Network analysis for complex neurodegenerative diseases

Current Genetic Medicine Reports (2020) 8:17–25
https://doi.org/10.1007/s40142-020-00181-z

BIOINFORMATICS (A PITTMAN, SECTION EDITOR)

Network Analysis for Complex Neurodegenerative Diseases

Claudia Manzoni[1,2] · Patrick A. Lewis[3,1,2] · Raffaele Ferrari[2]

Published online: 17 January 2020

Abstract

Purpose of Review Biomedicine is witnessing a paradigm shift in the way complex disorders are investigated. In particular, the need for big data interpretation has led to the development of pipelines that require the cooperation of different fields of expertise, including medicine, functional biology, informatics, mathematics and systems biology. This review sits at the crossroad of different disciplines and surveys the recent developments in the use of graph theory (in the form of network analysis) to interpret large and different datasets in the context of complex neurodegenerative diseases. It aims at a professional audience with different backgrounds.

Recent Findings Biomedicine has entered the era of big data, and this is actively changing the way we approach and perform research. The increase in size and power of biomedical studies has led to the establishment of multi-centre, international working groups coordinating open access platforms for data generation, storage and analysis. Particularly, pipelines for data interpretation are under development, and network analysis is gaining momentum since it represents a versatile approach to study complex systems made of interconnected multiple players.

Summary We will describe the era of big data in biomedicine and survey the major freely accessible multi-omics datasets. We will then introduce the principles of graph theory and provide examples of network analysis applied to the interpretation of complex neurodegenerative disorders.

Keywords Complex neurodegeneration · Network analysis · Protein-protein interactions · GWAS loci · Omics data · Data integration · Gene co-expression

Introduction

During the past decade, the field of human genetics has witnessed massive global improvements in the generation of high resolution genomics data as genotyping arrays and next generation sequencing (NGS, whole genome [WGS] or whole exome sequencing [WES]) have become time- and cost-effective techniques to assist the genetic study of health and disease [1]. Similarly, transcriptomics and proteomics (and also epigenomics and metabolomics) studies have benefited from rapid advancements in the technologies and methods for data generation and analysis [2•]. Alongside, bioinformatics tools and pipelines that are accessible and shared throughout the wider scientific community, together with ever improving computational environments, have supported an exponential growth in big data availability for basic and applied biomedical research [3••].

We are currently—probably for the first time in medical history—facing a paradoxical "abundance problem", i.e. having more data at hand than we can ever interpret and effectively translate into medical practice. Like for the Levinthal paradox on protein folding [4], the only way to tackle the current status quo is that of moving on from the classical way of analysing data one by one and changing the paradigm in which biomedical research operates. In truth, it appears that both reductionist

This article is part of the Topical Collection on *Bioinformatics*

✉ Claudia Manzoni
c.manzoni@reading.ac.uk

Patrick A. Lewis
plewis@rvc.ac.uk

Raffaele Ferrari
r.ferrari@ucl.ac.uk

[1] School of Pharmacy, University of Reading, Whiteknights, Reading RG6 6AP, UK

[2] Department of Neurodegenerative Diseases, University College London, 9-12 Russell Square House, London WC1B 5EH, UK

[3] Royal Veterinary College, Royal College Street, London NW1 OTU, UK

(classical) and holistic (novel) approaches cannot be treated as separate fields anymore and need to be considered on a convergent and cross-supportive path where systems biology, computational modelling, mathematics and informatics play a critical role [5•].

In the era of big data, it has become increasingly clear that advances can only result from collective efforts and data sharing [6]. Biomedicine has thus seen the rise of consortia, large-scale (mainly international) efforts aimed at sharing resources, maximizing both sample collection and data generation and harmonizing analytical strategies. In the field of neurodegeneration examples include the International Genomics of Alzheimer's Project (IGAP, http://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php), the International Parkinson's Disease Genomics Consortium (IPDGC, https://pdgenetics.org) and the International Frontotemporal Dementia Genomics Consortium (IFGC, https://ifgcsite.wordpress.com). These large-scale collaborative efforts are paving the way for a coherent understanding of the molecular mechanisms of complex neurodegenerative diseases. More in general, "resource" consortia together with international working committees and open access databases have been set up to promote international collaborations, standardize nomenclature, data storage and sharing in line with the highest standards and best practices (Table 1).

Complex Neurodegeneration and Network Analysis

In monogenic disorders, a mutation with high effect size in a specific gene that acts as pathogenic trigger and disease mechanism can be (directly) inferred through the functional analysis of that single mutated gene.

In the case of complex diseases, multiple genetic markers with small effect size contribute all together to the trait. In complex neurodegenerative diseases, the genetic component for the majority of cases (sporadic) is indeed defined by a plethora of variants, i.e. genetic architecture of disease, priming the individual to develop disease at a certain stage of life [30]. In a minority of complex neurodegeneration cases (familial), mutations in single genes are isolated. Even if these mutations hold strong causative effects, modifiers within the genetic architecture can modulate disease onset and progression. Reports of *PSEN1* mutation carriers who are resistant to or show a delayed onset for Alzheimer's disease (AD) due to their *APOE* genotype [31, 32] as well as the non-complete penetrance of *LRRK2* mutations in families affected by Parkinson's disease (PD) [33] are examples of how seemingly even monogenic cases of familial neurodegenerative diseases can be indeed classified as complex disorders. In addition to the genetic component, the environment also plays a role in complex disease pathogenesis, acting as an additional risk factor, e.g. inducing disease-relevant epigenetic changes and/or acting as disease trigger on a receptive genetic asset. The molecular mechanisms at the basis of complex neurodegenerative diseases are not straightforward to be read, since the genetic architecture of risk is difficult to be modelled and requires multiple causative markers to be analysed simultaneously (Fig. 1).

