Cell and Tissue Research
https://doi.org/10.1007/s00441-017-2734-5

REVIEW

Computational systems biology approaches for Parkinson's disease

Enrico Glaab[1]

Received: 24 July 2017 / Accepted: 6 November 2017

Abstract

Parkinson's disease (PD) is a prime example of a complex and heterogeneous disorder, characterized by multifaceted and varied motor- and non-motor symptoms and different possible interplays of genetic and environmental risk factors. While investigations of individual PD-causing mutations and risk factors in isolation are providing important insights to improve our understanding of the molecular mechanisms behind PD, there is a growing consensus that a more complete understanding of these mechanisms will require an integrative modeling of multifactorial disease-associated perturbations in molecular networks. Identifying and interpreting the combinatorial effects of multiple PD-associated molecular changes may pave the way towards an earlier and reliable diagnosis and more effective therapeutic interventions. This review provides an overview of computational systems biology approaches developed in recent years to study multifactorial molecular alterations in complex disorders, with a focus on PD research applications. Strengths and weaknesses of different cellular pathway and network analyses, and multivariate machine learning techniques for investigating PD-related omics data are discussed, and strategies proposed to exploit the synergies of multiple biological knowledge and data sources. A final outlook provides an overview of specific challenges and possible next steps for translating systems biology findings in PD to new omics-based diagnostic tools and targeted, drug-based therapeutic approaches.

Keywords Parkinson's disease · Systems biology · Pathway analysis · Network analysis · Bioinformatics

Introduction

Parkinson's disease (PD) is one of the most common age-related, neurodegenerative disorders. In spite of 200 years of research on PD since its first published description by James Parkinson (Parkinson 1817), the disease etiology is still not fully understood. No disease-modifying therapy is available and no reliable diagnostic and progression biomarkers have so far been identified. The lack of a detailed molecular understanding and comprehensive mechanistic models for disease initiation and progression may at least in part be explained by the striking heterogeneity and complexity of the disease, which is manifested by a wide variety of motor and non-motor symptoms (Jankovic 2008; Solla et al. 2012; Müller et al. 2013; Kalia and Lang 2015). Recent genetic and epidemiological findings suggest that this high clinical heterogeneity is also reflected by a multitude of diverse PD risk factors and complex interplays between them (Gorell et al. 2004; Dardiotis et al. 2013; Kieburtz and Wunderle 2013). Known genetic influences include more than 20 loci associated with familial forms of PD and several risk factor variants identified for idiopathic PD (Kalinderi et al. 2016). Since about 15% of patients have a first-degree relative with PD (Samii et al. 2004) and only about 6–7% of an estimated total heritability of around 27% can be explained by the currently known PD-associated genetic variants (Do et al. 2011), several further genetic or epigenetic alterations may be involved in PD. This heritable component of the disease is complemented by multiple environmental risk factors implicated in PD etiology by epidemiological or Mendelian randomization studies, including exposure to toxic environmental agents, head injuries, and various drugs and dietary factors (Bellou et al. 2016). In analogy to the 'dual-hit' hypothesis previously proposed for other complex disorders (Knudson 1971), interplays of different factors may cause the disease and modulate the onset and severity of symptoms.

While studies on the influences of individual causal and risk-associated factors still represent an important information

✉ Enrico Glaab
enrico.glaab@uni.lu

[1] Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 7 avenue des Hauts Fourneaux,
L-4362 Esch-sur-Alzette, Luxembourg

Published online: 29 November 2017

source, there is widespread agreement in the field that, in order to account for the ‘missing heritability’ in PD as well as the large proportion of idiopathic patients without a family history of PD, potential combinatorial effects of multiple genetic variations and/or environmental factors should be modeled and validated. Due to the large number of possible relevant molecular factors, an integrative modeling is not feasible using targeted experimental measurements and classical statistical methods alone, but additionally requires dedicated systems biology approaches, using high-throughput omics profiling techniques and bioinformatics approaches that exploit prior biological knowledge for data analysis.

This review presents a structured overview of current computational systems biology methods available for PD research, discusses their specific limitations and benefits, and highlights some of their recent applications in PD-related studies. First, methods for the analysis of PD-associated cellular pathway and molecular process alterations are compared, then related network analysis and causal reasoning approaches for identifying key regulatory factors are introduced. Next, machine learning approaches to build models for diagnostic sample classification and patient sub-group stratification are presented, including bioinformatics methods that exploit prior biological domain knowledge for integrative analyses. Because these approaches and their previous applications still suffer from several limitations, which have so far prevented the design of biomarker models for PD with sufficient accuracy, robustness and reproducibility, specific restrictions of prior work in this field are highlighted. As a final outlook, a discussion of PD-specific challenges and potential next steps for systems biology-based biomarker development and drug target identification is provided.

Analyzing disease-associated activity changes in cellular pathways

A common first step towards understanding systems-level changes in omics datasets for complex diseases like PD is the investigation of molecular activity alterations in the context of known cellular pathways and molecular processes. For this purpose, a multitude of manually curated pathway and process definitions are available in public databases, including the Kyoto Encyclopedia of Genes and Genomes (KEGG; Ogata et al. 1999), the Gene Ontology database (GeneOntologyConsortium 2004), BioCarta (Nishimura 2001), WikiPathways (Pico et al. 2008), Reactome (Joshi-Tope et al. 2005) and the Pathway Interaction Database (Schaefer et al. 2009). Moreover, in addition to these generic pathway repositories, disease-specific resources have been established in recent years, providing dedicated pathway maps for the neurodegenerative disorders PD (see the PDMap; Fujita et al. 2014) and Alzheimer’s disease (see AlzPathway; Mizuno et al. 2012). When using a large and generic pathway database rather than a smaller selection of putatively relevant pathways for identifying disease-associated cellular process changes in an omics dataset, one has to consider that the final significance scores for an analysis will need to be adjusted for the number of tested hypotheses (equal to the number of pathways) to prevent excessive false positive discoveries (Benjamini and Hochberg 1995). Accordingly, using prior biological knowledge to pre-filter the considered pathways can be an effective strategy to increase the statistical power for showing significant associations.

Apart from selecting a pathway collection, researchers also need to choose between a wide range of statistical analysis approaches. In general, these omics-based pathway and geneset enrichment analysis methods can be grouped into four main categories (combining classifications previously proposed by Huang et al. 2009 and Di Lena et al. 2015):

1. Over-representation analysis (ORA): These approaches quantify the statistical over-representation of a list of genes, proteins or metabolites among the members of a pathway using a statistical test (e.g., Fisher’s exact test). The input list usually corresponds to the biomolecules which displayed a differential abundance in an omics dataset between a condition of interest (e.g., a disease state) as compared to a control condition, according to a chosen test statistic and significance threshold.
2. Geneset enrichment analysis (GSEA): GSEA methods avoid the need for defining a significance threshold and instead assign ranking scores to all biomolecules in the analyzed omics data to test whether the members of a pathway are ranked unexpectedly high or low among them (e.g., using modified versions of the Kolmogorov–Smirnov test).
3. Network module-based pathway analysis (NMPA): These algorithms exploit prior knowledge from molecular interaction networks to improve the scoring of pathway associations for omics profiling data. Typical NMPA methods first identify dense sub-network regions enriched in biomolecules undergoing activity changes (called “modules”), and, in a second step, quantify associations of these network modules with known pathways.
4. Network topology-based pathway analysis (NTPA): Similar to NMPA approaches, NTPA methods exploit molecular network information to obtain more robust and sensitive pathway association scores, but they avoid the initial module identification step and directly quantify pathway associations using graph-based statistics to assess the network distances and multiplicity of interconnections between the biomolecules of interest and pathway members mapped onto the network.

Table 1 shows an overview of representative, publicly available software tools for each of these four pathway

Table 1 Publicly available software tools and web-applications for analyzing cellular pathway activity changes in omics datasets; some of the methods can be applied directly in the web browser (see column 4), and some of the tools provide advanced visualization features to facilitate the interpretation of the results (see column 5)

Method type	Software name	Availability	Web application	Visualization features	Reference
Over-representation analysis (ORA) tools	DAVID	https://david.ncifcrf.gov	Yes	No	Dennis et al. 2003
	GOstat	http://gostat.wehi.edu.au	Yes	Yes	Beißbarth and Speed 2004
	OntoExpress	http://vortex.cs.wayne.edu/ontoexpress	Yes	No	Draghici et al. 2003
	GoMiner	https://discover.nci.nih.gov/gominer	Yes	Yes	Zeeberg et al. 2003
	GOToolBox	http://genome.crg.es/GOToolBox	Yes	No	Martin et al. 2004
Geneset enrichment analysis (GSEA) tools	GSEA	http://software.broadinstitute.org/gsea	No	Yes	Subramanian et al. 2005
	GAGE	http://bioconductor.org/packages/release/bioc/html/gage.html	No	No	Luo et al. 2009
	GSA	http://statweb.stanford.edu/~tibs/GSA	No	No	Efron and Tibshirani 2007
	PAGE / PGSEA	https://www.bioconductor.org/packages/release/bioc/html/PGSEA.html	No	No	Kim and Volsky 2005
	GLOBALTEST	https://bioconductor.org/packages/release/bioc/html/globaltest.html	No	Yes	Goeman et al. 2004
	PADOG	http://bioconductor.org/packages/release/bioc/html/PADOG.html	No	No	Tarca et al. 2012
Network module-based pathway analysis (NMPA)	FunMOD	https://sourceforge.net/projects/funmodnetwork	No	Yes	Natale et al. 2014
	PINA	http://cbg.garvan.unsw.edu.au/pina	Yes	Yes	Cowley et al. 2012
	ReactomeFIViz	http://wiki.reactome.org/index.php/ReactomeFIViz	No	Yes	Wu et al. 2014
Network topology-based pathway analysis (NMPA)	PWEA	https://zlab.bu.edu/PWEA	No	Yes	Hung et al. 2010
	SPIA	http://bioconductor.org/packages/release/bioc/html/SPIA.html	No	Yes	Tarca et al. 2009
	PathNet	http://bioconductor.org/packages/release/bioc/html/PathNet.html	No	No	Dutta et al. 2012
	DeGraph	https://bioconductor.org/packages/release/bioc/html/DEGraph.html	No	Yes	Jacob et al. 2012
	EnrichNet	http://www.enrichnet.org	Yes	Yes	Glaab et al. 2012
	Ontologizer	http://ontologizer.de	Yes	Yes	Bauer et al. 2008
	SANTA	http://bioconductor.org/packages/release/bioc/html/SANTA.html	No	Yes	Cornish and Markowetz 2014
	ToPASeq	https://bioconductor.org/packages/release/bioc/html/ToPASeq.html	No	Yes	Ihnatova and Budinska 2015

analysis categories, including information on whether the tools are available as platform-independent web applications and whether they enable a visualization of the results.

When selecting a particular method among these choices, the following common limitations and benefits specific to different types of approaches should be considered: While the results of ORA methods are easy to calculate and interpret, they depend on the definition of a significance threshold and may not detect pathways enriched in many small molecular changes. By contrast, GSEA approaches do not require the specification of a significance cut-off and can identify pathways affected by strong cumulative effects of many small alterations. However, GSEA results are often difficult to interpret, and, as in ORA methods, the molecular network topology interconnecting the biomolecules of interest is not taken into consideration, since the statistics rely exclusively on available pathway annotations. This limitation is addressed by NMPA and NTPA approaches, which exploit information from gene regulatory, protein–protein or protein–metabolite interaction networks in order to increase the statistical power and robustness for identifying pathway-associated, co-ordinated network activity changes. Importantly, these software tools can account for the regulatory influences of biomolecules that have not yet been annotated for any known pathway. Moreover, they enable intuitive network visualizations, which can facilitate biological data interpretation. However, in contrast to pure network analysis methods (see the following section), NMPA and NTPA are hybrid approaches that combine aspects of both network and pathway analysis methods, and provide pathway rankings as the main output rather than altered sub-networks without known pathway annotations. As an important limitation, this also means that NMPA and NTPA approaches will not identify altered network regions that cannot be linked to any known cellular pathway. Moreover, a potential drawback in comparison to classical pathway analysis methods is that NMPA and NTPA statistics often rely heavily on the correctness and comprehensiveness of the underlying molecular interaction data. Similar to classical enrichment analyses, biases, noise, errors and incompleteness of the data used for network-based enrichment analyses can result in false-negative and false-positive findings. While sufficient high-quality molecular interaction data is typically available for the human species and common model organisms like mouse, rat, baker's yeast and fruit fly, corresponding interaction data resources for other studied organisms may still be too incomplete for an effective application of these network analyses. Finally, similar to ORA approaches, NMPA and NTPA methods rely on differential expression thresholds, which need to be defined by the user.

