Twin Research and Human Genetics (2020), **23**, 77–79
doi:10.1017/thg.2020.26

CAMBRIDGE UNIVERSITY PRESS

Article

Genetics of Biochemical Phenotypes

John B. Whitfield

Genetic Epidemiology, QIMR Berghofer Medical Research Institute, Brisbane, Australia

Abstract

Biomarkers diagnose, predict or assess the risk of disease, and studies of the effects of genetic variation on biomarker phenotypes in the general population complement studies on patients diagnosed with disease. This paper traces the evolution of studies on biomarker genetics over the past 40 years through examples drawn from the work of Professor Martin and his colleagues.

Keywords: Heritability; candidate gene studies; genetic linkage; genomewide association studies

(Received 19 February 2020; accepted 26 March 2020)

Our collaboration on genetic effects on biochemical characteristics began in about 1979 when Nick Martin was at the Australian National University, in Canberra, and exploring the possibility of conducting what became the Alcohol Challenge Twin Study (ACTS; Martin et al., 1985). He visited Sydney and came to see me at Royal Prince Alfred Hospital, mainly to ask about laboratory tests to assess subjects' alcohol intake. I had to tell him that the prospects for estimating alcohol intake accurately for an individual person were poor, but we went on to agree that doing a range of biochemical tests on twins and using the results to assess heritability would be valuable. At that time there were few studies of this kind, and they had mostly focused on lipids (particularly cholesterol) because of its relevance to cardiovascular disease.

Over the subsequent 40 years, our biochemical studies developed through a number of stages, as happened for other phenotypes of biomedical interest. From initial steps to establish the existence of genetic effects and estimate heritability using comparisons of monozygotic (MZ) and dizygotic (DZ) pair similarity, study size grew to allow consideration of genetic correlations between phenotypes. By about 1990, genotyping of variants in candidate genes was becoming possible and soon after that the typing of genomewide microsatellite markers led to (mostly unsuccessful) attempts to identify loci affecting quantitative variation by genetic linkage. Around 2005, the technical advances allowing manufacture of genotyping arrays, and the conceptual step from linkage to association testing (Risch & Merikangas, 1996), made genomewide association studies (GWAS) possible. Possible, that is, if one had access to samples for DNA extraction, phenotypic information, consent from study participants and funds to purchase the genotyping chips. Fortunately, we had the first three and this led gradually to the fourth.

The results of the GWAS revolution are still playing out, but developments so far include not only identification of loci affecting quantitative variation, but a greater understanding of the relationships between phenotypes (including between biomarkers and disease) and increasing use of genetic results to address questions of causation in epidemiology.

Heritability and Other Twin Pair Designs

Blood samples from the ACTS participants were used for a range of biochemical and hematological tests, and the results led to 10 papers that tended to estimate heritability and (because a subsample of ACTS participants were willing to return for a second time) repeatability. This combination clarified a fact that is still not sufficiently appreciated; when heritability and test–retest repeatability are similar, the long-term average of a diagnostic biomarker or risk factor is strongly dependent on genetic variation and environmental effects tend to be evanescent.

One of these biochemical studies (Whitfield & Martin, 1983) was an early example of integration of a genetic marker into a twin study. It had been known for a long time that serum alkaline phosphatase activity is affected by the ABO and Lewis blood groups, and ABO grouping was one of the tests used to confirm self-reported zygosity in the twin pairs. About 15% of the genetic variance in alkaline phosphatase activity was associated with ABO type — still a large effect even in the GWAS era, and the ABO locus has turned out to be significant (for reasons which are not clear) in GWAS of many phenotypes.

Another variation on twin studies was the use of MZ pairs, and those who participated twice, to assess postulated genetic effects on sensitivity to environmental variation. The hypothesis (Magnus et al., 1981) was that some variants, which might or might not affect mean values for a phenotype, would affect the response of the phenotype to environmental variation. By genotyping MZ twin pairs for the genetic variant (the MN blood group), and measuring the phenotype (cholesterol) in each twin or in the same person on more than one occasion, it would be possible to test the hypothesis that within-pair or within-person differences would be associated with genotype. Such gene–environment interaction would be of considerable importance if, say, some people obtained benefit from change in diet and others did not. As so often occurs, the original hypothesis was not strongly supported by results (Martin et al.,

Author for correspondence: John B. Whitfield, Email: John.Whitfield@qimrberghofer.edu.au

Cite this article: Whitfield JB. (2020) Genetics of Biochemical Phenotypes. *Twin Research and Human Genetics* **23**: 77–79, https://doi.org/10.1017/thg.2020.26

1983), but a slightly different one (of effects on triglycerides) emerged. Subsequent multicentre data (Surakka et al., 2012) suggested that gene-by-environment interaction for lipid levels might exist, with a just-significant result (but for a different locus and phenotype) from genomewide testing. I mention these studies as an example of an attractive hypothesis, worth some effort to test, not being supported in practice. More generally, G × E interaction has only been shown infrequently despite the large amount of GWAS data now available.

