Methods for comprehensive transcriptome analysis using next-generation sequencing and application in hypertrophic cardiomyopathy

Dissertation Advisor: Professors Jonathan and Christine Seidman Danos C Christodoulou

Methods for comprehensive transcriptome analysis using next-generation sequencing and application in hypertrophic cardiomyopathy

Abstract

Characterization of the RNA transcriptome by next-generation sequencing can produce an unprecedented yield of information that provides novel biologic insights. I describe four approaches for sequencing different aspects of the transcriptome and provide computational tools to analyze the resulting data. Methods that query the dynamic range of gene expression, low expressing transcripts, micro RNA levels, and start-site usage of transcripts are described.

Assessing changes in start-site usage can reveal major regulatory events that may be difficult to identify by standard gene expression analysis. By optimizing cDNA library construction steps to enhance for sequencing from 5' ends of transcripts, I developed a robust protocol to measure start-site usage of transcripts with high sensitivity. To identify genome-wide start-site usage changes between different biological specimens, I developed a computational approach that queries the distribution of reads at the 5' ends. This methodology is denoted as 5'RNAseq. Genome-wide 5'RNAseq of cardiac tissues from a mouse model of hypertrophic cardiomyopathy (HCM) identified Four-and-a-half LIM domains protein 1 (*Fhl1*) as the gene that exhibits the most marked change in start-site usage. Analysis of the specific cell populations in the heart revealed that the increased expression of *Fhl1* and change in 5’ regulation occurs in myocytes that become engulfed in fibrosis during disease progression. Further analyses of human ventricular specimens from subjects with a variety of cardiovascular pathologies revealed the identical *Fhl1* start-site switch occurs in both primary and secondary cardiomyopathies.

Genetic ablation of *Fhl1* in the mouse model of HCM resulted in markedly increased hypertrophy and histopathologic remodeling, indicating that *Fhl1* acts as an adaptive modifier of HCM. This result contrasts with data from an earlier report that suggest that Fhl1 attenuates hypertrophy caused by pressure overload, and suggests that *Fhl1* may be implicated in distinct responses to these different cardiac pathologies. As *Fhl1* is encoded on chromosome X, the potentially protective role of *Fhl1* in HCM may account for more severe clinical manifestations and outcomes in male patients with HCM and may provide novel therapeutic avenues to limit disease. More broadly, this work provides new tools that can be used to analyze transcriptomic information.

Table of contents

Acknowledgments vi

CHAPTER 1: Introduction 1

Inherited Cardiomyopathies 3

Hypertrophic Cardiomyopathy (HCM) and causal mutations 7

Cellular basis of HCM and non-sarcomere mutations causing HCM-like disease 11

The molecular and cellular hallmarks of HCM 13

Transcriptome analysis identifies pro-fibrotic molecules essential to the development of HCM 18

Processes impacted by a sarcomere mutation 20

Dissertation focus 26

CHAPTER 2: Quantification of Gene Transcripts with Deep Sequencing Analysis of Gene Expression (DSAGE) Using 1 to 2 µg Total RNA 36

CHAPTER 3: Construction of Normalized RNA-seq Libraries for Next-Generation Sequencing Using the Crab Duplex-Specific Nuclease 69

CHAPTER 4: Quantification of microRNA Expression with Next-Generation Sequencing 92

CHAPTER 5: Genome-wide transcriptional start-site analyses identifies Four-and-a-half LIM domains protein 1 as a modifier of hypertrophic cardiomyopathy 127

CHAPTER 6: Discussion 146

APPENDIX 1: Supplemental text and figures 157

APPENDIX 2: Published contributions 171

APPENDIX 3: Automatic download of data images from the UCSC browser 176

Acknowledgments

Graduate school was undoubtedly a very difficult trek, but it has been immensely rewarding. It has given me the chance to learn how to perform research in a setting characterized by great science and great people. I feel blessed for the opportunity to learn from these great mentors and peers and I treasure the opportunities that I had to contribute to the work and personal development of others.

I feel immensely indebted to my advisors, Professors Jonathan and Christine Seidman, who showed unbounded attention to my work and development, and provided me with an invaluable foundation for future work. The time I spent in their laboratories will be unforgettable and will forever guide my steps in the future. I am very grateful to Professor Rong Tian; the time that I spent in her laboratory was a true gift and my experiences there became an indispensable part of myself.

I would undoubtedly not have reached this point without a large multitude of people who kindly provided me with the essentials I needed to be able to graduate. This includes members of the Seidman and Tian laboratories, and I would like to particularly thank Hiroko Wakimoto and Georgios Karamanlides, as well as members of other labs and communities who played a fundamental role in my development. Only a very small number of these people are acknowledged by name in this section, in the separate chapters, or on the dissertation day.

I am very indebted to the Harvard and MIT communities for accepting me as a student and for the open research environment from which I benefited enormously despite starting with very naïve skills. My family has been a great support as well as a source of inspiration and courage. I would also like to acknowledge the contribution of

teachers and others in my country for providing me with early foundations which allowed me to pursue studies abroad.

I would like to thank everyone who has read, commented on and provided essential feedback for manuscript and dissertation write-up. I would like to thank my friends Jonathan Tejada, Joel Zhinzhong Yao, Brandon Gregory and Emil Apostolov for everything they have done for me in the past many years and for their continuous support and encouragement.

I would finally like to thank the world and all its beings for the unlimited opportunities for exploration, improvement and meaning.

CHAPTER 1

Chapter 1. Introduction:

Improving human health globally is one of the greatest achievable challenges of our time. A better understanding of human biology and disease enables development of medical treatments. Further improvement of public health relies on continued commitment to improving healthcare and strengthening of interdisciplinary research programs aiming to better understand human disease.

Accelerating advancements in biology and engineering offer new avenues to treat heart disease by illuminating the complex physiology of the heart, as well as elucidating the numerous causes of its failure. Despite improvements in care of heart disorders, heart disease remains the leading cause of death worldwide[1]. In the United States alone, heart disease was responsible for 600,000 deaths (approximately a quarter of total deaths) in 2011 [1]. Heart failure burdens sufferers with progressive disability and early death.

The heart is fundamentally a pump powered by the contraction of muscle tissue. The heart's defining function, the heartbeat, is felt throughout the entire body and is the life-defining characteristic of most animals. Survival requires that this muscle functions continuously without pause for the many decades of life. Cardiomyopathy is the dysfunction of this cardiac muscle tissue. Understanding the physiology of the heart and the multiple determinants of heart disease is key to finding new ways to effectively treat heart diseases.

Cardiomyopathies can be divided into two categories: primary and secondary. Primary cardiomyopathies are confined to the heart. Secondary cardiomyopathies are

[1] World Health Organization

due to a systemic disorder. While improved understanding of cardiac physiology has illuminated the general processes underlying most cardiomyopathies, the mechanistic basis of the most common primary cardiomyopathy, hypertrophic cardiomyopathy (HCM), remains elusive. This dissertation focuses on HCM, investigating its pathology and examining its relationship to other cardiomyopathies. Other cardiomyopathies are discussed in various extents throughout: discussion of HCM and primary cardiomyopathies is presented in Chapters 1, 5 and 6, whereas secondary cardiomyopathies are discussed in Chapters 5 and 6.

Inherited Cardiomyopathies

Inherited cardiomyopathies are marked by heterogeneity, in that the disease clinically manifests uniquely from person to person. Nevertheless, major clinical hallmarks can serve to classify cardiomyopathies into the following distinct conditions: hypertrophic cardiomyopathy (HCM), dilated cardiomyopathy (DCM), restrictive cardiomyopathy (RCM), arrthythmogenic right ventricular cardiomyopathy (ARVC), catecholaminergic polymorphic ventricular tachycardia (CPVT), and left ventricular noncompaction [2, 3, 4]. Precisely defining causal mutations will help to further our understanding of the genetics and pathophysiology of these diseases.

HCM is characterized by myocardial wall thickening, impaired relaxation of the muscle (diastolic dysfunction) and arrthythmias. It is a common disease, affecting 1:500 individuals [5, 6].

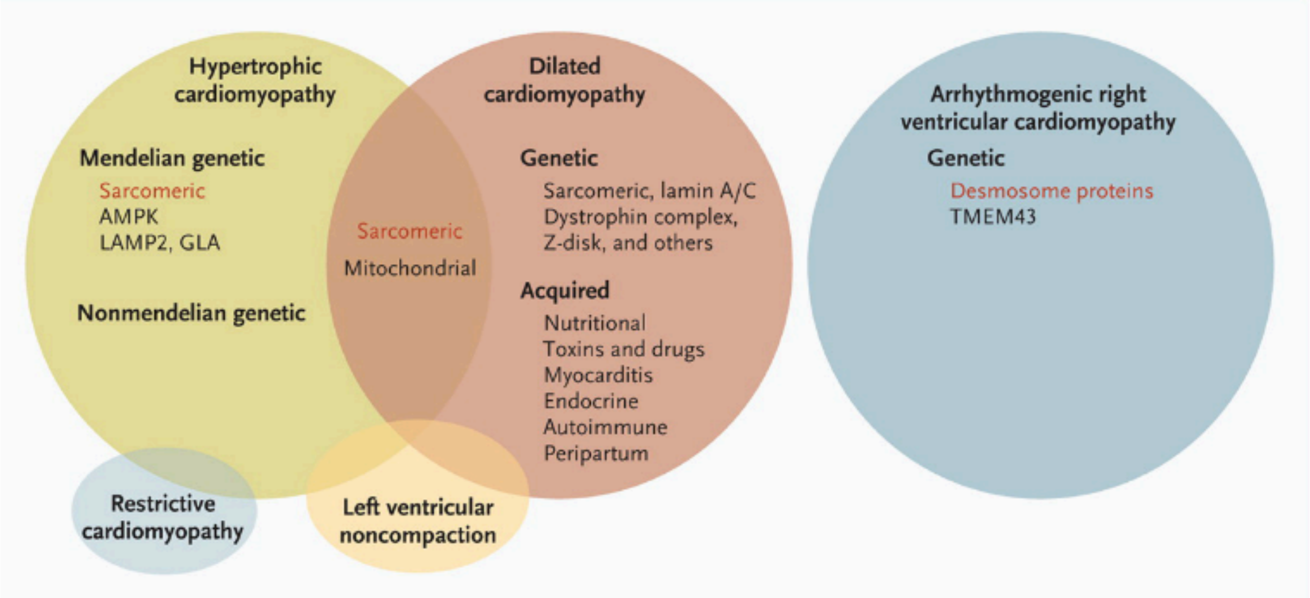


Figure 1.1. Overlap between genes with identified causal mutations for inherited cardiomyopathies. Reprinted from Watkins et al., 2011 [2] with permission from the copyright holder (Massachusetts Medical Society).

DCM is characterized by dysfunction during systole, dilatation and thinning of the myocardial wall, and a general impairment of the ability of the muscle to pump blood. Affecting at least 1:2500 individuals, DCM is familial in 20-35% of cases. Causal mutations have been identified in multiple cellular components for a considerable percentage of the cases [7].

RCM results from stiff myocardial walls, making the heart unable to stretch and fill with blood properly [8]. ARVC, on the other hand, is characterized by myocardial fibrofatty replacement and affects about 1:1200 individuals, with more than half of cases having familial inheritance. ARVC manifests with ventricular tachyarrthythmias and sudden cardiac death. The arrthythmias originate from the right ventricle. Causal mutations are found in desmosomes, proteins that hold cells together [9 - 13].

Among the less well-studied, CPVT arises from mutations in calcium handling proteins such as Ryr2 and CASQ2 and can result in ventricular syncope during exercise [14]. CPVT is a rare condition with unknown prevalence.

Left ventricular noncompaction originates from improper development of the myocardium and is characterized by a spongy left ventricular myocardium. Sarcomere gene mutations are causal for this disease [15, 16].

Delineating genotype-phenotype relationships can be challenging and cannot currently be done based on the amino-acid sequence change alone. Using molecular genetics, defining associations between mutation and clinical status drives our investigations to identify disease-causing mutations [17 - 20]. To complicate our understanding further, different mutations in the same gene can have different clinical manifestations. For example, mutations in the myosin-heavy chain can cause either HCM or DCM as in the instances of Arg403Gln (HCM), Arg719Trp (HCM), Ser532Pro (DCM) and Phe764Leu (DCM) [2, 21]. In addition, the severity of the disease can vary depending on the exact mutation. The resulting phenotypes are thought to constitute a continuum, sharing symptoms and histologic hallmarks [22].

The relationship among the various cardiomyopathies with respect to the genes in which causal mutations are found is shown (Figure 1.1). Cardiomyopathies overlap with respect to the genes where causal mutations are found and in terms of the symptoms associated with each of them. For example, HCM and DCM share similar histopathologic findings such as fibrosis and myocyte death, although to different degrees. However, the genetic causes of DCM are currently described as more diverse than those of HCM and include mutations in cellular compartments such as the cytoskeleton, nuclear envelope, calcium handling proteins, and elements involved in mechanotransduction [23].

In addition, mutations in a major sarcomere protein component, Titin, are now linked to DCM but not HCM. Next-generation sequencing has enabled the determination of causal mutations in Titin. In particular, a subgenome capture approach [24] was used to sequence Titin from DNA from multiple patients with HCM and DCM, in addition to normal controls. This was a considerable technological challenge, as Titin, is 33,000 amino acids in length, spanning the entire sarcomere. The frequency of Titin mutations in patient DNA was 54/203 for DCM, 3/231 for HCM and 7/249 for normal controls [25] (Figure 1.2). This indicates that Titin mutations are causal for DCM and that Titin mutations may constitute a minor proportion of all HCM-causing mutations. Studies with much larger cohorts may eventually lead to identification of specific Titin mutations as factors for HCM. Alternatively, it may eventually be shown that alterations in Titin predispose the patient to develop only DCM. One observation from the overall data discussed thus far is that a particular mutation tends to exhibit a well-defined clinical course, despite the observed heterogeneity in its expression (with the caveat that when evaluating large cohorts, genotyping typically followed selection of patients).

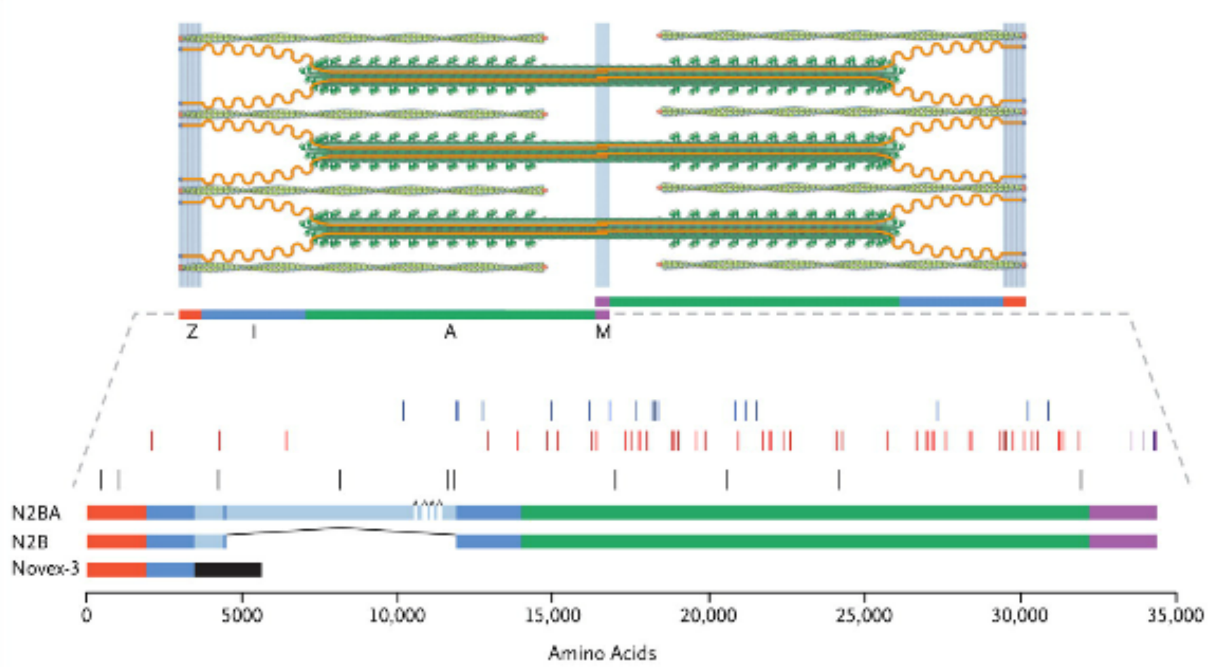


Figure 1.2. Distribution of identified Titin mutations on the protein shown as it spans from the Z line to the M line of the sarcomere. Red or blue: Location of mutations in DCM patients. Black: Location of mutations in HCM patients and healthy controls. Reprinted from Herman et al., 2012 [25] with permission from the copyright holder (Massachusetts Medical Society).

Hypertrophic Cardiomyopathy (HCM) and causal mutations

HCM is an inherited cardiomyopathy, with 50% of cases having a determined genetic etiology. Sarcomere genes with HCM associated mutations are shown (Figure 1.3). Mutations in 23 sarcomere and sarcomere-associated proteins are believed to cause HCM [22]. Interestingly, HCM appears to be a genetic disease caused by very defined mutations found mainly in two genes, MYH7 and MYBPC3, that account for ~3/4 of identified genetic causes. Mutations in 8 genes depicted (Figure 1.4) are definitively linked with pathology in patient cohorts and they have autosomal dominant inheritance [26, 27].

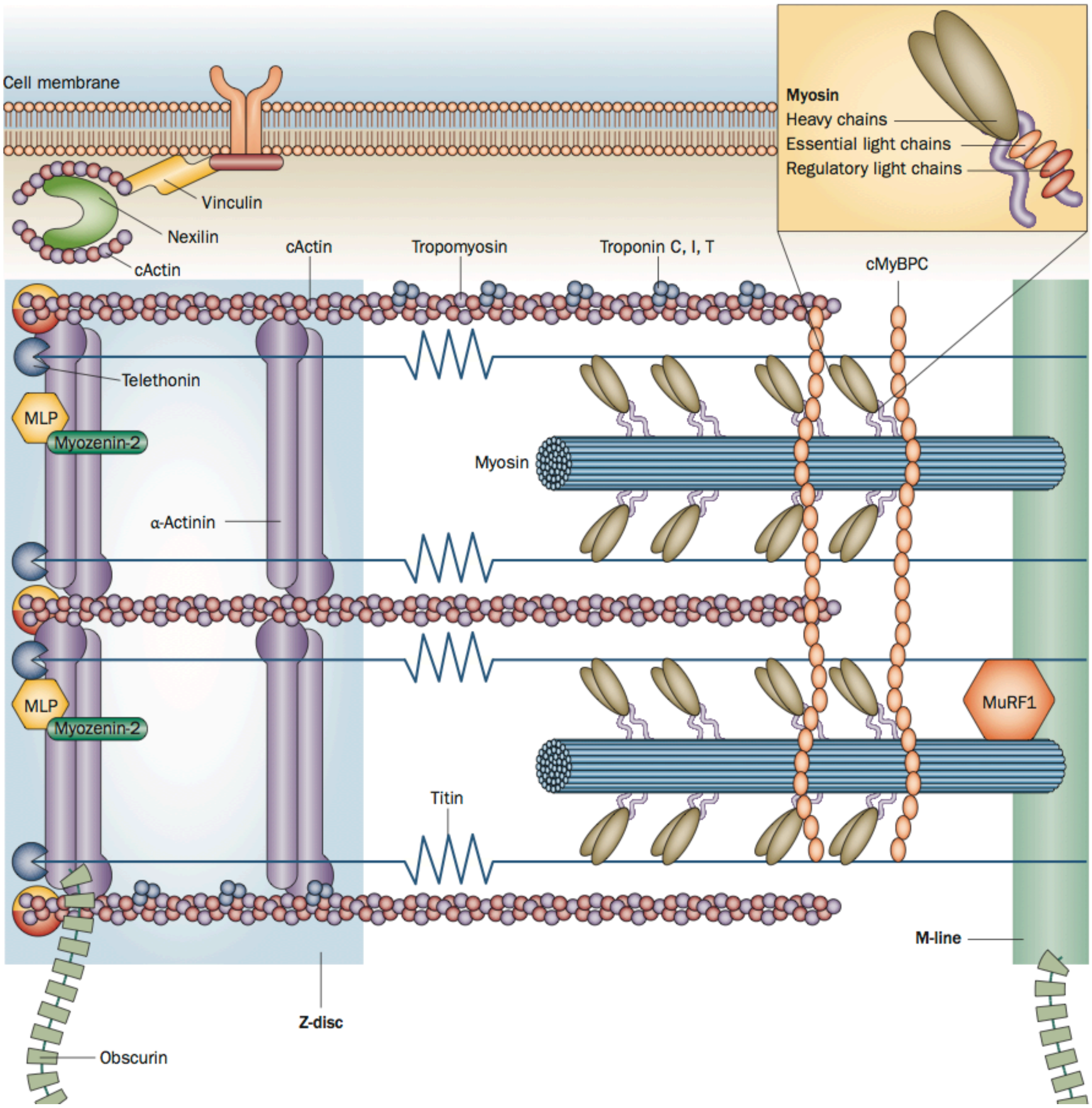


Figure 1.3. Depiction of the components of the sarcomere. A high magnification of the myosin and its components is shown on the upper right. Mutations in these sarcomere components are associated HCM as well as other primary cardiomyopathies. Reprinted from Frey et al., 2012 [22] with permission from the copyright holder.

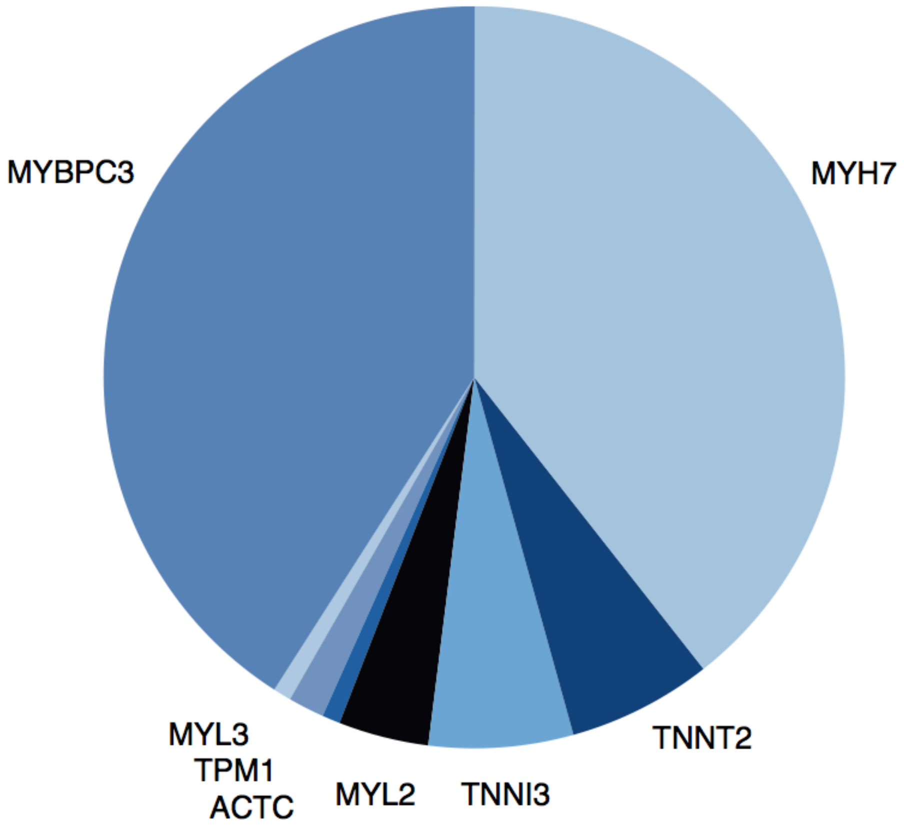


Figure 1.4. Distribution of HCM disease causing mutations in sarcomere genes. Mutations in these sarcomere genes are definitively linked with HCM. Reprinted from Marston, 2011 [26] with permission from the copyright holder.

The roles of the proteins with mutations that definitively cause or possibly cause HCM and are used for diagnostic testing (Partners Laboratory for Molecular Medicine) are briefly described below[2] [28, 29, 30]). These genes include energy production genes and genes responsible for organelle function, which cause HCM-like disease with profound cellular hallmarks.

[2] Information retrieved from the National Center for Biotechnology Information (NCBI) or associated references.

ACTC1	cardiac actin	one of the basic components of the sarcomere [31]
ACTN2	actinin	anchors actin filaments [32, 33]
CSRP3	cysteine and glycine-rich protein 3	mechanotrasduction at the Z-disk through a LIM domain [34]
GLA	galactosidase alpha	lysosome protein with mutations causing Fabry disease and late onset cardiac hypertrophy
LAMP2	lysosomal-associated membrane protein 2	lysosome protein with roles in autophagy causing severe hypertrophy presenting like HCM in young males
MYBPC3	cardiac myosin binding protein-C	associates myosin and modulates cardiac contraction [35]
MYH7	myosin-heavy chain beta	constitutes the motor of the sarcomere by binding and moving along the actin filament. Possesses ATP hydrolysis and actin binding domains
MYL2, MYL3	myosin regulatory/essential light chains	trigger contraction in response to calcium mediated phosphorylation.
MYOZ2	myozenin 2	localizes at the Z disk. Mutations in this protein may interfere with Z disk function [36]
NEXN	nexilin	actin binding protein
PLN	phospholamban	regulates calcium levels by modulating the sarco/endoplasmic reticulum Calcium – ATPase channel (abbreciated as SERCA)
PRKAG2	AMP-activated protein kinase subunit gamma-2	energy sensitive molecule which when mutated causes a glycogen storage cardiac hypertrophy
TNNC1	cardiac troponin C	subunit of troponin that when bound to calcium relieves inhibition of contraction [37 – 39]
TNNI3	troponin I type 3	inhibitory subunit of troponin blocking actin myosin interaction in the absence of calcium which accounts for the resting state
TNNT2	cardiac troponin T	part of the troponin complex, it stereochemically regulates tropomyosin inhibition of contraction [40, 41]
TPM1	tropomyosin	binds to actin and prevents myosin from binding (in the resting phase) locked position by troponins. This

		inhibition is relieved in the presence of calcium [42]
TTR	transthyretin	transports thyroxine, retinol (vitamin A), drugs etc.
ANKRD1	ankyrin repeat domain 1	interacts with Titin and may act as a transcription factor [43]
BAG3	BCL2-associated athanogene 3	associates with the sarcomere at the Z-disk location
CAV3	caveolin 3	scaffolding protein
LDB3	LIM domain binding 3	associates with the sarcomere at the Z-disk location, possibly providing stabilization
MYH6	alpha cardiac myosin heavy chain	the relative levels between alpha and beta myosin heavy chains fluctuate in development and disease [44]
MYLK2	myosin light chain kinase 2	calcium dependent protein
RYR2	ryanodine receptor 2	essential component of SERCA, important for regulation of calcium levels
TCAP	titin-cap	binds to titin and is important for sarcomere assembly
TTN	titin	large, essential cardiac protein provides passive force and active contractile force
VCL	vinculin	interacts with actin and helps anchor actin to intercalated discs [45, 46]

Cellular basis of HCM and non-sarcomere mutations causing HCM-like disease

HCM is characterized by hypertrophy of the cardiac tissue (including at times the right ventricle), myocyte disarray, fibrosis, and enhanced contractility of the sarcomere due to mutations and symptoms such as shortness of breath, angina and palpitations. Also, HCM can be asymptomatic in many disease carriers even with overt disease. Patients are at risk for sudden cardiac death, heart failure, and stroke. Change in the contractile function occurs early in the disease [47].

HCM-causing mutations in several sarcomere proteins such as MyBP-C, tropomyosin, troponin T, troponin I and troponin C can alter calcium sensitivity and lead to calcium accumulation during diastole. Increase of the affinity of troponin C with calcium can lead to a more gradual release of calcium that can persist during diastole. Increased affinity to calcium and higher than normal calcium levels during diastole can lead to an impairment of the cardiac muscle to properly relax [2].

Dysregulation of energy homeostasis can occur due to changes in contractile properties which can then lead to hypertrophic signaling. This is because of the very high energetic demands of the sarcomere which when altered can consume more ATP. In physiological state, ATP consumption by the sarcomere accounts for 70% of total ATP in the myocyte. Depletion of energy can also lead to deficient function of ion-pumps (such as the sarcoendoplasmic reticulum Ca2+ ATPase, SERCA) thus further contributing to calcium accumulation and predisposition to arrhythmias. Thus, increase in energy demand by a mutant sarcomere is thought to be a very prominent mechanism leading to pathology. Metabolic changes can also affect signaling leading to an overall change away from the physiological transcriptional state of the cell. In addition, mitochondrial disorders can lead to an HCM-like manifestation, further reinforcing the notion that such energy imbalances contribute to hypertrophic signaling [2, 22, 48 - 52].

Mutations in the energy sensing subunit of AMPK, Prkag2, and the lysosomal proteins Lamp2 and Gla present with unique hallmarks and left ventricular hypertrophy that is distinct from HCM. Prkag2 mutations lead to glycogen accumulation in the myocytes with electrophysiological abnormalities, in addition to hypertrophy. Lamp2 mutations cause massive cardiac hypertrophy in males, with autophagic vacuoles present at the histologic level. Gla mutations in the lysosomal galactosidase gene cause cardiac hypertrophy with other systemic effects. Mutant Lamp2 caused HCM typically manifests in adolescence, with a rapid progression to heart failure, while HCM-like disease from mutated Gla has a late onset [53-56].

The molecular and cellular hallmarks of HCM

The first identified genetic cause of HCM was an Arginine to Glutamine change at residue 403 (Arg403Gln) of the myosin heavy chain. This was accomplished by cloning, fine restriction enzyme site mapping and nucleotide sequence analysis of a family with familial HCM [57, 58] (Figure 1.5). This mutation is mapped to the globular head of the myosin near the ATPase site (Figure 1.6). It is noteworthy that a mutation in the adjacent residue 908 (just 15 nm further away from the motor domain) causes a milder form of HCM [59]. The functional location of the mutation as it relates to critical components of the myosin (Figure 1.7) may be directly associated with severity of the disease [60]. For example, the Arg403Gln mutation causes a severe, high-penetrant disease with 50% survival at 45 years [61].

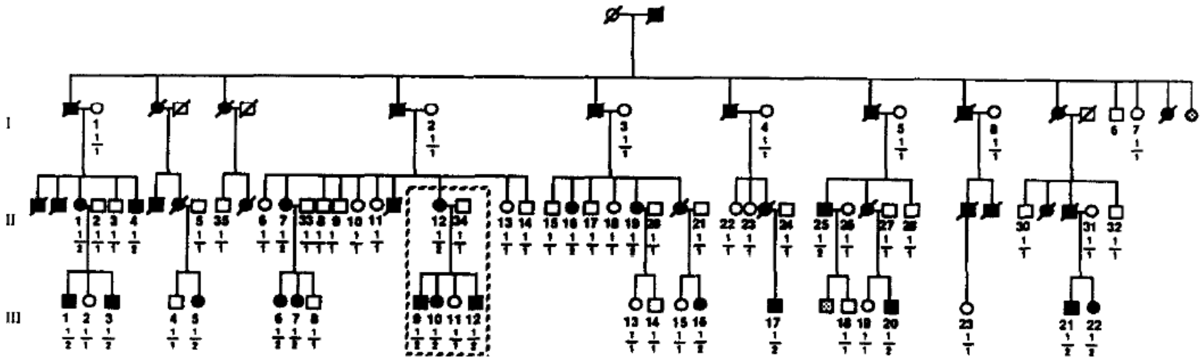


Figure 1.5. HCM co-segregating with Arg403Gln mutation with high penetrance. Squares: males. Circles: females. Filled: affected. Slash: deceased. 1/1: normal allele. ½: Arg403Gln. Reprinted from Geisterfer-Lowrance et al., 1990 [57] with permission from the copyright holder.

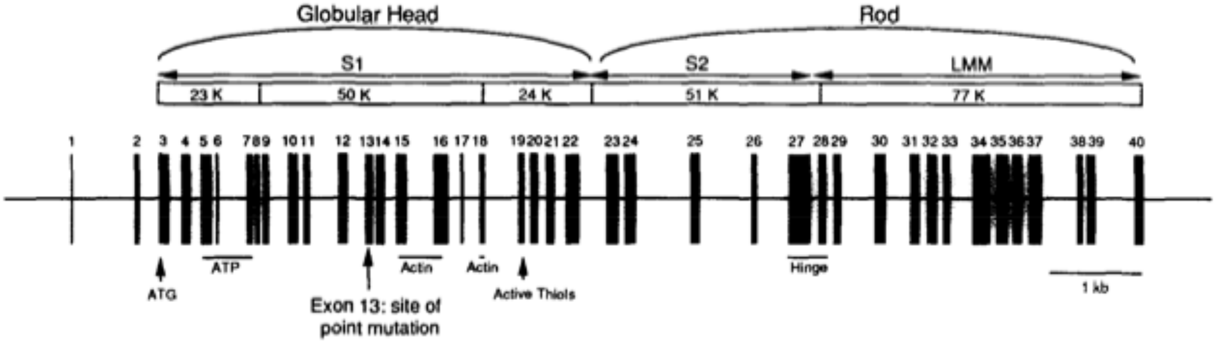


Figure 1.6. Exon composition of the globular head and rod region of myosin. The Arg403Gln mutation occurs at exon 13 in the globular head and is shown with an arrow. Reprinted from Geisterfer-Lowrance et al., 1990 [57] with permission from the copyright holder.

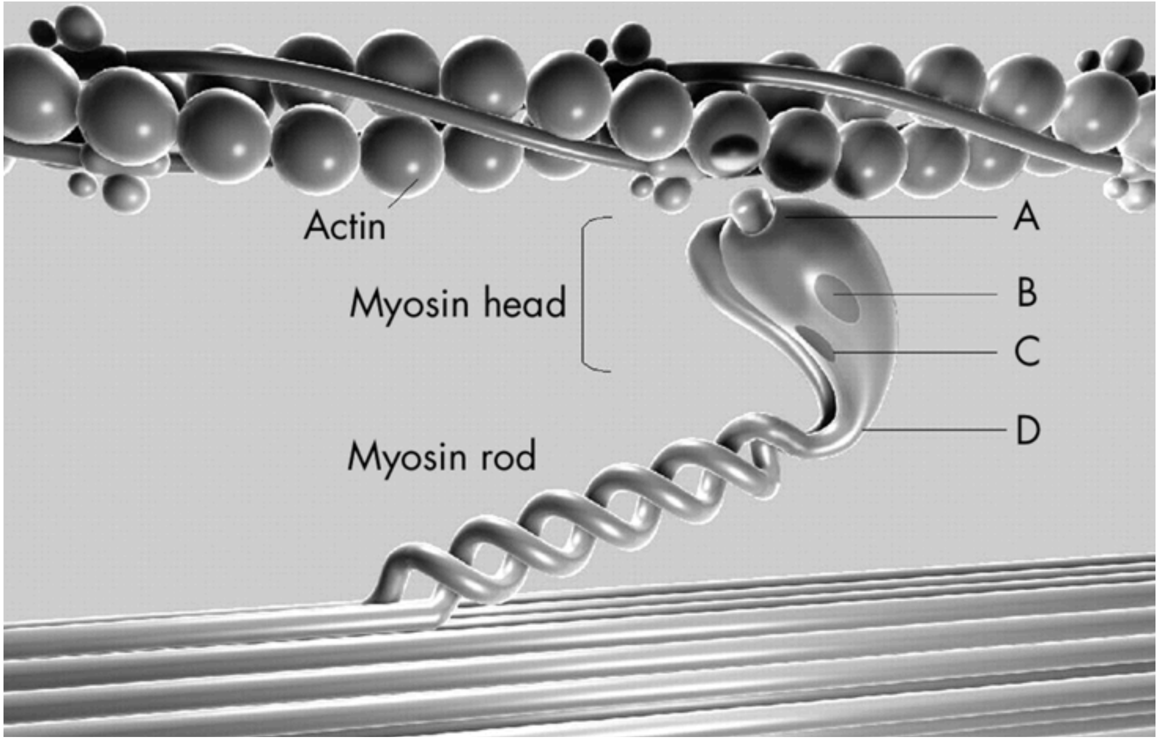


Figure 1.7. Depiction of myosin interacting with actin. Functional regions of the myosin head are detailed with A-D. A. Actin-binding region. B. Region of ATP hydrolysis. C. Myl3 binding region. D. Head-rod junction. Reprinted from Woo et al., 2003 [60] with permission from the copyright holder.

Treatments for HCM aim to manage symptoms and prevent sudden cardiac death and surgical therapies aim to relieve left ventricular outflow obstruction. β-adrenergic-receptor blockers, calcium-channel antagonists, or disopyramide are typically used. This treatment aims to lower contractility, decrease the heart rate and increase the duration of diastole. Diuretics can be used to relieve congestion. Surgical septal myectomy or alcohol ablation are the gold standard for symptom relief for individuals not responding to drug therapy when the left outflow tract is obstructed (Hypertrophic obstructive cardiomyopathy); this operation is associated with a high success rate and long term symptom relief. Implantable defibrillators are effective at reducing the risk of sudden cardiac death in those with severe arrhythmias [62-67]. Further work to improve the efficacy of available treatments is warranted.

Recapitulating HCM in animal models was an important achievement toward understanding the molecular and process determinants of the disease. The first mouse model of HCM was generated by introducing the Arg403Gln mutation in the respective mouse myosin heavy chain. Left atrial enlargement was observed in these mice (at 15 weeks), followed by cardiac hypertrophy and all classic histopathology landmarks, such as hypertrophied myocytes with large hyperchromatic nuclei, myocyte disarray, and fibrosis. Histology of these mice is shown (Figure 1.8). Also mirroring human symptoms, these mice were less tolerant to exercise than wild-type mice. In addition, females showed a slower progression to disease compared to males from the same inbred genetic background, providing evidence for genetically derived disease modifying processes, meaning that expressed differences between males and females can only account for the differences in disease expression. Such processes were postulated to potentially account for the observations that some human carriers died before reaching adulthood, while others had normal lifespans [68].

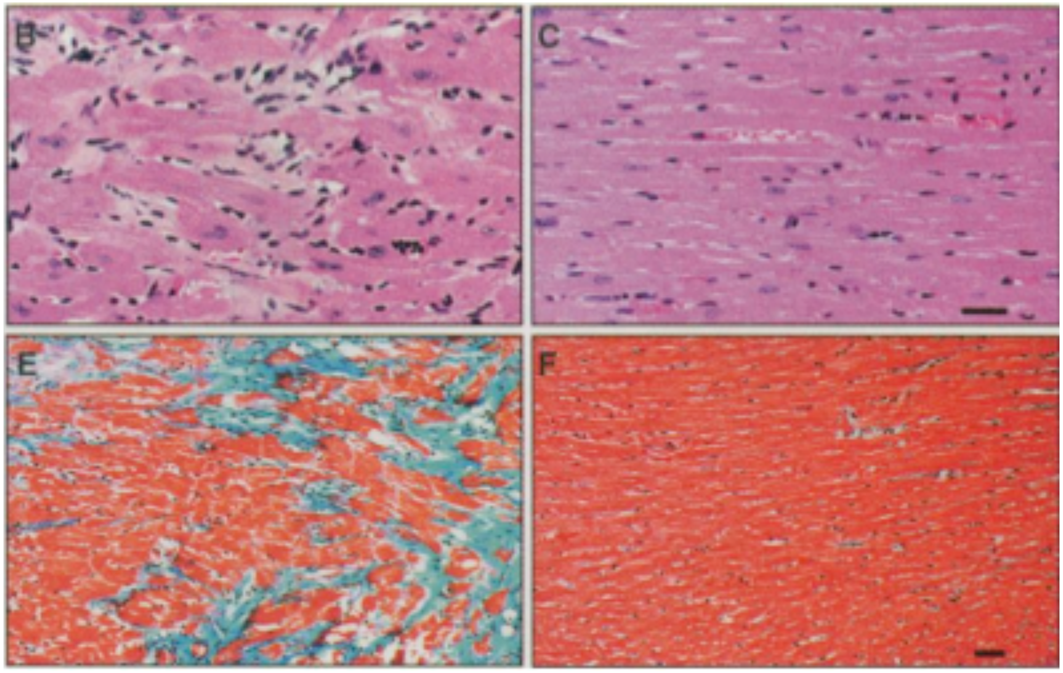


Figure 1.8. Myocyte disarray and fibrosis in Arg403Gln mice recapitulating human HCM histology. Staining: top panels, H&E; lower panels, masson's trichrome (staining fibrotic regions blue). (Left) Mice carriers of the sarcomere mutation. (Right) Wild-type controls. Reprinted from Geisterfer-Lowrance et al., 1996 [68] with permission from AAAS.