The choice of a suitable pathway analysis approach is further complicated by the fact that many methods additionally require a prior computation of differential expression or abundance scores for the individual biomolecules in the studied omics data. This can be achieved using classical statistical approaches (e.g., the parametric *t* test or the non-parametric Mann–Whitney *U* test) or moderated statistical tests with improved feature variance estimation (Smyth 2004; Demissie et al. 2008), and by subsequently adjusting the *P* value significance scores for multiple hypothesis testing (Benjamini and Hochberg 1995). Discussions of these statistics for assessing changes in individual biomolecules and benchmark comparisons have been provided previously (Cui and Churchill 2003; Rapaport et al. 2013). Additionally, for pathway analyses of GWAS and sequencing datasets, specific technical issues have to be addressed, e.g., biases related to linkage disequilibrium, gene length and geneset size (see the discussion and guidelines by Wang et al. 2011 and Rahmatallah et al. 2015).

As a general recommendation for omics-based pathway analyses, it may often be helpful to compare at least a few of the above-mentioned alternative types of approaches, in order to identify different forms of biologically relevant alterations (e.g., pathways affected by few changes with large effect size/high significance, by many changes with small effect size/low significance, or by co-ordinated alterations in a specific sub-network of a pathway).

One of the first prominent examples for the application of pathway analysis approaches for PD research was a GSEA-based case-control study of post-mortem transcriptomics data from the midbrain (*substantia nigra*), using a weighted meta-analysis to combine effect size estimates for pathway-representing genesets across multiple independent datasets (Zheng et al. 2010). This analysis identified significant PD-associated alterations in 28 pathways, including 10 pathways subsequently validated in early subclinical cases of PD and in other PD-affected brain regions. Since the underexpression of a geneset of PGC-1α–responsive genes was significantly associated with PD pathology in this meta-analysis, the authors investigated PGC-1α over-expression as a new therapeutic strategy and reported that it suppressed dopaminergic neuron loss in two cell culture models of PD (Zheng et al. 2010).

A similar integrative pathway analysis study combined ORA statistics for PD-related GWAS and brain transcriptomics data to identify consensus pathway alterations across these two data modalities (Edwards et al. 2011). The authors used an unweighted meta-analysis approach (Fisher's combined probability test) to integrate the significance scores for the genetic and gene expression data, and reported shared significant changes in multiple pathways, including the top three processes *axonal guidance*, *focal adhesion* and *calcium signaling*.

Apart from these studies focusing on human datasets, the integrated pathway-based analysis of data from PD-related animal models and human biospecimens has also been explored. By applying GSEA to 33 microarray datasets from human and animal model studies on PD, Oerton and Bender could show that the concordance across studies between summarized activity changes at the pathway-level was significantly higher than for individual differentially expressed genes (see fig. 2 in Oerton and Bender 2017). While only some animal model datasets revealed comparable changes to those in human studies, this study highlights that pathway analyses can help to address discrepancies between related omics studies at the level of single biomolecules due to technical and biological variance, and identify higher-level shared significant alteration signatures.

Finally, pathway analyses may also provide an effective means for cross-disease comparisons and for studying the molecular influences of factors associated with disease risk (e.g., aging, diet and toxin exposure). For example, an NTPA method revealed shared transcriptomics pathway alterations in the brain in PD and during adult aging (Glaab and Schneider 2015), and common inflammatory process changes in PD and Huntington's disease were recently identified in a comparative pathway analysis of mRNA-seq data using a GSEA approach (Labadorf et al. 2017).

In summary, a wide choice of pathway analysis tools is available to study systems-level alterations in complex diseases, and their previous applications to PD-related omics data have already led to new insights on the processes affected by disease-related changes.

Analyzing disease-associated molecular network alterations

While pathway-centric analyses can greatly facilitate the biological interpretation of omics data, the available public pathway definitions are often incomplete, may contain errors due to false-positive experimental discoveries, and inconsistencies can occur between subjectively defined boundaries for the same pathway across different databases (e.g., the "p53 signaling pathway" in KEGG differs significantly from the identically named pathway in the BioCarta database). As an alternative or extension to investigations based on pre-defined pathways, molecular network analyses have the potential to provide more detailed, comprehensive and novel findings for systems-level omics investigations. Network analyses do not require a time-consuming prior curation of cellular process annotations and avoid subjective judgments on the relevance of specific genes/proteins for a particular molecular function. They can exploit an extensive resource of interaction data from public databases, including STRING (Szklarczyk et al. 2015), BioGrid (Chatr-Aryamontri et al. 2015), IntAct (Kerrien et al. 2012), MINT (Licata et al. 2012), HPRD (Keshava Prasad et al. 2009) and HIPPIE (Schaefer et al. 2012), which cover significantly more biomolecular interactions than existing pathway databases.

However, a drawback of network analysis methods is that the results are often difficult to interpret; in particular, when the molecular changes of interest occur in a sub-network with few functionally annotated genes and no links to any known pathway. For this reason, hybrid approaches have been developed to combine the benefits of pathway and network analyses, e.g., algorithms to automatically extend existing pathway definitions via a graph-theoretic analysis of a surrounding genome-scale interaction network (Li et al. 2017). For network analyses in general, care must be taken to avoid biases: if data from small-scale protein interaction profiling studies is included in the network assembly, then frequently studied disease-related proteins may be biased to have larger numbers of identified interactors than other proteins. Therefore, either only data from genome-scale interaction profiling studies should be used or dedicated methods to reduce bias influences during the statistical sub-network analysis should be applied (e.g., see Ung et al. 2016).

Since a comprehensive discussion of biological network analysis approaches would extend beyond the scope of this review, only two of the most common method types are introduced here:

1. Network perturbation analyses (NPA): These methods aim to identify sub-networks within a genome-scale molecular or regulatory network that undergo co-ordinated activity changes in a biological condition of interest. Such co-ordinated network changes are characteristic for complex diseases, which tend to involve perturbations in the activity of entire molecular network regions rather than only in a few genes or proteins (Ideker and Sharan 2008; del Sol et al. 2010). NPA approaches typically consist of a search algorithm that heuristically explores the space of possible disease-affected sub-networks, and a scoring function that quantifies the overall significance and effect size of molecular changes in omics data mapped onto a sub-network. The final outcome of an NPA procedure is a ranking of the sub-networks with the most pronounced and robust alterations in the condition of interest as compared to a control condition.
2. Causal reasoning analyses (CRA): Causal reasoning (or causal network analysis) approaches use manually curated directional relationships, e.g., gene regulatory relationships or protein signaling cascades, to infer the root molecular causes for a set of observed condition-specific downstream changes in an omics dataset. While these directional relationships are often referred to as "causal relationships", the underlying data are mostly correlational rather than causal and have to be interpreted with caution. By constructing a signed, directed interaction graph (often referred to as "causal graph" in the literature) from a list of known directional relationships between interacting molecules, a CRA method can track back through the graph from the molecules that underwent measured activity alterations in the omics data to their known or putative upstream regulators. These regulators are then scored as potential drivers of the observed downstream changes by evaluating the overall consistency of the activating and inhibitory regulation patterns in the graph with the measured data (see Chindelevitch et al. 2012). CRA studies enable the discovery of key regulatory molecules controlling specific biological processes of interest, e.g., a disease-related process.

NPA and CRA methods are complementary methodologies with related purposes. NPA approaches help researchers to identify disease-associated co-ordinated activity changes across multiple biomolecules in a specific molecular network region, which may provide robust biomarker signatures for diagnostic applications. By contrast, CRA methods are mainly useful for identifying single upstream regulators with altered activity, which are responsible for a large fraction of observed downstream pathological changes and therefore of interest as potential drug targets for preclinical intervention studies. As a limitation, CRA software can only be applied to regulatory networks (represented by graphs with directed edges), whereas NPA tools are applicable to both regulatory and molecular interaction networks (represented by directed and undirected graphs). However, as illustrated in Fig. 1, for regulatory networks, the first steps of NPA and CRA approaches—consisting of the statistical omics data analysis, the network assembly and data mapping—are often identical or similar, so that NPA and CRA algorithms can be combined effectively within a single analysis pipeline. An overview of publicly available software tools for NPA and CRA is provided in Table 2, highlighting which of the tools provide network visualization features as opposed to a pure ranking functionality. Due to the high computational cost of most genome-scale network analyses, none of these tools are currently available as installation-free web applications; however, most of them can be installed on common desktop operating systems.

For network analyses in general, the result quality largely depends on how complete and correct the underlying network is, and corresponding methods may therefore not be applicable to model organisms with limited publicly available regulatory and interaction data. However, certain network analysis approaches, like the CRE method (see Table 2 and Chindelevitch et al. 2012), have been shown to be robust against considerable levels of noise in the input data.

In spite of the fact that many NPA and CRP algorithms have only been developed recently, multiple studies have already employed these approaches for PD molecular research. For example, Hu et al. (2017) manually curated the literature to define a set of 242 genes with previously reported genetic associations with PD, and investigated this geneset on a global human interactome using an NPA approach (the Steiner minimal tree algorithm, which is also implemented in the software *BioNet*; see Dittrich et al. 2008; Beisser et al. 2010). This resulted in the inference of a sub-network with PD-specific alterations, including new potential disease-related genes (Hu et al. 2017). In another study focusing on predictive network modeling and using transcriptomics data from the midbrain (*substantia nigra*) from PD cases and controls, the machine learning-based NPA approach, GenePEN, identified a connected sub-network signature within a genome-scale protein–protein interaction network with significant predictive power to distinguish between biospecimens from patients and unaffected subjects (Vlassis and Glaab 2015). As a further interesting application, PD-related brain transcriptomics data and an NPA method have been used to propose new functional links between microRNAs and PD, as well as new possible regulatory mechanisms for disease initiation and neuroprotection (Chandrasekaran and Bonchev 2013). Moreover, a

Fig. 1 Overview of common steps in molecular network analyses of disease-related omics data

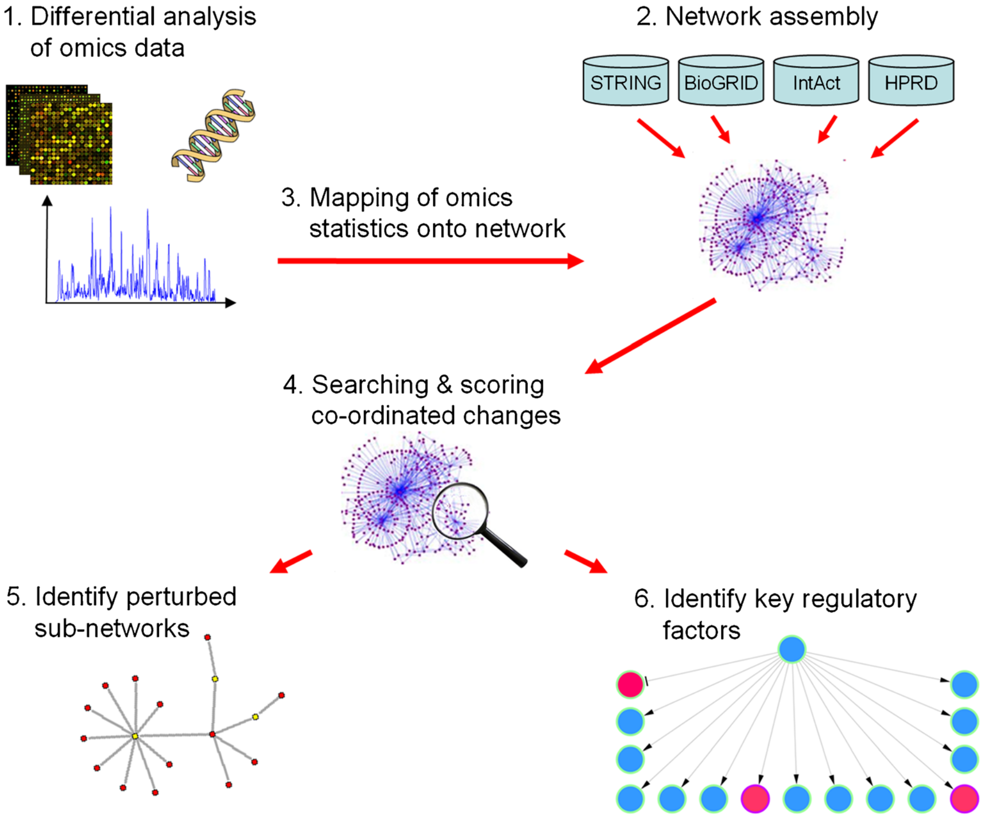
1. Differential analysis of omics data
2. Network assembly
STRING
BioGRID
IntAct
HPRD
3. Mapping of omics statistics onto network
4. Searching & scoring co-ordinated changes
5. Identify perturbed sub-networks
6. Identify key regulatory factors

Table 2 Publicly available software tools for identifying sub-network perturbations and key regulatory biomolecules in omics datasets; some of the methods can be applied directly in the web-browser (see column 4), and some of the tools provide advanced visualization features to facilitate the interpretation of the results (see column 5)

