Integration of ‘omics’ data in aging research: from biomarkers to systems biology

Jonas Zierer,1,2 Cristina Menni,1 Gabi Kastenmüller1,2 and Tim D. Spector1
1 Department of Twins Research and Genetic Epidemiology, King’s College London, London, London, United Kingdom
2 Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, Neuherberg, Germany

Summary
Age is the strongest risk factor for many diseases including neurodegenerative disorders, coronary heart disease, type 2 diabetes and cancer. Due to increasing life expectancy and low birth rates, the incidence of age-related diseases is increasing in industrialized countries. Therefore, understanding the relationship between diseases and aging and facilitating healthy aging are major goals in medical research. In the last decades, the dimension of biological data has drastically increased with high-throughput technologies now measuring thousands of (epi)genetic, expression and metabolic variables. The most common and so far successful approach to the analysis of these data is the so-called reductionist approach. It consists of separately testing each variable for association with the phenotype of interest such as age or age-related disease. However, a large portion of the observed phenotypic variance remains unexplained and a comprehensive understanding of most complex phenotypes is lacking. Systems biology aims to integrate data from different experiments to gain an understanding of the system as a whole rather than focusing on individual factors. It thus allows deeper insights into the mechanisms of complex traits, which are caused by the joint influence of several, interacting changes in the biological system. In this review, we look at the current progress of applying omics technologies to identify biomarkers of aging. We then survey existing systems biology approaches that allow for an integration of different types of data and highlight the need for further developments in this area to improve epidemiologic investigations.

Key words: data integration; graphical models; high-throughput data; omics; systems biology.

Introduction
Aging is often described as the progressive accumulation of changes with time leading to a loss of physiological aptitude and fertility, an increased susceptibility to disease and ultimately to death (Harman, 1988, 2001; Kirkwood & Austad, 2000; Vijg & Suh, 2005; López-Otín et al., 2013). Despite considerable effort and the development of many theories, the underlying process is still largely unknown (Kirkwood & Austad, 2000; Weinert & Timiras, 2003; Rattan, 2006).

Researchers distinguish between chronological and biological age. Chronological age is defined as the absolute time that an individual lives. In contrast, biological age is a broader concept that takes the individual physical and mental health into account, thus capturing individual differences of the aging process. Most aging studies search for associations of chronological age with clinical and molecular phenotypes (Warming et al., 2002). However, several studies used phenotypes, such as lung function, grip strength or bone mineral density, as proxies to investigate molecular changes in biological aging (Jackson et al., 2003; Bell et al., 2012; Levine, 2013). Researchers also investigated reasons of retarded biological aging and longevity by comparing centenarians with younger controls (Biagi et al., 2012; Sebastiani et al., 2012).

The life expectancy in the UK increased by 5.3 years for men and 4.7 for women over the last two decades and is predicted to further increase in the next twenty years (Oppean & Vaupel, 2002; Office for National Statistics 2014). With increasing life expectancy, age-related diseases are expected to rise dramatically (700 000 people suffered from dementia in 2000, 800 000 in 2012 and approximately 1 million people will be affected by dementia in 2021 (Alzheimer’s Society 2014)) with major impacts on healthcare costs. Thus, a better understanding of aging and its influence on disease is a long term public health goal and a hot topic of current medical research.

Omics technologies provide valuable tools to study aging on the molecular level. Reductionist data analyses, testing the measured variables separately for association with age, have been extensively applied. Such studies successfully identified hundreds of epigenetic mutations, gene expression levels, metabolite concentrations to be linked with chronological and/or biological age (see below for details). Even though these results improved our understanding of aging as a complex phenotype, the mechanisms underlying these associations and the impact of interactions between different biological entities remain elusive in most cases. In contrast to reductionist approaches, systems biology aims to analyse all components of a biological process simultaneously taking into account their interactions and their intrinsic hierarchical structure (Ideker et al., 2001; Barabási & Oltvai, 2004). With more and more high-throughput data becoming available, systems biology has led to many new methods and their successful application on age and age-related phenotypes (as outlined below).

In this review, we will briefly summarize the current progress in ‘omics’ technologies and their application in aging research. We will then highlight some problems of the reductionist approach and discuss how these may be overcome using systems biology. We present a selection of statistical methods used in systems biology along with their current and possible future applications in the field of aging research to move from biomarkers of aging to a more holistic understanding of the aging process.
Omics and aging

New technologies allow the measurement of ‘omics’ data and numerous association studies have been conducted. Valdes et al. (2013) thoroughly reviewed the application of these technologies to identify molecular markers of aging from each omics level. Therefore, the following section will only briefly highlight some key results and concentrate on recent findings.

Genomics

Genomics was the first omics field for which high-throughput measurements became available. Current chips are able to measure up to 5 million single nucleotide polymorphisms (SNPs) (Ha et al., 2014). Today, next-generation sequencing technology is slowly replacing the chip technology as the cost of sequencing has dropped below $0.10 per million bp (Liu et al., 2012). Thus, gene variation is nowadays often available at single nucleotide resolution.

