Review article

Navigating the non-coding genome in heart development and Congenital Heart Disease

Gulrez Chahal1,2, Sonika Tyagi1,3,4,*, Mirana Ramialison1,2,*

1 Australian Regenerative Medicine Institute (ARMI), 15 Innovation Walk, Monash University, Wellington Road, Clayton, 3800, VIC, Australia
2 School of Biological Sciences, Monash University, Wellington Road, Clayton, 3800, VIC, Australia
3 Systems Biology Institute (SBI), Wellington Road, Clayton, 3800, VIC, Australia
4 Australian Genome Research Facility, 305 Grattan Street, Melbourne, VIC, 3000, Australia

A R T I C L E   I N F O

Keywords: Epigenome Non-coding genome Regulatory elements High-throughput sequencing Congenital heart disease Heart development

A B S T R A C T

Congenital Heart Disease (CHD) is characterised by a wide range of cardiac defects, from mild to life-threatening, which occur in babies worldwide. To date, there is no cure to CHD, however, progress in surgery has reduced its mortality allowing children affected by CHD to reach adulthood. In an effort to understand its genetic basis, several studies involving whole-genome sequencing (WGS) of patients with CHD have been undertaken and generated a great wealth of information. The majority of putative causative mutations identified in WGS studies fall into the non-coding part of the genome. Unfortunately, due to the lack of understanding of the function of these non-coding mutations, it is challenging to establish a causal link between the non-coding mutation and the disease. Thus, here we review the state-of-the-art approaches to interpret non-coding mutations in the context of CHD and address the following questions:

What are the non-coding sequences important for cardiac function? Which technologies are used to identify them? Which resources are available to analyse them? What mutations are expected in these non-coding sequences? Learning from developmental process, what is their expected role in CHD?

1. Introduction

Congenital heart disease (CHD) refers to the group of abnormalities observed in the heart which disrupt its normal function and development. It is one of the most common birth defects observed in newborns in various parts of the world (Tennant et al., 2010), (Mitchell et al., 1971). The incidence of CHD-related defects, albeit variable, remains high in densely populated countries with high fertility rate, increasing the need of efficient diagnosis and clinical management (Hoffman, 2013), (van der Linde et al., 2011). CHD has two main causes, environmental and/or genetic (Vecoli et al., 2014).

Environmental exposures during fetal heart development, specifically fetal placental-maternal interaction between weeks 2 through 7 of gestation, play a crucial role in predicting CHD occurrence (Vecoli et al., 2014), (Jenkins et al., 2007). These exposures include teratogens (such as pesticides), and maternal therapeutics (such as barbituric acid or chemotherapeutic agents). Smoking, alcohol consumption, and gestational hypoxia, have also been associated with risk of CHD (Vecoli et al., 2014), (Jenkins et al., 2007), (Shi et al., 2016), (Patel and Burns, 2013).

Decades of extensive research on the molecular control of cardiac development has led to a deeper understanding of the genetic basis of CHD. In particular, the ability to generate in vivo and in vitro genetic models to explore cardiac phenotypes, enabled us to identify genes and developmental pathways essential for cardiogenesis and CHD pathogenesis (Moon, 2008). Several studies have established links between genetic variations in cardiac genes and CHD. Single gene variations in cardiac developmental genes were identified as a cause for about 3–5% of the disease. More recent evidence points out to the polygenic (or complex) nature of CHD (Blue et al., 2012).

Whether monogenic or complex, CHD occurs in sporadic as well as familial forms. Massively parallel sequencing technologies have indicated that de novo mutations contribute to ~10% of its sporadic form (Zaidi et al., 2013). These include deleterious mutations such as nonsense, frameshifts and splicing variants. Additionally, genetic screening...
of patients with familial CHD using gene panels identified variants in ~ 31–46% of the cohorts, while the cause remained unknown for the rest of the cohort (Blue et al., 2017), (Jia et al., 2015), (Blue et al., 2014). Furthermore, CHD may occur in association with extra-cardiac phenotypes, such as Down syndrome. Chromosomal anomalies associated with these syndromes contribute to 8–10% of the CHD cases (Blue et al., 2012).

Despite the large body of information available (McDermott et al., 2005), (Zahavich et al., 2017), (Granados-Riveron et al., 2012), the number of known genetic factors contributing to CHD remains limited. For instance, genetic screening platforms dedicated to diagnose CHD contain less than 100 genes (Cowan and Ware, 2015), and the majority of patients with familial CHD still lack a genetic diagnosis.

While our understanding of the etiology of CHD has slowly progressed, CHD management has drastically improved. Indeed, with advances in surgical procedures, prenatal and neonatal diagnosis, and improved guidelines for management of the disease, mortality rate has decreased (Khoshnood et al., 2005), (Baumgartner et al., 2010). Consequently, patients treated with surgery and carrying gene mutations linked with CHD may transmit the mutation to their offspring through the next generation, leading to an increase in incidence of the disease (Richards, 2010). It is therefore imperative to develop comprehensive genetic tests allowing patients to have access to complete information regarding the risk of CHD when planning a family. Hence, there is a fundamental need to identify genetic factors controlling to heart development in order to provide novel insights in the genetic bases of CHD.

One of the most important reasons for lack of our understanding of the genetic basis of CHD is limiting the focus on protein-coding regions of the genome. Most of the efforts in studying the human genome have been concentrating on screening for alterations in protein-coding genes. However, these compose only ~ 2% of our genome. The greater part of the genome comprises of non-coding regions which harbour functional non-coding DNA regulatory elements such as promoters and enhancers, and genomic regions encoding for small and long non-coding RNAs (ncRNA) (Pertea and Salzberg, 2010).

Mutations in the non-coding regions of the genome have been associated with a wide range of diseases such as cancer and various congenital anomalies (Scacheri and Scacheri, 2015), (Hornshoj et al., 2018). However, their role in CHD has been largely unexplored. Hence, studying non-coding mutations in the context of heart development is a very effective approach to identify novel genetic causes for CHD. This review focuses on our current knowledge of the cardiac non-coding genome and techniques and resources that have been recently made available to explore its role in CHD.

2. cis-Regulatory elements (cREs) in heart development and how to find them

2.1. Definition of cis-regulatory elements (cREs)

The machinery involved in spatio-temporal regulation of gene expression includes transcription factors (TFs) binding DNA at sequence-specific binding sites within cis-regulatory elements (cREs) (Maston et al., 2006), (Narlikar and Ovcharenko, 2009). Thus, cREs such as promoters, enhancers, silencers and insulators are stretches of open chromatin regions accessible to TFs to regulate the expression of their target genes. The binding of TFs to their respective transcription factor binding sites (TFBS) may occur in a cell type- or tissue-specific manner (Narlikar and Ovcharenko, 2009), (Spitz and Furlong, 2012). The composition of TFBS within the regulatory elements (also called cis-regulatory module) contributes to the tight spatio-temporal control of gene expression (Weingarten-Gabby and Segal, 2014), (Zinzen et al., 2009).

There are a limited number of studies which explore the role of cREs in CHD etiology. Since CHD is characterised by impaired heart formation, there is an urgent need to decipher the cis-regulatory mechanisms vital for cardiogenesis in order to identify novel cis-regulatory pathways contributing to CHD. In a bid to chart the regulatory blueprint of the developing heart, several studies have been undertaken in cardiac cells at different cardiac developmental stages as well as the whole heart (Martinez et al., 2015), (Freire-Pritchett et al., 2017), (Wamstad et al., 2014).

