Article Title: Systems analysis of dilated cardiomyopathy in the next generation sequencing era

Article Type:

○ OPINION	○ PRIMER	○ OVERVIEW
○ ADVANCED REVIEW	◉ FOCUS ARTICLE	○ SOFTWARE FOCUS

Authors:

First author

Name: Magdalena Harakalova

ORCID iD: 0000-0002-7293-1029

Affiliations: [1]Department of Cardiology, Division Heart and Lungs, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Email: m.harakalova@umcutrecht.nl

Conflicts of interest: none

*corresponding author

Second author

Name: Folkert W. Asselbergs

ORCID iD: 0000-0002-1692-8669

Affiliations:

[1]Department of Cardiology, Division Heart and Lungs, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

[2]Durrer Center for Cardiovascular Research, Netherlands Heart Institute, Utrecht, NL

[3]Institute of Cardiovascular Science, University College London, London, UK

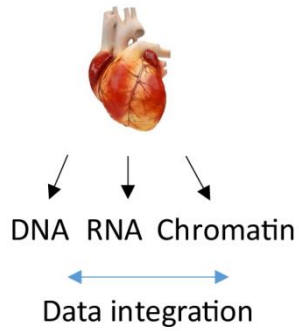
Email: f.w.asselbergs@umcutrecht.nl

Conflicts of interest: none

Abstract

Dilated cardiomyopathy (DCM) is a form of severe failure of cardiac muscle caused by a long list of etiologies ranging from myocardial infarction, DNA mutations in cardiac genes, to toxics. Systems analysis integrating next generation sequencing (NGS)-based omics approaches, such as the sequencing of DNA, RNA and chromatin, provide valuable insights into DCM mechanisms. The outcome and interpretation of NGS methods can be affected by the localization of cardiac biopsy, level of tissue degradation, and variable ratios of different cell populations, especially in the presence of fibrosis. Heart tissue composition may even differ between sexes, or siblings carrying the same disease causing mutation. Therefore, before planning any experiments, it is important to fully appreciate the complexities of DCM, and the selection of samples suitable for given research question should be an interdisciplinary effort involving clinicians and biologists. The list of NGS omics datasets in DCM to date is short. More studies have to be performed to contribute to public data repositories and facilitate systems analysis. Additionally, proper data integration is a difficult task requiring complex computational approaches. Despite these complications, there are multiple promising implications of systems analysis in DCM. By combining various types of data sets, e.g. RNAseq, ChIPseq, or 4C, deep insights into cardiac biology, and possible biomarkers and treatment targets, can be gained. Systems analysis can also facilitate the annotation of non-coding mutations in cardiac specific DNA regulatory regions that play a substantial role in maintaining the tissue and cell specific transcriptional programs in the heart.

Graphical/Visual Abstract and Caption



Caption: Systems analysis approach integrates various levels of information originating from cardiac tissue and translates knowledge into clinics.

Introduction

Before applying systems biology approaches on dilated cardiomyopathy (DCM), it is crucial to understand the complex character of the disease. DCM is defined as a severe systolic dysfunction and dilation of the (left) ventricle of the heart, leading to chronic end-stage heart failure or sudden cardiac death[1]. DCM exhibits a high morbidity and mortality rate, and it is the main indication for heart transplantation[2,3]. Based on the underlying etiology, DCM can be divided into ischemic (~70%) and non-ischemic (~30%) disease. Ischemic DCM is caused by myocardial infarction and this diagnosis is relatively easy to make[4]. Diagnostic categories of non-ischemic DCM, however, include myocarditis, valve disease, hypertension, diabetes, cardiotoxics (alcohol, drugs), and genetic disorders e.g. metabolic, neuromuscular, congenital heart defects, dysmorphic syndromes, and familial isolated forms with a known underlying gene[1]. However, in the majority of the primary (non-ischemic, non-valvular) DCM cases, the etiology is unknown. Genetics appear to play a role in 20-50% of non-ischemic, non-valvular DCM patients[1], and support a heterogeneous Mendelian inheritance model (autosomal dominant, autosomal recessive, X-linked and mitochondrial), with an incomplete penetrance, reduced expressivity, and even a possibility for digenic inheritance[5,6]. To date, over 100 genes that are encoding, among others, structural proteins of the sarcomere and desmosomes, calcium-metabolizing proteins, cell-signaling molecules, and mitochondrial enzymes, have been shown to cause DCM[1,5]. Among those, approximately 40 genes have been verified to be truly DCM-causing, mostly supported by convincing familial segregation and functional testing[1]. These discoveries have led to new recommendations for next generation sequencing (NGS)-based diagnostic evaluation to find the genetic cause of the patient's disease[7], predict disease prognosis (e.g. mutations in *LMNA* gene indicate a worse prognosis[7]), and to indicate treatment[8], or the use of cardiac devices[3,9]. In addition to the screening of the index patient, genetic counseling and cascade screening is offered to family members for an early recognition of individuals at risk[10]. Hence, the complex architecture of DCM does not end with its numerous etiologies. The clinical course of DCM is very variable and hard to predict, regardless the cause. It ranges from patients with little to no symptoms, to patients with an overt heart failure and a rapidly deteriorating ventricle, even between family members carrying the same mutation[11,12]. In addition, the clinical presentation, course, and outcome of DCM can differ between males and females[13]. Therefore, before

planning any experiments, it is important to fully appreciate the complexities of DCM, and the selection of samples suitable for given research question should be an interdisciplinary effort involving clinicians and biologists.

Systems biology is an emerging interdisciplinary field aiming at connecting information from various levels of complex biological systems[14]. Systems analysis integrating NGS-based omics approaches, such as the genome-wide sequencing of DNA, RNA, and chromatin (genomics, transcriptomics and regulomics, respectively)[15], might provide valuable insights into our understanding of DCM. Before the era of omics, numerous studies have provided crucial information on particular gene changes in DCM, and often have led to new treatment targets[16]. However, during both onset and manifestation of DCM, hundreds of genes undergo changes in their transcriptional regulation and expression leading to processes such as fibrosis, cell death or impairment of muscle function and lipid (energy substrate) metabolism[17-19]. Therefore the best advantage of systems analysis is integrating sequencing, regulatory and transcriptional information about all genes in given cells or tissue at the same time. These methods, however, need to be performed in a high degree of cardiac tissue- or even cardiomyocyte specificity[20], which might be problematic due to the limited availability of fresh patient and control cardiac tissue, the diffuse pattern of fibrosis, and the sensitivity of cardiomyocytes to culturing conditions[21,22]. It has recently been shown that tissue composition and physiological processes differ between various genetic forms of DCM that would otherwise be difficult to recognize clinically[22,23]. It is therefore very promising to explore the options of systems analysis, and translate the vast amount of obtained information into clinics. Depending on the setup of a research project, knowledge about the trigger of DCM, the mechanism of DCM progression, disease modifiers, or biomarkers, can be obtained well in line with the principles of precision medicine[24]. Consequently, the aim of this focus article is not to provide a catalogue of all kinds of methods and analyses implicated in systems biology in general, but rather to briefly discuss the real-world problems that researchers might encounter during such projects in DCM, how should they orient themselves in the long list of NGS techniques, and how could they utilize their results.

