PLOS GENETICS

OPINION PIECE

The great hairball gambit

Jonathan Flint[1]*, **Trey Ideker**[2]*

1 Department of Psychiatry and Biobehavioral Sciences, University of California Los Angeles, Los Angeles, California, United States of America, **2** Division of Genetics, Department of Medicine, University of California San Diego, La Jolla, California, United States of America

* jflint@mednet.ucla.edu (JF); tideker@ucsd.edu (TI)

It's generally agreed that the successful application of genome-wide association studies (GWAS) to the genetic dissection of psychiatric disease, resulting in the identification of hundreds of susceptibility loci for schizophrenia [1] and smaller numbers for depression [2, 3], autism [4], bipolar disorder [5], and anorexia [6], hasn't been matched by equal success in working out the biology of the same conditions. To get at biological mechanism, one thing that defenders of the GWAS methodology generally propose and instantiate in their papers is a description of relevant molecular pathways and interaction networks–a.k.a network analysis–which takes as its starting point risk loci occurring at or close to genes.

The construction of gene and protein networks, whether made from correlated expression of transcripts or the interaction partners of proteins, can either be lauded as a way to transform information about genetic risk loci from genes to etiological mechanisms, or derided as an uninformative exercise, flawed not only by the poverty of the data upon which it relies, but more fundamentally by its departure from reductionist explanations of how things work in biology. The divide between these two positions sometimes seems to be as wide as the gulf between Republican and Democrat, or between the supporters and opponents of Brexit. Somewhere, presumably, there is a bridge across. . .

Network analysis has the potential to shed light on GWAS loci due to interactions at multiple layers of cell biological information. First, multilayered networks involving cooperative and antagonistic combinations of histone modifications, sites of transcription factor occupancy, enhancers, promoters and genes, create co-regulation of gene transcription. As a consequence, any genetic variant that disturbs the expression of a transcription factor or histone modifier typically has knock-on effects on the expression of many other genes, and even variants in genes that are supposedly regulated downstream of all this can cause a readjustment of transcription in the larger system. Second, protein interaction networks are central to cell and tissue biology, otherwise there'd be no organization of proteins into large multimeric complexes, or co-localization in specific subcellular compartments. Currently at least 4,400 mammalian proteins have been placed in specific protein complexes [7] and greater than two-thirds of proteins have been mapped to highly organized spatial compartments within the nucleus, cytosol, mitochondrion or membranes [8]. Without protein interactions there would also be no basis for signaling networks and the transmission of inter and intracellular signals, for instance from ligand to receptor or from kinase to substrate. Finally, metabolic networks provide a means by which changes in the levels or activities of enzymes or metabolites can propagate to affect the levels or activities of many others. In short, these many layers of networks attest to the fact that most genes don't act in a vacuum, and thus to understand disease we need to know how individual effects alter larger biological processes modeled by networks.

Citation: Flint J, Ideker T (2019) The great hairball gambit. PLoS Genet 15(11): e1008519. https://doi.org/10.1371/journal.pgen.1008519

Editor: Gregory P. Copenhaver, The University of North Carolina at Chapel Hill, UNITED STATES

Published: November 26, 2019

Funding: This work was funded by grants from the National Institutes of Health (R01MH115979 to JF, DA037844 to TI). The funders had no role in the preparation of the article.

Competing interests: The authors have declared that no competing interests exist.

All of these networks (transcriptional, structural, signaling, metabolic) are perturbed by variants at genetic loci underlying psychiatric disease. As with other complex traits, most of the risk loci for psychiatric disease lie in regulatory regions of the genome [9], which means that rather than altering protein structure they alter mRNA production or mRNA stability via transcriptional networks. In turn, alterations to transcription affect protein networks, through changes in the stoichiometry of protein complexes and the activities of signaling cascades. They affect metabolic networks through changes in enzyme abundance, which then give rise to many indirect effects on metabolic levels and fluxes more globally. Ultimately, the collection of all network alterations in an individual translates to a distinct profile of effects on psychiatric development, behavior and therapeutic outcomes. Charting causal pathways through these networks, i.e. connecting causal genetic variants to impacts on transcriptional, protein, and/or metabolic networks to impacts on psychiatric phenotypes, is the essence of network analysis.

