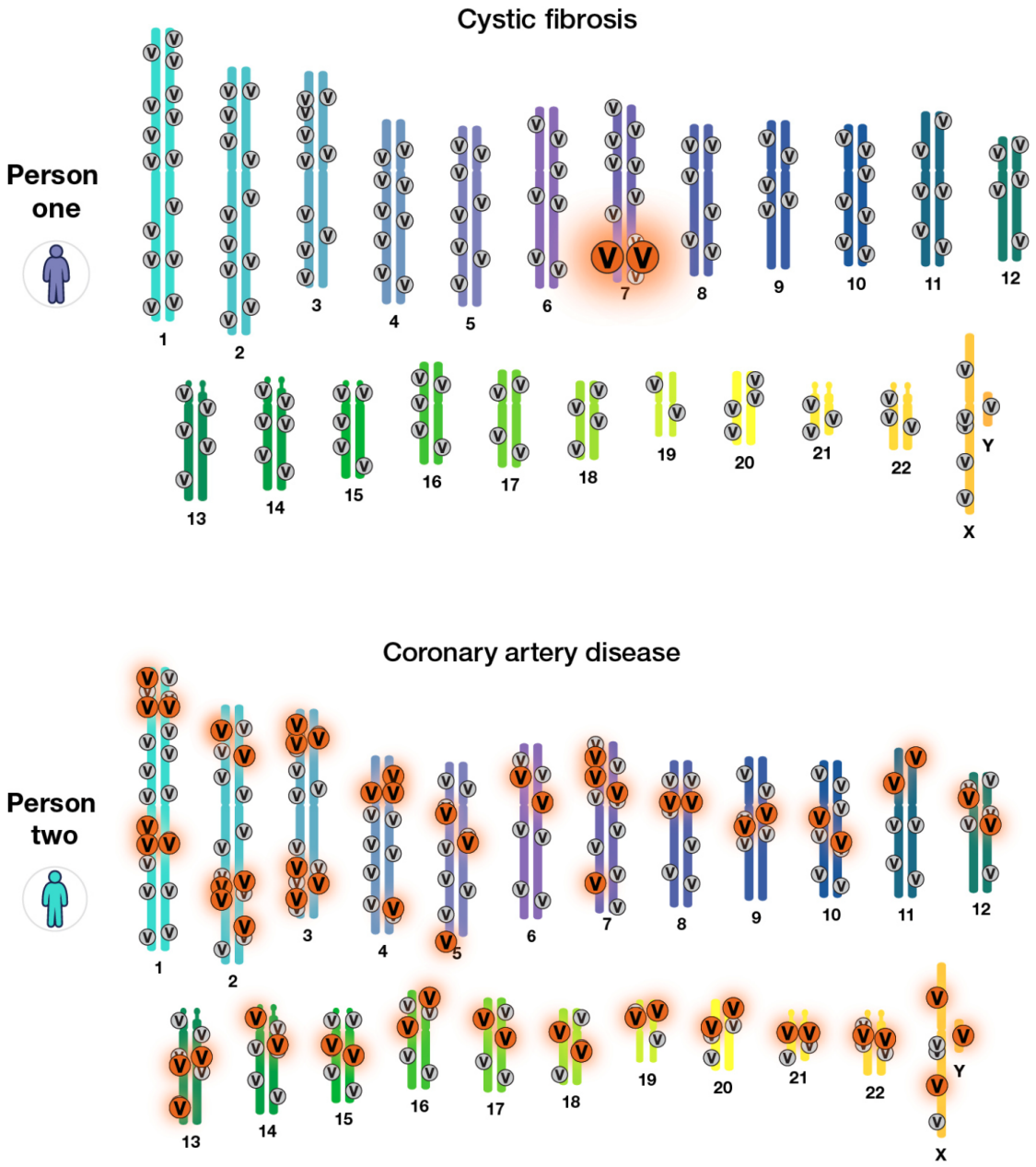

Figure 1: Example of monogenic and polygenic condition

In the case of cystic fibrosis, illustrated on the top, there is only one gene on chromosome 7 that when mutated causes cystic fibrosis. This makes cystic fibrosis a single-gene disease. On other hand, the second illustration shows all yet known variants that are highly associated with coronary artery disease which makes it a

complex disease controlled by many genetic variants which all need to be accounted for when a genetic risk is estimated.

The majority of genetically determined conditions have a complex or polygenic nature compared to the monogenic ones. Since complex traits and conditions involve hundreds of thousands of genes that control their onset, we need complex statistical machinery to estimate the risk they confer - **polygenic risk scoring**.

Polygenic risk score, also known as genetic risk score, polygenic score, or genome-wide score, is a type of genetic testing that gives an estimated relative risk score for how likely you are to develop a certain disease or trait. In other words, a PRS is a single number estimate that can tell you how genetically predisposed you are to a disease or trait.

The process of obtaining PRS for a given trait is a complex process consisting of several steps:

- Quality controls
- PRS building
- PRS calculation
- PRS testing and validation

In the next few articles, we will cover each of these in detail. For now, let's provide a short breakdown of these steps we follow to obtain PRS (Figure 2).

