
Genome-Wide Association Studies and Diet

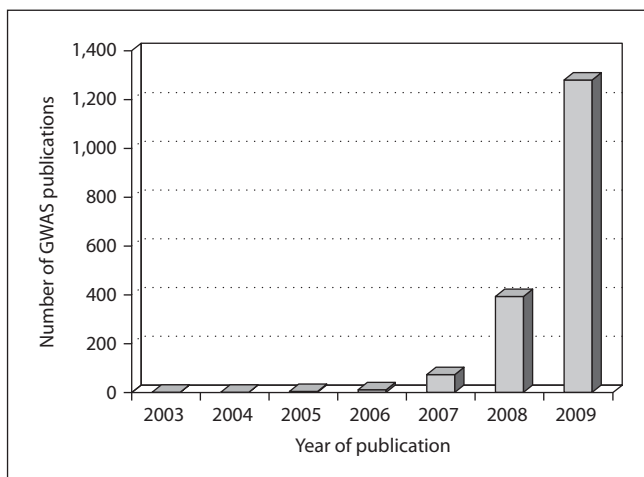
Lynnette R. Ferguson

Discipline of Nutrition, FM & HS, University of Auckland, and Nutrigenomics New Zealand, Auckland, New Zealand

Towards the end of the 20th century, we were successfully beginning to understand part of the genetic basis of some human diseases. Up to that time, progress had been relatively slow, largely depending upon establishing familial associations, and somewhat laboriously measured variations in candidate genes. This was usually in the form of single nucleotide polymorphisms (SNPs), measured one at a time with labor-intensive methods such as restriction fragment length polymorphism [1]. However, since the initial publication of work on the human genome [2], major advances in genotyping capability and reductions in cost, coupled with large collaborative population groups are enabling exponential advances in our understanding of human genetic variation. Genome-wide association studies (GWAS) are greatly increasing our understanding of the genetic basis of human disease, especially complex disease. Perhaps more importantly, they are more generally enhancing our knowledge of far more subtle differences between individuals, including behavioral characteristics, health, 'wellness' and performance. An analysis of GWAS publications since their first appearance in 2003 (fig. 1) emphasizes why Pennisi [3] described such studies as the breakthrough of the year in enabling knowledge of what makes each of us unique.

Since the earliest GWAS papers, there have been a number of commentaries in high-impact scientific journals, including a supplement in *Nature* (October 8), described as 'Human Genetics 2009'. The editorial points to the enhanced flow of human genetic information and the way that high density gene chip information is being utilized by online direct-to-consumer companies who claim to be predicting human health at the individual level. What is not being considered, however, is the key information that may be necessary to enable genetic data to predict human health or, more importantly, to develop strategies to modify the genetic predictions. That is, a parallel and integrated assessment of diet and environmental factors. In 2005, we commented on the need for international collaborative efforts in nutrigenomics to enable better understanding of the basis of human disease and 'wellness' characteristics that differ among individuals [4]. However, while there has been a proliferation

Fig. 1. Number of genome-wide association studies reported each year since 2003 in a PubMed literature search.



of studies on the genetic basis of disease (fig. 1), for the most part these have not been accompanied by stringent dietary and environmental data. It is suggested that this should become an essential input to such studies in the future.

Monogenic Disorders and Complex Disease

For much of the 20th century, knowledge of the genetic basis of human disease was limited to single-gene or Mendelian disease, where familial association is somewhat obvious. Most of these mutations are in the form of SNPs involving either missense or nonsense mutations [5]. Once the disease was accurately phenotyped and familial associations identified, it became relatively clear which gene was important. However, such diseases are relatively rare, and most of the common diseases are more complex, involving multiple genes and interactions with environment, including diet.

With complex disease, association studies are the only realistic approach, using large numbers of unrelated cases and controls, or family groupings such as trios involving 2 parents and an affected child [6]. Large numbers of markers that cover the genome are required to identify genes in complex disease. Perhaps more importantly, the presence or absence of a single gene variant is not usually sufficient to explain the disease phenotype. There is good reason to believe that genetic predisposition to complex disease is due to minor variations in a large number of genes, and their ability to interact with specific environmental factors. While these complex diseases are much more difficult to study, this knowledge may become increasingly important, since they are the most common cause of death in humans [7, 8]. There are, of-course, even more technical challenges in gaining an effective understanding of the human diet than of genes [9].

