CHAPTER 7

Conclusions and outlook.

7.1. Identifying candidate functional variants using maps of open chromatin

The biological interpretation of non-protein coding sequence variation associated with complex traits is challenging (Cooper & Shendure, 2011). It has been suggested that some of these non-coding variants influence phenotypic variation through regulation of gene expression.

Open chromatin assays provide an efficient and powerful tool for the identification of regulatory elements, as sites of open chromatin are general indicators of regulatory protein binding (**Section 1.7**). In recent studies, ~10% of the genome has been annotated as sites of open chromatin by DNase I- and/or FAIRE-seq across human cell types (Pennisi, 2011; Song et al., 2011). Despite the strong cross-validation of DNase I- and FAIRE-seq, NDRs specific to either assay are biologically relevant and functional, whereby DNase I- and FAIRE-specific sites tend to occur at TSSs and distal regions, respectively (Song et al., 2011). These differences may be the result of specific regulatory complexes that are bound in each NDR and influence the ability of formaldehyde to cross-link, or DNase I to cut.

However, open chromatin assays can neither directly reveal the function of the identified NDRs nor the transcription factors that are bound to them. Therefore, additional annotation data sets are required to corroborate the presence of a regulatory element and functionally classify it, e.g. enhancer vs. promoter, and type of transcription factor binding site. Here, *in silico* predictions may guide the identification of the specific transcription factor involved. In parallel, a multitude of high-resolution genome-wide annotation data sets across human cell types, developmental and disease states, as well as environmental conditions, will soon become publicly available through the efforts of large consortia including ENCODE and BLUEPRINT. These data sets can be used to further characterise regulatory elements that may be causally linked to phenotypic variation.

Open chromatin and ChIP protocols necessitate a large population of cells. In this thesis, at least 10 million cells were applied in each FAIRE assay (**Section 2.3**). Thus, the obtained regulatory maps have to be considered as an average of the chromatin status across a population of cells that may be heterogeneous. Often only a limited number of cells of a particular tissue or from a developmental stage are available. Protocols that require only a few thousand cells but still can be scaled to the whole genome are under development, with the aim of eventually probing single cells (Kalisky & Quake, 2011).

In **Chapter 3**, I intersected open chromatin maps in a megakaryocytic and an erythroblastoid cell line with CAD/MI risk alleles, but observed no overlap. One reason could be that the studied cell lines are not appropriate for studying the disease aetiology of that particular phenotype. Indeed, vascular smooth muscle cells, cardiomyocytes or other cell types may be more relevant. However, these cell types are difficult to obtain in adequate numbers from humans (especially from healthy individuals) for functional studies. Advances in induced pluripotent stem (iPS) cell technology may provide an attractive solution. This approach requires the heterologous overexpression of a few key transcription factors in mature cells for a period of few weeks. The mature cells, e.g. adult fibroblasts or keratinocytes, can be easily obtained through a skin biopsy and are then returned to an embryonic stem cell-like pluripotent state (de Souza, 2010). Even though differences compared to embryonic cells exist (Chin et al., 2009; Doi et al., 2009), iPS cell lines have the potential to differentiate into various different cell types. Therefore, this technology may hold great potential to generate essentially any cell type at various stages of cellular differentiation in large enough quantities to be used in experimental assays.

A key challenge in GWA follow-up studies is to link the putative regulatory element to a target gene or transcript. Candidate functional variants located within regulatory elements can be efficiently associated with transcript levels in eQTL studies. However, this method does not prove causality of the association. In contrast, allele-specific expression analyses overcome this limitation by directly linking risk alleles with transcript abundance, as shown at the *ZPBP2-GSDMB-ORMDL3* asthma and autoimmune disease risk locus (Verlaan et al., 2009). In GWA studies, the closest gene is typically reported as the most likely candidate. However, this assumption is rarely supported by experimental data in the absence of an eQTL, and many counter-examples in the literature exist (Spilianakis et al., 2005). Chromosome conformation capture techniques, i.e. 3C, 4C, 5C and Hi-C, offer an elegant way to link *cis*-regulatory elements with promoter regions of target genes (van Steensel & Dekker, 2010).

