Getting biological about the genetics of diabetes

Christopher B Newgard & Alan D Attie

New technology has provided methods for collecting large amounts of data reflecting gene expression, metabolite and protein abundance, and post-translational modification of proteins. Integration of these various data sets enable the genetic mapping of many new phenotypes and facilitates the creation of network models that link genetic variation with intermediate traits leading to human disease. The first round of genome-wide association studies has not accounted for common human diseases to the extent that was expected. New phenotyping approaches and methods of data integration should bring these studies closer to their promised goals.

In model organisms, a 'sensitized screen' is one in which a stress is applied (for example, a gene knockout or exposure to an inhibitor), followed by a second hit (such as random mutagenesis). The phenotypes that arise from these two hits would not have arisen without the initial sensitization step. For example, inhibition of cholesterol esterification makes cells far more susceptible to mutations that compromise cholesterol trafficking[1]. Twenty-five years ago, the US population began participating in a sensitized screen. The proportion of the population that is obese or overweight has doubled in that time frame, owing to changes in diet and other environmental factors. Obesity is now associated with type 2 diabetes, gallstones, hepatic steatosis, dyslipidemias, hypertension, cardiovascular disease, osteoarthritis, stroke, sleep disorders, polycystic ovary disease and various types of cancer.

Genetic factors have a strong influence on the susceptibility of obese individuals for development of these associated diseases and conditions. For example, particular racial groups (Hispanics, Native Americans and South Asians) have a lower body fat threshold for risk of type 2 diabetes than do Caucasians. However, unlike diseases such as cystic fibrosis or muscular dystrophy, in which a single gene is causative, or the rare monogenic forms of diabetes, known collectively as maturity onset diabetes of the young, obesity-related diseases, such as type 2 diabetes are polygenic in origin[2]. Thus, the specific genetic factors that link obesity to any of its comorbidities, including the most prevalent one—type 2 diabetes—remain largely undefined. This is in part because of the lack of precision and depth of the biological methods that have been used to understand the impact of genetic variability on phenotype.

Human genetics

The increased prevalence of obesity comorbidities has motivated numerous large-scale genome-wide association studies (GWASs). These studies have identified more than 30 genes that are associated with type 2 diabetes[3]. Most of these genes seem to affect beta cell function and/or mass rather than insulin signaling or peripheral metabolism[4], although there are notable exceptions, such as the recent discovery of a gene associated with hepatic steatosis in Hispanic populations[5]. Although informative, the results from GWASs have been disappointing in two respects. First, the sum of all the loci that were identified account for only a small part of the variation in the phenotype across the population. Second, the increased risk for disease of the susceptibility allele relative to the 'normal' allele (that is, the odds ratio) is small. For example, the strongest odds ratio, associated with variants of the gene encoding transcription factor 7–like 2 (*TCF7L2*), is just ~1.5 (ref. 6).

GWASs have limitations that might explain these results. With conventional high-throughput genotyping methods, these studies cannot detect rare disease alleles. In cases where extensive resequencing of genes associated with metabolic diseases has been carried out, multiple rare and common alleles, to varying degrees, seem to be responsible for disease susceptibility across the population[7–9]. GWASs require testing of many thousands of genetic markers to achieve a comprehensive coverage of the genome, increasing the likelihood of detecting genetic association even where none exists (false positives). This also results in a loss of statistical power to detect association between individual loci and disease phenotypes and, even worse, makes it difficult to explore the gene-gene interactions that are likely to contibute to complex diseases. In addition, in free-living populations it is difficult to control or account for the environment, making studies of epigenetics and their effect on disease occurrence difficult. Last, phenotypic analysis in humans can be conducted at only a minimally invasive level, owing to limited access to tissue samples and the inability to use methods commonly applied to model organisms, such as radioisotope tracer infusion.

