Beyond GWAS in atrial fibrillation genetics

Sander W. van der Laan, PhD[1], and Folkert. W. Asselbergs, MD PhD[2]

1 Central Diagnostics Laboratory, Laboratories, Pharmacy, and Biomedical Genetics, University Medical Center Utrecht, Utrecht University
2 Department of Cardiology, Division Heart & Lungs, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands;
3 Institute of Cardiovascular Science and Health Informatics, Faculty of Population Health Sciences, University College London, London, United Kingdom.

Atrial fibrillation (AF) is one of the leading causes of cardiovascular diseases. Despite tremendous progress in medicine, over 33 million individuals suffer from AF with a lifetime risk of ~1 in 3 for individuals from European ancestry[1]. The causes leading up to AF are myriad: environmental, behavioural, and genetic factors influence and modify the extent of disease[1].

So far, large-scale genome-wide association studies (GWAS) have identified over 100 genetic susceptibility loci for AF, but fail to explain more than 6.4% of the heritability (which is estimated to be 22%)[REF]. To improve the understanding of the biomolecular causes of AF, Wang and colleagues had previously explored the contribution of the transcriptome and methylome (of whole blood) on risk of atrial fibrillation[REF]. While these omics-studies revealed more of the biological complexities in disease etiology, there is a limit in terms of increasing sample size, and alternative methodologies are required to dissect the root causes of diseases.

In the current issue of *the journal*, Wang and colleagues[REF] hypothesized that integration of existing omics datasets should yield more information: the whole is greater than the sum of its parts. In their study three layers of biomolecular data, i.e. genetic variation (GWAS), DNA methylation at cytosine-guanine dinucleotides (CpGs) in an epigenome-wide association study (EWAS), and gene expression (transcriptome-wide association study [TWAS]), were integrated to improve the resolution of genes associated with AF. Per variant or per-CpG summary statistics were first collapsed to per-gene empirically derived associated p-values (Figure 1 in Wang et al.). These gene-based association statistics underscore the distinct information each biomolecular layer provides as there are no overlapping individual genes between the three datasets (Figure 2 in Wang et al.).
To integrate the data, Wang and colleagues performed a weighted meta-analysis to account for the different samples sizes between datasets. The resulting genesets were mapped to tissue-specific gene expression networks (heart and whole blood) and classified through machine learning models using NetWAS[2]. Next, they validated their integrated approach, that resulted in more predicted AF-related genes by comparing these with the genes associated to the latest GWAS on AF; they assessed the change in the C-statistic as a measure of predictive accuracy. They also quantified the contribution of the individual datasets to the overall integrated result by examining the effects of different (sample size) weights and gene

significance thresholds on the C-statistic. The results suggested that lenient gene significance thresholds and the integration with EWAS and TWAS data increased the yield of AF-related genes, which was not explained by a large GWAS sample size (as compared to EWAS or TWAS). Through this integrative approach 1,931 AF-related genes were identified, 288 of these were previously associated with disease. Subsequent pathway enrichment analyses revealed most of the 1,931 genes are involved in cardiac development and adrenergic signaling in cardiomyocytes. Tissue-specific gene expression analysis revealed highest expression in heart and skeletal muscles. Importantly, a few potential drug targets were identified, and gene-based heritability estimates were increased when comparing the contribution of the 1,931 to that estimated from genes solely derived from a GWAS.

Future perspective

One might think that with technological advancements and declining costs of omics-methodology, sample sizes will increase and all biological causes of AF - or CVD in general - will soon be discovered and heritability fully explained. However, the glass might not be that full just yet, as it has become clear that genetic variation can not solely explain disease susceptibility, and transcriptional and translational regulation is very complex. For example, DNA is wrapped in chromatin, acetylated or methylated and physically interacting with different elements across millions of base pairs, RNA can be non-coding and spliced, protein-coding or transcription regulatory, and proteins can be glycosylated and phosphorylated, and provide a transcriptional feedback by binding to DNA. And all these processes are occuring depending on external stimuli (the exposome) and are tissue specific. The old linear approach - focused on only one biomolecular aspect of disease - rapidly becomes obsolete, and Wang and colleagues explored an innovative integrative approach.
Theirs is not the only method to integrate more information, nor will it be the final one. Usually, GWAS are univariate, exploring only one phenotype at a time. Multivariate GWAS are based on the intuition that correlated phenotypes share biology, e.g. circulating lipids, and therefore are informative to each other[3]. Indeed, a multivariate GWAS of depressive disorders revealed more genetic loci than each individual univariate GWAS[3]. Likewise, gene regulatory network (GRN) analyses, based on genetic variants affecting tissue-specific gene expression, take a more holistic approach and consider the networks contributing to diseases, rather the individual genes. In a study of metabolic tissues obtained from coronary artery diseases (CAD) patients, GRNs explain more of the heritability of CAD than individual univariate GWAS could[4].