Low-throughput functional assays, e.g. EMSAs and luciferase reporter assays, are still the method-of-choice for establishing functionality of regulatory elements and variants, and studying conclusively their mechanisms (**Sections 5.3** and **6.3**). Following these *in vitro* experimental validation studies, regulatory elements may also be tested using *in vivo* assays. For example, the activity of tissue-specific enhancer sequences can be assessed in transgenic mouse assays (Visel et al., 2009).

Low-frequency and rare variants that impact phenotypic variation may reside on the same haplotype as the GWA tag SNP (**Section 1.1**). Therefore, the intersection of open chromatin or other regulatory maps with sequence variation depends on the availability of complete sequence information in order to make an informative assessment and reach a definite conclusion about causality. The sequence

catalogue provided by the 1000 Genomes Project has made targeted resequencing for common variants (e.g. as described in **Section 5.2**) virtually obsolete for many populations. Initiatives such as the UK10K Project (http://www.uk10k.org/) expand this catalogue of sequence variation down to 0.1% allele frequency in individuals with European ancestry. Importantly, low-frequency and rare variants tend to be population-specific. That is, if associated with complex traits, these variants may have different effects in the different ethnic groups (Gravel et al., 2011; Bustamante et al., 2011). These population-specific effects are due to differences in allele frequency of the genetic markers in different populations.

7.2. Translating candidate functional variants using gene regulatory network approaches

Functional sequence variation may impact complex traits and diseases through the perturbations they cause to transcriptional and other molecular, cellular, tissue and organism network states (Sieberts & Schadt, 2007; Schadt, 2009). As DNA is transcribed into RNA and RNA is subsequently translated into protein, the molecular effects of DNA sequence variation on complex physiologic processes are mediated by transcriptional networks. These networks can be studied through integrative analyses of sequence variation, and cell type- and tissue-specific transcriptional and phenotypic data. This information benefits our understanding of the molecular mechanisms that drive complex trait variation and disease.

Exemplified by ENCODE, integration of multidimensional annotation data sets of the regulatory non-coding portion of the genome into a unified and quantitative framework can also help to improve predictions of genome function (**Figure 7-1**).

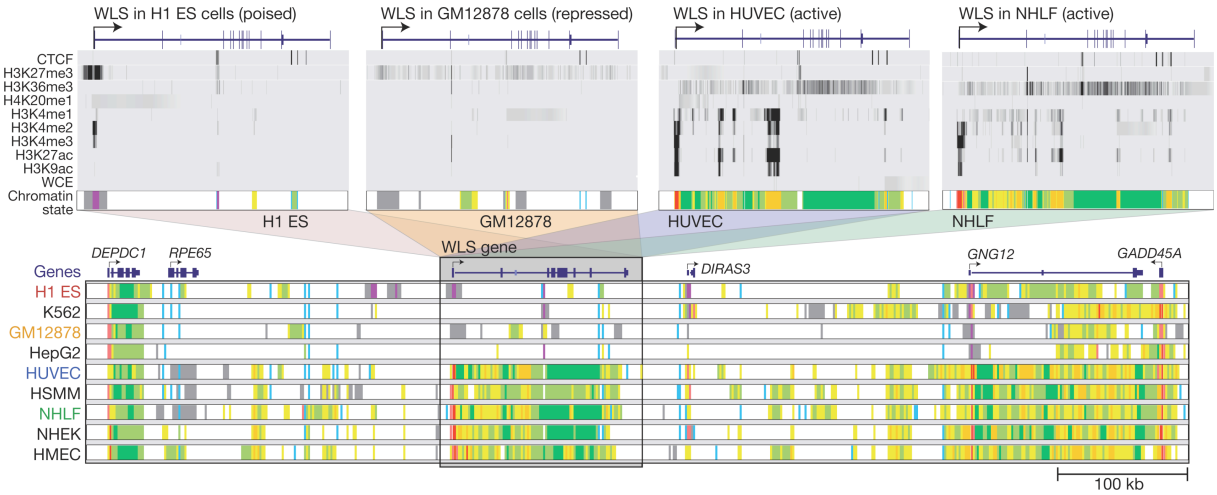

Figure 7-1. Example of the systematic genome annotation using chromatin maps across different cell types. Top: Annotation of the *WLS* gene locus using nine chromatin marks in four cell types. The complex chromatin status (grey scale) is summarised into functional modules and represented as a single, coloured annotation track. For example, red and orange indicate active promoters and strong enhancers, respectively. Bottom: Dynamic chromatin annotation of a 900 kb region centred on the *WLS* locus, showing activation and repression patterns for six genes with hundreds of gene regulatory elements. Figure taken from Ernst et al., 2011.