While aging (or rather longevity) itself was found to be only about 20% heritable (Murabito et al., 2012), many age-related diseases are highly heritable. For instance, Alzheimer’s disease (AD) shows a heritability above 70% (Gatz et al., 2006) and osteoarthritis (Ishimori et al., 2010) or cataract show 50% heritability (Hammond et al., 2001).

The GenAge database contains about 300 human candidate genes for aging based on homology with model organisms (Tacutu et al., 2013). Sebastiani et al. (2012) recently published a refined model consisting of 281 SNPs to distinguish between centenarians and younger controls in a cohort of 1715 people. One of these SNPs is located in ApoE, which is so far the only gene that has been reliably associated with longevity at genomewide significance level (Deelen et al., 2011; Nebel et al., 2011). Common genetic variants at this locus have been associated with accelerated aging and cognitive decline (Johnson, 2006; Davies et al., 2014), possibly by increasing the risk for coronary artery disease, stroke and AD (Smith, 2002). Even though some studies provided evidence that mutations of FOXO transcription factors are related to longevity (Willcox et al., 2008; Flachsbart et al., 2009), as well, GWASs failed to replicate this at the level of genomewide significance.

Epigenomics

Epigenomics describes the study of heritable changes in the genome that are not caused by DNA sequence mutations (Lodish, 2013). The most common epigenetic mechanism is DNA methylation, which is known to often silence gene expression. In contrast to the genome, which is the same in all cells, the epigenome is an important factor of cell differentiation leading to profound epigenetic differences across different cell types (Meissner, 2010). The current methylation chip by Illumina measures over 485 000 methylation sites and covers 99% of all RefSeq genes (Illumina 2011). However, it covers less than 10% of variable regions (Ziller et al., 2013).

The epigenome is influenced by environmental and lifestyle factors (Nakajima et al., 2010; Alegria-Torres et al., 2011; Breitling et al., 2011) and is associated with many complex diseases such as neurodegenerative disorders (reviewed by Portela & Esteller, 2010) and cancer (Ehrlich, 2002; Horvath, 2013). Nearly 500 differentially methylated regions were found to be associated with chronological age and age-related phenotypes such as lung function, cholesterol levels and maternal longevity (Bell et al., 2012). A recent study by Weidner et al. (2014) showed that methylation patterns of just three sites are sufficient to predict chronological age. Thus, many of the previously identified methylation sites might not be independently associated with age. Interestingly, variation in methylation with age is consistent across several tissues and cell types (Horvath, 2013). Together, they form a global pattern of hypomethylation in repetitive sequences, hypermethylation in promoter regions and higher intercell variability (Cevenini et al., 2008; Bacalini et al., 2014). Besides DNA methylation, other epigenetic changes, such as histone methylation and acetylation, have been found to be associated with longevity in model organisms (Dang et al., 2009; Greer et al., 2010). Investigating these modifications in humans could shed light on so far unknown mechanisms of aging.

Transcriptomics

Genes are transcribed into RNA molecules, which are further processed in a tightly controlled process. The entirety of the RNA transcripts is referred to as transcriptome. It can be divided in coding RNAs, which are further translated in proteins, and noncoding RNAs, which perform various functions, such as regulation of gene expression (Eddy, 2001). Transcript abundances can be measured by either chips or sequencing methods.

Similar to the epigenome, gene expression was shown to dramatically change with age. A pioneer study comparing postmortem human frontal cortex tissue samples between 30 individuals of different ages yielded 463 differentially expressed genes (Lu et al., 2004). Despite the small sample size, results were replicated in subsequent experiments. Four years later, Berchtold et al. (2008) identified several thousand age-related changes in gene expression in four different brain tissues. Later studies by different groups identified profound changes in the transcriptome with age in further tissues, such as skin, adipose tissue ($N = 865$) (Glass et al., 2013) and kidney ($N = 134$) (Rodwell et al., 2004). Most of these changes did not overlap in different tissues. A meta-analysis across different species and tissues revealed only 73 genes consistently associated with age (de Magalhães et al., 2009). This suggests that most observed age-related changes in the transcriptome are either species and tissue specific or false-positive discoveries (reviewed by Valdes et al., 2013). In their meta-analysis, genes related to immune response and lysosome tended to be overexpressed, while genes related to mitochondria and oxidative phosphorylation were underexpressed in elderly (de Magalhães et al., 2009).

Proteomics

Proteins are translated from coding transcripts. Due to alternative splicing and post-translational protein modifications, the number of proteins is estimated to be two orders of magnitudes higher than the number of genes (Ginsburg & Haga, 2006). However, current proteomic techniques based on immunomassays, protein arrays or mass spectrometry can measure only a small fraction of the proteome (up to 1000 proteins in a sample). The most comprehensive description of the human proteome across various tissues to date consists of 18 097 proteins (19 376 isoforms) collected from ten thousand mass spectrometry experiments (Wilhelm et al., 2014).