Method type	Software name	Availability	Visualization features	Reference
Network perturbation analysis (NPA)	BioNet / HEINZ	http://www.bioconductor.org/packages/release/bioc/html/BioNet.html	Yes	Dittrich et al. 2008; Beisser et al. 2010, Dennis et al. 2003
	WMAXC	http://combio.gist.ac.kr/WMAXC/WMAXC.html	No	Amgalan and Lee 2014, Beißbarth and Speed 2004
	jActiveModules	http://apps.cytoscape.org/apps/jactivemodules	Yes	Ideker et al. 2002
	PinnacleZ	http://apps.cytoscape.org/apps/pinnaclez	Yes	Chuang et al. 2007
	COSINE	http://cran.r-project.org/web/packages/COSINE	Yes	Ma et al. 2011
	GenePEN	http://lcsb-portal.uni.lu/software/index.html	No	Vlassis and Glaab 2015
	MCWalk	https://bitbucket.org/akittas/biosubg	Yes	Kittas et al. 2016
	ClustEx	http://bioinfo.au.tsinghua.edu.cn/member/jgu/clustex	Yes	Gu et al. 2010
	BMRF	https://sourceforge.net/projects/bmrfcjava/	Yes	Chen et al. 2013
Causal reasoning analysis (CRA)	CRE	R source code available upon request from the author	Yes	Chindelevitch et al. 2012
	Whistle	https://github.com/Selventa/whistle	No	Catlett et al. 2013
	CausalR	https://bioconductor.org/packages/release/bioc/html/CausalR.html	Yes	Bradley and Barrett 2017
	QuaternaryProd	https://www.bioconductor.org/packages/release/bioc/html/QuaternaryProd.html	No	Fakhry et al. 2016
	BayesCRE	source code available upon request from the author	Yes	Zarringhalam et al. 2013
	MCWalk	https://bitbucket.org/akittas/biosubg	Yes	Kittas et al. 2016
	SigNet	https://cbdd.clarivate.com/cbdd	Yes	Jaeger et al. 2014

first exemplary causal reasoning study subdivided differentially active pathways between brain transcriptomics samples from PD patients and controls into upstream and downstream processes, and ranked them hierarchically to propose new hypotheses on important upstream pathological alterations (Fu and Fu 2015). This approach suggested specifically that RNA metabolism pathology might be an upstream causal driver of PD pathogenesis.

Recently, network analysis techniques have also been employed as a means to compare PD to other complex diseases. Hypothesizing a relationship between PD and diabetes, Santiago and Potashkin (2013) mapped genes with known genome-wide significance in PD- and diabetes-related GWAS studies onto a human functional gene linkage network and identified a cluster of 478 genes closely associated with the seed genes for both diseases. Using a similar approach, Calderone et al. (2016) discovered shared and non-shared sub-networks associated with PD and Alzheimer’s disease, based on starting lists of genes derived from the public resources *PDMap* (Fujita et al. 2014) and *AlzPathway* (Mizuno et al. 2012). They then used functional and topological similarity measures to relate these sub-networks to biological processes in the Gene Ontology database, which pointed to associations with DNA repair, RNA metabolism and glucose metabolism, that could not be detected by a classical pathway enrichment analysis.

In summary, network perturbation and causal reasoning analyses are emerging as valuable complementary tools to conventional pathway analyses for the study of molecular changes in complex diseases. When significant pathological or protective activity changes occur in molecular sub-networks that still lack associated pathway annotations, only network analysis approaches are able to detect these alterations and predict new disease-associated processes and their main upstream regulators for subsequent experimental validation.

Generating predictive machine learning models and visualizing high-dimensional data

One of the primary goals behind systems-level analyses of omics data for complex diseases is to identify biomarker signatures for differential diagnosis, patient sub-group stratification or disease prognosis. Generic machine learning software for diagnostic sample classification and clustering for patient sub-group stratification can often be applied ‘out-of-the-box’, without adaptations in the algorithms, to preprocessed omics data. However, in recent years, machine learning approaches that exploit prior biological domain knowledge as additional information source have been developed, which tend to provide more accurate, robust and biologically interpretable models than the classical generic methods (Fang et al. 2006; Lottaz et al. 2007).

Predictive model building typically starts with a feature selection or feature transformation step, eliminating uninformative attributes from the input omics data (e.g., by removing biomolecules with low activity variation across the samples) or combining the original attributes into more robust derived features (e.g., pathway-representing features, using weighted sums of measurements for pathway member biomolecules). These approaches are also called "dimension reduction methods", because they reduce the number of dimensions of the input data (equal to the number of features) in order to address multiple common statistical issues during following analyses, previously summarized under the notion "curse of dimensionality" (Bellman 1961; Köppen 2000). Moreover, these methods enable the generation of low-dimensional visualizations of the data, e.g., 2D and 3D perspective plots, facilitating outlier detection and biological data interpretation.

Table 3 provides an overview of dedicated software tools for machine learning analyses of omics data, including multi-purpose tool sets for sample clustering (unsupervised analysis) and classification (supervised analysis), software centered around the ranking and selection of informative attributes, and data visualization approaches (since a great variety of algorithms and implementations are publicly available, the table only highlights a representative selection with a focus on tools designed for systems biology data analysis). To illustrate how different types of methods can be interlinked within one analysis pipeline, Fig. 2 shows a common generic workflow.

A major benefit of machine learning and visualization techniques for the analysis of omics data is their broad applicability. While different functional omics data types require different lower-level pre-processing methods, the higher-level machine learning and visualization tools discussed here are applicable across almost all types of pre-processed molecular data and often also support the integrative analysis of diverse omics types. Thus, given a pre-processed functional omics dataset, e.g., normalized microarray data, RNAseq read counts or mass-spectrometry-derived protein or metabolite abundance data, a wide variety of machine learning tools can be applied directly to identify clustering patterns (using unsupervised analyses) to build predictive models for the classification of new data samples (using supervised analyses), or to visually explore and interpret the data (using dimension reduction and visualization methods).

It is often recommendable to start the analysis of a normalized omics dataset with a simple visualization before using machine learning tools for automated pattern identification. Inspecting a 2D or 3D projection of the data can often reveal the presence of outliers, biases and other irregularities which are not always detected by automated quality control pipelines. The most commonly used approach for obtaining a low-dimensional visual representation of high-dimensional omics data is a principal component analysis (PCA). A benefit of PCA visualizations is that by design they tend to capture most of the variance in the data. However, in contrast to other dimension reduction methods like multidimensional scaling (MDS; Torgerson 1952), PCA is not designed to preserve original pairwise distances between data points when transforming the data to a low-dimensional space. Moreover, both PCA and MDS are linear approaches, which only tend to preserve distances between dissimilar data points in their low-dimensional representations, but in a linear data mapping it may often be impossible to keep highly similar data points close together (Van Der Maaten and Hinton 2008). Researchers may therefore want to consider some of the more recently developed non-linear data visualization approaches that focus on preserving local structure, e.g., locally linear embedding (Roweis and Saul 2000), Laplacian Eigenmaps (Belkin and Niyogi 2003) and t-SNE (Van Der Maaten and Hinton 2008).

After visual exploration of the omics data and the potential removal of outlier samples, the next steps for a machine learning analysis depend on the researcher's specific goals and the availability of condition labels or outcome measures for the samples: if only unlabeled data with no related outcome measures are available, or if the analysis goal is to find distinct sub-groups among the samples (e.g., to stratify patients with distinct molecular alteration patterns), then unsupervised clustering approaches should be applied. These methods will identify sub-groups of samples that are similar to each other in terms of their omics profiles but differ significantly from other identified sub-groups. The relevant algorithms can be grouped into hierarchical clustering methods (e.g., hybrid hierarchical clustering, Chipman and Tibshirani 2006; Bayesian hierarchical clustering, Heller and Ghahramani 2005; and self-organizing maps, Ritter and Kohonen 1989), partition-based approaches (e.g., k-Means, Hartigan and Wong 1979; k-Mediods, Kaufman and Rousseeuw 1987; partitioning around mediods, Kaufman and Rousseeuw 1990) and density-based techniques (e.g., DBSCAN, Ester et al. 1996; DENCLUE, Hinneburg and Keim 1998; Chameleon, Karypis et al. 1999). While a detailed discussion of these methods and their biomedical applications extends beyond the scope of this review, a corresponding overview and guideline for algorithm selection has been provided previously (Andreopoulos et al. 2009).

A significant limitation of these generic clustering analyses of high-dimensional omics data is that, even after filtering the attributes by variance, many uninformative clustering patterns may still occur in the data. These are not necessarily spurious patterns, but may reflect real biological differences of the studied biospecimens (e.g., differences in gender, age, and

Table 3 Overview of public software tools for predictive model building, clustering analysis and dimension reduction and visualization of omics data

Method type	Software name	Availability	Supported features[a]	Web application	Reference
Multi-purpose machine-learning analysis tool sets	CARMAWeb	https://carmaweb.genome.tugraz.at/carma/	N, P, C, D, V	Yes	Rainer et al. 2006
	ArrayMining	http://www.arraymining.net	N, P, C, D, V	Yes	Glaab et al. 2009
	mixOmics	https://cran.r-project.org/web/packages/mixOmics	P, D, V	No	Rohart et al. 2017
	Weka	http://www.cs.waikato.ac.nz/ml/weka	P, C, D, V	No	Hall et al. 2009
	Orange	https://orange.biolab.si	P, C, D, V	No	Demšar et al. 2013
	CMA	https://bioconductor.org/packages/release/bioc/html/CMA.html	P, D, V	No	Slawski et al. 2008
	MLSeq	https://bioconductor.org/packages/release/bioc/html/MLSeq.html	P, D	No	Zararsiz et al. 2014
Tools centered around feature ranking/feature selection	Limma	https://bioconductor.org/packages/release/bioc/html/limma.html	N, D, V	No	Smyth 2005
	RankProd	https://bioconductor.org/packages/release/bioc/html/RankProd.html	D, V	No	Hong et al. 2006
	ArrayPipe	http://www.pathogenomics.ca/arraypipe	N, D, V,	Yes	Hokamp et al. 2004
	RAP	https://bioinformatics.cineca.it/rap	N, D, V	Yes	D'Antonio et al. 2015
	EzArray	http://ezarray.com	N, D, V	Yes	Natale et al. 2014
Tools for low-dimensional data visualization	GGobi	http://www.ggobi.org	D, V	No	Temple Lang and Swayne 2001
	PlotViz	http://salsahpc.indiana.edu/plotviz	D, V	No	Choi et al. 2010
	RnavGraph	https://cran.r-project.org/web/packages/RnavGraph	V	No	Waddell and Oldford 2011
	Arena3D	http://arena3d.org	V	No	Secrier et al. 2012

[a] Column 3 highlights the supported features of the tools using the following codes: *N* normalization/preprocessing, *P* predictive model building, *C* unsupervised clustering, *D* dimension reduction (variable selection or feature transformation), *V* visualization. Tools available as web applications are highlighted in column 4

dietary habits reflected by different biomolecular profiles) that could overshadow unrelated biomedically relevant differences between patient sub-groups of interest (e.g., disease sub-types with different treatment responses). Therefore, more recent approaches integrate prior biological knowledge into the cluster analysis, e.g., using gene/protein functional annotations and information from disease-related pathways, in order to aggregate measurements for functionally related, disease-associated biomolecules and determine more robust and relevant clustering patterns (Fang et al. 2006; Lottaz et al. 2007). For the subsequent evaluation of clustering results, no standard approach is available, but a variety of cluster validity indices have been proposed and should ideally be considered in combination (Kovács et al. 2005; Rendón et al. 2011; Arbelaitz et al. 2013). In general, ideal clusterings of patient biospecimens are characterized by low within-cluster distances and high between-cluster distances, are biologically interpretable and biomedically relevant, and replicable across different cohorts.

