Multi-omics in research: epidemiology, methodology, and advanced data analysis

Faquih, T.O.

Chapter 1
General introduction

1 GENERAL INTRODUCTION: THE RISE OF OMICS

The human genome project was a scientific milestone for human biological understanding (1, 2). This achievement started an era of large genetic analyses of an assortment of diseases. Genetic studies, such as genome-wide association studies (GWAS), have since revolutionized our understanding of disease etiology, prognosis, and diagnosis, and have contributed to public health (3, 4). Cardiometabolic diseases, including obesity, cardiovascular disease (CVD), type 2 diabetes (T2D), hypertension, and liver disease such as non-alcoholic fatty liver disease (NAFLD), are prevalent diseases that have benefited from genomic studies. GWAS have resulted in the identification of thousands of single nucleotide polymorphisms (SNPs) associated with these diseases (5). Furthermore, these associated SNPs enabled genetic epidemiological studies that identified causal associations, expanded our understanding of the pathophysiology, and improved the prediction of these diseases. In the wake of rapid technological advancements, it became possible to perform extremely large genomic studies in millions of individuals.

2 PROTEOMICS AND METABOLOMICS

Technological developments have enabled the study of genome-wide gene expression (6), whole genome DNA modifications in various body tissues (7), and the large scale measurement of proteins and metabolites downstream of the genome (8). The large-scale study of biological measures is generally referred to as OMICs. When referring to proteomics and metabolomics, we are referring to modern methods of mass measurements of hundreds to thousands of proteins and metabolites from a single sample. These types of studies usually include a large number of individuals and therefore often require the collaboration of several cohorts.

2.1 Proteomics

Measurement and analysis of proteins has been possible for over 200 years (9). The start of modern proteomics can be dated back to the 1960s and 70s with the advent of two-dimensional gel electrophoresis (10) and the creation of protein databases (11). However, this process was slow, had low throughput, and was not easily reproducible. Advances in mass spectrometry (MS) in the 1990s provided a powerful tool for the identification of proteins that bypassed the limitations of previous methods (10). After the completion of the human genome project, the proteome became a new focus to complement the newly sequenced genome (10, 12).

Proteomics and proteome research aims to achieve several goals. First, to use high throughput technologies to enable the identification and quantification of all human proteins. The number of proteins that can be produced from genes is amplified by alternative RNA splicing and post-translational modifications. Whereas the genome is nearly identical in every cell of the body, the proteome can differ substantially. The variable expression of genes in different cell types as well as environmental influences determine when and in which cells proteins are produced (13). These facts make the proteome more dynamic than the genome. Moreover, it makes it difficult to pinpoint the total number of proteins in humans. Currently, the estimated number of human proteins varies from 10,000 to billions (13). Second, in addition to the quantification of proteins, proteomic research enables studying the functionality of proteins. This includes the association between protein function and disease. Assessing functionality is complicated by protein-protein interactions and protein-DNA interactions (14). Currently, proteomics research has yielded large protein atlases publicly available online, such as the human protein atlas (15). Diseases that have been studied using proteomics include T2D and CVD (16), Alzheimer's disease (17), NAFLD (18), osteoarthritis (19), and venous thrombosis (20), to name a few.

2.2 Metabolomics

Metabolomic research measures small biomolecules in biofluids that are often substrates and products of metabolism (21, 22). Metabolites can thus be consumed and produced by endogenous metabolic processes or acquired from external sources and subsequently modified. Metabolites include amino acids, fatty acids, cholesterols, nucleotides, triglycerides, lipids, lipoproteins, and externally acquired compounds such nicotine and its metabolites from smoking (22, 23). One of the earliest studies on metabolite measurements was reported in the 1940's by Willems et al. (24-26). This work was included in their pivotal publication in 1951 where they coined the concept of "metabolic profiles" (24, 26). In this work, the authors demonstrated the methodology of quantifying and estimating different metabolites from urine and saliva samples using paper chromatography. They further reported their extensive work on different metabolic profiles of alcoholics, schizophrenics, mentally deficient children, overweight and underweight individuals, and the metabolic profile of different diets (24).

In parallel with the aforementioned completion of the human genome project, advances in MS and nuclear magnetic resonance (NMR) technology and the expansion of proteomics in the early 2000s, metabolomics has followed suit and gained momentum. This advancement has been driven by the development of commercial and non-commercial resources for metabolomic measurements that have made this more accessible and affordable for researchers (27). Metabolomics data have been used to provide insight into the pathophysiology of several diseases. Examples include the lipoprotein and metabolic profile that have been associated with the risk of coronary artery disease (28), the metabolomic profile of healthy and unhealthy body weight and their associations with disease outcomes (29), and the metabolomic profile for depression (30) as well as numerous other diseases. Furthermore, metabolomics has been used for discovering disease biomarkers such as neurodegenerative diseases, NAFLD (31), and colorectal cancer (32).