Human disease genetics is also difficult to study in model organisms. The human population has a substantial amount of heterozygosity and therefore carries numerous harmful recessive alleles. The combinations of alleles in the population have undergone generations of selection in various environments and probably contain useful redundancies and biological buffering mechanisms. Understanding these processes is a key objective in biological research, but the redundancies and buff-

Christopher B. Newgard is at the Sarah W. Stedman Nutrition and Metabolism Center and Department of Pharmacology and Cancer Biology, Duke University Medical Center, Durham, North Carolina, USA. Alan D. Attie is in the Department of Biochemistry, University of Wisconsin–Madison, Madison,Wisconsin, USA.
e-mail: newga002@mc.duke.edu or adattie@wisc.edu

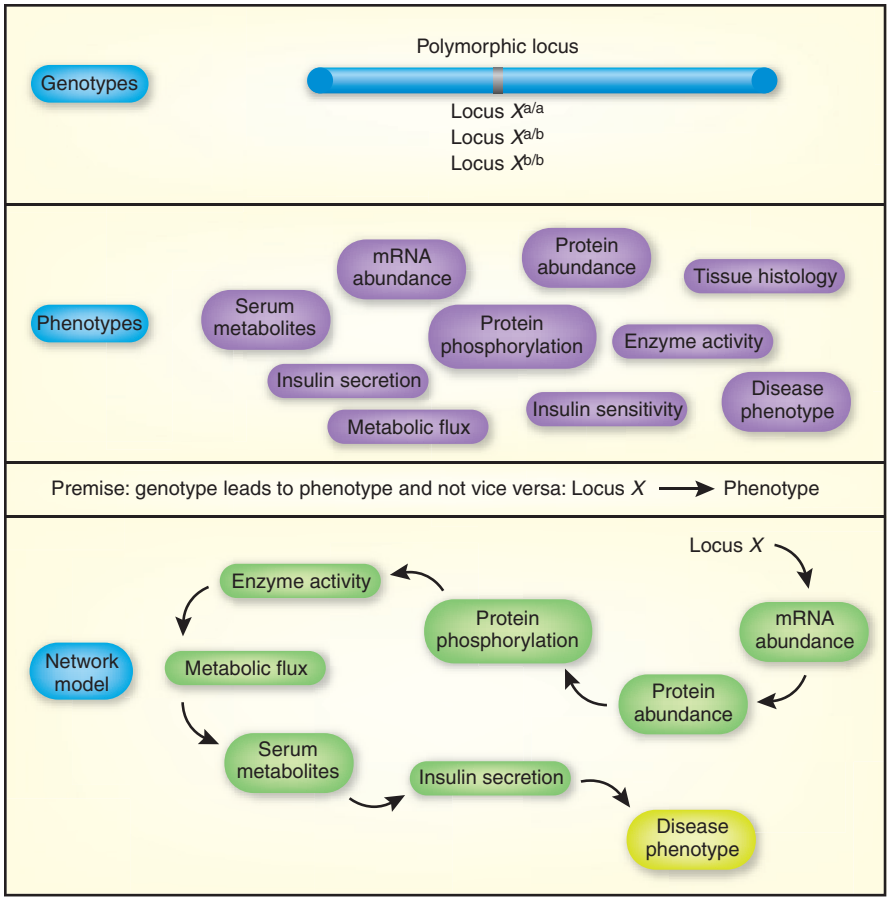

Figure 1 Constructing causal networks from genetic and molecular phenotyping data sets. A common problem with 'omics' data sets is that it is easy to compute many thousands of correlations but difficult to establish causal models. Because genetic variation gives rise to phenotype variation and not vice versa, genetic information (for example, whole-genome SNP analysis) provides an anchor point from which unidirectional causal network models can be generated to link genetic loci with changes in transcripts, proteins, metabolites and clinical variables, as shown here. Given the power of these models to integrate data from many types of sources, we predict that they will be crucial in understanding diseases such as type 2 diabetes, which result from the interactions between multiple genetic and environmental factors.

ering make the detection of heritability of phenotypes difficult. The recent emergence of advanced tools for measuring phenotypic variability, such as metabolomics, proteomics and gene expression analysis by microarray, coupled with the development of more sophisticated mouse genetics models, provides hope for advancement in this area.

Mouse genetics

Mouse genetics remedies some, but not all, of the problems of GWASs on humans. A typical mouse genetic study involves a two-way cross of inbred strains. In the offspring of these crosses, the two alleles at all polymorphic loci have a frequency of 50%, allowing their full impact on phenotype to be readily detected; in contrast, rare disease-causing alleles can be 'hidden' in the population in humans. Another advantage of working with inbred animals is that one can carry out multiple studies, exploring a wide range of phenotypes, with as much replication as is necessary for statistical power. As described below, these phenotypes can be integrated into network models that help to explain complex disease physiology and pathogenesis.