In this scenario, in silico systems biology approaches, for example network analysis, have the potential to revolutionise the translation of genetics information into functional understanding of the molecular basis of disease. The availability of large sets of well-curated omics data and the development of bioinformatics approaches based on graph theory are opening up the possibility, for the first time, to study complex diseases by simultaneously modelling the multiple genetic factors at play with a more holistic approach by studying networks [5•].

Networks, also called graphs, are mathematical objects that represent multiple data as a whole. Networks are composed of nodes (objects constituting the network) and edges (connections between those objects). One can visualize biological networks by using freely available tools such as Cytoscape [34] (https://cytoscape.org) and yED (https://www.yworks.com/products/yed), and study networks through the mathematical approaches offered by graph theory.

Transcriptomics or proteomics (both steady-state and time-series type of data) are used for building gene co-expression networks (GCNs) following the assumption that genes that are co-expressed are probably co-regulated and thus part of the same pathway [35]. The input dataset for GCNs needs to be statistically processed (different methods have been developed such as WGCNA, CLR, ARACNe, PCIT, GENIE3, SIRENE and GeCON [36••]) to generate the co-expression information, i.e. the relationships that are essential for building edges.

Protein interaction data, derived from a wide range of cellular and biochemical model systems, can be used for building protein-protein interaction networks (PINs). Generating PINs is relatively straightforward considering that the relationship between nodes and edges (i.e. protein interaction) directly reflects the type of information contained in the original datasets [37].

Finally, hybrid networks can be constructed by mixing different types of omics data. Gene regulatory networks (GRNs) are a type of complex network where nodes can be genes, proteins, metabolites. Pairs of nodes are connected by edges where one of the nodes in the pair influences (via inhibition or activation) the activity of the other [38]. The construction of GRNs is usually performed by applying statistical approaches based on inference algorithms (including Bayesian, artificial neuronal, and Boolean networks, regression-based model, ordinal differential equation and information theory) [39, 40]. These methods are all aimed at extracting the probability of the reciprocal regulation for all pairs of nodes within large datasets used as input (e.g. gene expression, protein-DNA interactions, transcription factors binding) mathematically generating edges between nodes.

Table 1 Open access, big data repositories

Resource	Website	Details	Ref.
Database of Genotypes and Phenotypes (dbGaP)	https://www.ncbi.nlm.nih.gov/gap	Catalogue of genetic datasets	[7]
European Genome-phenome Archive (EGA)	https://www.ebi.ac.uk/ega/home	Catalogue of genetic datasets	[8]
GWA catalogue	https://www.ebi.ac.uk/gwas	Catalogue of published GWA	[9]
Genome Reference Consortium	https://www.ncbi.nlm.nih.gov/grc/data	Controls or general population genome	[10]
1000 Genomes	https://www.internationalgenome.org	Controls or general population human genome	[11]
Exac	http://exac.broadinstitute.org	Controls or general population human genome	[12]
GnomAD	https://gnomad.broadinstitute.org	Controls or general population human genome	
Wellderly		Controls or general population human genome	[13]
NCBI	https://www.ncbi.nlm.nih.gov/genome	Open access genome browsers	[14]
UCSC	https://genome.ucsc.edu	Open access genome browsers	[14]
Ensembl	https://www.ensembl.org	Open access genome browsers	[14]
Gene Expression Omnibus (GeO)	https://www.ncbi.nlm.nih.gov/geo	Comprehensive collection of transcriptomics data and gene expression	[15]
Genotype-Tissue Expression project (GTEx)	https://gtexportal.org/home/index.html	Transcriptional profiles of multiple human tissues	[16]
Braineac	http://www.braineac.org	Transcriptional profiles of multiple human brain regions	[17]
The Encyclopedia of DNA Elements (ENCODE)	https://www.encodeproject.org	Catalogue of non-coding elements	[18]
Functional Annotation of the Mammalian Genome (FANTOM)	http://fantom.gsc.riken.jp	Catalogue of non-coding elements	[19]
ROADMAP	http://www.roadmapepigenomics.org	Catalogue of epigenetic changes	[20]
Uniprot	https://www.uniprot.org	Encyclopedic reference source for proteins	[21]
Molecular Exchange Consortium (IMEx)	https://www.imexconsortium.org	Protein-protein interaction (PPIs) are stored in manually curated databases the majority of which follow standardised guidelines for data processing and collection as defined by IMEx	[22]
ProteomeXchange	http://www.proteomexchange.org	Collection of proteomics repositories	[23]
Human Proteome Map	https://www.humanproteomemap.org/	Comprehensive proteomics data generated in multiple human tissues	[24]
Gene Ontology (GO)	http://geneontology.org	Catalogue of gene associated: biological processes & molecular functions & cellular components	[25]
Online Mendelian Inheritance in Man (OMIM)	https://www.omim.org	Gene-disease associations	[26]
DisGeNET	http://www.disgenet.org	Gene-disease associations	[27]
Reactome	https://reactome.org	Pathways repository	[28]
Kyoto Enciclopedia of Genes and Genomes (KEGG pathway)	https://www.genome.jp/kegg/pathway.html	Pathways repository	[29]

A key advantage of networks in an experimental context is that they are mathematical objects kept together by connections (i.e. relationships) between the nodes. Therefore, networks include multiple players simultaneously (nodes) that are analysed assessing their concurrent interactions within the global structure of the graph. This type of topological analysis is aimed at identifying relevant nodes and understanding how the information flows throughout the entire structure of the network [41]. Relevant nodes are, for example, hubs (highly connected nodes), i.e. essential genes within the network structure [42], and bottlenecks (shortcuts), i.e. non-essential genes that can be targeted (e.g. by drugs) to modify the flow of information within the network [43]. Assuming that nodes are genes and/or proteins connected in the network through functional relationships, the information contained in the network is a powerful aid for the prediction of disease pathways, key functional players, candidate genes for rare variant discovery or sites for therapeutic intervention. In this respect, one of the underlying assumptions when doing network analysis is the "guilt by association principle"; here, the function of a node is inferred from the functions of its connected nodes (neighbours) [44]. The "network parsimony principle" summarizes another important assumption used in network analysis, for which the shortest path across (disease) relevant nodes is supposed to be indicative of the disease molecular pathway. An additional approach to identify