In this section we discuss the current principles, technologies and approaches to identify cREs in heart development, based on their characteristic features as TF binding sequences and regions of open chromatin. Furthermore, we also discuss the resources made available in recent years to explore their implications in CHD.

2.2. Detection of cis-regulatory elements (cREs) as targets of cardiac transcription factors (TFs)

Unlike protein-coding genes, to date there is no known “regulatory code” allowing direct detection of cREs in the genome. Since cREs harbour transcription factor binding sites, identifying cardiac-specific TFs and designing experiments to recognise their target regions genome-wide provides the first step to discover cardiac-specific cREs.

2.2.1. cREs discovered from known cardiac TFs

Multiple transcription factor networks have been identified in cardiac tissues, which when altered, govern the development and etiology of heart disorders, including CHD. Table 3 presents a list of known TFs with a key role in heart development and CHD, collated from different studies (McCulley and Black, 2012), (Herrmann et al., 2012), (May et al., 2011). Notably, Nkx2-5, GATA4, and TBX5 have been extensively studied in heart development as well as in CHD. The Nkx2-5 gene, encoding a protein from the homeodomain family, has been demonstrated to have a central role throughout heart development, from proliferation of cardiac precursors to defining chamber identity in murine models (Prall et al., 2007). For these various processes, it interacts with several developmental regulators such as Isl1, Bmp2, Smad1, Irx4 and Me2c; indicating the complexity of transcriptional control in the heart (Prall et al., 2007), (Liberatore et al., 2002), (Vincenz et al., 2008), (Barth et al., 2010), (Schwartz and Olson, 1999). GATA4, encodes for a zinc finger protein which is involved in cardiac morphogenesis and proliferation of cardiomyocytes. It also interacts with several signalling pathway components including cyclin D2 and Cdk4, Tbx5, and MEF2 proteins (McCulley and Black, 2012), (Morin et al., 2000). Another critical and well explored TF, TBX5, is involved in cardiac morphogenesis and development, and has been commonly associated with cardiac defects occurring in Holt-Oram syndrome (Bruneau et al., 2001).

Genome-wide mapping assays can be performed to identify the target of these known cardiac TFs, thus indirectly identifying cardiac cREs. The most widely used technique to identify the binding sites of TFs is Chromatin immunoprecipitation (ChiP) (Table 1). Since the TF machinery is recruited at these accessible DNA sequences, an antibody specific to the TF of interest can be used to identify the bound DNA regions within the genome. These binding sites are deciphered using high-throughput sequencing (ChIP-seq) (Raha et al., 2010). In ChIP-seq, the DNA sequences obtained as “reads” are filtered based on various quality parameters and mapped to a reference genome. The enriched TF-specific precipitated regions are compared with the regions obtained by fragmenting DNA (or input DNA) to identify differential peaks representing binding regions for these TFs. In addition, an enzyme technology-based method, DamID (DNA adenine methyltransferase identification), has also been used to identify TF binding sites in the cardiac context (van Steensel and Henikoff, 2000), (Bouveret et al., 2015), (Ramaiison et al., 2017). *Escherichia coli* DNA adenine methyltransferase (Dam) is fused with a TF of interest. This Dam fusion binds to DNA and locally methylates the surrounding adenine, thereby capturing its target region (Bouveret et al., 2015), (Ramaiison et al., 2017).
These genome-wide mapping assays have been used to detect cREs upstream and downstream of cardiac genes (He et al., 2011, 2014), that are crucial for cardiac development. For instance, regulatory elements guiding spatial expression of Nkx2-5 in the cardiac crescent have been identified in mouse models (Reecy et al., 1999). Examples of cardiac cREs mutated in disease have also emerged such as the identification of mutation in an enhancer region in proximity of the TBX5 gene associated with the Holt–Oram syndrome (Smeno et al., 2012), (Brunel et al., 2001). Another study reported an interaction between the enhancer of Snr10A, a gene involved in Brugada syndrome in humans, with the promoter in Snr5A, to govern the expression of Snr5A in mice (van den Boogaard et al., 2014).

Furthermore, these genome-wide mapping assays revealed generic properties of cardiac transcription networks. For instance, ChIP-seq experiments with antibodies specific to Gata4, MeF2a, Nkx2-5 and Srf performed in mouse cardiomyocytes investigated the TFBSs lying 2.5 kb–10 kb upstream of the transcription start sites (Schlesinger et al., 2011). They revealed the combinatorial nature of cardiac TFs binding at their enhancer target sequences. The cardiac TF GATA4 in particular is shown to interact with histone acetyltransferase p300 and aids its recruitment to the chromatin (He et al., 2014). DamID has been used to identify cREs that are targets of NXX2-5, and novel binding sites bound by NXX2-5 mutants associated with CHD (Bouveret et al., 2015; Ramalison et al., 2017). These studies revealed that mutant TFs target a very distinct set of novel cREs, that are not the endogenous targets of wild-type TFs.

From this great wealth of information on cardiac TF binding, machine learning approaches can be used to infer whether there is a specific combinatorial TF signature that correlates with gene expression or cell-type specificity (Wang et al., 2018). One of the earliest applications of this technology used Support Vector machines to train on Drosophila melanogaster cardiac enhancers with known TFBS composition (obtained from ChIP datasets) and known activity (obtained from in situ hybridization on neighbouring genes). This training phase identified precise TFBS combinations that, not only were shared by enhancers driving similar gene expression, but also could be used to predict gene expression driven by an unknown enhancer (Zinzen et al., 2009). Further development of methods based on artificial intelligence combined with the accumulation of sequencing information will fast-forward our understanding of the regulatory code.

### 2.2.2. New cardiac TF prediction

Out of the ∼3000 TFs predicted in human genome (Brivanlou and Darnell, 2002), to date only 33 are known to play a key role in heart development and disease (Table 3). Hence, there are likely to be many more TFs having a role in the heart yet to be elucidated. Identifying these TFs and their genomic targets will greatly improve the detection of novel cREs essential for the regulation of cardiac development. Here we describe two methods to identify novel TFs.

#### 2.2.2.1. New cardiac TF prediction by database mining

The restricted expression of a TF in a tissue or organ-specific manner could indicate that this TF has a role in defining the identity of that particular tissue or organ. Following this rationale, traditionally TFs were selected to be studied in the context of heart development if they displayed cardiac-specific expression (Schott et al., 1998), (Garg et al., 2003). Nowadays, identification of novel cardiac TFs based on the same principle can be facilitated by database mining. Indeed, several databases exist that contain comprehensive spatio-temporal transcriptional information from different stages of heart development in mouse, human and pluripotent stem cells (Liu et al., 2008), (Wu et al., 2009), (Papatheodorou et al., 2018) (Table 2). Identification of novel cardiac-restricted genes can be achieved by performing differential gene expression analysis, not only from wild-type or healthy heart samples, but also from diseased cardiac datasets such as coronary artery disease, congestive heart failure and cardiomyopathy (Wu et al., 2009).