3 A SELECTION OF CURRENT PROTEOMICS AND METABOLOMICS MEASUREMENT PLATFORMS

A myriad of protein detection and quantification technologies have been developed over the last decades. These technologies include enzyme-linked immunosorbent assay (ELISA) and western blotting, two-dimensional gel electrophoresis, gas- and liquid chromatography, MS, NMR, and aptamer-based proteomics. Most metabolomics platforms use NMR or MS for metabolite detection and quantification. In this thesis we have used three techniques: for proteomics we have used the aptamer-based platform SomaScan, for metabolomics we have used the NMR based Nightingale Inc. platform and the ultra-high performance liquid chromatography - tandem mass spectrometry (UHPLC-MS/MS) based platform used by Metabolon Inc.

3.1 SomaScan for Proteomics

New techniques have been developed for high throughput protein measurements in recent years. One such method makes use of nucleotide based "aptamers" to identify and quantify proteins. A leading platform that uses aptamers for proteomic measurements is SomaScan. SomaScan (SomaLogic, Inc. Boulder, CO, USA) is a high throughput proteomics platform capable of simultaneous measurement of over a thousand proteins. Unlike traditional immunoassay instruments, SomaScan utilizes "Systematic Evolution of Ligands by Exponential enrichment" (SELEX), a biochemical technique used to create a library with a wide range of modified synthetic oligonucleotide ligands (i.e., the aptamers) bioengineered to bind to their respective protein targets. These aptamers are designed to emit a fluorescence signal only when they are bound to

their target protein. This fluorescence signature is then used to measure the relative concentrations of the target protein. This method provides several benefits over traditional immunoassay methods; Aptamers are inexpensive to produce, highly modifiable and are chemically stable. SomaLogic has developed a vast library of thousands of unique aptamers to detect proteins from a single biological sample (33). However, this technique is not perfect and some studies have noted that aptamers binding affinity can be affected by cross reactivity issues, genetic variations altering the protein structure, post-translational modifications, and the effects of the complexity and stability of the target protein structure (34). Nevertheless, SomaScan has been used for connecting genetics with the proteome profile of different diseases (35), creating a genomic atlas of the human proteome (36) and predicting coronary heart disease (37, 38) to name a few.

3.2 Metabolomics Measurement Techniques and Platforms

Metabolomic platforms are often divided into two categories: targeted and untargeted (also referred to as non-targeted). The targeted approach focuses on detection of predefined target metabolites (39). A benefit of using targeted metabolomics is its consistency and the possibility of generating absolute metabolite concentrations. Targeted metabolomic platforms, such as the Nightingale platform, currently measure several hundred metabolites (39, 40). Untargeted platforms, on the other hand do not fully target specific metabolites before the measurements. Instead, the platforms aim to detect and quantify as many metabolites as possible and subsequently identify them by cross referencing in large libraries of metabolites. The Metabolon platform is an example of untargeted platform that utilizes UHPLC-MS/MS technologies (41) and is capable of measuring thousands of metabolites including known and unknown metabolites.

3.2.1 Nightingale: Proton Nuclear Magnetic Resonance

The Nightingale metabolomics platform (Nightingale Health Plc., Helsinki, Finland) is a targeted platform that utilizes proton NMR (1H NMR) (39, 42). 1H NMR detects the hydrogen atoms of a prespecified selection of metabolites or macromolecules—that must be typically found in high concentrations in the samples—to capture their spectral characteristics. Due to these requirements 1H NMR is not as sensitive as other methods, such as MS (41). However, the properties of 1H NMR make it ideal for in depth quantitative measurement of lipoprotein particles (39). Moreover, it enables the reproducible quantification of absolute concentrations. The Nightingale platform utilizes three “molecular windows” for detecting different groups of metabolites. The lipoprotein (LIPO) window mainly detects the strong signal of lipoprotein subclasses and their lipid content. The low molecular weight molecules (LMWM) window filters out the signals from the LIPO window to detect the molecules with low molecular weight such as amino acids and glucose metabolites (42). Finally, the window for the lipids and lipids related molecules (LIPID) is used to specifically measure the saturated and unsaturated fatty acids, free and esterified cholesterol, sphingolipids, and phosphoglycerides (43). By using these three windows, 1H NMR Nightingale enables both in depth lipoprotein quantification and detection of several low molecular weight metabolites that are typically not easily measured using NMR. The Nightingale platforms provides approximately 230-245 metabolite measures including ratios and measurements of subfractions of lipoproteins (39).

3.2.2 Metabolon: Liquid Chromatography Tandem Mass Spectrometry

Chromatography and Mass Spectrometry

Chromatography is an important step for the separation of the biological molecules to enable the use of MS, particularly for metabolite and protein quantification (21). Chromatography is a

technique that applies high pressures to push the components of a biological sample through columns of silica, thus separating them. This requires the samples to be properly prepared and ionized before the procedure in order for the components to travel in the column. Moreover, the high pressure applied into the column must be consistent. This pressure can be supplied by either liquid or gas, respectively referred to as liquid chromatography (LC) and gas chromatography (GC). Exceptionally high pressure is in an advanced variant technique known as ultra-high performance liquid chromatography. The silica columns are designed to exhibit unique chemicals properties that have binding affinity with specific types of biological molecules. This interaction between the molecules and the column causes them to travel slower than those that do not bind to the column. In addition, the travel times differ between the different molecules in the column due to their unique affinities. The time it takes the molecules to pass through the column (referred to as the retention time) is key for inferring the molecular characteristics of the reacting molecules. Following this separation, mass spectrometry can be used to identify the molecules (21).