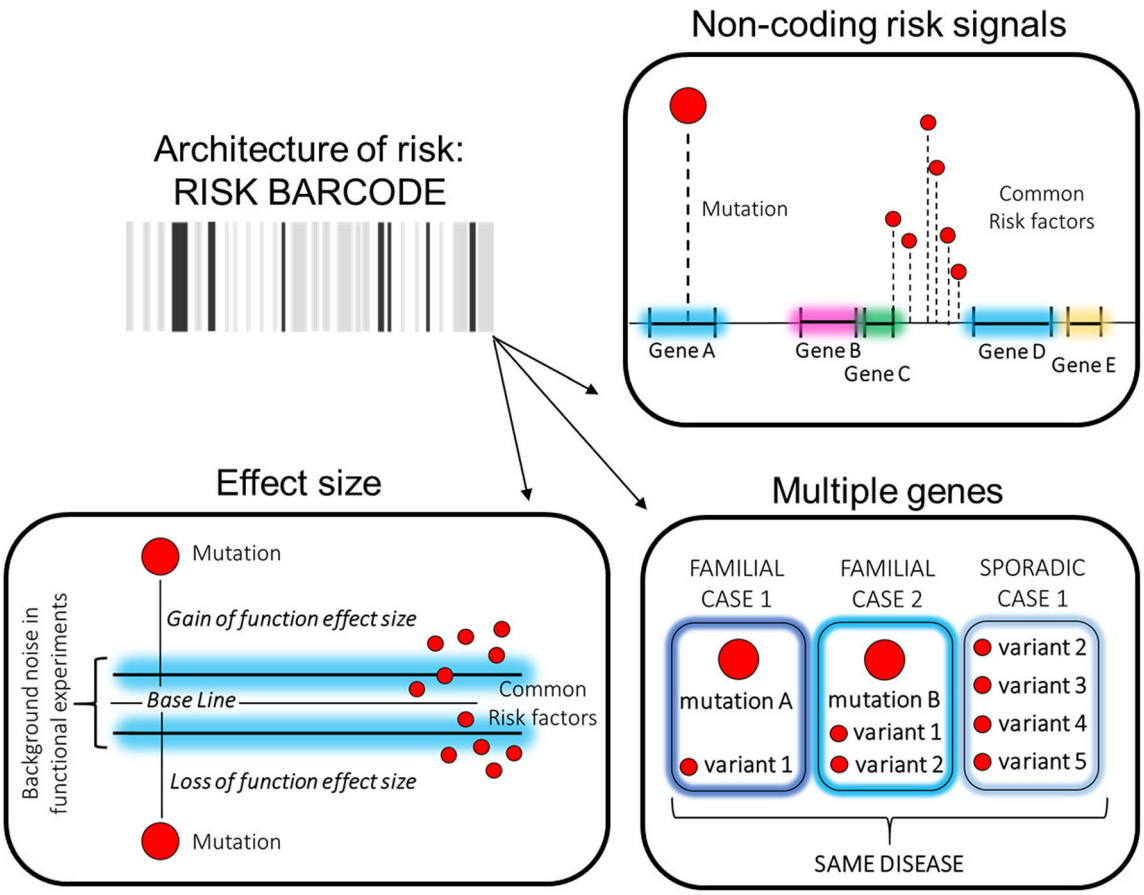

Fig. 1 The genetic architecture of disease can be graphically schematized by a "risk-barcode" where each line represents a risk factor that can be either a genetic variant or an environmental exposure. Lines have different thicknesses to represent different levels of contribution (strength or effect size) of each single component to the final disease risk. The principal problems in modelling the genetic architecture of risk with classical functional approaches are due to (i) common risk factors which are usually non-coding variants thus not immediately associated with any specific gene, (ii) common risk factors which have small effect sizes (strength) that are likely to fall below the sensitivity threshold of common functional experiments, (iii) modelling multiple risk factors concomitantly in the same model system which has proven challenging and sometimes impractical

relevant regions of the network for understanding how the flux of information moves within the graph and how this can be modified during disease is the detection/analysis of both motifs (peculiar concatenations of nodes) [45] and modules (representing portions of the network identified as discrete clusters because of shared homogenous characteristics). This lead to another network analysis principle called "local hypothesis", for which nodes involved in the same function (or disease) tend to share interactions and cluster within the same network module(s) [36••, 41].

Complex Neurodegenerative Diseases: Too Many Genes

As indicated above, in familial cases of complex neurodegenerative diseases, it is possible to identify mutation(s) with high effect size in so called Mendelian gene(s). It is noteworthy that different/multiple genes can be isolated in familial cases, and that all of them contribute to the pathogenesis of the same disease. For example, familial PD strongly associates with mutations in at least 7 different genes [46]; in familial frontotemporal dementia (FTD), at least 10 different mutated genes are associated with disease (despite some of them being extremely rare within the FTD population) [47]. It follows that a number of challenging questions arise, i.e why do many different (mutated) genes trigger a cascade of biological events that lead to the same clinical phenotype? One possibility is that, despite apparent differences, there is a limited number of common functions/pathways impacted in disease pathogenesis.