Enabling Technologies in GWAS

Although the first GWAS appeared in the literature in 2003, the initial tools did not cover a representative area of the human genome. In 2004, Ishkanian et al. [10] described a tiling resolution DNA microarray that they described as showing ‘complete coverage’ of the human genome. Such tools are essential starting blocks for GWAS studies, which are based on enabling genotype-phenotype correlations, using the same principles as candidate gene studies. However, such studies are hypothesis free, since the variants measured span the entire human genome. The international HapMap project recognized the need for studies of this sort, and sought to characterize the major SNPs across the genome, and in different human population groups [11]. The notion was that these would provide an idea of population structure on all ‘common’ SNPs (>5% frequency), in 2 phases of increasing density across the genome. This would be supplemented by deep re-sequencing as appropriate. This resource has provided the enabling technology for genetic variant assays using gene chips, now able to cover more than 1 million SNPs across the genome. The 2 main genotyping providers are Affymetrix, whose variants are randomly distributed, and Illumina, who have utilized more highly selected tagging SNPs. Either or both of these gene chips are ideal to measure a large number of SNPs and also copy number variants [12]. Deep re-sequencing techniques are also available for interrogating specific areas of the genome. Large collaborative databases are essential for providing the necessary statistical power for confidence in data interpretation.

GWAS: Why Are They Important?

GWAS provide an important mechanism for moving away from candidate gene studies, which select genes for study based on known or suspected disease mechanisms. Instead, GWAS permit a comprehensive scan of the genome in an unbiased fashion. By this means they have drawn out associations with genes not previously suspected of being related to the disease. They permit examination of inherited genetic variability at unprecedented levels of resolution, and have even picked up some associations in regions not even known to harbor genes. The methods continue to be refined. In their 2009 review, Ioannidis et al. [13] recommend large-scale exact replication across both similar and diverse populations, fine mapping and resequencing, determination of the most informative markers and multiple independent informative loci, incorporation of functional information, and improved phenotype mapping of the implicated genetic effects. Even where replication proves that an effect exists, definitive identification of the causal variant is often elusive. While these are all important points, it is of concern that even the excellent Ioannidis et al. review fails to consider diet as one of the missing variables.

Crohn's disease provides a good example of the power of this methodology. Candidate gene studies very slowly uncovered some of the genetic basis for this disease, with an initial report on the first disease gene, Nucleotide oligomerisation domain 2 (*NOD2*) in 2001 [14]. Other genes were slowly and sometimes unconvincingly revealed, including other immune recognition genes such as Toll-like receptor 4 (*TLR4*) [15]. However, the first publications of GWAS on this disease [16, 17] revealed the importance of SNPs in hitherto unsuspected genes, including the interleukin 23 receptor, *IL23R*, and the autophagy gene, *ATG16L1*. These genes both involve response to environmental factors, especially bacteria and diet. GWAS methodology continues to yield important findings on the genetic basis of this disease [18]. However, studies on dietary interactions are substantially lagging behind the genetic evidence.

Use of Gene Chips and GWAS Datasets in Personalized Health Predictions

The publication of this burgeoning number of datasets has led to a proliferation of online genetic testing companies, which purport to provide a measure of genetic risk to an individual, who has provided saliva or buccal swab samples for DNA isolation and genotyping. Inevitably, there has been concern expressed about their relevance. For example, Ng et al. [19], compared data provided by 2 different direct-to-consumer genetics-testing companies on a small number of individuals, to find quite significant differences in the predictions claimed. For their 5 test individuals, in predictions of 7 diseases only 50% or fewer of the predictions agreed between the 2 companies. Their information showed that the accuracy of the raw data was high. However, they questioned whether the predicted disease risks had clinical validity, and how well a genetic variant correlates with a specific disease or condition. They found that the companies showed very similar predictions for diseases where the genetic risk was convincing, and they concluded that companies should communicate high risks better than they are currently doing. They also suggested that test data would become more relevant to human health if the companies tested for drug response markers. They pointed to differences between the genetic basis of disease in different ethnicities, and suggested that the information gathered should include consideration of behavior. Surprisingly, however, they failed to highlight the potential importance of diet and/or environment. Their 9 recommendations are reproduced in table 1.

Celiac disease provides an example where Ng et al. [19] showed good agreement between direct-to-consumer testing companies. Celiac disease represents a major food intolerance, with a current prevalence rate of approximately 1 per 100 individuals in the population [20]. This disease is characterized by a lifelong intolerance to gluten, which is found in wheat, barley and rye, and products derived from them. The most effective treatment for celiac disease is nutritional [21] and remission of symptoms can be well maintained in the absence of gluten. At present, this disease is usually diagnosed phenotypically, once symptoms have developed, and requires an invasive

Table 1. Recommendations for improvement of direct-to-consumer genetic testing, as identified by Ng et al. [19].