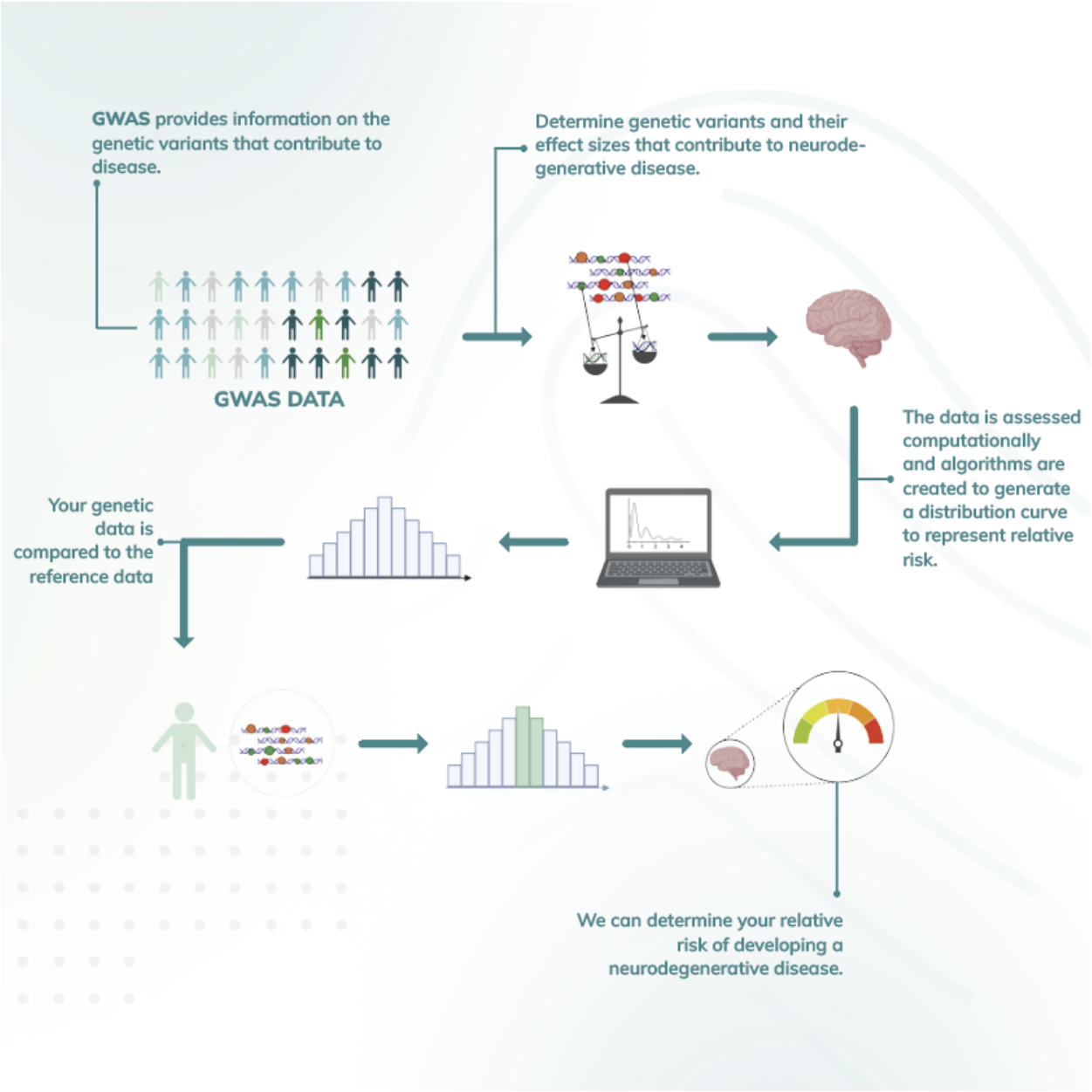

Figure 2. A brief overview of PRS determination of neurodegenerative disorder

1. We analyze thousands and millions of SNPs (genetic variants) across your genome. Each variant has its effect size corresponding to its impact on developing a trait or condition. We determine these in one of two ways:
    a. We select SNPs with already estimated and validated effect weights from polygenic scores (PGS) catalog.

    b. If the above is not feasible, we select SNPs from genome-wide association studies (GWAS) catalog and estimate our own weights.
2. We use appropriate genetic models to add up all effect size weights and get a final polygenic risk score for a given trait or condition.

The next article will explain GWAS and how data from GWAS is obtained and analyzed. Subsequently, we will move to a detailed explanation of all steps in PRS methodology.

References

1. Choi, S. W., Mak, T. S. H., & O'Reilly, P. F. (2020). Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols*, *15*(9), 2759-2772.
2. Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, *20*(8), 467-484.

Topic 3.2: What is GWAS & how does it work?

To calculate a PRS for a particular condition or disease, we must sum up individual risk estimates for all SNPs being associated with a given trait across your genome. But wait, how do we know what variants we should consider for any condition?

The answer lies in Genome-Wide Association Studies (GWAS) that serve as a reference. GWAS studies include millions of people from all over the world belonging to a specific ethnicity or population whose variants are tested for association with a given trait [1].

What is GWAS and why is it important for PRS?

In simple terms, GWAS studies use samples from many different people to identify and report genetic variants that are associated with the onset of a certain condition (also known as cases) compared to those individuals who lack them (also called controls). In other words, these studies identify and report on how alleles for genetic variants that are associated with a given disorder discriminate between affected and not affected individuals.

In this way, GWAS data provides information on which genetic variants (SNPs) we should pay attention to assess genetic risk for a condition of interest. By identifying these SNPs we can determine the underlying mechanisms that cause the trait and may even help us predict how much of the trait is caused by genetics and how much is caused by environmental factors. This clarifies how these alleles differ and are more common in people with particular traits and disorders. In that way, we know which variants we should look for in DNA samples submitted by our users.

What does everything have to do with PRS? In the context of the infinitesimal model (or polygenic model), we look for genetic variants, in this case SNPs, that have been found to explain the genetic variation of a given trait. This is generally done by GWAS. PRS model analyzes the effect size of each allele as obtained from GWAS and used to evaluate the SNP risk scores which are later summed for the final risk score value. It's important to note that the greater the effect size, the greater the weight given to the variant. Seems like too much information? Let's explain the procedure in more detail below!

How does it work?

As far as the sequence of nucleotides goes, humans share 99.5% of their genetic information [2]. It means that, for example, at a particular place the majority of people have a nucleotide A, while the remaining have nucleotide T at the same spot.

These forms are called **variants**. Since this location in human DNA can have multiple forms, it is called single nucleotide polymorphism (SNP). Therefore, the similarity of 99.5 % of 0.5% difference between two individuals does not refer to genes, but base pairs.

SNPs are the key point in understanding the genetic causes of human traits and conditions. Certain traits like aptitude for music or languages are environmental, while traits like eye-color are extremely heritable. Where SNPs help us is to determine and understand to what extent certain traits are due to genetics, or what biological mechanisms may be affecting this trait. To do this, we have to carry out **association analysis**.

Let's suppose we want to find genetic association with body mass index (BMI), or to find which SNPs (genetic variants) contribute to a person's BMI, such as genes that may increase or decrease metabolic activity of our body.

First, we need a large sample size of volunteers, preferably thousands of people, those with the same ethnicities are used to minimize the confounding effect of other factors on genetic variation. Other genetic models such as principal component analysis (PCA) can be used to account for differences in ethnicities and population structure. The next step is to have each participant genotyped, which means to have their nucleotides recorded at many known SNP locations. This results in obtaining information on millions of SNPs for each participant.

Next step is to record BMI for all participants. Once we have recorded genotype and phenotype (trait) data for a large number of people, we can proceed with computing the association between these two. This is done by means of a genome-wide association analysis program and a commonly used software called **PLINK**. This

software makes it possible to perform quality control filters on genetic datased, removing all individuals or SNPs that may not fulfill QC criteria

Afterwards, we perform a regression analysis for every SNP in the dataset with each individual being a data point. Let's say that we want to perform a regression analysis for SNP ID #1, which has alleles A and T (Figure 3). Each individual in the dataset has a number of T alleles for that SNP plotted against the trait of interest, in our case BMI. Person's DNA contains a copy from a father and a copy from a mother, meaning that person's combination for the SNP may be either AA, AT, or TT, which can be coded as 0, 1, and 2 respectively. Once each individual is plotted on the graph, the program tries to draw a line that estimates the relationship between the number of alleles and phenotype.