Classically, the effect of pathogenic mutations in familial genes has been investigated through knock-out/down models or in systems carrying one of the disease mutations (genetically modified models or patient-derived cells). Therefore, mutated genes have mainly been studied in isolation; rarely mixed models have been used to correlate the action of 2 or 3 genes. For example, *LRRK2* (frequently mutated in familial

PD) has been evaluated both in isolation (very frequently) and in hybrid models (rarely) in synergy with other familial PD genes such as *SNCA* [48] or *VPS35* [49, 50] showing that these genes might indeed be part of communal molecular patterns of disease. It must be noted that this type of studies can be expensive and technically challenging as classical functional biology is not well equipped to model multiple genes at the same time. Similarly, there are many mouse models for AD developed by modifying only one single gene, while very few models are available as double transgenic (to study concomitant mutations in *APP* and *PSEN1*) or triple transgenic (to study concomitant mutations in *APP*, *PSEN1* and *MAPT*) [51]. Network analysis has become an ever-increasing popular in silico approach to identify and prioritize communal pathways shared across "disease genes", thus helping to shed light onto molecular mechanisms of disease and assisting disease modelling. Results from network analyses still need confirmation in the functional environment; however, networks offer a time- and cost-effective approach to inform wet lab research. Specifically, networks allow for a more holistic support of disease modelling and help in focusing resources on the most promising functional targets.

Different network-based approaches have been developed, yet, generally, they can be categorised in 2 major groups. The bottom-up group comprises those approaches that "build the network up" starting from the genes under investigation. Conversely, the top-down methods build a larger and unbiased network in the first instance and then map the genes of interested onto it.

Our group has contributed to the bottom-up approaches by developing a pipeline named weighted protein-protein interaction network analysis (WPPINA); here, PPIs were used to build a multiple layers interactome for each of the familial genes for both FTD and PD. The single interactomes were subsequently merged into a final network (familial network for PD and familial network for FTD). Graph theory was applied to extract inter-interactome hubs (IIHs) that are those nodes responsible for keeping graph cohesion. IIHs were then used to successfully identify communal (and discriminative) pathways to disease via functional and pathway enrichment [52, 53].

Dervishi et al. applied a similar bottom-up approach to the study of amyotrophic lateral sclerosis (ALS). After selecting a number of distinct seed genes associated with disease, they used protein interactions (through Ingenuity, QIAGEN Comp, LA, USA) to build an ALS network used to: "suggest how different gene mutations converge into significant perturbations in protein interaction domains" [54]. Similarly, Beltran et al. applied the PIN approach through Ingenuity on an input set composed of copy number variations (CNVs) and additional genes differently associated with ALS. This was instrumental to identify a number of core genes in ALS-associated subnetworks, disease pathways and mechanisms to be further functionally validated [55]. A top-down approach was used in AD by building a whole human PPI network to function as background for inferring an AD-specific protein sub-network [56] for conjunct functional analysis of AD genes and prediction of additional AD gene candidates. Another top-down approach has been investigated by Kahle et al., in which they firstly generated a PIN for ataxia genes. Subsequently, they integrated the literature-derived information with primary interaction data, experimentally obtained for selected ataxia proteins. They parsed medical records of patients with ataxia to identity comorbidities and finally evaluated whether proteins implicated in the comorbid conditions were present within the ataxia interactome and how connections among these proteins were structured. Such strategy was instrumental in shedding light onto the biological origin of the comorbidity and shared mechanisms across diseases [57].

Ghiassian et al. investigated the concept of "disease module" and, by analysing how proteins involved in disease are typically linked together within the structure of the network, they developed a pipeline (DIAMOnD) to detect disease modules within PINs for specific disease phenotypes [58].

GCNs have been applied in the form of weighted gene co-expression network analysis (WGCNA [59]) to the study familial FTD. Here, expression profiles from different regions of the brain relevant for disease were analysed, and gene co-expression was analysed through permutation. Clusters (modules) of highly co-expressed genes were identified, and familial genes for FTD were mapped onto those modules prior topological and functional evaluation highlighting impacted biological pathways in different brain regions [60].

Gilman et al. [61] used a hybrid network approach (network-based analysis of genetic associations (NETBAG)) to simultaneously analyse all the genes affected by CNVs in autism to prioritize and suggest biological processes and pathways at the basis of the disorder. The hybrid network was built with the entire set of human genes as nodes and considering edges (connectivity) based on shared Gene Ontology annotations (from GO), KEGG pathways interaction partners and co-evolutionary patterns. Genes with autism-associated CNVs were then mapped on the network and used to identify strongly connected clusters to be studied. This permitted the evaluation of the entirety of CNVs alterations in one step and the assessment of their functional relevance in a genome-wide context.

Inferring Disease Genes from the Genetic Architecture of Risk

The genetic architecture at the basis of complex neurodegenerative diseases is difficult to model. Genome-wide association (GWA) analysis is, very frequently, the technique of choice to evaluate the contribution of small effect size variants (distributed in the entire genome) to a complex trait [62].

GWA findings have to be validated in model systems to provide functional information on mechanisms leading to

disease. However, the translation of genetics into the functional understanding of disease is challenging (Fig. 1). One of the issues with GWA types of studies is that signals do not necessarily pinpoint genes, rather regions of the genome (i.e. loci) that increase risk of disease, yet it is very difficult to understand which is the actual gene that is modulated by the disease-associated risk variants and what biological function is altered and eventually responsible for disease. Historically, researchers have suggested the open reading frame (ORF) closest to the risk signal to be the associated causal gene; this has, over the years, possibly generated type I error interpretation regarding the identity of the actual gene(s) modulated by the variant. Therefore, more recently, many groups are striving to establish pipelines to identify the real target genes of GWA variants.