PLANNING AN EXPERIMENT

Since the availability of human cardiac samples from DCM patients is restricted, the list of examples of NGS omics studies published to date is short. Especially in regards to methods that require fresh or freshly frozen tissue[25,26]. Under ideal circumstances, there would be enough fresh cardiac material obtained from a DCM patient to isolate sufficient amounts of DNA, RNA, and chromatin, and perform all suitable methods to obtain maximum systems analysis information. However, the access to human cardiac samples, as compared to the example of blood drawing or skin biopsies, is very limited. Systems analyses in DCM can be hindered by the lack of suitable (control) cardiac tissue, the quality of archived cardiac material, or the variable amount of cardiac fibrosis in biopsies[22]. Therefore it is crucial to appreciate the importance of several parameters, before starting an actual experiment.

Limited source of cardiac tissue

There are several limitations in obtaining heart tissue. Cardiac biopsies from patients are primarily acquired at the end-stage of DCM, already presenting with severe damage to cardiomyocytes and extracellular matrix[27]. Such samples can be obtained during diagnostic endomyocardial biopsies, heart transplantation, implantation of a left ventricular assist device (LVAD), or post mortem, where little can be learned about what has triggered heart failure. Most interesting group of patients are pathogenic mutation carriers with no or only mild DCM symptoms[28]. Since biopsies cannot be obtained at an early stage of DCM, iPSC-derived cardiomyocytes engineered from skin biopsies or from blood, might be the only possibility for pre-symptomatic mutation carriers[29]. It is further problematic to obtain fresh tissue from proper human controls, as control cardiac tissue is usually obtained from brain-death donors who would not qualify for heart transplantation, from abortion material, or post-mortem. More coordination is needed to collect fresh cardiac tissue more efficiently, especially in earlier stages of heart failure, or even in pre-symptomatic mutation carriers.

Variable quality of cardiac tissue

Variations in biobanking protocols might cause problems when comparing cohorts from different centers. Depending on the working protocol, there is variability in archiving cardiac biopsies as formalin fixed and paraphing-embedded (FFPE), or stored frozen at -80°C. This can later determine the

suitability of stored biopsy for NGS omics method, e.g. all RNA and chromatin-based methods need to be performed on fresh or frozen cardiac tissue[15]. In addition, DNA, RNA, and chromatin quality decreases over time, therefore post-mortem material, or material frozen after hours at room temperature or at 4°C, should be used with caution[30]. Quality control measures should always be taken to assess the degradation of material, such as the determination of RNA Integrity Number (RIN**, Figure 1A)** that might also give an approximate picture about the stability of DNA methylation or chromatin. RIN is based on a numbering system from 1 to 10, with 1 being the most degraded profile and 10 being the most intact. RNA samples with RIN>7 indicate the best quality[31].

Heterogeneous histological cell composition

In normal human cardiac tissue, cardiomyocytes occupy up to 85% of the tissue volume, hence they spend only 30-40% of the cardiac cell count[32]. In the remaining cell fraction, endothelial cells constitute up to 60%, fibroblasts up to 20%, and hematopoietic-derived cells 5%-10% of the non-myocytes in the heart[33]. During disease, the composition of cells might change, and even differ between left and right ventricle and the interventricular septum, especially between various DCM subtypes **(Figure 1B,C)**[22,34]. Therefore, it is also important to keep consistency in the localization of biopsies. Histopathological assessment, at least hematoxylin-eosin staining, should be performed for each included biopsy, and the expected amount of fibrosis in the sequenced area should be determined[22,34]. Results will differ, in case whole intact tissue slice is used, including all cell types present, if fibrotic subepicardial layer has macroscopically been removed with scalpel to enrich for cleaner areas of cardiomyocytes, or after cardiomyocyte nuclei selection using cardiomyocyte-specific antibody. However, the latter needs to be performed on fresh, unfrozen tissue[35]. Therefore, before we have reliable working protocols to isolate particular cardiac cells from archived tissue, the information about cell composition and localization of each cardiac biopsy is crucial for data interpretation, and should routinely be disclosed in scientific publications alongside data analysis. Importantly, the rapidly developing single cell sequencing brings the promise of solving cardiac tissue heterogeneity problem[36].

Homogeneous selection of DCM cohorts

The lack of proper clinical and genetic information including DNA subtype, stage of disease, age and sex, without proper match for localization of cardiac biopsy, and without performing histopathology of biopsy, will undoubtedly have a negative effect on study results and create too much heterogeneity. Where genetic background is expected, at least sequencing of DCM diagnostic gene panel would be preferred[5,7], including controls compared to a given DCM tested group, e.g. if *MYBPC3* mutation positive DCM group, control group should be *MYBPC3* mutation negative. In some cases it might be better to use end-stage DCM tissue with other type of DCM etiology instead of control tissue, e.g. *MYBPC3* vs. *TTN*-mutated samples or ischemic DCM etiology. Interestingly, in the case of ischemic DCM samples, the best type of control might be a non-infarcted (remote) area from the same heart. It is crucial to prepare a suitable research question and well-designed research plan taking into account the quality of available tissue and the expected etiology of DCM. Most importantly, regardless of the characteristics of tested cohorts, the right sample size, in order to obtain sufficient statistical power for a NGS-based omics experiment, should be used[37].

Avoiding technical batch effects

Various batch effects might impair data quality[38]. To avoid them, it is best to plan well in advance to include all available samples. DNA, RNA, and chromatin should be isolated, libraries should be prepared, and sequenced in mixed groups, to make sure that each batch contains a mix of samples from every tested group (e.g. controls should be processed in the same batch as patients, **Figure 2**). Otherwise it could be difficult to separate the effect of storage and collection condition from the actual biology. It is also good practice to avoid a direct comparison of a group of post-mortem obtained samples to a group of freshly frozen samples, since the degradation will be the strongest co-founder. Batch effects can be detected *in silico*[39] and removed by a vast range of bioinformatic tools[40]. However, those correction methods can impair downstream analysis[41], and if possible, it is best to avoid batch effects in the first place.