Candidate Genes, Linkage

Association studies involving candidate genes have proved to be a trap, and it is widely accepted that they have led to many false positives through lack of consideration of the multiple testing problem when claiming significant results. The positive aspect has been an increased awareness of the need to set stringent p values in genomewide studies and, as far as possible, to replicate results in independent cohorts. Linkage studies for quantitative phenotypes such as biochemical test results have mostly failed for a different reason, because the effect sizes (with few exceptions) are too small to be detectable. Our experience with candidate genes and linkage generally followed this pattern.

One successful candidate gene study was to evaluate the effects of variation at the homeostatic iron regulator (*HFE*) gene, newly found to be necessary (but not sufficient) for hemochromatosis, on serum iron and related measures of iron status in the general population (Whitfield et al., 2000). This integrated *HFE* genotype information with the twin study method and showed that although *HFE* variants had significant effects on iron status, they only accounted for a small proportion of the genetic variance — an early example of missing heritability.

Because we had suitable data on related study participants, at first DZ twin pairs and later nontwin siblings, we made a number of attempts to identify loci affecting lipids through linkage analysis, but association analyses soon displaced linkage. One successful attempt was for serum butyrylcholinesterse, where a linkage peak was found on chromosome 3, overlapping the *BCHE* gene location. This was later substantiated by GWAS, but it should be admitted that linkage also identified a peak on chromosome 5 which did not show association in the later GWAS.

Blood Lead, from h^2 to GWAS

Lead is toxic and widely distributed in the environment, largely because of human mining and industrial processes including previous use in house paints and as a petrol additive. It has been implicated in a range of phenomena from the fall of the Roman empire (now largely refuted; see Retief & Cilliers, 2006) to childhood behavior disorders and educational achievement (for which there is strong evidence of association; Bellinger, 2008, but many potential confounders that make causation uncertain). Because of the presence of lead in the environment, it was taken for granted that variation in blood lead would be 'environmental' rather than 'genetic'. A series of papers using data from twins and their relatives gave a different perspective.

First, the classical twin method (Whitfield et al., 2010) showed evidence for substantial heritability of blood lead concentration in adults ($h^2 \approx 40\%$), with no significant shared environment effect. Linkage analysis suggested that a region of chromosome 3 contained a variant affecting blood lead. This extensive region includes the solute carrier 4 member 7 (*SLC4A7*) gene, which codes for a transporter affecting lead influx into erythrocytes, which was very encouraging, but this linkage result was not supported by later GWAS results.

Given the evidence for heritability and the possible localization of a variant having substantial effects on blood lead, the next step was to conduct a GWAS. This was done in collaboration with Dave Evans and used the Avon Longitudinal Study of Parents and Children (ALSPAC) cohort from the UK in addition to our data. It found one significant locus (aminolevulinate dehydratase [*ALAD*]) using a combined sample size of 5400 people. However, there was no evidence for significant association at the chromosome 3 linkage region.

This example is important because it follows the stages of genetic investigation from heritability (of a seemingly environmental phenotype) to GWAS, with a diversion through linkage on the way. If it ever becomes possible to gather more data, a much larger GWAS for blood lead should identify more loci and permit the use of Mendelian randomization to assess whether associations between lead and childhood development are causal.

GWAS — Heart, Kidney, Liver

The panel of routine diagnostic tests which we ran on blood samples from twins and their families covered a number of organ systems or areas of risk — lipids for heart disease, creatinine, urea and uric acid for kidney function, enzyme tests for liver function, C-reactive protein (CRP) for inflammation.