Fibrosis, which is an important histological landmark in HCM, can be seen in patients histologically or with magnetic resonance imaging (MRI). In human carriers without or with overt hypertrophy, fibrosis is detected using gadolinium as a contrast agent. In addition to MRI showing thickening of the intraventricular septum (IVS), foci of fibrosis are observed (Figure 1.9). Propeptide of type I procollagen (PICP) precedes overt hypertrophy or MRI findings [69].

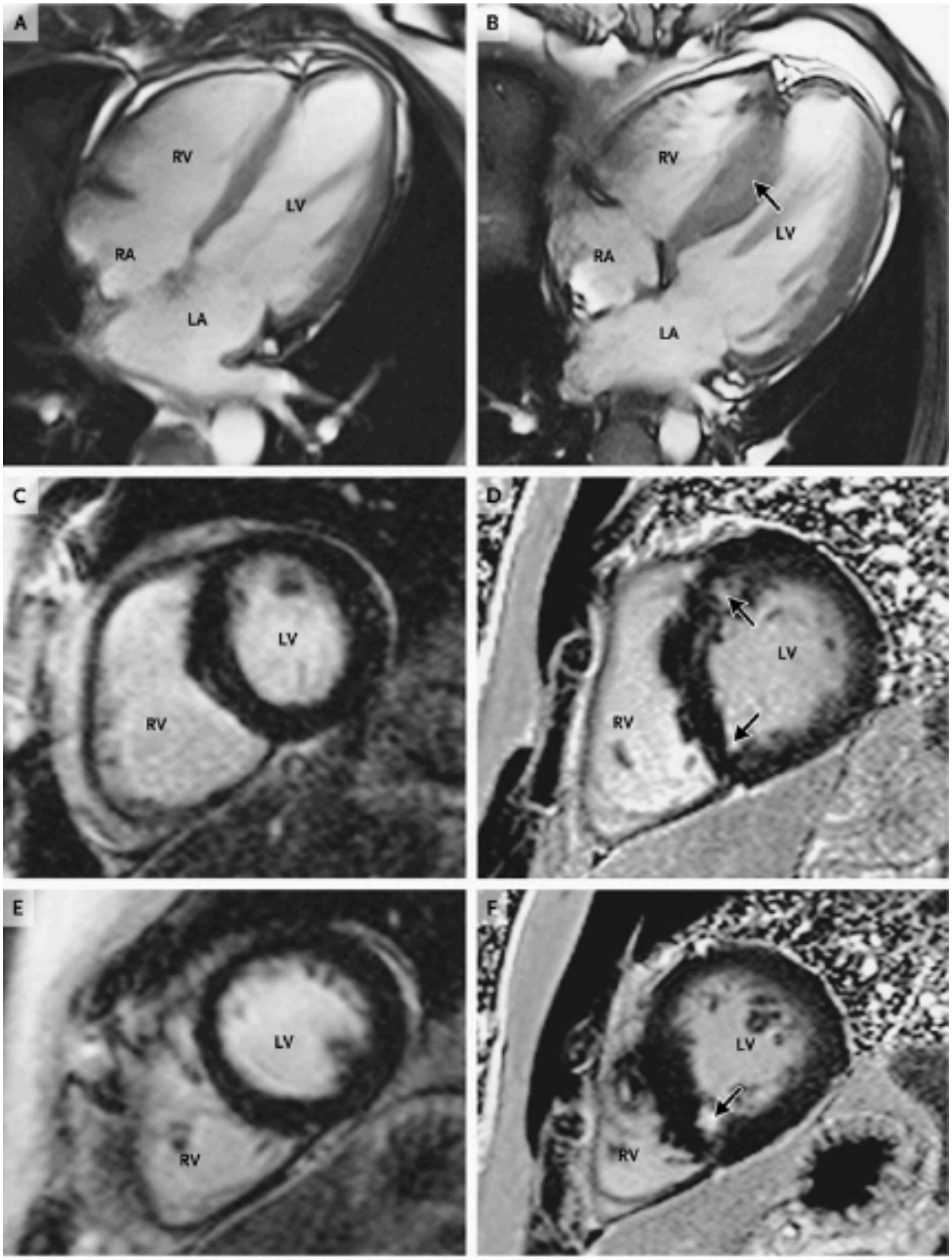


Figure 1.9. Magnetic resonance imaging can detect hallmarks of HCM in carriers of disease causing mutations. (Left panels) A carrier without overt hypertrophy. (Right panels) A carrier with overt hypertrophy. Reprinted from Ho et al., 2010 [69] with permission from the copyright holder (Massachusetts Medical Society).

Transcriptome analysis identifies pro-fibrotic molecules essential to the development of HCM

Early analysis of the transcriptome with high-throughput sequencing was performed in mice carriers of the Arg403Gln mutation. This new technology enabled a digital analysis of genes that changed expression in HCM (Figure 1.10). Of particular interest, this analysis identified specific profibrotic markers that were elevated in hearts of pre-hypertrophic mice, pointing to specific pathways altered in HCM such as the Transforming growth factor beta 1 (TGF-β1) pathway [70].

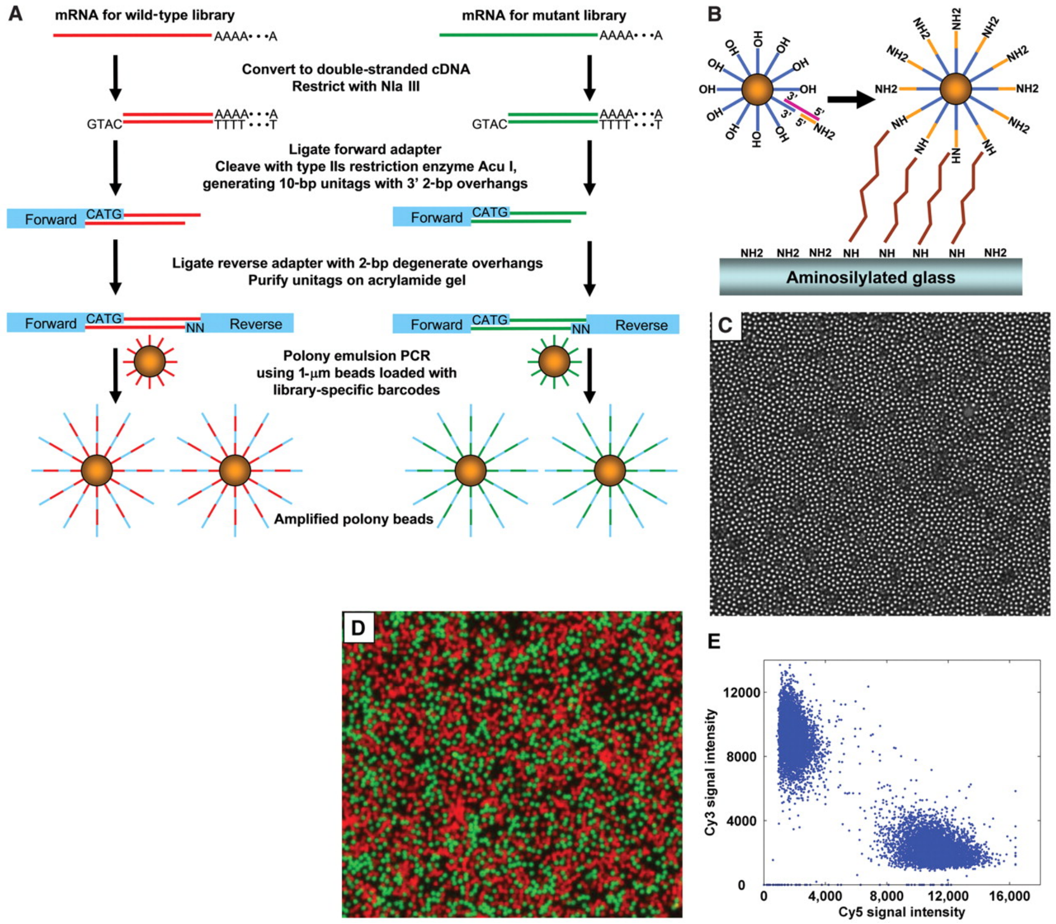

A
mRNA for wild-type library
mRNA for mutant library
AAAA• • •A
Convert to double-stranded cDNA
Restrict with Nla III
GTAC
AAAA• • •A
TTTT• • •T
Ligate forward adapter
Cleave with type IIs restriction enzyme Acu I,
generating 10-bp unitags with 3' 2-bp overhangs
Forward
CATG
Ligate reverse adapter with 2-bp degenerate overhangs
Purify unitags on acrylamide gel
NN
Reverse
Polony emulsion PCR
using 1-μm beads loaded with
library-specific barcodes
Amplified polony beads
B
OH
NH2
NH
3'
5'
Aminosilylated glass
C
D
E
Cy3 signal intensity
12000
8000
4000
0
0
4,000
8,000
12,000
16,000
Cy5 signal intensity

Figure 1.10. A high throughput sequencing approach utilizing parallel sequencing of cDNA tags is shown. A. Molecular biology steps to prepare and amplify cDNA tags amplified on polony beads. B-E. Sequencing of tags on glass by processing of fluorophore clusters. Application of this method identified the early fibrotic processes in HCM. Reprinted from Kim et al., 2007 [70] with permission from the copyright holder.

Experimental follow-up verified active proliferation of non-myocyte cells in HCM and allowed for testing of these pro-fibrotic factors. Interestingly, periostin, TGF-β1, and associated processes inhibited by losartan (an angiotensin inhibitor which is involved in TGF-β1 signaling) are essential for the development of HCM (Figure 1.11). This showed that targeting early prehypertrophic processes in non-myocytes identified by transcriptional analysis resulted in abrogation of the hypertrophic response from the sarcomere mutation [71].

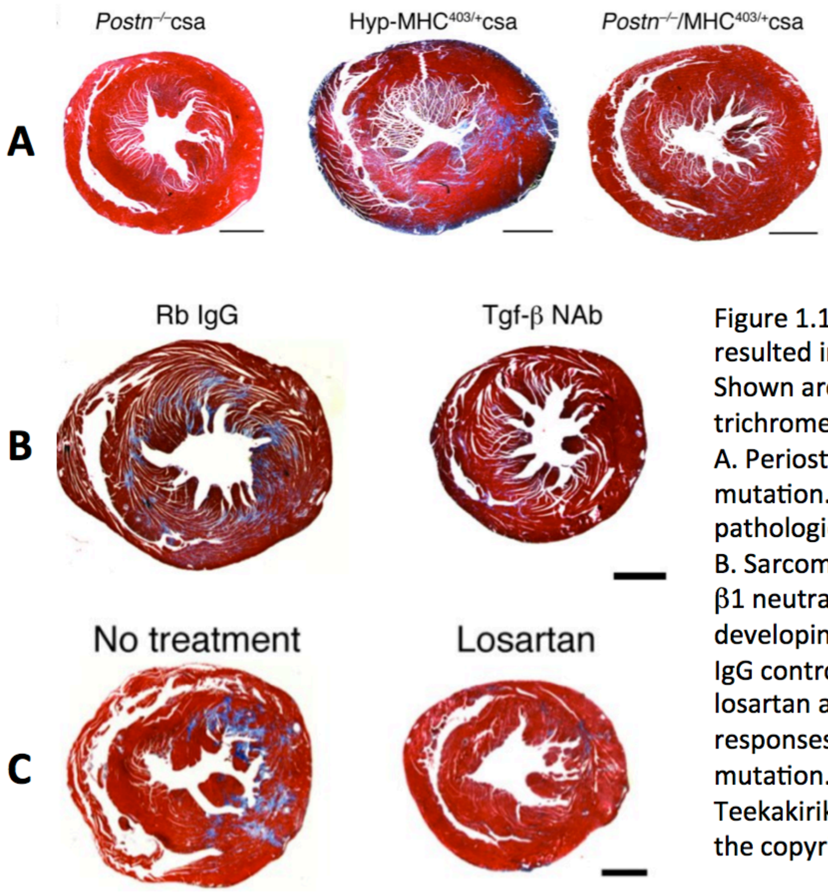


Figure 1.11. Targeting non-myocyte processes resulted in an abrogated hypertrophic response. Shown are sections stained with Masson's trichrome which colors blue the fibrotic regions. A. Periostin absence in the context of the causal mutation. Absence of periostin diminishes the pathologic response from the sarcomere mutation. B. Sarcomere mutation carriers treated with TGF-β1 neutralizing antibodies are protected from developing fibrosis or hypertrophy (compared to IgG control treatment). C. Administration of losartan also diminishes fibrotic or hypertrophy responses in the presence of the HCM causing mutation. Scale bars:1mm. Reprinted from Teekakirikul et al., 2010 [71] in accordance with the copyright.

Processes impacted by a sarcomere mutation

Models incorporating these findings provide a depiction of the critical pathways responsible for disease development arising from the initial mutation. Biomechanical stress sensing, followed by calcium sensitization, disturbed energy homeostasis, and downstream pathologic responses are shown as potential targets for drugs to ameliorate or prevent the onset of HCM (Figure 1.12). Calcium desensitizers (such as blebbistatin) or calcium channel inhibitors (such as L-type calcium channel antagonists) can target early disease processes in the myocytes (calcium blockers are already prescribed for alleviating the symptoms of HCM). Drugs that augment energy production such as perhexiline, which shifts energy production from fatty acids to glucose with higher ATP yield, utilize the notion that energy depletion is an important mediator. Metalocorticoid antagonists, statins, angiotensin II antagonists, and TGF-β1 inhibitors may prevent pathologic remodeling [22].

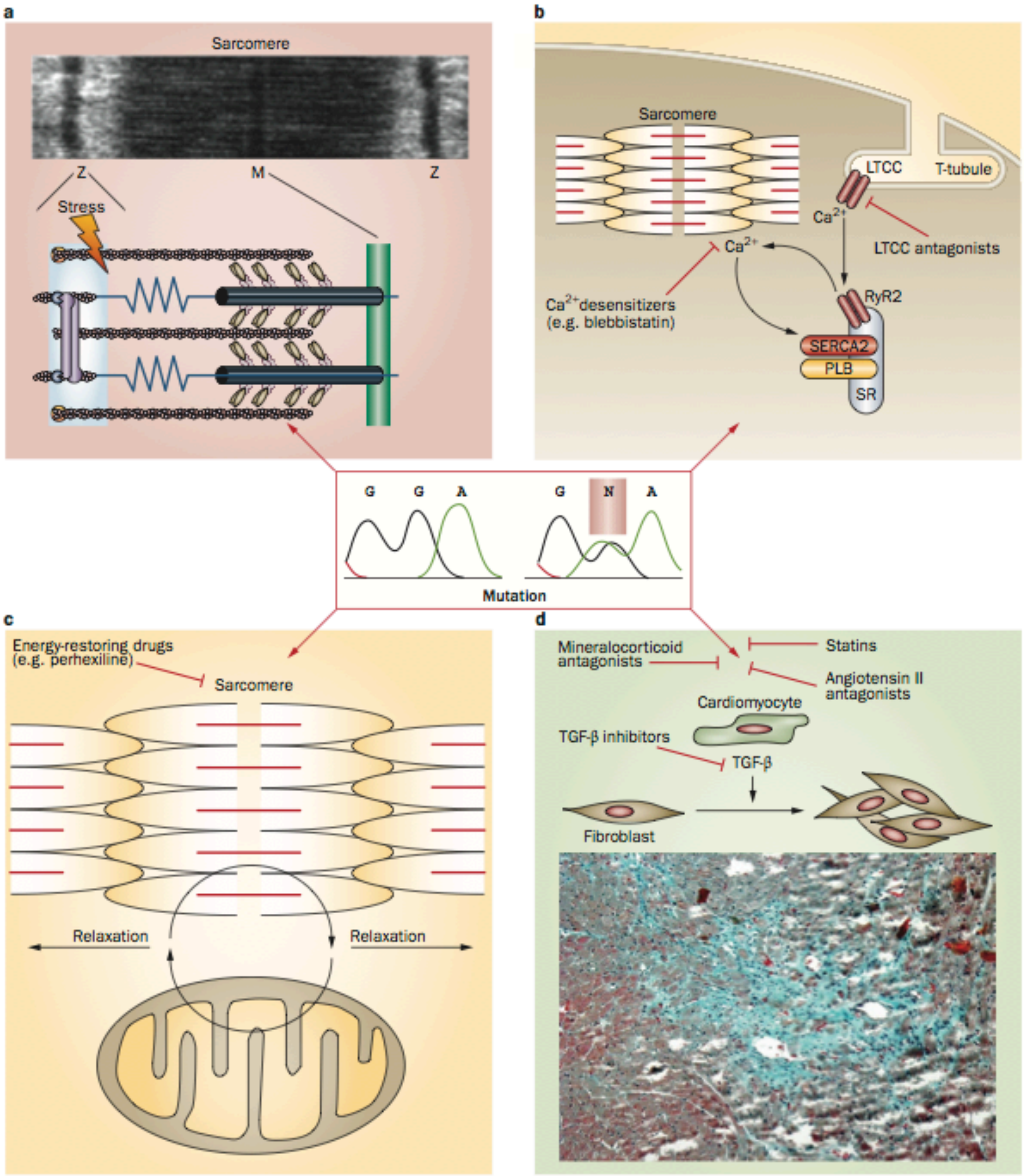


Figure 1.12. Depiction of processes and pathways responsible for HCM that can be potentially pharmacologically targeted. Reprinted from Frey et al., 2012 [22] with permission from the copyright holder.

A similar model depicting relationships between myocytes and fibroblasts is shown (Figure 1.13). Changes in ATP utilization, sliding velocity, force generation and calcium sensitivity are shown occurring at the myocyte level. Potential early and later interventions are shown. Stressed myocytes are shown to transduce signals to quiescent non-myocytes, causing the fibrotic response. This signal can be chemical (a secreted molecule) or physical (e.g. force transmission). TGF-β1 which is a proven essential mediator of the pathology is shown as a potential pharmacological target. Non-myocytes are shown engulfing myocytes, possibly contributing to myocyte death. Additional signaling or intercellular communication may add to overall myocyte stress [72].

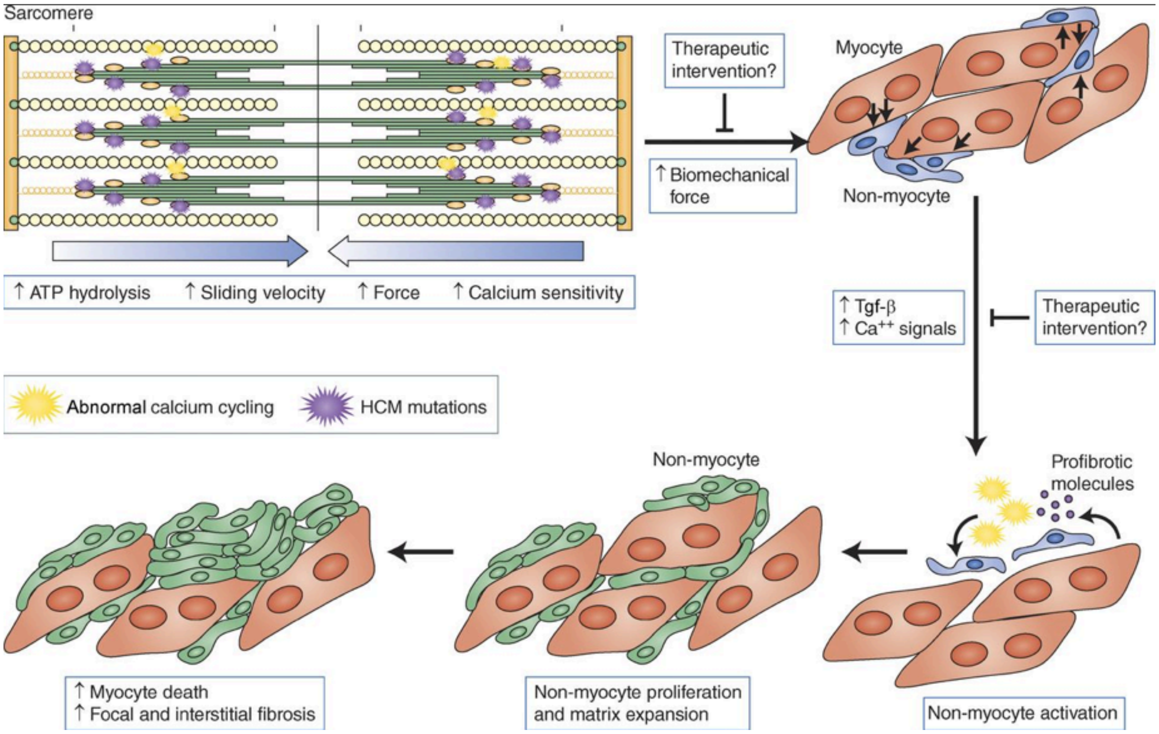


Figure 1.13. Myocyte and non-myocyte interactions are shown as crucial parts in the development of the pathology in HCM with opportunities for pharmacology intervention. Reprinted from Teekakirikul et al., 2012 [72] in accordance with the copyright.

Although myocytes account for the largest fraction of the heart by volume, non-myocytes may exceed myocytes in number. Non-myocytes present in the heart include fibroblasts, smooth muscle, endothelial cells and blood cells. Of particular importance is the cardiac fibroblast which has significant roles in cardiac physiology and disease. The cardiac fibroblast and its diverse functional roles in the heart and disease are shown (Figure 1.14). Physiologically, fibroblasts secrete and regulate collagen, which provides tensile strength and is essential for cardiac function. Fibroblasts are also recruited in areas of myocardial scaring providing structural integrity at those locations. Such broad roles can afford a high complexity in its function. Fibroblasts have roles in regulating physiological myocardial function and are essential for cardiac remodeling in HCM, left ventricular hypertrophy, myocardial infarction and heart failure. By secreting growth factors, they can facilitate wound healing. Such response processes are thought to contribute in maladaptive remodeling [73, 74].

In addition, a myofibroblast phenotype arising possibly because the fibroblast cell expresses contractile proteins exhibits potentially crucial effects. This cell has migratory and proliferative properties. Myofibroblasts are responsive to diverse, systemic signals including cytokines, vasoactive peptides, and hormones. Myofibroblasts can also secrete metalloproteases, which degrade the collagen, thus they directly regulate collagen turnover and matrix consistency [73].

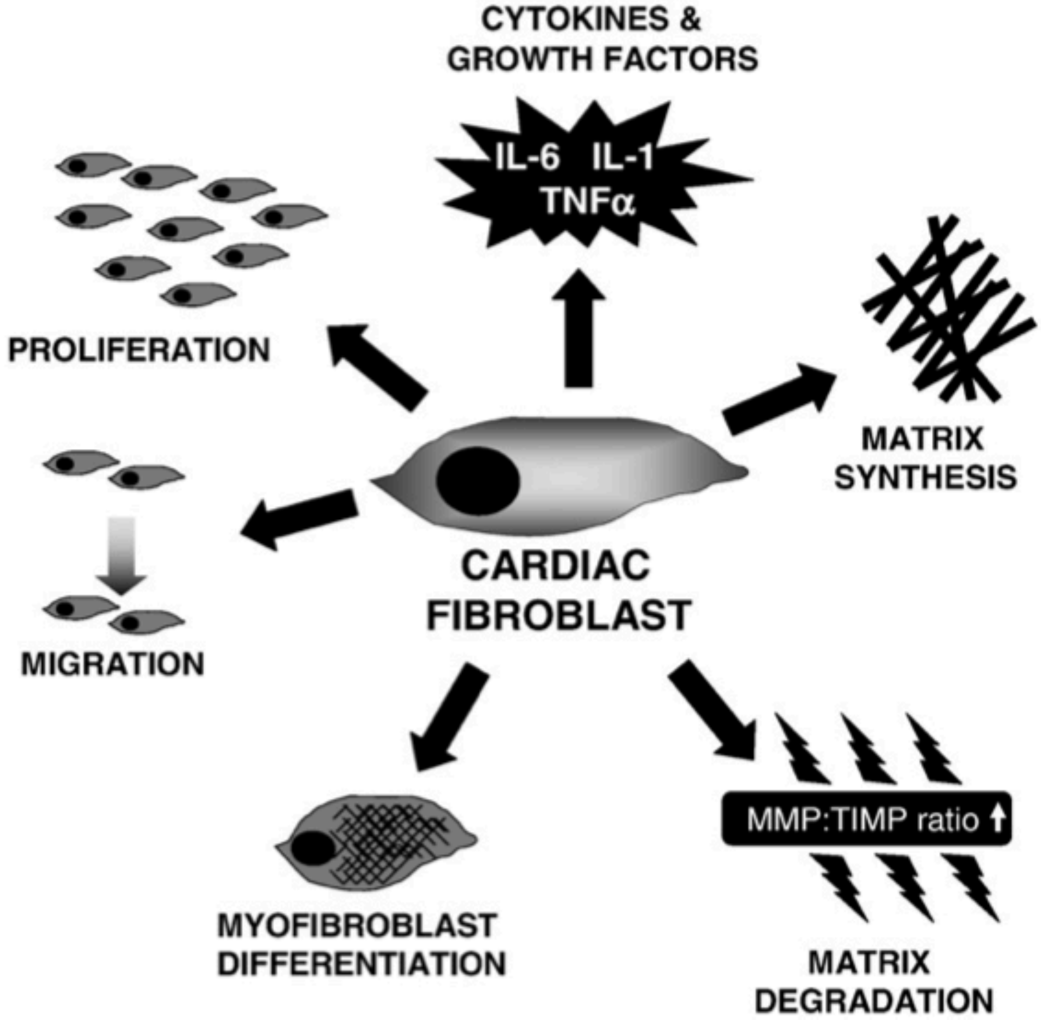


Figure 1.14. Depiction of the diverse functional roles of cardiac fibroblasts indicating a wealth of possible roles in normal physiology that can be found dysregulated in HCM. Reprinted from Porter and Turner, 2009 [73] with permission from the copyright holder.

The role of the myocyte in disease is of central importance in HCM as the sarcomere mutation is expressed in the myocyte. Thus, the myocyte is central to the first steps of HCM development. Enhanced biophysical/contractile properties of the mutant sarcomere were identified in HCM. This is measured as a sliding velocity between myosin and actin which appears to be quantal in nature, i.e. the myosin moves on the actin in discrete steps measured to be ~7nm, which may represent the distance between the myosin binding sites on the actin polymer. Increase of sliding velocity can be a direct effect of a reduction of the time that myosin remains on the actin (in the context of a myosin mutation), resulting in enhanced contractile properties of the sarcomere (Figure 1.15). A mutation potentially de-stabilizing actin binding to myosin could have such an effect. In both human and mouse HCM models, increased ATPase activity, force generation, and increased sliding velocities were observed [59, 75, 76].

In addition, stressed myocytes induce the expression of the transcription factor myocyte enahancement factor 2 (Mef2) possibly due to increased phosphorylated histone deacetylases and involvement from Ca2+/calmodulin-dependent protein kinase II, implicating Mef2 signaling in response processes in myocytes [77].

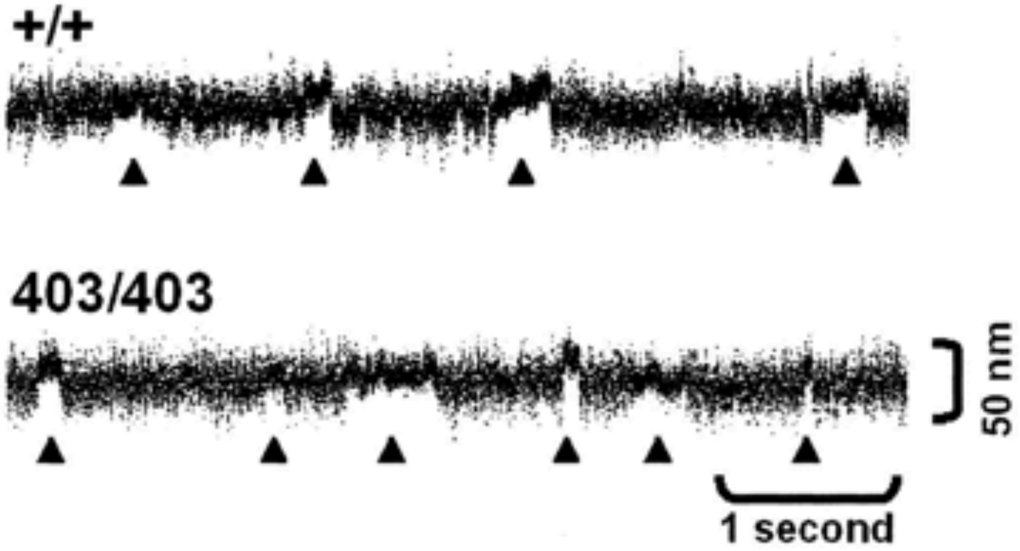


Figure 1.15. Biophysical differences between wild-type (+/+) and mutant (403/403) sarcomeres (representative data). Arrows: Actomyosin interaction resulting in a unitary displacement measured by an optical trap. Reprinted from Tyska et al., 2000 [76] with permission from the copyright holder.

Dissertation focus

The observation that HCM can be markedly heterogenenous even when it is caused by the same mutation points to the role of genetic, epigenetic and environmental factors in disease pathogenesis (Figure 1.16). Understanding these factors and processes can be of immense benefit in treating the disease. Compensatory factors can provide relief and provide new avenues for treatment. On the other hand, decompensatory environmental factors can be avoided [2].

The evolution of relevant technologies and methods allows wider and greater-depth assessment of disease processes. This dissertation undertook the task of developing methods based on next-generation sequencing to analyze the transcriptome of cells [78]. These methods have been applied in various cardiovascular or other biological contexts and in conjunction with follow-up analyses and experiments contributed to new insights [25, 79-82]. An early focus on genes that change start-site usage enabled usage of transcriptome data beyond the most commonly used gene expression analyses. As cells may utilize different start-sites to execute advanced transcriptional programs, we hypothesized that genes that change their 5' ends may serve important functions.

The remainder of this dissertation is organized as follows:

- Chapter 2 discusses the development of the Deep Sequencing Analysis of Gene Expression method and analytical tools.
- Chapter 3 discusses the development of normalizing cDNA libraries in order to facilitate increased sequencing of low expressing transcripts.

- Chapter 4 discusses the development of an approach to sequence miRNA, construct and compare miRNA profiles.
- Chapter 5 discusses the development of a complete framework involving library construction and analysis to assess start-site changes between samples. Applied in HCM, this method allowed the detection of Four-and-a-half LIM domains protein 1 as being critically regulated in HCM in stressed myocytes and providing for a compensatory function.
- Chapter 6 discusses conclusions and future directions.

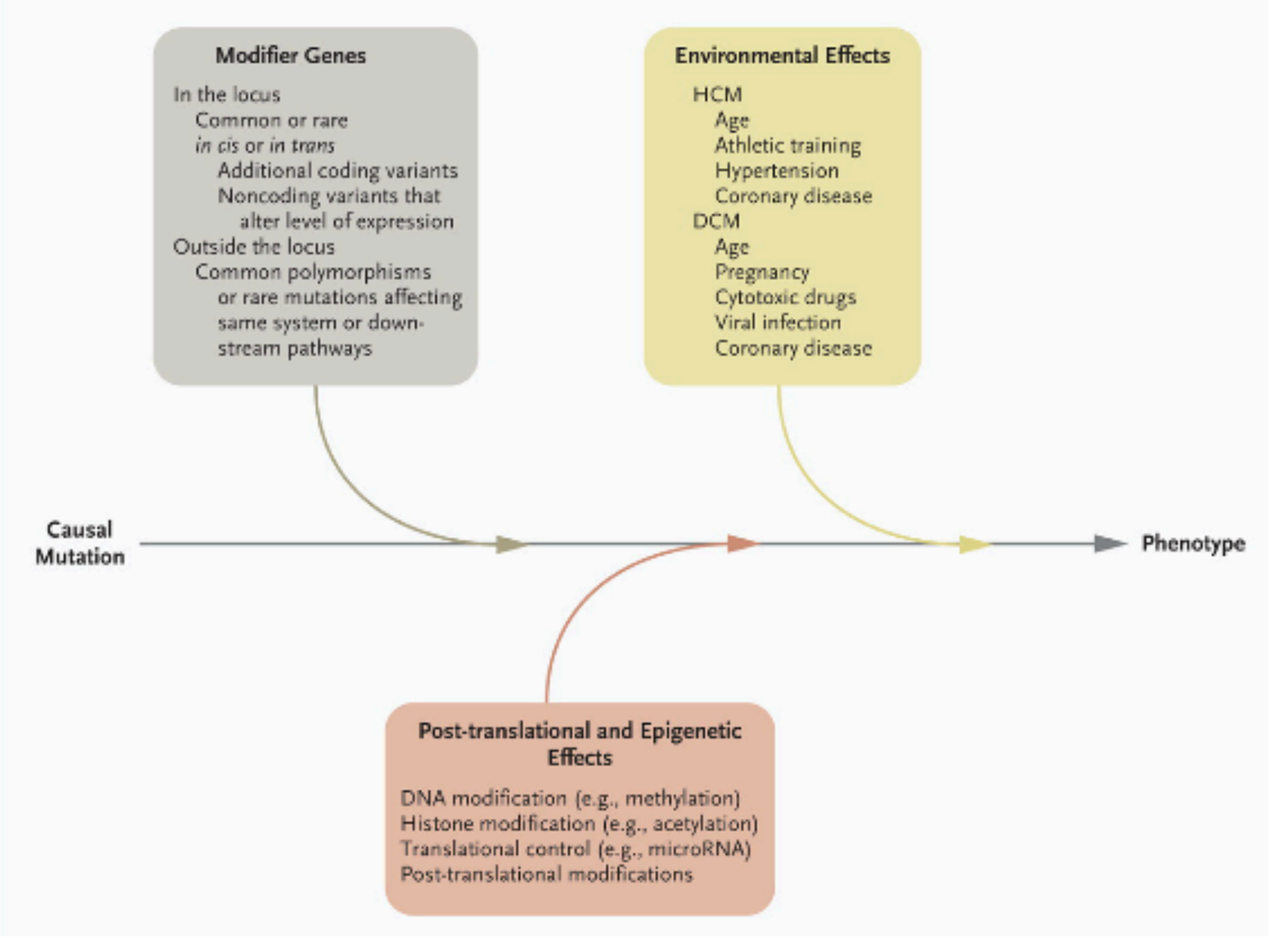

Modifier Genes
In the locus
Common or rare
in cis or *in trans*
Additional coding variants
Noncoding variants that alter level of expression
Outside the locus
Common polymorphisms or rare mutations affecting same system or down-stream pathways
Environmental Effects
HCM
Age
Athletic training
Hypertension
Coronary disease
DCM
Age
Pregnancy
Cytotoxic drugs
Viral infection
Coronary disease
Causal Mutation
Phenotype
Post-translational and Epigenetic Effects
DNA modification (e.g., methylation)
Histone modification (e.g., acetylation)
Translational control (e.g., microRNA)
Post-translational modifications

Figure 1.16. Genetic, epigenetic and environmental factors can affect the progression from causal mutation to HCM. Reprinted from Watkins et al., 2011 [2] with permission from the copyright holder (Massachusetts Medical Society).

References

1. Hoyert, D., J. Xu (2012). "Deaths: Preliminary Data for 2011." National Vital Statistics Reports 61(6).

2. Watkins, H., H. Ashrafian and C. Redwood (2011). "Inherited cardiomyopathies." N Engl J Med 364(17): 1643-56.

3. Morimoto, S. (2008). "Sarcomeric proteins and inherited cardiomyopathies." Cardiovasc Res 77(4): 659-66.

4. Wolf, C.M., C.I. Berul (2008). "Molecular mechanisms of inherited arrhythmias." Curr Genomics 9(3):160-8.

5. Elliott, P. and W. J. McKenna (2004). "Hypertrophic cardiomyopathy." Lancet

6. Ho, C. Y. (2010). "Hypertrophic cardiomyopathy." Heart Fail Clin 6(2): 141-59.

7. Maron, B.J., J.A. Towbin, G. Thiene, C. Antzelevitch, D. Corrado, D. Arnett, A.J. Moss, C.E. Seidman, J.B. Young; American Heart Association; Council on Clinical Cardiology, Heart Failure and Transplantation Committee; Quality of Care and Outcomes Research and Functional Genomics and Translational Biology Interdisciplinary Working Groups; Council on Epidemiology and Prevention (2006)."Contemporary definitions and classification of the cardiomyopathies: an American Heart Association Scientific Statement from the Council on Clinical Cardiology, Heart Failure and Transplantation Committee; Quality of Care and Outcomes Research and Functional Genomics and Translational Biology Interdisciplinary Working Groups; and Council on Epidemiology and Prevention." Circulation 113(14):1807-16.

8. Pruszczyk, P., A. Kostera-Pruszczyk, A. Shatunov, B. Goudeau, A. Draminska, K. Takeda, N. Sambuughin, P. Vicart, S. V. Strelkov, L. G. Goldfarb and A. Kaminska (2007). "Restrictive cardiomyopathy with atrioventricular conduction block resulting from a desmin mutation." Int J Cardiol 117(2): 244-53.

9. Peters, S. (2006). "Advances in the diagnostic management of arrhythmogenic right ventricular dysplasia-cardiomyopathy." Int J Cardiol 113(1):4-11.

10. Judge, D. P. (2011). "Arrhythmogenic right ventricular dysplasia/cardiomyopathy: a family affair." Circulation 123(23): 2661-3.

11. van Tintelen, J. P., R. M. Hofstra, A. C. Wiesfeld, M. P. van den Berg, R. N. Hauer and J. D. Jongbloed (2007). "Molecular genetics of arrhythmogenic right ventricular cardiomyopathy: emerging horizon?" Curr Opin Cardiol 22(3): 185-92.

12. Dalal, D., L. H. Molin, J. Piccini, C. Tichnell, C. James, C. Bomma, K. Prakasa, J. A. Towbin, F. I. Marcus, P. J. Spevak, D. A. Bluemke, T. Abraham, S. D. Russell, H. Calkins and D. P. Judge (2006). "Clinical features of arrhythmogenic right ventricular dysplasia/cardiomyopathy associated with mutations in plakophilin-2." Circulation 113(13): 1641-9.

13. Merner, N. D., K. A. Hodgkinson, A. F. Haywood, S. Connors, V. M. French, J. D. Drenckhahn, C. Kupprion, K. Ramadanova, L. Thierfelder, W. McKenna, B. Gallagher, L. Morris-Larkin, A. S. Bassett, P. S. Parfrey and T. L. Young (2008). "Arrhythmogenic right ventricular cardiomyopathy type 5 is a fully penetrant, lethal arrhythmic disorder caused by a missense mutation in the TMEM43 gene." Am J Hum Genet 82(4): 809-21.

14. Watanabe, H., B.C. Knollmann (2011). "Mechanism underlying catecholaminergic polymorphic ventricular tachycardia and approaches to therapy." J Electrocardiol 44(6):650-5.

15. Captur, G., P. Nihoyannopoulos (2010). "Left ventricular non-compaction: genetic heterogeneity, diagnosis and clinical course." Int J Cardiol 140(2):145-53.

16. Klaassen, S., S. Probst, E. Oechslin, B. Gerull, G. Krings, P. Schuler, M. Greutmann, D. Hürlimann, M. Yegitbasi, L. Pons, M. Gramlich, J.D. Drenckhahn, A. Heuser, F. Berger, R. Jenni, L. Thierfelder (2008). "Mutations in sarcomere protein genes in left ventricular noncompaction." Circulation 117(22):2893-901.

17. Morales, A., J. Cowan, J. Dagua and R. E. Hershberger (2008). "Family history: an essential tool for cardiovascular genetic medicine." Congest Heart Fail 14(1): 37-45.

18. Wang, L., J. G. Seidman and C. E. Seidman (2010). "Narrative review: harnessing molecular genetics for the diagnosis and management of hypertrophic cardiomyopathy." Ann Intern Med 152(8): 513-20, W181.

19. Rosenzweig, A., H. Watkins, D. S. Hwang, M. Miri, W. McKenna, T. A. Traill, J. G. Seidman and C. E. Seidman (1991). "Preclinical diagnosis of familial hypertrophic cardiomyopathy by genetic analysis of blood lymphocytes." N Engl J Med 325(25): 1753-60.

20. Erdmann, J., J. Raible, J. Maki-Abadi, M. Hummel, J. Hammann, B. Wollnik, E. Frantz, E. Fleck, R. Hetzer and V. Regitz-Zagrosek (2001). "Spectrum of clinical phenotypes and gene variants in cardiac myosin-binding protein C mutation carriers with hypertrophic cardiomyopathy." J Am Coll Cardiol 38(2): 322-30.

21. Kamisago, M., S.D. Sharma, S.R. DePalma, S. Solomon, P. Sharma, B. McDonough, L. Smoot, M.P. Mullen, P.K. Woolf, E.D. Wigle, J.G. Seidman, C.E. Seidman (2000). "Mutations in sarcomere protein genes as a cause of dilated cardiomyopathy." N Engl J Med 343(23):1688-96.