Due to these technicalities, ‘proteomics’ studies in aging research so far focused on smaller sets of proteins and small sample sizes. In an early study of protein abundance in the vastus lateralis muscle, Gelfi et al. (2006) observed higher abundance of several proteins involved in aerobic metabolism and a lower abundance of proteins involved in anaerobic metabolism in the elderly. Besides this, six transport proteins were consistently underexpressed in older individuals. However, only 12
samples were analysed in this study without replication. A recent study by our group analysed over 1000 proteins in 200 plasma samples using the SOMAscan assay (Menni et al., 2015). Eleven proteins were found to strongly associate with chronological age as well as age-related phenotypes such as lung function and blood pressure. The results were replicated in an independent cohort.

Even though comprehensive proteomics studies are still missing, proteins are likely to be associated with several age-related diseases. For instance, cardiovascular disease (Mehra et al., 2005) and AD (Svardfager et al., 2010) are consistently associated with elevated levels of pro-inflammatory cytokines.

Post-translational modifications – glycomics

Post-translational modifications are important elements of proteins, which can alter their biochemical properties such as protein structure, binding preferences and enzyme activity. There are many different modifications ranging from addition of small molecules (e.g. acetylation or phosphorylation), over addition of larger molecules such as lipids or sugar chains (e.g. palmitoylation, glycosylation), to the addition of whole proteins (e.g. ubiquitination).

The most common modification is glycosylation, which attaches sugar chains to proteins. The attached oligosaccharides – glycans – are supposed to mainly serve as structural elements of proteins or specific binding sites for other glycans or proteins (Varki et al., 2009). However, glycans are highly diverse and many of them are not yet characterized or annotated. Thus, glycans might have many additional functions. For example, glycans in the gut act as food for microbes (Koropatkin et al., 2012), which could be implicated in immune functions that are important in aging. Recent development allows the high-throughput measurement of glycans of either a single protein or all proteins simultaneously (Royle et al., 2008; Pucic et al., 2011).

The application of this technology on epidemiological cohorts revealed that glycan structures are stable for one individual over time (Gornik et al., 2009) but very diverse within a population (Knezevic et al., 2009; Pucic et al., 2011). Differences in glycomes were found to be related with various cancers (Fuster & Esko, 2005; Adamczyk et al., 2009). Differences in glycomes were found to be related with various cancers (Fuster & Esko, 2005; Adamczyk et al., 2009). However, glycans are highly diverse and many of them are not yet characterized or annotated. Thus, glycans might have many additional functions. For example, glycans in the gut act as food for microbes (Koropatkin et al., 2012), which could be implicated in immune functions that are important in aging. Recent development allows the high-throughput measurement of glycans of either a single protein or all proteins simultaneously (Royle et al., 2008; Pucic et al., 2011).

Metabolomics

Metabolomics investigates the low-molecular-weight molecules in a biological system. The measured molecules are often referred to as metabolites as many of them act as educts, products and intermediates of the cellular metabolism. Currently, the Human Metabolome Database (Wishart et al., 2013) contains more than 40 000 distinct metabolites from different tissues. Similar to proteomics, to date, there is no analytical method available to determine and quantify all metabolites in a single experiment. Current platforms, using either chromatography coupled with mass spectrometry or nuclear magnetic resonance, can measure roughly a thousand metabolites in untargeted settings and a smaller number using predefined targeted approaches. The restriction of the targeted approach comes with the advantages of higher sensitivity, absolute instead of relative quantification and straight-forward compound identification (Patti et al., 2012; Tzoulaki et al., 2014).

In 2008, the first metabolome-wide association study on age analysed the plasma metabolome of 269 individuals using an untargeted approach. The authors found 100 of 300 compounds to correlate with chronological age (Lawton et al., 2008). More recently, larger cohorts were employed to study the association of metabolites and age using both targeted and untargeted metabolomics platforms. Yu et al. (2012) analysed 131 targeted metabolites in 2162 individuals from the KORA study, while we analysed 280 untargeted metabolites in 6055 twins from the TwinsUK cohort (Menni et al., 2013b). Both studies identified the same metabolites to be associated with chronological age. Many of the those metabolites were also found to significantly correlate with age-related phenotypes such as lung function, bone mineral density and cholesterol levels (Menni et al., 2013b), cancer (reviewed by Teicher et al., 2012) and type 2 diabetes (N = 100) (Suhre et al., 2010; Menni et al., 2013a). One of those metabolites is C-glycosyltryptophan, a potential degradation product of glycosylated proteins.

Microbiomics

The human microbiome describes the complete set of microbial species (and their genomes) hosted by the human body. The largest microbial community resides in the gut, where microbial cells and their genes outnumber human cells (10:1) and genes (100:1) (Peterson et al., 2009; Zhu et al., 2010; The Human Microbiome Project 2014a). More than 10 000 different species with millions of protein-coding genes were identified by the Human Microbiome Project (Turnbaugh et al., 2007; Peterson et al., 2009; Biagi et al., 2012) and >1000 of these microbes have so far been fully sequenced (The Human Microbiome Project 2014b). Although twin studies have found a modest genetic influence on some phyla, most of the variation is environmental (Goodrich et al., 2014).

