Genome **Medicine**

COMMENTARY

Shedding new light on genetic dark matter

Nadine Melhem and Bernie Devlin*

Abstract

Discoveries from genome-wide association studies have contributed to our knowledge of the genetic etiology of many complex diseases. However, these account for only a small fraction of each disease's heritability. Here, we comment on approaches currently available to uncover more of the genetic 'dark matter,' including an approach introduced recently by Naukkarinen and colleagues. These authors propose a method for distinguishing between gene expression driven by genetic variation and that driven by non-genetic factors. This dichotomy allows investigators to focus statistical tests and further molecular analyses on a smaller set of genes, thereby discovering new genetic variation affecting risk for disease. We need more methods like this one if we are to shed a powerful light on dark matter. By enhancing our understanding of molecular genetic etiology, such methods will help us to understand disease processes better and will advance the promise of personalized medicine.

Background

The past three decades of studies have unveiled some of the genetic underpinnings of human disease. For complex diseases, those with obscure genetic roots, discoveries have accelerated recently owing to a bloom of genome-wide association studies (GWASs) [1]. Nevertheless, even for the most successful cases (such as inflammatory and ulcerative bowel disease [2,3]), discoveries account for only a fraction, often small, of the disease's heritability. These yet to be discovered genetic variants comprise the 'missing heritability' or the genetic 'dark matter' for disease.

State of dark matter

Heritability, the proportion of trait variability explained by genetic factors, has two somewhat different meanings. Narrow-sense heritability involves only the additive effects of genes. Broad-sense heritability involves both additive and non-additive effects. The difference between the two makes a difference when hunting for dark matter. If genetic variation were all to act additively, the best predictor of an offspring's trait value would be the average of his/her parents' values. Human height is an excellent example, after adjusting for gender. Hunting for dark matter for a trait such as human height will be more straightforward than for a disease such as schizophrenia, for which the evidence for substantial gene-gene interaction is compelling [4]. Yet when researchers refer to heritability of human height, they implicitly mean narrow-sense heritability; for schizophrenia, it is heritability in a much broader sense.

Why should we care about the genetic basis of disease? Greater understanding of the genetics equals greater understanding of molecular etiology and, with it, eventually more cogent treatments. However, the origins of some human diseases, especially those of the mind, can be mysterious. For diseases of the mind, few environmental or genetic risk factors are understood; instead the hope is that identified genetic factors will lead to a subtler understanding of why diseases such as schizophrenia arise and how they can be treated effectively. Even for cardiovascular disease, for which environmental risk factors are well characterized, new insights into its genetics could produce more targeted treatment. This leads to the other expectation - that greater genetic knowledge will pave the way for 'personalized' medicine. The rapid technological advances in genomics will soon make it feasible to sequence whole genomes at relatively low cost. The idea that each individual will have meaningful sequence variation in their medical records and will have interventions tailored to their risk profile and likely treatment response is quite appealing. The goal of personalized medicine, however, is hindered because so much molecular etiology remains in the dark.

One way to explain more of the dark matter is to develop more efficient ways to use existing data. Naukkarinen *et al.* [5] develop an innovative approach that integrates gene expression and genotype data. They apply these ideas to a GWAS of obesity, as measured by body mass index (BMI). Studies estimate BMI's heritability at 45 to 85%, but identified genetic variants

*Correspondence: devlinbj@upmc.edu
Department of Psychiatry, University of Pittsburgh School of Medicine, 3811 O'Hara St, Pittsburgh, PA 15213, USA

explain about 1% of the total variance [6]. To discover more variants, the authors [5] examined gene expression of adipose tissue in a sample of monozygotic (MZ) twins discordant for BMI and in a sample of unrelated individuals. Because MZ twins are genetically identical, or nearly so, the authors reasoned that genes showing expression differences between twins are 'reactive' genes with differences that are due to regulatory or epigenetic changes in response to environmental factors. By contrast, genes uncovered in unrelated individuals are a combination of reactive and genetically 'causal' genes. By contrasting results from the unrelated sample and discordant MZ twins, the authors identified 27 causal genes that were differentially regulated. They then tested 197 single nucleotide polymorphisms (SNPs) falling in and around these genes in a sample of 21,000 subjects. They discovered a significant excess of small *P*-values in this set of SNPs. Neither the set of SNPs defined by reactive genes nor the individual SNPs in the reactive set were associated with BMI. Notably, this work identifies a new gene, *F13A1*, which encodes the coagulation factor XIII A chain, with variation that affects BMI. This gene has also been identified by meta-analysis of 12 studies of venous thromboembolism [7]. Obesity is well known to predispose to vein thromboses; however, the study of Naukkarinen *et al*. [5] reveals a potential biological pathway for the relationship between obesity, thrombosis and cardiovascular risk.

The methods advanced by Naukkarinen and colleagues [5] require discordant MZ twins, which were available for BMI. This experimental design could prove highly informative for similar quantitative traits, for which extremes are easily identified and by which the pathology or phenotype of interest is defined. For some diseases, especially diseases of the brain, quantitative traits that map precisely onto risk are not yet available. In addition, because reactive genes are environment-dependent, successful implementation of this design might require a sample exposed to a homogeneous environment, limiting its generality. Regardless, this study shows how innovative research can cast more light on dark matter. Moreover, the study design could also inform us about pathways of correlated gene expression and how much these correlations are influenced by genetic and environmental variation.