22. Frey, N., M. Luedde, H.A. Katus (2011). "Mechanisms of disease: hypertrophic cardiomyopathy." Nat Rev Cardiol. 9(2):91-100.

23. Dellefave, L. and E. M. McNally (2010). "The genetics of dilated cardiomyopathy." Curr Opin Cardiol.

24. Herman, D.S., G.K. Hovingh, O. Iartchouk, H.L. Rehm, R. Kucherlapati, J.G. Seidman and C.E. Seidman CE (2009). "Filter-based hybridization capture of subgenomes enables resequencing and copy-number detection." Nat Methods 6(7):507-10.

25. Herman D.S., L. Lam, M.R. Taylor, L. Wang, P. Teekakirikul, D. Christodoulou, L. Conner, S.R. DePalma, B. McDonough, E. Sparks, D.L. Teodorescu, A.L. Cirino, N.R. Banner, D.J. Pennell, S. Graw, M. Merlo, A. Di Lenarda, G. Sinagra, J.M. Bos, M.J. Ackerman, R.N. Mitchell, C.E. Murry, N.K. Lakdawala, C.Y. Ho, P.J. Barton, S.A. Cook, L. Mestroni, J.G. Seidman and C.E. Seidman (2012). "Truncations of titin causing dilated cardiomyopathy." N Engl J Med 366(7):619-28.

26. Marston, S.B. (2011). "How do mutations in contractile proteins cause the primary familial cardiomyopathies?" J Cardiovasc Transl Res 4(3):245-55.

27. Seidman, C. E. and J. G. Seidman (2011). "Identifying sarcomere gene mutations in hypertrophic cardiomyopathy: a personal history." Circ Res 108(6): 743-50.

28. Morita, H., R. Nagai, J. G. Seidman and C. E. Seidman (2010). "Sarcomere gene mutations in hypertrophy and heart failure." J Cardiovasc Transl Res 3(4): 297-303.

29. Olivotto, I., F. Girolami, M. J. Ackerman, S. Nistri, J. M. Bos, E. Zachara, S. R. Ommen, J. L. Theis, R. A. Vaubel, F. Re, C. Armentano, C. Poggesi, F. Torricelli and F. Cecchi (2008). "Myofilament protein gene mutation screening and outcome of patients with hypertrophic cardiomyopathy." Mayo Clin Proc 83(6): 630-8.

30. Hershberger, R. E., J. Lindenfeld, L. Mestroni, C. E. Seidman, M. R. Taylor and J. A. Towbin (2009). "Genetic evaluation of cardiomyopathy--a Heart Failure Society of America practice guideline." J Card Fail 15(2): 83-97.

31. Olson, T. M., T. P. Doan, N. Y. Kishimoto, F. G. Whitby, M. J. Ackerman and L. Fananapazir (2000). "Inherited and de novo mutations in the cardiac actin gene cause hypertrophic cardiomyopathy." J Mol Cell Cardiol 32(9): 1687-94.

32. Chiu, C., R. D. Bagnall, J. Ingles, L. Yeates, M. Kennerson, J. A. Donald, M. Jormakka, J. M. Lind and C. Semsarian (2009). "Mutations in alpha-actinin-2 cause hypertrophic cardiomyopathy: a genome-wide analysis." J Am Coll Cardiol 55(11): 1127-35.

33. Knoll, R., R. Postel, J. Wang, R. Kratzner, G. Hennecke, A. M. Vacaru, P. Vakeel, C. Schubert, K. Murthy, B. K. Rana, D. Kube, G. Knoll, K. Schafer, T. Hayashi, T. Holm, A. Kimura, N. Schork, M. R. Toliat, P. Nurnberg, H. P. Schultheiss, W. Schaper, J. Schaper, E. Bos, J. Den Hertog, F. J. van Eeden, P. J. Peters, G. Hasenfuss, K. R. Chien and J. Bakkers (2007). "Laminin-alpha4 and integrin-linked kinase mutations cause human cardiomyopathy via simultaneous defects in cardiomyocytes and endothelial cells." Circulation 116(5): 515-25.

34. Geier, C., K. Gehmlich, E. Ehler, S. Hassfeld, A. Perrot, K. Hayess, N. Cardim, K. Wenzel, B. Erdmann, F. Krackhardt, M. G. Posch, K. J. Osterziel, A. Bublak, H. Nagele, T. Scheffold, R. Dietz, K. R. Chien, S. Spuler, D. O. Furst, P. Nurnberg and C. Ozcelik (2008). "Beyond the sarcomere: CSRP3 mutations cause hypertrophic cardiomyopathy." Hum Mol Genet 17(18): 2753-65.

35. Saltzman, A. J., D. Mancini-DiNardo, C. Li, W. K. Chung, C. Y. Ho, S. Hurst, J. Wynn, M. Care, R. M. Hamilton, G. W. Seidman, J. Gorham, B. McDonough, E. Sparks, J. G. Seidman, C. E. Seidman and H. L. Rehm (2010). "Short communication: the cardiac myosin binding protein C Arg502Trp mutation: a common cause of hypertrophic cardiomyopathy." Circ Res 106(9): 1549-52.

36. Osio, A., L. Tan, S. N. Chen, R. Lombardi, S. F. Nagueh, S. Shete, R. Roberts, J. T. Willerson and A. J. Marian (2007). "Myozenin 2 is a novel gene for human hypertrophic cardiomyopathy." Circ Res 100(6): 766-8.

37. Niimura, H., L. L. Bachinski, S. Sangwatanaroj, H. Watkins, A. E. Chudley, W. McKenna, A. Kristinsson, R. Roberts, M. Sole, B. J. Maron, J. G. Seidman and C. E. Seidman (1998). "Mutations in the gene for cardiac myosin-binding protein C and late-onset familial hypertrophic cardiomyopathy." N Engl J Med 338(18): 1248-57.

38. Hoffmann, B., H. Schmidt-Traub, A. Perrot, K. J. Osterziel and R. Gessner (2001). "First mutation in cardiac troponin C, L29Q, in a patient with hypertrophic cardiomyopathy." Hum Mutat 17(6): 524.

39. Konno, T., M. Shimizu, H. Ino, N. Fujino, K. Uchiyama, T. Mabuchi, K. Sakata, T. Kaneda, T. Fujita, E. Masuta and H. Mabuchi (2006). "A novel mutation in the cardiac myosin-binding protein C gene is responsible for hypertrophic cardiomyopathy with severe ventricular hypertrophy and sudden death." Clin Sci (Lond) 110(1): 125-31.

40. Anan, R., H. Shono, A. Kisanuki, S. Arima, S. Nakao and H. Tanaka (1998). "Patients with familial hypertrophic cardiomyopathy caused by a Phe110Ile missense mutation in the cardiac troponin T gene have variable cardiac morphologies and a favorable prognosis." Circulation 98(5): 391-7.

41. Ho, C. Y., H. M. Lever, R. DeSanctis, C. F. Farver, J. G. Seidman and C. E. Seidman (2000). "Homozygous mutation in cardiac troponin T: implications for hypertrophic cardiomyopathy." Circulation 102(16): 1950-5.

42. Coviello, D. A., B. J. Maron, P. Spirito, H. Watkins, H. P. Vosberg, L. Thierfelder, F. J. Schoen, J. G. Seidman and C. E. Seidman (1997). "Clinical features of hypertrophic cardiomyopathy caused by mutation of a "hot spot" in the alpha-tropomyosin gene." J Am Coll Cardiol 29(3): 635-40.

43. Arimura, T., J. M. Bos, A. Sato, T. Kubo, H. Okamoto, H. Nishi, H. Harada, Y. Koga, M. Moulik, Y. L. Doi, J. A. Towbin, M. J. Ackerman and A. Kimura (2009). "Cardiac ankyrin repeat protein gene (ANKRD1) mutations in hypertrophic cardiomyopathy." J Am Coll Cardiol 54(4): 334-42.

44. Carniel E., M.R. Taylor, G. Sinagra, A. Di Lenarda, L. Ku, P.R. Fain, M.M. Boucek, J. Cavanaugh, S. Miocic, D. Slavov, S.L. Graw, J. Feiger, X.Z. Zhu, D. Dao, D.A. Ferguson, M.R. Bristow and L. Mestroni "Alpha-myosin heavy chain: a sarcomeric gene associated with dilated and hypertrophic phenotypes of cardiomyopathy." Circulation 112(1):54-9.

45. Vasile, V.C., S.R. Ommen, W.D. Edwards and M.J. Ackerman. A missense mutation in a ubiquitously expressed protein, vinculin, confers susceptibility to hypertrophic cardiomyopathy. Biochem Biophys Res Commun. 2006 Jul 7;345(3):998-1003.

46. Vasile, V. C., M. L. Will, S. R. Ommen, W. D. Edwards, T. M. Olson and M. J. Ackerman (2006). "Identification of a metavinculin missense mutation, R975W, associated with both hypertrophic and dilated cardiomyopathy." Mol Genet Metab 87(2): 169-74.

47. Ho, C. Y. (2009). "Hypertrophic cardiomyopathy: preclinical and early phenotype." J Cardiovasc Transl Res 2(4): 462-70.

48. Szczesna-Cordary, D., M. Jones, J. R. Moore, J. Watt, W. G. Kerrick, Y. Xu, Y. Wang, C. Wagg and G. D. Lopaschuk (2007). "Myosin regulatory light chain E22K mutation results in decreased cardiac intracellular calcium and force transients." FASEB J 21(14): 3974-85.

49. Haim, T. E., C. Dowell, T. Diamanti, J. Scheuer and J. C. Tardiff (2007). "Independent FHC-related cardiac troponin T mutations exhibit specific alterations in myocellular contractility and calcium kinetics." J Mol Cell Cardiol 42(6): 1098-110.

50. Fatkin, D., B. K. McConnell, J. O. Mudd, C. Semsarian, I. G. Moskowitz, F. J. Schoen, M. Giewat, C. E. Seidman and J. G. Seidman (2000). "An abnormal Ca(2+) response in mutant sarcomere protein-mediated familial hypertrophic cardiomyopathy." J Clin Invest 106(11): 1351-9.

51. Baudenbacher, F., T. Schober, J. R. Pinto, V. Y. Sidorov, F. Hilliard, R. J. Solaro, J. D. Potter and B. C. Knollmann (2008). "Myofilament Ca2+ sensitization causes susceptibility to cardiac arrhythmia in mice." J Clin Invest 118(12): 3893-903.

52. Spindler, M., K. W. Saupe, M. E. Christe, H. L. Sweeney, C. E. Seidman, J. G. Seidman and J. S. Ingwall (1998). "Diastolic dysfunction and altered energetics in the alphaMHC403/+ mouse model of familial hypertrophic cardiomyopathy." J Clin Invest 101(8): 1775-83.

53. Arad, M., D. W. Benson, A. R. Perez-Atayde, W. J. McKenna, E. A. Sparks, R. J. Kanter, K. McGarry, J. G. Seidman and C. E. Seidman (2002). "Constitutively active AMP kinase mutations cause glycogen storage disease mimicking hypertrophic cardiomyopathy." J Clin Invest 109(3): 357-62.

54. Maron, B. J., W. C. Roberts, M. Arad, T. S. Haas, P. Spirito, G. B. Wright, A. K. Almquist, J. M. Baffa, J. P. Saul, C. Y. Ho, J. Seidman and C. E. Seidman (2009). "Clinical outcome and phenotypic expression in LAMP2 cardiomyopathy." JAMA 301(12): 1253-9.

55. Wolf, C. M., M. Arad, F. Ahmad, A. Sanbe, S. A. Bernstein, O. Toka, T. Konno, G. Morley, J. Robbins, J. G. Seidman, C. E. Seidman and C. I. Berul (2008). "Reversibility of PRKAG2 glycogen-storage cardiomyopathy and electrophysiological manifestations." Circulation 117(2): 144-54.

56. Konno, T., S. Chang, J. G. Seidman and C. E. Seidman (2010). "Genetics of hypertrophic cardiomyopathy." Curr Opin Cardiol.

57. Geisterfer-Lowrance, A. A., S. Kass, G. Tanigawa, H. P. Vosberg, W. McKenna, C. E. Seidman and J. G. Seidman (1990). "A molecular basis for familial hypertrophic cardiomyopathy: a beta cardiac myosin heavy chain gene missense mutation." Cell 62(5): 999-1006.

58. Jarcho, J. A., W. McKenna, J. A. Pare, S. D. Solomon, R. F. Holcombe, S. Dickie, T. Levi, H. Donis-Keller, J. G. Seidman and C. E. Seidman (1989). "Mapping a gene for familial hypertrophic cardiomyopathy to chromosome 14q1." N Engl J Med 321(20): 1372-8.

59. Palmiter, K. A., M. J. Tyska, J. R. Haeberle, N. R. Alpert, L. Fananapazir and D. M. Warshaw (2000). "R403Q and L908V mutant beta-cardiac myosin from patients with familial hypertrophic cardiomyopathy exhibit enhanced mechanical performance at the single molecule level." J Muscle Res Cell Motil 21(7): 609-20.

60. Woo, A., H. Rakowski, J. C. Liew, M. S. Zhao, C. C. Liew, T. G. Parker, M. Zeller, E. D. Wigle and M. J. Sole (2003). "Mutations of the beta myosin heavy chain gene in hypertrophic cardiomyopathy: critical functional sites determine prognosis." Heart 89(10): 1179-85.

61. Chung, M.W., T. Tsoutsman, C. Semsarian (2003). "Hypertrophic cardiomyopathy: from gene defect to clinical disease." Cell Res 13(1):9-20.

62. Ho, C.Y. and C.E. Seidman (2007). "Inherited Cardiomyopathies. Principles and Practice of Medical Genetics. " 53:1160-1183.

63. Maron, B. J. (2010). "Risk stratification and role of implantable defibrillators for prevention of sudden death in patients with hypertrophic cardiomyopathy." Circ J 74(11): 2271-82.

64. Maron, B. J., W. K. Shen, M. S. Link, A. E. Epstein, A. K. Almquist, J. P. Daubert, G. H. Bardy, S. Favale, R. F. Rea, G. Boriani, N. A. Estes, 3rd and P. Spirito (2000). "Efficacy of implantable cardioverter-defibrillators for the prevention of sudden death in patients with hypertrophic cardiomyopathy." N Engl J Med 342(6): 365-73.

65. Kimmelstiel, C. D. and B. J. Maron (2004). "Role of percutaneous septal ablation in hypertrophic obstructive cardiomyopathy." Circulation 109(4): 452-6.

66. McKenna, W. J. and E. R. Behr (2002). "Hypertrophic cardiomyopathy: management, risk stratification, and prevention of sudden death." Heart 87(2): 169-76.

67. Ho, C. Y. (2010). "Genetics and clinical destiny: improving care in hypertrophic cardiomyopathy." Circulation 122(23): 2430-40; discussion 2440.

68. Geisterfer-Lowrance, A. A., M. Christe, D. A. Conner, J. S. Ingwall, F. J. Schoen, C. E. Seidman and J. G. Seidman (1996). "A mouse model of familial hypertrophic cardiomyopathy." Science 272(5262): 731-4.

69. Ho, C. Y., B. Lopez, O. R. Coelho-Filho, N. K. Lakdawala, A. L. Cirino, P. Jarolim, R. Kwong, A. Gonzalez, S. D. Colan, J. G. Seidman, J. Diez and C. E. Seidman (2010). "Myocardial fibrosis as an early manifestation of hypertrophic cardiomyopathy." N Engl J Med 363(6): 552-63.

70. Kim, J. B., G. J. Porreca, L. Song, S. C. Greenway, J. M. Gorham, G. M. Church, C. E. Seidman and J. G. Seidman (2007). "Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy." Science 316(5830): 1481-4.

71. Teekakirikul, P., S. Eminaga, O. Toka, R. Alcalai, L. Wang, H. Wakimoto, M. Nayor, T. Konno, J. M. Gorham, C. M. Wolf, J. B. Kim, J. P. Schmitt, J. D. Molkentin, R. A. Norris, A. M. Tager, S. R. Hoffman, R. R. Markwald, C. E. Seidman and J. G. Seidman (2010). "Cardiac fibrosis in mice with hypertrophic cardiomyopathy is mediated by non-myocyte proliferation and requires Tgf-beta." J Clin Invest 120(10): 3520-9.

72. Teekakirikul, P., R.F. Padera, J.G. Seidman and C.E. Seidman. "Hypertrophic cardiomyopathy: Translating cellular cross talk into therapeutics." J Cell Biol. 2012 Oct 29;199(3):417-21.

73. Porter, K.E. and N.A. Turner (2009). "Cardiac fibroblasts: at the heart of myocardial remodeling." Pharmacol Ther. 123(2):255-78.

74. Shirani, J., R. Pick, W. C. Roberts and B. J. Maron (2000). "Morphology and significance of the left ventricular collagen network in young patients with hypertrophic cardiomyopathy and sudden cardiac death." J Am Coll Cardiol 35(1): 36-44.

75. Debold, E. P., J. P. Schmitt, J. B. Patlak, S. E. Beck, J. R. Moore, J. G. Seidman, C. Seidman and D. M. Warshaw (2007). "Hypertrophic and dilated cardiomyopathy mutations differentially affect the molecular force generation of mouse alpha-cardiac myosin in the laser trap assay." Am J Physiol Heart Circ Physiol 293(1): H284-91.

76. Tyska, M. J., E. Hayes, M. Giewat, C. E. Seidman, J. G. Seidman and D. M. Warshaw (2000). "Single-molecule mechanics of R403Q cardiac myosin isolated from the mouse model of familial hypertrophic cardiomyopathy." Circ Res 86(7): 737-44.

77. Konno, T., D. Chen, L. Wang, H. Wakimoto, P. Teekakirikul, M. Nayor, M. Kawana, S. Eminaga, J. M. Gorham, K. Pandya, O. Smithies, F. J. Naya, E. N. Olson, J. G. Seidman and C. E. Seidman (2010). "Heterogeneous myocyte enhancer factor-2 (Mef2) activation in myocytes predicts focal scarring in hypertrophic cardiomyopathy." Proc Natl Acad Sci U S A 107(42): 18097-102.

78. Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53-59.

79. Rios, X., A.W. Briggs, D. Christodoulou, J.M. Gorham, J.G. Seidman, G.M. Church (2012). "Stable gene targeting in human cells using single-strand oligonucleotides with modified bases. PLoS One 7(5).

80. Delgado-Olguín, P., Y. Huang, X. Li, D. Christodoulou, C.E. Seidman and J.G. Seidman, A. Tarakhovsky and B.G. Bruneau (2012). "Epigenetic repression of cardiac progenitor gene expression by Ezh2 is required for postnatal cardiac homeostasis." Nat Genet 44(3):343-7.

81. He, A., Q. Ma, J. Cao, A. von Gise, P. Zhou, H. Xie, B. Zhang, M. Hsing, D.C. Christodoulou, P. Cahan, G.Q. Daley, S.W. Kong, S.H. Orkin, C.E. Seidman, J.G. Seidman and W.T. Pu (2012). "Polycomb repressive complex 2 regulates normal development of the mouse heart." Circ Res 110(3):406-15.

82. Alon, S., F. Vigneault, S. Eminaga, D.C. Christodoulou, J.G. Seidman, G.M. Church and E. Eisenberg (2011). "Barcoding bias in high-throughput multiplex sequencing of miRNA." Genome Res 21(9).

CHAPTER 2

Chapter 2. Quantification of Gene Transcripts with Deep Sequencing Analysis of Gene Expression (DSAGE) Using 1 to 2 μg Total RNA:

Published in Current Protocols of Molecular Biology (Curr Protoc Mol Biol. 2011 Jan;Chapter 25:Unit25B.9). Reproduced with permission of Current Protocols and John Wiley & Sons, Inc. Copyright © 2011 by John Wiley & Sons, Inc.

Danos C. Christodoulou, Joshua M. Gorham, Masataka Kawana, Steven R. DePalma, Daniel S. Herman, Hiroko Wakimoto

Introduction

Deep sequencing analysis of gene expression (DSAGE) measures global gene transcript levels from only 1 to 2 μg total RNA by massively parallel sequencing of cDNA tags. First, 21-bp cDNA tags are generated by NlaIII digestion of the cDNA, followed by MmeI cleavage. The cDNA tags are then queried by massively parallel sequencing and aligned to a reference genome and transcriptome (or any available gene sequences) using Bowtie, an ultra-high-throughput short-read aligner, and Tophat, a fast splice-junction mapper. Analysis of 10 to 20 million tags, acquired using one lane of an Illumina Genome Analyzer II, provides sufficient depth to quantify gene expression and detect rare transcripts. Typically, we observe the expression of 15,000 genes in the cardiac left ventricle, including gene transcripts expressed at levels as low as one copy per cell. These expression profiles are highly reproducible (r > 0.99 between technical replicates), enabling sensitive detection of differences between experimental conditions as well as assessment of the relative transcript abundance of different genes. The significance of these differences is assessed, while accounting for multiple comparisons, using a false discovery rate approach. Thus, DSAGE can be used to quantify gene expression profiles and assess differential expression with high sensitivity, and requires small amounts of biological material.

Basic Protocol 1 describes the construction of 21-bp cDNA tag libraries appropriate for massively parallel sequencing and the analysis of the resulting sequence data. The adapter oligonucleotides used are optimized for sequencing with current Illumina massively parallel sequencers. A step-by-step implementation of the analysis protocol is described (see Basic Protocol 2), which is compiled into three steps (Alternate Protocol).

Basic Protocol 1: DSAGE Library Construction

Construction of the DSAGE library begins with extraction of 1 to 2 µg high-quality RNA. If it is desirable to minimize biological noise, RNA can be pooled from three to five sample replicates. RNA can be extracted using the RNAeasy kit (Qiagen) or a standard protocol using Trizol (Invitrogen). RNA quality can be assessed by running an aliquot on a Bioanalyzer (Agilent; Schroeder et al., 2006) or an agarose gel. Good-quality RNA has a ~2:1 ratio of 28S to 18S RNA.

An overview of the subsequent steps is presented in Figure 2.1. Poly(A) RNA is captured on Dynal(dT) beads. Reverse transcription, second strand synthesis, NlaIII digestion, and adapter 1 ligation are performed with the poly(A) tail of the original RNA attached to the beads. NlaIII cleaves cDNA at defined sites, and subsequent washes of the beads remove all but the 3’ fragment of the cDNA. Adapter 1 is then ligated to the overhang resulting from NlaIII digestion, creating an MmeI site at the overlap of the NlaIII site (cDNA) and the adapter sequence. MmeI cuts 18/20-bp downstream of its recognition site, releasing a 21-bp tag with a 2-bp overhang attached to adapter 1. Adapter 2 is then ligated to the released fragment, and the resulting molecules are amplified by PCR. To avoid amplification bias, a mock reaction is monitored by real-time PCR, and the subsequent library amplification is performed within the log phase of the reaction. The PCR product is then purified, quantified, and submitted for sequencing at a concentration of 10 nM library molecules. The sequencing is typically performed by a core facility and is described in detail in Bentley et al. (2008).

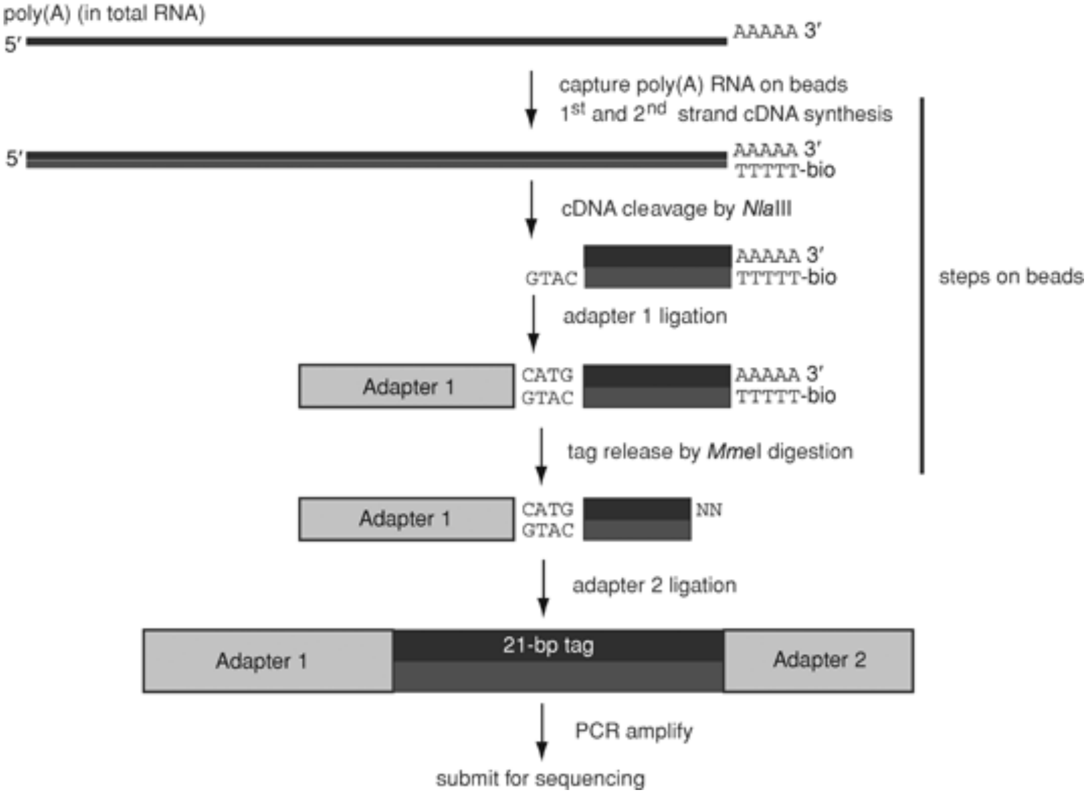


Figure 2.1 cDNA library construction for DSAGE. mRNA is selected from total RNA using oligo(dT) ferromagnetic beads, and double-stranded cDNA is synthesized while the mRNA is bound to the beads. cDNA is then cleaved with anchoring enzyme NlaIII, leaving a 4-bp 3'overhang (CATG), which is ligated to a forward adapter containing an MmeI restriction site. MmeI cuts 20 bp downstream of the non-palindromic recognition site, generating a 21-bp unique tag sequence for each mRNA transcript. Finally, the reverse adapter containing a 2-bp 3' degenerate overhang is ligated.

The adapter oligonucleotides are designed to work with standard Illumina Genome Analyzer flow cells and amplification primers. This protocol describes the use of a sequencing primer that overlaps the common NlaIII site and reduces the number of cycles needed for sequencing to 17.

Materials

- Dynabeads mRNA DIRECT Kit (Invitrogen, cat. #610-12), including oligo(dT) beads, lysis buffer, buffer A, and buffer B
- High-quality RNA, optionally from three to five sample replicates, extracted using RNAeasy kit (Qiagen, cat. #74104) or standard protocol using Trizol (Invitrogen, cat. #15596-018)
- SuperScript II reverse transcriptase (Invitrogen, cat. #18064-022), including First Strand Buffer, DTT, and 10 mM dNTPs
- Second-Strand Buffer (Invitrogen, cat. #10812-014)
- *E. coli* DNA polymerase I (Invitrogen, cat. #18010025)
- *E. coli* DNA ligase (Invitrogen, cat. #18052019)
- *E. coli* RNase H (Invitrogen, cat. #18021071)
- Buffers C, D, and E (see recipes)
- LoTE buffer (see recipe)
- 100× BSA (New England Biolabs, cat. #B9001S)
- NEBuffer 4 (New England Biolabs, cat. #B7004S)
- 0.5 M EDTA, pH 8.0 (APPENDIX 2)
- NlaIII (New England Biolabs, cat. #R0125L)

- Adapters 1 and 2 (see recipe)
- T4 DNA ligase, high concentration, with ligase buffer (Invitrogen, cat. #15224-041)
- MmeI (New England Biolabs, cat. #R0637L) with S-adenosylmethionine (SAM)
- Phenol/chloroform (Ambion, cat. #AM9732), adjusted to pH 7.5-8
- GlycoBlue (Ambion, cat. #AM9515)
- 7.5 M ammonium acetate (APPENDIX 2)
- 100% ethanol
- 10× BlueJuice gel loading buffer (Invitrogen, cat. #10816-015)
- Novex 20% TBE gels (Invitrogen, cat. #EC6315BOX)
- Low-molecular-weight DNA ladder (New England Biolabs, cat. #N3233S)
- SYBR Green I (Invitrogen, cat. #S7563)
- Glycogen (Roche, cat. #10901393001)
- MGB buffer (see recipe)
- Dimethyl sulfoxide (DMSO)
- 10 mM dNTP mix
- 100 μM GexPCR primers A and B (Integrated DNA Technologies):
  - 5’-CAAGCAGAAGACGGCATACGA
  - 5’-AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA
- Platinum Taq DNA polymerase (Invitrogen, cat. #10966034)
- Quant-iT dsDNA HS Assay Kit (Invitrogen, cat. #Q32851)
- 0.5- and 1.5-ml non-stick RNase-free microcentrifuge tubes (Ambion, cat. #12350 and 12450)

- DynaMag-2 magnet (Invitrogen, cat. #123-21D)
- 16°C water bath
- 50° and 60°C heating blocks
- UV light source
- 18-G needle
- Spin-X centrifuge tube filters (Corning, cat. #8161)
- Thermocycler
- Qubit fluorometer (Invitrogen, cat. #Q32857)
- Additional reagents and equipment for polyacrylamide (UNIT 2.7) and agarose (UNIT 2.5A) gel electrophoresis

Stage I: Capture RNA, reverse transcribe, and synthesize cDNA

1. Transfer 100 µl oligo(dT) Dynabeads to a 1.5-ml non-stick microcentrifuge tube. Capture the beads by placing the tube on a DynaMag-2 magnet, pipet out the buffer, and quickly resuspend the beads in 500 µl lysis buffer (supplied in the Dynabeads kit).

   Non-stick tubes are used throughout the experiment.
2. Add 1-2 µg RNA and incubate for 10 min with gentle agitation.
3. Capture the beads with the bound RNA on the magnet, remove the supernatant, and wash as follows:
   - Twice with 500 µl buffer A (Dynabeads kit)
   - Once with 500 µl buffer B (Dynabeads kit)
   - Twice with 200 µl 1× First Strand Buffer (SuperScript II kit)

4. Prepare the following mix:

    - 18 µl 5× First Strand Buffer
    - 9 µl 0.1 M DTT
    - 4.5 µl 10 mM dNTPs
    - 55.5 µl ddH_2O

    Vortex, briefly microcentrifuge to collect the solution at the bottom of the tube, and keep on ice.

5. Resuspend washed beads in the mix and place tubes at 37°C for 2 min.

6. Add 3 µl Superscript II reverse transcriptase, vortex gently, and incubate for 1 hr at 37°C. Mix beads every 20 min by gentle vortexing.

7. Place the tubes on ice and add the following components in the order shown:

    - 465 µl prechilled ddH_2O
    - 150 µl of 5× Second Strand Buffer
    - 15 µl 10 mM dNTPs
    - 5 µl *E. coli* DNA ligase
    - 20 µl *E. coli* DNA polymerase I
    - 5 µl *E. coli* RNase H

    Incubate 2 hr in a 16°C water bath. Vortex gently every 20 min.

8. During the 2-hr incubation, preheat buffer C to 75°C and prepare the NlaIII mix for each sample:

    - 172 µl LoTE
    - 2 µl 100× BSA

- 20 µl 10× NEBuffer 4
- Keep the above mix on ice.

9. Following second strand synthesis, add 45 µl of 0.5 M EDTA to stop the reaction. Capture the beads on the magnet and remove the reaction liquid.
10. Resuspend the beads in 450 µl preheated buffer C and incubate at 75°C for 12 min with intermittent gentle vortexing.
11. Capture the beads using the magnet and wash once more with preheated buffer C. Under these conditions, the beads tend to clump on the tube. If this occurs, scrape the clumped beads from the tube using a pipet tip and proceed quickly, since SDS tends to precipitate.
12. Capture beads, remove solution, and resuspend beads in 500 µl buffer D.
13. Transfer resuspended beads to a new tube, wash twice with 200 µl buffer D, then twice with 200 µl buffer E.

Stage II: Cleave cDNA with NlaIII and ligate adapter 1

14. Add 6 µl NlaIII to the prepared NlaIII mix from step 8. Resuspend beads in the mix and incubate for 1 hr at 37°C.
15. Wash beads two times with 450 µl buffer C (preheated to 37°C). Wash quickly since SDS tends to precipitate.
16. Wash twice with 200 µl buffer D and transfer the bead slurry to a new tube. At this stage the beads can be kept rotating overnight at 4°C, if desired.
17. Wash once more with buffer D, twice with 200 µl of 1× ligase buffer, and once with 50 µl of 1× ligase buffer.

At this stage, keep the beads in ligase buffer on ice if multiple samples are being handled.

18. Capture beads using the magnet, remove the solution, and resuspend in the following mix:
    - 11.5 µl LoTE
    - 4 µl 5× DNA ligase buffer
    - 2 µl adapter 1 (final 50 µM)
    - Incubate 2 min in a 50°C heating block, then 10 min at room temperature. Place the tubes on ice.
19. Add 2.5 µl T4 DNA ligase (high conc.) and incubate in a 16°C water bath for 2 hr. Mix beads every 20 min by gentle vortexing.

Stage III: Digest with MmeI, ligate adapter 2, and gel purify

20. During the ligase incubation (step 19), dilute 32 mM SAM (provided with MmeI) to make the following 10× SAM solution (prepare fresh):
    - 3 µl 32 mM SAM
    - 24 µl 10× NEBuffer 4
    - 213 µl ddH_2O
21. Prepare the MmeI mix just before use:
    - 20 µl 10× NEBuffer 4
    - 20 µl 10× SAM solution
    - 150 µl ddH_2O
22. After ligation, wash the beads three times with 200 µl buffer D, resuspend in 200 µl

buffer E, and transfer the bead slurry to a new tube. Wash once more with buffer E.

23. Preheat the MmeI mix for 2 min at 37°C. Capture the beads on the magnet and discard the buffer.

24. Quickly add 10 μl MmeI enzyme to the preheated MmeI mix and use to resuspend the beads (160 μl/sample). Vortex gently and incubate for 2 hr at 37°C. Mix beads every 20 min by gentle vortexing.

25. Microcentrifuge for 2 min at 14,000 rpm, room temperature. Transfer the supernatant to a new tube.

26. To collect any residual material, add 100 μl LoTE to the beads, mix, and recentrifuge. Place on the magnet and transfer the 100 μl supernatant to the previous tube (total 300 μl). Discard the beads.

27. To assess contamination, transfer a 40-μl aliquot to a new tube, label as the "ligase minus control", and adjust the volume of the remaining sample back to 300 μl with LoTE. Process the sample and control identically through step 30.

28. Add 300 μl phenol/chloroform, vortex, and microcentrifuge 2 min at 14,000 rpm. Transfer aqueous phase to a new tube.

29. To precipitate, add 2 μl Glycoblue and 133 μl of 7.5 M ammonium acetate and mix. Add 1 ml of 100% ethanol, vortex immediately, and precipitate at –80°C for at least 30 min.

    At this stage the sample can be kept overnight at –80°C.

30. Microcentrifuge 30 min at 14,000 rpm, 4°C. Wash two times with 500 μl cold 70% ethanol. Remove the second wash, allow the pellet to air dry, and resuspend in 2 μl

LoTE.

31. Add the following and mix by pipetting:
    - 2 µl 5× ligase buffer
    - 2 µl adapter 2 (final 50 µM)
    - 2 µl ddH_2O
    - 2 µl T4 DNA ligase (high conc.)

    Use water instead of ligase in the “ligase minus control”. Incubate both the sample and control overnight at 16°C in a thermocycler. Set control aside at –20°C until step 41.
32. Add 2 µl LoTE and 3 µl of 5× BlueJuice gel loading buffer to the sample ligation. Load sample in one lane of a 20% TBE gel, load low-molecular-weight DNA ladder in a separate lane, and run until the blue dye front reaches the bottom of the gel (~2 hr).
33. Stain with SYBR green for 15 min at room temperature. Visualize with a UV light and excise the gel band corresponding to 66 bp (Fig. 2.2).

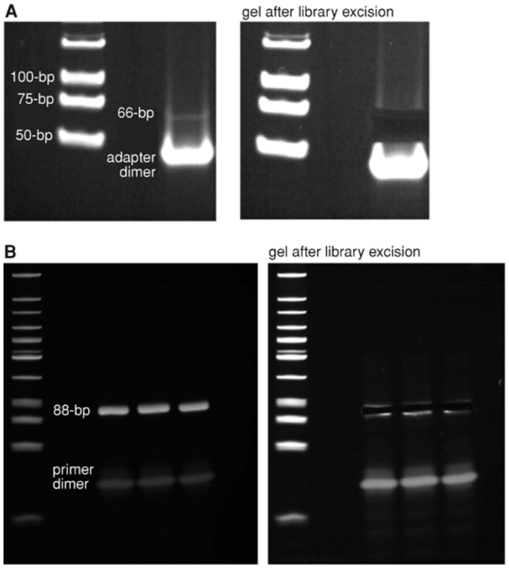


Figure 2.2 Library excision from a polyacrylamide gel. The 66-bp pre-PCR library (A) and the final amplified 88-bp library (B) are shown.

34. Place the excised band into a 0.5-ml tube with a hole punctured in the bottom using an 18-G needle, and place the 0.5-ml tube in a 1.5-ml tube. Crush the extracted gel by centrifuging 6 min at 14,000 rpm, collecting the crushed gel in the 1.5-ml tube.

35. Add 250 µl LoTE and 50 µl of 7.5 M ammonium acetate. Vortex and heat in a 60°C heating block for 15 min.

36. Pre-wet a SpinX column with 5 µl LoTE and add the sample. Centrifuge 5 min at 14,000 rpm, room temperature.

37. Add 3 µl glycogen and 133 µl of 7.5 M ammonium acetate, mix, and precipitate with 1 ml 100% ethanol at –80°C.

    The sample can be stored overnight at –80°C.

38. Centrifuge 15 min at 14,000 rpm, 4°C. Wash two times with 500 µl of 70% ethanol, air dry pellet, and resuspend in 14 µl LoTE.

    The library can be stored up to several days at –20°C.

Stage IV: Amplify library

39. To determine the optimal number of cycles, perform a mock reaction for qPCR including SYBR Green. For each sample assemble:
    - 5 µl 10× MGB PCR buffer
    - 2.5 µl DMSO
    - 3 µl 10 mM dNTP mix
    - 1 µl 100 µM GexPCR primer A
    - 1 µl 100 µM GexPCR primer B
    - 35 µl ddH_2O
    - 1 µl 5 U/µl Platinum Taq
    - 0.5 µl 10× SYBR Green
    - 1 µl library

    Amplification conditions:

1 cycle:	94°C for 1 min
24 cycles:	94°C for 30 sec
	57°C for 30 sec
	72°C for 1 min
1 cycle:	72°C for 5 min

40. Determine the optimal cycle number for final amplification. An optimal cycle can be chosen at the mid-upper part of the linear phase of the curve. Run on a 4% agarose gel to confirm that the size of the amplified product is 88 bp.

41. Prepare a new reaction mix as in step 39, but omit SYBR Green and increase water to 35.5 µl. Perform final library amplification using the predetermined optimal

number of cycles. At the same time, amplify the ligase minus control (step 31) for 40 cycles using the same conditions.

42. Run the control on a 4% agarose gel to assess contamination. If an 88-bp band is present, follow steps in the Table 2.1.
43. Add 10 µl of 6× BlueJuice gel loading buffer to the library and load in three lanes of a 20% TBE gel. Load low-molecular-weight ladder in a separate lane and electrophorese.
44. Repeat steps 32 to 38 to extract the 88-bp band (Fig. 2.2B).

    The added 22-bp sequence is used during sequencing.
45. Use a 1-µl aliquot to assess concentration on a Qubit fluorometer, and submit ~0.6 ng/µl for sequencing.

    Only 17 cycles are needed when the following sequencing primer is used:

    5'-CCGACAGGTTCAGAGTTCTACAGTCCGACATG

Basic Protocol 2: Generation and Comparison of Gene Expression Profiles

For the purpose of this analysis, it is assumed that the DNA sequences are available in FASTQ format. To recapitulate the full-length cDNA tag sequence, the NlaIII site is appended to the beginning of DNA sequences. The resulting 21-bp tag is aligned first to the genome and then the transcriptome using Bowtie and Tophat (Langmead et al., 2009; Trapnell et al., 2009). Here, Tophat is provided with a list of known splice junctions to enable alignment to the transcriptome. These junctions can be generated from a UCSC annotation table (see Support Protocol). The number of sequence tags aligned to each gene transcript is then tallied to generate gene expression profiles.