NGS-BASED OMICS METHODS

Thanks to low costs, short turnaround times, flexibility and genome-wide coverage, NGS has greatly accelerated biological and medical research and discovery. NGS has a broad spectrum of applications, which in principle always result from sequencing of a certain type of DNA library. The input can differ per biological question and depends on the isolation procedure of DNA, RNA, or chromatin[42].

DNA sequencing

Genome sequencing is based on protocols starting with genomic DNA, which is fragmented into small pieces subsequently used for DNA library preparation. DNA is often isolated from blood to investigate germline mutations, as it is the case in diagnostic DCM gene sequencing to determine the patient's genetic DCM etiology[1], or can be isolated directly from cardiac tissue, to identify somatic DNA mutations specifically acquired in cardiomyocytes[43]. Cardiac genomic DNA can be isolated from cells or a whole piece of cardiac tissue, including post-mortem tissue. It is possible to sequence the whole genome of DCM patient without additional modifications (whole genome sequencing – WGS)[44], or to enrich DNA libraries for specific regions of interest (genomic enrichment), which is the basis of diagnostic whole exome sequencing (WES) or DCM gene panels[1,5,7]. WES and WGS datasets can be used to detected point mutations, small insertions and deletions, but also large structural copy number variations implicated in DCM[45,46].

Methylome sequencing (Methyl-seq) for determination of methylation DNA status down to a single base resolution can be performed after additional bisulfite treatment of isolated DNA[47]. However, the usability of DNA methylation methods in DCM is limited by a relatively large amount of input cardiac tissue and high sequencing costs. Therefore, most of the knowledge about DNA methylation status so far has been gained from microarray-based methylation methods[47,48].

Importantly, the rapidly developing single cell DNA sequencing brings the promise of solving cardiac tissue heterogeneity problem, which bulk tissue sequencing brings[36]. The overall benefit of DNA sequencing techniques in general, with the exception of bisulfite-treated DNA, is that also post-mortem or FFPE tissue can be used after small modifications. It should be noted, that sequencing of germline DNA to determine the etiology of primary (non-ischemic, non-valvular) DCM patients and their family members, should be the first important step before moving on to RNA and chromatin sequencing methods.

RNA sequencing

Before the era of NGS, transcriptome analysis was relying on hybridization-based microarray techniques which have proven to be valuable in providing information about gene expression in cardiac tissue and cardiomyocytes[49,50]. However, those hybridization techniques provide only a limited view on the complexity and dynamics of the transcriptome, and are biased towards a priori knowledge about expression patterns[51].

After a total RNA or specific fraction of RNA (e.g. small RNAs, miRNAs) has been isolated from fresh or frozen material, NGS library preparation can proceed from complementary DNA (cDNA)[52]. Cardiac material with a longer post-mortem interval is not a suitable source of tissue for RNA sequencing (RNA-seq)[53]. The most frequently used protocols are the polyA-selection for selection of predominantly mRNA, but also lncRNAs, and small RNA species selection techniques for RNA-seq of miRNAs, snoRNAs, siRNAs, snRNAs, exRNAs, pRNAs, piRNAs, and scaRNAs[26,54,55].

Interestingly, tomo-seq has recently emerged as a new RNA-seq technique providing an extra layer of information by combining transcriptome analysis with spatial resolution[56,57]. Single cell RNA-seq sequencing has the potential to identify clusters of different cell types in cardiac tissue, even different subtypes of cardiomyocytes[35,58]. Ribosomal profiling, the deep sequencing of ribosome-protected mRNA fragments, represents a powerful tool for globally monitoring protein translation[59] and might especially be interesting in proving the actual damaging effect of DCM mutations[60]. Taken together, RNA-seq techniques provide unprecedented knowledge about specific expression of various types of RNA molecules from a whole piece of cardiac biopsy to a single cell resolution.

Chromatin sequencing

The change of chromatin landscape and activity (regulome) is an important hallmark of human disease[15]. It is the most unexplored part of epigenetic NGS omics methods in DCM, mostly because researchers often prefer to get a more direct information about expression of genes from RNA-seq methods. The basic principle of all chromatin-based NGS methods is the discovery of the genomic location and activity of DNA regulatory elements which are highly tissue, cell, and condition specific[15].

To understand the basic principles of the various chromatin sequencing methods and their potential for DCM research, it is first important to understand the characteristics of regulatory elements.

DNA regulatory elements

Regulatory elements encoded in DNA sequence, often referred to as promoters and enhancers, recruit transcription factors (TFs) to their DNA sequence through TF binding sites (TFBMs), allowing cells to precisely control the timing, location, and the level of gene transcription[61]. Promoters are classically recognized as proximate gene regions that initiate gene transcription, while enhancers as more distal regions acting through a physical contact with promoters via chromatin loops[62]. However, recent studies challenge the notion that promoters and enhancers are distinct entities and show evidence that also promoters can act over long distances[63,64]. It has been shown that regulatory regions display a high level of tissue specificity, and that enrichment of Gene Ontology (GO) terms based on the annotation of genes with the closest transcription start site has cell-specific patterns[61]. The genomic coordinates of regulatory elements indicate peaks and regions. The location of detected peaks can be used for *in silico* annotation of genes in the vicinity of the regions in the assumption that those genes are regulated, or to compare the intensity of signal between various conditions[65,66].

Probing chromatin activity

Sequencing of the regulatory features normally starts with obtaining a highly organized chromatin material in living or frozen cells[61]. The mostly used chromatin-based NGS technique is chromatin immunoprecipitation sequencing (ChIP-seq). It is based on hybridization of a DNA-binding antibody followed by sequencing of the precipitated chromatin product[67]. Various types of antibodies can be used depending on the research question, assuming the protein interacts with DNA, such as antibodies against polymerases, histone modifications, or transcription factors[20,68]. One of the most frequently used ChIP-seq techniques is based on probing the histone 3 lysine 27 acetylation (H3K27ac) mark, recognizing active from poised promoters and enhancers[69]. There are many other chromatin techniques, based on different principles, providing similar information about the location and activity of open chromatin regions, such as DNAse-seq, ATAC-seq or FAIRE-seq[70,71].

One of the largest obstacle of all NGS chromatin activity probing studies is to find out what gene is regulated by the actual regulatory region. To date, most of this annotation has been based on *in silico* analysis of the genomic vicinity of a gene to the regulatory region[65]. However, since those interactions can span several hundreds, or even thousands of kilobases and at the same time, and the regulated gene does not necessarily need to be the closest, there is a high change to create false positive and false negative associations. Therefore, the best way is to use chromatin conformation techniques, such as the 4C, HiC or ChIA-PET, which provide a unique view about the tight 3D chromatin interactions of regulatory elements even on longer distances[72-74].