Many other methods and designs are available to illuminate dark matter [8-15]. One appealing approach teams gene-expression results with genome-wide association data to produce targeted hypothesis tests [8]. One possibility is to organize tests by expression quantitative trait loci affecting genes in pathways meaningful for the disease. Statistical methods for targeted testing are available, whether on the basis of prior information of the likelihood of an association between a SNP and the phenotype or on the basis of plausible disease pathways [9,10]. Genetic variants with parental origin effects, or whose effects depend on the parent from whom they were inherited, could be part of the dark matter; methods are now available to determine the parental origin of alleles and haplotypes even in the absence of genotyped parents [13]. Studies of copy number variants and their inheritance in families could also reveal insight into plausible biological pathways for disease [14,15]. It is also safe to say that rare variants account for some of the dark matter [16], possibly the majority of it in some cases. Next-generation sequencing promises to fill some of our void in knowledge by identifying more penetrant but rarer variants.

Other approaches are less illuminating. Let's reconsider human height. We know numerous rare variants and about 50 common variants that have an impact on height. Thus far, known genetics account for roughly 5% of the variance. Using many SNPs from GWAS analysis that are not significantly associated with height, Yang *et al.* [17] estimated the proportion of variance in height explained by SNPs as 0.45 and even got close to the heritability estimate of 0.84 after correcting for incomplete linkage disequilibrium between SNPs genotyped and causal variants. In spirit, this approach [17] is similar to the allele score method [18], which seeks a predictive model for disease status on the basis of thousands of SNPs with modest evidence for association. If their results are correct, both studies [17,18] suggest that the effects of SNPs are small and will be difficult or impossible to detect from simple analyses of GWASs, at least for current sample sizes [19]. These intriguing approaches have some drawbacks: they shed no new light on the molecular etiology of phenotype; and inherent in the calculations are assumptions that could prove difficult to validate.

We all recognize the hidden biases that inflate estimates of heritability. There are other complex pathways for the transmission of a phenotype across generations without the transmission of a specific common or rare variant, namely through epigenetic factors that can result in the inheritance of gene expression patterns without an alteration of the DNA sequence [20]. Gene-environment interactions could also affect the estimates of heritability and when they are in play, they can explain as much of the variance in the phenotype as genetic factors [21].

Conclusions

Concerted effort will almost surely be required to understand the genetic architecture of most complex diseases. Naukkarinen *et al.*'s [5] novel study design illustrates the impact that concerted effort can have in advancing our knowledge of the genetic etiology of such diseases. There remains ample room for novel analytic methods and study designs to shed light on the genetic dark matter of disease. It is entirely possible, 10 years hence, that we will realize that much of the missing heritability was hiding in plain sight in common variants.

Abbreviations

GWAS, genome-wide association study; MZ, monozygotic; SNP, single nucleotide polymorphism.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

The authors contributed equally to the writing and preparation of this commentary.

Authors' information

BD is Associate Professor of Psychiatry and Human Genetics, University of Pittsburgh School of Medicine, Pittsburgh. His background and area of expertise is statistical genetics. NM is Assistant Professor of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh. Her background and areas of expertise are psychiatric epidemiology and statistical genetics.

Acknowledgements

This work was supported by R01 grants (MH057881, BD) and a K01 grant (MH077930, NM) from the National Institute of Mental Health. The National Institute of Mental Health did not participate in the preparation, review or approval of the manuscript.