Simply put, if there is no association between SNP and BMI, regression analysis will simply plot a horizontal line. However, if there is an association between the two, the line will have a slope. Effectiveness of the regression line in predicting the data points determines the p-value. A p-**value** is a statistical term, measuring a likelihood that the association between two variables was due to random chance, given that there is no association between the SNP and BMI. It means that the more data points clustered together around a sloped regression line, the less likely the association between variables is due to random chance, producing a small p-value. For each SNP we record the p-value and the slope of the regression line, which is also known as the effect size.

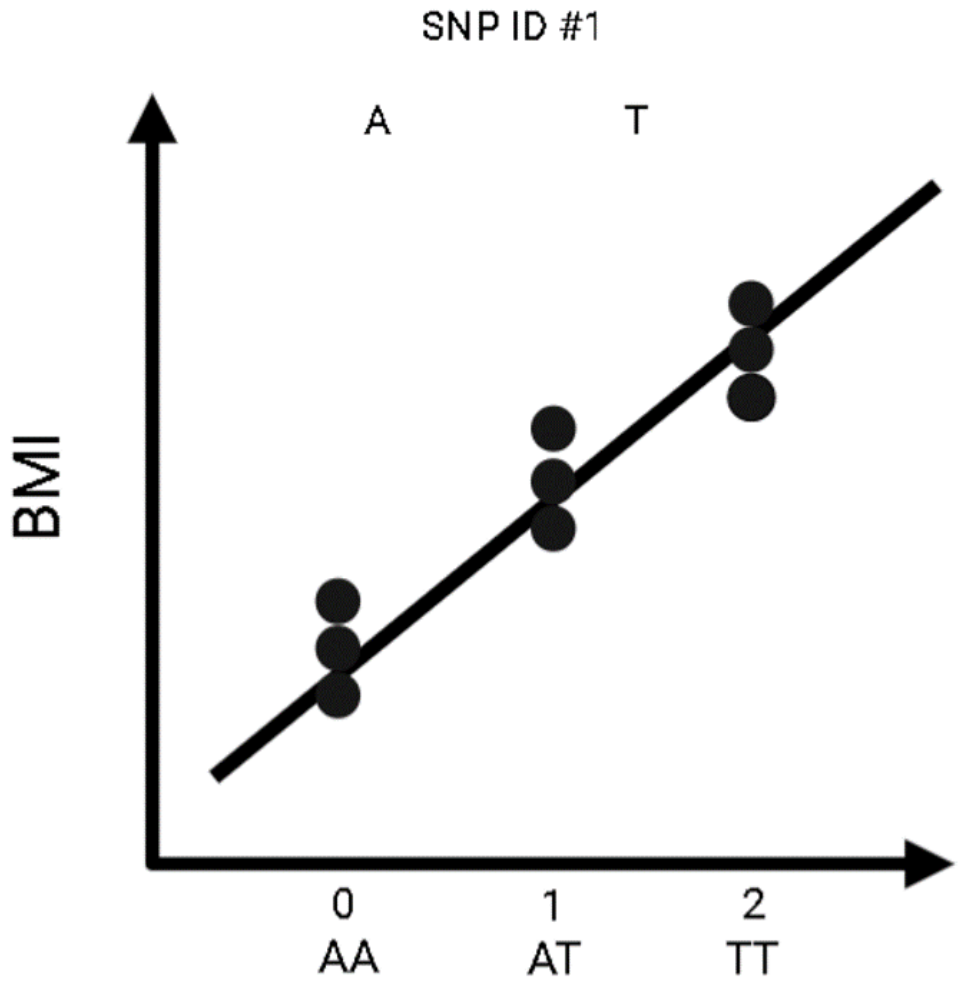

Figure 3: Association between SNP and BMI

This regression analysis is repeated for every single SNP in the dataset. If we work with millions of SNPs, it would take hours and days for the computer to execute results. However, programs like PLINK provide efficient processing such as multi-threading to finish the analysis faster.

Moreover, these programs also allow us to include **covariates,** which are other factors that may affect the phenotype or trait of interest. For example, BMI may be

hugely affected by the amount of exercise done weekly by a person, and having this information available for people in a dataset may significantly influence the slope of the regression line compared to considering genotype alone. Additionally, PCA is also used here to account for population substructure.

Finally, after calculating the p-values for all SNPs available, we can say if there is an association between SNP and trait, or if the association is significant. It has again to work with p-values, as the only values below 0.05 are considered to be statistically significant, which means that real association is present and not due to random chance.

However, when working with millions of SNPs, there is a possibility of producing thousands of false positives. This can be fixed by doing a Bonferroni correction which transforms the threshold required for achieving significance by taking the typical threshold of 0.05 and dividing it by the number of SNPs in the analysis. Quantitative genetics has adopted the value of 5E-8 as the default threshold for significance. The goal of applying Bonferroni correction is to reduce the number of false positives.

For visualization of results, a Manhattan plot is produced, where each SNP with its corresponding chromosome position is plotted on x-axis versus corresponding negative log of p-value on y-axis. Dots above the line indicating threshold represent SNPs that are significantly associated with the trait. Next, these SNPs are analyzed using SNP databases that indicate what genes are present for those regions. Identifying these genetic regions may be helpful for understanding biological mechanisms that may be useful in prevention and treatment of certain genetic diseases.

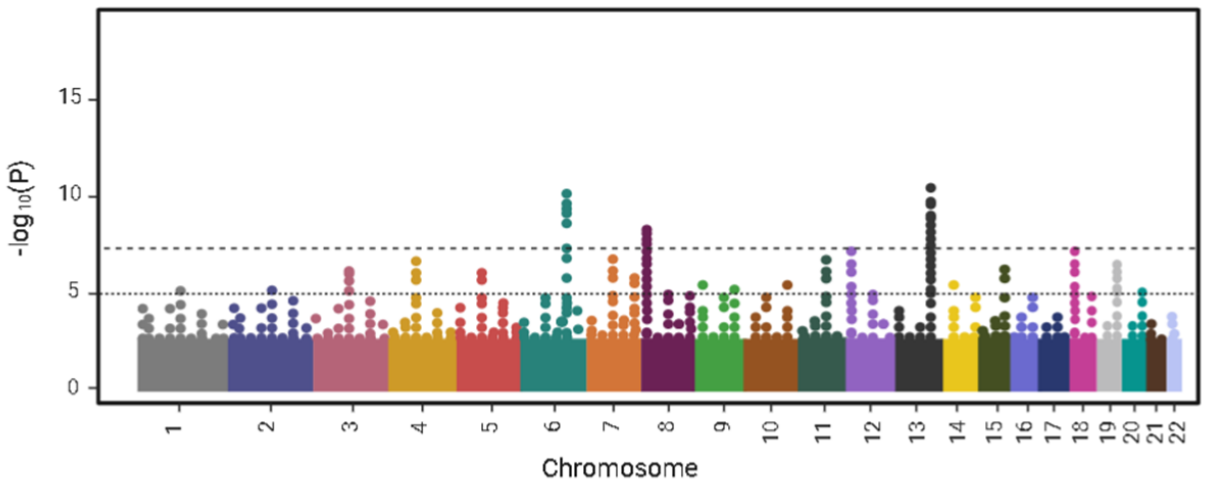

Figure 4: An example of a Manhattan plot

Key benefits and limitations of GWAS

There is no doubt that GWAS have revolutionized the approach and understanding of genetics behind the complex disease in recent decades, finding and reporting many significant associations between gene variants and complex diseases and traits. Some benefits that GWAS brought over the last several years are successful endeavors in the field of understanding complex diseases and genes that increase susceptibility to them, discovering new biological mechanisms underlying complex diseases, and translating them into clinical care.