Fragmentation of proteins and metabolites is an essential step preceding mass spectrometry. Fragmentation is commonly achieved by shooting beams of electrons that break the biological molecules. These fragments are subsequently detected by their electronic charge (z) and their mass (m). By combining these two values as the mass to charge (m/z) ratio, a relatively unique signature is assigned to the fragments. The collected data of m/z ratios of the fragments are then represented as "spectral data". For metabolomics, a second MS step is frequently applied. This technique of coupling a chromatographer with two mass spectrometers in tandem is known as chromatography MS/MS. Tandem MS are used in metabolomics to measure metabolites in the first MS and subsequently fragment them into smaller particles to be measured in the second MS. This substantially increases the sensitivity of the measured m/z ratios of metabolites (44). Subsequently, the retention time from the chromatography step and the spectral data of m/z ratios from both mass spectrometers are combined (21, 41). For final identification, these signatures are compared to a reference library containing a vast number of retention times and m/z ratios corresponding to annotated metabolites or proteins.

Ultra High-Performance Liquid Chromatography and Tandem Mass Spectrometry

Metabolon™ Discovery HD4 platform is an untargeted metabolomic platform at Metabolon Inc. (Durham, North Carolina, USA) that utilizes ultra high-performance liquid chromatography and tandem mass spectrometry (UHPLC-MS/MS). This platform uses four independent UHPLC-MS/MS platforms with different LC columns (41, 45). Two platforms use positive ionization reverse phase chromatography, one uses negative ionization reverse phase chromatography, and one uses hydrophilic interaction liquid chromatography negative ionization (45). Thus, the platform can measure a wide range of metabolites with different chemical properties and affinities with high sensitivity. The signal from the m/z ratio and retention time of the measured ionized molecules are subsequently cross referenced with an in-house large library of metabolite molecules. If the metabolite is known, it will be assigned the annotation from the library. Metabolites found in the library but without a full annotation will also be reported as novel unnamed metabolites (41, 46). The Metabolon™ HD4 platform currently measures up to 1400 metabolites from a single sample.

Xenobiotics

A feature of the Metabolon platforms is its ability to measure not only endogenous metabolites produced by the body but also externally acquired "xenobiotic", or exogenous metabolites in an untargeted fashion. Essential metabolites for the human body that are acquired from the diet are usually considered part of the endogenous metabolite group. Xenobiotics on the other

hand include nonessential metabolites from food consumption (such as caffeine) (8) Moreover, xenobiotics comprise metabolites and chemicals derived from environmental exposures (e.g., pollution or chemical contamination), nutrition and diet, lifestyle choices (e.g., smoking, applying cosmetics), and medication use. The study of the effect of these environmental and external exposures on individuals is also known as the study of the exposome. In addition to the factors above, the exposome also includes factors such as the general environment, social economic status, and climate related factors. Exposome research aims to elucidate the "external" causes of disease and improve their prevention by examining the environment of patients (47). Indeed, environmental exposures are suggested to have a stronger impact on health outcomes than genetic factors (48). Therefore, the Metabolon platform provides a glimpse of an individual's exposome in addition to the broad array of endogenous metabolites. In addition, novel "unnamed" metabolites are also measured which may belong to endogenous or xenobiotic sources. This enables the simultaneous study of both the internal and external metabolomic factors involved in disease etiology and clinical outcomes (8) and the identification of associations with novel, unknown metabolites.

4 GROWING PAINS: CHALLENGES IN EPIDEMIOLOGICAL STUDIES USING OMICS

4.1 Genomic Challenges: capturing structural variation and the missing heritability problem

Genomic studies were the first and are the most established of the contemporary OMICs fields. However, genomics studies have their shortcomings and limitations (49). One of which is their inherent inability to capture any effects from the exposome. Another issue with genomic studies is the observation that a very minor proportion, often less than 5%, of the heritability of many traits is explained by the genetic variants tested (50, 51). This is referred to as the missing heritability problem (52, 53). Despite the increase in the number of genomic studies and the sample sizes of the cohorts in these studies (often hundreds of thousands of subjects), the explained variance from these studies remains low (50). Basic genetic analyses can estimate the heredita-bility of diseases, especially in familial and twin studies that share large portions of the genome (54, 55). When GWAS became possible and widely available, it was expected that the identified loci and SNPs would fully account for the known hereditability estimates. Surprisingly, this was not the case. Even extensive GWAS with thousands of individuals still reported loci that combined explain a sub-portion of the hereditability of several phenotypes, such as T2D and height (53). It has been suggested that one of the reasons for the missing hereditability is that genetics alone do not capture the complete heritability of complex diseases, such as cardiometabolic diseases (52). The influence of environmental factors and their interaction with genetics could account for the missing heritability (52). Another potential reason for this problem is that GWAS do not usually include genetic variations outside of SNPs (52). SNPs are the most common type of variation in the genome and are defined as a germline substitution of a single nucleotide at a specific position in the genome. However, other mutation types exist such as copy number variations (CNV) or repeat expansions, in which large number of nucleotides or patterns of nucleotides are repeated in the genome. The intrinsic properties of SNPs make their detection using DNA sequencing techniques much easier then CNVs and other types of genetic variation. This has contributed to the focus on SNPs in GWAS. However, due to the rapid advances in genomic technology, it has become possible to readily detect CNVs and other structural variations in the genome (56). Indeed, these recent advances have enabled genomic research in examining CNVs

