Genomics 105 (2015) 83–89

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

RNA-Seq identifies novel myocardial gene expression signatures of heart failure

Yichuan Liu [a,*], Michael Morley [b], Jeffrey Brandimarto [b], Sridhar Hannenhalli [c], Yu Hu [a], Euan A. Ashley [d], W.H. Wilson Tang [e], Christine S. Moravec [e], Kenneth B. Margulies [b], Thomas P. Cappola [b], Mingyao Li [a,*], for the MAGNet consortium

[a] Department of Biostatistics and Epidemiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, United States
[b] Cardiovascular Institute, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, United States
[c] Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, United States
[d] Stanford Center for Inherited Cardiovascular Disease, Stanford University School of Medicine, Stanford, CA 94305, United States
[e] Department of Cardiovascular Medicine, Cleveland Clinic, Cleveland, OH 44195, United States

ARTICLE INFO

Article history:
Received 6 August 2014
Accepted 9 December 2014
Available online 17 December 2014

Keywords:
RNA-seq
Heart Failure
Disease classification

ABSTRACT

Heart failure is a complex clinical syndrome and has become the most common reason for adult hospitalization in developed countries. Two subtypes of heart failure, ischemic heart disease (ISCH) and dilated cardiomyopathy (DCM), have been studied using microarray platforms. However, microarray has limited resolution. Here we applied RNA sequencing (RNA-Seq) to identify gene signatures for heart failure from six individuals, including three controls, one ISCH and two DCM patients. Using genes identified from this small RNA-Seq dataset, we were able to accurately classify heart failure status in a much larger set of 313 individuals. The identified genes significantly overlapped with genes identified via genome-wide association studies for cardiometabolic traits and the promoters of those genes were enriched for binding sites for transcriptions factors. Our results indicate that it is possible to use RNA-Seq to classify disease status for complex diseases such as heart failure using an extremely small training dataset.

1. Introduction

Heart failure, defined as the inability of the heart to pump sufficient blood to meet the body's demands, is a syndrome associated with high morbidity and mortality. An estimated five million Americans are diagnosed with heart failure every year, causing more than 250,000 deaths annually. Heart failure is a complex disease that involves multiple genetic and environmental factors. Two of the most common subtypes of heart failure include ischemic heart disease (ISCH), which is caused by reduced blood supply to heart muscle, and dilated cardiomyopathy (DCM) in which the heart becomes weakened and enlarged despite normal blood flow [1]. Although ISCH and DCM can lead to similar symptoms of heart failure, emerging evidence suggest that the two subtypes may produce different structural and/or functional phenotypes and may respond differently to therapy [1–3]. In addition, patients with ISCH generally have reduced survival compared to those with DCM [1,2].

Most human genomic studies in heart failure are limited by insufficient clinical samples from patients with advanced heart failure [1]. As such, researchers have used animal models in combination with functional genomics to study the molecular underpinnings of heart failure [4,5]. Attempts have also been made to link gene signatures in human blood with heart failure outcomes [6,7]. Recently, several studies have been published based on human myocardium. Tan et al. [8] showed that end-stage heart failure is associated with an increase in expression levels for genes encoding for matrix/cytoskeletal and proteolysis/stress proteins based on a comparison of eight hearts from patients with end-stage heart failure and seven non-failing controls. Kittleson et al. [2] used microarray with a machine learning approach to distinguish patients with histological evidence of ischemic injury from those without a history of myocardial infarction, revascularization, or coronary artery disease. Using a much larger dataset derived from 185 failing and 14 non-failing hearts, Margulies et al. [9] identified 3088 differentially expressed transcripts with only a small subset demonstrating improvements that was correlated to the favorable remodeling observed during mechanical circulatory support. Using this dataset, Hannenhalli et al. [3] explored transcription factors that are associated with heart failure.