SYSTEMS ANALYSIS IMPLICATIONS IN DCM

Omics data integration is a difficult task requiring expensive and complex bioinformatic approaches[75,76]. The biggest obstacle is to simply visualize and interpret the biological relevance of hundreds to thousands differential genes (**Figure 3**). However, despite the complications, there are multiple promising implications of systems analysis in DCM. By combining various types of NGS omics data sets, e.g. RNAseq, ChIPseq, or 4C, deep insights into cardiac biology, and possible biomarkers and treatment targets, can be gained. Systems analysis can also facilitate the annotation of non-coding mutations in cardiac specific DNA regulatory regions that play a substantial role in maintaining the tissue and cell specific transcriptional programs in the heart.

Data integration

Proper data integration from different types of experiments, necessary for a wider understanding of the underlying molecular mechanisms that lead to onset and progression of DCM, is challenging. Each type of NGS omics dataset brings an enormous amount of information on its own. A single RNA-seq dataset already provides information about genes expressed in a certain cell or tissue, and would equal to thousands of qPCR experiments. For example, RNAseq data can be used to determine tissue specificity related to mutated genes in DCM to assess potential extracardiac manifestations[5], or the variation of allelic disbalance between the mutated and wild type allele[77]. However, integration of various other types of datasets would provide a much more complex picture.

The combination of ChIP-seq data providing information about the location of cardiac specific enhancer, together with RNA-seq information about expressed cardiac genes, and 4C signal from particular enhancer viewpoint, will inform about which gene is regulated by the tested enhancer. In case these data are supported by information from ChIPseq for specific transcription factor binding to the tested enhancer, and WGS to detect DNA mutation in the binding site of that transcription factor, such mutations might impair the enhancer activity of gene through impairing the binding affinity of this transcription factor[78].

Alternatively, in a recent epigenome-wide association study in DCM patients and controls, methylation array profiling was combined with RNA-seq and WGS to determine cardiac gene patterning and a novel class of biomarkers. DNA methylation was first mapped in left-ventricular biopsies and whole peripheral blood of living probands. RNA-seq was performed on the same samples in parallel to identify differentially expressed genes linked to methylation profiles. WGS of all patients allowed exclusion of promiscuous genotype-induced methylation calls. Using this staged multi-omics study design, 517 epigenetic loci were linked with DCM and cardiac gene expression[48].

Therefore, by combining various types of NGS omics data sets, deep insights into cardiac biology and possible biomarkers and treatment targets can be gained, especially with other omics techniques not discussed here, such as the proteome and the metabolome of the heart.

Annotation of coding and non-coding DNA variants in WGS

Until now, systematic approaches for the identification of causal and modifying mutations from genome-wide sequencing data have mainly been restricted to protein-coding DNA sequences. Regulatory regions that also play a substantial role in maintaining the tissue and cell specific transcriptional programs[15], have largely been ignored. Promoters and enhancers regulate the time, location and levels of gene expression. Sequence DNA variants in those regulatory elements can alter the binding affinity of transcription factors[79,80], thereby changing or even diminishing the expression of a regulated gene[78,81], even though the gene itself is not mutated. In addition, it is well documented, that long and small non-coding RNAs play crucial roles in translation, RNA splicing, DNA replication, gene regulation, or gene silencing[82,83].

Taken together, disease causing or modifying mutations might be located in classical protein-coding genes in nuclear and mitochondrial DNA encoding structural or functional proteins of cardiac cells[5], but also in RNA genes (e.g. lncRNA, miRNA, etc.)[84], or even non-coding regulatory elements (e.g. promoters, enhancers, and their transcription factor binding motifs)[85,86]. Although there are no known examples of non-coding mutations implicated in DCM yet, other genetic conditions with mutations in enhancers exist, such as isolated congenital heart defect in Holt-Oram syndrome (mutated *TBX5* enhancer)[87], familial agenesis of pancreas (mutated *PTF1A* enhancer)[88], or preaxial polydactyly (mutated SHH enhancer)[89,90].

A multi-omics stepwise approach to sift through the mountainous amount of data generated by systems analysis integrating various levels of NGS omics information is shown in **Figure 4** and **5.** The identification of coding mutations from WGS data should focus on exonic regions, which might seem straight-forward. The location of genes in general is already known and is easily accessible in public databases, such as Ensembl or USCS. However, information about which protein coding and RNA genes are expressed in cardiac tissue or cardiac cells during health and DCM is largely lacking. By integrating WGS data based on the sequencing of germline DNA of investigated DCM patients with RNA-seq data from available control or DCM cardiac biopsies, the searching space might significantly be narrowed-down to protein and RNA coding genes, which product might be present in the affected organ.

A more complicated approach should be taken for detection of biologically-meaningful variants overlapping with regulatory elements. Under an ideal scenario, a non-coding mutation detected by WGS should be located inside a regulatory element present in cardiac tissue or cardiac cells during health or DCM. Most promising are variants predicted to impair a crucial position inside a cardiac TF binding site motif, best if binding site confirmed via cardiac ChIPseq of the TF itself (**Figure 4** and **5**). However, ChIPseq datasets for TF binding in human cardiac tissue or cells are not available, therefore current approaches take into account only *in silico* predictions based on datasets from other species or tissues[91].

Searching for intrafamilial disease modifiers

The search for intrafamilial (epi)genetic modifiers might be one of the most translational examples of systems analysis utilization in DCM. Cascade genetic screening allows an early detection of many family members that are pathogenic carriers. However, a large portion of these mutation carriers do not

develop cardiomyopathy throughout life, or instead develop only a mild, subclinical phenotype[92]. This intrafamilial variability represents a significant problem, as asymptomatic mutation carriers may undergo unnecessary cardiologic testing, treatment with drugs that cause negative side-effects, or implantation of cardiac devices that are not necessary[93]. On the other hand, mutation carriers at risk, who are not identified properly, may be undertreated, and die suddenly at a young age[92].

It has been shown that patients with multiple protein-coding mutations show a more severe disease progression compared to patients with single mutation (gene-dose effect)[6,94]. Conclusively, the gene-dose effect in DCM patients and mutation carriers should also be considered for variants in non-coding DNA sequences[95], based on approaches described in **Figure 4** and **5.** Since DCM has can be caused by hundreds of mutations[96] it might be difficult to search for disease modifiers in single families with DCM mutation with an ultra-low frequency. Large founder mutation carrier populations[97] or multigenerational pedigrees with DCM might therefore be a better choice. Selection of family members should be carefully matched for extremes of phenotypes (younger individual with DCM vs. older family member with normal heart function), and if possible, should be sex matched. WGS datasets should be produced for the matched extremes and variants prioritized according to stepwise approaches indicated in **Figure 4** and **5**. Combining WGS with epigenomic maps and iPS-cell technologies to identify a disease-causing mutation in an enhancer element, has been already been proven useful for families with agenesis of pancreas, where a mutation in an enhancer of *PTF1A* gene has been detected[88]. Similar approach should be taken for intrafamilial disease modifying variants, where cardiomyocytes from iPS-cells originating from blood or skin biopsies of the included individuals, could be used to test the DCM-causing pathogenic mutation with the presence of absence of the newly detected DCM-modifying variant, including the correction of mutations using genetic engineering[98-100].