Although we accumulated biochemical data on around 17,000 adults (mostly with genotyping), the main value of this dataset came from collaboration with other groups who had similar data and from meta-analysis. Through these collaborations, sample sizes in the hundreds of thousands could be achieved, and discovery of significant variants has been far beyond what any single group could have managed (Ligthart et al., 2018; Tin et al., 2019; Willer et al., 2013; Wuttke et al., 2019). More importantly than listing significant variants, our data contributed to insights such as the causal role of triglycerides in coronary artery disease (Do et al., 2013); confirmation that most loci associated with kidney function assessed from creatinine results are also associated with urea and with diagnosed chronic kidney disease (Wuttke et al., 2019); and that genes containing variants that affect C-reactive protein concentration cluster in two groups, representing immune and metabolic pathways (Ligthart et al., 2018).

GWAS — Other Phenotypes

Apart from the widely available tests mentioned above, we measured a number of other biochemical phenotypes. Despite the limitations imposed by limited numbers (our studies plus one or just a few others), several important and/or interesting associations have been found.

- As well as blood lead (discussed above), the method for lead estimation also gave results for six other toxic or essential elements in blood cells (As, Cd, Cu, Hg, Se, Zn). These also showed significant heritability, and there was a notable genetic correlation between concentrations of As and Hg ($r_G = .83$, whereas $r_E = .34$). GWAS for the essential elements, and meta-analysis with similar data from the ALSPAC cohort, showed a number of significant loci for Cu, Se and Zn with, in many cases, probable explanations in terms of gene functions (Evans et al., 2013).
- An early study on iron and *HFE* genotypes was mentioned above. This was expanded to GWAS with our own data and then to meta-analysis of GWAS data from multiple groups, which

https://doi.org/10.1017/thg.2020.26 Published online by Cambridge University Press

included almost 50,000 participants. Eleven loci were identified as significant for one or more of the markers of iron status (Benyamin et al., 2014), and because of the biological importance of iron and its potential to cause tissue damage there have been a number of attempts to use the relevant genotypes as instrumental variables to test whether associations between iron and disease are causal.

- Plasma cholinesterase (butyrylcholinesterase, BCHE) is an enzyme whose activity is associated with obesity and other aspects of metabolic syndrome, but its function and the reasons for these associations are unknown. Because BCHE measurement was included in our test profile, we carried out a GWAS with the expectation that identification of genes affecting BCHE variation would shed light on its function and relationships with other phenotypes. By far the strongest associations were within or near the *BCHE* gene, and other significant loci were not associated with metabolic risk factors. On the other hand, Single Nucleotide Polymorphisms (SNPs) in genes associated with metabolic risk tended to have effects on BCHE, suggesting that BCHE variation is a consequence of metabolic abnormalities.
- Carbohydrate-deficient transferrin (CDT) comprises transferrin isoforms that have fewer than the usual four terminal sialic acid residues on their glycan sidechains, and their relative concentration in serum is increased in people with high alcohol intake. Because of our interest in markers of alcohol use, and as an example of variation affecting protein glycosylation, we conducted a GWAS for CDT (Kutalik et al., 2011). This identified two loci, the transferrin (*TF*) gene itself, and the phosphoglucomutase 1 (*PGM1*) gene, which catalyses an early step in synthesis of the carbohydrate side chains, showing that variation in both the protein structure and in formation of the glycan component can affect the product.
- Proteolytic cleavage of chromogranins leads to formation of a number of bioactive peptides including catestatin, which has a role in control of blood pressure. Collaboration with Dan O'Connor, the major player in study of chromogranins and related peptides, led us through heritability, linkage and GWAS stages to discovery of two loci affecting catestatin formation (Benyamin et al., 2017). Each locus contained a gene for a proteolytic enzyme involved in the intrinsic pathway of coagulation, and review of published literature showed that this process is important for formation of several peptide hormones from their precursors.

Conclusions

Studies on the genetics of biomarkers carry the expectation that because the biomarkers are associated with disease, results will be translatable to the genetics of disease. GWAS results in general may give insight into the mechanisms that regulate or influence the phenotype; they can (depending on the genetic architecture and on study size) predict the phenotype of an individual or stratify their risk of disease; and they can establish or refute causal relationships between apparent risk factors and disease. Genetic studies on biochemical phenotypes have grown and developed over the past 40 years, from 412 participants in our early twin studies to over a million in recent collaborative meta-analyses. It should be remembered that the justification for mega-GWAS studies came from initial, smaller GWAS, and the justification for the initial GWAS usually came from the knowledge that the phenotypes had significant heritability.

References

Bellinger, D. C. (2008). Very low lead exposures and children's neurodevelopment. *Current Opinion in Pediatrics*, *20*, 172–177.