The composition of the microbe flora varies a lot across individuals (Turnbaugh et al., 2007; Zhu et al., 2010) and even between different parts of the body (Kong, 2011). It has a huge influence on many biological processes such as immune response, metabolism and disease (Zhu et al., 2010; Grice & Segre, 2012). While the microbiome seems to be relatively stable during adulthood, it changes significantly in later life (Guigoz et al., 2008; Biagi et al., 2010; Claesson et al., 2011). Biagi et al. (2010) observed drastic changes in the gut microbiome of centenarians compared with young adults as well as elderly, namely a general loss of diversity and increased abundance of bacilli and proteobacteria. The latter were reported to promote inflammation under certain conditions (Round & Mazmanian, 2009). Similar findings were revealed in other elderly populations, which also considered the dietary and residential situation of elderly patients (Claesson et al., 2012).

Phenomics

Simultaneously with omics data, the dimension of clinical and lifestyle traits, particularly clinically used intermediate traits, keeps increasing. Epidemiological studies collected thousands of clinically relevant phenotypes beyond omics data types. These range from anthropometric measures to health and lifestyle questionnaires (Moayyeri et al., 2013). Collecting high-dimensional clinical data is important to unveil pleiotropy of genes and interactions amongst clinical phenotypes such as morbidities (Houle et al., 2010). Driven by omics technologies, statistical and bioinformatic methods to analyse high-dimensional data are becoming available. These facilitate the investigation of numerous clinical
Phenomics is especially important for aging research. Dozens of clinical phenotypes, such as Parkinson’s (Reeve et al., 2014), AD (McAuley et al., 2009), body mass index, blood pressure (Mungreiphy et al., 2011) and bone mineral density (Warming et al., 2002), as well as lifestyle parameters, such as nutrition (Wieser et al., 2011), smoking and physical activity, are strongly related to age (Harman, 1988; Wang et al., 2009). Composite measures such as the Rockwood frailty index (Rockwood & Mitnitski, 2007) combine several of those clinical traits to form a more homogenous phenotype – frailty – from its diverse appearance. Such frailty measures can be considered as measures for biological age (Mitnitski et al., 2013). Many of these (and other) clinical phenotypes correlate or even depend on each other (McAuley et al., 2009; Baylis et al., 2014). Only extensive collection of data and their joint analysis will help to unveil these dependencies and find causal relationships.

**From omics to systems biology**

Most of the studies summarized above concentrated on the bivariate associations of age (or age-related diseases) with one type of omics data. However, there are strong interdependencies within and between the different omics data (see Fig. 1).

Correlations can be observed practically between all levels of biological organization. Following the central dogma of molecular biology, genomics, transcriptomics and proteomics are correlated “by definition”. Furthermore, metabolite concentrations are influenced by genetic variants (Shin et al., 2014) and epigenetic factors (Petersen et al., 2014) mediated through changes in gene expression or enzyme activity. Methylation levels do not only influence the gene expression (Jaenisch & Bird, 2003), but are also correlated with gene variants (Bell et al., 2012) and environmental factors (Breitling et al., 2011). Our group has recently demonstrated that even the microbe composition is partly under host genetic influence (Goodrich et al., 2014). Similarly, all levels of omics data are influenced by genetics as well as by environment and aging. Correlations, however, do not only occur between but also within each type of data. For instance, in genomics linkage disequilibrium, the correlated occurrence of SNPs is a ubiquitous phenomenon. Transcription factors often coregulate the expression of multiple genes (Allocco et al., 2004), and methylation patterns of neighbouring CpG sites were reported to be correlated (Bell et al., 2012). Metabolites are linked by a network of biochemical reactions, causing strong correlations between them (Krumsieck et al., 2011). Even phenotypes often cluster. Comorbidities, the over proportional co-occurrence of diseases, were shown to affect many diseases possibly through shared underlying mechanisms (Goh et al., 2007).

These biological correlations can confound the associations and this is a major issue of current research. For instance, 153 metabolites were found by our group to be associated with age, but subsequent analyses showed that only 22 of them are associated with age independently (Menni et al., 2013b). Similarly, 21 of 24 measured IgG glycans were correlated with age, but only 3 of them explain 58% of the variance (Kristic et al., 2013). The same was found for epigenetic data (Weidner et al., 2014). Huge lists of associations with aging are being unveiled using all kinds of data, but the biologically interesting, causal associations are often obscured by this wealth of results. Approaches taking simultaneously information from all omics levels into account are needed to reconstruct the processes involved in aging on a systems level (Valdes et al., 2013).