in large studies. These studies have found that CNVs strongly contributed to the hereditability of various traits such as height and weight (52).

4.2 Challenges in Proteomics and Metabolomics

Many challenges may be encountered when using large metabolomics and proteomics data sets. Technical issues can occur during the preparation, processing, and quantification phase of metabolites and proteins. These issues can either be specific to the type of platform and technique used or can be general biological or chemical problems associated with the compound to be measured. For instance, if the biological sample and platform preparation is not performed properly, then the measurement and detection of the metabolites or proteins would be affected. This can occur if, for example, aptamers are not prepared properly or if the chromatography columns are contaminated due to overuse. Moreover, the efficiency of the measurement technique differs depending on the targeted biomarkers. Chemically complex biomarkers can be more difficult to quantify than simpler ones (57).

A more general issue that occurs in OMIC research is the batch and run day effects. OMIC platforms usually use the standard 96 well plate format to store and measure the samples. Therefore, they are limited in the number of samples that can be simultaneously quantified. Studies of hundreds or thousands of individuals requires sending the samples in batches at different time points. Additionally, each batch needs to be split to smaller groups to be measured by the platform over several days. The variation of how well each batch is stored and handled can affect the level of contamination and degradation in the samples. Moreover, the level of contamination may accumulate in the measurement platform itself after each batch run. Thus, the sensitivity of the platform and the samples quality can differ per run day and per batch, resulting in potential batch effects and measurement errors.

Another common problem during quantification occurs when some metabolites or proteins are in low concentrations in the sample. A limitation of most OMICs platforms is their inability to distinguish between different metabolites or proteins if their concentrations are below a certain cutoff range. This cutoff is referred to as the limit of detection (LoD). Metabolites or proteins below the LoD cannot be quantified and instead their concentrations are set to 0 or left blank (46). Run day and batch effects also contribute to the sensitivity of the platform and, in turn, the range of the LoD limits for the platforms (58).

After the physical quantification of the biological samples, computational post-processing steps are required. In these steps, further issues can occur. For example, correct matching of the m/z ratio and retention time from the UHPLC-MS/MS based Metabolon platform to the correct entry in the reference library is prone to machine and human errors. The signal matching procedure is usually automated by a software tool and then double checked manually. However, due to software errors or human errors, it is possible that a metabolite signal is not matched correctly. Similarity between metabolites or poor calibration of the platform during measurement also contribute to likelihood of these issues. Thus, a valid metabolite signature could be unmatched and, in worst case scenarios, be incorrectly matched with a completely different metabolite signature (46).

Once the quantification and postprocessing steps are complete, the generated data is used for statistical analysis. Here as well complications can occur. OMICs approaches have the benefit of measuring hundred to thousands of biochemicals from a single sample leading to high dimensional data. Often, the number of measurements is larger than the number of individuals in the study (N<P). If the sample size is too small, then this high dimensionality can decrease

the power of the study and lead to aberrant results. Furthermore, even if the sample size is sufficient, high dimensional data requires extra analytical procedures such as adjustment for multiple testing (59). Furthermore, beforementioned issues that may arise during quantification and postprocessing must be addressed during the statistical analysis. Indeed, transformation methods must be used to treat the variations between the batches and the run days. Likewise, missing measurements from LoD or other technical difficulties must also be addressed.

4.2.1 Missing Data in Metabolomics

Treating missing data is a common issue in epidemiological research. In metabolomics, handling and imputing missing values in metabolite measurements is a particularly important and challenging issue. As mentioned, missing values can occur due to contaminations in the platforms which in turn affects the sensitivity of the measurements. This issue is more common in untargeted metabolomics platforms (58). The reason for this is that the nature of untargeted metabolomics is to detect a large number of metabolites without prior selection. This approach may suffer from errors during the signal identification and signal matching steps (58). Naturally, the odds of these mistakes occurring increases as the number of metabolites measured expands, such as the case in untargeted metabolomics, and as the number of samples increases. Missing values can also occur due to the beforementioned technical difficulties, such as LoD (60), batch and run day effects, and mismatching issues (58). In addition, missing values could be truly missing and should not be imputed at all. This is the case with xenobiotic metabolites that are expected to be measured in specific individuals only. For example, imputing missing values for metabolites related to the metformin would imply that all participants are diabetics. These different issues and the high dimensionality of the data makes it statistically challenging to apply appropriate imputation methods (58).