### Table 1
Current high-throughput technologies, their principle and genomic information they capture as outcome.

<table>
<thead>
<tr>
<th>Sequencing technology</th>
<th>Methodology</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChIP-seq</td>
<td>Sequencing of specific isolated protein-DNA interaction. This protein can be transcription factors, co-factors or histones.</td>
<td>Putative enhancers, TFBS, respective histone marks</td>
</tr>
<tr>
<td>RNA-seq</td>
<td>Evaluates mRNA expressions for a given sample</td>
<td>Expression pattern of translated protein</td>
</tr>
<tr>
<td>ATAC-seq</td>
<td>Identifies open chromatin regions in the genome using hyperactive Tn5 transposase which inserts adapter sequences in the accessible regions.</td>
<td>TF binding sites, regulatory elements</td>
</tr>
<tr>
<td>DNase-seq</td>
<td>Identifies loosely packed DNase I hypersensitive sites in the chromatin by subjecting them to DNase enzyme digestion. These sites in the genome have been correlated to frequently contain regulatory elements and can be sequenced genome wide.</td>
<td>Regulatory elements, nucleosomes, DNA-binding proteins</td>
</tr>
<tr>
<td>MNase-seq</td>
<td>Identifies nucleosome positioning using micrococcal nuclease digestion followed by high throughput sequencing</td>
<td>DNase hypersensitive sites, transcriptional start sites, and active promoters</td>
</tr>
<tr>
<td>FAIRE-seq</td>
<td>Identifies nucleosome-depleted regions by using formaldehyde to cross-link the nucleosome-rich regions. The nucleosome-depleted regions can be separated by further chemical treatment and fluorescent labelling, followed by sequencing</td>
<td></td>
</tr>
</tbody>
</table>

### Table 2
Databases storing information on non-coding elements.

<table>
<thead>
<tr>
<th>Database Type</th>
<th>Transcription Factor Binding Site (TFBS)</th>
<th>mRNA</th>
<th>Non-coding variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensembl Regiation</td>
<td>Ensembl Regiation</td>
<td>RNAcentral</td>
<td>GWAS Central</td>
</tr>
<tr>
<td>FANTOMS</td>
<td>JASPAR</td>
<td>LncBase</td>
<td>NHGRI-EBI GWAS catalogue</td>
</tr>
<tr>
<td>VISTA Enhancer Browser</td>
<td>TRANSFAC</td>
<td>LNCipedia</td>
<td>GWASdb</td>
</tr>
<tr>
<td>TRANSFAC</td>
<td>HOCOMOCO</td>
<td>LncRNAWiki</td>
<td>ExAC</td>
</tr>
<tr>
<td>Expression Atlas</td>
<td>International Human Epigenome Consortium (IHEC)</td>
<td>miRBase</td>
<td></td>
</tr>
<tr>
<td>ENCODE</td>
<td></td>
<td>miRTarBase</td>
<td></td>
</tr>
<tr>
<td>NIH Roadmap Epigenomics Mapping Consortium</td>
<td></td>
<td>circNet</td>
<td></td>
</tr>
<tr>
<td>International Human Epigenome Consortium (IHEC)</td>
<td></td>
<td>PolyMiTS</td>
<td></td>
</tr>
<tr>
<td>ExAC</td>
<td></td>
<td></td>
<td>FANTOMS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Expression Atlas</td>
</tr>
</tbody>
</table>
Table 3
Key transcription factors specific to heart development and CHD, and the corresponding model organism for which DNA binding site information is available in JASPAR.

<table>
<thead>
<tr>
<th>Transcription Factors (human homologue)</th>
<th>Organism in JASPAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANKRD1</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>CITED2</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>ETS1</td>
<td>Homo sapiens, Mus musculus</td>
</tr>
<tr>
<td>FOG2/2FFM2</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>FOXC1</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>FOXG2</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>FOXH1</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>GATA1</td>
<td>Homo sapiens, Mus musculus, Arabidopsis thaliana</td>
</tr>
<tr>
<td>GATA4</td>
<td>Mus musculus</td>
</tr>
<tr>
<td>GATA6</td>
<td>Homo sapiens, Arabidopsis thaliana, Mus musculus</td>
</tr>
<tr>
<td>HAND1</td>
<td>Mus musculus</td>
</tr>
<tr>
<td>HAND2</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>HOXA1</td>
<td>Homo sapiens, Mus musculus</td>
</tr>
<tr>
<td>HNF1A</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>JAG1</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>MEF2C</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>MESP1</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>MESP2</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>MYOCD</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>NFATC1</td>
<td>Melogis galloprovo</td>
</tr>
<tr>
<td>NKX2-5</td>
<td>Mus musculus, Caenorhabditis elegans</td>
</tr>
<tr>
<td>NKX2-6</td>
<td>Mus musculus</td>
</tr>
<tr>
<td>PITX2</td>
<td>Mus musculus</td>
</tr>
<tr>
<td>PPARG</td>
<td>Homo sapiens, Mus musculus</td>
</tr>
<tr>
<td>SALL4</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>TBX1</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>TBX2</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>TBX20</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>TBX5</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>TFAP2B</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>ZBED2</td>
<td>Homo sapiens, Mus musculus</td>
</tr>
<tr>
<td>ZIC3</td>
<td>Homo sapiens, Mus musculus</td>
</tr>
</tbody>
</table>

For instance, Expression Atlas contains quality-controlled curated microarray and RNA-seq experiments from ArrayExpress or GEO (Kolesnikov et al., 2015). It allows baseline and differential analysis to identify gene products expressed in heart, and those that are differentially expressed in disease conditions. A meta-analysis study utilising 700 different mouse experiments downloaded from this resource predicted novel candidate genes with putative role in heart failure (Akavia and Benayahu, 2008). In summary, several databases are available to identify novel TFs by relying on their specificity of expression in the heart. However, this approach is limited as there is evidence that TFs that are broadly expressed can also play a specific role in heart (e.g. Elk TFs (Bouveret et al., 2015)). In addition, these databases may provide reliable in silico evidence of the role of new TFs in heart development, but this remains to be experimentally validated.

2.2.2.2. New cardiac TF prediction by motif discovery on targets of known TFs. Cardiac TFs bind to their target cREs in complexes, which include other transcriptional co-factors. Therefore, cREs contain multiple TF binding sites (TFBS) corresponding to the binding of the TFs in the complex. There are known TFs that are part of the regulatory complex, for example GATA4, TBX5, NKX2-5 and SRF have been shown to co-occupy the same cREs in HL1-cardiomyocytes (He et al., 2014). However, the regulatory complexes may contain other TFs that are not known to have a cardiac function. Thus, it is expected that the cREs targeted by these complexes will also contain the binding site of these novel TFs. With arduous experimental methods alone it is nearly impossible to identify these novel TFBS. Therefore, computational methods have been designed to predict these novel TFBS in the genome. By using in silico methods, it is possible to identify these novel TFBS, and predict novel TFs involved in heart development and disease that are binding to these novel TFBSs. Here we present methods to predict new TFs involved in heart development from predicted novel TFBS in cardiac cREs.