Published: 21 October 2010

References

1. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM: **Finding the missing heritability of complex diseases.** *Nature* 2009, **461:**747-753.
2. McGovern DP, Gardet A, Törkvist L, Goyette P, Essers J, Taylor KD, Neale BM, Ong RT, Lagacé C, Li C, Green T, Stevens CR, Beauchamp C, Fleshner PR, Carlson M, D'Amato M, Halfvarson J, Hibberd ML, Lördal M, Padyukov L, Andriulli A, Colombo E, Latiano A, Palmieri O, Bernard EJ, Deslandres C, Hommes DW, de Jong DJ, Stokkers PC, Weersma RK, *et al.*: **Genome-wide association identifies multiple ulcerative colitis susceptibility loci.** *Nat Genet* 2010, **42:**332-337.
3. Imielinski M, Baldassano RN, Griffiths A, Russell RK, Annese V, Dubinsky M, Kugathasan S, Bradfield JP, Walters TD, Sleiman P, Kim CE, Muise A, Wang K, Glessner JT, Saeed S, Zhang H, Frackelton EC, Hou C, Flory JH, Otieno G, Chiavacci RM, Grundmeier R, Castro M, Latiano A, Dallapiccola B, Stempak J, Abrams DJ, Taylor K, McGovern D; Western Regional Alliance for Pediatric IBD, *et al.*: **Common variants at five new loci associated with early-onset inflammatory bowel disease.** *Nat Genet* 2009, **41:**1335-1340.
4. Risch N: **Linkage strategies for genetically complex traits. I. Multilocus models.** *Am J Hum Genet* 1990, **46:**222-228.
5. Naukkarinen J, Surakka I, Pietiläinen KH, Rissanen A, Salomaa V, Ripatti S, Yki-Järvinen H, van Duijn CM, Wichmann HE, Kaprio J, Taskinen MR, Peltonen L, ENGAGE Consortium: **Use of genome-wide expression data to mine the "Gray Zone" of GWA studies leads to novel candidate obesity genes.** *PLoS Genet* 2010, **6:**e1000976.
6. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, Lango H, Rayner NW, Shields B, Harries LW, Barrett JC, Ellard S, Groves CJ, Knight B, Patch AM, Ness AR, Ebrahim S, Lawlor DA, Ring SM, Ben-Shlomo Y, Jarvelin MR, Sovio U, Bennett AJ, Melzer D, Ferrucci L, Loos RJ, Barroso I, Wareham NJ, *et al.*: **A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity.** *Science* 2007, **316:**889-894.
7. Wells PS, Anderson JL, Scarvelis DK, Doucette SP, Gagnon F: **Factor XIII Val34Leu variant is protective against venous thromboembolism: a HuGE review and meta-analysis.** *Am J Epidemiol* 2006, **164:**101-109.
8. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ: **Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS.** *PLoS Genet* 2010, **6:**e1000888.
9. Wang K, Li M, Bucan M: **Pathway-based approaches for analysis of genomewide association studies.** *Am J Hum Genet* 2007, **81:**1278-1283.
10. Roeder K, Wasserman L: **Genome-wide significance levels and weighted hypothesis testing.** *Stat Sci* 2009, **24:**398-413.
11. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH: **Missing heritability and strategies for finding the underlying causes of complex disease.** *Nat Rev Genet* 2010, **11:**446-450.
12. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK: **Understanding mechanisms underlying human gene expression variation with RNA sequencing.** *Nature* 2010, **464:**768-772.
13. Kong A, Steinthorsdottir V, Masson G, Thorleifsson G, Sulem P, Besenbacher S, Jonasdottir A, Sigurdsson A, Kristinsson KT, Jonasdottir A, Frigge ML, Gylfason A, Olason PI, Gudjonsson SA, Sverrisson S, Stacey SN, Sigurgeirsson B, Benediktsdottir KR, Sigurdsson H, Jonsson T, Benediktsson R, Olafsson JH, Johannsson OT, Hreidarsson AB, Sigurdsson G; DIAGRAM Consortium, Ferguson-Smith AC, Gudbjartsson DF, Thorsteinsdottir U, Stefansson K: **Parental origin of sequence variants associated with complex diseases.** *Nature* 2009, **462:**868-874.
14. Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S, Zhang H, Estes A, Brune CW, Bradfield JP, Imielinski M, Frackelton EC, Reichert J, Crawford EL, Munson J, Sleiman PM, Chiavacci R, Annaiah K, Thomas K, Hou C, Glaberson W, Flory J, Otieno F, Garris M, Soorya L, Klei L, Piven J, Meyer KJ, Anagnostou E, Sakurai T, *et al.*: **Autism genome-wide copy number variation reveals ubiquitin and neuronal genes.** *Nature* 2009, **459:**569-573.
15. Wellcome Trust Case Control Consortium, Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, Vukcevic D, Barnes C, Conrad DF, Giannoulatou E, Holmes C, Marchini JL, Stirrups K, Tobin MD, Wain LV, Yau C, Aerts J, Ahmad T, Andrews TD, Arbury H, Attwood A, Auton A, Ball SG, Balmforth AJ, Barrett JC, Barroso I, Barton A, Bennett AJ, Bhaskar S, *et al.*: **Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls.** *Nature* 2010, **464:**713-720.
16. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH: **Multiple rare alleles contribute to low plasma levels of HDL cholesterol.** *Science* 2004, **305:**869-872.
17. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM: **Common SNPs explain a large proportion of the heritability for human height.** *Nat Genet* 2010, **42:**565-569.
18. International Schizophrenia Consortium, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P: **Common polygenic variation contributes to risk of schizophrenia and bipolar disorder.** *Nature* 2009, **460:**748-752.
19. Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, Chatterjee N: **Estimation of effect size distribution from genome-wide association studies and implications for future discoveries.** *Nat Genet* 2010, **42:**570-575.
20. Nadeau JH: **Transgenerational genetic effects on phenotypic variation and disease risk.** *Hum Mol Genet* 2009, **18:**R202-R210.
21. Reed LK, Williams S, Springston M, Brown J, Freeman K, DesRoches CE, Sokolowski MB, Gibson G: **Genotype-by-diet interactions drive metabolic phenotype variation in *Drosophila melanogaster*.** *Genetics* 2010, **185:**1009-1019.

doi:10.1186/gm200
Cite this article as: Melhem N, Devlin B: **Shedding new light on genetic dark matter.** *Genome Medicine* 2010, **2:**79.