However, despite all of them, GWAS also have been subjected to many controversies and limiting factors. Let's try to summarize some key benefits and limitations of GWAS.

Benefits:

- GWAS have been very successful in identifying novel variant-trait associations. Thousands of GWAS have been published so far, with thousands of SNPs and associations reported including vast number of diseases and traits including major depressive disorder, anorexia nervosa, cancers and their subtypes, type II diabetes, coronary heart disease, schizophrenia, inflammatory bowel disease, insomnia, BMI etc.
- GWAS can lead to the discovery of novel biological mechanisms. For example, the role of autophagy in Crohn's disease was not known until SNPs associated with this disease were not discovered.
- GWAS findings have multiple benefits in clinical settings, where they help translate biological insights into medical advancements. GWAS may help for disease classification and subtyping. Genetic variants identified by GWAS can be used to identify individuals at high risk for developing a condition, which may provide clues in directing right prevention, treatment or diagnosis.
- GWAS can provide insight into ethnic variation of complex traits., since some risk SNPs and their locations on chromosomes show considerable ethnic differences in frequency and effect size.
- GWAS can be used to identify novel monogenic and oligogenic disease genes.
- Beyond gene identification, GWAS data may be used also for reconstruction of population history, ancestry and population substructure determination, fine-scale estimation of location of birth, estimation of SNPs heritability for complex traits, estimation of genetic correlations between traits, polygenic risk scores, forensic analyses etc.
- GWAS data generation, management, and analysis are straightforward as explained above.
- GWAS and their findings are easily available today and facilitate novel discoveries.

Limitations:

- There are concerns that most associations found in GWAS do not reflect functional variants, but variants in linkage disequilibrium with potential functional variants.
- SNPs used in GWAS account for a rather small fraction of heritability of complex traits. Heritability refers to a proportion of genetic variation due to genetic factors solely.
- Certain associations concluded in GWAS may rather be spurious, not pointing out causal variants and genes.
- Since most GWAS studies have been conducted on the European population, polygenic risk scores are mainly available for people of European ancestry. This affects the accuracy of PRS methodology if applied to other ethnic groups of non-European descent since there could be differences in their genetic variants and other confounding effects. However, the good news is that there are already large-scale GWAS studies that cover individuals from different ancestries and new ones are published daily.

These limitations have led to skepticism and hesitancy among non-geneticists and especially stakeholders and national funding agencies to fund new GWAS. However, there is considerable benefit from GWAS, as their associations have led to insights into the architecture of disease susceptibility, and to advances in clinical care and personalized medicine.

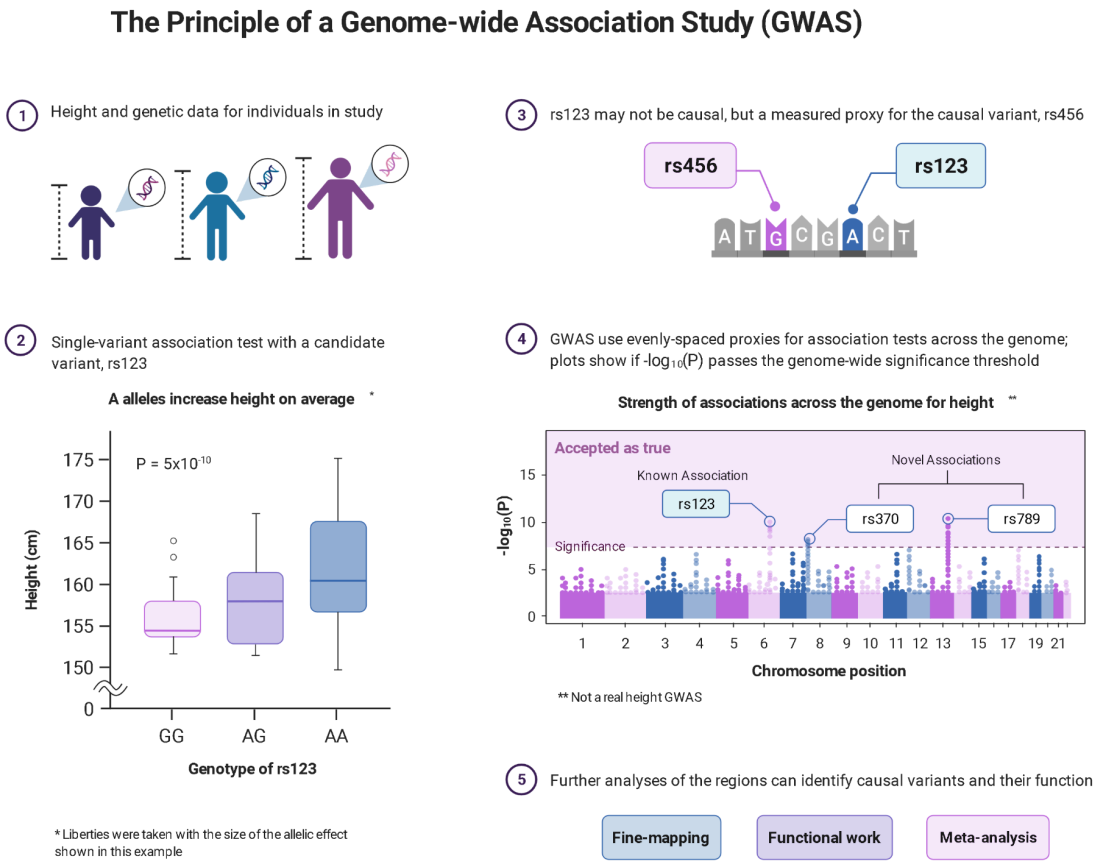

Figure 5: Overview of GWAS

References

1. Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, *20*(8), 467-484.

2. Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., ... & Venter, J. C. (2007). The diploid genome sequence of an individual human. *PLoS biology*, *5*(10), e254.

***Figures generated using BioRender**

Topic 3.3: How do we obtain PRS results? - A simplified Overview

PRS can be summarized into two main steps. First is the building of the PRS score from scratch and obtaining PRS effect weights and the second is to perform the necessary PRS calculation for a given individual.

The process of obtaining PRS results is rather long and complex. Here, we will show you a very simplified explanation of this process but we encourage you to dive deeper into this topic in Module 3: Additional Reading 1.

