Genetic variations in connection

SUMMARY

In this thesis solutions related to two aspects of the analysis of SNP data are presented. The first aspect regards the type of tools and technologies available for the analysis of SNPs in the context of biological pathways. The second aspect concerns the identification of potential biological function of the SNPs associated to a certain phenotype.

Currently, a variety of tools are used to perform analysis of genetic variants. **Chapter 2** presents an inventory and evaluation of tools that combine the analysis and visualization of variants linked to genes in pathways. We identified the advantages and the limitations of those tools. The purpose was to facilitate the work of both bioinformaticians and experimental biologists that are the first users of such tools. In addition, we propose new analytical features to add to software, such as the inclusion of new types of genetic interactions: edgetics, gene-environment interactions and epistasis relationships. **Chapter 3** presents an implementation of the BridgeDb identifier mapping database aimed at the improvement of interoperability in pathway analysis through mapping of genetic variants to genes. A gene-to-variant and variant-to-gene mapping using publicly available SNPs, is available in the BridgeDb repository. This has expanded the mapping domains available in that repository, that already included gene-products, metabolites, and reactions. The addition of the variants to gene mapping enables building a more complex data analysis in the different applications in which BridgeDB is integrated, such as R, PathVisio, and Cytoscape or can be used as a webservice e.g. in R or any programming language.

The second part of the thesis presents methods that are able to integrate multiple data in order to explain the genetic variation effect in diseases such as T2DM or in the development of obesity. In this respect **Chapters 4** and **5** show a workflow designed to further analyze a GWAS output, combining multiple data sources using pathway and network analysis. The result is a workflow that ends with a network construction, used as support for the interpretation of the data in a biological point of view. With the distinction that in **Chapter 4** the focus is on the SNPs that cause a protein coding effect, and in **Chapter 5** the main interest is on the understanding of the variants effects in non-coding area of the genome. The lesson learnt is pathway and network analysis are both valid methodologies that support data integration, and thereby allow us to dive deeper in the biological context. However, bottlenecks are still present concerning data re-usability. In addition, researchers need to select carefully the data that they wish to integrate, based on relevant biological connections with the genes affected by genetic variants.

Finally, **Chapter 6** reports an example in which re-using open data of different types such as co-expressed genes in diabetes, significant associations of SNPs with diabetes and pathways, provided biologically relevant outcomes. The purpose was to explore in more detail the molecular interactions of the ANGPTL8 gene with other genes and proteins. The results confirm that re-using data with a different purpose than the one for

which the original studies were designed, is possible. Indeed, we were able to identify what types of relationships exist between ANGPTL8 variants and T2D phenotypes in our populations, which can be validated in follow-up laboratories studies.