Even though high-throughput technologies are advancing and more and more data are becoming available, integration of omics remains a challenging problem. Besides the restricted availability of multi-omics data sets for the same samples, technical limitations hamper the integration process. While genomics and transcriptomics are able to measure the entire set of variants, other omics (e.g. proteomics and metabolomics) measure only a small fraction of all entities. Many high-throughput technologies suffer from considerable technical variation and strong batch effects. Stringent quality control and thorough data normalization are crucial when analyzing this type of data. Furthermore, the complexity of the organism has to be taken into account. While the genome is more or less stable, all other levels of omics change between cell types and over time. Many samples, such as whole blood, contain a mixture of different cell types with potentially different epigenomes, transcriptomes (Houseman et al., 2012; Jaffe & Irizarry, 2014). Finally, different organs and cells influence each other. The blood metabolome, for instance, is heavily influenced by processes occurring in the liver or in other organs, and multi-tissue samples are needed to fully understand these. This in turn is not always feasible in an epidemiological setting as collection of tissues often involves invasive procedures. Nevertheless, data integration is an important and active field of research. A first step of data integration is the integration and joint interpretation of separate results. The Digital Ageing Atlas (Craig et al., 2014) summarizes more than 4000 age-related changes across different technologies to facilitate systems-level analyses of aging.

**Introduction to systems biology**

The aim of systems biology is to understand the system and its functions as a whole rather than as separate components (Cassman, 2005), with the final objective to mathematically model biological systems and simulate their outcomes. As a first step, the complex interactions and dependencies between these components must be formally described to enable systematic analysis and simulation of the biological system of interest. A technique widely used in systems biology is to translate biological interactions into mathematically well-defined networks (graphs). For instance, metabolites interact in chemical reactions, thus forming a network in which nodes describe the metabolic compounds.
and edges indicate chemical reactions. Similarly, transcription factors bind DNA to control expression, forming the gene regulatory network (GRN) and interacting proteins build a protein interaction network (PPI) (cf. Fig. 2B). These networks interact, making data integration an important aspect of systems biology. One example for a phenotypic network was created by Goh et al. (2007) using diseases as nodes and connecting diseases with shared genetic risk factor by edges (cf. Fig. 2A). By doing so, they showed that many disorders share a set of underlying genetic risk variants and that similar diseases are caused by similar genes.

Graphs can be explored using a variety of established algorithms. One common task is the identification of modules, that is subgraphs in which nodes share certain properties. In biological networks, modules correspond to functional units, such as the glycolysis pathway in the metabolic network. The modules are usually interconnected and together form a hierarchical structure in which the distribution of node degrees — the number of edges per node — follows the power law (Barabási & Oltvai, 2004). Hence, most nodes have only few connections and few nodes have many connections. These highly connected nodes are called hubs (Albert et al., 2000; Jeong et al., 2001). Several other measures exist to describe the topology of networks and topological features of nodes. For example, the clustering coefficient measures how densely the neighbourhood of a node is connected and thus highlights nodes which are central within a cluster (e.g. LEPR in Fig. 2A). Another measure is the betweenness centrality, which measures the proportion of pairwise shortest paths containing a node. It thus quantifies the importance of a node for connecting other nodes from different modules (e.g. Parkinson's disease in Fig. 2A and APOD in Fig. 2B). The highly connected, central nodes are thought to be key players in the system, connecting several modules and controlling network fluxes. They were shown to be of particular importance for many diseases and survival of the organism (Barabási & Oltvai, 2004; Joy et al., 2005; Yu et al., 2007).

Many software packages for graph analysis and visualization are publicly available. For instance, the R package igraph (Csardi & Nepusz, 2006) or the standalone program Cytoscape (Shannon et al., 2003) can be used to analyse and visualize graphs. Cytoscape also provides easy integration of biological databases such as Gene Ontology (Ashburner et al., 2000), Reactome (Croft et al., 2014), the Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa & Goto, 2000) or BioGRID (Chatr-Aryamontri et al., 2013) by third-party apps. Several methods were developed to identify modules of nodes which are jointly affected by the condition of interest. Two publicly available examples are the Cytoscape plugin jActiveModules (Ideker et al., 2002) and the R package BioNet (Beisser et al., 2010).

Here, we present a selection of current methods to construct and analyse biological networks as an approach to systems biology and their impact on aging research.

**Enrichment and network topology analysis in predefined networks**

A popular approach to put the results of an association study in a systems biology context is projecting the variables of interest — such as age-related genes, proteins or metabolites — onto known biological (reference) networks. The neighbourhood of these target variables and their topological properties can then be assessed using the experimentally predefined PPI, GRN or metabolic networks. Instead of interpreting individual entities separately, a priori knowledge about their interactions and common functions can be used to identify modules that are jointly affected by the condition of interest.

Several databases offer a collection of experimentally identified interactions that can be used as predefined reference networks for enrichment and topology. In case of PPI, the Human Protein Reference Database provides more than 40 000 PPIs (Keshava Prasad et al., 2009), the Database of Interacting Proteins more than 7000 interactions (Xenarios et al., 2002) and the MIPS mammalian protein–protein database roughly 1000 hand-curated interactions of human proteins (Pagel et al., 2005). GRN are provided by the ChIPBase (Yang et al., 2013), which contains six million transcription factor binding sites from >300 experiments. Metabolic reactions are amongst others provided by KEGG.