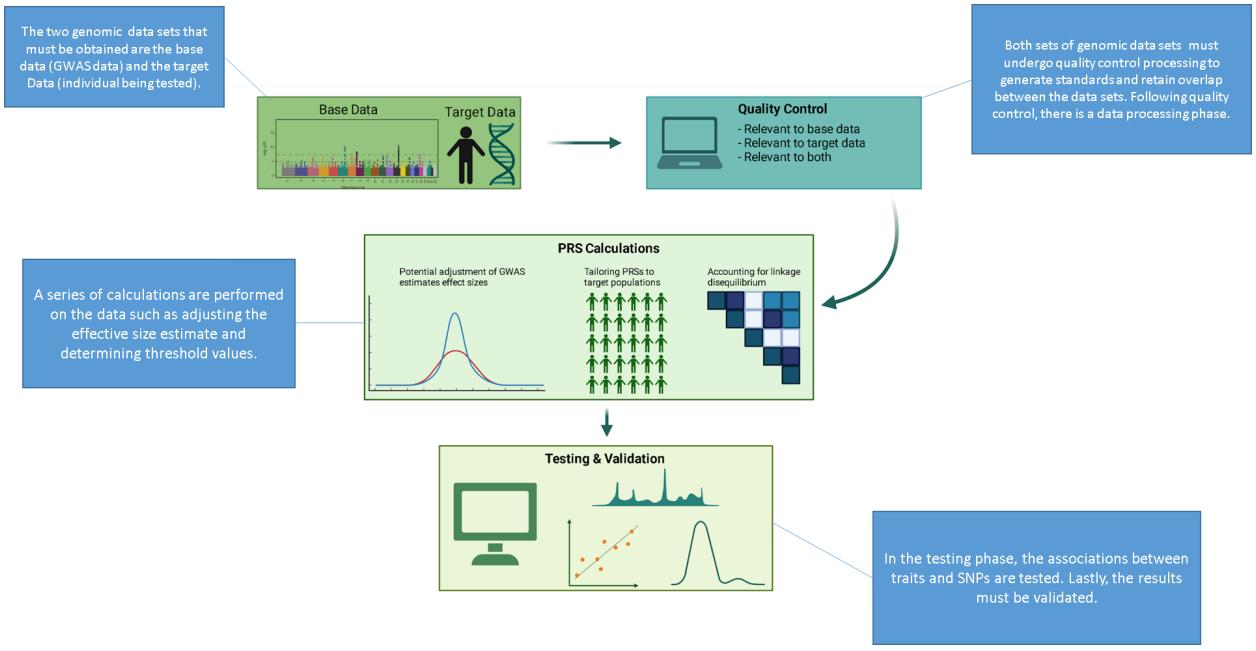

Figure 6: A simplified overview of obtaining PRS results

Topic 3.4: What do PRS results mean?

As a practitioner, understanding the meaning of PRS results is of great importance. When it comes to the PRS approach, this works on the basis that hundreds of thousands of gene variants accumulate to cause the disease phenotype. Therefore, we cannot practically present these results on a single gene basis; rather we present them from a statistical point of view and consider their summed weight effect.

Here are a few main differences between PRS and the standard genetic testing approach:

- It is **more comprehensive** as it takes into account all available genetic variants associated with the given disease
- It accounts for each genetic variant based on its effect and it **does not assume the equal effect**
- On the other hand, it is **more complex** and hard to set up
- Also, there is a **bias effect** depending on which population variants are detected to be associated with a given disease.

In other words, a PR score is providing relative genetic risk for a given trait or condition based on the (peer-reviewed) genetic variants that are proven to be associated with a given trait. Therefore, we are not talking about a single gene or set of genes anymore, but everything we know about the genetic mechanisms controlling that condition or trait.

If we look at the distribution of PRS scores across a population, we will see a normal distribution curve. This means that a small portion of the population will have low genetic risk scores, a small portion of the population will have high genetic risk

scores and the majority of the population will have an average genetic risk (Figure 7).

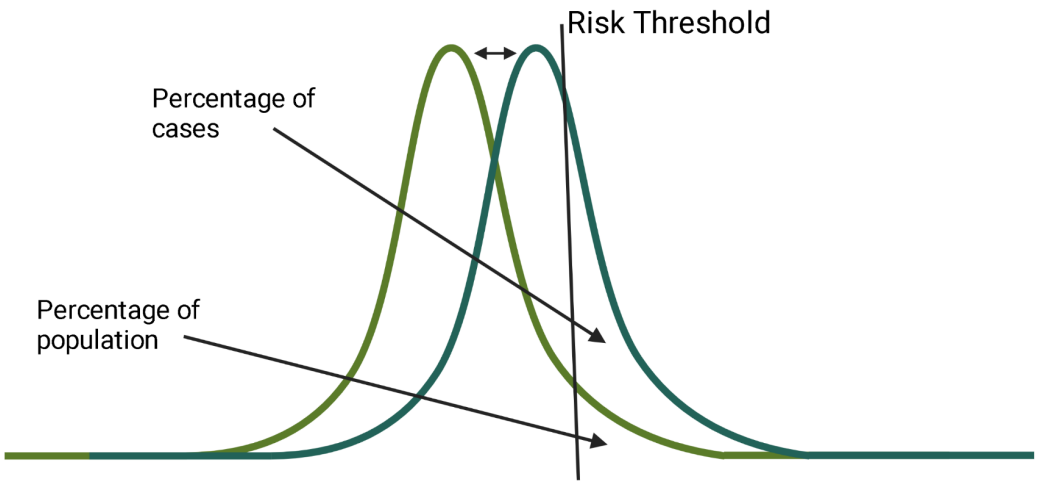

Figure 7: PRS within a population

However, when we look at the scores of a sample of individuals who have the disease of interest, we see a statistically significant shift in the distribution graph. This shift indicates there is a much higher relative risk for those with high PRS scores. We can also look at the risk threshold point where those with PRS greater than this point will be affected by the disease. As you can see in Figure 7 above, there is a greater portion of individuals who fall beyond the risk threshold in the cases with the disease [2, 3].

Getting your results from BioCeritca

In one of our articles, we have written about how we generate genomics reports at BioCertica. Here you can learn more about how we presented results before the introduction of PRS. Now we present PRS results on a scale and contextualize them by comparing them with the results of the general population. What it turns out is known as an individual's lifetime risk relative to population, and is expressed in percentages.

Genetic risk can fall into one of three categories: decreased, average or increased (Figure 8). However, we must consider how much of an impact genetics has on a given disease or trait, i.e. we are talking about heritability. For example, the genetic contribution to lung cancer is only about 8% [4]. Therefore, having an increased genetic risk for lung cancer does not equate to a high absolute risk of developing the disease as environmental factors are major influences.

However, other conditions have a much greater genetic contribution. The heritability of depression is approximately 40-50% [5], hence a decreased genetic risk of depression does not necessarily mean you have no chance of developing depression as other psychological or physical factors may cause an onset. Nevertheless, an increased genetic risk of depression does significantly increase the absolute risk of onset. Let's consider an example of the PRS results presentation below.