Among the most successful approaches, there is the integration of genetic with quantitative trait loci (QTL) data [63]. The most popular form of QTL is expression QTL (eQTL), where the expression of cis-genes in relation to susceptibility markers is assessed (Fig. 2). Other QTL approaches, such as splicing QTL (sQTL), have been applied to determine alterations in splicing induced by the risk variant; methylation QTL (m-QTL) to verify the epigenetic change in methylation profile of nearby genes, and protein QTL (p-QTL) where the protein levels, rather than the RNA levels (as per eQTL), are evaluated. Clearly, not all the GWA signals can be explained via QTL analysis. This is possibly due to incompleteness of the omics databases needed for the analysis (tissue- and disease-specific data availability) or by methodological restrictions (e.g. in evaluating trans-QTLs [64–66] or QTLs resulting from a combination of multiple variants [67]). In this respect, again, network approaches have been proposed to be combined with QTLs to improve prioritization of modulated genes at the GWA loci.

Our group applied PIN analysis to PD-GWA signals. We firstly identified pathways shared by familial genes for PD using PIN; we then mapped the ORFs in linkage disequilibrium (LD) with the GWA risk variants onto those pathways and the PIN. The rationale was that those ORFs whose protein product was present in the network, and was involved in at least one of those pathways, were to be prioritized as gene candidates [30].

Alternatively, Voineagu et al. generated a GCN for autism using RNA profiling of post-mortem brain tissues identifying 2 relevant modules, enriched in neuronal and glia markers, respectively, to be correlated with disease. Co-expression modules where then tested for enrichment of autism-associated signals showing that GWA data converged on the neuronal module only [68]. With a similar approach, Seyfried et al. reported GCNs (obtained through WGCNA applied to proteomic profiling of human brain cortical tissue) descriptive of expression changes of both asymptomatic and symptomatic AD. GWA signals were first linked to ORFs through gene set analysis (thus, generating a single *p* value for each ORF present at the risk loci). Then, significant ORFs were overlapped with GCN modules to infer specific pathways correlated with disease progression [69].

Future Directions

The pressing request for approaches able to handle increasingly large (omics) and complex (multi-omics) sets of data has been the driving force behind the development of tailored network analyses in biomedicine. This is consequence of networks being relatively simple yet powerful tools for biological data inference. Machine learning (ML) has started to support network analysis [70]. ML is referred to as a computational approach where a machine is set-up to recognize patterns in a dataset and to increase its accuracy by correction over process reiteration (learning) [71]. ML is used for building networks; many techniques for inferring edges in GRNs have been developed as ML approaches; for example, ML can power the identification of DNA patterns and transcriptional factors binding sites in large datasets. Alternatively, ML can be applied to the analysis of

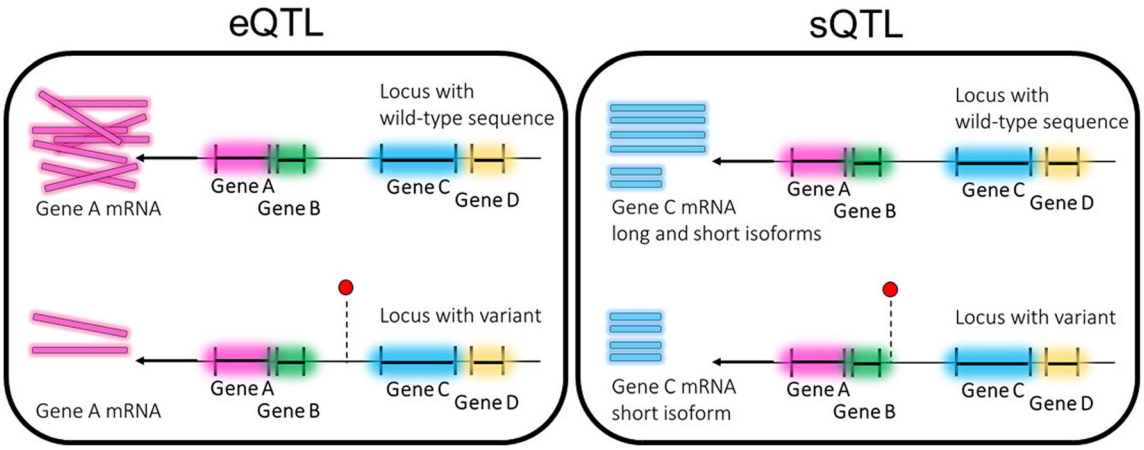

Fig. 2 Examples of a case of eQTL and a case of sQTL. eQTL, the presence of the variant at the locus affects the amount of mRNA (in this case reduction) produced for gene A, the variant is directly affecting the expression level of gene A. sQTL, the presence of the variant at the locus affects the splicing of gene C, in this case the long isoform is no more produced due to the presence of the variant

graphs. The potential of ML to efficiently detect re-occurring patterns of connections (motifs and network architecture) or identify similarities leading to node segregation (clustering) is starting to be investigated [72]. For example ML has been used to identify alterations in specific gene expression patterns indicative of candidate genes for cancer [73, 74] and predict PPIs based on protein pair features [75], as well as for dimensionality reduction after GO functional enrichment.

Conclusions

The research community is witnessing a very productive moment in biomedicine, experiencing an exponential growth in the amount of data that is generated with many initiatives taking place to improve the way we analyse data to extract biologically meaningful information to be translated for the benefit of medical practice. Of course, even if the computational power, the statistical approaches and the mathematics of graph theory are available, such paradigm shift in basic and applied research is still in its infancy. There still are levels of complexity that need to be overcome; for example, networks are more static than dynamic objects, where both edges and nodes can reconfigure themselves as in the real biological context [76•], and many omics datasets still lack that critical cell specificity type of information that would be necessary to draw more comprehensive functional conclusions. A specific initiative called Dialogue for Reverse Engineering Assessment and Methodology (DREAM) challenge (http://dreamchallenges.org) has been launched in 2006 as a crowdsourcing effort, where teams from all over the world are competing to develop the best performing pipelines to address compelling, big data problems in biomedicine. Analytical pipelines are being generated at a fast pace; however, these will need to stand the test of time; particularly, the next critical step will be validating the in silico findings, thus develop useful functional systems to model disease and highlight efficient endpoints for therapeutic drug intervention.