* Corresponding authors.
E-mail addresses: yichuan.edwdard.liu@gmail.com (Y. Liu), mmorley@mail.med.upenn.edu (M. Morley), bjeff@mail.med.upenn.edu (J. Brandimarto), sridhar@umiacs.umd.edu (S. Hannenhalli), huyu1@mail.med.upenn.edu (Y. Hu), euan@stanford.edu (E.A. Ashley), TANGW@ccf.org (W.H.W. Tang), MORAVEC@ccf.org (C.S. Moravec), Kenneth.Margulies@uphs.upenn.edu (K.B. Margulies), thomas.cappola@uphs.upenn.edu (T.P. Cappola), mingyao@mail.med.upenn.edu (M. Li).

http://dx.doi.org/10.1016/j.ygeno.2014.12.002
0888-7543/

All of the aforementioned studies were based on microarrays. Although microarrays have been the predominant method for gene expression studies due to their ability to measure thousands of transcripts simultaneously, they are subject to biases in hybridization strength, and potential for cross-hybridization to probes with similar sequences. Additionally, they are unable to identify novel genes or novel splicing events because of their reliance on existing gene models. RNA sequencing (RNA-Seq) is a newer approach for transcriptome profiling [10–12]. It is the first sequencing-based method that allows an unbiased survey of the entire transcriptome in a high-throughput manner. Briefly, RNA-Seq involves fragmenting poly-A selected RNA molecules into small fragments and converting into a cDNA library with adaptors attached to cDNA fragments. The cDNA library is then sequenced to obtain short sequences, which are subsequently aligned to a reference genome and/or transcriptome or assembled de novo without the reference sequence. The expression level for a gene is determined by counting the number of reads that are mapped to it. With RNA-Seq data, transcripts spanning multiple exons can be directly observed. Moreover, RNA-Seq has a greater dynamic range than microarrays, which suffer from non-specific hybridization and saturation biases [13].

Motivated by the advantages of RNA-Seq technology for gene expression profiling, we sequenced the transcriptomes of six human individuals' left ventricle tissue to identify genes that are associated with heart failure. Our study includes one ISCH patient, two DCM patients and three individuals with non-failing hearts (NF). Based on these six individuals, we identified genes that were differentially expressed between ISCH and NF, DCM and NF, and ISCH and DCM. A remarkable finding of our study is that using genes identified from this small RNA-Seq dataset, we were able to classify a much larger set of 313 individuals with failing or non-failing hearts. Our results suggest that, with highly accurately measured gene expression levels using RNA-Seq, it is possible to classify disease status for complex diseases such as heart failure using an extremely small training dataset.

2. Materials and methods

2.1. Sample collection

Samples of cardiac tissue (n = 6 for RNA-Seq, n = 313 for microarrays) were acquired from subjects from the MAGNet consortium (http://www.med.upenn.edu/magnet/). The heart was perfused with cold cardioplegia prior to cardiectomy to arrest contraction and prevent ischemic damage. Left ventricular free-wall tissue was harvested and snap frozen with liquid nitrogen at the time of cardiac surgery from subjects with heart failure undergoing transplantation and from unused donor hearts. The cause of heart failure (ISCH or DCM) was determined by medical history and pathological examination of the explanted hearts. All the samples were stored in −80 °C freezer until analyses. This study was approved by the University of Pennsylvania Institutional Review Board and the Cleveland Clinic Institutional Review Board. All participants were 18 years or older and provided written informed consent.

2.2. RNA extraction, library preparation and sequencing

RNAs for six selected individuals were extracted using RNeasy Lipid Tissue total RNA mini kit (Qiagen, Valencia, CA). Extracted RNA samples underwent quality control (QC) assessment using the Agilent Bioanalyzer (Agilent, Santa Clara, CA) and all RNA samples submitted for sequencing had an RNA Integrity Number (RIN) > 6, with a minimum of 1 μg input RNA. Poly-A library preparation and RNA sequencing were performed at the Penn Genome Frontiers Institute's High-Throughput Sequencing Facility per standard protocols. Briefly, we generated first-strand cDNA using random hexamer-primed reverse transcription, followed by second-strand cDNA synthesis using RNase H and DNA polymerase, and ligation of sequencing adapters using the TruSeq RNA Sample Preparation Kit (Illumina, San Diego, CA). Fragments of ~350 bp were selected by gel electrophoresis, followed by 15 cycles of PCR amplification. The prepared libraries were then sequenced using Illumina'sHiSeq 2000 with four RNA-seq libraries per lane (2 × 101 bp paired-end reads).