All this sounds very convincing, but to what extent has network analysis thrown new light on disease pathogenesis? There's certainly evidence that gene expression analysis identifies expression modules that contain GWAS signals, for example from autism spectrum disorder [10] and schizophrenia [11]. What's less certain is the utility of this information. You'll read that network analysis has indicated that autism and schizophrenia are disorders of the synapse [12, 13], and that autism is characterized by disorders of cortical projection neurons from mid-fetal layer 5/6 [14]. That doesn't, yet, amount to a mechanistic interpretation, other than in a broad sense. What we really want to know when we ask about the biological causes of psychosis is *what is wrong with the synapse*, *and how does that explain psychiatric phenomena*?

Perhaps a lack of detailed information on where and when in the brain genes are expressed limits the power of networks to identify causal mechanisms, a view that drives the collection of ever larger datasets, for instance taken from single cells at different developmental time points [15]. In the meantime, we can think of several major reasons for why networks may fall short of delivering the causal pathway explanations everyone is looking for.

First, the starting point for network analysis, a set of genes associated with a disease, does not fall out of the data like factors from a principal component analysis. Translation of genetic loci to genes is more often a matter of faith than of rigorous proof. Despite the presumptions of many bioinformatics programs (e.g. ALIGATOR, INRICH MAGENTA and MAGMA [16–20]), proximity to a GWAS signal is not a guarantee for gene identification. As a series of experiments have shown, regulatory elements containing SNPs don't always regulate the nearest gene [21–23]. Rather, the vast majority of GWAS loci are in regulatory regions that lie at a distance, sometimes many megabases, from their gene targets [9]. One estimate is that about half the targets are in fact the closest gene [24] whereas, roughly speaking, about half are incorrect when using genomic proximity as the sole indicator. Rigorous demonstration that the causative variant at a risk locus has been identified, the effect of that change confirmed, and the target gene found has been slow coming in the GWAS field, largely because it is such a difficult thing to do [21].

Second, even if we had accurate sets of genes from GWAS, we are still far from having complete interaction maps. Arguably, resources such as BioPlex [25] and the yeast-two-hybrid (Y2H) Human Reference Interactome [26] are nearing complete coverage of the human protein-protein interaction space, but these are either in one cell line (BioPlex is in HEK293T cells, workhorses of protein production but of unclear relevance for neurobiology) or, in the case of yeast two hybrid, not in a human cell type. Then there is the question of biological context, including the relevant growth conditions, tissue microenvironment, state of functional activation (especially relevant for the brain), and course of therapy. Even in a single biological context, network mapping technologies have typically large false negative rates [27] meaning

that many true interactions are missed, and methods that query interactions with a single partner introduce biases that at least have the virtue of being acknowledged.

As with protein network mapping, attempts at completeness of transcriptional profiling data are vitiated by the temporal, tissue and cell specific nature of what gets expressed. Except in constrained situations of analysing a single cell type, at the same point in the cell cycle, in unvarying environmental situations, with multiple measures taken over a uniform timeframe, we cannot obtain fully complete or reliable measures of transcriptional state. Furthermore, our dependence on what is most studied, such as cancer cells, contaminates annotations, and one can only wonder at what biases remain to be detected.

A third fundamental problem facing networks is their lack of specificity. In co-expression networks everything correlates with everything at some level, including both upstream causes and the multiple consequent downstream effects. The result is an enormous hairball in which, to a first approximation, every gene interacts with every other gene. Issues with specificity beset other network types too. For instance, protein interaction screens identify proteins by their affinity to a tagged "bait" which is typically overexpressed at dramatic levels. The degree to which such overexpression leads to specificity problems is not well characterized. Surely these specificity problems increase the difficulty of identifying a pathway causal for a disease?

A final, and implacable, enemy of network analysis is the genetic architecture of psychiatric disorders. Ever since David Goldstein plotted the distribution of effect sizes of loci for complex traits [28], it's been a source of worry to geneticists that explaining heritability is going to require a very large number of loci. Just how many still isn't known, but it's likely to be tens of thousands. For example, greater than 100,000 SNPs have been identified with independent effects on human height, perhaps extending upwards to as many as half of common SNPs [29]. There's every reason to believe that psychiatric disorders are similarly polygenic. In fact, one need hardly extrapolate to predict that the number of genes associated with such disorders will soon saturate at close to all 20,000 genes. Then what? What more will we really understand about the diseased genome, not to mention its molecular mechanisms or resulting physiology?