Funding Information This work was supported by the Medical Research Council (grant nos. MR/N026004/1; MR/L010933/1 to PAL); the Biomarkers Across Neurodegenerative Diseases Grant Program 2019, BAND3 (Michael J Fox Foundation, Alzheimer's Association, Alzheimer's Research UK and Weston Brain Institute, grant no. 18063 to CM and PAL); and Alzheimer's Society (grant no. 284 to RF).

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflicts of interest associated with this manuscript.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Papers of particular interest, published recently, have been highlighted as:
- • Of importance
- •• Of major importance

1. Wetterstrand KA. 2019. https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost.
2.• Manzoni C, Kia DA, Vandrovcova J, Hardy J, Wood NW, Lewis PA, et al. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Brief Bioinform. 2018;19(2):286–302. https://doi.org/10.1093/bib/bbw114 **Review paper surveing the basics of principal techniques, applications and pitfalls in genomics, transcriptomics and proteomics; accessible to readers with different background knowledge and students.**
3.•• Perez-Riverol Y, Zorin A, Dass G, Vu MT, Xu P, Glont M, et al. Quantifying the impact of public omics data. Nat Commun. 2019;10(1):3512. https://doi.org/10.1038/s41467-019-11461-w **Description of the Omics Discovery Index tool with interesting points of discussion regarding data availability and impact of datasets policies/ethics.**
4. Rooman M, Dehouck Y, Kwasigroch JM, Biot C, Gilis D. What is paradoxical about Levinthal paradox? J Biomol Struct Dyn. 2002;20(3):327–9. https://doi.org/10.1080/07391102.2002.10506850.
5.• Wang RS, Maron BA, Loscalzo J. Systems medicine: evolution of systems biology from bench to bedside. Wiley Interdiscip Rev Syst Biol Med. 2015;7(4):141–61. https://doi.org/10.1002/wsbm.1297 **Comprehensive review on the evolution of systems biology into systems medicine and systems pharmacology originated as integration and analysis of multiple-fields, large sets of data with translational examples.**
6. Vollstedt EJ, Kasten M, Klein C, Group MGGPsDS. Using global team science to identify genetic Parkinson's disease worldwide. Ann Neurol. 2019;86(2):153–7. https://doi.org/10.1002/ana.25514.
7. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, et al. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet. 2007;39(10):1181–6. https://doi.org/10.1038/ng1007-1181.
8. Lappalainen I, Almeida-King J, Kumanduri V, Senf A, Spalding JD, Ur-Rehman S, et al. The European genome-phenome archive of human data consented for biomedical research. Nat Genet. 2015;47(7):692–5. https://doi.org/10.1038/ng.3312.
9. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary

statistics 2019. Nucleic Acids Res. 2019;47(D1):D1005–D12. https://doi.org/10.1093/nar/gky1120.
10. Kay M, Clarke L, Santoyo-Lopez J, Maslen G, Siepel A, Cuomo C, et al. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):931–45.
11. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74. https://doi.org/10.1038/nature15393.
12. Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res. 2017;45(D1):D840–D5. https://doi.org/10.1093/nar/gkw971.
13. Erikson GA, Bodian DL, Rueda M, Molparia B, Scott ER, Scott-Van Zeeland AA, et al. Whole-genome sequencing of a healthy aging cohort. Cell. 2016;165(4):1002–11. https://doi.org/10.1016/j.cell.2016.03.022.
14. S SDaR. Encyclopedia of Bioinformatics and Computational Biology. Elsevier; 2019. p. 251–256.
15. Clough E, Barrett T. The gene expression omnibus database. Methods Mol Biol. 2016;1418:93–110. https://doi.org/10.1007/978-1-4939-3578-9_5.
16. Consortium GT. The genotype-tissue expression (GTEx) project. Nat Genet. 2013;45(6):580–5. https://doi.org/10.1038/ng.2653.
17. Ramasamy A, Trabzuni D, Guelfi S, Varghese V, Smith C, Walker R, et al. Genetic variability in the regulation of gene expression in ten regions of the human brain. Nat Neurosci. 2014;17(10):1418–28. https://doi.org/10.1038/nn.3801.
18. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 2018;46(D1):D794–801. https://doi.org/10.1093/nar/gkx1081.
19. Lizio M, Abugessaisa I, Noguchi S, Kondo A, Hasegawa A, Hon CC, et al. Update of the FANTOM web resource: expansion to provide additional transcriptome atlases. Nucleic Acids Res. 2019;47(D1):D752–D8. https://doi.org/10.1093/nar/gky1099.
20. Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, et al. Human body epigenome maps reveal noncanonical DNA methylation variation. Nature. 2015;523(7559):212–6. https://doi.org/10.1038/nature14465.
21. UniProt C. UniProt: a hub for protein information. Nucleic Acids Res. 2015;43(Database issue):D204–12. https://doi.org/10.1093/nar/gku989.
22. Orchard S, Kerrien S, Abbani S, Aranda B, Bhate J, Bidwell S, et al. Protein interaction data curation: the international molecular exchange (IMEx) consortium. Nat Methods. 2012;9(4):345–50. https://doi.org/10.1038/nmeth.1931.
23. Deutsch EW, Csordas A, Sun Z, Jarnuczak A, Perez-Riverol Y, Ternent T, et al. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. Nucleic Acids Res. 2017;45(D1):D1100–D6. https://doi.org/10.1093/nar/gkw936.
24. Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, et al. A draft map of the human proteome. Nature. 2014;509(7502):575–81. https://doi.org/10.1038/nature13302.
25. The Gene Ontology C. The gene ontology resource: 20 years and still GOing strong. Nucleic Acids Res. 2019;47(D1):D330–D8. https://doi.org/10.1093/nar/gky1055.
26. McKusick VA. Mendelian inheritance in man and its online version. OMIM Am J Hum Genet. 2007;80(4):588–604. https://doi.org/10.1086/514346.
27. Pinero J, Bravo A, Queralt-Rosinach N, Gutierrez-Sacristan A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2017;45(D1):D833–D9. https://doi.org/10.1093/nar/gkw943.
28. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The reactome pathway knowledge base. Nucleic Acids Res. 2018;46(D1):D649–D55. https://doi.org/10.1093/nar/gkx1132.
29. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30. https://doi.org/10.1093/nar/28.1.27.
30. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017;18(1):83. https://doi.org/10.1186/s13059-017-1215-1.
31. Arboleda-Velasquez FL JF, O'Hare M, Delgado-Tirado S, Marino C, Chmielewska N, Saez-Torres KL, et al. Resistance to autosomal dominant Alzheimer's disease in an APOE3 Christchurch homozygote: a case report. Nature Medicine. 2019. https://doi.org/10.1038/s41591-019-0611-3.
32. Velez JI, Lopera F, Sepulveda-Falla D, Patel HR, Johar AS, Chuah A, et al. APOE*E2 allele delays age of onset in PSEN1 E280A Alzheimer's disease. Mol Psychiatry. 2016;21(7):916–24. https://doi.org/10.1038/mp.2015.177.
33. Trinh J, Guella I, Farrer MJ. Disease penetrance of late-onset parkinsonism: a meta-analysis. JAMA Neurol. 2014;71(12):1535–9. https://doi.org/10.1001/jamaneurol.2014.1909.
34. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504. https://doi.org/10.1101/gr.1239303.
35. Serin EA, Nijveen H, Hilhorst HW, Ligterink W. Learning from co-expression networks: possibilities and challenges. Front Plant Sci. 2016;7:444. https://doi.org/10.3389/fpls.2016.00444.
36.•• Sonawane AR, Weiss ST, Glass K, Sharma A. Network medicine in the age of biomedical big data. Front Genet. 2019;10:294. https://doi.org/10.3389/fgene.2019.00294 **Comprehensive review on networks applied to different types of big-data with applicative examples to multiple aspects of biomedicine.**
37. V F. Encyclopedia of Bioinformatics and Computational Biology, vol. 1. Amsterdam: Elsevier; 2019. p. 915–21.
38. Delgado FM, Gomez-Vela F. Computational methods for gene regulatory networks reconstruction and analysis: a review. Artif Intell Med. 2019;95:133–45. https://doi.org/10.1016/j.artmed.2018.10.006.
39. Barbosa S, Niebel B, Wolf S, Mauch K, Takors R. A guide to gene regulatory network inference for obtaining predictive solutions: underlying assumptions and fundamental biological and data constraints. Biosystems. 2018;174:37–48. https://doi.org/10.1016/j.biosystems.2018.10.008.
40. Hawe JS, Theis FJ, Heinig M. Inferring interaction networks from multi-omics data. Front Genet. 2019;10:535. https://doi.org/10.3389/fgene.2019.00535.
41. Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12(1):56–68. https://doi.org/10.1038/nrg2918.
42. Costanzo M, Kuzmin E, van Leeuwen J, Mair B, Moffat J, Boone C, et al. Global genetic networks and the genotype-to-phenotype relationship. Cell. 2019;177(1):85–100. https://doi.org/10.1016/j.cell.2019.01.033.
43. Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS Comput Biol. 2007;3(4):e59. https://doi.org/10.1371/journal.pcbi.0030059.
44. Oliver S. Guilt-by-association goes global. Nature. 2000;403(6770):601–3. https://doi.org/10.1038/35001165.
45. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. Science. 2002;298(5594):824–7. https://doi.org/10.1126/science.298.5594.824.
46. Lill CM. Genetics of Parkinson's disease. Mol Cell Probes. 2016;30(6):386–96. https://doi.org/10.1016/j.mcp.2016.11.001.