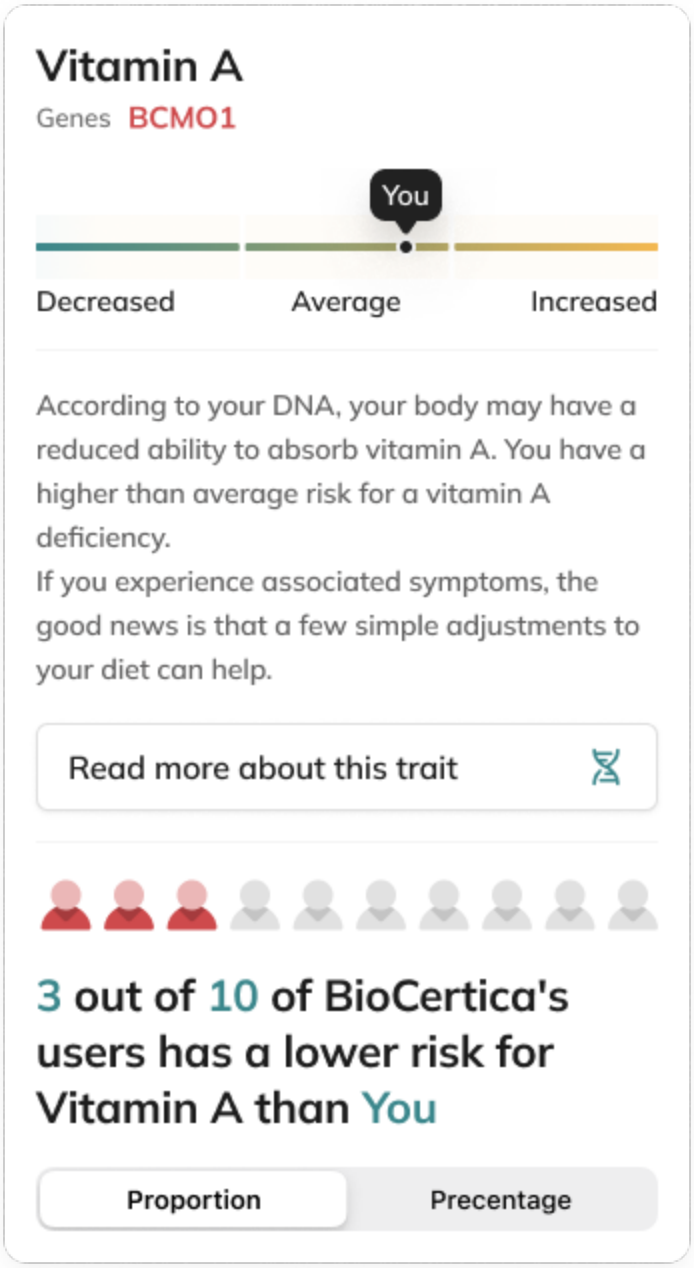

Figure 8: Example of BioCertica PRS result

In the case of the example above, the user has a high genetic risk of being prone to vitamin A deficiency.

Additionally, we also present users' results in the context of the population. We present this information as a percentage of how many users have lower or greater PRS scores.

It's important to consider that PRS results only show relevant risk and not absolute risk as they do not take into account lifestyle factors. For a more compressive look into their risk of developing the diseases and conditions assessed in our PRS scores, users have an option to consult any of our trained practitioners. Practitioners take into account the user's lifestyle and assess if they positively or negatively contribute to their risk of developing the disease.

References

[1] T. D. Thacher, P. R. Fischer, R. J. Singh, J. Roizen, and M. A. Levine, "CYP2R1 mutations impair generation of 25-hydroxyvitamin D and cause an atypical form of vitamin D deficiency," *J. Clin. Endocrinol. Metab.*, vol. 100, no. 7, pp. E1005–E1013, Jul. 2015.

[2] C. M. Lewis and E. Vassos, "Polygenic risk scores: From research tools to clinical instruments," *Genome Med.*, vol. 12, no. 1, pp. 1–11, May 2020.

[3] P. D. P. Pharoah, A. Antoniou, M. Bobrow, R. L. Zimmern, D. F. Easton, and B. A. J. Ponder, "Polygenic susceptibility to breast cancer and implications for prevention," *Nat. Genet. 2002 311*, vol. 31, no. 1, pp. 33–36, Mar. 2002.

[4] M. Kanwal, X. J. Ding, and Y. Cao, "Familial risk for lung cancer," *Oncol. Lett.*, vol. 13, no. 2, p. 535, Feb. 2017.

[5] "Major Depression and Genetics | Genetics of Brain Function | Stanford Medicine." [Online]. Available: https://med.stanford.edu/depressiongenetics/mddandgenes.html. [Accessed: 22-Feb-2022].

Topic 3.5: Factors that affect PRS estimation (Summary)