5 STUDY POPULATIONS

As aforementioned, OMIC research is typically performed with large sample sizes in large population-based studies to accommodate the large number of OMIC variables. The work in this thesis involved several population-based studies and collaborations with Dutch and international research groups.

5.1 NEO

The Netherlands Epidemiology of Obesity (NEO) study is an ongoing population-based, prospective cohort study of individuals aged 45–65 years, with an oversampling of individuals with overweight or obesity. Men and women aged between 45 and 65 years with a self-reported BMI of 27 kg/m^2 or higher, living in the greater area of Leiden (in the West of the Netherlands) were eligible to participate in the NEO study. In addition, all inhabitants aged between 45 and 65 years from one municipality (Leiderdorp) were invited, irrespective of their BMI. Recruitment of participants started in September 2008 and completed at the end of September 2012. In total, 6,671 participants have been included, of whom 5,217 with a BMI of 27 kg/m^2 or higher. Participants were invited to come to the NEO study center of the Leiden University Medical Center for one baseline study visit after an overnight fast of at least 10 hours. During the visit a blood sample of 108 mL was taken from the participants (61). The study was approved by the medical ethical committee of the Leiden University Medical Center.

5.2 THE-VTE

The Thrombophilia, Hypercoagulability and Environmental Risks in Venous Thromboembolism (THE-VTE) study — a multicenter case control study from Leiden (The Netherlands) and Cambridge (UK) (62). Inclusion took place between March 2003 and December 2008. In total, 626 patients were included, aged 18-75, with a first DVT or PE. Partners of the patients were invited as controls. Subsequent follow-up of the cases was performed to assess recurrence risk. The mean follow-up duration was 4.8 years after discontinuation of oral anticoagulant therapy. Blood samples were taken 2-3 months after discontinuation of anticoagulants. The study was approved by the Medical Ethics Committee of the Leiden University Medical Centre (Leiden, Netherlands) and the NHS Research Ethics Committee in Cambridge, UK.

5.3 PROSPER

The Prospective Study of Pravastatin in the Elderly at Risk (PROSPER) was a randomized, double-blind, placebo-controlled trial among 5,786 men and women between 70-82 years old with a pre-existing vascular disease or a raised risk for such a disease. The aim of the trial was to test the benefits of Pravastatin. Participants were recruited from three countries with 2,517 individuals from Scotland, 2,173 individuals from Ireland and 1,096 individuals from the Netherlands. Fasting blood sample were collected and stored at -80 degrees for later NMR metabolomics analysis (63). The study was approved by the institutional ethics review boards of all centers and written informed consent was obtained from all participants (64).

5.4 NESDA

The Netherlands Study of Depression and Anxiety (NESDA) is an ongoing longitudinal cohort study into the long-term course and consequences of depressive and anxiety disorders. The sample consists of 2,981 participants with depressive/anxiety disorders and healthy controls recruited from the general population, general practices, and secondary mental health centers (65). Blood samples were collected after an overnight fast at the baseline visit (2004-2007). The Ethical Committees of all participating universities approved the NESDA project, and all participants provided written informed consent (66)

5.5 The Rhineland Study

The Rhineland Study is an ongoing prospective population-based cohort study based in two geographically defined areas In Bonn, Germany. Participants were recruited via invitation beginning in 2016. The primary focus of the study is on aging and age-related brain disorders in adult life. The source population consists of all inhabitants aged 30 years or older in the specified Bonn area. Participation was only possible upon invitation and regardless of health status provided they had sufficient command in the German language to provide an informed consent. The ethics committee of the medical faculty of the University of Bonn approved the undertaking of the study and it was carried out according to the recommendations of the International Council for Harmonisation Good Clinical Practice standards. Written informed consent was acquired from all participants per the Declaration of Helsinki.

5.6 INTERVAL

INTERVAL is a prospective cohort study nested within a pragmatic randomized trial of blood donors enrolled from 25 static centers of NHS Blood and Transplant (67). Recruitment of about 50,000 male and female donors started in June 2012 and was completed in June 2014. Blood

donors 18 years and older were consented and recruited from 25 National Health Service Blood and Transplant (NHSBT) static donor centers across England. All participants fulfilled all normal criteria for blood donation (68). Therefore, participants included in the study were predominantly healthy. The INTERVAL study was approved by the Cambridge (East) Research Ethics Committee. Written informed consent was obtained from all participants.