Sample profiles are normalized to one million gene-aligned tags. Between each pair of samples, the fold-difference and significance of the difference are calculated (Audic and Claverie, 1997). An overview of the analysis pipeline is shown in Figure 2.3.

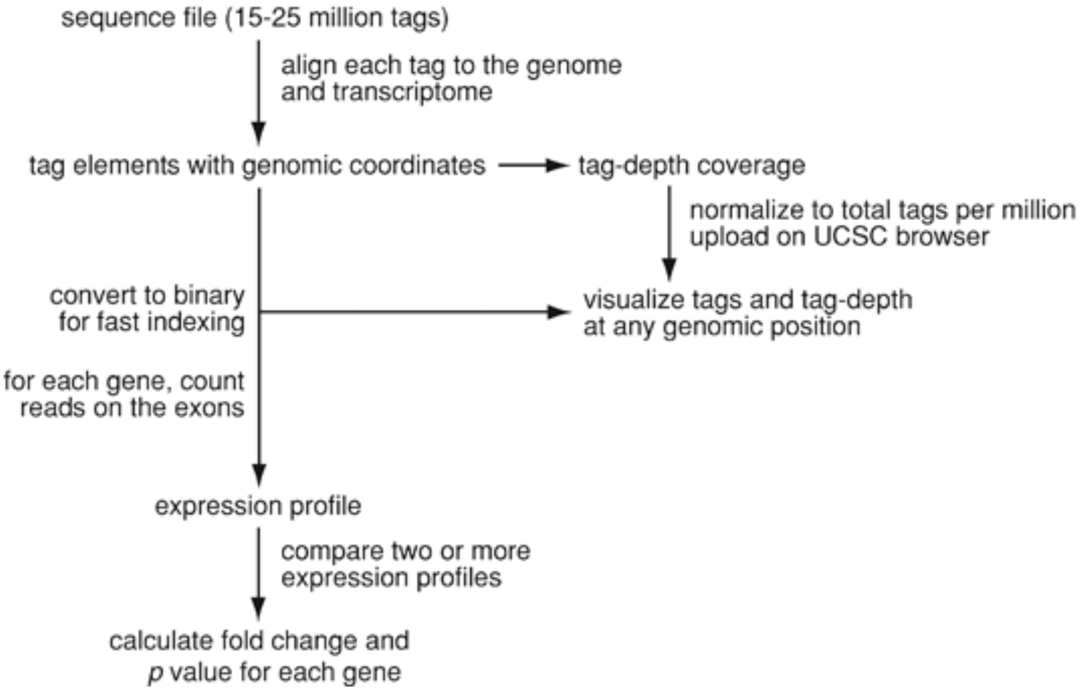

sequence file (15-25 million tags)
align each tag to the genome and transcriptome
tag elements with genomic coordinates
tag-depth coverage
normalize to total tags per million upload on UCSC browser
convert to binary for fast indexing
visualize tags and tag-depth at any genomic position
for each gene, count reads on the exons
expression profile
compare two or more expression profiles
calculate fold change and *p* value for each gene

Figure 2.3 Data analysis pipeline. Sequenced tags in the FASTQ file are aligned to the genome and transcriptome. Gene expression profiles are generated by tallying the aligned tags for each gene. The resulting expression profiles can be compared to assess fold change and significance in gene expression.

Materials

- Software pre-requisites (if using versions other than indicated, test for compatibility):
  - Bowtie 0.12.6
  - Samtools 0.1.11
  - Tophat 1.1.4
  - BioPerl-1.6.1

- Bio-SamTools 1.24
  - Perl v5.10.0
  - BioPerl module for SAGE comparison
  - Integrative Genomics Viewer
- Analysis software package includes:
  - Analysis programs written in Perl
  - Reference files
- Software package requires a Unix-based computer; a computing cluster is strongly recommended (run time ~1-2 hr using a 2.4-GHz computing node)

1. Download and install the following programs:
    1. Bowtie: http://bowtie-bio.sourceforge.net/index.shtml. Follow installation instructions and add the program to the PATH.
    2. Samtools: http://samtools.sourceforge.net/ (Li et al., 2009). Add the installed program to the PATH.
    3. Tophat: http://tophat.cbcb.umd.edu/. Follow the directions for working with Samtools. Add the installed program to the PATH.
    4. BioPerl: http://www.bioperl.org/.
    5. Bio-Samtools module: http://search.cpan.org/~lds/Bio-SamTools/. Add the module's path to the PERL library environment variable. (Use PERL5LIB for Perl 5 if the module cannot be installed in the standard BioPerl location.)
    6. Perl: http://www.perl.org/.
    7. BioPerl module for SAGE comparison by algorithm of Audic and Claverie,

    http://search.cpan.org/~scottzed/Bio-SAGE-Comparison-1.00/. Add the module's path to the PERL library environment variable (as above).

    8. Integrative Genomics Viewer (IGV): http://www.broadinstitute.org/igv.

2. Download the analysis package: http://seidman.med.harvard.edu/gs/DSAGE/. Unpack the files in a new directory (substitute “/dir/” below with the directory location).

   ```
   tar xvfz <downloaded package>
   ```

   Add the /dir/DSAGE/bin to the PATH.

   Add the following as environment variables to ~/.bash profile (assumes the bash shell is used).

   ```
   export DSAGE_TABLES=/dir/DSAGE/tables
   ```

3. Append the NlaIII sequence. Run:

   ```
   appendNlaIII.pl <Fastq file>
   ```

   The output will be <Fastq file>.NlaIII.

4. Align the tags from the appended FASTQ file with Tophat allowing 0 mismatches and by supplying known junctions:

   Run Tophat in the same folder (best if the job is submitted by bsub if a computer cluster is used).

   ```
   tophat –solexa1.3-quals -o <Sample name_folder> --segment-mismatches 0 -j
   $DSAGE_TABLES/<SPECIES>.juncs <REFGENOME> <Fastq file>.NlaIII
   ```

   For <REF GENOME>, follow the instructions on the bowtie website to generate or download the genome.

   In steps 4-7, for <SPECIES>, use mm9 for mouse, hg19 for human, or galGal3 for chicken. For <Sample name>, use a one-word description of the sample.

The output is a binary sequence alignment/map file (.bam).

5. Create an index file. Inside the Sample folder run:

    - samtools index accepted_hits.bam

6. Make a coverage file using the ‘wiggles’ program from the Tophat bin folder. Then normalize, name, and compress the coverage file:

    - samtools view accepted_hits.bam > accepted_hits.sam
    - wiggles accepted_hits.sam coverage.wig
    - normwig.pl coverage.wig <Sample name>

    The output file is <Sample_name>_norm.wig.gz, and can be uploaded to the UCSC browser as a custom track. The BAM file can also be uploaded.

    For more information on how to generate custom tracks, see http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html#CustomTracks.

    For genes that are highly expressed, IGV is recommended.

    At this stage, the aligned tags or the tag depth can be visualized on the UCSC Genome Browser (UNIT 19.9) as illustrated in Figure 2.4.

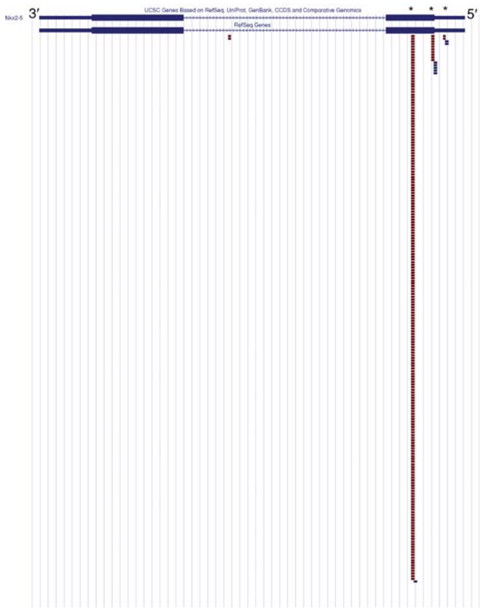


Figure 2.4 Tags aligned to the Nkx2-5 locus from mouse left ventricle. Nkx2-5 has three NlaIII sites found in the first exon (denoted by an asterisk). The majority of the sequence tags were aligned to the most 3’ NlaIII site, indicating that digestion with NlaIII was nearly complete.

7. Count sequence tags aligned to each gene transcript to construct the expression profile.

    - countreads.pl accepted_hits.bam $DSAGE_TABLES/<SPECIES>.ann

    The output file, which is the expression profile of the sample, is “genes.counts”.

8. Assemble, normalize, and compare expression profiles. Also append gene annotation. Read counts are normalized to one million reads.

    If running this command in the directory above the Tophat folders and if comparing two samples, run:

    expression_analyze.pl $DSAGE_TABLES/<SPECIES>.ann <Sample1> <Sample2>

    where the sample name is the name given previously to the Tophat folder.

    For multiple samples, run:

    expression_analyze.pl $DSAGE_TABLES/<SPECIES>.ann <Sample1> <Sample2>

```
<Sample 3> [       ] <Sample N>
```

In case the comparison involves samples run at different parent directories, the location of the Tophat folder can be included for any or all samples following the sample name. The location should be preceded with “dir:” (no spaces):

```
expression analyze.pl $DSAGE_TABLES/<SPECIES>.ann <Sample 1> dir:</dir1>

<Sample 2> dir:</dir2> [       ] <Sample N> dir:</dirN>
```

The output file is called <Sample1 Sample2.. SampleN>.expr.

Alternate Protocol: Automated Generation of Expression Profiles for Multiple Samples

While a step-by-step approach is needed to test individual components of the process (e.g., when running the analysis pipeline for the first time), automation is essential for the analysis of multiple samples (e.g., to troubleshoot or make modifications), as it can save time and minimize errors. This protocol offers tools to automate the generation of expression profiles for multiple samples into a single step.

1. Create a tab-delimited text input file, e.g., ‘Samples.txt’
    - First column: FASTQ file (full path if not running in the same directory)
    - Second column: Sample name
    - Third column: Species (e.g., mm9, hg19, galGal3).

    The sample name must match the Bowtie index genome.
2. To execute steps 3-7 from Basic Protocol 2:

- DSAGEanalyze_nobsub.pl <Text_file from Step 1>
- Example perl script for batching process to linux cluster:
- DSAGEanalyze_bsub.pl <Text_file from Step 1>

3. For sample comparisons and normalization of the data, see Basic Protocol 2, step 8.

Support Protocol: Generation of Gene-Based Annotation Tables from Expanded Transcript Tables

The steps outlined below can be used to make reference files for other species, or to remake the annotation tables to utilize updated gene annotations and/or genome assemblies.

1. Download a gene annotation table from the UCSC Genome Browser (http://genome.ucsc.edu/; also see UNIT 19.9).
    - Click the “Tables” link to proceed to the Table Browser.
    - Define genome and assembly.
    - Set group to ‘Genes and Gene Prediction Tracks'.
    - Set track to ‘UCSC genes'.
    - Set table to ‘knownGene’.
    - Set output to ‘selected fields from primary and related tables'.
    - Click ‘get output’.
    - Check the boxes for the first ten fields from knownGene (abbreviated: UCSC name, chr, strand, txstart, txend, cdsstart, cdsend, noexons, exonstarts,

exonends).

- Check the boxes for ‘geneSymbol’ and ‘Description’ from kgXref.
- Click ‘get output’ again.

2. To make a junctions file:

   makejunctions.pl <INPUT_UCSC_TABLE> <SPECIES>

3. To make a gene annotation file for gene expression using Refseq names:

   makeannexpression.pl <INPUT_UCSC_TABLE> <SPECIES>

4. Move the two generated files to the $DSAGE_TABLES folder.

   The <SPECIES> as named here should be used in Basic Protocol 2 and Alternate Protocol.

Reagents and Solutions

Use deionized, distilled water in all recipes and protocol steps. For common stock solutions, see APPENDIX 2; for suppliers, see APPENDIX 4.

Adapters 1 and 2

- Adapter set 1:
- 5’-p-TCGGACTGTAGAACTCTGAAC-NH_2
- 5’-ACAGGTTCAGAGTTCTACAGTCCGACATG
- Adapter set 2:
- 5’-CAAGCAGAAGACGGCATACGANN
- 5’-p-TCGTATGCCGTCTTCTGCTTG- NH_2

Adapter oligonucleotides were obtained from Integrated DNA Technologies.

For each adapter set, anneal oligos (each at 50 µM with LoTE) in a thermocycler using the following program:

- 2 min at 95°C
- – 0.1°C/sec to 65°C, then 10 min at 65°C
- – 0.1°C/sec to 37°C, then 10 min at 37°C
- – 0.1°/sec to 25°C, then 20 min at 25°C
- hold at 4°C.
- Store annealed adapters at –20°C until use (up to 1 month).

Bind and wash (BW) buffer, 2×

- 200 ml 5 M NaCl
- 2.5 ml 2 M Tris-Cl, pH 7.5 (APPENDIX 2)
- 1 ml 0.5 M EDTA (APPENDIX 2)
- 296.5 ml ddH_2O
- Store up to several months at 4°C

Buffer C

- 5 ml 10% SDS
- 25 ml 2× BW buffer (see recipe)
- 50 µl glycogen
- 20 ml ddH_2O
- Store up to several months at 4°C
- Heat prior use to dissolve SDS

Buffer D

- 25 ml 2× BW buffer (see recipe)
- 1 ml 100× BSA (New England Biolabs)
- 24 ml ddH_2O
- Store up to several months at 4°C

Buffer E

- 5 ml NEBuffer 4 (New England Biolabs)
- 1 ml 100× BSA (New England Biolabs)
- 44 ml ddH_2O
- Store up to several months at 4°C

LoTE buffer

- 150 µl 1 M Tris-Cl, pH 7.5 (APPENDIX 2)
- 20 µl 0.5 M EDTA (APPENDIX 2)
- 49.83 ml ddH_2O
- Store up to several months at 4°C

MGB buffer, 10×

- 8.3 ml 1 M $(NH_4)_2SO_4$
- 16.75 ml 2 M Tris-Cl, pH 8.8 (APPENDIX 2)
- 3.35 ml 1 M $MgCl_2$
- 0.351 ml β-mercaptoethanol
- 21.25 ml ddH_2O
- Store up to several months at 4°C

Commentary

Background Information

This unit describes a protocol for performing deep sequencing analysis of gene expression (DSAGE), a sequencing-based approach to quantifying gene transcript expression. Analog methods that involve hybridization of cDNA to microarrays of oligonucleotides have been used extensively to assess gene expression. However, such approaches only provide relative measurements with a small dynamic range, are limited to a set of known oligonucleotide probe sequences, and are hindered by hybridization biases (Draghici et al., 2006). In contrast, digital sequencing approaches to studying gene expression involve unbiased counting of observations of gene transcripts, offering quantification and more comprehensive analyses. Early sequencing-based methods, such as serial analysis of gene expression (SAGE; Velculescu et al., 1995; UNIT 25B.6), were limited in their depth to ~100,000 tags because of the high cost of Sanger dideoxy sequencing. We previously described polony multiplex analysis of gene expression (PMAGE), which utilizes a novel massively parallel sequencing method that facilitates a dramatic increase in sequencing throughput (to 5 million tags) and a substantial reduction in cost (Kim et al., 2007). DSAGE was adapted from PMAGE to use the Illumina Genome Analyzer II. Currently, 15 to 25 million tags can be sequenced with one lane using the current Illumina analyzer, and this number is dependent only on the state of the sequencing technology used.

Analysis tools developed for RNA-seq (UNIT 4.11), Tophat (Trapnell et al., 2009), and Bowtie (Langmead et al., 2009) efficiently assign DSAGE tags to their corresponding genes. This approach facilitates the integration of DSAGE data with other data types, including RNA-seq and genomic variation, and is able to map tags across exon splice junctions. This approach provides flexibility in the reference transcriptome to which tags are mapped, permitting study of organisms with incomplete genomes or gene sequences. The protocol described in this unit uses transcript annotations from the University of California Santa Cruz (UCSC) Genome Browser (Kent et al., 2002; UNIT 19.9), downloaded with the UCSC browser retrieval tool (Karolchik et al., 2004).

Critical Parameters and Troubleshooting

Table 2.1 describes some common problems encountered with DSAGE, as well as possible causes and recommended solutions.

Table 2.1 Troubleshooting DSAGE Library Construction and Data Analysis

Problem	Possible cause	Solution
PCR for no-ligase control produced an 88-bp band	Contamination from another library	Replace affected reagents. Keep separate bench areas (and materials) for pre- and post-amplification reactions.
The 66-bp pre-amplification library is not visible	Low yield	Use the marker to excise the 66-bp product.
No signal for low-expressing genes	Low-complexity library or insufficient sequence depth	Check activity of enzymes and purification steps. The amount of material before amplification correlates with the complexity of the

Table 2.1 (Continued)

		library. A total of 10-15 cycles should be sufficient for amplification.
Genes in the gene profiles display mostly null values	Genome assembly used for alignment may not match the annotation tables	Make sure that the genome assembly used for alignment matches the assembly of the reference tables. The reference tables provided with the analysis package are for assemblies hg19, mm9, and galGal3.
No signal for a gene that is expressed in the tissue	A small number of mRNAs do not have an NlaIII site	Verify that the gene has an NlaIII site by retrieving the mRNA sequence
Gene profiles of similar biological samples do not correlate strongly	Poor RNA quality	Use RNA with RNA integrity number (RIN) > 8
	Library may not be uniformly amplified	Keep amplification cycles within the exponential range
	Low-complexity library	Increase yield (see above)
	Contamination from another library	See above
Some program commands fail to return an output	Software may be incorrectly set up. Also insufficient memory or other problems with the computing node.	Test-run software using a small dataset

The key element in constructing a DSAGE library is the starting quality of the RNA. RNA with an RNA Integrity Number (Agilent) greater than 8 may be used for optimal results. The yield depends not only on the quality and quantity of the RNA used, but also on the state of the materials. An indicator of the yield is the presence of the 66-bp pre-amplification library, which is expected to be visible on the polyacrylamide gel after SYBR green staining. It should be noted that inability to visualize the band should not prompt discontinuation of the library preparation. In that case, the marker may be used to excise the band, carefully avoiding contamination with adapter dimer.

Contamination of the pre-amplified library with an amplified library can have magnified effects. Since DNA is stable at room temperature, avoiding such contamination requires maintaining a separate lab bench area to be used when processing the amplified library. Similarly, the instruments used for the amplified library (including refrigerator and freezer areas) should be marked as “post-PCR” and avoided when handling pre-amplified libraries. Any materials can be tested for contamination by performing the PCR amplification reaction as described for the no-ligase control.

The amount of the library obtained following amplification depends on the final amplification step and extraction from the gel. Cycles in the mid to upper part of the log-phase of the reaction should be sufficient to generate enough material to be used for sequencing (the cycles should never exceed the log portion of the reaction). Also, special care should be taken to crush the extracted polyacrylamide gel into fine pieces. If crushing is incomplete (i.e., the gel remains grainy), the process can be repeated. Lower-yield libraries may require additional cycles to amplify to sufficient levels for sequencing; however, the number of amplification cycles is not expected to exceed 15. Libraries requiring more amplification cycles would display low complexity when sequenced, indicating that the dynamic capture of the low-expressing genes would be compromised. Test the efficiency of the enzymes and purification steps. Note that NlaIII must be kept at –80°C for long-term storage (>3 months).

The software package pre-requisites must be installed and tested according to the developers' instructions. Although we recommend using Tophat version 10, as it appears to function best on our cluster with 21-bp reads, later versions can also be tested and used.

If the libraries are constructed properly, technical replicates are expected to yield highly reproducible results (r > 0.99). Similar biological samples are also expected to strongly correlate, as biological noise can be minimized by pooling RNA from biological replicates. If the expression profiles of such samples do not strongly correlate, the possibilities of poor RNA quality, poor amplification, low complexity of the amplified library, and the presence of contamination should be addressed.

As with PMAGE and SAGE, DSAGE requires the presence of an NlaIII site on the transcript. An NlaIII site is present on the majority of transcripts, enabling almost complete profiles to be generated. To verify, the mRNA sequence can be retrieved and queried for the NlaIII site.

In this analysis protocol, it is strongly recommended that alignment not allow any mismatches between the tag and the reference genome used. As a consequence, perfect matches are used to construct the transcriptional profiles. A complication of this, however, occurs when a polymorphism is present on the tag, in which case the tag will not be able to align to the genome. This may become an issue when comparing samples of different biological origin. However unlikely, this can be detected by identifying the NlaIII site on the transcript and identifying polymorphisms using the UCSC Genome Browser. Alternatively, Bowtie and Tophat could align the reads while allowing one or two mismatches.

Anticipated Results

The 66-bp library is expected to be visible on a polyacrylamide gel after SYBR Green staining when the yield of the previous reactions is optimal. For a library with good complexity, less than 15 cycles are required for amplification. The 88-bp amplified library should always be visible. Beyond primer dimers, nonspecific amplification products should not be present.

Current Illumina sequencers are expected to sequence more than 10 million tags. As an example, the authors typically observe expression of >10,000 genes in the mouse heart.

Time Considerations

Library preparation is expected to take 4 days. Sequencing is expected to take ~3 days, not taking into account scheduling queues at the sequencing facility. Data analysis should be completed within 1 day, given that run time is 1 hr per sample and parallel processing can be used with a computer cluster.

Acknowledgments

D.C. Christodoulou, J.M. Gorham, and M. Kawana contributed equally to this work. The authors are grateful to the principal investigators, Christine E. and Jonathan G. Seidman. This work was funded by SysCODE Grant #NIH PL1 HL092552 (C.E. and J.G. Seidman).

Literature Cited

- *Audic, S. and Claverie, J.M. 1997. The significance of digital gene expression profiles. Genome Res. 7:986-995.*

- *Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53-59.*

- *Draghici, S., Khatri, P., Eklund, A.C., and Szallasi, Z. 2006. Reliability and reproducibility issues in DNA microarray measurements. Trends Genet. 22:101-109.*

- *Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 32:D493-D496.*

- *Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. Genome Res. 12:996-1006.*

- *Kim, J.B., Porreca, G.J., Song, L., Greenway, S.C., Gorham, J.M., Church, G.M., Seidman, C.E., and Seidman, J.G. 2007. Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. Science 316:1481-1484.*

- *Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10:R25.*

- *Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R.; 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-2079.*

- *Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., Lightfoot, S., Menzel, W., Granzow, M., and Ragg, T. 2006. The RIN: An RNA integrity number for assigning integrity values to RNA measurements. BMC Mol. Biol. 7:3.*

- *Trapnell, C., Pachter, L., and Salzberg, S.L. 2009. TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics 25:1105-1111.*

- *Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. 1995. Serial analysis of gene expression. Science 270:484-487.*

Author contributions:

DCC wrote computational programs

JMG, MK, HW wrote molecular biology steps

SRD, DSH provided important computational information

HW, JGS, CES directed the project

CHAPTER 3

Chapter 3. Construction of Normalized RNA-seq Libraries for Next-Generation Sequencing Using the Crab Duplex-Specific Nuclease:

Published in Current Protocols of Molecular Biology (Curr Protoc Mol Biol. 2011 Apr;Chapter 4:Unit4.12.). Reproduced with permission of Current Protocols and John Wiley & Sons, Inc. Copyright © 2011 by John Wiley & Sons, Inc.

Danos C. Christodoulou, Joshua M. Gorham, Daniel S. Herman, J.G. Seidman

This unit describes the generation of a normalized RNA-seq library for next-generation sequencing by utilizing the preference of the crab duplex nuclease (DSN) for digesting double-stranded, rather than single-stranded DNA (UNIT 5.12; also see Zhulidov et al., 2004). In this approach (Basic Protocol), polyadenylated RNA is used to generate a complex RNA-seq library. The library is denatured and incompletely renatured, and then digested with DSN. The kinetics of DNA annealing are such that at any given time abundant DNA molecules are more likely to have re-annealed and become double stranded, while rare molecules are more likely to remain single stranded. Thus, preferentially digesting double-stranded DNA with DSN yields a library markedly enriched for more rare DNA species. Massively parallel sequencing of these normalized libraries allows for efficient and comprehensive assessment of the sequence and structure of the polyadenylated transcriptome.

The first step in the construction of a normalized RNA-seq library is to make a high-complexity RNA-seq library (i.e., a library that includes sufficient depth of sequence from all positions of all polyadenylated RNAs). This involves starting with ~40 to 400 ng polyadenylated RNA from ~2 to 20 μg total RNA. This method should also be suitable for non-polyadenylated RNA, in which case the protocol would be started at step 5. The quality of the RNA is crucial and can be determined using a Bioanalyzer (Agilent) (Schroeder et al., 2006) or an agarose gel. All steps in the protocol are performed with an excess amount of reagents to ensure each reaction approaches completion.

The resulting high-complexity RNA-seq library is amplified to generate 500 to 1200 ng DNA using 19- to 20-bp primers corresponding to the inner sequences of the Illumina Paired-End adapters. Normalization of the RNA-seq library is then achieved by denaturing and re-annealing at 68°C, followed by treatment with the duplex-specific nuclease (DSN). The annealing temperature of the resulting amplified RNA-seq library's adapter sequence is ~57°C, which minimizes nonspecific targeting by DSN. A second amplification and normalization is then performed. The normalized library is then amplified with the full-length Illumina paired-end primers. See Figure 3.1 for an overview of the library construction.

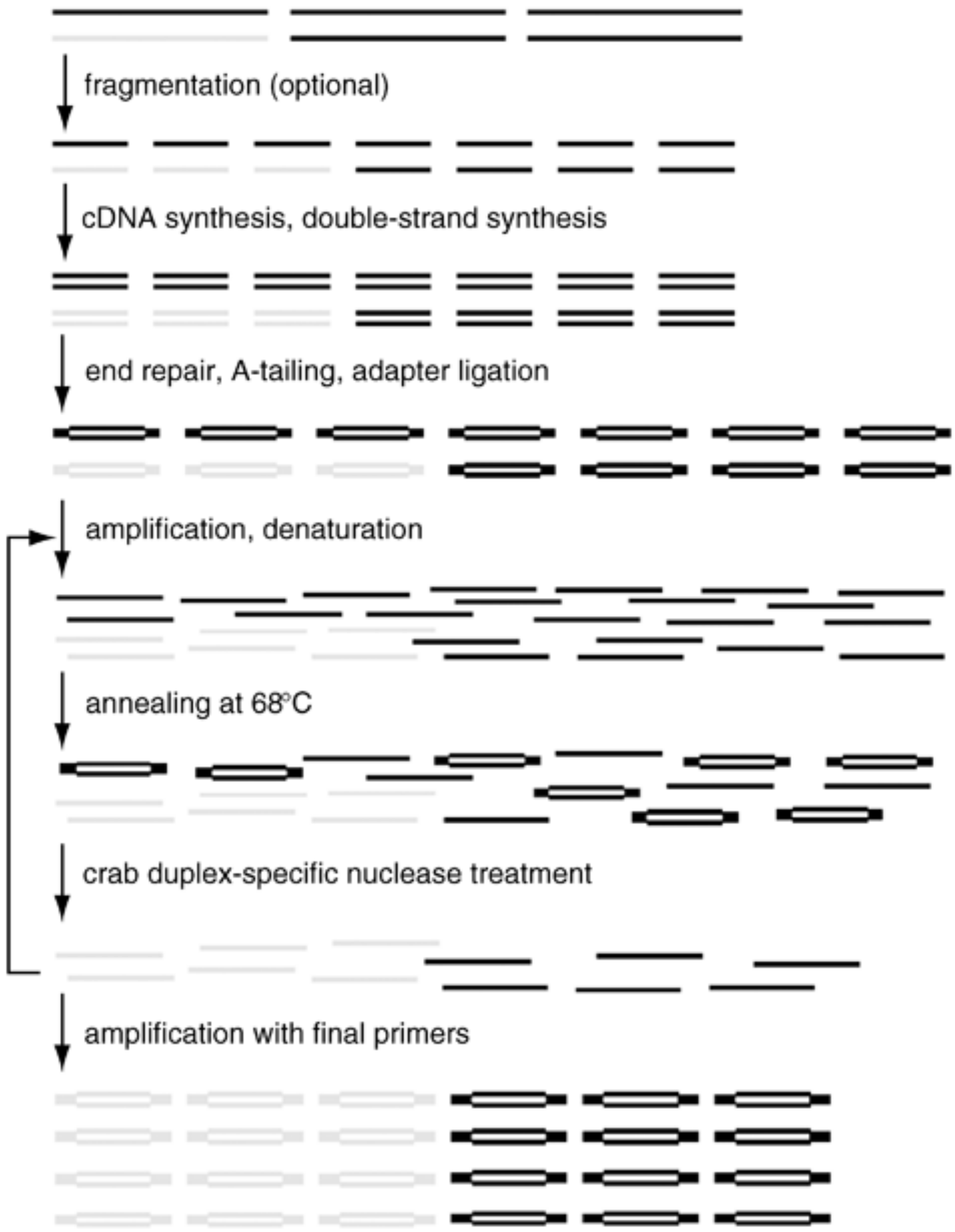

fragmentation (optional)
cDNA synthesis, double-strand synthesis
end repair, A-tailing, adapter ligation
amplification, denaturation
annealing at 68°C
crab duplex-specific nuclease treatment
amplification with final primers

Figure 3.1 Normalized RNA-seq library construction. Illustration of the normalization process starting from a rare (gray) and abundant (black) RNA transcript.

Materials

- Dynabeads mRNA DIRECT kit (Invitrogen, cat. no. 610-12) containing:
    - Lysis/binding buffer

  - Washing buffer B
- DynaMag-2 magnet (Invitrogen, cat. no. 123-21D)
- RNA, purified using TRIzol (Invitrogen, cat. no 15596-018) or RNeasy kit (Qiagen, cat. no. 74104)
- Quant-iT RNA assay kit (Invitrogen, cat. no. Q32852)
- 5× fragmentation buffer (see recipe)
- 3 M sodium acetate, pH 5.7
- Glycogen (Roche, cat. no. 10901393001)
- 70% (v/v) and 100% ethanol
- SuperScript III cDNA synthesis kit (Invitrogen, cat. no. 18080-051) containing:
  - 10 mM dNTPs
  - 0.1 M DTT
  - 10× RT buffer
  - RNase OUT
  - SuperScript III reverse transcriptase
  - Random hexamers
  - 25 mM $MgCl_2$
  - *E. coli* RNase H
- Second-strand buffer (Invitrogen, cat, no. 10812-014)
- *E. coli* DNA polymerase I (Invitrogen, cat. no. 18010025)
- QIAquick PCR purification kit (Qiagen, cat. no. 28104) containing:
  - Buffer EB

- End-It DNA End-Repair kit (Epicentre Biotechnologies, cat. no. ER81050) containing:
    - 10× End-repair buffer
    - ATP
    - dNTPs
    - End-repair enzyme mix
- dATP (Roche, cat. no. 11051440001)
- NEB2 buffer (New England Biolabs)
- 3'->5' KlenowExo (New England Biolabs, cat. no. M0212s)
- Quick Ligation kit (New England BioLabs, cat. no. M2200L) containing:
    - 2× ligation buffer
    - Quick T4 DNA ligase
- MinElute PCR purification kit (Qiagen, cat. no. 28004)
- 50-bp ladder (New England Biolabs, cat. no. N3236L)
- 25-bp ladder (Invitrogen, cat. no. 10597-011)
- SYBR Gold (Invitrogen, cat. no. S-11494)
- QIAquick Gel Extraction Kit (Qiagen, cat. no. 28004)
- Phusion High-Fidelity DNA Polymerase (New England Biolabs, cat. no. F-530S) containing 5× HF Phusion buffer
- Oligonucleotides (see Table 3.1)
- SYBR Green I (Invitrogen, cat. no. S7563)
- AMPure beads (Agencourt AMPure kit, cat. no. A29152)
- Quant-iT dsDNA HS Assay kit (Invitrogen, cat. no. Q32851)

- 4% to 20% TBE gel (Invitrogen, cat. no. EC62252BOX)
- 4× hybridization buffer (see recipe)
- Duplex-specific nuclease (Evrogen, cat. no. EVN-EA001-KI01) containing:
    - 10× master buffer
    - DSN enzyme
- EDTA
- UV transilluminator
- Vortex
- 1.5-ml nonstick RNase-free microcentrifuge tubes (Ambion, cat. no. 12450)
- Heat block or incubator at 61° to 73°C
- Qubit fluorometer (Invitrogen, cat. no. Q32857)
- Dark Reader Transilluminator (Clarechemical, cat. no. DR-88M)
- Thermal cycler
- Additional reagents and equipment for agarose gel electrophoresis (UNIT 2.5A)

Table 3.1 Oligonucleotide Sequences

Oligonucleotide	**Sequence**
PE adapter 1[a]	5’phosphate-GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
PE adapter 2[a]	5’ACACTCTTTCCCTACACGACGCTCTTCCGATCT
Inner PE primer1	5’CACGACGCTCTTCCGATCT
Inner PE primer2	5’CTGAACCGCTCTTCCGATCT

Table 3.1 (Continued)

Final PE primer 1	5’AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
Final PE primer 2	5’CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

To prepare the adapter, resuspend each adapter oligonucleotide to 200 µM using water or reduced Tris-EDTA buffer. Mix equal amounts of each adapter oligonucleotide. Using a thermal cycler, anneal using the program: 4 min at 95°C, –0.1°C/sec to 12°C, hold at 4°C. The annealed adapters can be stored at –20°C until use.

Stage I: polyA selection, RNA fragmentation, cDNA and second-strand synthesis, end repair, A-addition, adapter ligation, gel purification, and first amplification

Perform polyA selection

1. Vortex the oligo(dT) beads (Dynabeads) at a low speed to make homogeneous. Transfer 100 µl into a 1.5-ml tube (nonstick tubes are used throughout the experiment). Capture the beads by placing the tube on the DynaMag-2 magnet, pipet off most of the solution, and wash with 50 µl lysis/binding buffer (Dynabeads kit). Repeat the capture and liquid removal, and resuspend in 50 µl lysis/binding buffer.
2. Prepare the RNA on ice (previously extracted using an RNeasy kit or TRIzol and quantified) by adjusting the volume to 50 µl with ddH_2O. Add 50 µl lysis/binding buffer, heat at 61°C in a heat block for 2 min, and immediately place back on ice. Transfer the RNA to the washed beads and mix by pipetting. Incubate the binding reaction 10 min at room temperature with gentle agitation.

3. Capture the beads on the magnet and discard the supernatant. Wash twice with 200 µl buffer B (Dynabeads kit). Wash once with 50 µl washing buffer B, discard the buffer, and resuspend in 20 µl ddH$_2$O. Elute the RNA from the beads by incubating 1 to 2 min at 73°C in a heat block. Immediately capture the beads on the magnet and transfer the eluate to a new tube and place on ice.

   It is critical not to exceed the 73°C temperature during elution to prevent extensive RNA degradation.

4. Perform a second round of polyA selection by repeating steps 1 to 3 using the eluate from step 3 in place of total RNA in step 2 (adjust the volume to 50 µl by adding 30 µl of ddH$_2$O). As in step 3, elute into 20 µl of ddH$_2$O and place the tube on ice. Measure the mRNA concentration with the Quant-iT RNA assay kit and the Qubit fluorometer using 1 µl of the eluate.

   At this stage, the mRNA can be stored at –80°C until use.

RNA fragmentation (optional)

5. Use 40 to 400 ng mRNA, bring to a volume of 16 µl with ddH$_2$O, and add 4 µl of 5× fragmentation buffer. Incubate 2 to 5 min at 94°C in a thermal cycler and place on ice.

6. Precipitate the fragmented RNA by adding 8 µl 3 M sodium acetate, pH 5.7, and 1 µl of glycogen. Mix and then add 240 µl 100% ethanol. Vortex and incubate 30 to 50 min at –20°C. Pellet the precipitated RNA by centrifuging 15 min at maximum speed (>12,000 × g), 4°C. Remove the supernatant and wash the pellet with 800 µl 70% ethanol. Recentrifuge for 2 min at maximum speed (>12,000 × g), 4°C, completely

remove the ethanol, and dry the pellet for a few minutes at room temperature.

Synthesize the first strand

7. If continuing from step 6, resuspend the RNA pellet in 30 µl random hexamers (SuperScript III cDNA synthesis kit) and place on ice. If continuing from step 4, add 10 µl random hexamers to the polyA RNA to 30 µl. Add the following:

   - 8 µl 10 mM dNTPs
   - 2 µl ddH_2O.

   Mix and incubate 4 min at 65°C in a thermal cycler.

   Place on ice and add:

   - 8 µl DTT
   - 8 µl 10× RT buffer
   - 16 µl 25 mM $MgCl_2$
   - 2 µl RNase OUT
   - 6 µl SuperScript III RT.

   Incubate for 10 min at 25°C, 25 min at 42°C, 25 min at 50°C, and hold at 4°C in a thermal cycler.

Synthesize the second strand

8. Add the following as a mastermix (amounts per sample):

   - 204 µl ddH_2O
   - 80 µl 5× Second-strand buffer
   - 12 µl 10 mM dNTPs
   - 4 µl RNase H.

   Mix and add 15 µl DNA polymerase I. Mix again. Incubate 2.5 hr at 16°C and hold overnight at 4°C in a thermal cycler. Purify the resulting second-stranded product using a Qiagen PCR purification kit according to the kit's instructions. Elute with 90 µl buffer EB.

Following this step, the product can be stored overnight at 4°C or up to 6 months at –20°C.

Perform end repair

9. Using the end-repair kit, prepare a mastermix containing the following (amount per sample):
   - 20 µl 10× end-repair buffer
   - 20 µl 10 mM ATP
   - 20 µl dNTPs
   - 6 µl end-repair enzyme mix
   - 89 µl ddH_2O.

   Add 110 µl of end-repair mix to the 90-µl sample eluate from step 8 and incubate at 55 min 20°C and hold at 4°C in a thermal cycler. Use the QIAquick PCR kit to purify the sample and elute into 80 µl EB.

10. For A-tailing, prepare a mastermix containing the following and add to the sample:
    - 10 µl NEB2 buffer
    - 4 µl 5 mM dATP
    - 6 µl 3'-5' KlenowExo.

    Incubate 1 hr at 37°C in a thermal cycler. Use the QIAquick PCR kit to purify the sample and elute into 44 µl EB.

    A complete reaction is critical as A-tailing facilitates the addition of an adenosine overhang at the 3’ end, which prevents self-ligation during the next step.

Perform adapter ligation

11. First add 6 µl of 100 µM PE adapter and 56 µl 2× ligation buffer (NEB Quick Ligation kit) to the sample and mix. Add 6 µl quick ligase and mix well. Incubate 20 min at room temperature. Use the Qiagen minElute PCR purification kit according to the manufacturer's instructions and elute into 10 µl EB.

    Double eluting by passing the first eluate through the same column is recommended.

Perform the gel purification

12. Add gel loading dye to the above eluate and run on a 2% agarose gel (UNIT 2.5A) using the 50-bp ladder. Stain the gel with SYBR gold (1:10,000) for 10 to 15 min at room temperature and visualize with the Dark Reader Transilluminator.

    SYBR gold offers improved sensitivity compared to ethidium bromide and the transilluminator avoids exposing the DNA to UV radiation.

13. Extract a 100-bp band size between 180- to 350-bp (Fig. 3.2A). Use the Qiagen Gel extraction kit according to the manufacturer's instructions to extract the DNA and elute with 25 µl buffer EB. Pass the eluate through the column to double-elute.