Enrichment analysis is a convenient way to incorporate existing knowledge from biological reference networks without analysing graph topology directly. Therefore, predefined (functional) modules within the reference networks are used to test overrepresentation of associated genes, proteins or metabolites in these groups. When investigating genes, researchers usually use Gene Ontology to group genes based on biological processes, molecular functions or subcellular localization. For metabolites, the KEGG and Reactome databases provide curated information about biochemical pathways. The R packages GSEAbase, GAGE (Luo et al., 2009) and the web-service MSEA (Xia & Wishart, 2010)
are just some of many available implementations and variations in the original gene set enrichment analysis (Subramanian et al., 2005) algorithm.

In aging research, enrichment analysis unveiled an overexpression of genes involved in immune response, lysosome and glycoproteins and an underexpression of mitochondrial- and oxidative phosphorylation-related genes in old people compared with young (de Magalhães et al., 2009). In human brain tissue, oxidative stress/DNA repair and inflammation-related genes were shown to be enriched in the set of differentially expressed genes between young and old individuals (Lu et al., 2004). Enrichment analysis facilitates the identification of pathways that are important for the aging process. It thus helps to make sense out of the individual associations and find biological interpretations for the observed molecular changes.

To become independent of predefined module annotation and to enable more detailed network analysis, the variables of interest can also be mapped directly on the known PPI, GRN or metabolism networks. Modules can then be identified dynamically based on the measured data. Moreover, additional topological properties of the variables of interest can be assessed.

Studying human PPI networks revealed that genes that are associated with aging by homology have higher node degrees and higher betweenness centrality compared with other genes (Bell et al., 2009). Furthermore, aging-related genes are not spread throughout the interactome, but cluster in few tightly connected modules. These modules were enriched in DNA damage repair and stress response genes (Kriete et al., 2011). The high connectivity of aging genes was used by Tacutu et al. (2012) to select neighbours of longevity-related genes in a PPI network as longevity-gene candidates. Subsequent experiments in C. elegans revealed 30 new longevity-associated genes, proving the potential of network biology for candidate gene selection. Using a modified PPI network, Wang et al. (2009) showed a tight connection of the genetic causes of aging and disease. These results indicate that aging does not occur due to random errors but is an organized process. Another PPI-based approach to data integration was developed by West et al. (2013). They incorporated epigenomic data by assigning DNA methylation sites to each protein in the graph and then identifying modules of differentially methylated genes/proteins in the resulting network. By doing so, they avoided predefined gene sets as used by enrichment analysis. The analysis revealed three differentially methylated modules, which were replicated across several tissues. Two of them contained mainly transcription regulating genes, while the third one contained genes related to stem cell differentiation.

A drawback of experimentally derived PPI or GRN is that such methods detect up to 50% false positives while many true interactions are missed (Huang & Bader, 2009; Marbach et al., 2012). Even more importantly, those reference networks completely ignore the temporal-spatial properties of the interactions. This restricts results to already observed, possibly inactive interactions. One method to overcome the static nature of PPI networks are Negative–Positive (NP) networks (Xia et al., 2006). These integrate the PPI network with transcriptomics data by restricting it to edges between (anti-)correlated proteins/genes. Therefore, only those interactions (edges) that are active under the observed condition are further analysed. Xue et al. (2007) applied this method to the previously mentioned data set of brain gene expression and unveiled two anticorrelated modules containing cell proliferation- and cell differentiation-related proteins. Two other modules consisting of protein processing and immunity-related genes, respectively, were found to be slightly correlated with the cell proliferation module. A recent study went one step further and restricted a PPI network to highly expressed genes in different stages of aging for each sample separately, thus generating a set of dynamic binding networks instead of a single network. Even though the global properties of all those graphs were very similar, the centrality of several genes correlated with age (Faisal & Milenković, 2014).

Incorporating biological networks to analyse aging-related changes showed the tight connection of aging and disease on a molecular level. Furthermore, it has been shown that aging affects central genes, which are important for the network integrity (Bell et al., 2009). While network-based enrichment and analysis using PPI networks is common for genetic and transcriptomics data, it has not been applied to aging studies using metabolomics data. This could be a promising approach to systematically identify metabolic pathways jointly affected by the aging process.

**Analysis of data-derived networks**

Despite their successful applications, all approaches presented so far rely on predefined, static networks. To overcome the limitations of such networks, inferring networks directly from the measured data is the next step.

**Weighted gene co-expression network analysis**

The weighted gene co-expression network analysis (WGCNA) (Zhang & Horvath, 2005) infers gene–gene interaction networks directly from transcriptomics data. Miller et al. (2008) applied this method to the previously mentioned gene expression data set of 30 human frontal cortex samples at different ages and then compared the results with a network derived from an AD transcriptomics study. It revealed significant overlap between healthy aging and AD, suggesting that there might be a shared molecular basis for both processes. Three AD network modules overlapped with aging network modules, containing mostly synapses- and transport- and transcriptional regulation-related genes.