If class labels or quantitative outcome measures are available for the studied omics samples, reflecting biological conditions of interest (e.g., patient vs. control, or known disease sub-types) or measures of disease severity (e.g., scores from the Unified Parkinson's Disease Rating Scale; Goetz 2003), then predictive models for diagnostic biospecimen classification can be built from the data by applying supervised machine learning approaches (relevant software tools are listed in Table 3 and highlighted by the code "P" for "prediction" in the third column). These algorithms use a set of omics data, called the "training set", with known values for a chosen outcome variable, to identify patterns that enable a prediction of the outcome for new, unlabeled omics samples. By first applying a machine learning approach on the training set to generate a predictive mathematical function that relates patterns in the data to the outcome measure of interest, and then testing this predictive model on an independent set of omics samples with known outcomes (called the "test set"), the accuracy, sensitivity and specificity of the model can be

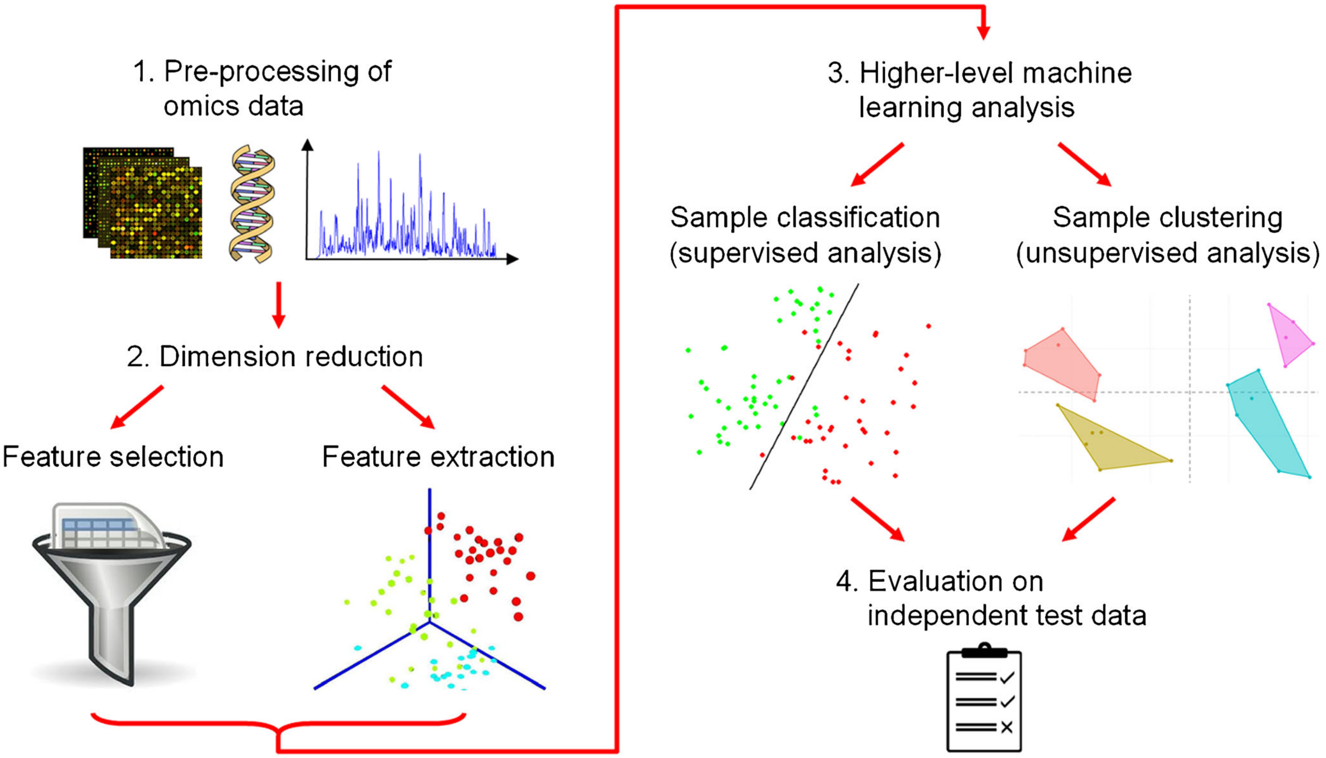
1. Pre-processing of omics data
2. Dimension reduction
Feature selection
Feature extraction
3. Higher-level machine learning analysis
Sample classification (supervised analysis)
Sample clustering (unsupervised analysis)
4. Evaluation on independent test data

Fig. 2 Common generic workflow for a machine learning analysis of omics data, including steps to reduce the dimension of the data through feature selection or feature extraction, higher-level machine learning analysis for classifying omics data samples (a supervised analysis) or clustering the samples (an unsupervised analysis), and evaluation of the obtained machine learning models on external test data

estimated. More detailed guidelines on how to optimize machine-learning models using cross-validation and bootstrap procedures and how to evaluate the model performance on external tests set have already been provided elsewhere (Browne 2000; Braga-Neto and Dougherty 2004). Importantly, in particular when combining attribute selection methods with predictive machine learning, care must be taken to avoid selection biases (Wood et al. 2007).

In addition to classical generic statistical learning methods, new machine learning approaches guided by prior domain knowledge have been developed in recent years. These algorithms use dedicated multi-objective optimization approaches, which optimize the generated prediction models both by minimizing the training set error and by maximizing the consistency of the model with prior biological knowledge, or exploit biology-inspired data structures like ontology graphs or molecular networks for a structured data integration of multiple omics datasets. Representative examples for these types of approaches include the sparse overlapping Group Lasso approach for integrative multi-omics analysis (Park et al. 2015), which identifies driver genes in a biomedical omics datasets based on prior biological knowledge derived from predefined overlapping groups of features (e.g., gene functions in the Gene Ontology database), the network-constraint regularization approach for machine learning analysis of omics data by Li and Li (2008), and the multi-omics analysis approach by Mosca and Milanesi (2013), using a multi-objective optimization procedure to drive the identification of network regions enriched in molecular alterations across multiple omics data sources. A review by Li et al. (2016b) provides a more detailed overview on corresponding methods that exploit prior biological knowledge for integrative machine learning analysis of omics data.

In PD research, previous machine learning applications on omics data have mainly focused on diagnostic biomarker discovery in cerebrospinal fluid (CSF) and whole-blood samples. These efforts were motivated by the observation that even after the onset of visible motor symptoms, the currently used clinical diagnostic criteria for PD (UK Parkinson's Disease Society Brain Bank criteria) only reached around 76% specificity in recent studies (increasing to 82% with retrospective application and 90% at death in a follow-up study; Berg et al. 2013). Importantly, while omics-based biomarker signatures for PD could in principle enable a more objective and accurate diagnosis, it is important to highlight that the previously proposed signatures have mostly not been reproduced or do not provide sufficient sensitivity and specificity for practical diagnostic purposes. This may largely be explained by limitations arising from small sample sizes, high biological and technical variance in the data, biases in the experimental procedures and instruments used for omics profiling (e.g., related to the machine, kit, experimenter, library or lane), no filtering of treatment/medication effects, no adjustment for common confounding factors, missing control samples for other neurodegenerative disorders, and the application of inadequate model building and validation techniques that result in over-fitted prediction models. Therefore, previous research on omics-based biomarker models for PD still represents preliminary work that has to be interpreted cautiously, and major technical and methodological challenges still have to be overcome to obtain clinically useful biomarker signatures.

A first representative metabolomics profiling study on blood samples from 66 PD patients and 25 unaffected controls reported a signature with 100% correct separation (Bogdanov et al. 2008). However, no cross-validation and no independent test set validation was performed, and, although some of the metabolite markers could be linked to known PD-associated processes, e.g., oxidative stress, the robustness and replicability of the prediction model has not yet been verified. Therefore, further study is needed, also in order to evaluate the extent to which the overall signature reflects PD-specific or generic disease-associated changes (e.g., blood inflammation-related markers are altered in many disorders).

Recently, a new metabolomics signature in CSF was proposed by Trezzi et al., using a non-targeted gas chromatography-mass spectrometry approach to study the CSF metabolome of 44 early-stage, untreated idiopathic PD patients in comparison with 43 age- and gender-matched unaffected controls (Trezzi et al. 2017). By applying a logistic regression approach, a machine learning model was trained to discriminate between patients and controls and tested on two independent validation sets ($n = 18$ and $n = 38$). The model involved the three metabolites mannose, threonic acid, and fructose as predictive features and was reported to provide a sensitivity of 0.79 and a specificity of 0.8. Additional studies including patients with other neurological disorders and larger numbers of samples from multiple cohorts are still needed to assess the predictive value of the signature for differential diagnosis.

Apart from metabolomics and proteomics signatures, gene expression changes in blood have also been considered as possible biomarkers for PD. Molochnikov et al. investigated gene transcription in blood samples from 62 early-stage PD patients and 64 unaffected controls and built a predictive model using stepwise multivariate logistic regression (Molochnikov et al. 2012). The resulting five-gene classification model was tested on an independent cohort of 30 advanced stage PD patients and 29 Alzheimer's disease patients and separated them with 100% accuracy. However, no unaffected controls and no atypical forms of PD as disease control were included in the study validation set. Since Alzheimer's disease is not associated with any motor symptoms similar to PD, assessing the potential of the proposed model for differential diagnosis of similar movement disorders will require additional investigations, including more disease conditions and larger sample sizes.

A further transcriptomics signature for PD was proposed by Scherzer et al., who used whole-blood microarray expression data from 105 subjects, covering 50 patients with early motor-stage PD, 33 control subjects with other neurological disorders and 22 unaffected controls (Scherzer et al. 2007). Their multigene marker was built by ranking and selecting genes in terms of their absolute Pearson correlation with binary sample class labels (representing PD vs. all controls), forming a template for each class from the mean values of the discriminating genes, and then defining a combined risk score for new biospecimen measurements corresponding to their Pearson correlation with the PD template minus its Pearson correlation with the non-PD template. The resulting signature was further validated in 39 independent test samples, but has so far not been replicated by independent research groups.

While most PD biomarker discovery approaches focus on data from idiopathic PD (IPD) patients, an interesting alternative approach using an integrative analysis of whole-blood gene expression data from IPD patients, familial PD patients with the *LRRK2* G2019S mutation and different mouse models was presented by Chikina et al. (2015). By first identifying differentially expressed genes between four groups of mice (overexpressing wild-type LRRK2, overexpressing G2019S LRRK2, LRRK2-knockout and wild-type mice) and combining them with previously proposed PD marker genes from the literature, a panel of 113 candidate marker genes was assembled and their expression measured for 34 symptomatic PD patients (both wild-type LRRK2 and G2019S LRRK2) and 32 asymptomatic controls using a digital gene expression platform. This led to the discovery of a subset of 14 markers discriminating between PD patients and asymptomatic controls with a reported accuracy of 79%. However, similar to other PD biomarker studies, no neurological disorder controls were included in the analysis, and further studies are required to determine whether the gene signature provides a significant informative value for differential diagnosis or whether it reflects a more generic inflammation response that may also occur in other disorders.

More recently, Shamir et al. presented a whole-blood gene expression signature for idiopathic PD, derived from microarray data analysis of 486 subjects ($n = 205$ PD, $n = 233$ controls, $n = 48$ other neurodegenerative diseases) (Shamir et al. 2017). Using batch-effect reduction and cross-validation procedures to prevent overfitting, their machine learning model included signatures of 100 genetic probes and was reported to reach a significant predictive performance on an independent validation cohort [area under the curve (AUC) = 0.79, $P =$ 7.13E-6] and a subsequent independent test cohort (AUC = 0.74, $P = 4.2$E-4). The model was trained to differentiate between PD and unaffected controls rather than between PD and other neurologic disorders, and further analyses are needed to evaluate the potential of extending the model towards differential diagnostic applications, reducing the number of required genetic probes and increasing the generalization performance.

Multivariate machine learning methods have not only been applied for the analysis of PD-specific omics data but also for cross-disease comparisons between PD and other neurodegenerative disorders. In an exemplary study by Potashkin et al., splice variant-specific microarrays were used to find markers discriminating between whole-blood samples from 51 PD patients, 17 patients with multiple systems atrophy (MSA), 17 patients with progressive supranuclear palsy (PSP) and 39 unaffected controls (Potashkin et al. 2012). When applying a linear discriminant analysis to test the predictive accuracy of a signature of 13 selected differentially expressed, PD patients were reported to be distinguished from all controls with 96% sensitivity and 90% specificity and from the combined MSA and PSP patients with 94% sensitivity and 96% specificity. Seven of the 13 candidate markers were later confirmed to be dysregulated in PD on an independent set of whole-blood samples from 50 PD patients and 46 unaffected controls as part of a follow-up study by the authors (Santiago et al. 2013). While a major benefit of this work is that the baseline study considered two atypical forms or parkinsonism, MSA and PSP, in addition to PD, the signature has not yet been replicated by independent investigators and larger sample sizes for the neurodegenerative disorder controls will be required in future studies to evaluate the utility of the signature for differential diagnosis more precisely and robustly.

A further cross-disease comparative machine learning analysis presented by Abdi et al. (2006) involved a multiplex quantitative proteomics method, iTRAQ (isobaric tagging for relative and absolute protein quantification), applied in conjunction with multidimensional chromatography, followed by tandem mass spectrometry (MS/MS). This experimental procedure was used to compare the cerebrospinal fluid (CSF) proteome in patients with PD ($n = 10$), Alzheimer's disease ($n = 10$), dementia with Lewy body ($n = 5$) and unaffected controls ($n = 10$). The authors determined a multifactorial marker signature using logistic regression, which was reported to provide a sensitivity of 78% and a specificity of 95% for discriminating between PD and the other disorders. Given the limited sample sizes in this study and the lack of an external replication, the authors acknowledge that their preliminary findings will have to be validated in a larger and independent population of patients.