Addressed to	Recommendation
Industry	Report the genetic contribution for the markers tested Focus on high-risk predictions Directly genotype risk markers Test pharmacogenomic markers Agree on strong-effect markers
Community	Monitor behavioral outcomes Carry out prospective studies Replicate associated markers in other ethnicities Sequence rather than genotype

small intestinal biopsy for diagnosis. However, twin studies provide good evidence for a genetic basis of the disease, with 10% of first-degree relatives being affected, and 75% concordance between monozygotic twins [22]. Several genes are clearly involved, but the most consistent genetic component depends on the variants in HLA-DQ (DQ2 and/or DQ8) genes [23]. The main genes in celiac disease lead to a 7-fold increased disease risk, and can be diagnosed fairly consistently, either on a phenotype or genotype level. More recent GWAS also give insight into the other relevant genes for this disease [24]. There would seem to be a case for earlier genetic diagnosis in susceptible families, avoiding the inevitable suffering associated with the presence of symptoms.

Where slightly different interpretations occurred was where one company had kept up with the very latest literature but other companies had not. Crohn's disease provided an example where there were inconsistencies in diagnosis. Even though this disease has proved remarkably tractable in GWAS studies, with very strong probabilities of accurate diagnosis of genes [25], individually these have very low relative risks. This may suggest the importance of environmental interactions.

Gene-Diet Interactions: Crohn's Disease

The genetic basis of Crohn's disease has not been as easy to characterize as celiac disease. The genetic basis of the disease is, as with celiac disease, supported by twin studies. For example, Tysk et al. [26] have shown strong familial associations and around 44% concordance between monozygotic twins. Although key genes have been revealed by GWAS and other approaches, the odds ratios associated with individual risk alleles are not spectacular [25]. Furthermore, although key dietary items have been revealed, unlike celiac disease, there is no 'one size fits all' solution. For example, in our own studies, wheat products, dairy foods, red wine, corn, mushrooms, soy

milk and yoghurt are all examples of foods for which a number of individuals with the disease report an exacerbation of risk. However, there are also a proportion of individuals who consistently report that they appear to benefit from regularly eating one or more of these foods. We have been able to demonstrate that at least some of these apparently inconsistent data, for example with mushrooms, are a result of gene-diet interactions [27]. In this example, a genetic variant in a solute transporter molecule, OCTN1, appeared important for Crohn's disease risk in some overseas populations; it did not show statistically significant association with disease risk in a New Zealand population. However, when the ability to tolerate mushrooms was factored in, those individuals with strong mushroom intolerance showed significantly enhanced levels of the OCTN1 variant as compared with the control population.

The experience with Crohn's disease leads to caution in interpreting dietary information in such a complex disease. There is a considerable effort to increase the sensitivity and accuracy of dietary information [9]. However, more accurate dietary questionnaires reveal a typical eating pattern for the individual. From an analysis of such data for an individual with Crohn's disease, one might conclude that a deficiency of wheat products, dairy foods, red wine, corn, mushrooms, soy milk and yoghurt has led to the development of the disease. However, the actual picture is likely to be the complete converse of this. The observation is that, when he or she actually eats these dietary items, disease symptoms develop, and thus the individuals learn to avoid foods that trigger symptoms. This means that it is the presence rather than the absence of these items that actually led to the establishment of symptoms of the disease. This is the complete converse of traditional dietary interpretation, and may lead the way to different thinking about dietary studies in association with GWAS. Certainly, effective methods are increasingly becoming available [28]. However such studies are performed, it is essential that they are done if we are to uncover the true role of genetic variants, and the interplay with diet and environment, in the etiology of complex disease.

Acknowledgments

NutriGenomics New Zealand is a collaboration between AgResearch Ltd., Plant & Food Research and the University of Auckland, with funding through the Foundation for Research Science and Technology.

References

- 1 Breen G, Harold D, Ralston S, Shaw D, St Clair D: Determining SNP allele frequencies in DNA pools. *Biotechniques* 2000;28:464–466.
- 2 Venter JC, Adams MD, Myers EW, et al: The sequence of the human genome. *Science* 2001; 291:1304–1351. Erratum: *Science* 2001;292: 1838.
- 3 Pennisi E: Breakthrough of the year: Human genetic variation. *Science* 2007;318:1842–1843.
- 4 Kaput J, Ordovas JM, Ferguson L, et al: The case for strategic international alliances to harness nutritional genomics for public and personal health. *Br J Nutr* 2005;94:623–632.

