1 INTRODUCTION
A fundamental aspect of disease research involves the understanding of biological processes that underpin observed phenotypes. In order to achieve this level of understanding, diseases need to be described as collections of measured phenotypes and these phenotypes need to be analysed in relation to their genetic causes, genomic effects and linked with information on molecular interactions. One consequence of these efforts could be the ability to produce predictive models of phenotypes from genomic profiles with the aim of describing diseases more accurately. Such models will be helpful in understanding the genetic basis and molecular mechanisms leading to complex or rare developmental diseases, the process of ageing, as well as the characterisation and progression of cancer types. In particular, models built from model organism datasets can be translated into insights on humans in areas such as disease gene identification and drug target testing.

Methods for assigning genotypes to phenotypes have been developed and used intensively (Ramanan, 2012), with the example of genome wide association studies (GWAS) applied for identifying causative genotypes for various conditions and phenotypes. These studies are usually followed by functional experiments, trying to unravel the biological mechanisms that could influence the phenotypes given the observed genotype. Moreover, there have been numerous efforts to link gene expression to phenotype. Data classification methods have been used extensively to characterise healthy or diseased tissue from the context of gene expression. In the above examples, although the genetic and genomic outcomes of the disease can be associated with phenotypes, the biological events leading to the phenotype at the systems level is not discovered. Signalling and metabolic pathway analyses can inform on the specific mechanisms of the genetic/ causes of the phenotypes.

Here we identify areas of research that need to be further developed in order to facilitate computational prediction of the biological mechanisms that link genotypes to phenotypes. We break down the areas into three different themes (phenotype characterisation, gene expression to pathways and pathways to phenotypes) and describe their current status and future challenges.

2 ONTOLOGICAL CHARACTERISATION OF PHENOTYPES
In order to understand complex biological systems, reasoning chains reaching from a molecular level to the entire individual need to be built. With the availability of data from several model organisms, options are not only limited to human systems but may include data across different species. As a consequence, three major aspects of data integration need to be addressed: the integration across the different levels of complexity within an organism, the integration across species and the frequencies of occurrences of phenotypes (quantification). To facilitate data integration, numerous ontologies have been developed that define the meaning of biological concepts, such as the Gene Ontology (GO).

Available ontologies in the biomedical domain cover the different levels of complexity, i.e. ontologies that represent gene function (GO) as well as tissue information (e.g. BRENDA tissue ontology) or phenotypes (e.g. the Mammalian Phenotype Ontology or the Human Phenotype Ontology). However, the integration of data across the different levels of complexity is ongoing work (Hoehndorf, 2013; Oellrich, 2014). Solutions to combine phenotype data from different ontologies include the development of so called Entity-Quality (EQ) statements that enable the composition of phenotypes using species-independent ontologies (Mungall, 2010), e.g. GO (for the representation of processes) or UBERON (a cross-species anatomy ontology), but the application relies on manually curated EQ statements that are so far only available for a small selection of species and genotypes.

The quantification of phenotype data is slowly on its way: databases such as OrphaNet describe disease phenotypes with additional quantifiers, e.g. the phenotype dwarfish is very frequent in patients with a 12q14 micro deletion syndrome. While clinical databases already work on the inclusion of quantified phenotype data, model organism databases lag behind by not providing this information. Thus, quantified phenotype information cannot yet be used for data analysis and computational modelling.

In conclusion, more work is required to integrate data across the different complexity levels and to quantify phenotypes using the existing ontologies, in order to build reasoning chains that can be used for reliable, automated predictions.
3 GENE EXPRESSION TO PHENOTYPES

The ease of obtaining whole genome expression datasets has enabled more thorough classification of phenotypes associated with the expression of sets of genes. For example, classifications of tumour types from high-throughput gene expression and/or copy number profiles have helped unravel the complexity of different cancer types and better understanding of cancer progression, as well as the identification of new diagnostic biomarkers (Marisa, 2013). Given enough data sets, existing data mining methods can assign patterns of gene expression to the phenotypes under study.