Sidebar 1: Individual NGS omics data shared publicly enhance systems analysis but raise ethical issues. The submission of raw (fastq files) or mapped (bam files) individual NGS data into public repositories is currently required by majority of scientific journals. These files contain sequencing information reusable by others to test new methods, perform meta-analyses, or test new research questions[101]. Informed consent of the patient or their family member is required, when applicable, and it is routinely preformed for DNA sequencing studies, such as WGS or WES. However, ethical research approvals for methods like Methyl-seq, RNAseq or ChIPseq often do not include specific permission for

putting NGS data in public repositories. Although patient data must be de-identified, the genotyping information itself contains several identifiers. There is a high potential of discovering a person via bam file by for example sex match (active promoter or expression of *XIST* gene on chromosome X for females, and active promoter or expression for *EIF1AY* on chromosome Y for male, or simply the presence or absence of reads on chromosome Y) or even a rare (population-specific) pathogenic DCM mutation[97]. More effort has to be invested into reevaluating the present rules and creating file types that would at least prevent detailed genotyping information, while preserving information about genomic coordinates[102]. It also has to be ensured that insurance companies and governmental organizations are not allowed to access and use patients' genetic information, to avoid that carriership will lead to higher insurance fees[103].

Conclusion

Systems analysis approach has limitless future applications that could improve the understanding of DCM. These applications are already used in other fields, such as oncology and immunology, but many of them still need to be applied to DCM. Thus far, DNA sequencing methods, such as targeted gene enrichment, WES, or WGS, are the most frequently NGS methods utilized in DCM research and diagnostics. The number of RNA-seq studies and other novel RNA-seq techniques, such as single cell RNA-seq, tomo-seq, or ribosomal profiling, is currently on the rise. However, more studies need to be produced to investigate the implications of methylome and regulome in DCM. We are still at the stage when simply a repository of NGS omics DCM-related datasets is needed to assist substantial number of researchers, especially to annotate genome-wide DNA studies, such as WGS, but also GWAS[104]. Future applications of NGS omics methods in DCM will depend on our ability to isolate and culture cardiomyocytes from human cardiac biopsies, especially the archived ones, or create iPSC that are representative for adult cardiomyocytes. Efficient 3D cardiomyocyte cultures from pre-symptomatic or symptomatic DCM mutation carriers will facilitate the development of druggable targets based on systems analysis approaches. We also have to develop better bioinformatic tools that would allow to understand the complex interplay between genes and proteins in a changing environment during disease, and for their intuitive visualization. Taken together, the field of systems analysis in DCM is yet to be started.

Acknowledgments

We would like to thank Ema Nagyova and Jiayi Pei for providing Bioanalyzer images. Magdalena Harakalova is supported by Wilhelmina Children's Hospital research funding OZF/14 and by personal research fund VENI ZonMW/NVO2016 - 016.176.136. Folkert W. Asselbergs is supported by UCL Hospitals NIHR Biomedical Research Centre. We acknowledge the support from the Netherlands Cardiovascular Research Initiative: An initiative with support of the Dutch Heart Foundation (CVON2015-12 eDETECT and CVON2014-40 DOSIS). Part of this work is funded through ERA-CVD program, jointly funded by Dutch Heart Foundation and Netherlands Organization for Health Research and Development (DETECTIN-HF, 2016T096).

Further Reading

The ENCODE project: http://www.nature.com/encode/#/threads

Human heart proteome: http://www.proteinatlas.org/humanproteome/heart

American Heart Association approved data repositories (AHA): https://professional.heart.org/professional/ResearchPrograms/UCM_461443_AHA-Approved-Data-Repositories.jsp

References

References

1 Hershberger, R. E., Hedges, D. J. & Morales, A. Dilated cardiomyopathy: the complexity of a diverse genetic architecture. *Nature reviews. Cardiology* **10**, 531-547, doi:10.1038/nrcardio.2013.105 (2013).

2 Sammani, A. *et al.* Thirty years of heart transplantation at the University Medical Centre Utrecht. *Netherlands heart journal : monthly journal of the Netherlands Society of Cardiology and the Netherlands Heart Foundation*, doi:10.1007/s12471-017-0969-0 (2017).

3 Wilde, A. A. & Behr, E. R. Genetic testing for inherited cardiac disease. *Nature reviews. Cardiology* **10**, 571-583, doi:10.1038/nrcardio.2013.108 (2013).

4 Felker, G. M., Shaw, L. K. & O'Connor, C. M. A standardized definition of ischemic cardiomyopathy for use in clinical research. *J Am Coll Cardiol* **39**, 210-218 (2002).