In silico methods to predict cREs in the genome initially relied on searching for DNA motifs with specific nucleotide composition (e.g. TATA-box like sequences and CpG dinucleotide abundance), usually in proximity to the transcription start site (TSS). However, this lead to a large number of false positives due to the abundance of such sequences with these simple signatures in the genome (Fickett and Hatzigeorgiou, 1997). Phylogenetic footprinting based algorithms were then developed to identify putative functional cREs. They are based on the principle that functional regions of the genome accumulate mutations more slowly compared to regions with no function. Thus, comparison of non-coding sequences across genomes may reveal highly conserved segments indicating a role in transcriptional regulation (Lenhard et al., 2003). However, it is non-trivial to identify stretches of highly conserved non-coding regions, in particular, cardiac enhancers have been shown to display less sequence conservation than brain enhancers for example (Blow et al., 2010). More recent methods are based on, de novo motif discovery to predict novel TFBS without any prior knowledge of TFBS composition. (Lenhard et al., 2003), (Ben-Gal et al., 2005), (Sandve and Drablos, 2006), (Narlikar et al., 2010), (Wong, 2017), (Hestand et al., 2008), (Mathelier et al., 2015), (Wei and Yu, 2007), (Dang et al., 2018), (Zhou and Troyanskaya, 2015), (Wang et al., 2018). These algorithms detect short DNA sequences (motifs) that are significantly enriched in a set of cREs (for instance obtained from ChIP-seq datasets), in comparison to random genomic sequences. In a ChIP-seq experiment, it is expected that the binding site of the targeted TF will appear as the highest-ranked predicted motif. Any additional enriched motifs may correspond to the binding sites of novel co-factors of that TF. To determine which TFs bind to these newly identified motifs, databases holding information about known TF binding sites can be interrogated. JASPAR is an open source collection of TF binding sites presented as position weight matrices (PWMs), where the JASPAR core version is a non-redundant collection of experimentally validated mammalian TFBSs (Vlieghe et al., 2006), (khan et al., 2018). TRANSFAC is a commercially available database which covers a collection of experimentally determined TFBSs stored as PWMs (Vlieghe et al., 2006). The Homo sapiens comprehensive model collection (HOCOMOCO) and UniPROBE are other popular PWM-based resource which respectively integrate curated TFBSs from various experiments for over 400 human TFs (Kulakovskiy et al., 2013, 2018) or homebox families of TFs (Newburger and Bulyk, 2009). However, the majority of transcription factors encoded in our genomes still have an unknown DNA binding site, due to the lack of experimentally validated TFBS in different organisms and disease models (Wasserman and Sandelin, 2004). Of the 3000 TFs predicted in the human genome, only a third has been captured in the databases above.

An example of cofactor identification using experimentally derived TF-associated sequence information is a discovery from DamID experiments targeting the NKX2-5 cardiac TF. For mutant NKX2-5 protein lacking the DNA-binding domain, a novel motif was enriched in the associated sequences that was not detectable when using the wild-type protein. Using TFBS databases, this motif was identified as the binding site of the Elk1/4 transcription factors. Hence, these studies revealed the role of Elk1/4 as novel co-factor of Nkx2-5 (Bruneau et al., 2001, Raha et al., 2010). Similar studies, based on de novo motif prediction on enhancers of taste receptor genes, revealed cardiac TFs regulating genes usually expressed in the tongue, to also be expressed in the heart (Foster et al., 2015).

2.3. Detection of cREs using epigenetic marks

Histone modifications are well known post-translational...
modifications which regulate expression of genes by strengthening or weakening the interaction between DNA and histones. They regulate the binding of TFs to target regions based on strength of packaging of the chromatin, with the help of chromatin-modifying enzymes.

Histone modifications include histone (mono-, di- or tri-) methylation, acetylation, phosphorylation, ADP-ribosylation, ubiquitination and sumoylation. Histone H3 methylation (H3K4me3 and H3K27me3), and acetylation (H3K27ac) have been widely explored with respect to regulating gene expression, where the former is involved in both transcriptional activation or repression, the latter is associated with transcriptional activation (Vallaster et al., 2012). Various histone modifications govern the expression of different cardiac genes during early heart development and disease, including CHD (Zhou et al., 2011a), (Wamstad et al., 2012), (Heintzman et al., 2009), (Heintzman et al., 2007). ChIP-seq is a widely used technique to determine the location of histone with specific modifications (or histone marks) in the genome (O’Geen et al., 2011). One of the commonly used methods to identify enhancers is to perform ChIP-seq using antibodies against H3K4me1, H3K4me2, H3K27ac and the histone acetyltransferase p300 (Zhou et al., 2011b). Comparing and overlapping the resulting peaks of ChIP-seq experiments using different antibodies, allows a more accurate detection of cREs.

Another class of epigenetic marks associated with transcriptional regulation is DNA methylation. It involves covalent transfer of a methyl group to the C-5 position of the cytosine ring of DNA by DNA methyltransferases (DNMTs), at CG dinucleotides in the genome, resulting in epigenetic silencing of transcription. The CpG-rich regions within promoters and enhancers, tend to have a differential pattern of DNA methylation when compared to non-regulatory regions of the genome (Zhang et al., 2015), (Jin et al., 2011). Whole-genome bisulfite sequencing is the standard technique to identify differentially methylated cREs. Treatment of DNA with bisulfite mediates the conversion of cytosine into uracil, which is then read as thymine in the sequencing downstream. However, the methylated cytosine-5 are resistant to this conversion and are thus retained as cytosine. Thus, differentially methylated regulatory regions are identified by comparing the bisulfite treated with the untreated DNA sample (Kurdyukov and Bullock, 2016).

These technologies have been widely used in the context of heart development and disease to identify relevant cREs. For instance, Gilshbach et al. provided a comprehensive resource of enhancers active in human cardiomyocytes in various developmental stages based on histone mark locations and differential CpG methylation during prenatal vs postnatal heart development. These enhancers regions were strongly associated with genes involved in cardiomyocyte development, highlighting their functional relevance in cardiogenesis. In addition, investigation of the activity of these enhancers in chronic heart failure revealed that dynamic changes to histone marks, including H3K27ac and H3K36me3, were concordant with disease-relevant gene expression, thereby supporting the role of enhancers in cardiac disease, implicating them as regulators of gene expression in cardiac development and disease (Gilshbach et al., 2018). Another genome-wide studies have expanded our knowledge of cardiac enhancers in adult and fetal human heart using this principle (Wamstad et al., 2014), (Blow et al., 2010), (van Duijvenboden et al., 2014), (May et al., 2011). A comprehensive study across more than 35 epigenetic data sets for pre- and post-natal mouse and human models predicted more than 80,000 putative enhancer regions (Dickel et al., 2016). May D et al. predicted ~6200 candidate enhancers enriched near genes involved in heart development and disorders including CHD, using the results from ChIP-seq of enhancer-associated proteins p300/CFP. The p300/CFP-bound regions were classified based on their vicinity to transcription start sites of heart-specific genes, including KCNQ1, KCNE2, MYL2, ACTN2 and TNNC1, GATA4, NKX2-5 and TBX5 involved in CHD. Enhancer sequences located in the vicinity of embryonic heart developmental genes such as GATA4, GATA6 and MEF2C were experimentally validated in vivo using transgenic mouse model enhancer assay (May et al., 2011).

Finally, there is evidence that de novo mutations in the genes contributing to histone modifications also contribute to CHD patients. Indeed mutations have been observed in genes such as CHD7, KDM5A, KDM5B and RNF20, which affect histone 3 lysine 4 (H3K4) methylation, and ubiquitination of H2BK120, required for H3K4 methylation (Zaidi et al., 2013), (Zhang and Liu, 2015). Another recent study demonstrated elevated catalytic activity of histone deacetylase (HDAC Class I, IIa, and IIb) in children with single ventricle heart disease of their right ventricle (Blakeslee et al., 2017). These few yet relevant studies further strengthen the evidence that histone modifications and associated regulatory pathways play a role in the pathogenesis of CHD.