The study of the agouti lethal yellow mouse is an example of how genetics delivers an entirely new biological pathway to solve a riddle that would have been virtually impossible to solve with ordinary hypothesis-testing experimentation. In this case, the project identified a molecule that simultaneously blocks the action at two receptors, the melanocortin-1 receptor, which controls skin pigmentation, and the melanocortin-4 receptor, which controls food intake[10–12]. Mutations in the latter receptor are now the most prevalent known monogenic genetic cause of obesity in humans[13].

Likewise, the discovery of the hormone leptin shows how an insightful physiology experiment produced a hypothesis that then required a genetic study to advance the underlying physiology to a molecular mechanism. Parabiotic experiments posited that a circulating factor responsible for suppressing hunger is mutated in the *ob/ob* mouse and that a receptor for that factor is defective in the *db/db* mouse[14]. This prescient prediction was borne out by the discovery, through positional cloning, of leptin and its receptor[15,16]. Although these proteins are rarely responsible for human obesity[17], their profound role in energy balance and metabolism is now broadly recognized in humans and animal models[18]. So mouse genetics, although often used to find a direct cause of a phenotype, can also improve our understanding of complex biology.

Although many advances in the understanding of normo- and pathophysiology have been brought about by mouse genetics, its use has also exposed its limitations. For example, intercrosses between inbred mouse strains enable mapping of complex traits but only to low resolution. This is because allowing for just one or two generations of meiotic recombination produces large segments of the genome that are inherited as one block. To enhance resolution, mouse geneticists typically create congenic strains, in which particular genomic segments from one parental background are selected and retained through backcrossing into the other parental background. As these strains do not always replicate the genetic context in which the phenotype was mapped in the original intercross, phenotypes can be lost or altered in ways that were not initially predicted. Creation of inbred mouse strains has also resulted in selection against all recessive lethal alleles and may therefore not fully model important disease-causing alleles in humans or other outbred populations. To overcome this problem, outbred stocks have been used successfully to map disease traits to small genomic regions[19]. However, these studies suffer from the disadvantage of not enabling biological replication of results in genetically identical animals.

In an attempt to overcome some of the problems with conventional mouse genetics, a project termed 'Collaborative Cross' has been initiated. Eight founder strains are being intercrossed, and their offspring are being inbred to homozygosity, essentially producing recombinant inbred lines containing three to seven alleles[20]. Because of the number of generations of breeding, more recombinations will occur, thereby enabling higher-resolution mapping. As the mice will be fully inbred, they will allow for biological replication and, therefore, for the application of new tools for assessing molecular phenotypes, as described below. In short, the project will generate a highly diverse population

of mice that need to be genotyped just once to enable phenotypes to be mapped.

The hope is that a broad community of biologists will apply sophisticated phenotyping technologies to the unique resource provided by the Collaborative Cross. Indeed, it can be argued that mouse genetics has not been broadly embraced by many of the scientists who have the best capabilities for in-depth phenotyping. This includes physiologists, biochemists and pathologists who study processes such as metabolic regulation, fuel partitioning, signaling pathways, transcriptional and post-transcriptional regulatory processes, organelle biology and the many biological processes that reflect genetic diversity and contribute to human disease. The full potential of the Collaborative Cross will not be realized without the engagement of this full spectrum of biologists, coupled with skillful integration of the complex data sets that are likely to emerge.

Applying microarrays and metabolomics

For the full potential of new mouse genetics models to be realized, and to foster progress in our understanding of human disease, application of more precise and detailed methods for phenotypic analysis will be required. For example, the ability to interrogate the entire transcriptome with microarray technology has already had a profound effect. Comprehensive metabolic profiling, or metabolomics, is at an earlier stage of development, but it has been increasingly integrated with genetic analyses to gain new insights[21].

Microarray technologies have shown how defined conditions lead to coordinated regulation of clusters of genes, helping to identify the functions of unannotated genes or new functions of known genes. For example, microarray analysis of fat tissue from adipose-specific solute carrier family 2 (facilitated glucose transporter), member 4 (*Slc2a4*; also known as *Glut4*) knockout mice revealed upregulation of retinol binding protein-4 (RBP-4), a carrier protein for retinol in the circulation[22]. Subsequent studies showed that experimental elevation of RBP-4 concentrations in mice resulted in insulin resistance and that RBP-4 levels are elevated in insulin-resistant humans[22,23]. Similarly, microarray analysis of pancreatic islets from humans with type 2 diabetes compared to normal controls revealed a 90% decrease in expression of aryl hydrocarbon receptor nuclear translocator (ARNT; also known as HIF1b), and beta cell–specific knockout of ARNT in mice resulted in impaired insulin secretion, glucose intolerance and changes in beta cell gene expression[24]. Unfortunately, studies such as these, which move beyond the initial microarray analysis to actually pinpoint the biological effects of individual genes, are rare. This needs to change if comprehensive phenotyping is to make a true impact on our understanding of genetic diseases.