**Gaussian graphical models**

Despite the successful application of WGCNA on transcriptomics data, Krumsiek et al. (2011) showed that ordinary correlations are not suitable to analyse metabolomics data from large cohort studies. They analysed metabolite concentrations of >1000 samples and found that more than half of all pairs of 151 metabolites correlated significantly, even when using a restrictive Bonferroni correction at an alpha level of 0.01. This is largely due to indirect associations, which cannot be distinguished from direct associations by the Pearson correlation coefficient. Graphical models (GMs), also known as conditional independence graphs, were proposed to overcome this problem and infer biological meaningful networks from metabolomics (Steuer, 2006; Krumsiek et al., 2011) as well as other omics data (de la Fuente et al., 2004; Yuan et al., 2011; Mangin et al., 2012). GMs are probabilistic models where an edge between two variables illustrates their conditional dependence given all other variables in the model. Implicitly, the absence of an edge represents the conditional independence of the according variables. Several algorithms to infer GMs from purely binary data are publicly available as R packages (Wainwright et al., 2006; Höfling & Tibshirani, 2009; Guo et al., 2010; Ravikumar et al., 2010). Their counterparts for purely continuous data are Gaussian graphical models (GGMs), which use partial correlations to infer graphs. A partial correlation of two variables X and Y conditioned on a set of variables Z quantifies the portion of the correlation between X and Y which cannot be attributed to Z. Several algorithms exist to infer GGMS (d’Aspremont et al., 2006; Meinshausen & Bühlmann, 2006; Yuan & Lin, 2007; Friedman et al., 2008).
Several of them, such as the well-established graphical lasso (Friedman et al., 2008; Mazumder & Hastie, 2012), use regularization to further reduce the number of edges in the graph. This allows researchers to concentrate on fewer high-confidence interactions.

Gaussian graphical models can reconstruct biological pathways from metabolomics and transcriptomics data, but have not yet been applied in aging research. However, their application could help reduce the ‘overabundance’ of results to fewer, meaningful associations. The major drawback of GMs is that they can only be used for pure Gaussian or pure binary data. Shin et al. (2014) overcame this problem by first constructing a GGM from metabolite concentrations and then adding gene variants as nodes and connecting them with associated metabolites. The resulting network illustrates the genetic control of the metabolism in an intuitive way. However, it is no longer a GM, and edges do not indicate conditional independence any more.

**Mixed graphical models**

Recent developments allow the integration of different types of data while maintaining the favourable properties of GGMs, namely mixed graphical models (MGMs) (Tur & Castelo, 2012; Chen et al., 2013; Fellinghauer et al., 2013; Lee & Hastie, 2015). Fellinghauer et al. (2013) proposed a very flexible algorithm based on stability selection (Meinshausen & Bühlmann, 2010). It makes use of established methods such as random forests or regression models to rank interactions between variables of different types. Thus, it can handle many different data types such as disease states, metabolite levels and gene variants. Due to the usage of stability selection, it has an intrinsic error control. MGMs provide a powerful tool for multivariate analyses of high-dimensional data, but have not been applied in biological research, yet. Their application could shed light on the complex relationship between aging and disease.

Gaussian graphical models as well as MGMs are undirected models. Therefore, neither of them can be used to infer causal direction. In epidemiological research, Mendelian randomization is a common approach to infer causality from observational data. It takes advantage of the invariability of gene variants to separate the study population in groups, thus mimicking a randomized controlled trial (for further details, see Brion et al., 2014). Mendelian randomization can be used to further investigate edges of interest that were previously identified by GMs. However, it relies on stable associations with genetic variants and assumes that this genetic variant is not related to any other potential confounding factor. Due to these restrictions, it is not suitable to infer large-scale networks.

**Bayesian networks**

Another approach that allows inferring causality from observational data under certain assumptions is based on Bayesian networks (BNs). Similar to GGMs, BNs are probabilistic models in which edges represent the conditional independence between variables. However, BNs are DAGs, thus distinguishing between an influence of X on Y and the influence of Y on X. In return, the acyclicity of the causal graph is an assumption which might not hold true for biological networks. The application of BNs on high-throughput transcriptomics data by Friedman et al. (2000) demonstrated the potential of this method to extract biological meaningful associations without prior knowledge. Several different methods are available to estimate the structure of BNs from binary, continuous and even mixed data such as the R packages bnlearn (Scutari, 2010) (Table 1).

The methods presented here are just a selection of the available methods for graph inference. Several other methods such as Boolean networks (Shmulevich et al., 2002) or differential equation systems (Chen et al., 1999; Lorenz et al., 2009) are commonly used for modelling biological networks.

The development of new techniques facilitates graph inference from high-dimensional data, and the presented studies illustrate their usefulness in biological research. However, most graph inference methods rely on large sample sizes and usually more samples than variables are needed. When analysing omics data, particularly genomics or transcriptomics, this is often not feasible and it is referred to as the \( n \ll p \) problem. Another common problem is overfitting of models due to the high number of parameters. Some techniques such as regularization have been proposed to relax these constraints and reduce overfitting. Nevertheless, stringent cross-validation and replication in independent cohorts should be employed to avoid spurious results. Finally, many high-throughput methods suffer from considerable technical variation and strong batch effects. Researchers should carefully normalize all mea-

---

© 2015 The Authors. Aging Cell published by the Anatomical Society and John Wiley & Sons Ltd.
measurements according to current standards before integrating different data sets.