2.3. Analysis of RNA-Seq data

The RNA-Seq data were aligned to the hg19 reference genome using Tophat with default options [14]. In order to eliminate mapping errors and reduce potential mapping ambiguity due to homologous sequences, several filtering steps were applied. Specifically, we required (1) the mapping quality score of each read is ≥30, (2) reads from the same pair were mapped to the same chromosome with expected orientations and the mapping distance between the read pair was <500,000 bp, and (3) each read was uniquely mapped to the genome. All subsequent analyses were based on filtered alignment files.

Transcripts were assembled using Cufflinks [15,16]. For each gene, we compared the expression levels between two individuals for each of the three categories, including ISCH vs. NF, DCM vs. NF, and ISCH vs. DCM. To test for differential expression, Cufflinks first computes the logarithm of the ratio of Fragments Per Kilobase of exon per Million fragments mapped (FPKMs) between the two subjects, and then uses delta method to estimate the variance of the log ratio. The test statistic is log ratio of the FPKMs divided by the standard deviation of the log ratio. It is possible to estimate the standard deviation based on a single subject because of the availability of multiple reads per subject. To ensure reliable expression estimates, we required the FPKM value to be greater than or equal to 3 for at least one of the two individuals under comparison [17]. A gene was considered differentially expressed if the FDR adjusted p-value was <0.05.

For differentially expressed genes, we carried out functional annotation analysis using DAVID [18,19]. Differentially expressed genes were used as input gene list, and all human genes that were expressed in the heart were used as the background. We looked for enrichment for genetic association with disease class, KEGG pathways, and biological processes in Gene Ontology (GO). Multiple testing was adjusted using Benjamini approach, and enrichment was declared if Benjamini adjusted p-value was less than 0.05.

To search for evidence of over representation of transcription factor binding sites in heart failure, we used a computational approach previously developed by Hannenhalli et al. [3]. First, a set of cardiac genes was determined from RNA-Seq data by selecting genes with FPKM > 3. Each cardiac gene was then mapped to its corresponding promoter region sequence, defined as the 5 kb of genomic sequence upstream from the transcription start site, based on the RefSeq annotation. Transcription factor binding sites were determined within these promoters with the TRANSFAC database [20] of vertebrate transcription factor binding sites, with a focus on promoter regions that show human-mouse evolutionary sequence conservation. We then determined which binding sites were statistically over represented among genes that showed altered expression in heart failure.

For differentially expressed genes, we further examined whether they were more likely to overlap with GWAS findings. Our analysis was based on all GWAS signals summarized in the NHGRI GWAS catalog (http://www.genome.gov/gwastudies). We only considered GWAS signals for cardiometabolic traits (Supplementary Table 6). The enrichment analysis was investigated using Fisher's exact test.

2.4. RNA preparation and processing of microarray data

RNAs for 313 subjects, including 95 ISCH patients, 82 DCM patients, and 136 individuals with normal hearts were hybridized with Affymetrix Human Exon ST1.1 arrays using the manufacturer's instructions. The resulting CEL files were normalized with the robust multiarray analysis using Bioconductor to generate transcript-level

intensity estimates [21]. To remove residual batch effect, expression values were further adjusted using ComBat [22,23], an empirical Bayes method that estimates parameters for location and scale adjustment of each batch for each gene independently. Probe sets were removed if the $\log_2$-transformed expression values were less than four on all arrays. This filtering yielded sets of genes present well above background levels in the human heart. For the remaining probe sets, their Affymetrix probe annotations were cross checked by mapping probe sequences to the hg19 reference genome. Only uniquely mapped probes with no mismatches were kept for subsequent analysis.