Even so, we need to set the bar higher and point to three key challenges in these large-scale omics-studies. First, we need to explore different methods to map (epi)genetic variation to genes. Often this is based on physical location, i.e. the statistically most significant variant or CpG is mapped to the closest gene within a certain base pair boundary like Wang and colleagues did. However, this could also be based, for example, on cell- and trait specific chromatin-protein interaction boundaries. The current study identified genes through transcriptomic and methylomic profiles derived from whole blood, but advances in single-cell sequencing has made it possible to increase the resolution to the individual cellular constituents of blood. Associating genetic variation with transcriptional, regulatory, or translational data at single-cell resolution could be more informative. Recently, the

CVgenes@target Consortium (http://cvgenesattarget.eu/) mapped genetic variation at over 200 loci associated to CAD based on physical location, the coding consequences, their (predicted) deleterious effects, effects (of putative genes) in murine atherosclerotic models, and the effects on tissue-specific gene expression[5]. Through this integrative approach we prioritized 68 putatively causal genes with a high druggability potential for CAD. But most importantly, these efforts underscore the need to converge to a standardized approach of mapping (epi)genetic variation to genes.

Second, to fulfill the promise of systems biology - the true integration of biomolecular information - sharing and standardization of data, analytical models, pipelines, and quality control will be key. The elements for this are in place. The UK Biobank (https://www.ukbiobank.ac.uk/) is an exemplar for collaborative data science opening up the data of over 500,000 participants, including genetic variation and over 4,700 different traits and measurements, to every scientist across the globe. However, disease or trait definitions are not uniform across studies using UK Biobank data. The cardiovascular disease knowledge portal (http://www.broadcvdi.org/) provides standardized access to cardiovascular genetic data, while keeping data integrity and privacy, but lacks the ability for in depth (re-)analysis. Code sharing and maintaining through platforms like GitHub (www.github.com) increase quality of code while enabling more scientists to apply the same method to different datasets, but GitHub lacks datasets. A logical next step would be the creation of a platform where 1) data is kept locally - to preserve data integrity and privacy - but is accessible to others through federated learning, and 2) scientist work simultaneous on codes to increase quality, interoperability and uniformity of definitions (e.g. similar to Wikipedia).

Third, all of these integrative approaches are *in silico* and often lack experimental evidence showing that the discovered genes are indeed causative to disease. In that respect, the study by Wang was hypothesis generating - as are many of the above described *in silico* methods. Technological advancements provide the opportunity to examine each step in the process from DNA to RNA to protein by means of various sequencing and mass-spectrometry methods[6]. In parallel, laboratory techniques are available to manipulate cells and tissues to the nucleic-acid level with CRISPR-Cas9[7], and visualize spatial expression in tissues at single-cell resolution[8]. For better interpretation of *in silico* integrative methods, high-throughput screening of putative targets in relevant assays are needed. Induced pluripotent stem cells (iPSC), derived from skin or circulating cells, are an unlimited source of any type of human cells and enables the study of disease specific variants and their effects in disease mechanisms.

Wang and colleagues have shown a great example of how to join different types of data, but true progression will be the integration of different types of scientists (e.g. data, biomedical and clinical scientist - Figure 1). The days of the homo universalis that can disrupt and break through barriers in different scientific areas have long passed. Nowadays, the c in science stands for collaboration holding the key to push the boundaries and shift paradigms.