5 Harakalova, M. *et al.* A systematic analysis of genetic dilated cardiomyopathy reveals numerous ubiquitously expressed and muscle-specific genes. *Eur J Heart Fail* **17**, 484-493, doi:10.1002/ejhf.255 (2015).
6 Roncarati, R. *et al.* Doubly heterozygous LMNA and TTN mutations revealed by exome sequencing in a severe form of dilated cardiomyopathy. *European Journal of Human Genetics* **21**, 1105-1111 (2013).
7 Van Spaendonck-Zwarts, K. Y. *et al.* Genetic analysis in 418 index patients with idiopathic dilated cardiomyopathy: Overview of 10 years' experience. *European Journal of Heart Failure* **15**, 628-636 (2013).
8 Harakalova, M. *et al.* Dominant missense mutations in ABCC9 cause Cantu syndrome. *Nat Genet* **44**, 793-796, doi:10.1038/ng.2324 (2012).
9 Kummeling, G. J., Baas, A. F., Harakalova, M., van der Smagt, J. J. & Asselbergs, F. W. Cardiovascular genetics: technological advancements and applicability for dilated cardiomyopathy. *Netherlands heart journal : monthly journal of the Netherlands Society of Cardiology and the Netherlands Heart Foundation* **23**, 356-362, doi:10.1007/s12471-015-0700-y (2015).
10 Hershberger, R. E. *et al.* Genetic evaluation of cardiomyopathy--a Heart Failure Society of America practice guideline. *J Card Fail* **15**, 83-97, doi:10.1016/j.cardfail.2009.01.006 (2009).
11 Aleksova, A. *et al.* Natural history of dilated cardiomyopathy: from asymptomatic left ventricular dysfunction to heart failure--a subgroup analysis from the Trieste Cardiomyopathy Registry. *Journal of cardiovascular medicine (Hagerstown, Md.)* **10**, 699-705, doi:10.2459/JCM.0b013e32832bba35 (2009).
12 Fatkin, D. Familial dilated cardiomyopathy: Current challenges and future directions. *Glob Cardiol Sci Pract* **2012**, 8, doi:10.5339/gcsp.2012.8 (2012).
13 Meyer, S., van der Meer, P., van Tintelen, J. P. & van den Berg, M. P. Sex differences in cardiomyopathies. *Eur J Heart Fail* **16**, 238-247, doi:10.1002/ejhf.15 (2014).
14 Ayers, D. & Day, P. J. Systems Medicine: The Application of Systems Biology Approaches for Modern Medical Research and Drug Development. *Mol Biol Int* **2015**, 698169, doi:10.1155/2015/698169 (2015).
15 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).
16 Leask, A. Potential therapeutic targets for cardiac fibrosis: TGFbeta, angiotensin, endothelin, CCN2, and PDGF, partners in fibroblast activation. *Circ Res* **106**, 1675-1680, doi:10.1161/CIRCRESAHA.110.217737 (2010).
17 Fukushima, A., Milner, K., Gupta, A. & Lopaschuk, G. D. Myocardial Energy Substrate Metabolism in Heart Failure : from Pathways to Therapeutic Targets. *Curr Pharm Des* **21**, 3654-3664 (2015).
18 Harvey, P. A. & Leinwand, L. A. The cell biology of disease: cellular mechanisms of cardiomyopathy. *J Cell Biol* **194**, 355-365, doi:10.1083/jcb.201101100 (2011).
19 Louzao-Martinez, L. *et al.* Characteristic adaptations of the extracellular matrix in dilated cardiomyopathy. *Int J Cardiol* **220**, 634-646, doi:10.1016/j.ijcard.2016.06.253 (2016).
20 Levo, M. & Segal, E. In pursuit of design principles of regulatory sequences. *Nature reviews. Genetics* **15**, 453-468, doi:10.1038/nrg3684 (2014).
21 Louch, W. E., Sheehan, K. A. & Wolska, B. M. Methods in cardiomyocyte isolation, culture, and gene transfer. *J Mol Cell Cardiol* **51**, 288-298, doi:10.1016/j.yjmcc.2011.06.012 (2011).
22 Sepehrkhouy, S. *et al.* Distinct fibrosis pattern in desmosomal and phospholamban mutation carriers in hereditary cardiomyopathies. *Heart Rhythm* **14**, 1024-1032, doi:10.1016/j.hrthm.2017.03.034 (2017).
23 Bollen, I. A. E. *et al.* Genotype-specific pathogenic effects in human dilated cardiomyopathy. *J Physiol* **595**, 4677-4693, doi:10.1113/JP274145 (2017).

24 Lu, Y. F., Goldstein, D. B., Angrist, M. & Cavalleri, G. Personalized medicine and human genetic diversity. *Cold Spring Harb Perspect Med* **4**, a008581, doi:10.1101/cshperspect.a008581 (2014).

25 Herrer, I. *et al.* RNA-sequencing analysis reveals new alterations in cardiomyocyte cytoskeletal genes in patients with heart failure. *Laboratory investigation; a journal of technical methods and pathology* **94**, 645-653, doi:10.1038/labinvest.2014.54 (2014).

26 Liu, Y. *et al.* RNA-Seq identifies novel myocardial gene expression signatures of heart failure. *Genomics* **105**, 83-89, doi:10.1016/j.ygeno.2014.12.002 (2015).

27 Lok, S. I. *et al.* Myocardial fibrosis and pro-fibrotic markers in end-stage heart failure patients during continuous-flow left ventricular assist device support. *European journal of cardio-thoracic surgery : official journal of the European Association for Cardio-thoracic Surgery* **48**, 407-415, doi:10.1093/ejcts/ezu539 (2015).

28 Lakdawala, N. K. *et al.* Subtle abnormalities in contractile function are an early manifestation of sarcomere mutations in dilated cardiomyopathy. *Circulation. Cardiovascular genetics* **5**, 503-510, doi:10.1161/CIRCGENETICS.112.962761 (2012).

29 Del Alamo, J. C. *et al.* High throughput physiological screening of iPSC-derived cardiomyocytes for drug development. *Biochim Biophys Acta* **1863**, 1717-1727, doi:10.1016/j.bbamcr.2016.03.003 (2016).

30 Gupta, S., Halushka, M. K., Hilton, G. M. & Arking, D. E. Postmortem cardiac tissue maintains gene expression profile even after late harvesting. *BMC Genomics* **13**, 26, doi:10.1186/1471-2164-13-26 (2012).

31 Schroeder, A. *et al.* The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC molecular biology* **7**, 3, doi:10.1186/1471-2199-7-3 (2006).

32 Zhou, P. & Pu, W. T. Recounting Cardiac Cellular Composition. *Circulation research* **118**, 368-370, doi:10.1161/CIRCRESAHA.116.308139 (2016).

33 Pinto, A. R. *et al.* Revisiting Cardiac Cellular Composition. *Circulation research* **118**, 400-409, doi:10.1161/CIRCRESAHA.115.307778 (2016).

34 Gho, J. M. *et al.* High resolution systematic digital histological quantification of cardiac fibrosis and adipose tissue in phospholamban p.Arg14del mutation associated cardiomyopathy. *PLoS One* **9**, e94820, doi:10.1371/journal.pone.0094820 (2014).

35 See, K. *et al.* Single cardiomyocyte nuclear transcriptomes reveal a lincRNA-regulated de-differentiation and cell cycle stress-response in vivo. *Nat Commun* **8**, 225, doi:10.1038/s41467-017-00319-8 (2017).

36 Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nature reviews. Genetics* **17**, 175-188, doi:10.1038/nrg.2015.16 (2016).

37 Ching, T., Huang, S. & Garmire, L. X. Power analysis and sample size estimation for RNA-Seq differential expression. *RNA* **20**, 1684-1696, doi:10.1261/rna.046011.114 (2014).

38 Goh, W. W. B., Wang, W. & Wong, L. Why Batch Effects Matter in Omics Data, and How to Avoid Them. *Trends Biotechnol* **35**, 498-507, doi:10.1016/j.tibtech.2017.02.012 (2017).

39 Jiang, L. *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome research* **21**, 1543-1551, doi:10.1101/gr.121095.111 (2011).