Finally, a comparative machine learning analysis across multiple neurodegenerative disorders has also been performed by Ishigami et al., who used MALDI-TOF profiling of CSF peptides and proteins from 37 PD patients, 32 MSA patients and 26 control subjects with other neurological disorders (OND) (Ishigami et al. 2012). They applied a PCA for dimension reduction in combination with a support vector machine algorithm for supervised sample classification, and reported average cross-validated classification accuracies of 90.2% for distinguishing PD versus MSA and 98.2% for PD versus OND. The authors acknowledged that the sample size was small, and no independent replication has so far been conducted. Thus, additional external validation is still required to assess the generalization performance of this model.

Unsupervised machine learning approaches for patient sub-group identification in PD research have so far mainly been applied to clinical data. A first data-driven approach to characterize the heterogeneity in PD via clustering techniques was presented by Graham and Sagar (1999), who collected clinical information for 176 idiopathic patients and applied k-Means clustering to the normalized continuous variables. Their analysis suggested a separation of patients into three sub-groups at a disease duration of 5.6 years, and two sub-groups at 13.4 years. The identified sub-groups mainly differed in terms of measures of motor control and complications, age at onset and the degree of cognitive impairment. Similar studies by other research groups suggested a variety of different patient sub-groups: A two-group separation into rapid and slow progression (Gasparoli et al. 2002), a mild and a severely impaired group in terms of motor dysfunction and cognition (Dujardin et al. 2004), a young and an old onset group (Schrag et al. 2006), a three-group separation (Post et al. 2008), four alternative clusterings into four groups (Lewis et al. 2005; Reijnders et al. 2009; Liu et al. 2011; Van Rooden et al. 2011) and one five-group clustering (Lawton et al. 2015). The differences in the number and characteristics of the estimated clusters in these previous studies may mainly be explained by differences in the underlying patient cohorts and the considered features (e.g., only the study by Dujardin included SPECT measurements, and the clinical variables used across the studies differed significantly). While some of the proposed sub-types were reported to be reproduced in independent cohorts (Lewis et al. 2005; Reijnders et al. 2009; Van Rooden et al. 2011), in most studies no quantitative cluster validity index analyses were provided. Overall, further study is still warranted to derive and evaluate clinically relevant classification algorithms for PD patient sub-groups. A detailed review of stratification analysis results obtained so far, including recommendations on how to translate the gained knowledge into PD clinical research, has been provided by Marras and Lang (2013).

In summary, the previous application of machine learning methods for stratification and biomarker profiling analysis of PD suggest that multiple distinct sub-groups are present, and that significant disease-associated alterations occur in both CSF and blood. Given the limited sample sizes and restrictions in the types of neurodegenerative disorder controls available for biomarker profiling, further assessments are needed to determine whether the proposed signatures can be translated into clinically relevant tests for differential diagnosis with high robustness, sensitivity and specificity. Similarly, current stratification studies are partly hampered by restrictions in terms of the number and types of quantitative features considered, and in terms of the external statistical validation of clustering results. Future studies could address these limitations by combining further data types for robust cluster pattern identification, by assessing cluster correlations with independent clinically relevant variables and using additional quantitative external validations.

Outlook on challenges and possible next steps for systems biology-based biomarker development and drug target discovery for PD

In recent years, the discovery of multiple PD-causing mutations and risk factor variants and the growth of public data resources for PD research, e.g., through the Parkinson's Progression Markers Initiative (www.ppmi-info.org), have provided new means to pinpoint the main affected cellular pathways and gain a more detailed understanding of pathological changes in the disease. In order to translate the resulting knowledge and research efforts into improved diagnostic models and preclinical drug intervention studies, a variety of challenges still have to be overcome.

Since the midbrain (*substantia nigra*) is regarded as the main affected tissue in PD and only post-mortem omics data are available for this brain region, one of the main challenges for omics-based biomarker modeling is to identify reliable surrogate markers in peripheral tissues or body fluids. One possible strategy to address this in the future could be to use the non-lesional access to the brain during deep brain stimulation (DBS) surgery, by capturing cells spontaneously adhering to the DBS stylet for omics profiling (Zaccaria et al. 2016), and correlating these profiles to corresponding molecular measurements in blood samples. Using pathway and network analysis approaches discussed in this review, blood–brain correlations could not only be assessed at the level of individual biomolecules but also via pathway or sub-network activity scores to establish more robust correlations. A further strategy to explore could be the combined analysis of measurements for peripheral markers with limited specificity, e.g., biomarkers for oxidative stress in blood, with neuroimaging and clinical data using integrative machine learning methods. Classification models trained on individual data types could be combined via model averaging techniques (Dietterich 2000), or standardized features from the different data sources could be used to train a single, integrative prediction model. This synergistic modeling may help to address limitations of the individual data modalities and provide more robust and sensitive diagnostic models. Moreover, integrative analyses may reveal new interrelations across the different data types.

In order to obtain clinically relevant and reliable biomarker signatures for differential diagnosis, a further important objective for the future is to compare omics measurements for PD to sufficiently large sample sizes for atypical forms of PD and related neurodegenerative disorders. While it is challenging to recruit large numbers of atypical PD patients for a study, ongoing work on integrating information across different disease cohorts may help to address this issue.

A related hurdle in diagnostic model building is the general lack of statistical power in many studies. The high heterogeneity among PD patients, discussed earlier in this review, increases the variance in omics measurements and decreases the power to detect significant differences between patients and controls. Moreover, for PD, the number of publicly available omics data samples is much smaller as compared to Alzheimer's disease and many cancer diseases. Larger sample sizes in combination with dimension reduction techniques and integrative analyses of multiple omics types will help to increase the power to identify new statistically significant PD-associated alterations and build more accurate prediction models. Moreover, computational approaches for combinatorial selection of biospecimens from a biobank for molecular profiling, designed to attain an optimal matching between patient and control samples in terms of multiple known confounding factors (age, gender, body-mass index, co-morbidities, smoking and dietary habits), provide a further possibility to increase the statistical power for comparative analyses at no added cost, but they are rarely used in practice. Bioinformatics methods can also facilitate biomarker discovery for the early pre-motor phase of PD, e.g., by combining analyses of molecular data from in vitro an in vivo models of early-stage PD with measurements from biospecimen of untreated de novo patients to pinpoint shared early-stage disease-associated changes. These integrative analyses could complement ongoing studies on the follow-up of at-risk cohorts, applying omics profiling analyses to biospecimens collected prior to the conversion to PD to discover presymptomatic molecular dysregulations.

For the specific goal of modeling and estimating the future progression of PD using omics data, time-series measurements from longitudinal studies will need to be collected in larger quantities and probably also at smaller time intervals. Current longitudinal studies for PD typically involve between one to two follow-up investigations per year. While important changes in the disease course may occur during shorter time periods, the burden for the patient through blood draws and clinical assessments needs to be minimized. Since many longitudinal studies so far only collect limited molecular data, a more comprehensive molecular phenotyping may currently deserve higher priority than narrowing the time interval between follow-up investigations. For the statistical analysis of time series data from corresponding studies, similar strategies to increase the statistical power can be applied as discussed for cross-sectional analyses in this review, e.g., using dimension reduction approaches, prior knowledge on interrelationships between biomolecules from pathways/networks and the literature, and integrative omics analyses to identify coordinated alteration trends over time. Representative examples of relevant time series analysis approaches for omics data have been presented by Wachter and Beißbarth (2014) and Lee et al. (2016).

Apart from the exploration of new omics measurements for biomarker modeling, the same data will also provide an important resource for systems biology analyses dedicated to the discovery, validation and characterization of PD drug targets. Network analyses including the causal reasoning approaches discussed in this review can help to identify pathological activity alterations in key regulatory proteins, and provide a starting point to prioritize candidate protein drug targets for further analyses. These investigations can be integrated with other more generic *in silico* target prioritization approaches (Aerts et al. 2006; Chen et al. 2009; Isik et al. 2015) and algorithms for scoring protein druggability via automated analyses of their molecular surface cavities in crystal structures (An et al. 2005; Volkamer et al. 2012). A limitation in the subsequent validation of pre-selected candidate targets using in vitro and in vivo disease models is that the current model systems for PD only reflect small subsets of the pathological features of PD as opposed to more established models for other complex disorders like Alzheimer’s disease (Beal 2001; Antony et al. 2011). Strategies involving the combined use of multiple complementary disease models, as well as ongoing projects on developing models with more robust pathological changes (e.g., using double knockouts of PD-mutated genes), will help to address these shortcomings. In this context, omics profiling and computational systems biology approaches will help to compare different disease models in terms of pathological and protective pathway activity changes and to assess their similarity to corresponding alterations in biospecimens from PD patients.

Finally, apart from their possible roles in the identification and preclinical validation of a drug target, systems biology approaches can also support the discovery of relevant drug-like small molecule binders. For example, a variety of systems-level approaches for drug repositioning have been developed (Dudley et al. 2011; Li and Lu 2013; Napolitano et al. 2013; Wu et al. 2013; Wang et al. 2014; Li et al. 2016a; Xu and Wang 2016), which can be complemented by virtual screening methods to identify new small molecule ligands for pre-selected targets (Stahura and Bajorath 2004; McInnes 2007). Further compound filtering is required due to the specific challenge for brain disorders that candidate drug-like molecules need to pass the blood–brain barrier (BBB). However, for compounds with unknown BBB permeability, dedicated *in silico* methods to predict this property are available as a prior filter for subsequent experimental testing (Kortagere et al. 2008; Muehlbacher et al. 2011; Carpenter et al. 2014). A more problematic common bottleneck is that extensive preclinical validation experiments for drug targets and their small-molecule binders are often not feasible in an academic setting in terms of the associated cost and resources, preventing promising target and compound discoveries from moving forward towards clinical development and testing. Projects that incentivize an earlier and more intensive collaboration between industry and academia on experimental target validation and preclinical drug development, e.g., the European Lead Factory (Mullard 2013), as well as the establishment of shared hardware and software infrastructures for systems biology (Athey et al. 2013; Auffray et al. 2016), will therefore be key facilitators for bridging the gap between new biomedical discoveries and their clinical translation.

In summary, computational systems biology approaches support experimental biomedical investigations by helping to prioritize candidate biomarkers, drug targets and binding compounds for subsequent validation, and providing insights into the mechanisms of molecular network and pathway dysregulations. For PD research specifically, integrative and comparative omics analyses that exploit prior biological knowledge can help to address current limitations in the available disease models and omics sample sizes, and to find surrogate markers for molecular changes in the brain. These computational systems-level analyses do not represent an alternative to targeted experimental studies of individual genes and proteins, but rather both targeted and systems-level approaches provide complementary information that will, collectively, help to pave the way towards improved biomarker signatures and new viable drug targets.

Acknowledgements Acknowledgement is made for support by the Fonds Nationale de la Recherche (FNR) Luxembourg, through the National Centre of Excellence in Research (NCER) on Parkinson’s disease, I1R-BIC-PFN-15NCER), and as part of the project MitoPD, under the auspices of the bilateral e:Med program by the German Federal Ministry of Education and Research and the FNR (INTER/BMBF/13/04).

Compliance with ethical standards

Conflict of interest The author declares that he has no conflict of interest.