None of this is to say that biological explanations eschew networks, but how can we advance molecular network maps along a productive path for doing human genetics? First off, at least in their present form co-expression data may simply be the wrong type of network information for the job of reading the genome. No matter how complete they are, they tell you very little about mechanism, since they are dominated by knock-on effects and massive gene-expression regulons. The causal change in expression, the one that can illuminate the pathway from gene to disease, is lost in a haystack of thousands of other changes, or, worse still, the causal pathway is just simply unobservable at the level of gene expression. Furthermore, simply computing correlations in gene expression over all available samples, as is often the practice, is confounded by issues of physiological context and cell type–what may be really going on is that the 'co-expressed' genes are specifically active in one cell type, the frequency of which varies across experimental subjects.

Co-expression networks are widely used because they are conveniently generated from high-throughput sequencing, not because they are necessarily the right structure for genetic interpretation. Strikingly, in a systematic assembly of autophagy pathways [30] mRNA co-expression could be dispensed with *entirely* without loss of information, whereas other network types such as protein-protein and synthetic-lethal interactions were critical for reconstructing current knowledge of pathway relationships and functions. Will the same conclusions hold when different network data types are evaluated for their power in interpreting psychiatric GWAS? Thus far it appears that networks integrating multiple types of evidence are significantly more useful for interrelating disease loci than those representing one data set or type only [31]—good to know but hardly surprising.

Alongside network type are considerations of biological conditions. Networks should be mapped in the cellular and environmental conditions most relevant to psychiatric disease, while continually working in contexts in which systematic datasets can be readily generated. Whether this means selecting a suitable panel of neuronal cell lines or organoids, or going straight to brain or other tissues, remains to be determined. Moreover, in making these decisions we need more information about the stability of transcriptional, protein, and metabolic networks to genetic and environmental perturbations, as well as to experimental errors. At one extreme, network architecture might be so ephemeral as to require separate maps in each individual, or worse yet, every clonal cell population within each individual. At the other extreme, perhaps one or a few reference networks are enough to understand the etiology of most cases of a disease. Regardless, having *an initial* draft set of network reference maps for neuropsychiatric disease would represent a big leap forward from where we are now, as has recently been undertaken by consortia like SynGO [32].

Given that only a dozen years ago almost nothing was known about the genetic loci underlying behavior, the success of GWAS is remarkable: 145 loci for schizophrenia [1], 102 for depression [3], 202 for insomnia [33] and 1,271 for educational attainment [34]. Similar troves of loci are accumulating for most other complex genetic diseases, including cancer, cardiovascular disorders, diabetes, and rheumatoid arthritis. Access to large biobanks [35] together with cheap genotyping and data sharing has made it relatively straightforward to find loci. Perhaps that very success has encouraged the view that the functional interpretation of genetic signals has become equally routine.

On the contrary, our opinion is that the difficulties of integrating network and genetic data are under appreciated. The risk is that many of the networks currently reported to be responsible for psychiatric disease will turn out to be the equivalent of the false-positive findings that emerged from the early genetic studies of behavior: incomplete results, obtained from the wrong type of data in the wrong cellular conditions, using the wrong genes, conspire to create hairballs of questionable value. Before we can obtain robust functional interpretations of GWAS findings, we'd like to see a comprehensive collection of network and pathway datasets for psychiatric and other complex disorders, boosting network coverage and accuracy, and integrating data from multiple sources, rather than the common practice of relying solely on expression data.

Acknowledgments

We thank Abraham Palmer for his valuable feedback on the manuscript.