- 5 Stenson PD, Ball EV, Mort M, et al: Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* 2003;21:577–581.
- 6 Nowotny P, Kwon JM, Goate AM: SNP analysis to dissect human traits. *Curr Opin Neurobiol* 2001;11: 637–641.
- 7 Bell J: Predicting disease using genomics. *Nature* 2004;429:453–456.
- 8 Frazer KA, Murray SS, Schork NJ, Topol EJ: Human genetic variation and its contribution to complex traits. *Nat Rev Genet* 2009;10: 241–251.
- 9 Poslusna K, Ruprich J, de Vries JH, Jakubikova M, van't Veer P: Misreporting of energy and micronutrient intake estimated by food records and 24 hour recalls, control and adjustment methods in practice. *Br J Nutr* 2009;101(suppl 2):S73–S85.
- 10 Ishkanian AS, Malloff CA, Watson SK, et al: A tiling resolution DNA microarray with complete coverage of the human genome. *Nat Genet* 2004;36:299–303.
- 11 Couzin J: Human genome: HapMap launched with pledges of \$100 million. *Science* 2002;298:941–942.
- 12 Shelling AN, Ferguson LR: Genetic variation in human disease and a new role for copy number variants. *Mutat Res* 2007;622:33–41.
- 13 Ioannidis JP, Thomas G, Daly MJ: Validating, augmenting and refining genome-wide association signals. *Nat Rev Genet* 2009;10:318–329.
- 14 Ogura Y, Bonen DK, Inohara N, et al: A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* 2001;411:603–606.
- 15 Browning BL, Huebner C, Petermann I, et al: Has Toll-like receptor 4 been prematurely dismissed as an inflammatory bowel disease gene? Association study combined with meta-analysis shows strong evidence for association. *Am J Gastroenterol* 2007; 102:2504–2512.
- 16 Duerr RH, Taylor KD, Brant SR, et al: A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* 2006;314: 1461–1463.
- 17 Cummings JRF, Cooney R, Pathan S, et al: Confirmation of the role of ATG16L1 as a Crohn's disease susceptibility gene. *Inflamm Bowel Dis* 2007;13:941–946.
- 18 Marquez A, Cenit MC, Nunez C, et al: Effect of BSN-MST1 locus on inflammatory bowel disease and multiple sclerosis susceptibility. *Genes Immunity* 2009;10:631–635.
- 19 Ng PC, Murray SS, Levy S, Venter JC: An agenda for personalized medicine. *Nature* 2009;461:724–726.
- 20 McGough N, Cummings JH: Coeliac disease: a diverse clinical syndrome caused by intolerance of wheat, barley and rye. *Proc Nutr Soc* 2005;64:434–450.
- 21 Greco L, Romino R, Coto I, et al: The first large population based twin study of coeliac disease. *Gut* 2002;50:624–628.
- 22 Di Gilio A, Fusco M, Mazzacca G: Grown-up coeliac children: the effects of only a few years on a gluten-free diet in childhood. *Aliment Pharmacol Ther* 2005;21:421–429.
- 23 Romanos J, van Diemen CC, Nolte IM, et al: Analysis of HLA and non-HLA alleles can identify individuals at high risk for celiac disease. *Gastroenterology* 2009;137:834–840.
- 24 Garner CP, Murray JA, Ding YC, et al: Replication of celiac disease UK genome-wide association study results in a US population. *Hum Mol Genet* 2009; 18:4219–4225.
- 25 Wellcome Trust Case Control Consortium: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007;447:661–678.
- 26 Tysk C, Lindberg E, Järnerot G, Flodérus-Myrhed B: Ulcerative colitis and Crohn's disease in an unselected population of monozygotic and dizygotic twins: a study of heritability and the influence of smoking. *Gut* 1988;29:990–996.
- 27 Petermann I, Triggs CM, Huebner C, et al: Mushroom intolerance: a novel diet-gene interaction in Crohn's disease. *Br J Nutr* 2009;102:506–508.
- 28 Bureau A, Diallo MS, Ordovas JM, Cupples LA: Estimating interaction between genetic and environmental risk factors: efficiency of sampling designs within a cohort. *Epidemiology* 2008;19:83–93.

Lynnette R. Ferguson
 Discipline of Nutrition, FM&HS, University of Auckland,
 Auckland (New Zealand)
 Tel. +64 9 373 7599, Fax +64 9 303 5963, E-Mail l.ferguson@auckland.ac.nz