Although “gene expression signatures” to phenotypic associations are an important step into determining the causal link between genes and phenotypes, it is still difficult to determine the underlying biological mechanism from gene expression data sets alone. In an experimental setting where a gene mutation is introduced and phenotypes and whole mRNA are profiled, the mRNA profile will include primary as well as secondary effects of the mutation, reflect tissue-specific expression, as well as developmental/cell-cycle specific gene expression. Tissue-specific and temporal based gene expression with matching phenotype measurements could be dealt with appropriate experimental controls, however, these are often absent or impractical to implement in large-scale phenotyping assays or in cases of meta-analyses from already available data sets (Oellrich, 2014).

4 PATHWAYS TO PHENOTYPES

Deriving the underlying mechanism of the phenotype, given the initial mutations and resulting gene expression involves the integration of knowledge on protein interactions and pathways (Khatri, 2012). This may come from direct protein level measurements, therefore enabling the use of computational simulations for the formulation of predictive hypotheses that can subsequently be tested experimentally. Such approaches have the potential to produce predictive mathematical models describing the underlying mechanisms at high-levels of detail (Petelenz-Kurdziel, 2013). However, they are not easy to implement on a large scale and are mostly useful when there is already substantial knowledge of the biological process involved. In cases where the biological process involved is unknown or poorly defined, high-throughput protein interaction data or high-level pathway information from pathway databases can help disentangle the mechanisms that are responsible for or induced by the observed gene expression. Boolean logic and other logic-based approaches (Papatheodorou, 2012) have been used successfully for qualitative pathway analyses, to generate hypotheses that link gene expression, pathways and phenotypes.

Further work needs to focus on linking the different levels of information, protein levels, gene expression and metabolic and signalling pathways into computational models that can handle qualitative and quantitative pathway parameters. Integrating different kinds of data sets from different species to solve a single, common biological process is an invaluable step in pathway analyses, but remains a difficult task. Advances in text-mining methods, as well as more accurate orthologous relationships between the genes of different species will help overcome these problems. Finally, recent efforts on multi-scale models of organs attempt to bridge the gap between molecular pathways and physiology through projects such as the Virtual Physiological Human (Coveney, 2013). Such efforts will facilitate a better understanding of the relationship between genes and phenotypes.

5 CONCLUSIONS

High-resolution gene expression data sets are providing more insight into the functional consequences of the genotype as well as clues into the mechanisms that might control the phenotype. At the same time, research utilising pathway analysis and data integration has been increasingly important in explaining the biological mechanisms under which genotypes (and gene expression) influence phenotypes. Some form of pathway analysis is routinely part of gene expression studies, however, this is hindered by the lack of detailed pathway maps and quantitative information on the reactions. From the perspective of phenotype characterisation, the development of different types of ontologies and links between them, is increasingly improving the integration of gene, tissue, anatomical and disease data sets within and between species. In recent years computational methods in mapping and organising these relationships have improved significantly, creating the basis for more detailed associations between genes, pathways and phenotypes in the future.

REFERENCES

Hochndorf, R. et al. (2013) Systematic analysis of experimental phenotype data reveals gene functions. PLoS ONE.

Marisa, L. et al. (2013) Gene expression classification of colon cancer into molecular subtypes: characterisation, validation and prognostic value. PLoS Medicine.

Mungall, C. J. et al. (2010) Integrating phenotype ontologies across multiple species. Genome Biology, 11(1), R2.

Oellrich, A. et al. (2014) Linking tissues to phenotypes using gene expression profiles. Database.

Khatri, P. et al. (2012) Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Computational Biology.

Petelenz-Kurdziel, E. et al. (2013) Quantitative analysis of glycerol accumulation, glycolysis and growth under osmotic stress. PLoS Computational Biology.

Papatheodorou, I. et al. (2012) Using answer set programming to integrate RNA expression with signalling pathway information to infer how mutations affect ageing. PLoS ONE.

Ramanan, V. K. et al. (2012) Pathway analysis of genomic data: concepts, methods, and prospects for future development. Trends Genetics 28(7): 323-332.

Coveney, P. V. et al. (2013) Integrative approaches to computational biomedicine. INTERFACE FOCUS 3(2).