In other studies, changes in gene expression can be subtle when examined at a single-gene level, but they become substantial when analyzed in conjunction with other genes in a given biological pathway. An example is the finding of reduced expression of a group of mitochondrial oxidative phosphorylation genes in the muscle of subjects with diabetes[25,26]. Similarly, early applications of metabolomics technology to the diabetes field have led to a number of insights about the roles of individual metabolites or clusters of metabolites (defined by statistical tools such as principal components analysis) in disease processes. Examples are described in a recent review[21], and they include the demonstration of mitochondrial lipid overload in diet-induced obesity and insulin resistance[27], the identification of a specific fatty acid, palmitoleate (C16:1), with the capacity to enhance insulin sensitivity[28], the identification of *N*-acylphosphatidylethanolamines as unique satiety signals produced in the gut during ingestion of high-fat diets[29], the demonstration of a link between nutrient availability and regulation of the ghrelin-modifying enzyme ghrelin *O*-acyl transferase[30] and the identification of a branched-chain amino acid–related metabolite signature associated with human insulin resistance[31,32].

Combining genetics with comprehensive gene expression and metabolite profiling technologies enables the construction of causal networks. This takes advantage of the unidirectional flow of information from DNA variation to phenotype. Thus, the genotype at a marker can be correlated with a change in mRNA or metabolite abundance, but we might assume *a priori* that mRNA or metabolite abundance cannot affect genotype. Notably, when transcript and metabolite profiling is performed with rigor, mRNA levels and metabolite levels are heritable traits that can be mapped to specific regions of the genome in animals and humans as 'expression quantitative trait loci' or 'metabolite quantitative trait loci'. This provides an anchor point to develop causal network models in which a genomic region is proposed to influence the levels of transcripts, proteins and metabolites (**Fig. 1**).

As explained by Schadt[33], one way to view complex disease is that it is an "emergent property of networks;" that is, the interrelationships between various phenotypes (for example, mRNA abundance, metabolites or physiological traits) together constitute the functional unit that must be examined to understand the link to human disease. These methods were used to define a locus on mouse chromosome 1 that controls a wide range of metabolic syndrome traits (such as glucose, abdominal fat, plasma cholesterol and triglycerides)[34]. Using the expression data from an intercross, they constructed a network that connected the locus to these traits. The network is highly enriched in macrophage-derived inflammatory genes, several of which had already been causally associated with obesity traits—protein phosphatase 1–like (*Ppm1l*), lipoprotein lipase (*Lpl*) and lactamase-β (*Lactb*). Knockouts in all three of these genes produced a body-weight phenotype, validating the predictions of the network model.

A related approach to finding differences in the relationship between genes is to examine the correlation structure of the gene expression data[35,36]. By performing gene set correlation analysis, one can find changes in the correlations of genes with one another, which often includes genes that have not changed their absolute level of expression. This change in correlation structure can create a signature of a physiological state. This approach enables a gene to be associated with disease in a context-dependent fashion and can explain why its absolute association can be weak when interrogated in a GWAS.

When sets of genes respond together to a set of physiological conditions or to genetic variation, one can hypothesize that they are coregulated. In the context of diabetes, this approach has been used to investigate how beta cells can mount a proliferative response to overcome the demand for more insulin brought about by insulin resistance in obese animals. So, using a method that filters gene sets with a tight correlation structure[37], an obesity-regulated module of cell-cycle genes was identified in pancreatic islets[38]. Recent studies show that this module is genetically controlled by distinct quantitative trait loci (M. Keller, Y. Choi, P. Wang, D. David, M. Rabaglia *et al.*, unpublished observations).