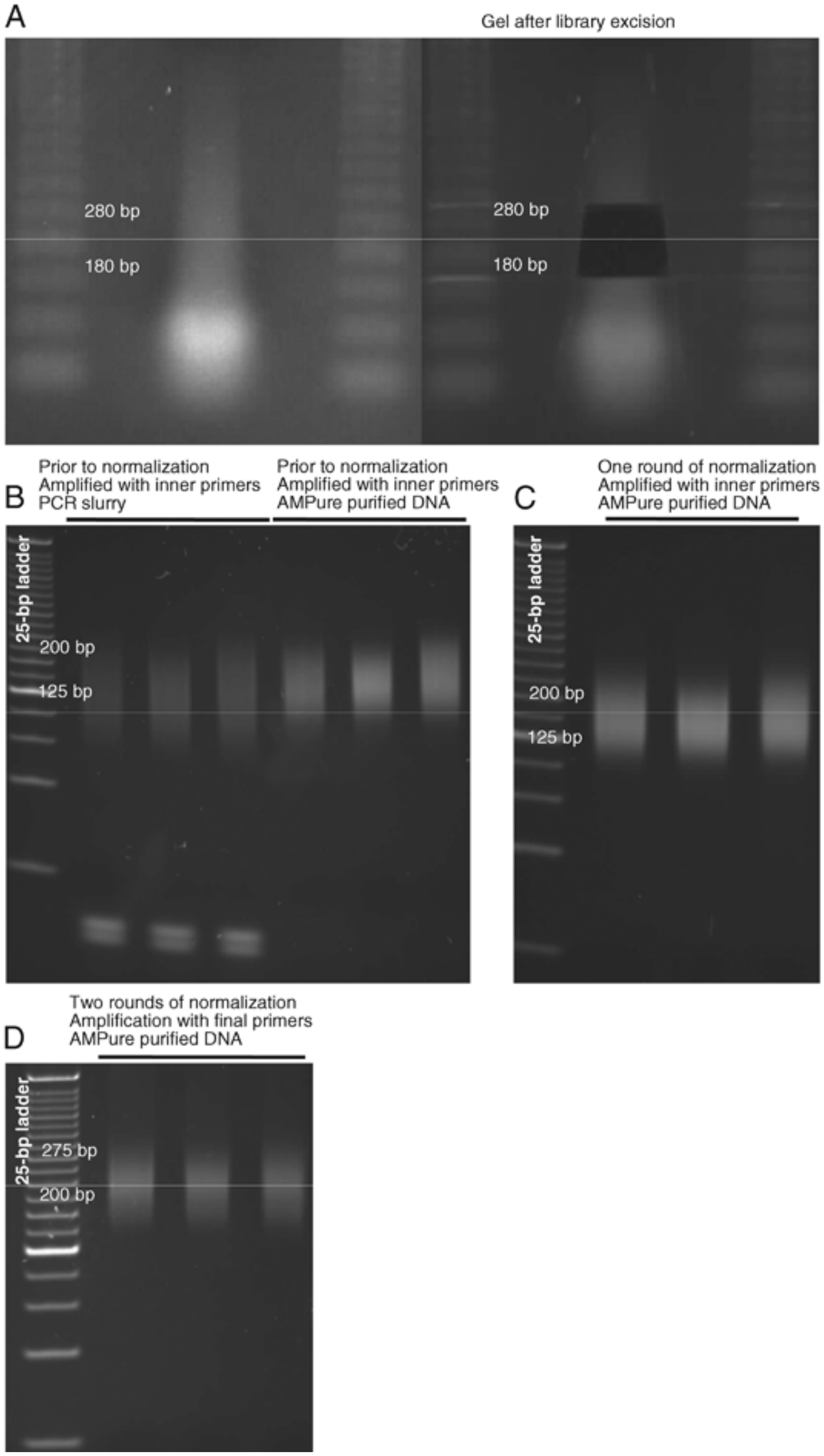


Figure 3.2 (A) Library excision from a 2% agarose gel stained with SYBR gold (step 12). Here, an 180- to 280-bp fragment was excised. B-D. Aliquots from different stages of library preparation. (B) Amplified library prior to normalization (step 18), (C) library after one round of normalization (step 21), (D) library after two rounds of normalization amplified with the final Paired-End primers (step 22).

Perform the first amplification

14. For each sample, prepare the following PCR mix:

    - 5 µl 5× HF Phusion buffer
    - 1 µl of 10 µM Inner primer 1
    - 1 µl of 10 µM Inner primer 2
    - 1.25 µl 10 mM dNTPs
    - 0.4 µl Phusion polymerase
    - 0.25 µl SYBR Green I (diluted 1:1000)
    - 15.6 µl ddH$_2$O
    - 0.5 µl RNA-seq library (end-product of step 13).

15. Carry out the amplification in a thermal cycler using the following conditions:

Initial step:	30 sec	98°C	(initial denaturation)
24 cycles:	10 sec	98°C	(denaturation)
	30 sec	59°C	(annealing)
	30 sec	72°C	(extension)
1 cycle:	5 min	72°C	(final extension).

Determine the optimal cycle number for the final PCR amplification by selecting the point before the real-time PCR reaction saturates. That is, use the graphed fluorescence versus cycle number from the real-time PCR. The optimal cycle number is at the end of the exponential phase but before the reaction reaches a plateau.

Choosing cycles within the exponential range facilitates uniform amplification of the library and minimizes amplification errors.

16. Use the sample generated in step 13 to create twenty replicate reactions prepared as in steps 14 and 15 (substituting SYBR green with ddH_2O) with the number of cycles determined in step 15.

    This aims at amplifying about half the sample (the remainder can be saved to be used if needed to repeat the amplification or to make a nonnormalized library).

17. Following completion of the reaction, combine the product from all tubes. Save 1 to 2 µl to be run on the polyacrylamide gel (step 18), and purify the remainder using 2× volume AMPure beads according to the manufacturer's instructions. Elute with 20 µl ddH_2O.

    This setup creates enough material for normalization, while using most of the initial material for amplification. This aims at amplifying a high complexity library.

18. Dilute 1 µl from the purified product 1:4 with water. Use 1 µl of the diluted product to estimate the library concentration using the Quant-iT dsDNA HS Assay kit and the Qubit fluorometer. Run 1 to 2 µl of the diluted product on a 4% to 20% polyacrylamide gel (UNIT 2.5A; include a lane of a 1-µl aliquot from prior to the DNA purification). Stain with SYBR Gold (1:10,000) for 10 min and visualize using a UV transilluminator (Fig. 3.2B).

    It is critical that the library appears as a smear and not a discrete band.

Stage II: Perform first normalization, second amplification, purification, and second normalization

19. Bring 500 to 1200 ng amplified DNA to a volume of 12 µl with ddH_2O. Add 4 µl of 4× hybridization buffer, mix, denature at 98°C for 2 min and re-anneal at 68°C for 6 hr in a thermal cycler.
20. For this step always work near the thermal cycler and keep the sample on the thermal cycler at 68°C, except when spinning down the contents of the tube. First, prewarm the 2× master buffer (dilute from the 10× provided in the Duplex-specific nuclease kit) on the thermal cycler and add 20 µl to the sample. Mix and incubate for 10 min. Add 3 µl of DSN enzyme (reconstitute and test in accordance to the manufacturer's instructions), mix well, spin down with a microcentrifuge for a few seconds at room temperature, and incubate for 25 min at 68°C. Add 1 µl of 125 mM EDTA, mix well, spin down, and incubate for 5 min. Following this reaction, immediately place the tube on ice and freeze if needed.

    It is critical that the hybridization temperature (68°C) be maintained throughout this procedure. Note that the DSN enzyme may still be active—keep the reaction cold or frozen to avoid nonspecific degradation.
21. Follow steps 14 to 20 to re-amplify the library with the inner primers and re-normalize. Use 1 µl of the product slurry from step 20 as the template. Before normalizing, run a 1-µl aliquot on a polyacrylamide gel.

    The library should appear as a smear and not a discrete band when run on a polyacrylamide gel (Fig. 3.2C).

Stage III: Final amplification and purification

22. Use the final PE primers to amplify the product from step 21. Follow steps 14 to 18 to amplify the library using 1 µl of 5 µM final PE primer and 1 µl of the product slurry from step 21. Purify with 1.8× volume AMPure beads and elute with 25 µl ddH_2O. Measure the concentration and assess the DNA library size as in step 18.

    When running on a polyacrylamide gel, the library should appear as a smear (Fig. 3.2D). The higher size of the smear is due to the added sequences from using the full-length paired-end primers.

23. Submit the library for end-sequencing at a concentration recommended by the sequencing facility.

Reagents and Solutions

Use deionized, distilled water in all recipes and protocol steps. For common stock solutions, see APPENDIX 2 ; for suppliers, see APPENDIX 4.

Fragmentation buffer, 5×

- 200 mM Tris acetate, pH 8.2
- 500 mM potassium acetate
- 150 mM magnesium acetate
- Store up to 1 year at –20°C

Hybridization buffer, 4×

- 200 mM HEPES, pH 7.5
- 2 M NaCl
- Store up to 1 year at –20°C

Commentary

Background Information

RNA-seq is transforming our understanding of transcriptomes by revealing important information about transcript diversity, sequence, and structure. Comparison of these RNA transcriptomes between various cellular states has also facilitated study of transcript function and regulation (Cloonan et al., 2008; Mortazavi et al., 2008; Nagalakshmi et al., 2008; Sultan et al., 2008; Wang, 2008; Tang et al., 2009; Trapnell et al., 2010). However, despite the improved power of next-generation sequencing technologies, the very broad dynamic range of gene expression has been an obstacle to studies of RNA transcript structural changes in genes expressed at low levels, such as transcription factors, which, even at low levels, are of profound physiologic importance. Sequencing-based strategies to assess expression profiles, such as Polony Multiplex Analysis of Gene Expression (PMAGE) and Deep Sequencing Analysis of Gene Expression (DSAGE; UNIT 25B.9 and Kim et al., 2007), have demonstrated that in the cardiac left ventricle the 50 most highly expressed genes, which consist of mitochondrial, structural, sarcomere, and energy production genes, comprise greater than 40% of the polyadenylated RNAs. In addition, the most abundant RNA can comprise as much as 10% of the polyadenylated RNAs. By contrast, low-expressing RNAs are often found to comprise as few as 1 molecule per million, indicating that the range of RNA transcript expression levels spans at least five orders of magnitude. This large range impedes comprehensive characterization of transcripts expressed at low levels.

Critical Parameters and Troubleshooting

An important consideration in constructing a normalized RNA-seq library is the complexity of the library prior to normalization. Lower complexity libraries can still be normalized, but will result in non-uniform sequencing of lower-expressing genes, as well as a high number of sequenced duplicate reads.

Library cross-contamination is a major concern in this procedure, because such problems will be exaggerated by the normalization and PCR amplification steps. Contamination can be minimized by physically separating amplification steps from post-amplified products and washing any laboratory equipment that comes in direct contact with the libraries, especially gel chambers and gel excision equipment.

RNA degradation is also a significant concern. During RNA isolation steps, significant effort is required to avoid RNA degradation. In addition to working quickly, keeping the RNA on ice and using RNase-free glass- or plasticware, RNase inhibitors can be added to the sample. The RNA should be handled at the indicated temperatures, as higher temperatures or prolonged incubation can increase degradation. Degradation during the polyA selection will result in loss of transcript 5’ sequences, because selection is performed using the 3¢ polyA tail.

Confirm that the DSN enzyme is active within a week prior to use, per the manufacturer's instructions. Also, see Table 3.2 for other critical parameters and troubleshooting tips.

Table 3.2 Troubleshooting for Normalized RNA-seq Library Construction

Problem	**Possible cause**	**Solution**
RNA reads are not uniform	RNA degradation (the 5' of the transcript is more susceptible since the polyA selection is done from the 3' end)	Start with high-quality RNA; use RNA with RNA integrity number >8; reduce fragmentation time; begin with a single RNA sample and work quickly.
Excessive (>15) PCR cycles are needed for amplifying the library before the normalization step	Low starting RNA concentration	Test RNA concentration and quality at the beginning.
	Poor library generation, perhaps due to inactive enzymes/reagents	Start with a new batch of reagents
	Wrong primers	Make sure that the inner primers are being used for the first amplifications
A discrete band appears instead of a smear	Low complexity library or PCR cycles exceeded the linear phase	Check reactions; carefully assess the number of the real-time PCR cycles needed to amplify the library.
Coverage enrichment only affects moderately expressed genes	Low complexity library	Start with a greater amount of polyA RNA. Utilize more templates by setting up more amplification reactions.
		If starting with total RNA, polyA-select prior library construction.
Gene expression profiles do not appear normalized	Inactive DSN enzyme or another problem with the DSN reaction. In addition, the library may not have been sufficiently amplified.	Test DSN enzyme activity according to the manufacturer's instructions. Assess the effect of DSN by including a no-enzyme tube (and compare overall digestion in the sample with real-time PCR).
Excessive number of cycles are needed to amplify the library after it was normalized (but not before normalization); if sequenced, low complexity library	DSN may have degraded the library.	Make sure that the library is well denatured. Check the quality of the thermal cycler or increase the denaturing time. During normalization, do not allow the tubes to be at room temperature beyond the short time needed to centrifuge the contents. Always work on ice when setting up follow-up amplifications after normalization, as the DSN enzyme may still be active. Also, make sure the buffer

Table 3.2 (continued)

		concentrations used during normalization are correct.
Unexpected read sequences (e.g., from a different organism or unexpected genomic regions)	Contamination	Make sure the library is not contaminated with another library. Keep a separate area with equipment and reagents for work during the pre-amplification phase of the library. Use a negative control during the real-time PCR amplification and run on a gel to check if a curve is visible.

Anticipated Results

Amplification cycles are expected to be in the range of 8 to 12 cycles and 12 to 15 cycles for the final amplification. When amplified within the exponential phase, by using real-time PCR, these products are expected to appear as a smear rather than as discrete bands, as in Figures 3.2B-D. The library smears resulting from the amplifications using the inner primers are expected to be smaller than the gel-extracted library size, since they only amplify the inner part of the universal adapter sequence. Following amplification with the long Illumina final primers, the library size is expected to increase (Fig. 3.2).

In the normalized libraries, highly expressed RNA transcripts are expected to be decreased in proportion in the library by about 10-fold compared to their original proportion (the highest expressing gene is reduced 50× in proportion). Lower-expressed RNA transcripts are expected to be enriched ~10-fold. Additionally, some previously undetectable RNA transcripts and some pre-mRNAs are expected to be present at nonnegligible levels. Untranscribed regions of the genome should not be enriched, but may contain rare interspersed reads not exceeding 1 to 2 reads depth.

Time Considerations

The RNA-seq library preparation takes ~4 days and normalization and final amplification takes ~3 days.

Acknowledgments

This work was funded by LeDucq Fondation, NIH U54 Syscode PL1 HL092552 and NHLBI 1U01HL098166 grants awarded to Jonathan and Christine Seidman.

Literature Cited

- *Cloonan, N., Forrest, A.R., Kolle, G., Gardiner, B.B., Faulkner, G.J., Brown, M.K., Taylor, D.F., Steptoe, A.L., Wani, S., Bethel, G., Robertson, A.J., Perkins, A.C., Bruce, S.J., Lee, C.C., Ranade, S.S., Peckham, H.E., Manning, J.M., McKernan, K.J., and Grimmond, S.M. 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat. Methods 5:613-619.*

- *Kim, J.B., Porreca, G.J., Song, L., Greenway, S.C., Gorham, J.M., Church, G.M., Seidman, C.E., and Seidman, J.G. 2007. Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. Science 316:1481-1484.*

- *Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5:621-628.*

- *Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320:1344-1349.*

- *Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., Lightfoot, S., Menzel, W., Granzow, M., and Ragg, T. 2006. The RIN: An RNA integrity number for assigning integrity values to RNA measurements. BMC Mol. Biol. 7:3.*

- *Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., Schmidt, D., O'Keeffe, S., Haas, S., Vingron, M., Lehrach, H., and Yaspo, M.L. 2008. A global view of* *gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321:956-960.*

- *Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., Lao, K., and Surani, M.A. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods 6:377-382.*

- *Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28:511-515.*

- *Wang, E.T. 2008. Alternative isoform regulation in human tissue transcriptomes. Nature 456:470-476.*

- *Zhulidov, P.A., Bogdanova, E.A., Shcheglov, A.S., Vagner, L.L., Khaspekov, G.L., Kozhemyako, V.B., Matz, M.V., Meleshkevitch, E., Moroz, L.L., Lukyanov, S.A., and Shagin, D.A. 2004. Simple cDNA normalization using kamchatka crab duplex-specific nuclease. Nucleic Acids Res. 32:e37.*

Author contributions:

DCC designed and implemented the project

JMG and DSH provided important protocol optimizations

JGS conceived and directed the project

CHAPTER 4

Chapter 4. Quantification of microRNA Expression with Next-Generation Sequencing:

This manuscript has been accepted for publication in Current Protocols of Molecular Biology. Reproduced with permission of Current Protocols and John Wiley & Sons, Inc.

AUTHORS

Seda Eminaga[1*], Danos C. Christodoulou[1*], Francois Vigneault[1,2,3*] , George M. Church[1,2], Jonathan G. Seidman[1]

AFFILIATIONS

1 Department of Genetics, Harvard Medical School, Boston, MA 02115, USA

2 Wyss Institute for Biologically Inspired Engineering, Boston, MA 02115, USA

3 Ragon Institute of MGH, MIT and Harvard, Massachusetts, MA 02129, USA

* These authors contributed equally to this work.

ABSTRACT

Rapid advancement of next generation sequencing technologies has made it possible to study expression profiles of microRNAs (miRNAs) comprehensively and efficiently. We have previously shown that multiplexing miRNA libraries by barcoding can significantly reduce sequencing cost per sample without compromising library quality [Alon et al. 2011, Vigneault et al 2012]. In this unit, we provide a step-by-step protocol to isolate miRNAs and construct multiplexed miRNA libraries. We also describe a custom computational pipeline designed to analyze the multiplexed miRNA library sequencing reads generated by Illumina-based technology.

INTRODUCTION

miRNAs are short, 17-25 nucleotide non-coding RNAs that have emerged as critical regulators of gene expression in various physiological and pathophysiological processes [Mendell and Olson, 2012] due to their ability to influence the stability/translation of a large number of RNAs. Our understanding of roles of miRNAs and their targets has been aided by miRNA expression profiling studies. While high-throughput platforms including multiplex PCR and microarrays have proven to be valuable tools, next generation sequencing technology has quickly emerged as the preferred platform for studying miRNA expression. One of the advantages of next-generation sequencing is the ability to pool and sequence multiple samples in one lane of a sequencer, significantly lowering costs, without compromising the ability to construct comprehensive expression profiles for every assessed sample.

A typical miRNA library construction protocol involves 3’ and 5’ adapter ligation to miRNAs, followed by reverse transcription to generate cDNA libraries which are then amplified by PCR and purified by gel extraction prior to sequencing [Motameny et al. 2010]. Multiplexing can be achieved by addition of barcodes (unique nucleotide tag sequences incorporated into the adapter or PCR primer) during library construction. However, barcodes on the adapter confer bias during ligation and thus must be avoided, whereas barcodes/indexes introduced during the amplification step is a safe alternative (Alon et al., 2011).

One of the challenges of next generation sequencing is to sift through large data sets of millions of short reads generated by sequencers requiring computational analysis. Reads from miRNA sequencing are first processed to remove the 3’ adapter sequences and then matched to the reference sequence for identification of mature and/or novel miRNAs. Identifying miRNAs with differential expression among samples by quantifying the number of reads per miRNA helps infer relevant biological processes.

The purpose of this unit is to provide a step-by-step protocol of total RNA extraction to generation of miRNA expression profiles in tissues of interest. First, we describe a protocol to isolate miRNAs and construct multiplexed miRNA libraries for Illumina sequencing (Alon et al 2011, Vigneault et al, 2012: CPHG Unit 11.12.1-10). Next, we describe a custom bioinformatics pipeline designed to construct comprehensive miRNA expression profiles and assess differential expression between samples. Any measured changes can be quantified with precision as each miRNA is expected to be sampled a large number of times during this process. [Although this pipeline is tested for Illumina reads, it can be adapted for any other platform.]

BASIC PROTOCOL 1:

ISOLATION OF TOTAL RNA CONTAINING miRNAs

Total RNA containing miRNAs can be isolated from tissues either by Trizol/Phenol/Chloroform extraction or with commercially available kits. One important determinant when choosing a commercially available RNA isolation kit is whether they retain small RNAs (<200 nucleotides). We have successfully used Ambion's miRVana miRNA isolation kit to isolate good quality total RNA, specifically from mouse heart tissue. Before starting isolation, make sure to clean all equipment and bench-top with RNase Zap and frequently change gloves to prevent any RNase contamination.

Materials

RNase Zap (Ambion, cat. no AM9780)

RNAlater (Qiagen, cat. no. 76106)

RNase-free 1.5 ml microcentrifuge tubes

RNase-free tips

miRVana miRNA isolation kit (Ambion, cat. no. AM1560)

ACS grade 100% Ethanol

Microcentrifuge

Bioanalyzer 2100 (Agilent)

1. Isolate total RNA according to manufacturer’s protocol (miRVana isolation kit or Trizol).

We recommend isolating RNA from fresh tissue or tissue stored in RNAlater. For good quality RNA, the time between tissue dissection and RNA isolation should be minimized. We prefer homogenizing with steel beads using TissueLyser (Qiagen) as it allows processing of several samples simultaneously, and prevents cross-contamination between samples. Alternatively, if TissueLyser is not available, we prefer homogenizing using rotor/stator type homogenizer; however, equipment should be thoroughly cleaned between processing of different samples to prevent contamination. If the tissue will not be immediately used for RNA isolation, it should be snap-frozen in liquid nitrogen and stored at -80 C. When isolating RNA from frozen tissue, tissue should first be grinded to powder using prechilled mortar/pestle with liquid nitrogen to prevent thawing of tissue before lysis. Basic Protocol 2 for miRNA library construction works well on total RNA, but if enrichment is desired, an aliquot of total RNA should be saved for quality control and miRVana kit can be used to enrich for miRNAs. Although we have not directly compared expression profiles prepared from miRNA-enriched vs. total RNA, there may be some differences and we recommend that when different samples are compared, they should be prepared with the same method to prevent any bias.

2. Confirm quality of RNA.

We strongly recommend that RNA quality is confirmed by Bioanalyzer (Agilent) and we have observed that RNA with RNA Integrity Number (RIN) >8 provides high quality libraries. Alternatively, RNA quality may be checked by running an agarose gel to visualize 28S and 18S RNA. A 2:1 ratio of 28S to 18S indicates good quality RNA.

3. Store RNA in small aliquots at -80 C.

Repeated freeze-thaw of RNA should be avoided to minimize RNA degradation, as degraded fragments can adversely affect multiple steps and possibly introduce biases.

BASIC PROTOCOL 2:

MULTIPLEX microRNA LIBRARY CONSTRUCTION FOR ILLUMINA SEQUENCING

[Basic Protocol 2 has previously been published and incorporated from Vigneault et al in Current Protocols in Human Genetics, CPHG Unit 11.12.1-10]

This procedure describes a method for constructing multiplexed miRNA libraries. A miRNA library is made (Figure 4.1) from each RNA sample by 3’ adapter ligation, 5’ RT primer annealing, 5’ adapter ligation, reverse transcription, and PCR amplification. Although the forward PCR primer is the same, a different reverse PCR primer with a unique barcode is used for each RNA sample. The different libraries can then be pooled into a single sequencing reaction at the end of the library construction. The following instructions are for the preparation of one sample, so users must scale-up according to the specific number of samples they are preparing. All incubations are conducted in a thermal cycler.

Materials

RNase Zap (Ambion, AM9780)

Nuclease-free water (Ambion, AM9937)

Starting RNA

10x T4 RNA Ligase 2tr Buffer (Enzymatics, L607)

3' rApp-adapter (see Table 4.1)

100% DMSO (Sigma, D9170)

RNase Inhibitor (Enzymatics, Y924L)

RT primer (see Table 4.1)

5' RNA adapter (see Table 4.1)

ATP (Enzymatics, N207-10-L)

T4 RNA Ligase 1 (Enzymatics, L605L)

dNTPs (Enzymatics, N205L)

Superscript III First-Strand Synthesis System (Invitrogen, 180080-051)

Phusion High-Fidelity DNA Polymerase (NEB, M0530S)

BCmiRNA_PCR1 (see Table 4.1)

BCmiRNA_PCR2_BC (see Table 4.1)

AgencourtAMPure XP 5 mL Kit (Beckman Coulter Genomics, A63880)

E-Gel EX Gel, 2% (Invitrogen, G4020-02)

70% (v/v) ethanol

25 bp Ladder (Invitrogen, 10597-011)

100 bp Ladder (Invitrogen, 15628-019)

MinElute Reaction Cleanup Kit (Qiagen, 28204)

Agilent High Sensitivity DNA Kit (Agilent, 5067-4626)

Thermal Cycler (for all incubations)

E-Gel I-Base Power System (Invitrogen, G6400)

E-Gel Safe Imager Real-Time Transilluminator (Invitrogen, G6500)

Dynamag-2 Magnet (Invitrogen, 123-21D)

Recommended: Iceless Cold Pack (Eppendorf 022510509)

Recommended: Agilent 2100 Bioanalyzer

Optional: Nanodrop Spectrophotometer 2000

PROCEDURE

Ligation of 3' adenylated adapter

Make sure to clean surfaces and instruments with RNase Zap and maintain RNase-free conditions throughout the protocol. While as few as 100 ng of total RNA is sufficient, we recommend starting with at least 1 μg of total RNA (one can also use the equivalent fraction of enriched for small RNAs if desired). We recommend verifying RNA quality using the Agilent Bioanalyzer RNA nano or pico chip and using samples of RIN value of 7 or above.

1| Dilute the starting RNA to 200 ng/μL in nuclease-free dH_2O, if possible.

2| Set up a ligation reaction in a 200 ul PCR tube.

Component	**Volume (μl)**	**Final Concentration**
200 ng/μlRNA in dH_2O	5	1 μg total
10x T4 RNA Ligase 2tr Buffer	1	1x
10 μM 3'rApp-adapter	1	10 pmoles total
100% DMSO	1	10%

3| Denature for 30 sec at 90°C, then for at least 30 sec at 4°C.

4| Add the following directly to the ligation reactions on ice:

Component	**Volume (μl)**	**Final Concentration**
RNase Inhibitor (40 U/μl)	0.5	2 U/μl

T4 RNA Ligase 2tr (200 U/ μl)	1.5	30 U/μl

5| Incubate for 1 h at 22°C.

We recommend using Enzymatics buffer, as its composition gave us significantly higher yield than other commercially available T4 RNA ligase 2 truncated buffers.

Annealing of RT primer

6| Add the following directly to each reaction on ice:

Component	**Volume (μl)**	**Final Concentration**
10μM RT Primer	1	10 pmoles total

The final amount of RT primer must be at equimolar ratio (10 pmoles) with the starting amount of 3’rApp-adapter (10 pmoles) for each sample.

7| Incubate for 30 sec at 90°C, then for 5 min at 65°C, then for at least 30 sec at 4°C.

Ligation of 5’ RNA adapter

8| Prepare the 5’ RNA adapter by incubating ~5μl in a 200 ul PCR tube at 70°C for 2 min, then 4°C for at least 30 sec.

An excess of volume is prepared to account for evaporation and facilitate the pipetting of the proper volume at the next step.

9| Spin down the ligation mixture by centrifuging 10 sec at ~2000 x g, room temperature, using a microcentrifuge and add the following reagents directly to it:

Component	**Volume (µl)**	**Final Concentration**
10 mM ATP	1.5	1µM
10 µM 5' RNA Adapter	1	10 pmoles total
T4 RNA Ligase 1 (20 U/µl)	1.5	2 U/µl

10| Incubate for 1 h at 20°C.

Reverse transcription of captured MicroRNAs

The previous steps result in a reaction volume of 15 µl. Only 5µl is used in the subsequent RT-PCR step, and so the remaining can be stored (-80°C) as a backup (highly recommended) for two more runs. However, the rest of the protocol below can be scaled up 3 times and the full 15µl may be processed at once if you need to achieve higher yield (for example, when starting with lower amounts of RNA).

11| Prepare the following reactionin a 200 ul PCR tube:

Component	**Volume (µl)**	**Final Concentration**
Ligated miRNAs	5	-
5x First strand buffer	2	1x
12.5mM dNTP mix	0.5	625 µM

100mM DTT	1	10 mM
RNase Inhibitor (40 U/µl)	0.5	2 U/µl
Superscript III (200 U/µl)	1	20 U/µl

12| Incubate for 30 min at 48°C.

As the RT primer was annealed earlier, do not denature or conduct an annealing cycle at this stage but go directly to the reverse transcriptase incubation temperature (48°C, as shown above).

Limited PCR Amplification

13| Prepare the PCR reaction in a 200 ul PCR tube:

Component	**Volume (µl)**	**Final concentration**
dH2O	27	To 50 ul total
Reverse Transcribed-miRNAs	10	-
5x HF buffer	10	1x
25mM dNTPs	0.5	0.5 mM
25 μM BCmiRNA_PCR1	1	0.5 μM
25 μM BCmiRNA_PCR2_BC*	1	0.5 μM
Phusion DNA pol. (2 U/µl)	0.5	1 U

BCmiRNA_PCR2_BC refers to the bar-coded primer,where for each unique starting RNA sample a unique bar-code primer needs to be used (see Table 4.1). To limit* *bar-code / samples aerosol contamination, it is recommended to only open and close one tube of primer at a time.*

14| Cycle the PCR reaction as follows in a thermal cycler:

1- 98°C for 30 sec

2- 98°C for 10 seconds

3- 60°C for 20 seconds

4- 72°C for 20 seconds go to step 2, 11 more times

5- 72°C for 5 min

6- 4°C pause

The number of cycles can be varied according to the amount of microRNA present in the starting sample. In our hands, a total of 12 cycles generally results in the best yield while limiting unnecessary cycling. We recommend not exceeding 15 cycles as this will increase non-specific background amplification and reduce optimal yield of the desired products. Instead, additional starting RNA should be prepared in parallel and combined at the final stage to increase yield.

PCR Clean-Up with AMPure XP beads

15| Transfer PCR reactions to a new 1.5 mL tube.

16| Vigorously mix the AMPure XP beads and then add 90 µl of beads to each 50 µl PCR reaction. Pipet the beads slowly.

17| Vortex for 30 seconds, and then incubate on bench for 5 min.

18| Quick spin, and place on the magnetic rack for 5 min.

19| With the tubes still on the magnet, aspirate and discard the liquid from the reaction.

20| With the tubes still on the magnet, add 400 µL 70% EtOH to the beads and leave for 30 sec. Then discard the 70% EtOH.

21| Repeat the previous step for a second wash.

22| Quick spin on a microfuge to collect last traces of EtOh. Put tubes back on magnet and remove any last drops of EtOH at the bottom of the tube.

23| Leave the tube open to air dry for 2 min.

24| Remove the tube from the magnet and add 45 µL of nuclease-free water.

25| Vortex for 30 sec.

26| Place the tubes on the magnet and leave for 1 min.

27| With the tubes on the magnet, transfer 42 µL to new 1.5 mL tubes.

Gel extraction of microRNA library

28| Prepare a 2% Agarose Gel EX following the manufacturer's protocol.

29| Dilute the 25 bp and 100 bp ladders 1:20 in water and load 20 µL of each.

30| Split each microRNA library prepared above across 2 lanes by loading 20 ul per well.

31| Run the gel for 14 minutes on the Invitrogen I-Base using the 2% E-Gel settings.

32| When the run is complete, take a picture of the gel.

The migration patterns of DNA on E-gels are affected by the total amount and salts present in the loaded sample, and sometime one may observe a shift in migration of the expected product in relation to the ladder.

33| Pry open the E-Gel by cracking open each side.

34| Using a clean razor blade for each sample, cut between 125 and 175 bp to capture the two dominant miRNA bands.

35| Gel extract using the Mini Elute Qiagen Gel-Extraction Kit following the manufacturer's protocol, conducting the final elution in 15 ul of dH2O

Melt the gel bands at 37°C instead of the recommended 55 °C. The MinElute columns have a tendency to trap residual EtOH from the wash steps. To avoid this issue, dry spin the column for 1 minute at maximum speed and then turn the column 180 degrees and repeat the spin for another 1 minute. Then transfer the column to a recovery tube and leave the column open for 3 min to air dry prior to adding the elution buffer.

Library QC and Mixing

36| The library can now be mixed at equimolar concentration, prior to submission for sequencing. We strongly recommend analyzing the quality and concentration of each final library using the Agilent Bioanalyzer DNA high sensitivity chip in order to combine the different libraries at equimolar ratios into a single multiplexed library. Although less accurate, a Nanodrop spectrophotometer would also work to a decent degree for this step if an Agilent Bioanalyzer is inaccessible. For high throughput project with high amount of samples, the Bioanalyzer can be used to combine the libraries prior to gel extraction.

37| Submit library for sequencing using standard Illumina genomic primer or Truseq primer with 75 bp single-read. Alternatively, a custom indexing primer can also be used if desired (Table 4.1).

BASIC PROTOCOL 3:

BIOINFORMATICS ANALYSIS OF MULTIPLEXED miRNA LIBRARY SEQUENCING DATA

Analysis pipeline presented here aims at measuring miRNA expression and assessing differential expression between samples. It also provides an efficient way to construct comprehensive miRNA expression profiles. Here, we use annotated mature miRNAs deposited regularly in miRBase as a reference. (http://www.mirbase.org). In our analysis pipeline, first, the 3' adapter sequences are removed and then the reads are assigned to separate output files based on different barcodes used. Next, miRNAs are identified by aligning the reads to miRBase (version18) (Kozomara A and Griffiths-Jones S 2011) and the reads are tallied to generate total counts for each miRNA. Finally, statistical significance (p-value) between 2 or more samples is calculated to generate differential expression profiles. The analysis programs are written in Perl and the steps described can be performed through the command line. The workflow of analysis pipeline is shown in Figure 4.2.

Materials

Hardware pre-requisites:

Linux, Unix or Mac OS X installed computer or Cygwin with Windows.

A computer cluster may not be necessary although recommended.

No significant memory requirements.

Software pre-requisites:

Perl v5.10.0

BioPerl module for SAGE comparison

Text-editor program (e.g. TextWrangler)

Analysis package:

Analysis programs are all written in Perl and can be found at http://seidman.med.harvard.edu/fgs/software/mirna_soft/mirna_soft.tar.gz

Files needed:

Sequence file in FASTQ format (e.g. reads.fastq)

A FASTQ file from a sequencing run includes 4 lines per read. The first line starts with "@" followed by a unique identifier, second line includes the sequencing read, third line may contain additional sequencing run information and fourth line includes the quality scores for each nucleotide in the sequencing read.

mature.fa.gz (http://www.mirbase.org/ftp.shtml) reads.fastq is provided in the package for a test run.

1. Download and install the following programs:
   a. Perl: http://www.perl.org
   b. BioPerl module for SAGE comparison by algorithm described by Audic and Claverie (1997) http://search.cpan.org/~scottzed/Bio-SAGE-Comparison-1.00/ Once downloaded, unpack the module. Type:

      ```
      tar xvfz Bio-SAGE-Comparison-1.00.tar.gz
      ```

      Then, add the module's path to the PERL library environment variable. Use PERL5LIB for Perl 5. Type:

      ```
      export PERL5LIB=/location_to_downloaded_module/Bio-SAGE-Comparison-1.00/lib
      ```

   c. The analysis package: http://seidman.med.harvard.edu/fgs/software/mirna_soft/mirna_soft.tar.gz which includes the following scripts: separatebarcodes.pl, countreads_mirna.pl and cmp_mirna.pl . To unpack the modules, type:

      ```
      tar xvfz mirna_soft.tar.gz
      ```

2. Download the latest mature.fa.gz from http://www.mirbase.org/ftp.shtml containing all the mature miRNA sequences.

3. Remove the 3' adapter sequence and separate reads into individual files by barcode. Type:

```
perl separatebarcodes.pl  reads.fastq
```

A section of separatebarcodes.pl *is shown below and MUST be modified according to the samples/barcodes/adapter used by using a text editor program such as TextWrangler (Mac) or equivalent.*

```
my $adaptor="ACGGGCTAATATTTATCGGTGGAGC"; ## specific adapter sequence
my %barcodes = (
        CGTGAT => "Sample1", #BC1 ## these are associating each barcode with
        ACATCG => "Sample2", #BC2 ## corresponding sample
        GCCTAA => "Sample3", #BC3
        TGGTCA => "Sample4", #BC4
        CACTGT => "Sample5", #BC5
        ATTGGC => "Sample6", #BC6
        GATCTG => "Sample7", #BC7
        TCAAGT => "Sample8", #BC8
);
```

*This perl script is designed to analyze the sequencing reads obtained from multiplexed miRNA libraries constructed using the adapters and PCR amplification primers shown in Table 4.1. Usually, the first part in the sequence corresponds to the miRNA sequence. The miRNA sequence is then followed by the adapter sequence “*ACGGGCTAATATTTATCGGTGGAGC*”, which is followed by a 6-nucleotide barcode (Table 4.1*, **bold-underlined***). First, the script opens the sequencing file* reads.fastq*, removes the 3' adapter, identifies each sample by barcode and creates one output file per sample based on the barcodes. The user is advised to confirm the output file by '*less*' or '*more*' command or by opening with any text-editor program. The output file should look like* Samplename.fq. *If Illumina TruSeq small RNA adapters/primers are used, the user can easily modify the script to reflect the appropriate adapter and barcode sequences.*

4. Uncompress mature.fa.gz downloaded from miRBase. Type:

```
gunzip mature.fa.gz
```

Open the file mature.fa and select the entries corresponding to the appropriate species (e.g. *mus musculus* for mouse). Save as species_mature.fa in a separate file to use in the next step. Alternatively, from the command line type (for mouse):

```
grep –A1 musculus  mature.fa > mouse_mature.fa
```

Open and verify final file.

5. Match the reads to reference miRBase and tally the counts of reads per microRNA to generate an expression profile per sample. Here, output files generated in step 3 are used as input. Type:

```
perl countreads_mirna.pl species_mature.fa OUTPUT_STEP3
```

*This perl script is applied to each output file from step 3 (*OUTPUT_STEP3, e.g. Samplename.fq) *after processing with the first script. First, the sequences shorter than 17 nucleotides are eliminated. Then, the script aligns the sequence to reference mature miRNA sequences, downloaded from miRBase. There are 2 output files created per sample: The first file* Samplename.m *contains information for aligned miRNAs: "miRNA name" "Sequence of the miRNA", "Unique matches", "Multi-matches". "Unique matches" are reads matching specifically to that miRNA out of the total list. "Multi-matches" includes reads with multiple reference miRNA sequences and are not subsequently used. To maximize read counts and incorporate reads from putative isomiRs, an unmatched full length read is trimmed at the ends before another matching attempt. The second file* Samplename.unm *contains the reads that did not match to any miRNA in the miRBase. These unmatched reads often contain small inserts that may not qualify as miRNAs such as fragmented RNAs or bad quality reads, and any unannotated novel miRNAs. To identify novel miRNAs,* Samplename.unm *can be matched to other species' miRNA sequences using the method described here, and they can also be* *aligned to the reference genome depending on the uniqueness and length of the sequence.*

The output file can be confirmed by using ‘less’ *or* ‘more’ *commands or can be opened by any text-editor program.*

6. Normalize and compare expression profiles of miRNAs. Type:

```
perl cmp_mirna.pl Samplename_1.m Samplename_2.m …Samplename_8.m
```

This script can be applied to two or more output files from step 5 to make a comparison. The read counts are normalized to counts per million by dividing the total read counts of a miRNA by the total read counts of the sample and multiplying this number by 10^6*. The script uses a Bayesian comparison based on Audic and Claverie (1997) to calculate the p-value. The output file for two samples will include “miRNA ID”, “Tag Sequence“ “Normalized Counts for Sample1”, “Normalized Counts for Sample2”, ”Raw reads for Sample 1”, “Raw reads for Sample2”, and “p-value” , while for more samples, it will include corresponding additional fields.*

7. Analyze the final output file.

We generally consider p-value ≤ 0.01 as an appropriate cut-off given the list of miRNAs. However, more stringent cut-off may be used.

COMMENTARY

Background Information

Owing to the development of high-throughput methods to study miRNA expression profiles, there has been an exponential increase in the amount of data generated in the miRNA field over the last decade. Compared to microarray or qPCR-based miRNA exression profiling techniques, next generation sequencing technology offers several advantages, including high sensitivity to measure miRNA levels over a wide dynamic range, ability to identify novel miRNAs and to detect miRNA expression levels in species for which complete genomes are not yet available. In addition, next-generation sequencing is able to detect miRNAs that differ by just one nucleotide [Pritchard et al. 2012]. Last but not the least, next-generation sequencing allows multiplexing of samples by tagging libraries with barcodes during library preparation.

Even though different sequencing platforms require different protocols and adapters/primers, they typically follow similar steps. First step of miRNA library construction is to capture miRNAs using specific adapters. miRNAs have a 5' phosphate and a 3' hydroxyl group as a result of RNAse III activity of Drosha and Dicer. To prevent self ligation and circularization of miRNAs during adapter ligation, pre-adenylated 3' adapters are used with truncated T4 RNA ligase 2 (which does not require ATP). Next, miRNAs are ligated to a 5' adapter, which contains the binding site for sequencing primer. The resulting adapter-captured miRNAs are reverse transcribed with a primer complementary to the 3' adapter to generate cDNA library. To provide enough yield for sequencing, cDNA library then needs to be PCR-amplified. Finally, the amplified library is gel extracted and quality-checked prior to sequencing. To sequence multiple libraries in one lane of a flow cell, each miRNA library can be barcoded by including a unique tag sequence as part of the adapter or PCR amplification primer during construction. However, we and others have reported previously that including barcodes in the adapter sequence creates significant ligation bias [Alon et al. 2011, Hafner et al. 2011], e.g. same biological sample tagged with 2 different barcodes show significant differences in miRNA expression levels, presumably as a result of bias by T4 RNA ligase-mediated ligation [Zhuang et al. 2012, Jayaprakash et al. 2011]. In the current protocol, we add barcode to each library during PCR amplification step, which helps avoid any ligation bias and therefore, provides a significant advantage (Alon et al 2011, Vigneault et al. 2012 CPHG Unit 11.12.1-10). With this method, up to 12 libraries can be prepared in parallel and sequenced in one lane of Illumina Genome Analyzer /HiSeq.