2.4. Detecting cREs using chromatin accessibility and organisation

Chromatin state governs the accessibility of regulatory elements to TFs. Differential chromatin state (open or close) can indicate the presence and the activity of cREs in the DNA. Various high-throughput techniques that enable to directly profile the chromatin accessibility include DNase-seq, FAIRE-seq, ATAC-seq and MNase-seq. Comparing and overlapping the resulting regions of DNase-seq experiments can identify relevant cREs in the genome. These sites in the genome have been correlated to frequently contain regulatory elements and can be sequenced genome-wide (DNase-ChIP and DNase-seq) (Crawford et al., 2008), (Crawford et al., 2004), (Song and Crawford, 2010). Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq), is another method to identify open chromatin. It involves hyperactive Tn5 transposase which inserts adapter sequences in the accessible regions and is used to identify TF binding regions and nucleosome positions (Buenrostro et al., 2015). These accessible regions are captured as sequences which are then filtered for quality and adapter sequences, and are aligned to the reference genome. Peaks or collection of aligned reads in a genomic region indicate the region that is open, and therefore accessible for DNA-binding proteins such as TFs (Miskimen et al., 2017). The FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) sequencing method uses formaldehyde to crosslink nucleosome rich regions, and involves recovery of the non-nucleosome bound DNA which corresponds to regulatory components of the genome such as DNase hypersensitive sites, transcription start sites and active promoters (Giresi et al., 2007). Finally, MNase-seq is a method to identify nucleosome free regions of the genome, which involves micrococcal nuclease based digestion of nucleosome bound regions in the DNA, followed by sequencing (Rizzato et al., 2012). Of all these methods, ATAC-seq has been more often used to interrogate chromatin accessibility in the cardiac context and has provided a great wealth of information to complete the mapping of cardiac cREs, from bulk tissues or single cells (Yuan et al., 2018), (Jia et al., 2018).

The three-dimensional (3D) organisation of chromatin adds another layer of complexity in identifying cREs and their target genes. Distal enhancers have been shown to physically interact with gene promoters located thousands of base pairs apart. Thus, understanding the 3D organisation of chromatin helps to reveal additional REs by reducing the search space to interacting genomic regions. Chromatin conformation capture technologies (3C) and its derived methods (4C, 5C and HiC) are used to uncover three-dimensional loops of cREs resulting from the interaction between TFs bound to one cRE and their co-factors bound to another cRE. These methods involve chromatin fixation, followed by restriction enzyme-based cutting to produce DNA fragments with sticky ends. These sticky ends are re-ligated to capture long distance fragment interactions (Denker and de Laat, 2016), (de Wit and de Laat, 2012), (Tolhuis et al., 2002). HiCwas used to map chromatin interactions in cardiac cells and tissues (Montefiori et al., 2018), (Schmitt et al., 2016) and has revealed the crucial role of chromatin remodeling in cardiac myocytes and hypertrophy (Rosa-Garrido et al., 2017). While to date there have been not any such studies in the specific context of human CHD, a 4C-seq study in mouse heart has revealed the interaction of a
cardiac enhancer located in Scn10a (encompassing SCN10A variant associated with slow conduction) with the promoter of Scn5a (van den Boogaard et al., 2014).

2.5. Available resources

Numerous online databases and web portals provide experimental datasets and computational analyses that can be applied to identification of cREs in cardiac systems (Table 2). EnsEMBL Regulation at the European Bioinformatics Institute (EBI) provides resources for studying gene expression and its regulation in human and mouse. Their database contains annotations for regulatory features such as promoters, enhancers, binding sites of CTCFs (ubiquitous chromatin structural proteins) and open chromatin regions and transcription start sites (TSSs), supported by ChIP-seq and DNase-seq (Zerbin et al., 2016). EnsEMBL provides the data mining tool BioMart to access and retrieve information from its databases, which can be accessed via web browser or programmatically (e.g. for R Bioconductor packages and other programming languages) through application programming interfaces (APIs) (Smedley et al., 2015). Data from VISTA Enhancer Browser containing TSS and enhancer data have also been implemented (Visel et al., 2007). It does a pairwise alignment of sequences with a sliding window across the sequences to generate the nucleotide identity plot (Wasserman and Sandelin, 2004). A comparison of VISTA-predicted heart-specific enhancers with those identified in human cardiomyocytes in a ChIP-seq study, revealed an overlap of 49% of enhancers. VISTA enhancers that were not identified in the cardiomyocytes were adjacent to genes with a role in artery morphogenesis and organisation of collagen. This overlap increased to 90% by narrowing the VISTA list to confirmed mouse embryonic heart-specific enhancers (Gilsbach et al., 2018). This study demonstrates the modest efficiency of cRE prediction, at the same time highlights the need for more experimentally validated datasets to improve its efficiency.

One of the breakthroughs in exploring the human epigenome has been the Encyclopedia of DNA Elements (ENCODE) project aimed at capturing genome wide functional elements (Table 2). This is a comprehensive catalogue of non-coding functional elements including TFBSs, CTCF-occupied sites, non-coding RNAs and also loci associated with a dozen of histone modifications. The project quantified that 80.4% of the genome is involved in at least one biochemical functional event in at least a single cell type, highlighting our limited understanding of the roles of non-coding parts of the genome (Consortium, 2012). The NIH Roadmap Epigenomics Mapping Consortium is an additional project aimed at mapping functional epigenomic elements from publicly accessible, tissue-specific sequencing data (such as ENCODE) (Romanski et al., 2015). As per ENCODE Portal (2007-Present), out of 340 heart-related datasets from human and mouse in the data matrix, 258 from ENCODE and 82 from NIH Roadmap are currently available. Another major effort undertaken by the International Human Epigenome Consortium (IHEC) is the generation of more than 1000 epigenomes for normal vs diseased human cell types. In addition, ENCODE and NIH Roadmap data are accessible via the IHEC portal (Bae, 2013), (Albrecht et al., 2016). These resources integrate genome-wide experimental data to improve the current knowledge of non-coding elements including cREs. However, it is important to keep in mind that these databases annotate data over an entire cell population which might be non-homogenous, and may result in a biased genetic or physiological representation (A user’s guide to the encyclopedia, 2011). Furthermore, developing improved user-friendly analyses and visualization tools will facilitate the wider use of these comprehensive resources.

2.6. Functional validation of cREs

Application of the above discussed techniques and data resources has uncovered several putative cREs which raises the need for functional validation to confirm which of these are truly functional in cells or tissues. Classical reporter assays such as fluorescence in situ hybridization (FISH) and luciferase assays have been used to study the enhancer activity. In luciferase assays, the enhancer is introduced into a reporter construct with a minimal promoter to activate the transcription of the reporter gene (luciferase), which can be quantified based on light signal capture. Additionally, transgenic reporter assays have been used to test cell-specific enhancers in vivo (Kwon, 2015). However, they can test only one individual candidate at a time making it arduous to validate multiple candidates, spurring the development of more high-throughput techniques such as massively parallel reporter assays (MPRA) for enhancer testing. Self-Transcribing Active Regulatory Region sequencing (STARR-seq) is one such technique where candidate enhancers are cloned downstream of a core promoter and within the 3′ UTR of a reporter gene. These candidates will be transcribed as a part of the resulting reporter transcript when active and can be deep sequenced (Muerdter et al., 2015). This technique has been used to generate a Drosophila genome-wide quantified enhancer map (Arnold et al., 2013). However, STARR-seq might not be a preferred technique in mammalian genomes due to their large size. Functional Identification of Regulatory Elements Within Accessible Chromatin (FIREWACH) another high-throughput functional assay has been used to evaluate over 80,000 DNA fragments simultaneously in mammalians cells (Babbitt et al., 2015). This technique is able to handle large genomes by focusing the analyses to the most relevant 2% of the accessible chromatin within the cell. These MPRA techniques have yet to be implemented in cardiac cells to identify heart-relevant cREs on a genomic scale. Thus, despite all these significant milestones in annotating the non-coding part of the genome, key non-coding elements that are functionally proven to be implicated in CHD etiology still remain unknown.