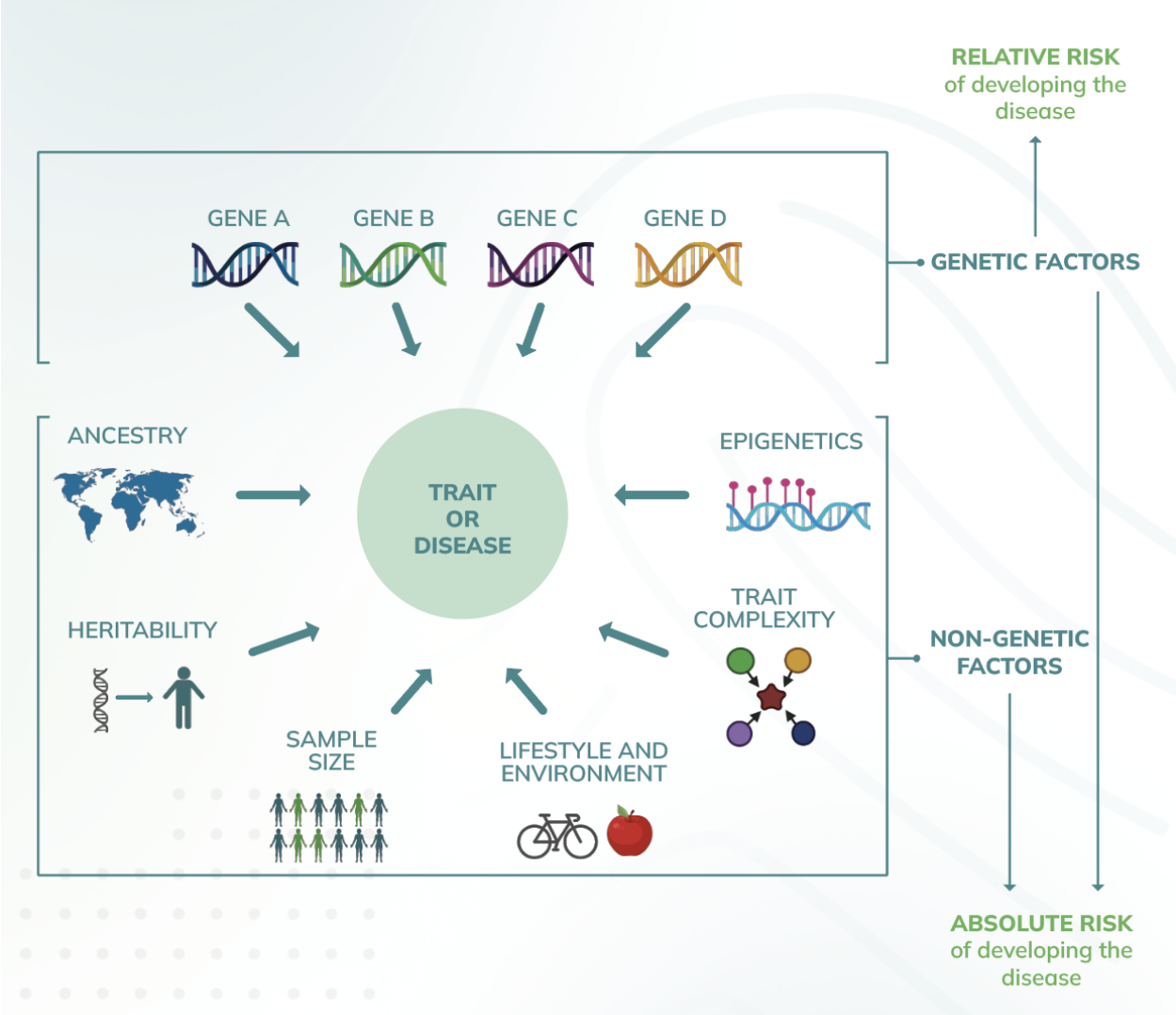

Figure 9: Factors that affect PRS estimation

As a consultant, we do not expect you to understand the details of the factors that affect PRS estimation but we would like you to be aware that genetics alone cannot determine a trait. Genetic factors indicate the relative risk of a trait, one piece of the puzzle. However, the whole picture or absolute risk is determined by genetic factors and non-genetic factors. The important non-genetic factors include:

Heritability:

Heritability is defined as the proportion of total phenotypic variation in a population due to genetic factors. In other words, how much of the trait is due to genetics and how much is due to the environment.

Trait complexity:

How traits are determined is often the result of multiple molecular and cellular pathways. Some traits are more complex than others and therefore the genetics behind them is more extensive.

Sample size:

One of the major limiting factors in PRS is the reference data that it is based on, and the target sample size. Additionally, there is much bias in current GWAS studies as most of these come from people of European ancestry. Therefore this reference data does not show a true reflection of all populations.

Ancestry:

This bias is further reflected in the next factor influencing PRS estimation-ancestry. PRS estimations can differ greatly in admixture individuals. Additionally, this worsens demographic inequalities in access to healthcare, and we see a decreased performance of PRS scores in people of non-European ancestry. However, advancements are being made to rectify this as ancestry bias is being addressed through large-scale, diverse cohort recruitment and sharing of ancestry-specific GWAS summary statistics.

Lifestyle and environmental factors:

Many traits are multifactorial and have an environmental component. In short, all traits will have some lifestyle/environment component but how much lifestyle affects the risk depends on the nature of the disease. Some diseases have a strong lifestyle influence that can be considered preventable despite genetic risk factors.

Despite the influences of these factors, PRS is the best way to go for estimating genetic risk and much better in comparison to single gene test reporting.

Topic 3.6: Applications of PRS in clinical medicine (Summary)

The main advantage of PRS is that it provides individuals with the knowledge of their genetic susceptibility to a condition or a disease. This knowledge empowers an individual to take measures towards lowering the risk of developing a disease or detecting it early before its onset.

Some may argue that PRS testing is not ready to be used in a clinical setting at this point. However, the counterargument is that despite some limitations to the PRS testing, it is still a useful tool in healthcare. Large-scale studies are being conducted to investigate how useful PRS is clinically. PRS testing is becoming increasingly commercially available from direct-to-consumer companies worldwide.

PRS testing has major potential in clinical applications and advancements in medical care and has the potential to be a great supplementary tool to cover the four basic areas of medicine: predictive, preventative, personalized, and participatory. PRS testing can dive individuals towards risk-reducing behaviors such as lifestyle changes, regular screening, and preventive medical intervention. The more available the testing becomes, the greater their impact.