References

Abdi F, Quinn JF, Jankovic J, McIntosh M, Leverenz JB, Peskind E, Nixon R, Nutt J, Chung K, Zabetian C, Samii A, Lin M, Hattan S, Pan C, Wang Y, Jin J, Zhu D, Li GJ, Liu Y, Waichunas D, Montine TJ, Zhang J (2006) Detection of biomarkers with a multiplex quantitative proteomic platform in cerebrospinal fluid of patients with neurodegenerative disorders. J Alzheimers Dis 9:293–348

Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent L-C, De Moor B, Marynen P, Hassan B, Carmeliet P, Moreau Y (2006) Gene prioritization through genomic data fusion. Nat Biotechnol 24:537–544. https://doi.org/10.1038/nbt1203

Amgalan B, Lee H (2014) WMAXC: a weighted maximum clique method for identifying condition-specific sub-network. PLoS ONE 9: e104993. https://doi.org/10.1371/journal.pone.0104993

An J, Totrov M, Abagyan R (2005) Pocketome via comprehensive identification and classification of ligand binding envelopes. Mol Cell Proteomics 4:752–761. https://doi.org/10.1074/mcp.M400159-MCP200

Andreopoulos B, An A, Wang X, Schroeder M (2009) A roadmap of clustering algorithms: finding a match for a biomedical application. Brief Bioinform 10:297–314. https://doi.org/10.1093/bib/bbn058

Antony PMA, Diederich NJ, Balling R (2011) Parkinson's disease mouse models in translational research. Mamm Genome 22:401–419. https://doi.org/10.1007/s00335-011-9330-x

Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM, Perona I (2013) An extensive comparative study of cluster validity indices. Pattern Recogn 46:243–256. https://doi.org/10.1016/j.patcog.2012.07.021

Athey BD, Braxenthaler M, Haas M, Guo Y (2013) tranSMART: an open source and community-driven informatics and data sharing platform for clinical and translational research. Proc AMIA Jt Summits Transl Sci 2013:6–8

Auffray C, Balling R, Barroso I, Bencze L, Benson M, Bergeron J, Bernal-Delgado E, Blomberg N, Bock C, Conesa A, Del Signore S, Delogne C, Devilee P, Di Meglio A, Eijkemans M, Flicek P, Graf N, Grimm V, Guchelaar H-J, Guo Y-K, Gut IG, Hanbury A, Hanif S, Hilgers R-D, Honrado Á, Hose DR, Houwing-Duistermaat J, Hubbard T, Janacek SH, Karanikas H, Kievits T, Kohler M, Kremer A, Lanfear J, Lengauer T, Maes E, Meert T, Müller W, Nickel D, Oledzki P, Pedersen B, Petkovic M, Pliakos K, Rattray M, Màs JR I, Schneider R, Sengstag T, Serra-Picamal X, Spek W, LAI V, van Batenburg O, Vandelaer M, Varnai P, Villoslada P, Vizcaíno JA, JPM W, Zanetti G (2016) Making sense of big data in health research: towards an EU action plan. Genome Med 8:71. https://doi.org/10.1186/s13073-016-0323-y

Bauer S, Grossmann S, Vingron M, Robinson PN (2008) Ontologizer 2.0–a multifunctional tool for GO term enrichment analysis and data exploration. Bioinformatics 24:1650–1651. https://doi.org/10.1093/bioinformatics/btn250

Beal MF (2001) Experimental models of Parkinson's disease. Nat Rev Neurosci 2:325–334. https://doi.org/10.1038/35072550

Beißbarth T, Speed TP (2004) GOstat: find statistically overrepresented gene Ontologies within a group of genes. Bioinformatics 20:1464–1465. https://doi.org/10.1093/bioinformatics/bth088

Beisser D, Klau GW, Dandekar T, Müller T, Dittrich MT (2010) BioNet: an R-package for the functional analysis of biological networks. Bioinformatics 26:1129–1130. https://doi.org/10.1093/bioinformatics/btq089

Belkin M, Niyogi P (2003) Laplacian Eigenmaps for dimensionality reduction and data representation. Neural Comput 15:1373–1396. https://doi.org/10.1162/089976603321780317

Bellman RE (1961) Adaptive control processes: a guided tour. Princetown University Press, Princetown

Bellou V, Belbasis L, Tzoulaki I, Evangelou E, Ioannidis JPA (2016) Environmental risk factors and Parkinson's disease: an umbrella review of meta-analyses. Parkinsonism Relat Disord 23:1–9. https://doi.org/10.1016/j.parkreldis.2015.12.008

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57:289–300. https://doi.org/10.2307/2346101

Berg D, Lang AE, Postuma RB, Maetzler W, Deuschl G, Gasser T, Siderowf A, Schapira AH, Oertel W, Obeso JA, Olanow CW, Poewe W, Stern M (2013) Changing the research criteria for the diagnosis of Parkinson's disease: obstacles and opportunities. Lancet Neurol 12:514–524

Bogdanov M, Matson WR, Wang L, Matson T, Saunders-Pullman R, Bressman SS, Beal MF (2008) Metabolomic profiling to develop blood biomarkers for Parkinson's disease. Brain 131:389–396. https://doi.org/10.1093/brain/awm304

Bradley G, Barrett SJ (2017) CausalR-extracting mechanistic sense from genome scale data. Bioinformatics. https://doi.org/10.1093/bioinformatics/btx425

Braga-Neto UM, Dougherty ER (2004) Is cross-validation valid for small-sample microarray classification? Bioinformatics 20:374–380. https://doi.org/10.1093/bioinformatics/btg419

Browne M (2000) Cross-validation methods. J Math Psychol 44:108–132. https://doi.org/10.1006/jmps.1999.1279

Calderone A, Formenti M, Aprea F, Papa M, Alberghina L, Colangelo AM, Bertolazzi P (2016) Comparing Alzheimer's and Parkinson's diseases networks using graph communities structure. BMC Syst Biol 10:25. https://doi.org/10.1186/s12918-016-0270-7

Carpenter TS, Kirshner DA, Lau EY, Wong SE, Nilmeier JP, Lightstone FC (2014) A method to predict blood-brain barrier permeability of drug-like compounds using molecular dynamics simulations. Biophys J 107:630–641. https://doi.org/10.1016/j.bpj.2014.06.024

Catlett NL, Bargnesi AJ, Ungerer S, Seagaran T, Ladd W, Elliston KO, Pratt D (2013) Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data. BMC Bioinformatics 14:340. https://doi.org/10.1186/1471-2105-14-340

Chandrasekaran S, Bonchev D (2013) A network view on Parkinson's disease. Comput Struct Biotechnol J 7:e201304004. https://doi.org/10.5936/csbj.201304004

Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, Breitkreutz A, Kolas N, O'Donnell L, Reguly T, Nixon J, Ramage L, Winter A, Sellam A, Chang C, Hirschman J, Theesfeld C, Rust J, Livstone MS, Dolinski K, Tyers M (2015) The BioGRID interaction database: 2015 update. Nucleic Acids Res 43: D470–D478. https://doi.org/10.1093/nar/gku1204

Chen J, Bardes EE, Aronow BJ, Jegga AG (2009) ToppGene suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. https://doi.org/10.1093/nar/gkp427

Chen L, Xuan J, Riggins RB, Wang Y, Clarke R (2013) Identifying protein interaction subnetworks by a bagging Markov random field-based method. Nucleic Acids Res. https://doi.org/10.1093/nar/gks951

Chikina MD, Gerald CP, Li X, Ge Y, Pincas H, Nair VD, Wong AK, Krishnan A, Troyanskaya OG, Raymond D, Saunders-Pullman R, Bressman SB, Yue Z, Sealfon SC (2015) Low-variance RNAs identify Parkinson's disease molecular signature in blood. Mov Disord 30:813–821. https://doi.org/10.1002/mds.26205

Chindelevitch L, Ziemek D, Enayetallah A, Randhawa R, Sidders B, Brockel C, Huang ES (2012) Causal reasoning on biological networks: interpreting transcriptional changes. Bioinformatics 28: 1114–1121. https://doi.org/10.1093/bioinformatics/bts090

Chipman H, Tibshirani R (2006) Hybrid hierarchical clustering with applications to microarray data. Biostatistics 7:286–301. https://doi.org/10.1093/biostatistics/kxj007

Choi JY, Bae S-H, Qiu X, Fox G (2010) High performance dimension reduction and visualization for large high-dimensional data analysis. 2010 10th IEEE/ACM Int Conf Clust cloud grid Comput 331–340. doi: https://doi.org/10.1109/CCGRID.2010.104

Chuang H-Y, Lee E, Liu Y-T, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. Mol Syst Biol 3:140

Cornish AJ, Markowetz F (2014) SANTA: quantifying the functional content of molecular networks. PLoS Comput Biol. https://doi.org/10.1371/journal.pcbi.1003808

Cowley MJ, Pinese M, Kassahn KS, Waddell N, Pearson J V., Grimmond SM, Biankin A V., Hautaniemi S, Wu J (2012) PINA v2.0: mining interactome modules. Nucleic Acids Res. doi: https://doi.org/10.1093/nar/gkr967

Cui X, Churchill GA (2003) Statistical tests for differential expression in cDNA microarray experiments. Genome Biol 4:210. https://doi.org/10.1186/gb-2003-4-4-210

D'Antonio M, D'Onorio De Meo P, Pallocca M, Picardi E, D'Erchia AM, Calogero RA, Castrignanò T, Pesole G (2015) RAP: RNA-Seq analysis pipeline, a new cloud-based NGS web application. BMC Genomics 16:S3. https://doi.org/10.1186/1471-2164-16-S6-S3

Dardiotis E, Xiromerisiou G, Hadjichristodoulou C, Tsatsakis AM, Wilks MF, Hadjigeorgiou GM (2013) The interplay between environmental and genetic factors in Parkinson's disease susceptibility: the evidence for pesticides. Toxicology 307:17–23. https://doi.org/10.1016/j.tox.2012.12.016

del Sol A, Balling R, Hood L, Galas D (2010) Diseases as network perturbations. Curr Opin Biotechnol 21:566–571

Demissie M, Mascialino B, Calza S, Pawitan Y (2008) Unequal group variances in microarray data analyses. Bioinformatics 24:1168–1174. https://doi.org/10.1093/bioinformatics/btn100

Demšar J, Curk T, Erjavec A, Hočevar T, Milutinovič M, Možina M, Polajnar M, Toplak M, Starič A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B (2013) Orange: data mining toolbox in python. J Mach Learn Res 14:23492353

Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane CH, Lempicki RA, Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA (2003) DAVID: database for annotation, visualization, and integrated discovery. Genome Biol 4:R60. https://doi.org/10.1186/gb-2003-4-9-r60

Di Lena P, Martelli PL, Fariselli P, Casadio R (2015) NET-GE: a novel NETwork-based gene enrichment for detecting biological processes associated to Mendelian diseases. BMC Genomics 16:S6. https://doi.org/10.1186/1471-2164-16-S8-S6

Dietterich TG (2000) Ensemble methods in machine learning. Mult Classif Syst 1857:1–15. https://doi.org/10.1007/3-540-45014-9

Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Müller T (2008) Identifying functional modules in protein-protein interaction networks: an integrated exact approach. Bioinformatics. https://doi.org/10.1093/bioinformatics/btn161

Do CB, Tung JY, Dorfman E, Kiefer AK, Drabant EM, Francke U, Mountain JL, Goldman SM, Tanner CM, Langston JW, Wojcicki A, Eriksson N (2011) Web-based genome-wide association study identifies two novel loci and a substantial genetic component for parkinson's disease. PLoS Genet. https://doi.org/10.1371/journal.pgen.1002141

Draghici S, Khatri P, Bhavsar P, Shah A, Krawetz SA, Tainsky MA (2003) Onto-tools, the toolkit of the modern biologist: onto-express, onto-compare, onto-design and onto-translate. Nucleic Acids Res 31:3775–3781. https://doi.org/10.1093/nar/gkg624

Dudley JT, Deshpande T, Butte AJ (2011) Exploiting drug-disease relationships for computational drug repositioning. Brief Bioinform 12:303–311. https://doi.org/10.1093/bib/bbr013

Dujardin K, Defebvre L, Duhamel A, Lecouffe P, Rogelet P, Steinling M, Destée A (2004) Cognitive and SPECT characteristics predict progression of Parkinson's disease in newly diagnosed patients. J Neurol 251:1383–1392. https://doi.org/10.1007/s00415-004-0549-2

Dutta B, Wallqvist A, Reifman J (2012) PathNet: a tool for pathway analysis using topological information. Source Code Biol Med 7:10. https://doi.org/10.1186/1751-0473-7-10

Edwards YJK, Beecham GW, Scott WK, Khuri S, Bademci G, Tekin D, Martin ER, Jiang Z, Mash DC, Ffrench-Mullen J, Pericak-Vance MA, Tsinoremas N, Vance JM (2011) Identifying consensus disease pathways in Parkinson's disease using an integrative systems biology approach. PLoS ONE. https://doi.org/10.1371/journal.pone.0016917

Efron B, Tibshirani R (2007) On testing the significance of sets of genes. Ann Appl Stat 1:107–129. https://doi.org/10.1214/07-AOAS101

Ester M, Kriegel HP, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. Proc 2nd Int Conf Knowl Discov Data Min 226–231. Doi: 10.1.1.71.1980

Fakhry CT, Choudhary P, Gutteridge A, Sidders B, Chen P, Ziemek D, Zarringhalam K (2016) Interpreting transcriptional changes using causal graphs: new methods and their practical utility on public networks. BMC Bioinformatics 17:318. https://doi.org/10.1186/s12859-016-1181-8

Fang Z, Yang J, Li Y, Luo Q, Liu L (2006) Knowledge guided analysis of microarray data. J Biomed Inform 39:401–411. https://doi.org/10.1016/j.jbi.2005.08.004

Fu LM, Fu KA (2015) Analysis of Parkinson's disease pathophysiology using an integrated genomics-bioinformatics approach. Pathophysiology 22:15–29. https://doi.org/10.1016/j.pathophys.2014.10.002

Fujita KA, Ostaszewski M, Matsuoka Y, Ghosh S, Glaab E, Trefois C, Crespo I, Perumal TM, Jurkowski W, Antony PMA, Diederich N, Buttini M, Kodama A, Satagopam VP, Eifes S, Sol A, Schneider R, Kitano H, Balling R (2014) Integrating pathways of Parkinson's disease in a molecular interaction map. Mol Neurobiol 49:88–102