3. Role of non-coding RNAs in cardiac gene regulation

3.1. Definition

Non-coding RNAs are functional RNAs that are transcribed yet not translated into protein, and are involved in the regulation of gene-expression at both transcriptional and post-transcriptional level. Non-coding RNAs (ncRNAs) can be broadly categorised into three groups based on their size, i.e. small ncRNAs (∼16–31 nucleotides), medium size ncRNA (20–200 nucleotides) and long non-coding RNAs (lncRNA > 200 nucleotides) (Esteller, 2011).

3.2. miRNAs

microRNAs (miRNAs) are the best studied class of small ncRNAs that have been shown to be involved in various cardiovascular disorders (Papageorgiou et al., 2012), (Schulte and Zeller, 2015), (Wang et al., 2016a) and specifically in CHD (Smith et al., 2015), (Xie et al., 2013), (Xie et al., 2016). miRNAs generally downregulate their target mRNA by inhibiting translation or degrading it. A single miRNA commonly affects multiple target genes, therefore changes in expression of a miRNA can perturb an entire gene network resulting in complex pathological states. miRNAs such as mir-1 and mir-133a/b have been identified for their roles in embryonic cardiogenesis (Bostjanic et al., 2015), (Kwon et al., 2005). These small RNAs have presumably multiple effects in development as well as in disease progression (Sayed and Abdellatif, 2011). For instance, high expression of mir-126 has been observed in zebrafish heart and vasculature. Deletion of mir-126 resulted in defective neangiogenesis in mice models when subjected to myocardial infarction (Wang et al., 2008). mirR-1-2 deletion has been associated with the development of ventricular septum defects and increased cardiomyocyte proliferation (Zhao et al., 2007). Also, SNPs have been identified in the 3′ untranslated region (3′UTR) of CHD-relevant genes such as GATA4 which alter miRNA binding sites in CHD patients, indicating their role in the disease development (Sabina et al., 2013).
incRNAs are a class of regulatory RNAs that individually have been implicated in varied roles such as chromatin modification (Gupta et al., 2010), transcriptional (Huarte et al., 2010) and post-transcriptional gene expression regulation and RNA transport (Geisler and Coller, 2013). IncRNAs function by interacting with DNA, RNA and regulatory proteins. Similar to protein-coding genes, IncRNAs could be contributing to tissue-specific response via their role in epigenetic regulation. They appear to be necessary for proper targeting of histone modifying complexes and play a role in DNA methylation and recruitment of regulatory proteins to the cREs of its target genes (Peschansky and Wahlestedt, 2013). Wang et al. referred to this class of epigenetic regulatory IncRNA as Epi-IncRNAs (Wang and Wang, 2015). In recent years, several studies have emerged highlighting the role of IncRNAs in cardiac development (Schonrock et al., 2012). Some examples are Myosin Heavy Chain Associated RNA Transcripts (My-HEART or Mhrt), Cardiac hypertrophy-associated transcript (Chast) (Viereck et al., 2016), Cardiac-hypertrophy-associated epigenetic regulator (Chaeor) (Wang et al., 2016b) and Braveheart (Bvht) (Schmitz et al., 2016), (Klattenhoff et al., 2013). Mhrt have been identified as cardiac-specific IncRNA, with low level of transcripts in fetal hearts, and increased abundance as the heart developed with increased Myh7 expression; and have been shown to protect heart from cardiac hypertrophy in mice (Han et al., 2014). Bvht is involved in activation of cardiac gene regulatory network by regulating the expression of Meq1 gene, and also interacts with protein SUZ12 involved in cardiomyocyte differentiation (Klattenhoff et al., 2013). Two other IncRNAs, Chast and Chaer, have recently been shown to be essential in the development of cardiac hypertrophy (Wang et al., 2016b), (Viereck et al., 2016). The latter directly interacts with chromatin remodeling factor PCR2 to interfere with normal gene expression regulation.

Another class of non-coding RNAs are circRNAs which are studied for their role in cardiovascular diseases (Fan et al., 2017). circRNAs are stable, covalently closed circular RNA molecules, formed initially by pairing of complementary motifs present at the two ends (which have often been observed to be Alu repeats and are expressed widely in human cells) after which the ends of the loop are covalently ligated through a “back-splicing” event. They regulate miRNA target gene expression by competitively binding to miRNAs, reducing their ability to down-regulate the expression of their target genes. Thus, they have been termed microRNA sponges (Fan et al., 2017). Recent studies in various cardiovascular diseases have demonstrated the regulatory role of circRNA. Wang et al. demonstrated the role of Heart-related circRNA (HRCR), which can sequester mir-223, (targeting the ARC gene), as a positive regulator of cardiac hypertrophy (Wang et al., 2016c). Similarly, other circRNAs have been identified in myocardial infarction and cardiac senescence (Geng et al., 2016), (Du et al., 2017). Since circRNAs represent another level of gene regulation and have been associated with various heart diseases, it would be worth investigating whether some circRNAs play a role in CHD.

3.5. Techniques to explore ncRNAs

ncRNAs are often profiled in whole-transcriptome studies by applying high-throughput RNA sequencing technologies to total RNA preparations after targeted depletion of rRNA. Small RNAs in the range of 19–33 bp are isolated with specialized protocols and profiled using high-throughput sequencing, most often with the aim of detecting miRNAs (Hafner et al., 2008), (Lu et al., 2007). However, the results of RNA-seq experiments have to be interpreted carefully while identifying ncRNAs as they are short sequences (<100 bp in length) and more often repetitive elements, which might map to multiple regions in the genome increasing the false positives (Uchida, 2017), (Weirick et al., 2016). Ultra deep sequencing will be required to detect lowly expressed ncRNAs in general.

IncRNA are harder to isolate since they vary both in length and expression level. The IncRNAs are found preferentially proximal (within ~10 kb) to a protein-coding locus and are often expressed in a tissue-specific manner (Cabilia et al., 2011). Heart-specific localisation of IncRNA might also be indicative of its role in cardiac development and disease. Generally, to view the localisation of ncRNAs, in situ hybridisation assay has been commonly used (Soares et al., 2018). For instance, myocardium-specific expression of IncRNA Mhrt was observed using in situ hybridization in mice indicating its role in heart development (Han et al., 2014). In order to identify the targets of IncRNAs, various hybridization techniques such as Capture Hybridization Analysis of RNA Targets (CHART) are applied to detect genome-wide binding locations of IncRNAs. The RNA bound chromatin across the genome is enriched using biotinylated complementary antisense oligonucleotides (22–28 nucleotides) that bind to the accessible target RNA. RNA-chromatin complexes are then immobilised on beads and eluted using RNAse H over multiple washes. The targets of IncRNA in the genome are then determined by sequencing the associated DNA and mapping it genome-wide (Vance, 2017), (Simon, 2013). These hybridization-based methods are used to determine genome-wide associations of known IncRNA.