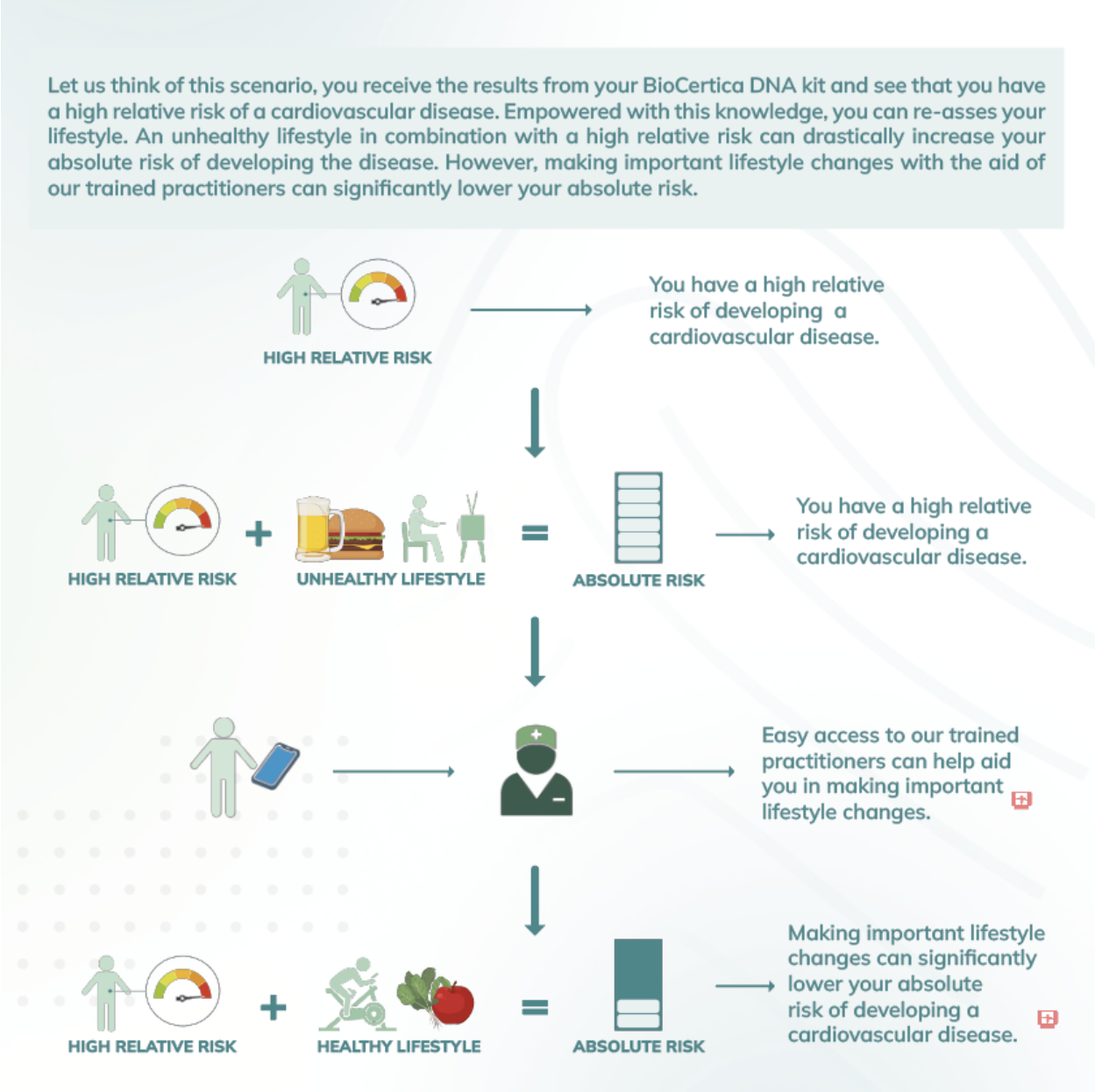

Figure 10: Application of PRS testing through BioCertica.

Topic 3.7: Key challenges for PRS in the future (Summary)

In previous articles, we have discussed some of the limitations of PRS. Fortunately, there is much research and progression towards eliminating these challenges for future PRS.

The 4 main challenges that currently exist in PRS are:

- Diversity bias
- Absolute risk prediction
- Understanding heritability of traits
- Standardizing PRS

Diversity Bias

The challenge: GWAS studies do not reflect the diversity of all populations. About 88% of genetic information in GWAS studies has been obtained from individuals of European descent; an ancestral bias is created.

The solution for the future: Fortunately, studies are already addressing this issue, and large-scale diverse cohort recruitment and sharing of ancestry-specific GWAS summary statistics are underway.

Absolute risk prediction

The challenge: PRS is a prediction tool rooted in genetic information. However, non-genetic factors play a major role in influencing the onset and progression of diseases or conditions.

The solution for the future: to solve this, there needs to be an integration of environmental factors into an overall clinical model for absolute risk. Multiple research teams have developed models with the potential to aid in building an absolute risk prediction.

Understanding heritability of the trait

The challenge: a continuing mystery in genetics has been termed the “missing heritability” problem or the common disease-common variant hypothesis. In short, one would expect genetic variants commonly found in a population would contribute to the commonly found diseases in that population. However, this is not the case.

However, this heritability may not necessarily be missing but rather has not been detected yet.

The solution for the future: we have to find those low-frequency variants. Machine learning has been used to develop functional disease-associated SNPs prediction (FDSP), which statistically analyzes millions of SNPs from GWAS studies and determines predicted positive risk SNPs for further validation experiments.

Standardizing PRS

The challenge: Multiple direct-to-consumer companies provide PRS testing. There are no standards or agreements for clinical reports that include PRSs. Additionally, there is no regulation for the best practice to report PRS results to patients.

The solution for the future: to address these issues, standards must be defined and adopted for PRS testing, and further efforts must be made to standardize data workflows and clinical pathways. Polygenic Risk Score Reporting Standards (PRS-RS) that have been proposed include a range of categories from population studies and data to data transparency and availability.

Fortunately, the current key challenges in PRS are being addressed and we anticipate major improvements in PRS testing. PRS testing is still in its early phases and evolving rapidly to become a supplementary tool for use in clinical practices and diagnostics.

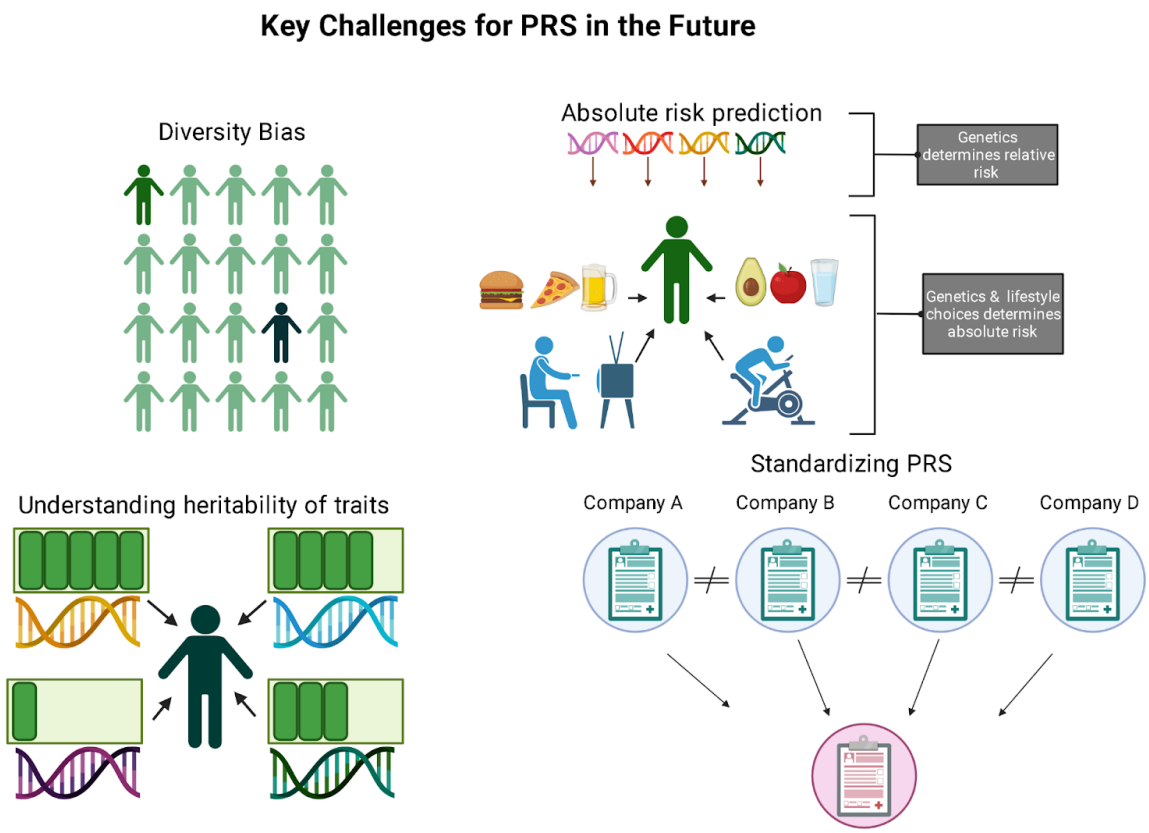

Figure 11: Key Challenges in PRS