6 OUTLINE

In this thesis, we will look at the methodological challenges and epidemiological applications of genomics, proteomics, and in particular metabolomics. Part I focuses on methodological challenges and applications of proteomics and metabolomics research. In Part I, **Chapter 2**, we demonstrate the application of measures of agreement to compare contemporary large-scale aptamer-based proteomics with standardized clinical measurements in the THE-VTE study. Part I, **Chapter 3** explores the challenges of treating missing data in metabolomics and describes a workflow for imputing the different types of missing data. In Part I, **Chapter 4**, we present a metabolomic age prediction model based on 826 UHPLC-MS/MS measured metabolites from the INTERVAL study. We also report our evaluation of the model using several comorbidities in the NEO study. Part II of this of this thesis focuses on the etiological applications of metabolomics and genomics. In Part II, **Chapter 5**, we combine regression analysis and network analysis to investigate the association between UHPLC-MS/MS metabolite measurements with hepatic triglyceride content in the NEO study. In addition, we illustrate the results as an interactive online atlas. Part II, **Chapter 6**, focuses on the effects of genetic tandem repeat mutations in the *HTT* gene on NMR metabolite measurements in the NEO, PROSPER, and NESDA studies. We further explore the role of BMI mediation on the associations. In Part II, **Chapter 7**, we present the results for the effect of the environmental contaminant Per- and polyfluoroalkyl substances (PFAS) on the metabolic and lipoprotein profile of the general populations in Germany (The Rhineland Study) and the Netherlands (NEO study). In Part III, **Chapter 8**, we discuss our findings and offer our thoughts on the future evolution of multi-OMIC in research.

7 REFERENCES

1. Rood JE, Regev A. The legacy of the Human Genome Project. 2021;373(6562):1442-3.
2. Gibbs RA. The Human Genome Project changed everything. Nature Reviews Genetics. 2020;21(10):575-6.
3. Khoury MJ, Bowen MS, Clyne M, Dotson WD, Gwinn ML, Green RF, et al. From public health genomics to precision public health: a 20-year journey. Genetics in Medicine. 2018;20(6):574-82.
4. Molster CM, Bowman FL, Bilkey GA, Cho AS, Burns BL, Nowak KJ, et al. The Evolution of Public Health Genomics: Exploring Its Past, Present, and Future. Frontiers in public health. 2018;6:247.
5. Atanasovska B, Kumar V, Fu J, Wijmenga C, Hofker MH. GWAS as a Driver of Gene Discovery in Cardiometabolic Diseases. Trends in Endocrinology & Metabolism. 2015;26(12):722-32.
6. PHG Foundation. What is transcriptomics? : PHG Foundation; 2022 [updated 2022/08/16/. Available from: https://www.phgfoundation.org/blog/what-is-transcriptomics.
7. CDC. What is Epigenetics? | CDC Centers for Disease Control and Prevention 2022 [updated 2022/08/15/. Available from: https://www.cdc.gov/genomics/disease/epigenetics.htm.
8. Rattray NJW, Deziel NC, Wallach JD, Khan SA, Vasiliou V, Ioannidis JPA, et al. Beyond genomics: understanding exposotypes through metabolomics. Hum Genomics. 2018;12(1):4.
9. Hartley H. Origin of the Word 'Protein'. Nature. 1951;168(4267):244-.
10. Pandey A, Mann M. Proteomics to study genes and genomes. Nature. 2000;405(6788):837-46.
11. Strasser BJ. Collecting, comparing, and computing sequences: the making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure, 1954-1965. Journal of the history of biology. 2010;43(4):623-60.
12. Cox J, Mann M. Is Proteomics the New Genomics? Cell. 2007;130(3):395-8.
13. Ponomarenko EA, Poverennaya EV, Ilgisonis EV, Pyatnitskiy MA, Kopylov AT, Zgoda VG, et al. The Size of the Human Proteome: The Width and Depth. International journal of analytical chemistry. 2016;2016:7436849.
14. Gonzalez MW, Kann MG. Chapter 4: Protein interactions and disease. PLoS computational biology. 2012;8(12):e1002819.
15. Sjöstedt E, Zhong W, Fagerberg L, Karlsson M, Mitsios N, Adori C, et al. An atlas of the protein-coding genes in the human, pig, and mouse brain. 2020;367(6482):eaay5947.
16. Ferrannini G, Manca ML, Magnoni M, Andreotti F, Andreini D, Latini R, et al. Coronary Artery Disease and Type 2 Diabetes: A Proteomic Study. Diabetes Care. 2020;43(4):843-51.
17. Whelan CD, Mattsson N, Nagle MW, Vijayaraghavan S, Hyde C, Janelidze S, et al. Multiplex proteomics identifies novel CSF and plasma biomarkers of early Alzheimer's disease. Acta neuropathologica communications. 2019;7(1):169.
18. Niu L, Geyer PE, Wewer Albrechtsen NJ, Gluud LL, Santos A, Doll S, et al. Plasma proteome profiling discovers novel proteins associated with non-alcoholic fatty liver disease. Molecular systems biology. 2019;15(3):e8793.
19. Tardif G, Paré F, Gotti C, Roux-Dalvai F, Droit A, Zhai G, et al. Mass spectrometry-based proteomics identify novel serum osteoarthritis biomarkers. Arthritis research & therapy. 2022;24(1):120.
20. Edfors F, Iglesias MJ, Butler LM, Odeberg J. Proteomics in thrombosis research. Research and practice in thrombosis and haemostasis. 2022;6(3):e12706.
21. Alonso A, Marsal S, Julia A. Analytical methods in untargeted metabolomics: state of the art in 2015. Front Bioeng Biotechnol. 2015;3:23.
22. Ryals J, Lawton K, Stevens D, Milburn M. Metabolon, Inc. Pharmacogenomics. 2007;8(7):863-6.
23. Roethig HJ, Munjal S, Feng S, Liang Q, Sarkar M, Walk RA, et al. Population estimates for biomarkers of exposure to cigarette smoke in adult U.S. cigarette smokers. Nicotine & tobacco research : official journal of the Society for Research on Nicotine and Tobacco. 2009;11(10):1216-25.
24. Willems RJ. Individual metabolic patterns and human disease : an exploratory study utilizing predominantly paper chromatographic methods. The University of Texas Publication. 1951;BIOCHEMICAL INSTITUTE STUDIES IV