40 Oytam, Y. *et al.* Risk-conscious correction of batch effects: maximising information extraction from high-throughput genomic datasets. *BMC Bioinformatics* **17**, 332, doi:10.1186/s12859-016-1212-5 (2016).

41 Nygaard, V., Rodland, E. A. & Hovig, E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics* **17**, 29-39, doi:10.1093/biostatistics/kxv027 (2016).

42 Reuter, J. A., Spacek, D. V. & Snyder, M. P. High-throughput sequencing technologies. *Mol Cell* **58**, 586-597, doi:10.1016/j.molcel.2015.05.004 (2015).

43 Blokzijl, F. *et al.* Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260-264, doi:10.1038/nature19768 (2016).

44 Hastings, R. *et al.* Combination of Whole Genome Sequencing, Linkage, and Functional Studies Implicates a Missense Mutation in Titin as a Cause of Autosomal Dominant Cardiomyopathy With Features of Left Ventricular Noncompaction. *Circulation. Cardiovascular genetics* **9**, 426-435, doi:10.1161/CIRCGENETICS.116.001431 (2016).
45 Straver, R., Weiss, M. M., Waisfisz, Q., Sistermans, E. A. & Reinders, M. J. T. WISExome: a within-sample comparison approach to detect copy number variations in whole exome sequencing data. *European journal of human genetics : EJHG* **25**, 1354-1363, doi:10.1038/s41431-017-0005-2 (2017).
46 Norton, N. *et al.* Genome-wide studies of copy number variation and exome sequencing identify rare variants in BAG3 as a cause of dilated cardiomyopathy. *American journal of human genetics* **88**, 273-282, doi:10.1016/j.ajhg.2011.01.016 (2011).
47 Haas, J. *et al.* Alterations in cardiac DNA methylation in human dilated cardiomyopathy. *EMBO Mol Med* **5**, 413-429, doi:10.1002/emmm.201201553 (2013).
48 Meder, B. *et al.* Epigenome-Wide Association Study Identifies Cardiac Gene Patterning and a Novel Class of Biomarkers for Heart Failure. *Circulation* **136**, 1528-1544, doi:10.1161/CIRCULATIONAHA.117.027355 (2017).
49 Barrans, J. D., Allen, P. D., Stamatiou, D., Dzau, V. J. & Liew, C. C. Global gene expression profiling of end-stage dilated cardiomyopathy using a human cardiovascular-based cDNA microarray. *The American journal of pathology* **160**, 2035-2043, doi:10.1016/S0002-9440(10)61153-4 (2002).
50 Colak, D. *et al.* Left ventricular global transcriptional profiling in human end-stage dilated cardiomyopathy. *Genomics* **94**, 20-31, doi:10.1016/j.ygeno.2009.03.003 (2009).
51 Zhao, S., Fung-Leung, W. P., Bittner, A., Ngo, K. & Liu, X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One* **9**, e78644, doi:10.1371/journal.pone.0078644 (2014).
52 Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics* **10**, 57-63, doi:10.1038/nrg2484 (2009).
53 Zhu, Y., Wang, L., Yin, Y. & Yang, E. Systematic analysis of gene expression patterns associated with postmortem interval in human tissues. *Sci Rep* **7**, 5435, doi:10.1038/s41598-017-05882-0 (2017).
54 Akat, K. M. *et al.* Comparative RNA-sequencing analysis of myocardial and circulating small RNAs in human heart failure and their utility as biomarkers. *Proc Natl Acad Sci U S A* **111**, 11151-11156, doi:10.1073/pnas.1401724111 (2014).
55 Burke, M. A. *et al.* Molecular profiling of dilated cardiomyopathy that progresses to heart failure. *JCI Insight* **1**, doi:10.1172/jci.insight.86898 (2016).
56 Lacraz, G. P. A. *et al.* Tomo-seq Identifies SOX9 as a Key Regulator of Cardiac Fibrosis During Ischemic Injury. *Circulation*, doi:10.1161/CIRCULATIONAHA.117.027832 (2017).
57 Kruse, F., Junker, J. P., van Oudenaarden, A. & Bakkers, J. Tomo-seq: A method to obtain genome-wide expression data with spatial resolution. *Methods Cell Biol* **135**, 299-307, doi:10.1016/bs.mcb.2016.01.006 (2016).
58 DeLaughter, D. M. *et al.* Single-Cell Resolution of Temporal Gene Expression during Heart Development. *Dev Cell* **39**, 480-490, doi:10.1016/j.devcel.2016.10.001 (2016).
59 Brar, G. A. & Weissman, J. S. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat Rev Mol Cell Biol* **16**, 651-664, doi:10.1038/nrm4069 (2015).
60 Schafer, S. *et al.* Titin-truncating variants affect heart function in disease cohorts and the general population. *Nat Genet* **49**, 46-53, doi:10.1038/ng.3719 (2017).
61 Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43-49, doi:10.1038/nature09906 (2011).
62 Marsman, J. & Horsfield, J. A. Long distance relationships: enhancer-promoter communication and dynamic gene transcription. *Biochim Biophys Acta* **1819**, 1217-1227, doi:10.1016/j.bbagrm.2012.10.008 (2012).

63 Dao, L. T. M. *et al.* Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nat Genet* **49**, 1073-1081, doi:10.1038/ng.3884 (2017).

64 Diao, Y. *et al.* A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat Methods* **14**, 629-635, doi:10.1038/nmeth.4264 (2017).

65 Steinhauser, S., Kurzawa, N., Eils, R. & Herrmann, C. A comprehensive comparison of tools for differential ChIP-seq analysis. *Brief Bioinform* **17**, 953-966, doi:10.1093/bib/bbv110 (2016).

66 Goldstein, I. & Hager, G. L. Dynamic enhancer function in the chromatin context. *Wiley Interdiscip Rev Syst Biol Med*, doi:10.1002/wsbm.1390 (2017).

67 He, A., Kong, S. W., Ma, Q. & Pu, W. T. Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. *Proc Natl Acad Sci U S A* **108**, 5632-5637, doi:10.1073/pnas.1016959108 (2011).

68 Blow, M. J. *et al.* ChIP-Seq identification of weakly conserved heart enhancers. *Nat Genet* **42**, 806-810, doi:10.1038/ng.650 (2010).

69 Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931-21936, doi:10.1073/pnas.1016071107 (2010).

70 Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R. & Lieb, J. D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome research* **17**, 877-885, doi:10.1101/gr.5533506 (2007).

71 Meyer, C. A. & Liu, X. S. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature reviews. Genetics* **15**, 709-721, doi:10.1038/nrg3788 (2014).