Network models can be expanded to include metabolic phenotypes. For example, we have recently integrated microarray and metabolomics analyses of liver samples from mice derived from two generations of breeding of diabetes-resistant C57BL/6 *ob*/*ob* mice with diabetes-susceptible BTBR *ob*/*ob* mice[39]. Integration of these data sets with whole-genome analysis of single-nucleotide polymorphisms (SNPs), and the application of advanced statistical techniques[40], revealed correlations between genetic loci, transcripts and metabolites. This analysis predicted a previously undescribed metabolic regulatory network in which glutamine regulates the gene for the key gluconeogenic enzyme

phosphoenolpyruvate carboxykinase (*Pck1*) through intermediate nodes composed of the genes encoding the amino acid–metabolizing enzymes alanine-glyoxylate aminotransferase (*Agxt*) and arginase-1 (*Arg1*). Experimental support for this network came from the demonstration of strong upregulation of *Agxt*, *Arg1* and *Pck1* in response to addition of glutamine to mouse hepatocytes[39]. Also, glutamine concentrations and *Pck1* mRNA levels are increased in the livers of diabetes-susceptible BTBR-*ob*/*ob* animals and Zucker diabetic fatty rats relative to nondiabetic C57BL/6 *ob*/*ob* and Zucker rats, respectively (C. Ferrara, D. Gupta, A.D.A. and C.B.N., unpublished observations), suggesting that dysregulation of this network may contribute to poorly controlled gluconeogenesis and to the development of diabetes, at least in murine models.

In another study, nuclear magnetic resonance–based metabolite profiling was applied to second-generation rats from a cross of diabetic Goto-Kakizaki and normoglycemic Brown Norway rats and integrated with information about physiological quantitative trait loci in the same cross[41]. Glucose and benzoate were among the metabolites associated with the most significant quantitative trait loci. Subsequent transcriptomic analysis revealed that Goto-Kakizaki rats lack transcripts encoding an enzyme that metabolizes benzoate and other xenobiotics in mammals—UDP-glucuronosyl-transferase-2b (Ugt2b). The absence of Ugt2b expression was subsequently found to result from a chromosomal deletion in the Goto-Kakizaki strain, demonstrating the ability of metabolomics to uncover otherwise undetected chromosomal abnormalities.

Last, a recent study provides a glimpse of the early stages of the application of integrated phenotypic analysis tools to human genetic diseases and conditions. Mass spectrometry–based metabolic profiling was used to detect significant associations between frequent SNPs and changes in metabolites in human subjects[42]. Specifically, polymorphisms in genes encoding metabolic enzymes (fatty acid desaturase-1 (*FADS1*), hepatic lipase (*LIPC*) and acyl-CoA dehydrogenase, C-2 to C-3 short chain (*ACADS*; also known as *SCAD*) and medium chain (*ACADM*; also known as *MCAD*)) were linked to perturbations in the metabolic pathways in which the enzymes are known to reside. These recent examples in animal models and humans demonstrate the potential for understanding the full phenotypic consequence of genetic variation, thereby potentially leading to more informed and personalized disease therapeutic strategies.

Conclusions and opportunities

The development of high-throughput, high-resolution technologies for assessing genetic variation on a whole-genome scale has been rapid and impressive and will soon be improved further as new technologies such as RNA sequencing, deep sequencing at the genomic level and chromatin immunoprecipitation combined with sequencing (ChIP-Seq) become cheaper and more accessible to a broader number of investigators. However, the application of available genomic technologies to type 2 diabetes has revealed two fundamental concerns: polymorphic loci identified to date account for only a small part of the diabetes phenotype across the population, and obesity-related type 2 diabetes is probably explained by a combination of common and rare alleles, the corresponding phenotypes of which are influenced by the environment, and all of the elements of this dynamic are simply not captured without controlling or accounting for the environmental effects.

Here we have attempted to summarize the emergent strategies for combining GWASs with comprehensive transcriptomic and metabolomic phenotyping, as well as early examples of the integration of some of these tools for pathway identification and insights into disease pathogenesis. Metabolomics and transcriptomics now have the potential to redefine the term 'phenotype' as it applies to genetic studies. The need to apply multiple 'omics' tools is underscored by the fact that the variance explained by gene loci on mRNA abundance traits is likely to be far higher than on distal phenotypes such as blood analyte levels. This may also be true of surveys of protein abundance (proteomics).