Gasparoli E, Delibori D, Polesello G, Santelli L, Ermani M, Battistin L, Bracco F (2002) Clinical predictors in Parkinson's disease. Neurol Sci 23(Suppl 2):S77–S78. https://doi.org/10.1007/s100720200078

GeneOntologyConsortium (2004) The gene ontology (GO) database and informatics resource. Nucleic Acids Res 32:258D–2261. https://doi.org/10.1093/nar/gkh036

Glaab E, Baudot A, Krasnogor N, Schneider R, Valencia A (2012) EnrichNet: network-based gene set enrichment analysis. Bioinformatics 28:i451–i457. https://doi.org/10.1093/bioinformatics/BTS389

Glaab E, Garibaldi JM, Krasnogor N (2009) ArrayMining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization. BMC Bioinformatics 10:358

Glaab E, Schneider R (2015) Comparative pathway and network analysis of brain transcriptome changes during adult aging and in Parkinson's disease. Neurobiol Dis 74:1–13. https://doi.org/10.1016/j.nbd.2014.11.002

Goeman JJ, Van de Geer S, De Kort F, van Houwellingen HC (2004) A global test for groups fo genes: testing association with a clinical outcome. Bioinformatics 20:93–99. https://doi.org/10.1093/bioinformatics/btg382

Goetz CC (2003) The unified Parkinson's disease rating scale (UPDRS): status and recommendations. Mov Disord 18:738–750

Gorell JM, Peterson EL, Rybicki BA, Johnson CC (2004) Multiple risk factors for Parkinson's disease. J Neurol Sci 217:169–174. https://doi.org/10.1016/j.jns.2003.09.014

Graham JM, Sagar HJ (1999) A data-driven approach to the study of heterogeneity in idiopathic Parkinson's disease: identification of three distinct subtypes. Mov Disord 14:10–20. https://doi.org/10.1002/1531-8257(199901)14:1<10::AID-MDS1005>3.0.CO;2-4

Gu J, Chen Y, Li S, Li Y (2010) Identification of responsive gene modules by network-based gene clustering and extending: application to inflammation and angiogenesis. BMC Syst Biol 4:47. https://doi.org/10.1186/1752-0509-4-47

Hall MA, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. SIGKDD Explor 11:10–18. https://doi.org/10.1145/1656274.1656278

Hartigan JA, Wong MA (1979) A K-means clustering algorithm. Appl Stat 28:100–108. https://doi.org/10.2307/2346830

Heller K A, Ghahramani Z (2005) Bayesian hierarchical clustering. Proc 22nd Int Conf Mach Learn 297–304. https://doi.org/10.1145/1102351.1102389

Hinneburg A, Keim D (1998) DENCLUE: An efficient approach to clustering in large multimedia databases with noise. Proceedings of the

4th International Conference on Knowledge Discovery and Datamining, New York, p 58–65.

Hokamp K, Roche FM, Acab M, Rousseau ME, Kuo B, Goode D, Aeschliman D, Bryan J, Babiuk LA, Hancock REW, Brinkman FSL (2004) ArrayPipe: a flexible processing pipeline for microarray data. Nucleic Acids Res. https://doi.org/10.1093/nar/gkh446

Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J (2006) RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. Bioinformatics 22:2825–2827. https://doi.org/10.1093/bioinformatics/btl476

Hu Y, Pan Z, Hu Y, Zhang L, Wang J (2017) Network and pathway-based analyses of genes associated with Parkinson's disease. Mol Neurobiol 54:4452–4465. https://doi.org/10.1007/s12035-016-9998-8

Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37:1–13. https://doi.org/10.1093/nar/gkn923

Hung J-H, Whitfield TW, Yang T-H, Hu Z, Weng Z, DeLisi C (2010) Identification of functional modules that correlate with phenotypic difference: the influence of network topology. Genome Biol 11:R23. https://doi.org/10.1186/gb-2010-11-2-r23

Ideker T, Ozier O, Schwikowski B, Siegel AF (2002) Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics 18(Suppl 1):S233–S240. https://doi.org/10.1093/bioinformatics/18.suppl_1.S233

Ideker T, Sharan R (2008) Protein networks in disease. Genome Res 18: 644–652

Ihnatova I, Budinska E (2015) ToPASeq: an R package for topology-based pathway analysis of microarray and RNA-Seq data. BMC Bioinformatics 16:350. https://doi.org/10.1186/s12859-015-0763-1

Ishigami N, Tokuda T, Ikegawa M, Komori M, Kasai T, Kondo T, Matsuyama Y, Nirasawa T, Thiele H, Tashiro K, Nakagawa M (2012) Cerebrospinal fluid proteomic patterns discriminate Parkinson's disease and multiple system atrophy. Mov Disord 27: 851–857. https://doi.org/10.1002/mds.24994

Isik Z, Baldow C, Cannistraci CV, Schroeder M (2015) Drug target prioritization by perturbed gene expression and network information. Sci Rep 5:17417. https://doi.org/10.1038/srep17417

Jacob L, Neuvial P, Dudoit S (2012) More power via graph-structured tests for differential expression of gene networks. Ann Appl Stat 6: 561–600. https://doi.org/10.1214/11-AOAS528

Jaeger S, Min J, Nigsch F, Camargo M, Hutz J, Cornett A, Cleaver S, Buckler A, Jenkins JL (2014) Causal network models for predicting compound targets and driving pathways in cancer. J Biomol Screen 19:791–802. https://doi.org/10.1177/1087057114522690

Jankovic J (2008) Parkinson's disease: clinical features and diagnosis. J Neurol Neurosurg Psychiatry 79:368–376. https://doi.org/10.1136/jnnp.2007.131045

Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L (2005) Nucleic Acids Res. https://doi.org/10.1093/nar/gki072

Kalia LV, Lang AE (2015) Parkinson's disease. Lancet 386:896–912. https://doi.org/10.1016/S0140-6736(14)61393-3

Kalinderi K, Bostantjopoulou S, Fidani L (2016) The genetic background of Parkinson's disease: current progress and future prospects. Acta Neurol Scand 134:314–326. https://doi.org/10.1111/ane.12563

Karypis G, Han E-H, Kumar V (1999) Chameleon: hierarchical clustering using dynamic modeling. Computer (Long Beach Calif) 32:68–75. https://doi.org/10.1109/2.781637

Kaufman L, Rousseeuw PJ (1987) Clustering by means of medoids. Stat. Data Anal. Based L 1-Norm Relat. Methods. First Int Conf. 405–416416

Kaufman L, Rousseeuw PJ (1990) Partitioning around Medoids (program PAM). Find Groups Data An Introd to Clust Anal:68–125. https://doi.org/10.1002/9780470316801.ch2

Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U, Jandrasits C, Jimenez RC, Khadake J, Mahadevan U, Masson P, Pedruzzi I, Pfeiffenberger E, Porras P, Raghunath A, Roechert B, Orchard S, Hermjakob H (2012) The IntAct molecular interaction database in 2012. Nucleic Acids Res. https://doi.org/10.1093/nar/gkr1088

Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A (2009) Human protein reference database–2009 update. Nucleic Acids Res 37:D767–D772. https://doi.org/10.1093/nar/gkn892

Kieburtz K, Wunderle KB (2013) Parkinson's disease: evidence for environmental risk factors. Mov Disord 28:8–13

Kim S-Y, Volsky DJ (2005) PAGE: parametric analysis of gene set enrichment. BMC Bioinformatics 6:144

Kittas A, Delobelle A, Schmitt S, Breuhahn K, Guziolowski C, Grabe N (2016) Directed random walks and constraint programming reveal active pathways in hepatocyte growth factor signaling. FEBS J 283: 350–360. https://doi.org/10.1111/febs.13580

Knudson AG (1971) Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci U S A 68:820–823. https://doi.org/10.1073/pnas.68.4.820

Köppen M (2000) The curse of dimensionality. 5th online world conf soft Comput Ind Appl 1:4–8. doi: https://doi.org/10.1200/JCO.2010.30.1986

Kortagere S, Chekmarev D, Welsh WJ, Ekins S (2008) New predictive models for blood-brain barrier permeability of drug-like molecules. Pharm Res 25:1836–1845. https://doi.org/10.1007/s11095-008-9584-5

Kovács F, Legány C, Babos A (2005) Cluster validity measurement techniques. Proc 6th Int Symp Hungarian res. Comput Intell 2006:1–11. https://doi.org/10.7547/87507315-91-9-465

Labadorf A, Choi SH, Myers RH (2017) Comparative Huntington and Parkinson disease mRNA analysis reveals common inflammatory processes. bioRxiv. doi: https://doi.org/10.1101/139451

Lawton M, Baig F, Rolinski M, Ruffman C, Nithi K, May MT, Ben-Shlomo Y, Hu MTM (2015) Parkinson's disease subtypes in the Oxford Parkinson disease centre (OPDC) discovery cohort. J Parkinsons Dis 5:269–279. https://doi.org/10.3233/JPD-140523

Lee J, Jo K, Lee S, Kang J, Kim S (2016) Prioritizing biological pathways by recognizing context in time-series gene expression data. BMC Bioinformatics 17:477. https://doi.org/10.1186/s12859-016-1335-8

Lewis SJG, Foltynie T, Blackwell AD, Robbins TW, Owen AM, Barker RA (2005) Heterogeneity of Parkinson's disease in the early clinical stages using a data driven approach. J Neurol Neurosurg Psychiatry 76:343–348. https://doi.org/10.1136/jnnp.2003.033530

Li C, Li H (2008) Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics 24:1175–1182. https://doi.org/10.1093/bioinformatics/btn081

Li J, Lu Z (2013) Pathway-based drug repositioning using causal inference. BMC Bioinformatics 14:S3. https://doi.org/10.1186/1471-2105-14-S16-S3

Li J, Zheng S, Chen B, Butte AJ, Swamidass SJ, Lu Z (2016a) A survey of current trends in computational drug repositioning. Brief Bioinform 17:2–12. https://doi.org/10.1093/bib/bbv020

Li Y, Jourdain AA, Calvo SE, Liu JS, Mootha VK (2017) CLIC, a tool for expanding biological pathways based on co-expression across thousands of datasets. PLoS Comput Biol 13:1–29. https://doi.org/10.1371/journal.pcbi.1005653

Li Y, Wu F-X, Ngom A (2016b) A review on machine learning principles for multi-view biological data integration. Brief bioinform bbw113. doi: https://doi.org/10.1093/bib/bbw113

Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardozza AP, Santonico E, Castagnoli L, Cesareni G (2012) MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. https://doi.org/10.1093/nar/gkr930

Liu P, Feng T, Wang Y-J, Zhang X, Chen B (2011) Clinical heterogeneity in patients with early-stage Parkinson's disease: a cluster analysis. J Zhejiang Univ Sci B 12:694–703. https://doi.org/10.1631/jzus.B1100069

Lottaz C, Toedling J, Spang R (2007) Annotation-based distance measures for patient subgroup discovery in clinical microarray studies. Bioinformatics 23:2256–2264. https://doi.org/10.1093/bioinformatics/btm322

Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ (2009) GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinformatics 10:161. https://doi.org/10.1186/1471-2105-10-161

Ma H, Schadt EE, Kaplan LM, Zhao H (2011) COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method. Bioinformatics 27:1290–1298. https://doi.org/10.1093/bioinformatics/btr136

Van Der Maaten L, Hinton G (2008) Visualizing Data using t-SNE. J Mach Learn Res 620:267–284. https://doi.org/10.1007/s10479-011-0841-3

Marras C, Lang A (2013) Parkinson's disease subtypes: lost in translation? J Neurol Neurosurg Psychiatry 84:409–415. https://doi.org/10.1136/jnnp-2012-303455

Martin D, Martin D, Brun C, Brun C, Remy E, Remy E, Mouren P, Mouren P, Thieffry D, Thieffry D, Jacq B, Jacq B (2004) GOToolBox: functional analysis of gene datasets based on gene ontology. Genome Biol 5:R101. https://doi.org/10.1186/gb-2004-5-12-r101

McInnes C (2007) Virtual screening strategies in drug discovery. Curr Opin Chem Biol 11:494–502

Mizuno S, Iijima R, Ogishima S, Kikuchi M, Matsuoka Y, Ghosh S, Miyamoto T, Miyashita A, Kuwano R, Tanaka H (2012) AlzPathway: a comprehensive map of signaling pathways of Alzheimer's disease. BMC Syst Biol 6:1–10. https://doi.org/10.1186/1752-0509-6-52