References

1. Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP,

Chamberlain AM, Chang AR, Cheng S, Das SR, Delling FN, Djousse L, Elkind MSV, Ferguson JF, Fornage M, Jordan LC, Khan SS, Kissela BM, Knutson KL, Kwan TW, Lackland DT, Lewis TT, Lichtman JH, Longenecker CT, Loop MS, Lutsey PL, Martin SS, Matsushita K, Moran AE, Mussolino ME, O'Flaherty M, Pandey A, Perak AM, Rosamond WD, Roth GA, Sampson UKA, Satou GM, Schroeder EB, Shah SH, Spartano NL, Stokes A, Tirschwell DL, Tsao CW, Turakhia MP, VanWagner LB, Wilkins JT, Wong SS, Virani SS, American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association. *Circulation*. 2019;139:e56–e528.

2. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, Zhang R, Hartmann BM, Zaslavsky E, Sealfon SC, Chasman DI, FitzGerald GA, Dolinski K, Grosser T, Troyanskaya OG. Understanding multicellular function and disease with human tissue-specific networks. *Nat Genet*. 2015;47:569–576.

3. Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, Nguyen-Viet TA, Wedow R, Zacher M, Furlotte NA, Magnusson P, Oskarsson S, Johannesson M, Visscher PM, Laibson D, Cesarini D, Neale BM, Benjamin DJ, 23andMe Research Team, Social Science Genetic Association Consortium. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat Genet* [Internet]. 2018;Available from: http://dx.doi.org/10.1038/s41588-017-0009-4

4. Zeng L, Talukdar HA, Koplev S, Giannarelli C, Ivert T, Gan L-M, Ruusalepp A, Schadt EE, Kovacic JC, Lusis AJ, Michoel T, Schunkert H, Björkegren JLM. Contribution of Gene Regulatory Networks to Heritability of Coronary Artery Disease. *J Am Coll Cardiol*. 2019;73:2946–2957.

5. Lempiäinen H, Brænne I, Michoel T, Tragante V, Vilne B, Webb TR, Kyriakou T, Eichner J, Zeng L, Willenborg C, Franzen O, Ruusalepp A, Goel A, van der Laan SW, Biegert C, Hamby S, Talukdar HA, Foroughi Asl H, CVgenes@target consortium, Pasterkamp G, Watkins H, Samani NJ, Wittenberger T, Erdmann J, Schunkert H, Asselbergs FW, Björkegren JLM. Network analysis of coronary artery disease risk genes elucidates disease mechanisms and druggable targets. *Sci Rep*. 2018;8:3434.

6. Nurnberg ST, Zhang H, Hand NJ, Bauer RC, Saleheen D, Reilly MP, Rader DJ. From Loci to Biology: Functional Genomics of Genome-Wide Association for Coronary Disease. *Circ Res*. 2016;118:586–606.

7. Fellmann C, Gowen BG, Lin P-C, Doudna JA, Corn JE. Cornerstones of CRISPR-Cas in drug discovery and therapy. *Nat Rev Drug Discov*. 2017;16:89–100.

8. Asp M, Salmén F, Ståhl PL, Vickovic S, Felldin U, Löfling M, Fernandez Navarro J, Maaskola J, Eriksson MJ, Persson B, Corbascio M, Persson H, Linde C, Lundeberg J. Spatial detection of fetal marker genes expressed at low level in adult human heart tissue. *Sci Rep*. 2017;7:12941.

Acknowledgement

SWvdL is supported by The Netherlands Heart Foundation (CVON2017-20: Generating the best evidence-bases pharmaceutical targets and drugs for atherosclerosis [GENIUS II]). FWA is supported by UCL Hospitals NIHR Biomedical Research Centre. The European Research

Area Network on Cardiovascular diseases (ERA-CVD, grant number 01KL1802) is kindly acknowledged.

Legend Figure 1

Different types of scientists will collaborate and integrate different types of data to push the boundaries of medical science. Technological advancements have made it possible to interrogate the human condition in sickness or in health at every level (A), but sharing of codes and methods, standardization of analytical approaches, and interoperability and uniformity of definitions will be key. High-throughput screening of putative causal genes at the cellular and model organism level and integration with data from human Mendelian diseases and traits are to be pursuit to discern disease relevant targets (B). Ultimately this will yield innovative new drugs, clinical intervention regimes, and lifestyle changes (C).