47. Ferrari R, Manzoni C, Hardy J. Genetics and molecular mechanisms of frontotemporal lobar degeneration: an update and future avenues. Neurobiol Aging. 2019;78:98–110. https://doi.org/10.1016/j.neurobiolaging.2019.02.006.
48. Bae EJ, Kim DK, Kim C, Mante M, Adame A, Rockenstein E, et al. LRRK2 kinase regulates alpha-synuclein propagation via RAB35 phosphorylation. Nat Commun. 2018;9(1):3465. https://doi.org/10.1038/s41467-018-05958-z.
49. Inoshita T, Arano T, Hosaka Y, Meng H, Umezaki Y, Kosugi S, et al. Vps35 in cooperation with LRRK2 regulates synaptic vesicle endocytosis through the endosomal pathway in drosophila. Hum Mol Genet. 2017;26(15):2933–48. https://doi.org/10.1093/hmg/ddx179.
50. Mir R, Tonelli F, Lis P, Macartney T, Polinski NK, Martinez TN, et al. The Parkinson's disease VPS35[D620N] mutation enhances LRRK2-mediated Rab protein phosphorylation in mouse and human. Biochem J. 2018;475(11):1861–83. https://doi.org/10.1042/BCJ20180248.
51. Myers A, McGonigle P. Overview of transgenic mouse models for Alzheimer's disease. Curr Protoc Neurosci. 2019;89(1):e81. https://doi.org/10.1002/cpns.81.
52. Ferrari R, Kia DA, Tomkins JE, Hardy J, Wood NW, Lovering RC, et al. Stratification of candidate genes for Parkinson's disease using weighted protein-protein interaction network analysis. BMC Genomics. 2018;19(1):452. https://doi.org/10.1186/s12864-018-4804-9.
53. Ferrari R, Lovering RC, Hardy J, Lewis PA, Manzoni C. Weighted protein interaction network analysis of frontotemporal dementia. J Proteome Res. 2017;16(2):999–1013. https://doi.org/10.1021/acs.jproteome.6b00934.
54. Dervishi I, Gozutok O, Murnan K, Gautam M, Heller D, Bigio E, et al. Protein-protein interactions reveal key canonical pathways, upstream regulators, interactome domains, and novel targets in ALS. Sci Rep. 2018;8(1):14732. https://doi.org/10.1038/s41598-018-32902-4.
55. Beltran S, Nassif M, Vicencio E, Arcos J, Labrador L, Cortes BI, et al. Network approach identifies pacer as an autophagy protein involved in ALS pathogenesis. Mol Neurodegener. 2019;14(1):14. https://doi.org/10.1186/s13024-019-0313-9.
56. Hu YS, Xin J, Hu Y, Zhang L, Wang J. Analyzing the genes related to Alzheimer's disease via a network and pathway-based approach. Alzheimers Res Ther. 2017;9(1):29. https://doi.org/10.1186/s13195-017-0252-z.
57. Kahle JJ, Gulbahce N, Shaw CA, Lim J, Hill DE, Barabasi AL, et al. Comparison of an expanded ataxia interactome with patient medical records reveals a relationship between macular degeneration and ataxia. Hum Mol Genet. 2011;20(3):510–27. https://doi.org/10.1093/hmg/ddq496.
58. Ghiassian SD, Menche J, Barabasi AL. A DIseAse MOdule detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. PLoS Comput Biol. 2015;11(4):e1004120. https://doi.org/10.1371/journal.pcbi.1004120.
59. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559. https://doi.org/10.1186/1471-2105-9-559.
60. Ferrari R, Forabosco P, Vandrovcova J, Botia JA, Guelfi S, Warren JD, et al. Frontotemporal dementia: insights into the biological underpinnings of disease through gene co-expression network analysis. Mol Neurodegener. 2016;11:21. https://doi.org/10.1186/s13024-016-0085-4.
61. Gilman SR, Iossifov I, Levy D, Ronemus M, Wigler M, Vitkup D. Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. Neuron. 2011;70(5):898–907. https://doi.org/10.1016/j.neuron.2011.05.021.
62. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747–53. https://doi.org/10.1038/nature08494.
63. Nica AC, Dermitzakis ET. Expression quantitative trait loci: present and future. Philos Trans R Soc Lond Ser B Biol Sci. 2013;368(1620):20120362. https://doi.org/10.1098/rstb.2012.0362.
64. Brynedal B, Choi J, Raj T, Bjornson R, Stranger BE, Neale BM, et al. Large-scale trans-eQTLs affect hundreds of transcripts and mediate patterns of transcriptional co-regulation. Am J Hum Genet. 2017;100(4):581–91. https://doi.org/10.1016/j.ajhg.2017.02.004.
65. Clyde D. Disease genomics: transitioning from association to causation with eQTLs. Nat Rev Genet. 2017;18(5):271. https://doi.org/10.1038/nrg.2017.22.
66. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013;45(10):1238–43. https://doi.org/10.1038/ng.2756.
67. Zeng B, Lloyd-Jones LR, Holloway A, Marigorta UM, Metspalu A, Montgomery GW, et al. Constraints on eQTL fine mapping in the presence of multisite local regulation of gene expression. G3 (Bethesda). 2017;7(8):2533–44. https://doi.org/10.1534/g3.117.043752.
68. Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S, et al. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature. 2011;474(7351):380–4. https://doi.org/10.1038/nature10110.
69. Seyfried NT, Dammer EB, Swarup V, Nandakumar D, Duong DM, Yin L, et al. A multi-network approach identifies protein-specific co-expression in asymptomatic and symptomatic Alzheimer's disease. Cell Syst. 2017;4(1):60–72 **e4**. https://doi.org/10.1016/j.cels.2016.11.006.
70. Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ. Next-generation machine learning for biological networks. Cell. 2018;173(7):1581–92. https://doi.org/10.1016/j.cell.2018.05.015.
71. Xu C, Jackson SA. Machine learning and complex biological data. Genome Biol. 2019;20(1):76. https://doi.org/10.1186/s13059-019-1689-0.
72. Oyelade J, Isewon I, Oladipupo F, Aromolaran O, Uwoghiren E, Ameh F, et al. Clustering algorithms: their application to gene expression data. Bioinform Biol Insights. 2016;10:237–53. https://doi.org/10.4137/BBI.S38316.
73. Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, et al. An integrated approach to uncover drivers of cancer. Cell. 2010;143(6):1005–17. https://doi.org/10.1016/j.cell.2010.11.013.
74. Ghanat Bari M, Ung CY, Zhang C, Zhu S, Li H. Machine learning-assisted network inference approach to identify a new class of genes that coordinate the functionality of cancer networks. Sci Rep. 2017;7(1):6993. https://doi.org/10.1038/s41598-017-07481-5.
75. Chen KH, Wang TF, Hu YJ. Protein-protein interaction prediction using a hybrid feature representation and a stacked generalization scheme. BMC Bioinformatics. 2019;20(1):308. https://doi.org/10.1186/s12859-019-2907-1.
76.• Bassett DS, Sporns O. Network neuroscience. Nat Neurosci. 2017;20(3):353–64. https://doi.org/10.1038/nn.4502 **Comprehensive review on integration of neurobiology knowledge using network neuroscience, particularly interesting for the discussion on the current frontiers in network neuroscience research, unmet needs and future directions.**