25. Williams RJ, Berry LJ, Beerstecher E. Individual Metabolic Patterns, Alcoholism, Genetotrophic Diseases*. 1949;35(6):265-71.
26. Gates SC, Sweeley CC. Quantitative metabolic profiling based on gas chromatography. Clinical chemistry. 1978;24(10):1663-73.
27. Miggiels P, Wouters B, van Westen GJP, Dubbelman A-C, Hankemeier T. Novel technologies for metabolomics: More for less. TrAC Trends in Analytical Chemistry. 2019;120:115323.
28. Tikkanen E, Jägerroos V, Holmes MV, Sattar N, Ala-Korpela M, Jousilahti P, et al. Metabolic Biomarker Discovery for Risk of Peripheral Artery Disease Compared With Coronary Artery Disease: Lipoprotein and Metabolite Profiling of 31 657 Individuals From 5 Prospective Cohorts. 2021;10(23):e021995.
29. Cirulli ET, Guo L, Leon Swisher C, Shah N, Huang L, Napier LA, et al. Profound Perturbation of the Metabolome in Obesity Is Associated with Health Risk. Cell Metabolism. 2019;29(2):488-500.e2.
30. Bot M, Milaneschi Y, Al-Shehri T, Amin N, Garmaeva S, Onderwater GLJ, et al. Metabolomics Profile in Depression: A Pooled Analysis of 230 Metabolic Markers in 5283 Cases With Depression and 10,145 Controls. Biological psychiatry. 2020;87(5):409-18.
31. Masoodi M, Gastaldelli A, Hyötyläinen T, Arretxe E, Alonso C, Gaggini M, et al. Metabolomics and lipidomics in NAFLD: biomarkers and non-invasive diagnostic tests. Nature reviews Gastroenterology & hepatology. 2021;18(12):835-56.
32. Zhang A, Sun H, Yan G, Wang P, Han Y, Wang X. Metabolomics in diagnosis and biomarker discovery of colorectal cancer. Cancer Letters. 2014;345(1):17-20.
33. Gold L, Ayers D, Bertino J, Bock C, Bock A, Brody EN, et al. Aptamer-based multiplexed proteomic technology for biomarker discovery. PLoS One. 2010;5(12):e15004.
34. Joshi A, Mayr M. In Aptamers They Trust: The Caveats of the SOMAscan Biomarker Discovery Platform from SomaLogic. Circulation. 2018;138(22):2482-5.
35. Suhre K, Arnold M, Bhagwat AM, Cotton RJ, Engelke R, Raffler J, et al. Connecting genetic risk to disease end points through the human blood plasma proteome. Nature Communications. 2017;8(1):14357.
36. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, et al. Genomic atlas of the human plasma proteome. Nature. 2018;558(7708):73-9.
37. Ganz P, Heidecker B, Hveem K, Jonasson C, Kato S, Segal MR, et al. Development and Validation of a Protein-Based Risk Score for Cardiovascular Outcomes Among Patients With Stable Coronary Heart Disease. Jama. 2016;315(23):2532-41.
38. Cuvelliez M, Vandewalle V, Brunin M, Beseme O, Hulot A, de Groote P, et al. Circulating proteomic signature of early death in heart failure patients with reduced ejection fraction. Scientific Reports. 2019;9(1):19202.
39. Soininen P, Kangas AJ, Wurtz P, Suna T, Ala-Korpela M. Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. Circ Cardiovasc Genet. 2015;8(1):192-206.
40. Emwas A-HM, Salek RM, Griffin JL, Merzaban J. NMR-based metabolomics in human disease diagnosis: applications, limitations, and recommendations. Metabolomics. 2013;9(5):1048-72.
41. Evans A, Bridgewater B, Liu Q, Mitchell M, Robinson R, Dai H, et al. High resolution mass spectrometry improves data quantity and quality as compared to unit mass resolution mass spectrometry in high-throughput profiling metabolomics. Metabolomics. 2014;4(2):1.
42. Soininen P, Kangas AJ, Würtz P, Tukiainen T, Tynkkynen T, Laatikainen R, et al. High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism. The Analyst. 2009;134(9):1781.
43. Fuertes-Martín R, Correig X, Vallvé JC, Amigó N. Human Serum/Plasma Glycoprotein Analysis by (1) H-NMR, an Emerging Method of Inflammatory Assessment. Journal of clinical medicine. 2020;9(2).
44. Heiles S. Advanced tandem mass spectrometry in metabolomics and lipidomics—methods and applications. Analytical and Bioanalytical Chemistry. 2021;413(24):5927-48.
45. Rhee EP, Waikar SS, Rebholz CM, Zheng Z, Perichon R, Clish CB, et al. Variability of Two Metabolomic Platforms in CKD. Clinical Journal of the American Society of Nephrology. 2019;14(1):40.