References

1. Pardinas AF, Holmans P, Pocklington AJ, Escott-Price V, Ripke S, Carrera N, et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. Nat Genet. 2018; 50(3):381–9. https://doi.org/10.1038/s41588-018-0059-2 PMID: 29483656
2. Wray NR, Ripke S, Mattheisen M, Trzaskowski M, Byrne EM, Abdellaoui A, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat Genet. 2018; 50(5):668–81. https://doi.org/10.1038/s41588-018-0090-3 PMID: 29700475
3. Howard DM, Adams MJ, Clarke T, Hafferty J, Gibson J, Shirali M, et al. Genome-wide meta-analysis of depression in 807,553 individuals identifies 102 independent variants with replication in a further 1,507,153 individuals. Nature Neuroscience. 2018.
4. Grove J, Ripke S, Als TD, Mattheisen M, Walters RK, Won H, et al. Identification of common genetic risk variants for autism spectrum disorder. Nat Genet. 2019; 51(3):431–44. https://doi.org/10.1038/s41588-019-0344-8 PMID: 30804558.
5. Stahl EA, Breen G, Forstner AJ, McQuillin A, Ripke S, Trubetskoy V, et al. Genome-wide association study identifies 30 loci associated with bipolar disorder. Nat Genet. 2019; 51(5):793–803. https://doi.org/10.1038/s41588-019-0397-8 PMID: 31043756.

6. Watson HJ, Yilmaz Z, Thornton LM, Hubel C, Coleman JRI, Gaspar HA, et al. Genome-wide association study identifies eight risk loci and implicates metabo-psychiatric origins for anorexia nervosa. Nat Genet. 2019. https://doi.org/10.1038/s41588-019-0439-2 PMID: 31308545.
7. Giurgiu M, Reinhard J, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, et al. CORUM: the comprehensive resource of mammalian protein complexes-2019. Nucleic Acids Res. 2019; 47(D1):D559–D63. https://doi.org/10.1093/nar/gky973 PMID: 30357367
8. Thul PJ, Akesson L, Wiking M, Mahdessian D, Geladaki A, Ait Blal H, et al. A subcellular map of the human proteome. Science. 2017; 356(6340). https://doi.org/10.1126/science.aal3321 PMID: 28495876.
9. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012; 337(6099):1190–5. Epub 2012/09/08. https://doi.org/10.1126/science.1222794 PMID: 22955828.
10. Parikshak NN, Luo R, Zhang A, Won H, Lowe JK, Chandran V, et al. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. Cell. 2013; 155(5):1008–21. Epub 2013/11/26. https://doi.org/10.1016/j.cell.2013.10.031 PMID: 24267887
11. Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FCP, et al. Comprehensive functional genomic resource and integrative model for the human brain. Science. 2018; 362(6420). https://doi.org/10.1126/science.aat8464 PMID: 30545857.
12. Fromer M, Pocklington AJ, Kavanagh DH, Williams HJ, Dwyer S, Gormley P, et al. De novo mutations in schizophrenia implicate synaptic networks. Nature. 2014; 506(7487):179–84. Epub 2014/01/28. https://doi.org/10.1038/nature12929 PMID: 24463507.
13. Ebert DH, Greenberg ME. Activity-dependent neuronal signalling and autism spectrum disorder. Nature. 2013; 493(7432):327–37. https://doi.org/10.1038/nature11860 PMID: 23325215
14. Willsey AJ, Sanders SJ, Li M, Dong S, Tebbenkamp AT, Muhle RA, et al. Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. Cell. 2013; 155(5):997–1007. https://doi.org/10.1016/j.cell.2013.10.020 PMID: 24267886
15. Parikshak NN, Gandal MJ, Geschwind DH. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. Nat Rev Genet. 2015; 16(8):441–58. https://doi.org/10.1038/nrg3934 PMID: 26149713
16. Lee PH, O'Dushlaine C, Thomas B, Purcell SM. INRICH: interval-based enrichment analysis for genome-wide association studies. Bioinformatics. 2012; 28(13):1797–9. https://doi.org/10.1093/bioinformatics/bts191 PMID: 22513993
17. Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. Am J Hum Genet. 2007; 81(6):1278–83. https://doi.org/10.1086/522374 PMID: 17966091
18. Holmans P, Green EK, Pahwa JS, Ferreira MA, Purcell SM, Sklar P, et al. Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. Am J Hum Genet. 2009; 85(1):13–24. https://doi.org/10.1016/j.ajhg.2009.05.011 PMID: 19539887
19. Segre AV, Consortium D, investigators M, Groop L, Mootha VK, Daly MJ, et al. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. PLoS Genet. 2010; 6(8). https://doi.org/10.1371/journal.pgen.1001058 PMID: 20714348
20. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol. 2015; 11(4):e1004219. https://doi.org/10.1371/journal.pcbi.1004219 PMID: 25885710
21. Claussnitzer M, Dankel SN, Kim KH, Quon G, Meuleman W, Haugen C, et al. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. N Engl J Med. 2015; 373(10):895–907. https://doi.org/10.1056/NEJMoa1502214 PMID: 26287746.
22. Smemo S, Tena JJ, Kim KH, Gamazon ER, Sakabe NJ, Gomez-Marin C, et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature. 2014; 507(7492):371–5. Epub 2014/03/22. https://doi.org/10.1038/nature13138 PMID: 24646999.
23. Lettice LA, Horikoshi T, Heaney SJ, van Baren MJ, van der Linde HC, Breedveld GJ, et al. Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. Proc Natl Acad Sci U S A. 2002; 99(11):7548–53. https://doi.org/10.1073/pnas.112212199 PMID: 12032320.
24. Won H, de la Torre-Ubieta L, Stein JL, Parikshak NN, Huang J, Opland CK, et al. Chromosome conformation elucidates regulatory relationships in developing human brain. Nature. 2016; 538(7626):523–7. https://doi.org/10.1038/nature19847 PMID: 27760116.
25. Huttlin EL, Ting L, Bruckner RJ, Gebreab F, Gygi MP, Szpyt J, et al. The BioPlex Network: A Systematic Exploration of the Human Interactome. Cell. 2015; 162(2):425–40. https://doi.org/10.1016/j.cell.2015.06.043 PMID: 26186194