Benyamin, B., Esko, T., Ried, J. S., Radhakrishnan, A., Vermeulen, S. H., Traglia, M., … **Whitfield, J. B.** (2014). Novel loci affecting iron homeostasis and their effects in individuals at risk for hemochromatosis. *Nature Communications*, *5*, 4926.

Benyamin, B., Maihofer, A. X., Schork, A. J., Hamilton, B. A., Rao, F., Schmid-Schonbein, G. W., … **O'Connor, D. T.** (2017). Identification of novel loci affecting circulating chromogranins and related peptides. *Human Molecular Genetics*, *26*, 233–242.

Do, R., Willer, C. J., Schmidt, E. M., Sengupta, S., Gao, C., Peloso, G. M., … **Kathiresan, S.** (2013). Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nature Genetics*, *45*, 1345–1352.

Evans, D. M., Zhu, G., Dy, V., Heath, A. C., Madden, P. A., Kemp, J. P., … **Whitfield, J. B.** (2013). Genome-wide association study identifies loci affecting blood copper, selenium and zinc. *Human Molecular Genetics*, *22*, 3998–4006.

Kutalik, Z., Benyamin, B., Bergmann, S., Mooser, V., Waeber G., Montgomery, G. W., … **Whitfield, J. B.** (2011). Genome-wide association study identifies two loci strongly affecting transferrin glycosylation. *Human Molecular Genetics*, *20*, 3710–3717.

Ligthart, S., Vaez, A., Vosa, U., Stathopoulou, M. G., de Vries, P. S., Prins, B. P., … **Alizadeh, B. Z.** (2018). Genome analyses of >200,000 individuals identify 58 loci for chronic inflammation and highlight pathways that link inflammation and complex disorders. *American Journal of Human Genetics*, *103*, 691–706.

Magnus, P., Berg, K., Borresen, A. L., & Nance, W. E. (1981). Apparent influence of marker genotypes on variation in serum cholesterol in monozygotic twins. *Clinical Genetics*, *19*, 67–70.

Martin, N. G., Perl, J., Oakeshott, J. G., Gibson, J. B., Starmer, G. A., & Wilks, A. V. (1985). A twin study of ethanol metabolism. *Behavior Genetics*, *15*, 93–109.

Martin, N. G., Rowell, D. M., & Whitfield, J. B. (1983). Do the MN and Jk systems influence environmental variability in serum lipid levels? *Clinical Genetics*, *24*, 1–14.

Retief, F. P., & Cilliers, L. (2006). Lead poisoning in ancient Rome. *Acta Theologica*, *26*, 147–164.

Risch, N., & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, *273*, 1516–1517.

Surakka, I., Whitfield, J. B., Perola, M., Visscher, P. M., Montgomery, G. W., Falchi, M., … **GenomEUtwin Project**. (2012). A genome-wide association study of monozygotic twin-pairs suggests a locus related to variability of serum high-density lipoprotein cholesterol. *Twin Research and Human Genetics*, *15*, 691–699.

Tin, A., Marten, J., Halperin Kuhns, V. L., Li, Y., Wuttke, M., Kirsten H, … **Köttgen A.** (2019). Target genes, variants, tissues and transcriptional pathways influencing human serum urate levels. *Nature Genetics*, *51*, 1459–1474.

Whitfield, J. B., & Martin, N. G. (1983). Determinants of variation in plasma alkaline phosphatase activity: a twin study. *American Journal of Human Genetics*, *35*, 978–986.

Whitfield, J. B., Cullen, L. M., Jazwinska, E. C., Powell, L. W., Heath, A. C., Zhu, G., … **Martin, N. G.** (2000). Effects of HFE c282y and h63d polymorphisms and polygenic background on iron stores in a large community sample of twins. *American Journal of Human Genetics*, *66*, 1246–1258.

Whitfield, J. B., Dy, V., McQuilty, R., Zhu, G., Heath, A. C., Montgomery, G. W., & Martin, N. G. (2010). Genetic effects on toxic and essential elements in humans: arsenic, cadmium, copper, lead, mercury, selenium and zinc in erythrocytes. *Environmental Health Perspectives*, *118*, 76–82.

Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., … **Global Lipids Genetic Consortium.** (2013). Discovery and refinement of loci associated with lipid levels. *Nature Genetics*, *45*, 1274–1283.

Wuttke, M., Li, Y., Li, M., Sieber, K. B., Feitosa, M. F., Gorski, M., … **Pattaro, C.** (2019). A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nature Genetics*, *51*, 957–972.

https://doi.org/10.1017/thg.2020.26 Published online by Cambridge University Press