3.6. Available resources

RNA-related databases such as RNAcentral store datasets of different classes of RNAs, including non-coding RNA (Petrov et al., 2017). The RNAcentral Consortium is formed by 40 expert databases, 23 of which have already been imported into RNAcentral. It has a wide variety of IncRNA sources (IncBase (Paraskevopoulou et al., 2016), LNCipedia (Volders et al., 2013), LncRNAWiki (Ma et al., 2015), LncRNAdb (Quek et al., 2015), GENCODE (Harrow et al., 2012)), miRNAs and their targets (including miRBase (Griffiths-Jones et al., 2006), miRtarBase (Chou et al., 2018)) and snoRNAs (NONCODE (Liu et al., 2005)) datasets. For circRNAs, circNet database contains tissue-specific information on circRNAs and their miRNA partners (Liu et al., 2016). PolymiRTS (Polymorphism in microRNAs and their Target Sites) is a functional analysis tool to explore miRNAs and their target sites in disease contexts, including coronary heart disease and heart failure (Bhattacharya et al., 2014). Finally the FANTOM consortium has built an atlas of various IncRNAs across more than 1800 human cell- and tissue-types including cardiac, based on Cap analysis gene expression (CAGE) data integrated with a collection of transcript models (Iyer et al., 2015).

Although there are increasingly large collections of ncRNA datasets becoming available for various disease and physiological states from human tissues, more specific heart or CHD examples are very few. Because ncRNAs operate at various regulatory levels (chromatin tertiary structure, splicing, transcription, RNA stability and translation) and mechanistic levels (interacting with DNA, protein and/or other RNAs), more integrative approach to find associations between genomic datasets from different experiment types are required to find new ncRNA players in heart and CHD-specific regulatory machinery.
4. Alterations to be expected in non-coding regulatory elements

4.1. Point mutations

4.1.1. Definition

A point mutation is a substitution of a single nucleotide in the DNA sequence with another nucleotide which may or may not have a pathogenic effect. Several point mutations have been observed in both syndromic and non-syndromic CHD. Here, we focus on SNPs that have been identified in non-coding regions.

Genome-wide association studies (GWAS) have been conducted in CHD patients to pinpoint CHD associated loci including point mutations (Cordell et al., 2013a), (Blue et al., 2017). GWAS aims to identify genetic variants that are associated with a phenotype (typically, disease) in large population cohorts, most commonly using single nucleotide polymorphism arrays (SNP arrays). These arrays allow to estimate the frequency of known variants or loci across a diseased genome, in comparison with their frequency of occurrence in healthy controls, in order to identify variants with a statistically significant association with the disease (Visscher et al., 2012). These studies have been widely used to identify SNPs associated with CHD risk (Hu et al., 2013), (Cordell et al., 2013b), (Zhao et al., 2014), (Lin et al., 2015).

Strikingly, ~88% of SNPs identified in GWAS lie in non-coding regions (Edwards et al., 2013), but they are often not pursued for further investigation. This could be on account of linkage disequilibrium observed in families which leads to indirect association of variants with the disease, making it difficult to narrow down the causal variant (Bush and Moore, 2012). Often, to reduce the number of variants under consideration, coding variants have been prioritized because it is easier to predict their immediate biological effect (e.g. producing a truncated protein). Furthermore, even statistically significant association with disease does not resolve the underlying biological mechanism of the variant's pathogenesis, which may be important in the long term for developing new potential therapies.

4.1.2. Relevant studies and techniques

Cordell and colleagues have identified the SNP rs870142 at a non-coding locus on chromosome 4p16, which might be involved in transcriptional regulation, to be associated with an ostium secundum atrial septal defect (ASD) in individuals of European ancestry (Cordell et al., 2013b). The same locus association was replicated in Han Chinese population with ASD by Zhao et al. (2014). In addition, Lin and colleagues identified SNPs upstream of genes including TUBX3 and TUBX5 (SNP rs2433752), SMARCA2 (SNP rs7863990), EDNRA (SNP rs1400558) as CHD susceptibility loci in the Chinese population. These genes are known to have a role in various stages of heart development (Lin et al., 2015). These studies reinforce the need to understand the role of such loci in regulating the transcription of various genes to eventually predict CHD risks in the clinic.

The roles of point mutations in non-coding regions in syndromic disease are still unexplored. This can be attributed to the fact that a large set of alterations known in syndromic CHD constitute chromosomal aberrations where a large proportion of the chromosome is either translocated or inserted/deleted. In this scenario, it becomes very difficult to identify single point mutations (both in intronic and genic regions) that could underlie the pathology. However, genes involved in syndromic form of the disease such as TUBX5 (involved in Holt-Oram syndrome) have been observed to interact with various regulatory elements to control the expression of several other genes (Waldron et al., 2016). A non-coding pathogenic single nucleotide mutation may affect gene regulation by altering TF binding. A study identified a non-coding single nucleotide mutation present ~90 kb downstream of the TUBX5 gene in one of the enhancer elements in patients with non-syndromic ventricular septal defect. This variant was demonstrated to alter the capacity of the enhancer to drive cardiac-specific gene expression in transgenic mice and zebrafish during heart development (Smemo et al., 2012).

Whole genome sequencing (WGS) unlike exome sequencing which captures only coding regions (around 2% of the entire genomic information), covers both coding and non-coding regions of the genome. Furthermore, WGS allows more reliable detection of copy number variants (CNVs) with a good read depth. The falling cost of WGS and improving turnaround time of data analysis will most likely drive it to overtake WES in the near future. However, the challenge to narrow down to the causal variants and avoid a bias in the filtering process based on what is already known in the genome remains a limitation in both approaches (Dorn et al., 2014). Indeed, similarly to cancer genomes (Gan et al., 2018), a large number of single nucleotide variations (SNVs) are likely to be observed in the non-coding regions of healthy versus CHD genomes. Features such as experimental observations of local epigenomic states, conservation of motifs, their vicinity to heart-relevant genes and integration of gene expression data can be used to narrow down the list of SNV's relevant for disease (May et al., 2011). However, prediction of these CHD-relevant SNPs still poses a challenge in terms of limited CHD-specific data sets, scarce conservation of these regulatory elements across animal models and humans, and associative nature of GWAS studies.

4.1.3. Available resources

Several GWAS databases are available which aim at integrating genotype-to-phenotype associations in various diseases including CHD (Brookes and Robinson, 2015). These include NHGRI-EBI GWAS catalogue and GWASdb (Li et al., 2012), (MacArthur et al., 2017), (Welter et al., 2014). GWAS Central (the former Human Genome Variation database of Genotype-To-Phenotype information) collates data from public portals of both small and large population studies and includes CHD datasets such as Tetralogy of Fallot (Beck et al., 2014). Moreover, the CHARGE consortium (Cohorts for Heart and Aging Research in Genomic Epidemiology) is one of the biggest large population GWAS studies in cardiovascular disease for which data is available at the database of Genotypes and Phenotypes (dbGaP), NIH. DNA banks of CHD patients like CONCOR (Netherlands) emphasize on genetic studies to identify markers to predict CHD risks and survival (van der Velde et al., 2005). Potential limitations of these datasets are biased participants’ selection criteria and genotyping errors.

ClinVar integrates point mutations from dbSNP, archives variants including those in splice sites, UTRs, near gene regions and their clinical significance (Landrum et al., 2018). The Exome Aggregation Consortium (ExAC), though majorly known for exome datasets from large sequencing studies, covers various genes with CHD consequence and their variants, as well as intronic and promoter sequences [http://exac.broadinstitute.org]. Another widely used database is 1000 Genomes which catalogues the largest human variants and genotype data which is currently maintained by EMBL-EBI and provides access to the variant calling files [http://www.internationalgenome.org/about]. The Human Gene Mutation Database (HGMD) is another comprehensive repository of clinically identified variations including those in regulatory regions for inherited diseases including heart disorders [http://www.hgmd.cf.ac.uk/ac/index.php]. The Cardiovascular Research Grid (CVRG) Project is a collective effort by NIH and other institutions, aimed at integrating multi-scale data including genomics, proteomics, electrocardiographic, structural and functional cardiovascular clinical data, and is of great use for GWAS studies [http://cvrgrid.org/].