Molochnikov L, Rabey JM, Dobronevsky E, Bonucelli U, Ceravolo R, Frosini D, Grünblatt E, Riederer P, Jacob C, Aharon-Peretz J, Bashenko Y, Youdim MBH, Mandel SA (2012) A molecular signature in blood identifies early Parkinson's disease. Mol Neurodegener 7:26. https://doi.org/10.1186/1750-1326-7-26

Mosca E, Milanesi L (2013) Network-based analysis of omics with multi-objective optimization. Mol BioSyst 9:2971. https://doi.org/10.1039/c3mb70327d

Muehlbacher M, Spitzer GM, Liedl KR, Kornhuber J (2011) Qualitative prediction of blood-brain barrier permeability on a large and refined dataset. J Comput Aided Mol Des 25:1095–1106. https://doi.org/10.1007/s10822-011-9478-1

Mullard A (2013) European lead factory opens for business. Nat Rev Drug Discov 12:173–175. https://doi.org/10.1038/nrd3956

Müller B, Assmus J, Herlofson K, Larsen JP, Tysnes OB (2013) Importance of motor vs. non-motor symptoms for health-related quality of life in early Parkinson's disease. Park Relat Disord 19:1027–1032. https://doi.org/10.1016/j.parkreldis.2013.07.010

Napolitano F, Zhao Y, Moreira VM, Tagliaferri R, Kere J, D'Amato M, Greco D (2013) Drug repositioning: a machine-learning approach through data integration. J Cheminform. https://doi.org/10.1186/1758-2946-5-30

Natale M, Benso A, Di Carlo S, Ficarra E (2014) FunMod: a Cytoscape Plugin for identifying functional modules in undirected protein-protein networks. Genomics Proteomics Bioinforma 12:178–186. https://doi.org/10.1016/j.gpb.2014.05.002

Nishimura D (2001) BioCarta. Biotech Softw Internet Rep 2:117–120. https://doi.org/10.1089/152791601750294344

Oerton E, Bender A (2017) Concordance analysis of microarray studies identifies representative gene expression changes in Parkinson's disease: a comparison of 33 human and animal studies. BMC Neurol 17:58. https://doi.org/10.1186/s12883-017-0838-x

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M (1999) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 27:29–34

Park H, Niida A, Miyano S, Imoto S (2015) Sparse overlapping group lasso for integrative multi-Omics analysis. J Comput Biol 22:73–84. https://doi.org/10.1089/cmb.2014.0197

Parkinson J (1817) An essay on the shaking palsy. J Neuropsychiatry Clin Neurosci 14:223–236

Pico AR, Kelder T, Van Iersel MP, Hanspers K, Conklin BR, Evelo C (2008) WikiPathways: pathway editing for the people. PLoS Biol 6:1403–1407

Post B, Speelman JD, de Haan RJ (2008) Clinical heterogeneity in newly diagnosed Parkinson's disease. J Neurol 255:716–722. https://doi.org/10.1007/s00415-008-0782-1

Potashkin JA, Santiago JA, Ravina BM, Watts A, Leontovich AA (2012) Biosignatures for Parkinson's disease and atypical parkinsonian disorders patients. PLoS ONE. https://doi.org/10.1371/journal.pone.0043595

Rahmatallah Y, Emmert-Streib F, Glazko G (2015) Gene set analysis approaches for RNA-seq data: performance evaluation and application guideline. Brief Bioinform 17:1–15. https://doi.org/10.1093/bib/bbv069

Rainer J, Sanchez-Cabo F, Stocker G, Sturn A, Trajanoski Z (2006) CARMAweb: comprehensive R- and bioconductor-based web service for microarray data analysis. Nucleic Acids Res. https://doi.org/10.1093/nar/gkl038

Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, Betel D (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol 14:R95. https://doi.org/10.1186/gb-2013-14-9-r95

Reijnders JSAM, Ehrt U, Lousberg R, Aarsland D, Leentjens AFG (2009) The association between motor subtypes and psychopathology in Parkinson's disease. Park Relat Disord 15:379–382. https://doi.org/10.1016/j.parkreldis.2008.09.003

Rendón E, Abundez I, Arizmendi A, Quiroz EM (2011) Internal versus external cluster validation indexes. Int J Comput Commun 5:27–34

Ritter H, Kohonen T (1989) Self-organizing semantic maps. Biol Cybern 61:241–254. https://doi.org/10.1007/BF00203171

Rohart F, Gautier B, Singh A, Le Cao K-A (2017) MixOmics: an R package for 'omics feature selection and multiple data integration. bioRxiv 108597. doi: https://doi.org/10.1101/108597

Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290(80):2323–2326. https://doi.org/10.1126/science.290.5500.2323

Samii A, Nutt JG, Ransom BR (2004) Parkinson's disease. Lancet 363:1783–1793

Santiago JA, Scherzer CR, Potashkin JA (2013) Specific splice variants are associated with Parkinson's disease. Mov Disord 28:1724–1727. https://doi.org/10.1002/mds.25635

Santiago JA, Potashkin JA (2013) Integrative network analysis unveils convergent molecular pathways in Parkinson's disease and diabetes. PLoS ONE. https://doi.org/10.1371/journal.pone.0083940

Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH (2009) PID: the pathway interaction database. Nucleic Acids Res. https://doi.org/10.1093/nar/gkn653

Schaefer MH, Fontaine JF, Vinayagam A, Porras P, Wanker EE, Andrade-Navarro MA (2012) Hippie: integrating protein interaction networks with experiment based quality scores. PLoS ONE. https://doi.org/10.1371/journal.pone.0031826

Scherzer CR, Eklund AC, Morse LJ, Liao Z, Locascio JJ, Fefer D, Schwarzschild MA, Schlossmacher MG, Hauser MA, Vance JM, Sudarsky LR, Standaert DG, Growdon JH, Jensen RV, Gullans SR

(2007) Molecular markers of early Parkinson's disease based on gene expression in blood. Proc Natl Acad Sci U S A 104:955–960

Schrag A, Quinn NP, Ben-Shlomo Y (2006) Heterogeneity of Parkinson's disease. J Neurol Neurosurg Psychiatry 77:275–276

Secrier M, Pavlopoulos GA, Aerts J, Schneider R (2012) Arena3D: visualizing time-driven phenotypic differences in biological systems. BMC Bioinformatics 13:45. https://doi.org/10.1186/1471-2105-13-45

Shamir R, Klein C, Amar D, Vollstedt E-J, Bonin M, Usenovic M, Wong YC, Maver A, Poths S, Safer H, Corvol J-C, Lesage S, Lavi O, Deuschl G, Kuhlenbaeumer G, Pawlack H, Ulitsky I, Kasten M, Riess O, Brice A, Peterlin B, Krainc D (2017) Analysis of blood-based gene expression in idiopathic Parkinson disease. Neurology. https://doi.org/10.1212/WNL.0000000000004516

Slawski M, Daumer M, Boulesteix A-L (2008) CMA: a comprehensive bioconductor package for supervised classification with high dimensional data. BMC Bioinformatics 9:439. https://doi.org/10.1186/1471-2105-9-439

Smyth G (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3:article 3. doi: https://doi.org/10.2202/1544-6115.1027

Smyth G (2005) Limma: linear models for microarray data. In: Bioinformatics and computational biology solutions using R and bioconductor. Springer, New York, pp 397–420

Solla P, Cannas A, Ibba FC, Loi F, Corona M, Orofino G, Marrosu MG, Marrosu F (2012) Gender differences in motor and non-motor symptoms among Sardinian patients with Parkinson's disease. J Neurol Sci 323:33–39. https://doi.org/10.1016/j.jns.2012.07.026

Stahura FL, Bajorath J (2004) Virtual screening methods that complement HTS. Comb Chem High Throughput Screen 7:259–269. https://doi.org/10.2174/1386207043328706

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102:15545–15550

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, Von Mering C (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res 43:D447–D452. https://doi.org/10.1093/nar/gku1003

Tarca A, Draghici S, Bhatti G, Romero R (2012) Down-weighting overlapping genes improves gene set analysis. BMC Bioinformatics 13:136. https://doi.org/10.1186/1471-2105-13-136

Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS, Kim CJ, Kusanovic JP, Romero R (2009) A novel signaling pathway impact analysis. Bioinformatics 25:75–82. https://doi.org/10.1093/bioinformatics/btn577

Temple Lang D, Swayne DF (2001) GGobi meets R: an extensible environment for interactive dynamic data visualization. Proc 2nd Int Work Distrib Stat Comput 11

Torgerson WS (1952) Multidimensional scaling: I. Theory Method Psychometrika 17:401–419. https://doi.org/10.1007/BF02288916

Trezzi JP, Galozzi S, Jaeger C, Barkovits K, Brockmann K, Maetzler W, Berg D, Marcus K, Betsou F, Hiller K, Mollenhauer B (2017) Distinct metabolomic signature in cerebrospinal fluid in early parkinson's disease. Mov Disord (in press)

Ung MH, Liu C-C, Cheng C (2016) Integrative analysis of cancer genes in a functional interactome. Sci Rep 6:29228. https://doi.org/10.1038/srep29228

Van Rooden SM, Colas F, Martínez-Martín P, Visser M, Verbaan D, Marinus J, Chaudhuri RK, Kok JN, Van Hilten JJ (2011) Clinical subtypes of Parkinson's disease. Mov Disord 26:51–58. https://doi.org/10.1002/mds.23346

Vlassis N, Glaab E (2015) GenePEN: analysis of network activity alterations in complex diseases via the pairwise elastic net. Stat Appl Genet Mol Biol 14:221–224

Volkamer A, Kuhn D, Rippmann F, Rarey M (2012) Dogsitescorer: a web server for automatic binding site prediction, analysis and druggability assessment. Bioinformatics 28:2074–2075. https://doi.org/10.1093/bioinformatics/bts310

Wachter A, Beißbarth T (2014) PwOmics: an R package for pathway-based integration of time-series omics data using public database knowledge. Bioinformatics 31:3072–3074. https://doi.org/10.1093/bioinformatics/btv323

Waddell A, Oldford RW (2011) RnavGraph: a visualization tool for navigating through high-dimensional data. Proc 58th World Stat Congr 1852–1860

Wang L, Jia P, Wolfinger RD, Chen X, Zhao Z (2011) Gene set analysis of genome-wide association studies: methodological issues and perspectives. Genomics 98:1–8

Wang W, Yang S, Zhang X, Li J (2014) Drug repositioning by integrating target information through a heterogeneous network model. Bioinformatics 30:2923–2930. https://doi.org/10.1093/bioinformatics/btu403

Wood IA, Visscher PM, Mengersen KL (2007) Classification based upon gene expression data: bias and precision of error rates. Bioinformatics 23:1363–1370. https://doi.org/10.1093/bioinformatics/btm117

Wu C, Gudivada RC, Aronow BJ, Jegga AG (2013) Computational drug repositioning through heterogeneous network clustering. BMC Syst Biol 7:1–9. https://doi.org/10.1186/1752-0509-7-S5-S6

Wu G, Dawson E, Duong A, Haw R, Stein L (2014) ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis. F1000Research. Doi: 10.12688/f1000research.4431.2

Xu R, Wang Q (2016) A genomics-based systems approach towards drug repositioning for rheumatoid arthritis. BMC Genomics 17(Suppl 7):518. https://doi.org/10.1186/s12864-016-2910-0

Zaccaria A, Bouamrani A, Chabardès S, El Atifi M, Seigneuret E, Lobrinus JA, Dubois-Dauphin M, Berger F, Burkhard PR (2016) Deep brain stimulation-associated brain tissue imprints: a new in vivo approach to biological research in human Parkinson's disease. Mol Neurodegener 11:12. https://doi.org/10.1186/s13024-016-0077-4

Zararsiz G, Goksuluk D, Korkmaz S, Eldem V, Duru IP, Ozturk A, Unver T (2014) Classification of RNA-Seq data via bagging support vector machines. bioRxiv. doi: https://doi.org/10.1101/007526

Zarringhalam K, Enayetallah A, Gutteridge A, Sidders B, Ziemek D, Kelso J (2013) Molecular causes of transcriptional response: a Bayesian prior knowledge approach. Bioinformatics 29:3167–3173. https://doi.org/10.1093/bioinformatics/btt557

Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC, Weinstein JN (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol 4:R28. https://doi.org/10.1186/gb-2003-4-4-r28

Zheng B, Liao Z, Locascio JJ, Lesniak KA, Roderick SS, Watt ML, Eklund AC, Zhang-James Y, Kim PD, Hauser MA, Grünblatt E, Moran LB, Mandel SA, Riederer P, Miller RM, Federoff HJ, Wüllner U, Papapetropoulos S, Youdim MB, Cantuti-Castelvetri I, Young AB, Vance JM, Davis RL, Hedreen JC, Adler CH, Beach TG, Graeber MB, Middleton FA, Rochet J-C, Scherzer CR (2010) PGC-1α, a potential therapeutic target for early intervention in Parkinson's disease. Sci Transl Med 2:52ra73