Several bioinformatics tools have been developed to analyze miRNA sequencing reads (Li et al. 2012) and they generally follow similar steps with some variations. Typically, after the first step of 3' adapter removal, the sequences are aligned against a reference sequence to identify annotated miRNAs as well as novel ones. One of the important parameters for bioinformatics analysis of miRNA sequencing reads is the alignment criteria. Several groups have identified isomiRs (variants from the reference) for a given mature miRNA and reported that the most abundant isomiR may not always be same as the mature miRNA [Morin et al. 2008, Wyman et al. 2011, Lee et al. 2010]. IsomiRs, which differ mainly at the 3' end, and to a lesser extent at the 5' end, may also include nucleotide substitutions or 3' non-template addition of nucleotides. While some studies have quantified the most abundant isomiR as the most representative of the miRNA level [Morin et al. 2008], others have quantified miRNAs with 100% length and sequence match with mature miRNA [Wang et al. 2009]. In a recent study, expression levels of isomiRs were reported to highly correlate with that of annotated mature miRNAs [Cloonan et al. 2011]. While isomiRs may well be biologically relevant, abundance of some isomiRs above that of mature miRNA may also result from T4 RNA ligase bias in capturing small RNAs [Hafner et al. 2011, Jayaprakash et al. 2011, Zhuang et al. 2012]. Since the biological significance of isomiRs is largely unknown and their biogenesis is not understood, currently, our analysis pipeline counts miRNAs with exact matches as well as those with some variation to the annotated mature miRNA in miRBase (version 18). Our analysis pipeline also keeps a separate file of reads not aligning to miRBase, and these reads can be mapped against the reference genome to identify putative novel miRNAs.

Critical Parameters and Troubleshooting

One of the most critical parameters for a high quality library is the quality of starting RNA. Therefore, it is important to maintain an RNAse-free workspace throughout RNA isolation (Basic Protocol 1) and library construction protocol (Basic Protocol 2: steps 1-12). Another critical parameter is to avoid contaminating pre-amplified library with amplified library, as even the slightest contamination will have detrimental effects for downstream analysis. Therefore, we recommend designating separate equipment, reagents and bench-space for all the steps prior to PCR-amplification of libraries (Basic

Protocol 2: Step 14) as “pre-PCR” and for all the steps after PCR-amplification as “post-PCR”. We strongly recommend including a “no-ligase” control reaction (where T4 RNA ligase is omitted) and processing in parallel to the experimental samples to assess for any possible contamination, which will be evident after PCR amplification step. If an amplified library is detected in the control, the prepared libraries should be discarded and new reagents should be used to start over. Therefore, to prevent wasting of reagents, we recommend that oligos, and all other library construction reagents are stored in small aliquots.

During library construction, 3’ and 5’ adapters are used in excess to ensure efficient capture of miRNAs and to prevent their self-circularization. As a result, the adapters can ligate to each other and the undesired adapter-adapter dimer band is often observed in the agarose gels during library preparations. In this protocol, annealing the RT primer to the 3’ adapter-ligated product, prior to ligation of 5’ adapter, can significantly reduce the adapter dimer formation and therefore, we do not expect to see the dimer band (114bp) on an agarose gel. However, if an adapter dimer band is observed, one or two rounds of denaturing PAGE extractions should be performed (as described in Alon et al. 2011) to remove the adapter dimer and to prevent unnecessary sequencing of this undesired ligation product. Denaturing gel extraction ensures that any dimer fraction that may have annealed to the full-length library fragments is removed.

Another critical parameter is to quantify the final multiplexed library yield (Basic Protocol 2: step 36) right before submission, as low yield library may fail to cluster. If low yield libraries are obtained, we recommend starting with either more RNA, or preparing multiple libraries from the same RNA in parallel. In addition, PCR amplification cycles can be increased (but not more than 15, as this may result in low complexity library).

The multiplexed miRNA libraries prepared according to our protocol can be sequenced on Illumina Genome Analyzer II or HiSeq using either the standard Illumina primer for genomic libraries or primer for Truseq. It is critical to perform a 75bp single-read sequencing to ensure that insert miRNA (<30bp), 3' adapter (25bp) and barcode (6bp) are all sequenced. Alternatively, the libraries can be sequenced using a custom indexing primer (Table 4.1), however, this may be more expensive, as the whole flow cell has to be cycled twice, which may not be required for other lanes in the flow cell.

For bioinformatics analysis of sequencing reads, we highly recommend that for the first-time analysis, the user runs each script successively to generate the final output file. This will ensure that each module is running successfully, and if there is a problem, it will be easier to identify. If an extremely high-throughput is desired, the steps can be automated for more efficient processing of the sequencing data.

Even though we expect strong correlation (r >0.99) in technical replicates, in order to minimize differences due to technical variation, we recommend that the libraries that are to be directly compared, be handled and prepared in parallel. It is also expected that the biological replicates show a strong correlation; however, if they do not, this can be due to several factors including poor RNA quality (we recommend RIN>8), low complexity library (we recommend increasing the library yield before the PCR amplification step by starting with more RNA or processing multiple libraries from the same RNA and keep PCR amplification to no more than 15 cycles) or contamination with another library. We have also observed that there is a strong correlation between libraries prepared from “pooled” RNA samples (pooling RNA from 3 biological replicates) and average of 3 libraries prepared separately from RNA of 3 biological replicates. We believe that “pooled” RNA approach can help reduce the biological variation, allow user to pool RNA when sources are limited and also reduce the cost of sequencing.

As a first follow-up experiment, we highly recommend that the sequencing data is confirmed by quantitative RT-PCR. There are several commercially available qPCR kits to quantify miRNAs, and in our hands, Applied Biosystems’ stem-loop qPCR primers (Taqman) work well.

Anticipated Results

Technical and/or biological replicates are expected to have a strong correlation (r >0.99) with sequence counts ranging from 10 to 100000< counts per million. We advise ignoring miRNAs with less than 10 counts per million in all the groups being compared, since they can be due to sequencing errors.

Time Considerations

Total RNA isolation can be completed in under 1 hour for few to several samples. The library construction protocol of up to 12 samples can be completed in one day. The bioinformatics analysis can be completed within one to several hours depending on the number of total sequences.

Table 4.1. List of all oligos used for multiplexed miRNA library preparation for Illumina sequencing. **Bold-underlined** bases represent 6-nucleotide barcodes. (Alon et al. 2011).

Name	Sequence (5’-3’)
BCPCR_3'rApp-adapter	/5rApp/ACGGGCTAATATTTATCGGTGG/3SpC3/
BCPCR_5'RNA-adapter	rUrCrCrCrUrArCrArCrGrArCrGrCrUrCrUrUrCrCrGrArUrCrUrC
BCPCR_RT primer	GCTCCACCGATAAATATTAGCCCGT
BCPCR_PCR1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
BCPCR_PCR2-BC1	CAAGCAGAAGACGGCATACGAGAT**CGTGAT**GCTCCACCGATAAATATTAGCCCGT
BCPCR_PCR2-BC2	CAAGCAGAAGACGGCATACGAGAT**ACATCG**GCTCCACCGATAAATATTAGCCCGT
BCPCR_PCR2-BC3	CAAGCAGAAGACGGCATACGAGAT**GCCTAA**GCTCCACCGATAAATATTAGCCCGT
BCPCR_PCR2-BC4	CAAGCAGAAGACGGCATACGAGAT**TGGTCA**GCTCCACCGATAAATATTAGCCCGT
BCPCR_PCR2-BC5	CAAGCAGAAGACGGCATACGAGAT**CACTGT**GCTCCACCGATAAATATTAGCCCGT
BCPCR_PCR2-BC6	CAAGCAGAAGACGGCATACGAGAT**ATTGGC**GCTCCACCGATAAATATTAGCCCGT

Table 4.1 (continued)

BCPCR_PCR2-BC7	CAAGCAGAAGACGGCATACGAGAT**GATCTG**GCTCCACCGATAAATATTAGCCCGT
BCPCR_PCR2-BC8	CAAGCAGAAGACGGCATACGAGAT**TCAAGT**GCTCCACCGATAAATATTAGCCCGT
BCPCR_PCR2-BC9	CAAGCAGAAGACGGCATACGAGAT**CTGATC**GCTCCACCGATAAATATTAGCCCGT
BCPCR_PCR2-BC10	CAAGCAGAAGACGGCATACGAGAT**AAGCTA**GCTCCACCGATAAATATTAGCCCGT
BCPCR_PCR2-BC11	CAAGCAGAAGACGGCATACGAGAT**GTAGCC**GCTCCACCGATAAATATTAGCCCGT
BCPCR_PCR2-BC12	CAAGCAGAAGACGGCATACGAGAT**TACAAG**GCTCCACCGATAAATATTAGCCCGT
BC_Custom_Indexing	ACGGGCTAATATTTATCGGTGGAGC (optional)

Notes: All oligonucleotides can be ordered through Integrated DNA Technologies (IDT; http://www.idtdna.com). The oligos should be ordered with HPLC purification. The adenylated adapter can be ordered from IDT, or if on a budget, made as previously described (Vigneault, Sismour, & Church, 2008).

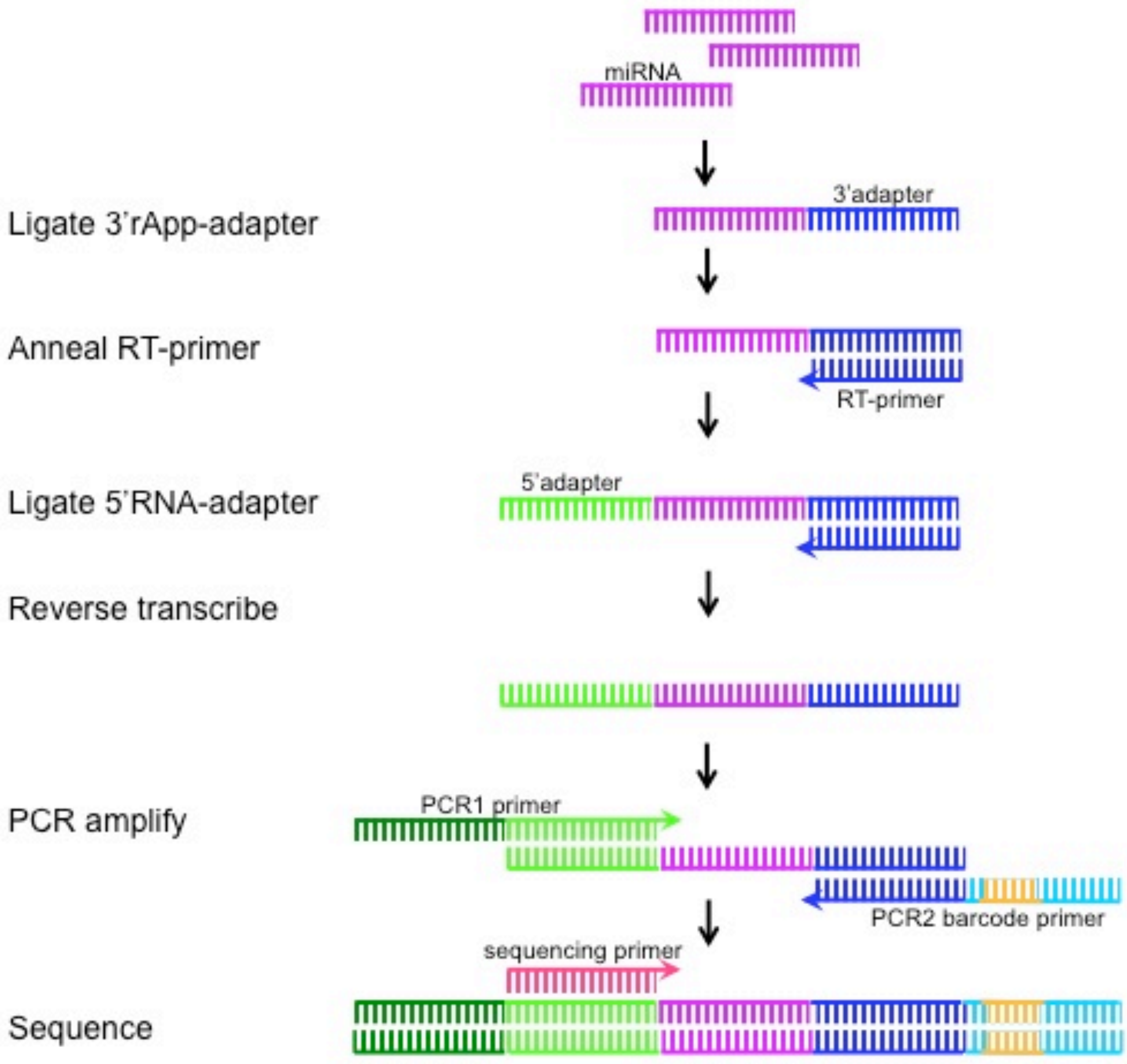


Figure 4.1. Schematic for miRNA library construction for next-generation sequencing

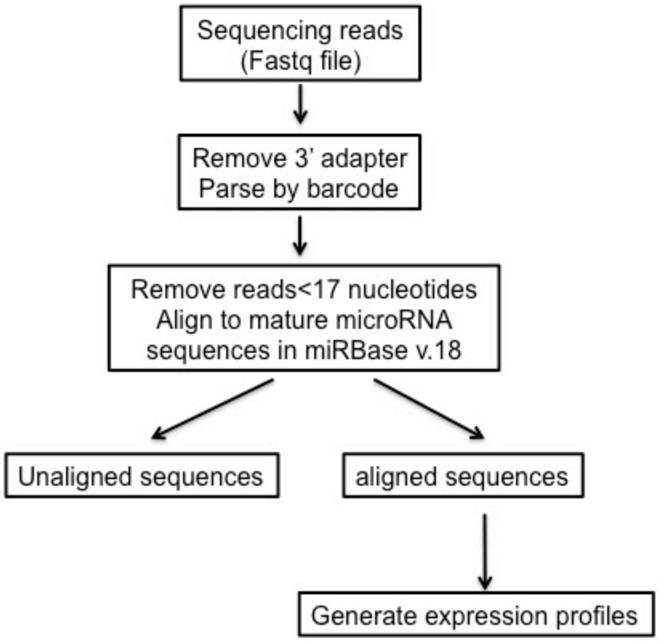


Figure 4.2. Workflow for bioinformatics analysis of multiplex mRNA library sequencing data

FUNDING

This work is supported by the Fondation LeDucq, NIH U54 Syscode PL1 HL092552 and NHLBI 1U01HL098166 (JG Seidman) and Canadian Institutes of Health Research and Ragon Institute Fellowship (F Vigneault).

REFERENCES

Audic S and Claverie JM. The significance of digital gene expression profiles. *Genome Res*. 1997 Oct;7(10):986-95.

Alon S, Vigneault F, Eminaga S, Christodoulou DC, Seidman JG, Church GM, Eisenberg E. Barcoding bias in high-throughput multiplex sequencing of miRNA. *Genome Res.* 2011 Sep;21(9):1506-11.

Cloonan N, Wani S, Xu Q, Gu J, Lea K, Heater S, Barbacioru C, Steptoe AL, Martin HC, Nourbakhsh E, Krishnan K, Gardiner B, Wang X, Nones K, Steen JA, Matigian NA, Wood DL, Kassahn KS, Waddell N, Shepherd J, Lee C, Ichikawa J, McKernan K, Bramlett K, Kuersten S, Grimmond SM. MicroRNAs and their isomiRs function cooperatively to target common biological pathways. *Genome Biol*. 2011 Dec 30;12(12):R126.

Hafner M, Renwick N, Brown M, Mihailović A, Holoch D, Lin C, Pena JT, Nusbaum JD, Morozov P, Ludwig J, Ojo T, Luo S, Schroth G, Tuschl T. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA*. 2011 Sep;17(9):1697-712.

Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R. Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic Acids Res*. 2011 Nov;39(21):e141. Epub 2011 Sep 2.

Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res*. 2011 Jan;39(Database issue):D152-7.

Lee LW, Zhang S, Etheridge A, Ma L, Martin D, Galas D, Wang K. Complexity of the microRNA repertoire revealed by next-generation sequencing. *RNA*. 2010 Nov;16(11):2170-80.

Li Y, Zhang Z, Liu F, Vongsangnak W, Jing Q, Shen B. Performance comparison and evaluation of software tools for microRNA deep-sequencing data analysis. *Nucleic Acids Res*. 2012 May 1;40(10):4298-305.

Mendell JT, Olson EN. MicroRNAs in stress signaling and human disease. *Cell.* 2012 Mar 16;148(6):1172-87. Review.

Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A, Prabhu AL, Zhao Y, McDonald H, Zeng T, Hirst M, Eaves CJ, Marra MA. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res.* 2008 Apr;18(4):610-21.

Motameny S, Wolters S, Nurnberg P, Schumacher B. Next Generation Sequencing of miRNAs – Strategies, Resources and Methods. *Genes*. 2010. 1(1), 70-84.

Pritchard CC, Cheng HH, Tewari M. MicroRNA profiling: approaches and considerations. *Nat Rev Genet.* 2012 Apr 18;13(5):358-69.

Vigneault F, Sismour AM, Church GM. Efficient micoRNA capture and bar-coding via enzymatic oligonucleotide adenylation. *Nat Methods*. 2008 Sep;5(9):777-9.

Vigneault F, Ter-Ovanesyan D, Alon S, Eminaga S, C Christodoulou D, Seidman JG, Eisenberg E, M Church G. High-throughput multiplex sequencing of miRNA. *Curr Protoc Hum Genet.* 2012 Apr;Chapter 11:Unit 11.12.1-10.

Wang WC, Lin FM, Chang WC, Lin KY, Huang HD, Lin NS. miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression. *BMC Bioinformatics.* 2009 Oct 12;10:328.

Wyman SK, Knouf EC, Parkin RK, Fritz BR, Lin DW, Dennis LM, Krouse MA, Webster PJ, Tewari M. Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity. *Genome Res*. 2011 Sep;21(9):1450-61.

Zhuang F, Fuchs RT, Sun Z, Zheng Y, Robb GB. Structural bias in T4 RNA ligase-mediated 3'-adapter ligation. *Nucleic Acids Res*. 2012 Apr;40(7):

Author contributions

SE implemented RNA isolation steps and all steps discussed in this manuscript

DCC and JGS wrote computational programs

FV and GMC are responsible for the molecular biology steps and the initial conception of this protocol

CHAPTER 5

Chapter 5. Genome-wide transcriptional start-site analyses identifies Four-and-a-half LIM domains protein 1 as a modifier of hypertrophic cardiomyopathy:

Manuscript is in submission. Supplemental tables and supplemental image data can be provided upon request. Supplemental text and figures are in Appendix 1. The supplemental software package can be downloaded from the following location:

http://seidman.med.harvard.edu/fgs/software/RNAseq-5pr-package.tar.gz

Danos C. Christodoulou[1], Hiroko Wakimoto[1], Kenji Onoue[1], Seda Eminaga[1], Joshua M. Gorham[1], Steve R. DePalma[1], Daniel S. Herman[1], David A. Conner[1], David M. McKean[1], Anton Aboukhalil[2], Stephen Chang[1], Gyan Srivastava[3], Martha L. Bulyk[2], Jochen D. Muehlschlegel[4], Christine E. Seidman[1,5] & Jonathan G. Seidman[1]

1. Department of Genetics, Harvard Medical School, Boston, MA 02115

2. Division of Genetics, Brigham and Women's Hospital, Boston, MA 02115

3. Harvard Institutes of Medicine, Brigham and Women's Hospital, Boston, MA 02115

4. Department of Anesthesiology, Perioperative and Pain Medicine, Brigham and Women's Hospital, Boston, MA 02115

5. Howard Hughes Medical Institute, Division of Cardiovascular Medicine, Brigham and Women's Hospital, Boston, MA 02115

We report a transcriptome sequencing (RNA-seq) strategy that detects genome-wide changes in transcript levels and start-site usage. Using this methodology we interrogated the pathologic responses to sarcomere gene mutations in hypertrophic cardiomyopathy (HCM). We show that Four-and-a-half LIM domains protein 1 (*Fhl1*) undergoes robust transcriptional activation and altered start-site usage in HCM, resulting in a distinct protein isoform in myocytes. The *Fhl1* isoform switch is conserved in mice and humans with primary and secondary cardiomyopathies. By analyses of genetic ablation of *Fhl1* in mice with HCM, we demonstrate that *Fhl1* transcriptional activation and isoform switching has salutary effects on cardiomyopathy. These data implicate *Fhl1* as an adaptive modifier to cardiomyopathy stress. Encoded on chromosome X, stress-induced transcriptional changes in *Fhl1* may contribute to gender differences in the clinical severity of HCM.

Pathogenic mutations produce broad changes in cell biology, in part by impacting gene expression. Changes in gene expression occur by altering transcription levels and/or diversify the transcriptome through the alternative use of 5' start-sites, alternative exon splicing, and differential use of polyadenylation sites. Several methods for high throughput sequencing of cDNAs have been recently developed to interrogate the transcriptome (RNAseq)[1-4].

By modifying RNAseq protocols, we describe a strategy to define genome-wide assessment of RNA levels and structure, inclusive of quantitative assessments of 5' start-site usage. Using this methodology we studied the mouse cardiac transcriptome in a model of human hypertrophic cardiomyopathy (HCM) and identified transcriptional diversity in *Fhl1* that influences the clinical severity of this disorder.

We first optimized steps for constructing cDNA libraries for RNAseq to minimize RNA or cDNA fragmentation and we incorporated random hexamer priming to complete cDNA synthesis (**Supplemental Methods**). After adding adaptors, we used size selection and cDNA polarity to enrich the sequencing of small fragments that include transcripts with 5' ends (**Fig. 5.1a**, **Supplemental Fig. 5.1** and **Supplemental Fig. 5.2**). The gene expression profiles derived from this approach, termed as 5'RNA-seq, had high technical reproducibility (**Supplemental Fig. 5.3**). Additionally, there was a strong correlation (r=0.9) between the gene set with altered transcription (including low expression genes) identified by sequenced based transcriptional profiling and this RNAseq methodology (**Supplemental Fig. 5.4** and **Supplemental Table 5.1**).

To identify differences in start-site usage among isoforms, we developed computational approaches that detect shifts in read-depth distribution at the start-site regions of genes and to quantify the extent of change in start-site usage (**Fig. 5.1b** and **Supplemental Methods**). The sum of instances with detected shifts at 5' ends was converted into a RNAseq score for each gene. We provide this bioinformatic approach as open-source, free to use computational tools with configurable parameters to provide concurrent computation of gene expression profiles, quantification of 5' changes, and a tool for obtaining image data of gene profiles (**Supplemental software package**).

We used two strategies to validate this approach for defining alternative start-site usage. First we visually assessed RNAseq reads in 800 genes and confirmed that reads corresponded to alternative transcriptional start-sites annotated in UCSC Genome browser (**Supplemental Image Data**). The rank order of RNAseq scores significantly correlated with alternative start-site usage **(**P-value=1.07E-11 calculated with ROC curve analysis; **Supplemental Fig. 5.5).** We also performed 5' RACE on genes RNAseq scores (between 30 and 257), which validated an alternative start-site usage in each.

We employed 5'RNAseq to assess the transcriptional diversity in the hearts of MHC$^{403/+}$ mice, which carry a human HCM missense mutation, Arg403Gln, in the myosin heavy chain gene[5-7]. Male MHC$^{403/+}$ mice recapitulate the human disease and develop increased left ventricular (LV) wall thickness (hypertrophy) with myocyte hypertrophy and disarray as well as increased amounts of myocardial fibrosis. Prior comparison of the LV transcriptomes from adult male wildtype and MHC$^{403/+}$ mice showed hundreds of genes with altered levels of expression[8,9]. We observed very high correlation (r=0.9) in the altered expression (increased or decreased) of genes defined by RNAseq (**Supplemental Table 5.1**) and these earlier studies.

From the UCSC Genome browser we identified 8,000 annotated mouse genes that have more than one transcriptional start-site (**Supplemental methods**). Genome-wide assessment of start-site usage by 5'RNAseq analyses of MHC$^{403/+}$ hearts revealed 97 genes with scores ≥20, indicative of significant fold changes in transcriptional start-site usage (**Fig. 5.1c, Supplemental Table 5.2**, **Supplemental Table 5.3** and **Supplemental Image Data**). Most genes with altered start-site usage also had increased expression in MHC$^{403/+}$ compared to wildtype hearts (P value = 1.1E-74 calculated by Chi-square**).**

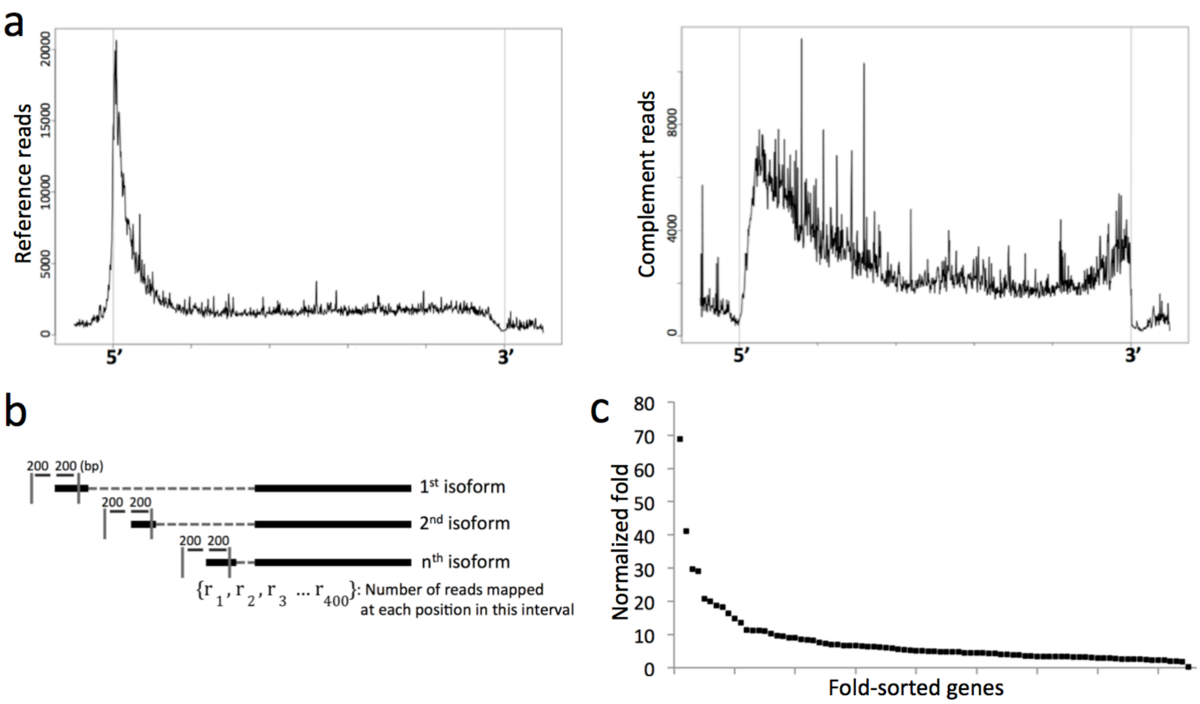


Figure 5.1 RNA-seq without RNA or cDNA fragmentation allows sensitive assessment of start-site changes. (a) Enhanced sequencing at 5' ends of transcripts and directional information. Reads with transcript reference (left panel) or complement strand (right panel) plotted at positions normalized for transcript length. (b) Bioinformatic approach assessing changes in distributions at gene start-sites. (c) Distribution of fold of start-site changes in HCM distinguishes Four-and-a-half LIM domains protein 1 having the most robust change. Plotted are changes for genes that are statistically significant (p-value<0.05 calculated by Chi-square).

Fhl1, which encodes four-and-a-half LIM domains protein-1 had both markedly increased expression in MHC$^{403/+}$ hearts and exhibited the most significant change in start-site usage (69-fold; P value= 4.8E-24). *Fhl1* transcriptional profiles **(Fig. 5.2a)** revealed that wildtype mice primarily used a start-site (denoted *b*Fhl1, basal) located 5' to an alternative start-site (denoted *s*Fhl1, stress-induced) that was predominant in MHC$^{403/+}$ mice. Since normalized transcriptional read-depths are similar for *b*Fhl1 in wildtype and MHC$^{403/+}$, the markedly increased *Fhl1* expression in mutant hearts reflects use of the *s*Fhl1 start-site. The *s*Fhl1 translational site is predicted to incorporate 16 amino acids (NCBI accession: NP_001070830) that are absent from *b*Fhl1 transcripts.

We performed 5'RACE to validate and provide semi-quantitative assessment of *Fhl1* start-site usage in the hearts of wildtype and MHC$^{403/+}$ mice (**Fig. 5.2b** and **Supplemental Fig. 5.6**). Dideoxy sequences of 5'RACE fragments indicated that *b*Fhl1 predominated in wildtype hearts and *s*Fhl1 in MHC$^{403/+}$ hearts (**Supplemental Fig. 5.7,** which shows 5'RACE data confirming 5'RNA-seq start-site usage assessment). To confirm the impact of transcriptional changes in *Fhl1* on protein expression, we also performed Western blots of cardiac lysates. Consistent with increased and altered start-site usage, the *Fhl1* protein levels were markedly increased (**Fig. 5.2c**) and slightly larger (~2KDa) in lysates from mutant compared to wildtype mice.

We considered whether altered *Fhl1* expression in MHC$^{403/+}$ hearts reflected transcriptional changes in myocytes, which express the mutant sarcomere protein, or in non-myocytes, which proliferate and contribute to increased fibrosis in MHC$^{403/+}$ hearts. RNAseq of isolated myocytes and non-myocytes revealed *Fhl1* expression in both cardiac cell populations (**Supplemental Table 5.3**), albeit at lower levels than was detected in whole hearts. Non-myocytes isolated from wildtype or MHC$^{403/+}$ mice expressed comparable levels of *b*Fhl1 transcripts and no *s*Fhl1 transcripts. In contrast, the *Fhl1* transcriptional profiles from myocytes paralleled those in whole hearts: MHC$^{403/+}$ myocytes had higher *Fhl1* levels, with predominantly *sFhl1* transcripts in comparison to wildtype myocytes (**Fig. 5.2d** and **Supplemental Fig. 5.8**).

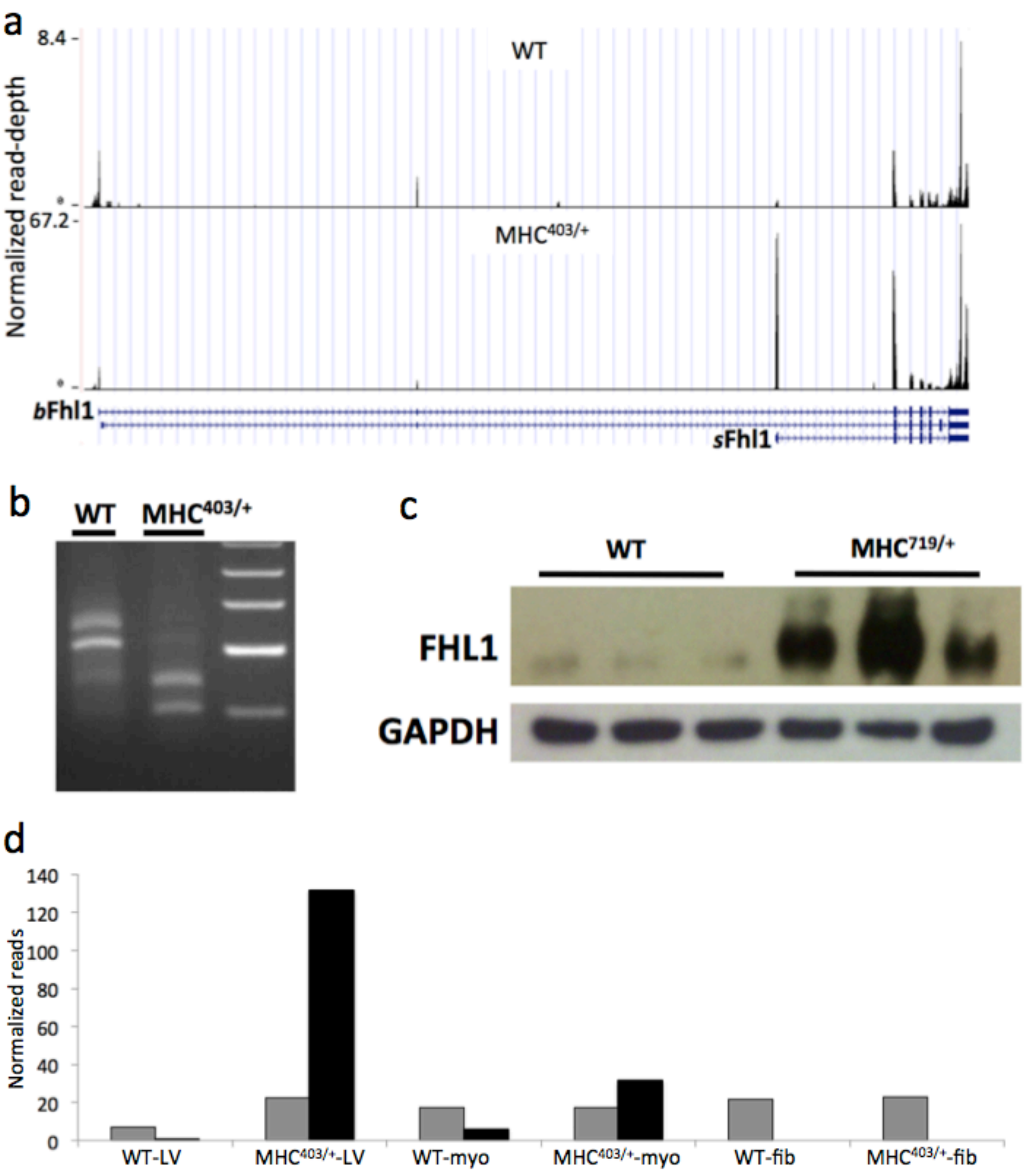


Figure 5.2 Marked upregulation of a low-expressing isoform in myocytes accounts for *Fhl1* transcriptional switch and results in increased protein levels. (a) Normalized read depth for the *Fhl1* gene locus. *b*Fhl1 (basal) and *s*Fhl1 (stressed-induced) isoforms are indicated. *s*Fhl1 is markedly upregulated in MHC$^{403/+}$ hearts. (b) Fhl1 5'RACE product run on an agarose gel is consistent with RNA-seq data (Analysis shown in Supplemental Fig. 5.7). (c) Western blot showing *Fhl1* increased levels in HCM. A mild increase in the size of the upregulated protein is accounted by an earlier translational start in the *s*Fhl1 isoform. (d) *Fhl1* start-site switch occurs in the myocytes (myo). *s*Fhl1 isoform is not expressed in non-myocytes (fib) from either wild-type or MHC$^{403/+}$ hearts. Reads were normalized to total reads per million excluding mitochondrial reads. (Grey bar: *b*Fhl1, Black bar: *s*Fhl1.)

We hypothesized that induction of the *sFhl1* isoform might have functional roles in the myocyte responses to HCM mutations. To consider the consequences of *Fhl1* ablation in HCM mice we crossed *Fhl1*-null mice that express lacZ under the endogenous *Fhl1* promoter[10] with MHC$^{403/+}$ mice. As *Fhl1* is encoded on chromosome X we studied hemizygous male (Fhl1^{lacZ}) and homozygous female (Fhl1$^{lacZ/lacZ}$) mice. The LV dimensions and function of adult male and female *Fhl1*-null mice were comparable to wildtype controls (**Table 5.1**). In contrast male MHC$^{403/+}$/Fhl1^{lacZ} mice and female MHC$^{403/+}$/ Fhl1$^{lacZ/lacZ}$ mice developed more marked HCM hypertrophy and histopathology than MHC$^{403/+}$ mice **(Table 5.1)**.

Table 5.1. Echocardiographic measurements of left ventricular wall thickness (LVWT). To assess the effect of *Fhl1* in hypertrophy in the presence of a sarcomere mutation, p-values were computed (student's two-tailed t-test). Shown also are *Fhl1 null* control mice that do not develop hypertrophy (one-tailed test).

Genotype	**Sex**	**Numbe r**	**LVWT**	***P*value (vs MHC$^{403/+}$)**
MHC$^{403/+}$	M	12	0.99±0.33	
MHC403$^{/+}$/*Fhl1 null*	M	12	1.31±0.37	0.039
Fhl1 null	M	6	0.72±0.05	0.036
MHC$^{403/+}$	F	11	0.95±0.39	
MHC$^{403/+}$/*Fhl1 null*	F	7	1.44±0.38	0.017
Fhl1 null	F	4	0.77±0.05	0.2
Wild-type	M	3	0.70±0.03	
Wild-type	F	3	0.69±0.09	

Using RNA-seq we confirmed that compound wildtype/ Fhl1$^{lacZ/lacZ}$ and MHC$^{403/+}$/ Fhl1$^{lacZ/lacZ}$ mice maintained the *Fhl1* transcriptional patterns defined in wildtype and MHC$^{403/+}$ mice, respectively (**Supplemental Fig. 5.9)**. We also confirmed that both the *b*Fhl1 and *s*Fhl1 transcripts encoded lacZ sequences in frame (not shown). β-galactosidase (β-gal) assays detected low baseline expression in cardiac tissues from compound wildtype / Fhl1$^{lacZ/lacZ}$ mice. Young, pre-hypertrophic male MHC$^{403/+}$/ Fhl1^{lacZ} mice had minimal and focal β-gal expression in myocytes adjacent to fibrotic regions, that was demarcated by Sirius red (**Fig. 5.3a, bottom left panel**). Consistent with increased transcription of the *s*Fhl1 isoform, male MHC$^{403/+}$/ Fhl1^{lacZ} mice with overt hypertrophy had robust myocyte β-gal expression.

As prior studies identified increased *Fhl1* expression in the LV of mice with pressure overload hypertrophy[11] as well as in human HCM hearts[12] we examined *Fhl1* transcriptional changes in other cardiac pathologies. LV RNAseq data from LV samples derived from patients with HCM, dilated cardiomyopathy, aortic stenosis, and heart failure each showed increased *Fhl1* expression due to induction of *s*Fhl1 transcripts (**Fig. 5.3b**).