4.2. Structural abnormalities

4.2.1. Definition

Structural variation refers to a variety of microscopic and sub-microscopic alterations which encompass genomic DNA with size larger than 1 kb: quantitative (copy number variants (CNVs) comprising deletions, insertions and duplications) and/or positional (translocations) or orientational (inversions) [reviewed in (Scherer et al., 2007)]. These
variations initially were limited to cytogenetic detection, until genome sequencing technologies were introduced which narrowed this detection method-based distinction (Freeman et al., 2006), (Sharp et al., 2005). In this section we discuss the structural variants that are commonly observed in syndromic CHD, techniques used to detect them, and resources storing such data. In addition, we highlight the limited exploration of non-coding regions in these alterations, which warrant further investigation.

4.2.2. Relevant studies and techniques

A variety of structural aberrations have been associated with CHD. However, the studies reporting these aberrations have been focusing on the characterisation of the coding regions within the structural variant. A large population-based study (4430) on infants with CHD, demonstrated the occurrence of chromosomal abnormalities in 12.3% of the population, with the common ones being trisomy 21, trisomy 18, 22q11.2 deletion, and trisomy 13 (Hartman et al., 2011). These structural variations have been frequently associated with syndromic CHD. Williams syndrome is another disorder caused by a microdeletion of chromosome 7q11.23. The deletion affects 26–28 genes spanning 1.5–1.8 Mb and results in multiple defects including supravalvular aortic stenosis (Adams and Schmaier, 2012). Another study reported de novo and germline aberrations to have a role in CHD. The aberrations are of median size varying from 2.74 Mb to 6.21 Mb with around 17–20 genes involved. Some of these aberrations span regions genes with well-known cardiac development role, such as NKX2-5, while others affect novel regions, whose significance in CHD is yet to be explored (Thienpont et al., 2007). These structural variants, including CNVs, apart from altering gene dosage, can also modify the topological interaction of regulatory elements lying within chromatin loops (Ibn-Salem et al., 2014). 3C and HiC techniques have captured topological associating domains (TADs) in chromatin structure where regulatory elements, such as promoters, enhancers and insulators interact. The boundaries of these domains have been found to play a role in inhibiting these interactions and are called TAD boundaries. CNVs especially large deletions leading to disruption of TAD boundaries have been associated with congenital diseases (Ibn-Salem et al., 2014), (Kyrchanova and Georgiev, 2014). However, such studies need to be applied to CHD.

Cytogenetic analyses including fluorescence in situ hybridization (FISH) and metaphase karyotype analysis, have a resolution of 5 Mb and higher, and are commonly used in the clinic to detect chromosomal aberrations in CHD syndromes such as Williams-Beuren, DiGeorge, and Alagille (Chai et al., 2016). Probes in FISH assay can be designed to detect alterations in a known DNA sequence or even RNA including lncRNAs and miRNAs (Zhang et al., 2012), (Lee et al., 2011), (Soares et al., 2018). Multiplex ligation-dependent probe amplification (MLPA) has been used in detecting chromosomal deletions and duplications in CHD patients including Tetralogy of Fallot and chromosomal 8p23 deletion syndrome (Zhu et al., 2016), (Guida et al., 2010) as well as the detection of presence or absence of an anomaly in a known genetic region and/or at intron-exon boundaries (Yoshida et al., 2016). MLPA is most useful in cases of CHDs that are genetically well-characterised and hence provide an opportunity for better clinical management of the disease. However, the drawback of these techniques lies in the fact that they are targeted probing, where aberrations in any novel region cannot be detected (Stuppia et al., 2012).

A more commonly used high-throughput detection method is array-based comparative genomic hybridization (aCGH), also known as chromosomal microarray analysis (CMA) (Haeri et al., 2016). This method detects CNVs genome-wide albeit with a resolution lower than that of FISH. However, aCGH has been used to detect other anomalies including deletions, duplications and translocations. Thienpont et al. evaluated the diagnostic value of this technique in detecting such aberrations (Thienpont et al., 2007). Defects including atrial septal defect, ventricular septal defect, pulmonary valve stenosis, aortic regurgitation, sinus venous atrial septal defect, have been successfully detected with submicroscopic aberrations using aCGH in fetuses with abnormal cardiac ultrasound results (Yan et al., 2014). The ability to identify structural variants by high-throughput sequencing using methodologies such as CNV-seq (Zhu et al., 2016) has led to development of bioinformatics tools to call these variants with high specificity and sensitivity (Tattini et al., 2015). DELLY, is one of the earliest tools developed to call various structural variants including deletions, tandem duplications and balanced rearrangements like inversions or reciprocal translocations (Rausch et al., 2012). Other tools such as Lumpy and Soft search are now available for calling structural variants, while others are also integrating these to design pipelines with a combinatorial parameter approach (Tattini et al., 2015), (Layer et al., 2014), (Hart et al., 2013). For instance, genome STRIP (Genome STRucture In Populations) is a collection of tools for identifying and genotyping structural variants (Handscape et al., 2015).

4.2.3. Available resources

Curated databases have collated structural aberrations in numerous diseases including CHD. The European Bioinformatics Institute (EBI) and the National Centre for Biotechnology Information (NCBI) maintain public archives dbVAR and DVGa for structural variants in the genome (Lappalainen et al., 2013). dbVAR (also available under ClinVar database) contains variants majoritarily greater than 50 bp in length and offers freely available raw data (Harrison et al., 2016). DECIPHER (Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources) allows access to analysis and identification of rare pathogenic aberrations in patient whole-genome and exome data and uses Ensembl resources. There are datasets of more than 1700 patients with cardiovascular diseases including phenotypical details of fetal patients (Firth et al., 2009).

Given the large number of resources and technologies available to detect CNVs, clinical interpretation and identification of syndromic and non-syndromic CHD still need clear genetic marker boundaries. Breckpot and colleagues proposed at step-by-step interpretation of these CNVs based on various features including size of the aberration and insertions. They also reported that the frequency of causal CNVs in non-syndromic CHD is lower than in syndromic CNVs (Breckpot et al., 2011).

5. Conclusion

In this article, we have discussed the role of non-coding regions in the etiology of CHD and available resources to help predict and interpret their function. We reviewed examples where alterations in various non-coding elements including enhancers, promoters and ncRNAs are associated to CHD development. However, unlike other diseases such as cancer, the exploration of non-coding regions in CHD still remains limited.

Considering the number of putative enhancers identified at various stages of heart development, there is a need to collate and annotate them to understand their role in pathogenesis of CHD. In this review, we have discussed the techniques involved in the identification of these regions in the genome including high-throughput sequencing, and we have described resources with non-coding variant data sets. These resources collate multi-scale genetic information of different stages of heart development in humans as well as other model organisms, and across various CHD populations. However, due to the weakly conserved nature of these sequences across various animal models, the identification of functional enhancers still poses a challenge. Having a comprehensive set of coding and non-coding genome markers will assist to improve the efficiency of CHD diagnosis, while further investigations into the mechanisms of pathogenesis of these variants will boost our understanding of the underlying causes of CHD.