72 Haitjema, S. *et al.* Additional Candidate Genes for Human Atherosclerotic Disease Identified Through Annotation Based on Chromatin Organization. *Circulation. Cardiovascular genetics* **10**, doi:10.1161/CIRCGENETICS.116.001664 (2017).

73 Zhang, Y. *et al.* Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature* **504**, 306-310, doi:10.1038/nature12716 (2013).

74 Stadhouders, R. *et al.* Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. *Nature protocols* **8**, 509-524, doi:10.1038/nprot.2013.018 (2013).

75 Mayer, G. *et al.* Omics-bioinformatics in the context of clinical data. *Methods Mol Biol* **719**, 479-497, doi:10.1007/978-1-61779-027-0_22 (2011).

76 Schneider, M. V. & Orchard, S. Omics technologies, data and bioinformatics principles. *Methods Mol Biol* **719**, 3-30, doi:10.1007/978-1-61779-027-0_1 (2011).

77 Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome Biol* **16**, 195, doi:10.1186/s13059-015-0762-6 (2015).

78 Mokry, M. *et al.* Many inflammatory bowel disease risk loci include regions that regulate gene expression in immune cells and the intestinal epithelium. *Gastroenterology* **146**, 1040-1047, doi:10.1053/j.gastro.2013.12.003 (2014).

79 Li, M. J., Yan, B., Sham, P. C. & Wang, J. Exploring the function of genetic variants in the non-coding genomic regions: approaches for identifying human regulatory variants affecting gene expression. *Brief Bioinform* **16**, 393-412, doi:10.1093/bib/bbu018 (2015).

80 Kamanu, F. K. *et al.* Mutations and binding sites of human transcription factors. *Front Genet* **3**, 100, doi:10.3389/fgene.2012.00100 (2012).

81 Arbiza, L. *et al.* Genome-wide inference of natural selection on human transcription factor binding sites. *Nat Genet* **45**, 723-729, doi:10.1038/ng.2658 (2013).

82 Quinn, J. J. & Chang, H. Y. Unique features of long non-coding RNA biogenesis and function. *Nature reviews. Genetics* **17**, 47-62, doi:10.1038/nrg.2015.10 (2015).

83 Jonas, S. & Izaurralde, E. Towards a molecular understanding of microRNA-mediated gene silencing. *Nature reviews. Genetics* **16**, 421-433, doi:10.1038/nrg3965 (2015).

84 Han, P. *et al.* A long noncoding RNA protects the heart from pathological hypertrophy. *Nature* **514**, 102-106, doi:10.1038/nature13596 (2014).

85 Ward, L. D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol* **30**, 1095-1106, doi:10.1038/nbt.2422 (2012).

86 Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum Mol Genet* **24**, R102-110, doi:10.1093/hmg/ddv259 (2015).

87 Smemo, S. *et al.* Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. *Human molecular genetics* **21**, 3255-3263, doi:10.1093/hmg/dds165 (2012).

88 Weedon, M. N. *et al.* Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nature genetics* **46**, 61-64, doi:10.1038/ng.2826 (2014).

89 Petit, F. *et al.* The disruption of a novel limb cis-regulatory element of SHH is associated with autosomal dominant preaxial polydactyly-hypertrichosis. *European journal of human genetics : EJHG* **24**, 37-43, doi:10.1038/ejhg.2015.53 (2016).

90 Scacheri, C. A. & Scacheri, P. C. Mutations in the noncoding genome. *Current opinion in pediatrics* **27**, 659-664, doi:10.1097/MOP.0000000000000283 (2015).

91 Yevshin, I., Sharipov, R., Valeev, T., Kel, A. & Kolpakov, F. GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic acids research* **45**, D61-D67, doi:10.1093/nar/gkw951 (2017).

92 van Rijsingen, I. A. *et al.* Outcome in phospholamban R14del carriers: results of a large multicentre cohort study. *Circulation. Cardiovascular genetics* **7**, 455-465, doi:10.1161/CIRCGENETICS.113.000374 (2014).

93 van Rijsingen, I. A. *et al.* Risk factors for malignant ventricular arrhythmias in lamin a/c mutation carriers a European cohort study. *J Am Coll Cardiol* **59**, 493-500, doi:10.1016/j.jacc.2011.08.078 (2012).

94 Ingles, J. *et al.* Compound and double mutations in patients with hypertrophic cardiomyopathy: implications for genetic testing and counselling. *J Med Genet* **42**, e59, doi:10.1136/jmg.2005.033886 (2005).

95 Weedon, M. N. *et al.* Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat Genet* **46**, 61-64, doi:10.1038/ng.2826 (2014).

96 Walsh, R. *et al.* Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genetics in medicine : official journal of the American College of Medical Genetics* **19**, 192-203, doi:10.1038/gim.2016.90 (2017).

97 van Tintelen, J. P., Wilde, A. A. & Jongbloed, J. D. Recurrent and founder mutations in inherited cardiac diseases in the Netherlands. *Netherlands heart journal : monthly journal of the Netherlands Society of Cardiology and the Netherlands Heart Foundation* **17**, 407-408 (2009).

98 Karakikes, I. *et al.* Correction of human phospholamban R14del mutation associated with cardiomyopathy using targeted nucleases and combination therapy. *Nature communications* **6**, 6955, doi:10.1038/ncomms7955 (2015).

99 Zhang, Y. *et al.* CRISPR-Cpf1 correction of muscular dystrophy mutations in human cardiomyocytes and mice. *Science advances* **3**, e1602814, doi:10.1126/sciadv.1602814 (2017).

100 Johansen, A. K. *et al.* Postnatal Cardiac Gene Editing Using CRISPR/Cas9 With AAV9-Mediated Delivery of Short Guide RNAs Results in Mosaic Gene Disruption. *Circulation research* **121**, 1168-1181, doi:10.1161/CIRCRESAHA.116.310370 (2017).

101 Nekrutenko, A. & Taylor, J. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature reviews. Genetics* **13**, 667-672, doi:10.1038/nrg3305 (2012).

102 Mittelstadt, B. D. & Floridi, L. The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. *Sci Eng Ethics* **22**, 303-341, doi:10.1007/s11948-015-9652-2 (2016).
103 Fulda, K. G. & Lykens, K. Ethical issues in predictive genetic testing: a public health perspective. *J Med Ethics* **32**, 143-147, doi:10.1136/jme.2004.010272 (2006).
104 Mokry, M., Harakalova, M., Asselbergs, F. W., de Bakker, P. I. & Nieuwenhuis, E. E. Extensive Association of Common Disease Variants with Regulatory Sequence. *PLoS One* **11**, e0165893, doi:10.1371/journal.pone.0165893 (2016).