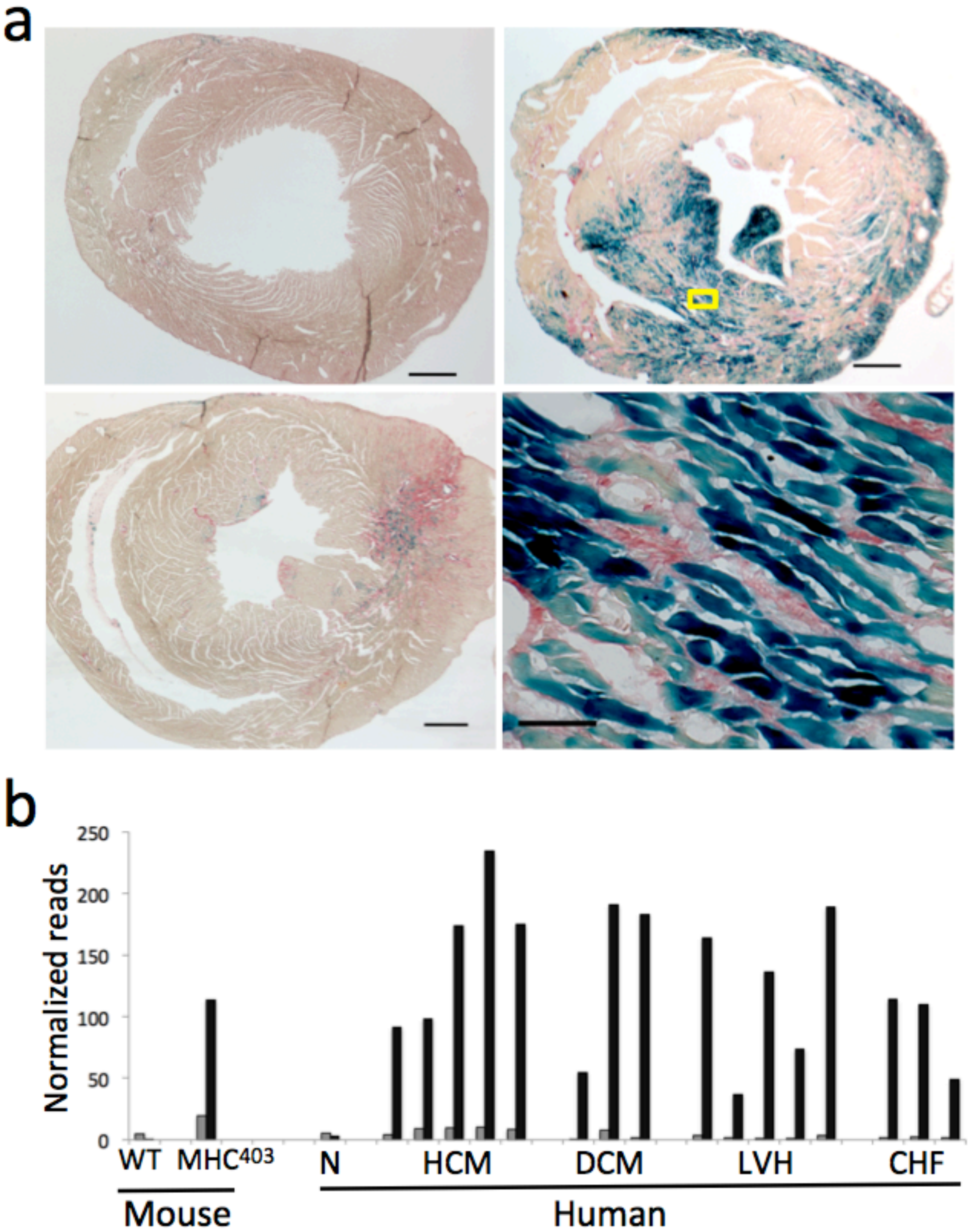


Figure 5.3 *Fhl1* upregulation co-occurs with fibrosis, and is universally found in other cardiomyopathies. (a) Histologic analysis co-localizes *Fhl1* upregulation (blue staining from β-gal with X-gal) with fibrosis (red staining from collagen with Sirius red). All sections carry β-gal driven by the *Fhl1* promoter. Experimental sections were processed in parallel with control sections. (upper left) Control mouse with wild-type myosin heavy chain and no upregulated *Fhl1* or fibrosis. (lower left) Prehypertrophic MHC$^{403/+}$ exhibit *Fhl1* upregulation in myocytes in the

(Figure legend 5.3, Continued) prehypertrophic scar region. (upper right) *Fhl1* upregulation in myocytes co-occurs with fibrosis in MHC$^{403/+}$ with overt hypertrophy. (lower right) Magnified region of upper right panel (indicated in a yellow box). (Bars: 0.5mm for left and upper right panels; 0.05mm for lower right panel). (b) Transcriptional switch from *s*Fhl1 upregulation occurs in primary and secondary cardiomyopathies. WT: wild-type; N: normal; HCM: hypertrophic cardiomyopathy; DCM: dilated cardiomyopathy; LVH: pressure-overload left ventricular hypertrophy; CHF: congestive heart failure. The isoform switch is conserved between mouse and human where it also occurs between the corresponding isoform orthologues. Grey bar: *b*Fhl1, Black bar: *s*Fhl1. Libraries from mouse samples were constructed from pooled replicates; human samples were constructed from individual replicates.

Collectively these data demonstrate that *s*Fhl1 transcriptional changes are a prominent myocyte response to HCM and other pathologies. Increased *Fhl1* expression in myocytes, by activation of an alternative transcriptional start-site, implicated *s*Fhl in either the pathogenic pathways triggered by a sarcomere gene mutation or in compensatory responses to HCM. Gene ablation studies clarified this ambiguity; MHC$^{403/+}$/ *Fhl1*-null mice had significantly worse HCM, therein defining induction of *s*Fhl1 as beneficial to myocytes. Moreover, as this transcript was increased in a wide range of cardiac pathologies, *s*Fhl1 induction appears to be a common adaptive response to myocyte stress.

Recently three rare human variants in *Fhl1* have been hypothesized to cause HCM[13] that exhibited incomplete penetrance in females and an associated skeletal myopathy in some males. That *Fhl1*-null mice had normal cardiac dimensions and function challenges this conclusion, and raises the possibility that human variation in *Fhl1*, in particular those that cause a loss of function, prevent the salutary effects that increased levels and isoform switching of *Fhl1* have on cardiac disease.

Given its location on chromosome X, the protective effects of *s*Fhl1 may contribute to the gender-specific outcomes observed in HCM. Although autosomal dominant sarcomere gene mutations cause HCM, clinical studies demonstrate more severe LV dysfunction in men compared to women with HCM[14], and more sudden death events during athletic participation.

Our data shows a strong trend whereby a large fraction of female MHC$^{403/+}$/ Fhl1-null mice develop severe hypertrophy (43% females versus 25% males; LVWT ≥ 1.6), even though due to the small sample size it does not reach statistical significance. We speculate that dosage of *s*Fhl1 and perhaps human variation contributes to these important clinical differences.

In conclusion, development of a novel RNAseq strategy that provides genome-wide ascertainment of transcriptional diversity uncovered dynamic regulation of *s*Fhl1 in HCM myocytes. To our knowledge, *s*FHl1 represents the first beneficial genetic modifier in HCM. We suggest that elucidation of the differences in protein-protein interactions conveyed by the 16 amino acids and identification of molecules that activate *s*Fhl warrants further study.

ACKNOWLEDGMENTS

We are very grateful to Polakit Teekakirikul, Alain Watts, Victor Soukoulis, Michael Springer, Suzanne White, Lena Liu, Kelly Maloney (Harvard Medical School).

AUTHOR CONTRIBUTIONS

DCC, HW, JGS, CES designed the study; DCC, JMG, DMM optimized library construction steps; DCC and JMG constructed libraries; DCC and SRD wrote computational programs and analyzed sequence data; HW and DCC performed echocardiography; KO, DCC, HW performed cell dissociation analyses; DCC, SE, HW performed protein and histological analyses; GS, DCC contributed with the mathematical description; DAC designed genotyping primers; SC, DSH and JDM contributed with samples and feedback; AA, MLB, DCC performed ROC analyses.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

METHODS

Model Organisms. All study protocols were approved by the Animal Care and Use Committee of Harvard Medical School. aMHC$^{403/+}$ in 129/SvEv and *Fhl1* null in Black Swiss crossed mice were studied. Newly designed primers for genotyping *Fhl1* null mice were used: AAACCAGGCAAAGCGCCATTCG, GGCCCACTTGTCCTCTGAGTCAGC, TGCAAGGGGTCCCTGCAGTAGTGA. Used in the same reaction, the combination of these three primers produce a ~370bp (wild-type), ~510bp (knock-out) bands. Cyclosporin A (CsA) was administered in chow (100mg CsA/100g chow) to accelerate hypertrophic remodeling[9]. Hypertrophic MHC$^{403/+}$ mice were used (assessed by echordiography) unless when indicated.

Echocardiographic studies. Mice were anesthetized with an isoflurane vaporizer (VetEquip) and each limb was placed on the ECG leads on a Vevo Mouse Handling Table (VisualSonics Inc.) maintaining the body temperature at 37ºC during the study. Transthoracic echocardiography was performed using Vevo 770 High-Resolution In Vivo Micro-Imaging System and RMV 707B scan-head (VisualSonics Inc.) with the heart rate at 500-550 bpm. The images were acquired as 2D (left parasternal long and short axes) and M-mode (left parasternal short axis). Measurements averaged from 3 consecutive heart beats of M-mode tracings were used for LV end-diastolic diameter (LVEDD), LV end-systolic diameter (LVESD), and wall thickness (interventricular septum, IVS and left ventricular posterior wall, LVPW). LV fractional shortening (%) was calculated as follows: (LVEDD– LVESD)/LVEDD × 100. All echocardiographic measurements were done blinded to mouse genotype.

RNA-seq library construction and analysis. To reduce biological variation, RNA was pooled from biological replicates. RNA-seq was performed as previously described with no RNA or cDNA fragmentation and without any normalization[15]. Uniform amplification of the cDNA library was achieved with amplification cycles before the reaction reached saturation determined with qPCR. Following sequencing, alignment of reads was performed with Tophat[16]. Read-depth profiles were constructed with Tophat's "wiggles" tool and values were normalized to total aligned reads and uploaded on the UCSC browser[17] or IGV[18]. Read-depths shown here were adapted after visualizing on UCSC Browser. Gene expression profiles are constructed by tallying reads on gene loci and comparisons utilized a Bayesian p-value[19,20]**.** Samtools[21,22] was used to generate and process binary alignment files. Plots of directional reads were generated by assigning gene reads to bins based on the position of the read on the transcript including upstream and downstream regions.

Start-site analysis algorithm. For the genome with 'm' genes $\{x_1, x_2, x_3 \ldots x_m\}$, let's define a typical gene model and mathematical algorithm. Let's say gene 'x' has 'n' isoforms and each isoform has a unique 5' position (Fig. 5.1b).

Let's say total number of reads for gene 'x' is

x^a for sample 'a'

x^b for sample 'b'

As shown (Fig. 5.1b), each isoform has a 5' interval of 400 bp and let's say each position has some read coverage and is represented as

$R_i^{x,a} = \{r_1, r_2, r_3 \ldots r_{400}\}_i^{x,a}$

where $R_i^{x,a}$ = vector of reads at each position in ith isoform for 'x' gene in 'a' sample.

We normalize each vector with corresponding total reads:

NRD$_{i}^{x,a}$ = R$_{i}^{x,a}$/x^{a} where 'a' = sample 'a'

x = gene 'x'

x^{a}= total reads for gene 'x'

NRD$_{i}^{x,a}$ = normalized read depth for ith isoform

for gene 'x' in sample 'a'

We calculate a normalized read-depth ratio (NRDR) vector:

NRDR$_{i}^{x}$ = {NRD$_{i}^{x,a}$/NRD$_{i}^{x,b}$} = {nrdr$_{1}$, nrdr$_{2}$, nrdr$_{3}$ … nrdr$_{400}$}$_{i}^{x}$

After calculating NRDR$_{i}^{x}$, we calculate p-values based on Bayesian test[19]. We get a p-value vector for 'x' gene at 'i' isoform:

P$_{i}^{x}$ = {p$_{1}$, p$_{2}$, p$_{3}$ … p$_{400}$}$_{i}^{x}$

Using these NRDR$_{i}^{x}$ and P$_{i}^{x}$, we assign a binary vector to each isoform 'i' of gene 'x' using formula below, where 'j' is the position in the 5' interval:

f(b$_{j}^{i,x}$) = { 1 if nrdr$_{j}^{i,x}$ ≥ foldcut-off

or nrdr$_{j}^{i,x}$ ≤ 1/foldcut-off

AND $p^{i,x}_{j}$ < p-valuecut-off

0 otherwise }

Using this function, we prepare a matrix of binary vectors for each isoform ‘i’ of each gene ‘x’ in the genome and assign a score S_{x} as below:

$$S_{x} = \sum_{i=1}^{n} \sum_{j=1}^{400} b^{i,x}_{j}$$

This way for each gene ‘x’ from the genome, we get a score vector for all ‘m’ genes:

$$S = \{S_{x}: \forall\ x \in [1, m]\}$$

We implemented this mathematical concept using Perl as shown (Supplemental software package).

Myocyte and non-myocyte isolation. Cells were isolated using the Langendorff heart preparation as previously performed[9]. In addition, the final myocyte pellet was resuspended in MEM with 5% FBS and 2mM L-glutamine and plated in laminin-coated culture dishes (in 2% CO_2 incubator at 37°C). After a 1-hour incubation, cells were rinsed with sterile PBS to remove nonadherent cells and debris and TRIzol (Invitrogen) was added to extract total RNA according to manufacturer protocols.

5’RACE. 5’RACE was performed using a commercially available kit (Ambion, cat#AM1700) with amplification performed using rTth DNA polymerase (Applied Biosystems, cat#N808-0188). For

Fhl1, an outer primer with sequence TCCAGATGTGATGGCCTTGTTGCACTT and an inner primer with sequence ACATGGTGCCCACCTTATAGCTGGA were used.

Western blotting and histology staining. Western blot was performed as previously described using mice harboring another mutation (MHC$^{719/+}$) that causes hypertrophic cardiomyopathy[9]. An *Fhl1* mouse monoclonal antibody was used (Abcam cat. Ab58067). For histology, frozen sections were prepared and incubated with X-gal for 30 minutes. Sections were then washed in PBS, fixed in 4% PFA for 40 minutes, washed in PBS again, and stained with Sirius red as previously described[23] and mounted with Permount.

References

1. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* **5**, 621-628 (2008).

2. Sultan, M.*, et al.* A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321**, 956-960 (2008).

3. Nagalakshmi, U.*, et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344-1349 (2008).

4. Trapnell, C.*, et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* **7**, 562-578 (2012).

5. Geisterfer-Lowrance, A.A.*, et al.* A mouse model of familial hypertrophic cardiomyopathy. *Science* **272**, 731-734 (1996).

6. Seidman, J.G. & Seidman, C. The genetic basis for cardiomyopathy: from mutation identification to mechanistic paradigms. *Cell* **104**, 557-567 (2001).

7. Geisterfer-Lowrance, A.A.*, et al.* A molecular basis for familial hypertrophic cardiomyopathy: a beta cardiac myosin heavy chain gene missense mutation. *Cell* **62**, 999-1006 (1990).

8. Kim, J.B.*, et al.* Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science* **316**, 1481-1484 (2007).

9. Teekakirikul, P.*, et al.* Cardiac fibrosis in mice with hypertrophic cardiomyopathy is mediated by non-myocyte proliferation and requires Tgf-beta. *The Journal of clinical investigation* **120**, 3520-3529 (2010).

10. Sheikh, F.*, et al.* An FHL1-containing complex within the cardiomyocyte sarcomere mediates hypertrophic biomechanical stress responses in mice. *The Journal of clinical investigation* **118**, 3870-3880 (2008).

11. Song, H.K., Hong, S.E., Kim, T. & Kim do, H. Deep RNA sequencing reveals novel cardiac transcriptomic signatures for physiological and pathological hypertrophy. *PloS one* **7**, e35552 (2012).

12. Lim, D.S., Roberts, R. & Marian, A.J. Expression profiling of cardiac genes in human hypertrophic cardiomyopathy: insight into the pathogenesis of phenotypes. *Journal of the American College of Cardiology* **38**, 1175-1180 (2001).

13. Friedrich, F.W.*, et al.* Evidence for FHL1 as a novel disease gene for isolated hypertrophic cardiomyopathy. *Human molecular genetics* **21**, 3237-3254 (2012).

14. Page, S.P.*, et al.* Cardiac myosin binding protein-C mutations in families with hypertrophic cardiomyopathy: disease expression in relation to age, gender, and long term outcome. *Circulation. Cardiovascular genetics* **5**, 156-166 (2012).

15. Christodoulou, D.C., Gorham, J.M., Herman, D.S. & Seidman, J.G. Construction of normalized RNA-seq libraries for next-generation sequencing using the crab duplex-specific nuclease. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]* **Chapter 4**, Unit4 12 (2011).

16. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009).

17. Kuhn, R.M., Haussler, D. & Kent, W.J. The UCSC genome browser and associated tools. *Briefings in bioinformatics* (2012).

18. Robinson, J.T.*, et al.* Integrative genomics viewer. *Nature biotechnology* **29**, 24-26 (2011).

19. Audic, S. & Claverie, J.M. The significance of digital gene expression profiles. *Genome research* **7**, 986-995 (1997).

20. Christodoulou, D.C.*, et al.* Quantification of gene transcripts with deep sequencing analysis of gene expression (DSAGE) using 1 to 2 microg total RNA. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]* **Chapter 25**, Unit25B 29 (2011).

21. Li, H.*, et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

22. Ramirez-Gonzalez, R.H., Bonnal, R., Caccamo, M. & Maclean, D. Bio-samtools: Ruby bindings for SAMtools, a library for accessing BAM files containing high-throughput sequence alignments. *Source code for biology and medicine* **7**, 6 (2012).

23. Fujisawa, G., Dilley, R., Fullerton, M.J. & Funder, J.W. Experimental cardiac fibrosis: differential time course of responses to mineralocorticoid-salt administration. *Endocrinology* **142**, 3625-3631 (2001).

CHAPTER 6

Chapter 6. Discussion:

Built upon the biological and engineering advancements of the past century, the recently developed ability to sequence millions of DNA fragments simultaneously with relative ease is revolutionary for the field of biology. This technological breakthrough creates opportunities to measure fundamental biological properties with unprecedented resolution, offering new avenues to investigate the underlying processes that are responsible for certain diseases. While the high throughput of DNA sequencing (now on the order of 100 million sequences at a cost of less than 3-4 thousand dollars and falling)[3] allows for high resolution measurements, similar analysis of the transcriptome offers a substantial challenge, as transcripts are expressed in a very wide dynamic range (some transcripts are found at significant proportions while others of importance such as transcription factors are found in proportions as low as 1-2 copies per million) (Chapter 3). While transcripts comprising the mRNA are typically longer than 200bp, miRNAs are tiny, with sizes around ~20bp (Chapter 4). Thus, comprehensive sequencing of RNA species is currently not possible by using one method alone.

This dissertation is concerned with the development of approaches to query different aspects of the transcriptome offering the potential to provide new biological information. Deep sequencing analysis of gene expression (Chapter 2) offers the ability to quantify the dynamic range of gene expression. Normalization with the duplex-specific nuclease offers the ability to dramatically increase the sequencing of low-expressing transcripts by degrading the high-expressing species (Chapter 3). Additionally, sequencing of miRNAs also requires development of dedicated protocols

[3] National Human Genome Institute

given the small size of miRNAs (Chapter 4). Although not pursued experimentally or discussed here further, additional work led to computational methods to compare exon read coverage, identify mutations in transcriptional data, and analyze ChIP-seq data.

Lastly, gene expression changes can be assayed accurately with RNA-seq without RNA or cDNA fragmentation (Chapter 5). The chemistry of the library construction favors sequencing from the 5' ends of transcripts, providing for a greater depth of sequencing at these regions. This is beneficial as it gives a specific signature of the location of the sequencing, resulting in another level of quality validation (in particular for low-expressing transcripts), as well as leading to the ability to assess changes at start-site usage.

These methods can be applied to ubiquitous biological settings. Application of gene expression analysis on cells that were genomically edited or mouse hearts in which histone methylation was perturbed is briefly described (Appendix 2).

Application of this new approach in Hypertrophic cardiomyopathy (HCM) led to the identification of Four-and-a-half LIM domains protein 1 (Fhl1) as the gene displaying the most marked change in start-site usage. This start-site usage change is very robust (~70 fold), is accurately conserved between mouse and human, and occurs in all cardiomyopathies assessed. Further, it co-occurs histologically with fibrosis which may explain why start-site change appears to be global in cardiomyopathy. Interestingly, Fhl1 genetic ablation worsens the hypertrophic response in mouse carriers of the sarcomere mutation. In other words, the sarcomere mutation in the absence of Fhl1 expresses the HCM pathology more strongly, indicating that Fhl1 as an adaptive modifier of HCM.

The role of Fhl1 in HCM is contrary to the one identified in pressure overload hypertrophy, where Fhl1 ablation alleviated the hypertrophic response (Chapter 5). The root causes of pathologic hypertrophy in pressure overload and of HCM caused by a sarcomere mutation are different. While left ventricular hypertrophy caused by pressure overload is characterized by increased hemodynamic load, a sarcomere mutation results in enhanced biophysical properties of the contracting sarcomere. This would imply, assuming the two cardiomyopathies converge at later stages, that Fhl1 has a very early role in the pathology. Particularly, a sarcomere gene mutation in the early stages of HCM leads to enhanced myocyte contractile properties which results in disease, possibly through disturbance of energy homeostasis, i.e. overall energy depletion (Chapter 1). Fhl1, which is thought to act as a biomechanical sensor in the heart [1, 2], may inhibit the contractile properties of myocytes, an effect that would be beneficial in the context of an overactive sarcomere in HCM but not in pressure overload (Chapter 5). An association of Fhl1 with sarcomere components could result in energy transmission from the sarcomere to Fhl-mediated downstream signals. Another possible explanation is that Fhl1 insertion in the sarcomere could modify the myosin-actin interaction [3], also resulting in slowing down of the overactive sarcomere. Some evidence that Fhl1 mutations can exert effects on contractility has been shown [4]. If Fhl1 does not change the contractility of myocytes, its compensatory effect in HCM may be due to downstream signaling. This is supported by the identification of multiple binding partners including DNA [3, 5 - 9]. Thus, Fhl1 may also exert effects in multiple locations. As reviewed (Chapter 1), different pathways are thought to be mechanistically involved in the development of HCM and Fhl1 could be having unrecognized roles in some of these pathways. Fhl1 mutations have been firmly associated with muscular diseases [10-20]. Evidence also supports a role of Fhl1 in the cardiovascular system [1, 3, 21, 22].

This dissertation identifies Fhl1 as a compensatory modifier of HCM and discusses its possible modes of action. Generalizing this observation, a factor that brings the biophysical properties closer to physiological levels or alleviates downstream pathologic processes can be an adaptive modifier. Conversely, a factor that increases the contractile properties of the sarcomere or the downstream consequences of the sarcomere mutation and that adds additional cellular strain (e.g. by limiting energetics further) has the potential to be a disease-enhancing modifier.

The importance of modifier genes cannot be overstated, as modifier genes can provide therapeutic avenues to prevent the onset of disease in many people carrying pathogenic mutations. As discussed (Chapter 1), knowledge of modifier genes and the factors that lead to their induction or suppression can lead to the identification of potential therapeutic regimens. Modifier genes are genes that can modify the expression of a disease-causal allele. This can be seen when variants in these genes -- knocking or changing the function of the protein -- have a demonstrable effect in disease expression. Mouse models may provide a setting for testing potential modifier genes. E.g., knock-out mice recaptitulate loss-of-function variants and provide an understanding of the basic properties of the gene. As Fhl1 is on the X chromosome, one hit is sufficient to affect its function in males.

While we performed preliminary experiments to identify upstream factors that upregulate Fhl1 in HCM and assessed the role of estrogen signaling or serum response factor, such factors still remain elusive. Two possibilities appear to have potential for future investigation: Mef2 signaling [23] and DNA methylation [24].

Three future directions could be helpful for further elucidating disease processes discussed in this dissertation. One would be full identification of modifier genes by comparing DNA from individuals with deleterious sarcomere mutations who developed or did not develop HCM. In addition, this can be complemented with mixed background mouse models crossed with the sarcomere mutation. With the decreasing cost of exome and whole genome sequencing, this analysis will become possible in the near future. Second, high-throughput screens measuring changes in contractility can be performed to identify modifying factors. Predicting which factors may have modifying effects on contractility can be useful clinically [25]. Such an approach can be used to delineate the effect of multiple Fhl1 mutations. Also, additional screens can be conceived focusing on established critical downstream processes of the sarcomere mutation. Third, a mechanistic follow-up on Fhl1 can provide a real example of how a modifier gene can prevent disease pathology. The signals that affect Fhl1 expression in HCM can be investigated. This also provides a setting to assess whether a genetic modifier can prevent or reverse the onset of pathology by using in vivo overexpression techniques of the Fhl1 isoforms.

Future work might address the role of Fhl1 in disease while carefully examining the role of each isoform separately. For example, the role of the 16 amino-acid sequence on the N-terminus of the induced isoform can be assessed. Mice overexpressing both isoforms can be generated. These mice can be used to assess whether overexpression of Fhl1 can inhibit HCM further and whether the effects from overexpressing either isoform differ. The 16 amino-acid sequence on the longer isoform is conserved between mouse and human, which should predict a physiologic function, i.e. a potential biochemical difference. In addition, such mouse models can be used to assess Fhl1 roles in other tissues and disease settings. Fhl1 may have distinct roles in the various tissues it is putatively expressed in, such as spleen, thymus, testis ovary, brain, placenta, lung, liver, kidney and pancreatic tissue [9], that could also be ascertained.

In addition, Fhl1 has nine zinc fingers with which it may be involved in biomechanical sensing and possibly in integrating multiple signals. Elucidating this further in the context of an overactive sarcomere and the two Fhl1 isoforms is possible by using biophysical models and by mutating key amino-acid substrates. As zinc fingers are important protein domains that mediate critical cellular interactions, such as protein-protein or protein-nucleic acid, the presence of nine zinc fingers could potentially result in binding to a large number of proteins. Also, the combination of these domains on a single protein molecule may create a structural domain that may be functionally important in cardiovascular physiology.

Examining the role of Fhl1 in other cardiomyopathies for which mouse models exist might provide useful information. Such experiments may define the role of Fhl1 in other cardiomyopathies, providing a functional assessment of common and divergent processes. Biophysical experiments have shown that sarcomere mutations result in worsened contractile properties in Dilated Cardiomyopathy (DCM), which is the opposite compared to HCM [26]. This suggests that Fhl1 ablation may ameliorate DCM in that setting. However, it is unclear what the Fhl1 effect would be in other settings such as DCM resulting from a non-sarcomere mutation or in HCM-like cardiomyopathies.

Increasing our understanding of the role of Fhl1 in cardiomyopathies and the involvement of Fhl1 in cardiac processes can provide additional mechanistic insights. Pathways that Fhl1 participates in can be dissected with transcriptional profiling using knock-out and possibly transgenic models. These pathways are likely to be affected in models of cardiomyopathy in similar and different ways.

Future work exploring such avenues has the potential to address novel questions of clinical importance and to increase understanding of the role of differential isoform regulation in disease as well as in other disease settings. For example, assessment of the role of Fhl1 in other cardiomyopathies may provide a clinically/therapeutically meaningful axis for further categorizing and understanding cardiomyopathies. Combined with results from identifying other modifiers of cardiomyopathy, this can contribute to a multi-gene approach that can be used to silence disease pathology in multiple settings even when the causal mutation causes severe effects and cannot be targeted with other approaches.

This dissertation offers four methods to analyze the transcriptome and identifies Four-and-a-half LIM domains protein 1 as an adaptive modifier in HCM. Further identification of genetic modifiers and clarification of their exact mode of action, with Fhl1 providing such an early opportunity, can improve our understanding of the underlying mechanisms of cardiomyopathies and potentially reveal avenues to novel treatments.

References

1. Sheikh, F., A. Raskin, P. H. Chu, S. Lange, A. A. Domenighetti, M. Zheng, X. Liang, T. Zhang, T. Yajima, Y. Gu, N. D. Dalton, S. K. Mahata, G. W. Dorn, 2nd, J. H. Brown, K. L. Peterson, J. H. Omens, A. D. McCulloch and J. Chen (2008). "An FHL1-containing complex within the cardiomyocyte sarcomere mediates hypertrophic biomechanical stress responses in mice." J Clin Invest 118(12): 3870-80.

2. Raskin, A., S. Lange, K. Banares, R. C. Lyon, A. Zieseniss, L. K. Lee, K. G. Yamazaki, H. L. Granzier, C. C. Gregorio, A. D. McCulloch, J. H. Omens and F. Sheikh (2012). "A novel mechanism involving four-and-a-half LIM domain protein-1 and extracellular signal-regulated kinase-2 regulates titin phosphorylation and mechanics." J Biol Chem 287(35): 29273-84.

3. Shathasivam, T., T. Kislinger and A. O. Gramolini (2010). "Genes, proteins and complexes: the multifaceted nature of FHL family proteins in diverse tissues." J Cell Mol Med 14(12): 2702-20.

4. Friedrich, F. W., B. R. Wilding, S. Reischmann, C. Crocini, P. Lang, P. Charron, O. J. Muller, M. J. McGrath, I. Vollert, A. Hansen, W. A. Linke, C. Hengstenberg, G. Bonne, S. Morner, T. Wichter, H. Madeira, E. Arbustini, T. Eschenhagen, C. A. Mitchell, R. Isnard and L. Carrier (2012). "Evidence for FHL1 as a novel disease gene for isolated hypertrophic cardiomyopathy." Hum Mol Genet 21(14): 3237-54.

5. Chu, P. H. and J. Chen (2011). "The novel roles of four and a half LIM proteins 1 and 2 in the cardiovascular system." Chang Gung Med J 34(2): 127-34.

6. Sharma, P., T. Shathasivam, V. Ignatchenko, T. Kislinger and A. O. Gramolini (2011). "Identification of an FHL1 protein complex containing ACTN1, ACTN4, and PDLIM1 using affinity purifications and MS-based protein-protein interaction analysis." Mol Biosyst 7(4): 1185-96.

7. Yang, Z., C. F. Browning, H. Hallaq, L. Yermalitskaya, J. Esker, M. R. Hall, A. J. Link, A. J. Ham, M. J. McGrath, C. A. Mitchell and K. T. Murray (2008). "Four and a half LIM protein 1: a partner for KCNA5 in human atrium." Cardiovasc Res 78(3): 449-57.

8. Lee, J. Y., I. C. Chien, W. Y. Lin, S. M. Wu, B. H. Wei, Y. E. Lee and H. H. Lee (2012). "Fhl1 as a downstream target of Wnt signaling to promote myogenesis of C2C12 cells." Mol Cell Biochem 365(1-2): 251-62.

9. Kadrmas, J.L., M.C. Beckerle (2004). "The LIM domain: from the cytoskeleton to the nucleus." Nat Rev Mol Cell Biol. 5(11):920-31.

10. Bonne, G., F. Leturcq and R. Ben Yaou (1993). Emery-Dreifuss Muscular Dystrophy. GeneReviews. R. A. Pagon, T. D. Bird, C. R. Dolan, K. Stephens and M. P. Adam. Seattle (WA).

11. Cowling, B. S., D. L. Cottle, B. R. Wilding, C. E. D'Arcy, C. A. Mitchell and M. J. McGrath (2011). "Four and a half LIM protein 1 gene mutations cause four distinct human myopathies: a comprehensive review of the clinical, histological and pathological features." Neuromuscul Disord 21(4): 237-51.

12. Gueneau, L., A. T. Bertrand, J. P. Jais, M. A. Salih, T. Stojkovic, M. Wehnert, M. Hoeltzenbein, S. Spuler, S. Saitoh, A. Verschueren, C. Tranchant, M. Beuvin, E. Lacene, N. B. Romero, S. Heath, D. Zelenika, T. Voit, B. Eymard, R. Ben Yaou and G. Bonne (2009). "Mutations of the FHL1 gene cause Emery-Dreifuss muscular dystrophy." Am J Hum Genet 85(3): 338-53.

13. Windpassinger, C., B. Schoser, V. Straub, S. Hochmeister, A. Noor, B. Lohberger, N. Farra, E. Petek, T. Schwarzbraun, L. Ofner, W. N. Loscher, K. Wagner, H. Lochmuller, J. B. Vincent and S. Quasthoff (2008). "An X-linked myopathy with postural muscle atrophy and generalized hypertrophy, termed XMPMA, is caused by mutations in FHL1." Am J Hum Genet 82(1): 88-99.

14. Selcen, D., M. B. Bromberg, S. S. Chin and A. G. Engel (2011). "Reducing bodies and myofibrillar myopathy features in FHL1 muscular dystrophy." Neurology 77(22): 1951-9.

15. Schreckenbach, T., W. Henn, W. Kress, A. Roos, M. Maschke, W. Feiden, U. Dillmann, J. B. Schulz, J. Weis and K. G. Claeys (2012). "Novel FHL1 mutation in a family with reducing body myopathy." Muscle Nerve.

16. Chen, D. H., W. H. Raskind, W. W. Parson, J. A. Sonnen, T. Vu, Y. Zheng, M. Matsushita, J. Wolff, H. Lipe and T. D. Bird (2010). "A novel mutation in FHL1 in a family with X-linked scapuloperoneal myopathy: phenotypic spectrum and structural study of FHL1 mutations." J Neurol Sci 296(1-2): 22-9.

17. Shalaby, S., Y. K. Hayashi, I. Nonaka, S. Noguchi and I. Nishino (2009). "Novel FHL1 mutations in fatal and benign reducing body myopathy." Neurology 72(4): 375-6.

18. Komagamine, T., M. Kawai, N. Kokubun, S. Miyatake, K. Ogata, Y. K. Hayashi, I. Nishino and K. Hirata (2012). "Selective muscle involvement in a family affected by a second LIM domain mutation of fhl1: an imaging study using computed tomography." J Neurol Sci 318(1-2): 163-7.

19. Schessl, J., S. Feldkirchner, C. Kubny and B. Schoser (2011). "Reducing body myopathy and other FHL1-related muscular disorders." Semin Pediatr Neurol 18(4): 257-63.

20. Waddell, L. B., J. Tran, X. F. Zheng, C. G. Bonnemann, Y. Hu, F. J. Evesson, M. Lek, S. Arbuckle, M. X. Wang, R. L. Smith, K. N. North and N. F. Clarke (2011). "A

study of FHL1, BAG3, MATR3, PTRF and TCAP in Australian muscular dystrophy patients." Neuromuscul Disord 21(11): 776-81.

21. Binder, J. S., F. Weidemann, B. Schoser, M. Niemann, W. Machann, M. Beer, G. Plank, A. Schmidt, E. Bisping, I. Poparic, I. Lafer, T. Stojakovic, S. Quasthoff, J. B. Vincent, R. Rienmueller, M. R. Speicher, A. Berghold, B. Pieske and C. Windpassinger (2012). "Spongious Hypertrophic Cardiomyopathy in Patients With Mutations in the Four-and-a-Half LIM Domain 1 Gene." Circ Cardiovasc Genet 5(5): 490-502.

22. Knoblauch, H., C. Geier, S. Adams, B. Budde, A. Rudolph, U. Zacharias, J. Schulz-Menger, A. Spuler, R. B. Yaou, P. Nurnberg, T. Voit, G. Bonne and S. Spuler (2010). "Contractures and hypertrophic cardiomyopathy in a novel FHL1 mutation." Ann Neurol 67(1): 136-40.

23. Konno, T., M. Shimizu, H. Ino, N. Fujino, K. Uchiyama, T. Mabuchi, K. Sakata, T. Kaneda, T. Fujita, E. Masuta and H. Mabuchi (2006). "A novel mutation in the cardiac myosin-binding protein C gene is responsible for hypertrophic cardiomyopathy with severe ventricular hypertrophy and sudden death." Clin Sci (Lond) 110(1): 125-31.

24. Matsumoto, M., K. Kawakami, H. Enokida, K. Toki, R. Matsuda, T. Chiyomaru, K. Nishiyama, K. Kawahara, N. Seki and M. Nakagawa (2010). "CpG hypermethylation of human four-and-a-half LIM domains 1 contributes to migration and invasion activity of human bladder cancer." Int J Mol Med 26(2): 241-7.

25. Teekakirikul, P., R.F. Padera, J.G. Seidman and C.E. Seidman. "Hypertrophic cardiomyopathy: Translating cellular cross talk into therapeutics." J Cell Biol. 2012 Oct 29;199(3):417-21.

26. Debold EP, Schmitt JP, Patlak JB, Beck SE, Moore JR, Seidman JG, Seidman C, Warshaw DM (2007). Hypertrophic and dilated cardiomyopathy mutations differentially affect the molecular force generation of mouse alpha-cardiac myosin in the laser trap assay. Am J Physiol Heart Circ Physiol. 293(1):H284-91.

APPENDIX 1: Supplemental text and figures

Supplemental text:

5' RNA-seq

Below we describe the steps to construct RNA-seq libraries that yield increased proportions of 5' end sequences and computational approaches to identify genes with altered transcriptional initiation. We also provide an approach for obtaining data images from the UCSC browser that can be sorted for analyses (Appendix 3).

Library preparation:

The steps for construction of libraries from RNA with increased proportions of fragments from 5' ends of genes (Supplemental Fig. 5.1) are adapted from published methodologies [1, 2] and notably do not include normalization or fragmentation.

Step 1: Isolate RNA. Total RNA can be isolated with Trizol or column-based methods. Trizol has some recognized advantages such as it can isolate RNA molecules of a large size spectrum and results in a high yield of RNA amount.

Step 2: Assess RNA quality. Assay an aliquot of the transcriptome on a gel and assess quality of 18S and 28S using Bioanalyzer (Agilent) or the Tapestation (Agilent). Broken RNAs will have artificial 5' ends as RNA hydrolysis may also occur at defined places. In addition, broken RNA will result in loss of the real 5' ends since selection is done from the 3' end (Step 3). Libraries constructed from broken RNA will have diminished enrichment of reads at 5' ends.

Step 3: Perform two-rounds of selecting RNAs possessing poly-A tails by annealing them to poly-T oligos bound to beads (Invitrogen) so as to remove rRNA. After annealing the RNAs to the poly-T-beads, wash unbound material, and then elute the polyA species by melting the annealed structures at high temperature. It is crucial that temperatures do not exceed 72^{0}C to avoid extensive RNA hydrolysis. After 2 rounds of polyA selection, the rRNA should be ~5% of total RNA.

Step 4: Reverse transcribe (RT) mRNAs using random hexamers (Invitrogen). Be sure reaction is complete to reach 5' ends. Random primers can be of any length. Large random primers (10-20bp) can exhibit higher specificities during annealing. This can result in preference to transcribing low expressing species, as the primer pool annealing to high expressing species would be depleted. While offering such an advantage, using long random primers may skew the measurement of the distribution of transcripts stronger than shorter random primers.

Step 5: Perform double-stranded DNA (dsDNA) synthesis by adding Pol I and RNA-se H to the reacted products from Step 4 and incubate overnight. It is vital that dsDNA synthesis be performed as completely as possible, as inefficiencies will result in loss of 5' ends. Priming at the end of the fragment is needed to preserve the information at the 5' end, from either hydrolyzed RNA fragments resulting from RNA-se H digestion or leftover random hexamer from Step 4. Incubating the reaction overnight may aid in completion of the reaction.The resulting dsDNA is a polar molecule with directional information. (Discussed further at Step 9.)

Step 6: Perform end-repair. This step typically uses addition of enzymes such as polymerases and exonucleases serving to obtain dsDNA fragments with no overhangs.

Step 7: Perform Adenosine addition at 3' ends of both DNA strands and then ligate appropriate next-generation sequencing adapters. The adenosine overhangs at 3' ends prevent self-ligation of dsDNA fragments during adapter ligation. An excess of adapter dimers may further reduce self-ligation events.

Step 8: Perform size selection by fractionating the DNA library in an agarose gel to excise a size of about ~150-400bp. This can be achieved by running an electrophoresis chamber manual or by using specialized equipment such as Pippin Prep (Sage science).

Size selection allows for appropriate size molecules for next-generation sequencing and enriches fragments that resulted from annealing of random hexamers near the 5' ends of transcripts. Failure to optimally execute this step can result in substantial loss of the material. Increase in yield can be achieved by making thinner gels or with the Pippin Prep.

Step 9: Perform a uniform amplification of the library with PCR. The number of cycles of the PCR should be carefully selected so that the reaction does not reach saturation, but yields sufficient material for next-generation sequencers. To achieve uniform amplification of all fragments, stop the reaction at the stage of exponential amplification. A qPCR amplification reaction mirroring final amplification conditions can be used to map the amplification of the library. Cycle numbers at the upper range of the exponential phase can be selected for the final amplification. Multiple parallel reactions can be used to provide the necessary amount for sequencing. Using the majority of the library material in this process is important to maximize the complexity of the library. Note that lower than optimal cycles will lead to insufficient material and higher than optimal cycles will skew the data and introduce sequencing errors (such as when the abundance of dNTP becomes limiting).

A library constructed and sequenced with this approach contains directional information owing to the polarity of the cDNA molecule. Sequences will have one of two configurations: 1) when the fragment is sequenced from the end that corresponds to the upstream part of the reference transcript, the resulting sequence preserves the same strand and direction of the reference transcript. 2) Conversely, when the fragment is sequenced from the end that corresponds to the 3' end of the reference transcript, the resulting sequence is in reverse/complement relationship with the reference transcript. This information provides a signature of 5' end sites (Supplemental Fig. 5.2).

Computational analysis:

Gene/transcript definitions: Gene and transcript definitions are retrieved from the UCSC genome browser using the Table browser. Thus, UCSC transcript definitions that are assigned to RefSeq names are used and these definitions are further processed to obtain annotation tables. For example, to evaluate overall gene expression, all exons of isoforms for each gene are merged. A second annotation file is made that maintains the 5' end locations of isoforms of genes to process 5' end differences.

Using these UCSC transcript definitions, we estimate that 8,000 of the 30,000 annotated mouse genes have at least one additional annotated 5' end. The "count_5pr_isoforms.pl" is provided (Supplemental software package).

Expression analysis: Normalize libraries to control for differences in sequencing depth. For similar samples comparing total RNA, housekeeping genes are good reference points. For dissimilar samples, a TATA box binding protein gene Tbp can provide the reference gene.

Reads at 5' end regions are expected to correspond to actual transcript levels (1 read:1 transcript relationship). The total reads derived from other regions of transcripts have a dependency on transcript length and possibly RNA structure.

A p-value is computed using Bayesian statistics [3] to compare the read numbers for a gene from two samples. This takes into account the number of reads as evidence and their relative proportion in the assessed population. When a library is well constructed, this p-value is an accurate reflection of the comparison of the RNA species from the original sample.