The characterisation of regulatory elements using chromatin maps contributes to the definition of the nuclear-based phenotype of the cell. In addition, nuclear-based phenotypes comprise other molecular interactions that occur on the chromatin level, such as transcript abundance and protein binding. The synergistic effect of nuclear-based phenotypes is expressed as cytoplasmic phenotypes. For example, transcript abundance results in protein abundance, but this depends on post-translational modifications. Other examples include metabolites or side-products that result from signalling or biochemical cascades (Dermitzakis, 2012).

The key aim of such integrative analysis is to identify relevant genes and effector cell types, and ultimately gain knowledge about pathways pertinent to the complex trait or disease of interest (**Figure 7-2**). This may benefit identification of possible drug targets or classification of the complex trait into sub-phenotypes, potentially resulting in strategies for disease diagnosis, prevention and therapy.

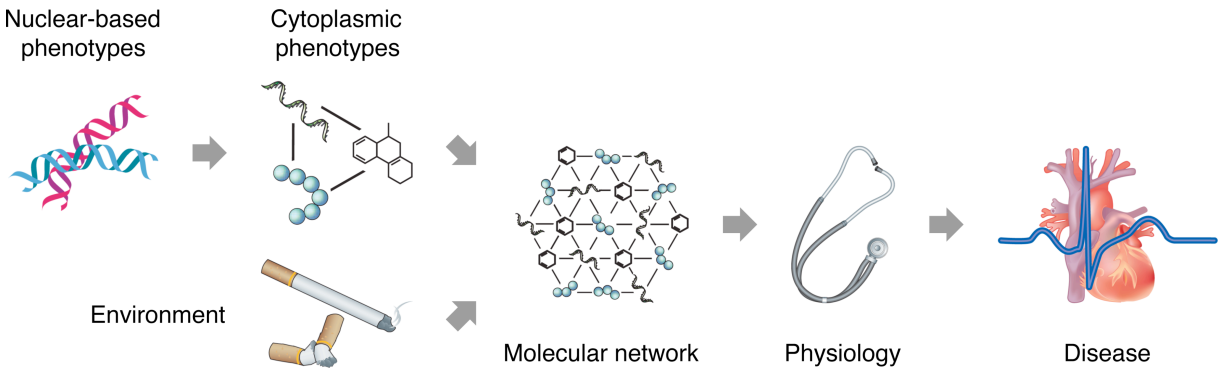

Figure 7-2. Integrative genomics and molecular networks. Molecular networks informed by nuclear-based and cytoplasmic phenotypes, such as DNA variation and protein abundance, define physiological states of human disease. Figure adapted from Schadt, 2009.

PIK3CG was identified as potential target gene underlying the association with platelet volume and function (**Chapter 5**). Indeed, platelets respond to a range of G-protein-coupled receptor agonists that activate PI3K signalling, including thrombin, thromboxane and ADP. Therefore, PI3Kγ may represent an attractive target in antiplatelet therapies (Rückle et al., 2006; Michelson, 2010). Our protein-protein interaction network, reported in **Section 5.6**, highlighted additional signalling pathways that may be investigated with respect to platelet characteristics and function.

A recent analysis investigated how many of the genes at GWA loci were amendable to pharmacological modulation using small molecules or biopharmaceuticals (i.e. therapeutic antibodies or protein therapeutics). The authors found that these genes were significantly more likely to be potentially 'druggable' and 'biopharmable' compared to the entire genome. The study also indicated opportunities for drug-repositioning (Sanseau et al., 2012).

Another route to translating disease-related genetic variants into patient benefits involves the identification of diagnostic biomarkers to inform disease processes. While variants in GWA studies generally have small effects and are therefore not directly suitable for diagnostics, variants identified in exome studies may be of interest. Indeed, the discovery of the genetic basis of TAR, described in **Chapter 6**, will make it simpler to more accurately diagnose future cases with a DNA test. In fact, such a test is currently being developed for the NHS as part of the international ThromboGenomics initiative (https://haemgen.haem.cam.ac.uk/thrombogenomics/). This diagnostic platform may lead to improvements in genetic counselling and patient care.