Model biological systems

The ultimate goal of systems biology is not only the qualitative exploration, but the quantitative modelling of the organism, facilitating in silico experiments, hypotheses generation and predictions. The first – and so far only - attempt to model a whole organism was conducted by Karr et al. (2012). They created a model of a mycoplasma genitalium cell simulating cell cycle and predicting metabolite concentrations. However, the model is far from perfect (Freddolino & Tavazoie, 2012) and too primitive to be adapted to more complex organisms. Currently, modelling eukaryotic cells or even whole organisms is not feasible. Also, processes like aging are too complex to be entirely modelled. However, some effort has been undertaken to create network representations of smaller subsystems as well as certain aspects of the aging process. For instance, Gillespie et al. (2004) simulated aging of yeast based on the accumulation of extrachromosomal ribosomal DNA circles. Also, Oda & Kitano (2006) summarized results from several hundred studies to create a model of the Toll-like receptor (TLR) signalling network. The same group also created a similar model for epidermal growth factor receptor signalling (Oda et al., 2005). Both studies revealed a bowtie-like global structure with one important key regulator. However, both networks are only qualitative descriptions without kinetic parameters. Thus, they cannot be used for computer simulations.

Other groups concentrated on even smaller subsystems to facilitate quantitative modelling. One study investigated the influence of increased cortisol levels on hippocampus activity (McAuley et al., 2009). A quantitative model was created to simulate the decline in hippocampal output with age and the acceleration of this process due to acute and chronic increases in cortisol levels. Simulations using ordinary differential equations suggested that chronic increase in cortisol levels leads to faster decline in hippocampal output than acute bursts, but could be treated more efficiently. Sozou & Kirkwood (2001) modelled cell senescence based on telomere shortening and oxidative stress. The same group also described the influence of chaperones and accumulation of misfolded proteins on aging (Proctor et al., 2005). Other groups investigated various further aspects of the aging process, such as mitochondrial fusion and fission events and accumulation of defective mitochondria (Kowald et al., 2005; Figge et al., 2012), incomplete replication of epigenetic information (Przybilla et al., 2014) and age-related alterations in the lipid metabolism (McAuley & Mooney, 2015). Adjusting the kinetics of such models to correspond to experimental observations allows to come up with plausible hypotheses about the causes of aging.

In contrast to earlier presented networks, which inferred large-scale networks from data (top-down approach), these approaches model small subsystems in high details based on expert a priori knowledge (bottom-up approach). Such bottom-up models allow mechanistic insights into the processes of aging that cannot be generated by individual association studies. Moreover, they facilitate the development of new hypothesis and testing the plausibility of current hypothesis.

Conclusions and challenges

The major recent advances of omics technologies are now enabling the simultaneous measurement of millions of biochemical entities. Association studies have revealed many associations of omics data with aging and age-related diseases. After decades of reductionist studies, network analysis and integrated omics data analysis have begun to target the aging process at a systems level. As a result, some studies take into account also the interaction effects between variables. However, given the complexity of aging, new methods are needed to further unveil the multiple interactions.

Systems biology already provides such methods, but their application on real biological problems lags behind. For example, GGMs have been adapted to mixed data types and could readily be applied in aging research. Also, several studies developed models of processes that contribute to aging. These provide detailed knowledge about important components of the aging process and their interactions. Building on these results, future studies should aim to integrate these different parts to gain a more systems-level understanding of aging.

However, in many cases, the available data limit the possibilities. Problems such as incomplete data, asynchronous experiments, strong batch effects and insufficient sample sizes have to be dealt with. Another issue is the limited availability of multi-omics data sets, which complicates replication of results in this field. A variety of different methods, protocols and platforms further hampers reproducible results. As replication of results is crucial to prevent spurious results and validation, methods like splitting the available data into discovery and replication sets should be considered more often.

Despite these obstacles, there are several large population studies in existence with multi-omics data available which could be explored using systems biology approaches. For instance, the GTex project aims to collect gene expression and methylation data from multississue samples (The Gtex Consortium 2013). Simultaneously, the development of new methods should help to analyse real, partially incomplete data sets and facilitate analysis of multiissue and multi-organ data, thus enabling the investigation of real systems-level effects. Addressing these problems and developing integrated models of aging should improve our understanding of the aging process, thus allowing the development of strategies to improve health in old age.

Funding

This work was supported by the EU Framework Programme 7 small-scale focused research collaborative project EurHEALTHAging [277849]; TwinsUK was funded by the Wellcome Trust; European Community’s Seventh Framework Programme [FP7/2007-2013]. The study also receives support from the National Institute for Health Research (NIHR) Clinical Research Facility at Guy’s & St Thomas’ NHS Foundation Trust and NIHR Biomedical Research Centre based at Guy’s and St Thomas’ NHS Foundation Trust and King’s College London. TDS is a NIHR senior research fellow.

Conflict of interest

None declared.

References