Comparison of read-depth distributions at 5' end regions: Reads are first aligned to the genome and transcriptome. The read-depths at every base-pair position at 5' regions of genes are retrieved and compared between samples one gene at a time, after normalization to the respective total gene expression. This comparison calculates the fold of the normalized read-depths and a p-value using the Bayesian statistic. An assessment of a shift in read distribution is made when the calculated fold exceeds a parametrically defined fold-threshold (3-fold is used), and the p-value is lower compared to a p-value threshold (0.01 is used). Thus, changes in start-site usage is defined by quantifying the instances at 5' ends at which there is a change in distribution of reads on the gene locus. The sum of these instances is used as a score, it is tabulated for all genes, and it is interpreted as a positive signal that can recognize shifts in distribution in either direction. Genes with identified changes in distributions are queried visually on the UCSC browser or Integrative Genome Viewer (IGV) from highest to lowest score to visually identify the change in read-depth profiles and to assess whether these changes correspond to changes in 5' end usage. Evidence from UCSC browser sources (ESTs, mRNA annotations, conservation of sequence, and aligned locations of transcripts identified in other species) can be used to support the recognition of unannotated 5' ends.

The Supplemental software package provides a strategy to retrieve data images from the UCSC browser, and for sorting and organization of image data. Obtaining the coordinates of 5' regions using the UCSC browser allows retrieval of information about aligned reads in an automatic approach for multiple genes and for assessment of significance.

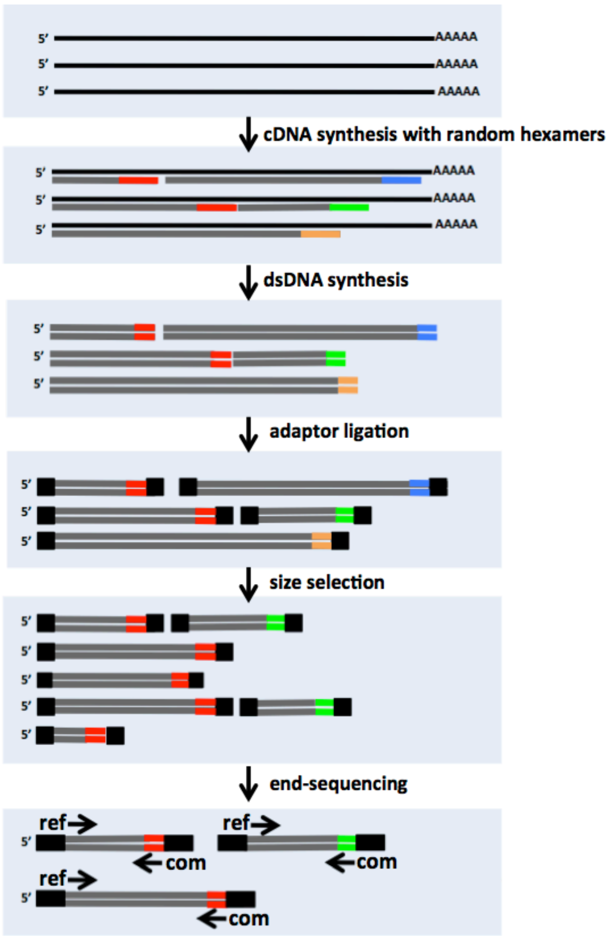
Data on 5’ end usage is best when there is sufficient sequencing, when the 5’ end uses a unique exon, and when the reads are distributed on the gene locus with a strong 5’ end distribution. When applied genome-wide, this can provide a new dimension of information for transcriptional data.

Additional references:

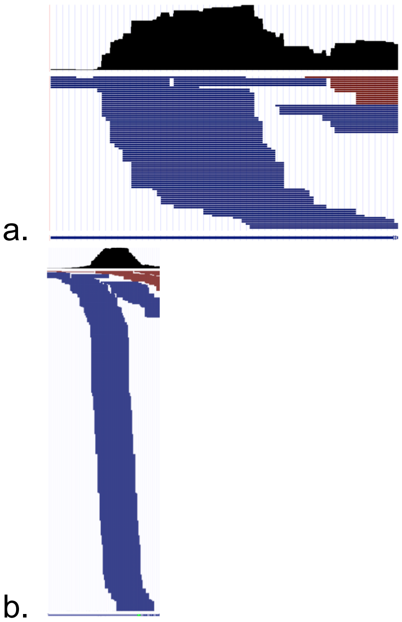
1. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics* **10**, 57-63 (2009).

2. Christodoulou, D.C., Gorham, J.M., Herman, D.S. & Seidman, J.G. Construction of normalized RNA-seq libraries for next-generation sequencing using the crab duplex-specific nuclease. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]* **Chapter 4**, Unit4 12 (2011).

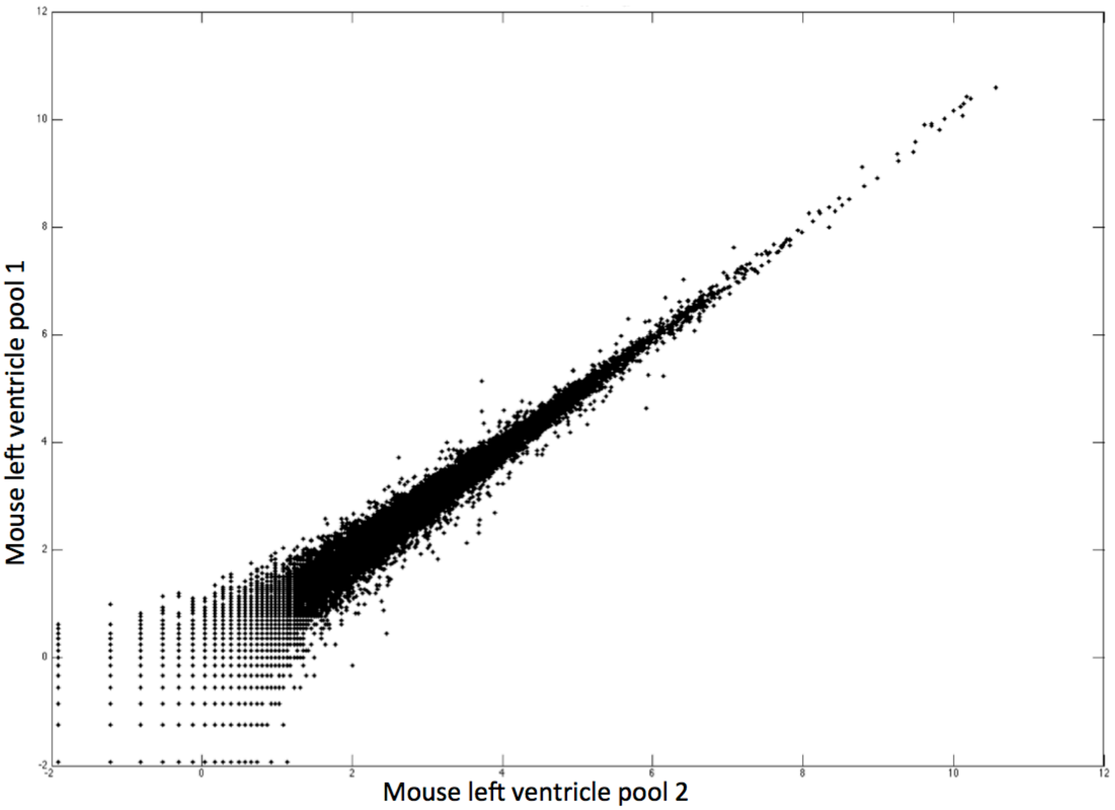
3. Audic, S. & Claverie, J.M. The significance of digital gene expression profiles. *Genome research* **7**, 986-995 (1997).



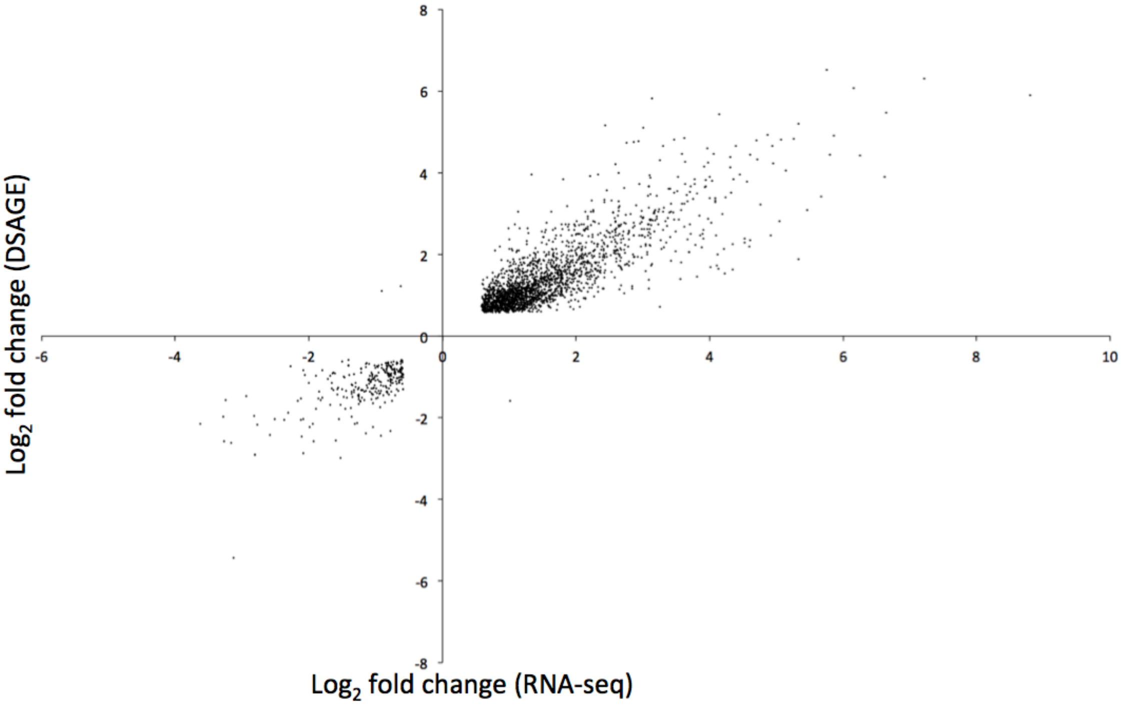
Supplemental Figure 5.1 cDNA library construction steps to enhance 5' sequencing. Random priming near 5' ends of transcripts (red) results in the synthesis of smaller fragments that are selected for sequencing.



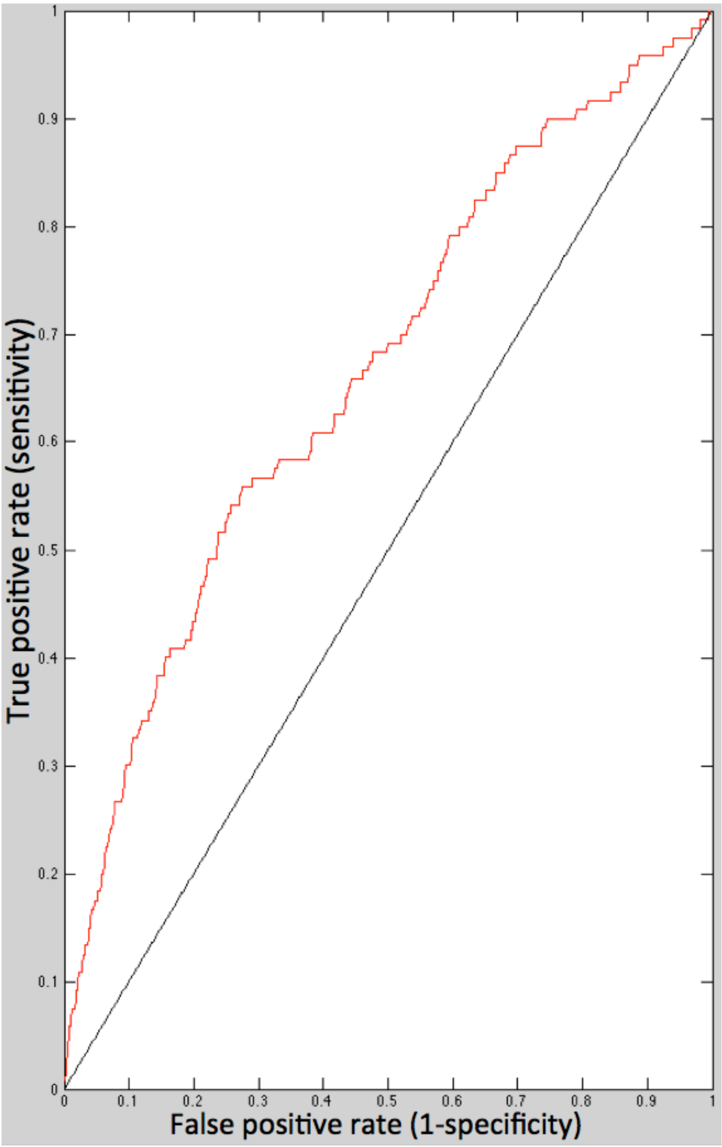
Supplemental Figure 5.2 Read depth and reads of first exons of Fhl1 of (a) *b*Fhl1 and (b) *s*Fhl1 isoforms. Reads from the transcript (+, blue) strand precede reads on complement (-, red) strand, and illustrate directional information at 5’ regions.



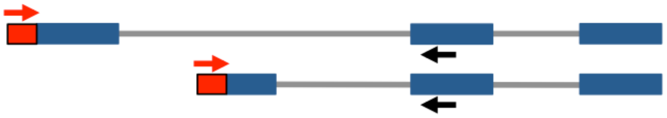
Supplemental Figure 5.3 High technical replication of RNAseq libraries. Log values (base 10) of normalized expression values from two biological replicate samples of wildtype heart tissue are shown (R=0.994).



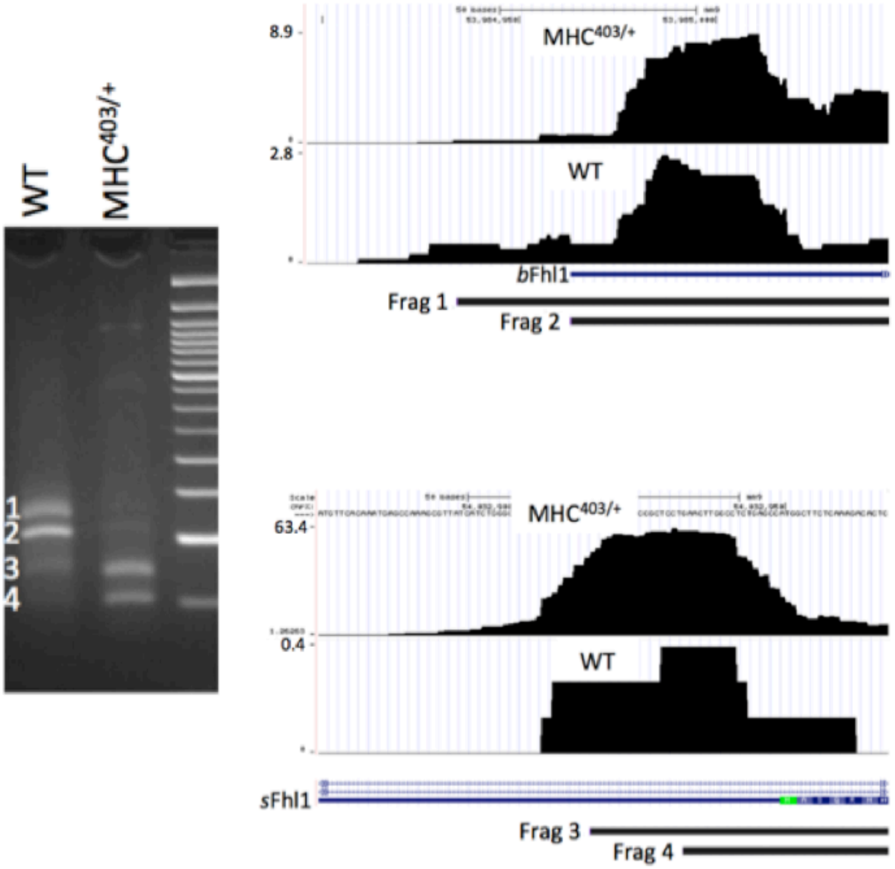
Supplemental Figure 5.4 Comparison of RNA-seq with sequencing analysis of gene expression. The fold change (base 2) of gene expression that show agreement (first and third quadrants) or disagreement (second and fourth quadrants) between the two methods are plotted for all genes (a minimum of 10 reads was required). Agreement was defined when both methods concurred that expression was increased or decreased by 1.5 fold, p-value 0.001. Disagreement was defined when gene expression was discordant by 1.5 fold, p-value 0.001. There is strong correlation between these methods R=0.9.



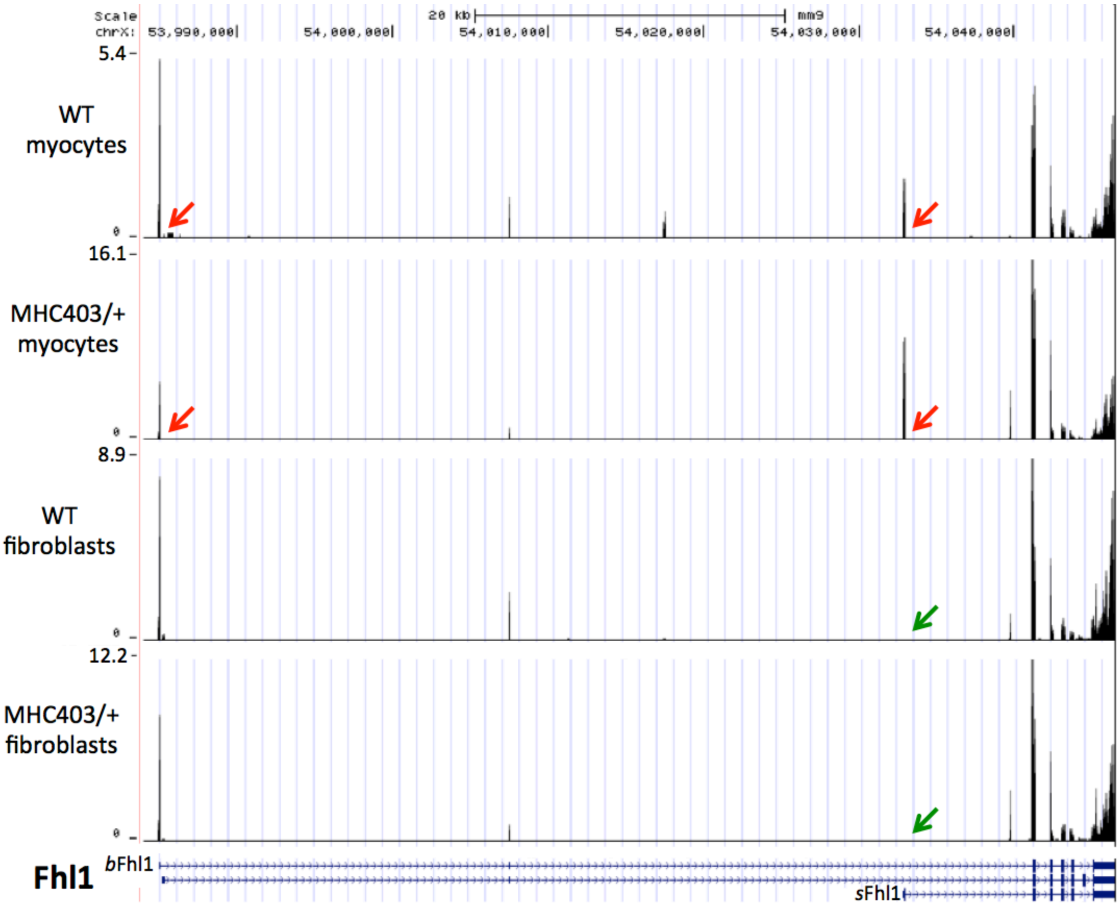
Supplemental Figure 5.5 Receiver operating characteristic (ROC) curve illustrating the performance of the altered 5' start-site algorithm for genes with RNA-seq score of 20 or higher. The area under the curve (AUC) corresponds to a p-value of 1.07E-11.



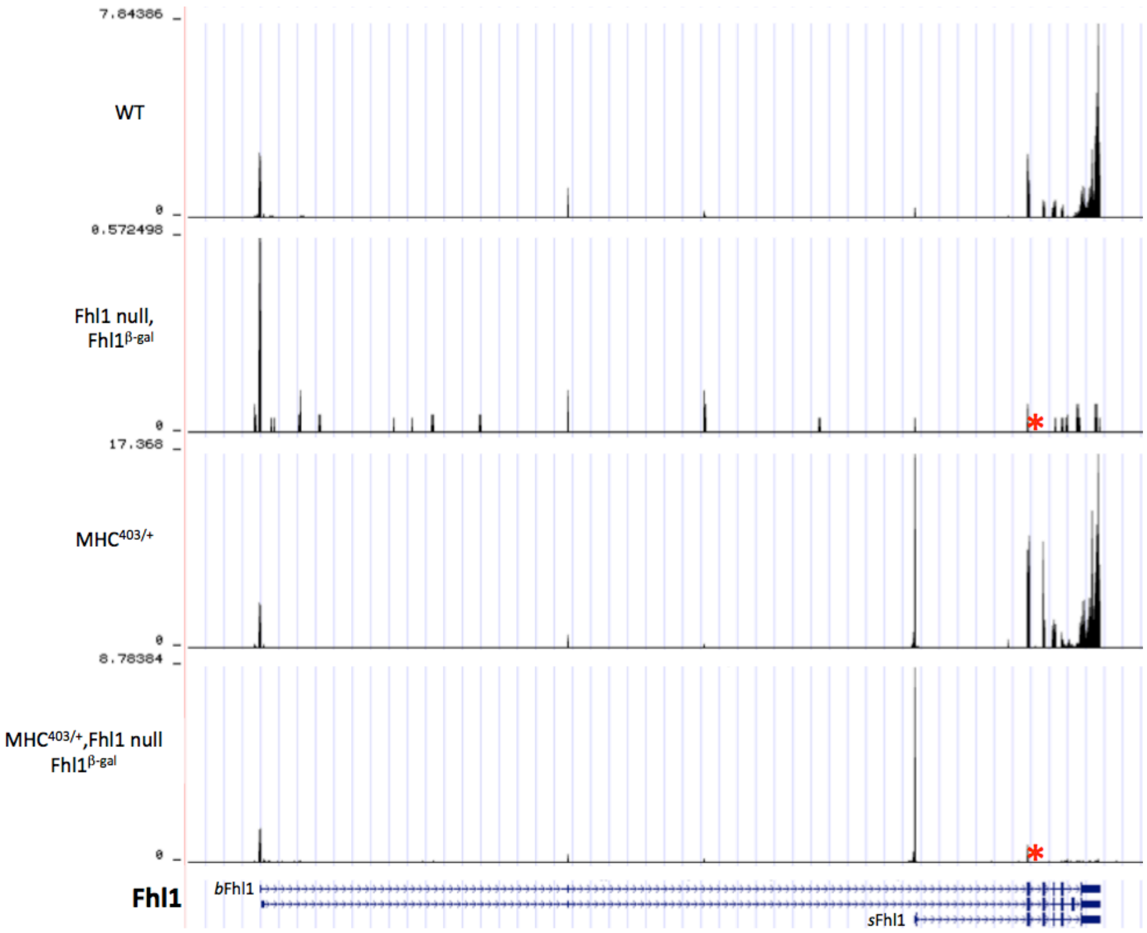
Supplemental Figure 5.6 Primer design scheme for semi-quantitative 5'RACE. In the same reaction, the relative concentration of queried isoforms can be assessed. Red arrows indicate primers that correspond to the ligated adapter at the 5' end (supplied by company). Black arrows indicate primers designed to correspond to shared sequences in different isoforms.



Supplemental Figure 5.7 5'RACE validates Fhl1 5' end changes using primers that correspond to a common downstream exon of the isoforms (Methods). Fragments (numbered) were sequenced and aligned to the genome with Blat (UCSC browser) and shown with RNA-seq data. Alignment of these products corresponds to RNA-seq reads. The same total RNA was used for 5'RACE and RNA-seq.



Supplemental Figure 5.8 Transcriptional profiles of isolated myocytes and non-myocytes/fibroblasts were assessed with RNA-seq. Transcriptional switch of Fhl1 occurs in myocytes (red arrows) while isoform 2 is not expressed in non-myocytes (green arrows).



Supplemental Figure 5.9 Fhl1 expression in left ventricles tissue derived from wildtype, Fhl1$^{LacZ/LacZ}$ (Fhl1-null) , MHC$^{403/+}$, MHC$^{403/+}$/ Fhl1$^{LacZ/LacZ}$. The location of the inserted LacZ sequences is indicated (*). Note that no encoding Fhl1 exon sequences are absent while 5' regulatory elements are expressed in Fhl1-null mice. The LacZ sequences were detected by manual sequence assembly (not shown) and encode the protein in-frame.

APPENDIX 2: Published contributions[4]

My main contribution for the work summarized below was the optimization of the RNA-seq steps (as shown in Chapter 5) and the development and application of an automatic analysis workflow. For the Titin paper, I also performed verification of a Titin splice isoform that was originally visualized from RNA-seq data. For the micro-RNA study, I provided an initial analysis of the sequence data. I am very grateful to have had the opportunity to work with these researchers and to have been able to contribute to this work.

Stable gene targeting in human cells using single-strand oligonucleotides with modified bases. Rios X, Briggs AW, Christodoulou D, Gorham JM, Seidman JG, Church GM. PLoS One. 2012;7(5):e36697.

Abstract

Recent advances allow multiplexed genome engineering in E. coli, employing easily designed oligonucleotides to edit multiple loci simultaneously. A similar technology in human cells would greatly expedite functional genomics, both by enhancing our ability to test how individual variants such as single nucleotide polymorphisms (SNPs) are related to specific phenotypes, and potentially allowing simultaneous mutation of multiple loci. However, oligo-mediated targeting of human cells is currently limited by low targeting efficiencies and low survival of modified cells. Using a HeLa-based EGFP-rescue reporter system we show that use of modified base analogs can increase targeting efficiency, in part by avoiding the mismatch repair machinery. We investigate the effects of oligonucleotide toxicity and find a strong correlation between the number of phosphorothioate bonds and toxicity. Stably EGFP-corrected cells were generated at a frequency of ~0.05% with an optimized oligonucleotide design combining modified bases and reduced number of phosphorothioate bonds. We provide evidence from comparative RNA-seq analysis suggesting cellular immunity induced by the oligonucleotides might contribute to the low viability of oligo-corrected cells. Further optimization of this method should allow rapid and scalable genome engineering in human cells.

[4] Information obtained from NCBI

Truncations of titin causing dilated cardiomyopathy. Herman DS, Lam L, Taylor MR, Wang L, Teekakirikul P, Christodoulou D, Conner L, DePalma SR, McDonough B, Sparks E, Teodorescu DL, Cirino AL, Banner NR, Pennell DJ, Graw S, Merlo M, Di Lenarda A, Sinagra G, Bos JM, Ackerman MJ, Mitchell RN, Murry CE, Lakdawala NK, Ho CY, Barton PJ, Cook SA, Mestroni L, Seidman JG, Seidman CE.vN Engl J Med. 2012 Feb 16;366(7):619-28.

Abstract:
BACKGROUND:
Dilated cardiomyopathy and hypertrophic cardiomyopathy arise from mutations in many genes. TTN, the gene encoding the sarcomere protein titin, has been insufficiently analyzed for cardiomyopathy mutations because of its enormous size.
METHODS:
We analyzed TTN in 312 subjects with dilated cardiomyopathy, 231 subjects with hypertrophic cardiomyopathy, and 249 controls by using next-generation or dideoxy sequencing. We evaluated deleterious variants for cosegregation in families and assessed clinical characteristics.
RESULTS:
We identified 72 unique mutations (25 nonsense, 23 frameshift, 23 splicing, and 1 large tandem insertion) that altered full-length titin. Among subjects studied by means of next-generation sequencing, the frequency of TTN mutations was significantly higher among subjects with dilated cardiomyopathy (54 of 203 [27%]) than among subjects with hypertrophic cardiomyopathy (3 of 231 [1%], P=3×10(-16)) or controls (7 of 249 [3%], P=9×10(-14)). TTN mutations cosegregated with dilated cardiomyopathy in families (combined lod score, 11.1) with high (>95%) observed penetrance after the age of 40 years. Mutations associated with dilated cardiomyopathy were overrepresented in the titin A-band but were absent from the Z-disk and M-band regions of titin (P≤0.01 for all comparisons). Overall, the rates of cardiac outcomes were similar in subjects with and those without TTN mutations, but adverse events occurred earlier in male mutation carriers than in female carriers (P=4×10(-5)).
CONCLUSIONS:
TTN truncating mutations are a common cause of dilated cardiomyopathy, occurring in approximately 25% of familial cases of idiopathic dilated cardiomyopathy and in 18% of sporadic cases. Incorporation of sequencing approaches that detect TTN truncations into genetic testing for dilated cardiomyopathy should substantially increase test sensitivity, thereby allowing earlier diagnosis and therapeutic intervention for many patients with dilated cardiomyopathy. Defining the functional effects of TTN truncating mutations should improve our understanding of the pathophysiology of dilated cardiomyopathy. (Funded by the Howard Hughes Medical Institute and others.).

Epigenetic repression of cardiac progenitor gene expression by Ezh2 is required for postnatal cardiac homeostasis. Delgado-Olguín P, Huang Y, Li X, Christodoulou D, Seidman CE, Seidman JG, Tarakhovsky A, Bruneau BG. Nat Genet. 2012 Jan 22;44(3):343-7.

Abstract

Adult-onset diseases can be associated with in utero events, but mechanisms for this remain unknown(1,2). The Polycomb histone methyltransferase Ezh2 stabilizes transcription by depositing repressive marks during development that persist into adulthood(3-9), but its function in postnatal organ homeostasis is unknown. We show that Ezh2 stabilizes cardiac gene expression and prevents cardiac pathology by repressing the homeodomain transcription factor gene Six1, which functions in cardiac progenitor cells but is stably silenced upon cardiac differentiation. Deletion of Ezh2 in cardiac progenitors caused postnatal myocardial pathology and destabilized cardiac gene expression with activation of Six1-dependent skeletal muscle genes. Six1 induced cardiomyocyte hypertrophy and skeletal muscle gene expression. Furthermore, genetically reducing Six1 levels rescued the pathology of Ezh2-deficient hearts. Thus, Ezh2-mediated repression of Six1 in differentiating cardiac progenitors is essential for stable gene expression and homeostasis in the postnatal heart. Our results suggest that epigenetic dysregulation in embryonic progenitor cells is a predisposing factor for adult disease and dysregulated stress responses.

Polycomb repressive complex 2 regulates normal development of the mouse heart. He A, Ma Q, Cao J, von Gise A, Zhou P, Xie H, Zhang B, Hsing M, Christodoulou DC, Cahan P, Daley GQ, Kong SW, Orkin SH, Seidman CE, Seidman JG, Pu WT. Circ Res. 2012 Feb 3;110(3):406-15.

Abstract

RATIONALE:
Epigenetic marks are crucial for organogenesis, but their role in heart development is poorly understood. Polycomb repressive complex 2 (PRC2) trimethylates histone H3 at lysine 27, which establishes H3K27me3 repressive epigenetic marks that promote tissue-specific differentiation by silencing ectopic gene programs.
OBJECTIVE:
We studied the function of PRC2 in murine heart development using a tissue-restricted conditional inactivation strategy.
METHODS AND RESULTS:
Inactivation of the PRC2 subunit Ezh2 by Nkx2-5(Cre) (Ezh2(NK)) caused lethal congenital heart malformations, namely, compact myocardial hypoplasia, hypertrabeculation, and ventricular septal defect. Candidate and genome-wide RNA expression profiling and chromatin immunoprecipitation analyses of Ezh2(NK) heart identified genes directly repressed by EZH2. Among these were the potent cell cycle inhibitors Ink4a/b (inhibitors of cyclin-dependent kinase 4 A and B), the upregulation of which was associated with decreased cardiomyocyte proliferation in Ezh2(NK). EZH2-repressed genes were enriched for transcriptional regulators of noncardiomyocyte expression programs such as Pax6, Isl1, and Six1. EZH2 was also required for proper spatiotemporal regulation of cardiac gene expression, because Hcn4, Mlc2a, and Bmp10 were inappropriately upregulated in ventricular RNA. PRC2 was also required later in heart development, as indicated by cardiomyocyte-restricted TNT-Cre inactivation of the PRC2 subunit Eed. However, Ezh2 inactivation by TNT-Cre did not cause an overt phenotype, likely because of functional redundancy with Ezh1. Thus, early Ezh2 inactivation by Nk2-5(Cre) caused later disruption of cardiomyocyte gene expression and heart development.
CONCLUSIONS:
Our study reveals a previously undescribed role of EZH2 in regulating heart formation and shows that perturbation of the epigenetic landscape early in cardiogenesis has sustained disruptive effects at later developmental stages.

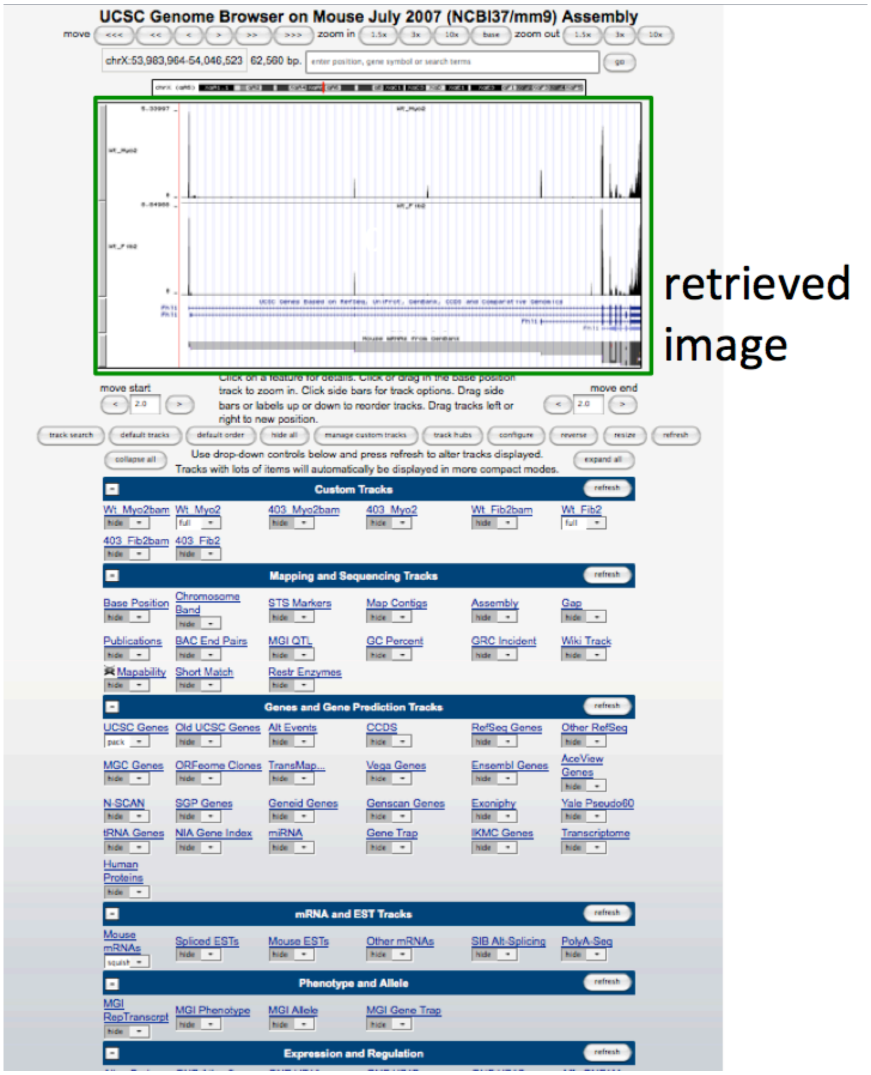
Barcoding bias in high-throughput multiplex sequencing of miRNA. Alon S, Vigneault F, Eminaga S, Christodoulou DC, Seidman JG, Church GM, Eisenberg E. Genome Res. 2011 Sep;21(9):1506-11.

Abstract

Second-generation sequencing is gradually becoming the method of choice for miRNA detection and expression profiling. Given the relatively small number of miRNAs and improvements in DNA sequencing technology, studying miRNA expression profiles of multiple samples in a single flow cell lane becomes feasible. Multiplexing strategies require marking each miRNA library with a DNA barcode. Here we report that barcodes introduced through adapter ligation confer significant bias on miRNA expression profiles. This bias is much higher than the expected Poisson noise and masks significant expression differences between miRNA libraries. This bias can be eliminated by adding barcodes during PCR amplification of libraries. The accuracy of miRNA expression measurement in multiplexed experiments becomes a function of sample number.

APPENDIX 3: Automatic download of data images from the UCSC browser

This allows collection of data images using the UCSC genome browser platform which can be then be imported automatically and in the desired sorted order on Microsoft Powerpoint (Detailed instructions in README.txt file, Supplemental software package, Chapter 5: http://seidman.med.harvard.edu/fgs/software/RNAseq-5pr-package.tar.gz)



UCSC Genome Browser on Mouse July 2007 (NCBI37/mm9) Assembly

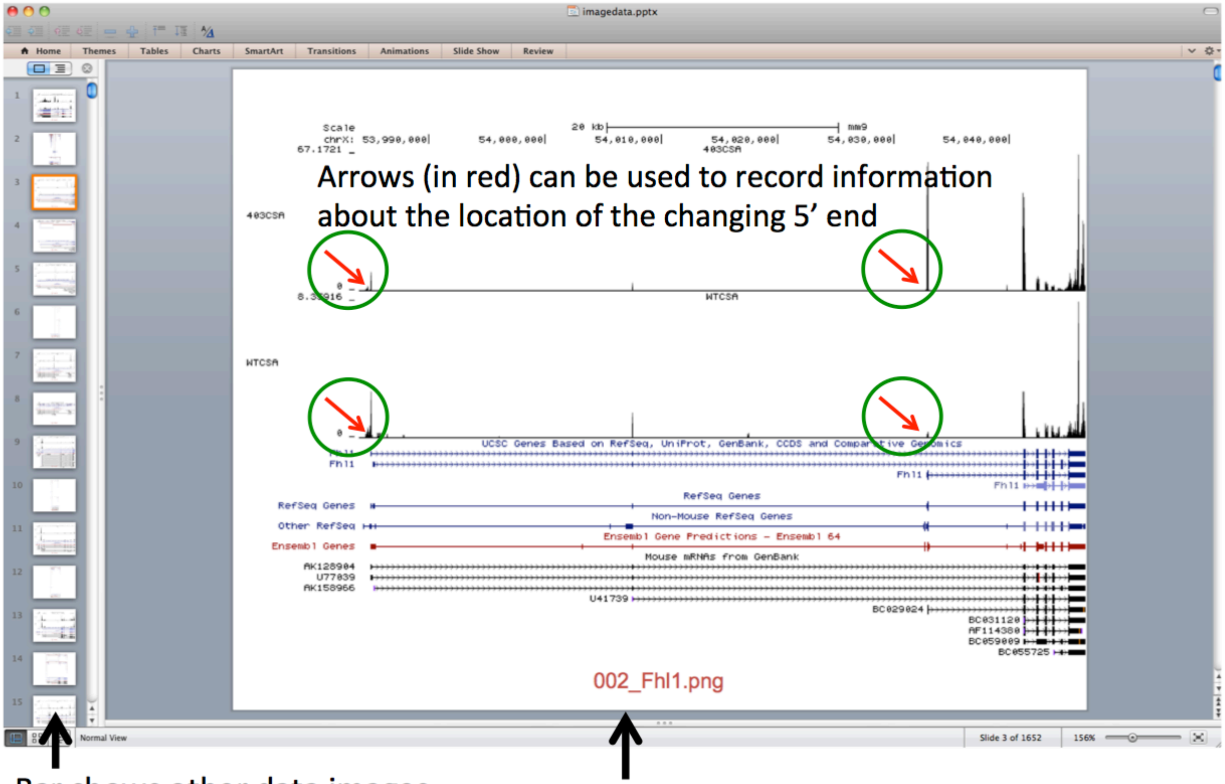
chrX:53,983,964-54,046,523 62,560 bp.

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

Use drop-down controls below and press refresh to alter tracks displayed. Tracks with lots of items will automatically be displayed in more compact modes.

retrieved image

Figure Appx. 3.1. Data browse mode using the UCSC genome browser (http://genome.ucsc.edu). The retrieved data image is shown in a green box.



Arrows (in red) can be used to record information about the location of the changing 5' end

002_Fhl1.png

Bar shows other data images that were simultaneously loaded (~1600 are loaded here)

File-name which corresponds to its sorted position and gene name

Figure Appx. 3.2. Shown is image data imported on Microsoft powerpoint after the images were downloaded.

Attributions: Danos Christodoulou and Steve DePalma co-developed this approach.