46. Dehaven C, Evans A, Dai H, Lawton K. Software Techniques for Enabling High-Throughput Analysis of Metabolomic Datasets. 2012.
47. Vrijheid M. The exposome: a new paradigm to study the impact of environment on health. Thorax. 2014;69(9):876-8.
48. DeBord DG, Carreón T, Lentz TJ, Middendorf PJ, Hoover MD, Schulte PA. Use of the "Exposome" in the Practice of Epidemiology: A Primer on -Omic Technologies. American journal of epidemiology. 2016;184(4):302-14.
49. Rochfort S. Metabolomics reviewed: a new "omics" platform technology for systems biology and implications for natural products research. Journal of natural products. 2005;68(12):1813-20.
50. Matthews LJ, Turkheimer E. Three legs of the missing heritability problem. Studies in history and philosophy of science. 2022;93:183-91.
51. Altmüller J, Palmer LJ, Fischer G, Scherb H, Wjst M. Genomewide scans of complex human diseases: true linkage is hard to find. American journal of human genetics. 2001;69(5):936-50.
52. Génin E. Missing heritability of complex diseases: case solved? Human Genetics. 2020;139(1):103-13.
53. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747-53.
54. Falconer DS. The inheritance of liability to certain diseases, estimated from the incidence among relatives. 1965;29(1):51-76.
55. Young AI. Solving the missing heritability problem. PLOS Genetics. 2019;15(6):e1008222.
56. Gordeeva V, Sharova E, Arapidi G. Progress in Methods for Copy Number Variation Profiling. International journal of molecular sciences. 2022;23(4).
57. Joshi R, Wannamethee G, Engmann J, Gaunt T, Lawlor DA, Price J, et al. Establishing reference intervals for triglyceride-containing lipoprotein subfraction metabolites measured using nuclear magnetic resonance spectroscopy in a UK population. Annals of clinical biochemistry. 2021;58(1):47-53.
58. Do KT, Wahl S, Raffler J, Molnos S, Laimighofer M, Adamski J, et al. Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. Metabolomics : Official journal of the Metabolomic Society. 2018;14(10):128-.
59. Tzoulaki I, Ebbels TMD, Valdes A, Elliott P, Ioannidis JPA. Design and Analysis of Metabolomics Studies in Epidemiologic Research: A Primer on -Omic Technologies. American journal of epidemiology. 2014;180(2):129-39.
60. Redestig H, Kobayashi M, Saito K, Kusano M. Exploring matrix effects and quantification performance in metabolomics experiments using artificial biological gradients. Anal Chem. 2011;83(14):5645-51.
61. de Mutsert R, den Heijer M, Rabelink TJ, Smit JW, Romijn JA, Jukema JW, et al. The Netherlands Epidemiology of Obesity (NEO) study: study design and data collection. Eur J Epidemiol. 2013;28(6):513-23.
62. van Hylckama Vlieg A, Baglin CA, Luddington R, MacDonald S, Rosendaal FR, Baglin TP. The risk of a first and a recurrent venous thrombosis associated with an elevated D-dimer level and an elevated thrombin potential: results of the THE-VTE study. Journal of thrombosis and haemostasis : JTH. 2015;13(9):1642-52.
63. Delles C, Rankin NJ, Boachie C, McConnachie A, Ford I, Kangas A, et al. Nuclear magnetic resonance-based metabolomics identifies phenylalanine as a novel predictor of incident heart failure hospitalisation: results from PROSPER and FINRISK 1997. Eur J Heart Fail. 2018;20(4):663-73.
64. Shepherd J, Blauw GJ, Murphy MB, Bollen EL, Buckley BM, Cobbe SM, et al. Pravastatin in elderly individuals at risk of vascular disease (PROSPER): a randomised controlled trial. Lancet. 2002;360(9346):1623-30.
65. Penninx BW, Beekman AT, Smit JH, Zitman FG, Nolen WA, Spinhoven P, et al. The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. Int J Methods Psychiatr Res. 2008;17(3):121-40.

66. de Kluiver H, Jansen R, Milaneschi Y, Bot M, Giltay EJ, Schoevers R, et al. Metabolomic profiles discriminating anxiety from depression. Acta Psychiatr Scand. 2021;144(2):178-93.
67. Moore C, Sambrook J, Walker M, Tolkien Z, Kaptoge S, Allen D, et al. The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. Trials. 2014;15:363-.
68. NHS. NHS Blood and Transplant criteria for giving blood 2014 [updated 2022/08/11/. Available from: https://www.blood.co.uk/who-can-give-blood.