We have therefore arrived at a point where advanced phenotyping methods are poised to provide more precise readouts of the impact of genetic variability on obesity-related diabetes pathogenesis and progression, both in more sophisticated mouse models such as the Collaborative Cross and in human populations. These tools have the potential to contribute to a fuller understanding of the genetic basis of type 2 diabetes.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Tinkelenberg, A.H. *et al. J. Biol. Chem.* **275**, 40667–40670 (2000).
2. O'Rahilly, S. *Nature* **462**, 307–314 (2009).
3. Zeggini, E. *et al. Nat. Genet.* **40**, 638–645 (2008).
4. Prokopenko, I., McCarthy, M.I. & Lindgren, C.M. *Trends Genet.* **24**, 613–621 (2008).
5. Romeo, S. *et al. Nat. Genet.* **40**, 1461–1465 (2008).
6. Tong, Y. *et al. BMC Med. Genet.* **10**, 15 (2009).
7. Cohen, J.C. *et al. Science* **305**, 869–872 (2004).
8. Kathiresan, S. *et al. Nat. Genet.* **41**, 56–65 (2009).
9. Romeo, S. *et al. Nat. Genet.* **39**, 513–516 (2007).
10. Lu, D. *et al. Nature* **371**, 799–802 (1994).
11. Michaud, E.J. *et al. Proc. Natl. Acad. Sci. USA* **91**, 2562–2566 (1994).
12. Bultman, S.J., Michaud, E.J. & Woychik, R.P. *Cell* **71**, 1195–1204 (1992).
13. Loos, R.J. *et al. Nat. Genet.* **40**, 768–775 (2008).
14. Coleman, D.L. *Diabetologia* **14**, 141–148 (1978).
15. Zhang, Y. *et al. Nature* **372**, 425–432 (1994).
16. Tartaglia, L.A. *et al. Cell* **83**, 1263–1271 (1995).
17. Farooqi, S. & O'Rahilly, S. *Endocr. Rev.* **27**, 710–718 (2006).
18. Farooqi, I.S. & O'Rahilly, S. *Am. J. Clin. Nutr.* **89**, 980S–984S (2009).
19. Talbot, C.J. *et al. Nat. Genet.* **21**, 305–308 (1999).
20. Chesler, E.J. *et al. Mamm. Genome* **19**, 382–389 (2008).
21. Bain, J.R. *et al. Diabetes* **58**, 2429–2443 (2009).
22. Yang, Q. *et al. Nature* **436**, 356–362 (2005).
23. Graham, T.E. *et al. N. Engl. J. Med.* **354**, 2552–2563 (2006).
24. Gunton, J.E. *et al. Cell* **122**, 337–349 (2005).
25. Mootha, V.K. *et al. Nat. Genet.* **34**, 267–273 (2003).
26. Patti, M.E. *et al. Proc. Natl. Acad. Sci. USA* **100**, 8466–8471 (2003).
27. Koves, T.R. *et al. Cell Metab.* **7**, 45–56 (2008).
28. Cao, H. *et al. Cell* **134**, 933–944 (2008).
29. Gillum, M.P. *et al. Cell* **135**, 813–824 (2008).
30. Kirchner, H. *et al. Nat. Med.* **15**, 741–745 (2009).
31. Newgard, C.B. *et al. Cell Metab.* **9**, 311–326 (2009).
32. Huffman, K.M. *et al. Diabetes Care* **32**, 1678–1683 (2009).
33. Schadt, E.E. *Nature* **461**, 218–223 (2009).
34. Schadt, E.E. *et al. PLoS Biol.* **6**, e107 (2008).
35. Choi, Y. & Kendziorski, C. *Bioinformatics* **25**, 2780–2786 (2009).
36. Lai, Y., Wu, B., Chen, L. & Zhao, H. *Bioinformatics* **20**, 3146–3155 (2004).
37. Zhang, B. & Horvath, S. *Stat. Appl. Genet. Mol. Biol.* **4**, Article17 (2005).
38. Keller, M.P. *et al. Genome Res.* **18**, 706–716 (2008).
39. Ferrara, C.T. *et al. PLoS Genet.* **4**, e1000034 (2008).
40. Chaibub Neto, E., Ferrara, C.T., Attie, A.D. & Yandell, B.S. *Genetics* **179**, 1089–1100 (2008).
41. Dumas, M.E. *et al. Nat. Genet.* **39**, 666–672 (2007).
42. Gieger, C. *et al. PLoS Genet.* **4**, e1000282 (2008).