26. Luck K, Kim D-K, Lambourne L, Spirohn K, Begg BE, Bian W, et al. A reference map of the human protein interactome. bioRxiv. 2019:605451. https://doi.org/10.1101/605451

27. Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, et al. An empirical framework for binary interactome mapping. Nat Methods. 2009; 6(1):83–90. https://doi.org/10.1038/nmeth.1280 PMID: 19060904

28. Goldstein DB. Common genetic variation and human traits. N Engl J Med. 2009; 360(17):1696–8. Epub 2009/04/17. https://doi.org/10.1056/NEJMp0806284 PMID: 19369660.

29. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017; 169(7):1177–86. https://doi.org/10.1016/j.cell.2017.05.038 PMID: 28622505

30. Kramer MH, Farre JC, Mitra K, Yu MK, Ono K, Demchak B, et al. Active Interaction Mapping Reveals the Hierarchical Organization of Autophagy. Mol Cell. 2017; 65(4):761–74 e5. https://doi.org/10.1016/j.molcel.2016.12.024 PMID: 28132844

31. Huang JK, Carlin DE, Yu MK, Zhang W, Kreisberg JF, Tamayo P, et al. Systematic Evaluation of Molecular Networks for Discovery of Disease Genes. Cell Syst. 2018; 6(4):484–95 e5. https://doi.org/10.1016/j.cels.2018.03.001 PMID: 29605183

32. Koopmans F, van Nierop P, Andres-Alonso M, Byrnes A, Cijsouw T, Coba MP, et al. SynGO: An Evidence-Based, Expert-Curated Knowledge Base for the Synapse. Neuron. 2019; 103(2):217–34 e4. https://doi.org/10.1016/j.neuron.2019.05.002 PMID: 31171447

33. Jansen PR, Watanabe K, Stringer S, Skene N, Bryois J, Hammerschlag AR, et al. Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. Nat Genet. 2019; 51(3):394–403. https://doi.org/10.1038/s41588-018-0333-3 PMID: 30804565.

34. Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat Genet. 2018; 50(8):1112–21. https://doi.org/10.1038/s41588-018-0147-3 PMID: 30038396.

35. Collins R. What makes UK Biobank special? Lancet. 2012; 379(9822):1173–4. https://doi.org/10.1016/S0140-6736(12)60404-8 PMID: 22463865.