2.5. Classification of disease status using gene signatures identified from RNA-Seq

Our goal was to use those differentially expressed genes identified from RNA-Seq as feature vectors to classify disease status for the 313 individuals with microarray data. In order to classify the ISCH/NF (n = 231) individuals, we used genes that were differentially expressed in all pairwise comparisons (defined as globally differentially expressed) of ISCH vs. NF in RNA-Seq as the feature vector. Similarly, globally differentially expressed genes were used as the feature vectors to classify DCM/NF individuals (n = 218), and ISCH/DCM individuals (n = 177). After the feature vectors were determined, the K-means clustering algorithm implemented in R's "amap: Another Multidimensional" package was used to classify the individuals into two groups, and Pearson correlation distance metric was used in the clustering with a maximum of 50 iterations.

2.6. Data access

RNA-Seq and microarray data have been deposited in the Gene Expression Omnibus (GEO) database (accession number GSE57345).

3. Results

3.1. RNA-Seq data alignment

The RNA-Seq data were aligned and filtered as described in the Materials and methods section. We obtained a high mapping rate with 76–83% of reads mapped to the reference genome, and 66–71% were uniquely mapped, properly filtered, and used in subsequent analysis (Supplementary Table 1). All RNA-Seq samples passed FastQC's basic statistics test (Supplementary Fig. 1).

3.2. Analysis of differential expression using RNA-Seq

First, we compared the gene expression profiles of the ISCH and NF individuals. Our RNA-Seq experiment includes one ISCH patient and three individuals with non-failing hearts, yielding three possible pairwise comparisons for differential expression analysis. Using Cufflinks, we identified 492, 522 and 418 differentially expressed genes in the three pairs, respectively (Table 1; Supplementary Tables 2A–2C). Union of these gene lists gave 983 genes that were differentially expressed in at least one of the three pairs, among which 531 (54%) had higher expression levels in ISCH and 452 (46%) had higher expression levels in NF (Supplementary 3A). By intersecting differentially expressed genes across all three pairs, 70 genes were retained and we call these genes as globally differentially expressed and used them as feature vector for the K-means clustering of the 231 ISCH/NF individuals with microarray data (Supplementary Table 4A).

Next, we compared DCM and NF individuals. With two DCM and three NF individuals, there were six possible pairwise comparisons for differential expression analysis. Using Cufflinks, we identified 361, 393, 491, 482, 343 and 491 differentially expressed genes, respectively, for the six pairs (Table 1; Supplementary Tables 2D–2I). Union of these gene lists gave 1109 genes that were differentially expressed in at least one of the six pairs (Supplementary 3B). Among these genes, 844 (76%) had higher expression levels in DCM and 265 (24%) had higher expression levels in NF. By intersecting differentially expressed genes across all six pairs, we identified 12 genes that were globally differentially expressed (Supplementary 4B). These genes were used as a feature vector in the K-means clustering of the 218 DCM/NF individuals with microarray data.

We also compared the two subtypes of heart failure. Two possible combinations were considered based on one ISCH and two DCM individuals. We found 484 and 492 differentially expressed genes in the two pairs (Table 1; Supplementary Tables 2J–2K), respectively, yielding a total of 825 differentially expressed genes in at least one pair, including 476 (58%) with higher expression levels in ISCH and 349 (42%) with higher expression levels in DCM (Supplementary Table 3C). The interaction of the gene lists yielded 129 genes that were differentially expressed in both pairs and they were used as the feature vector in the K-means clustering of the 177 ISCH/DCM individuals with microarray data (Supplementary Table 4C).

3.3. Categories of differentially expressed genes

To investigate what categories of genes were differentially expressed, we carried out functional annotation analysis using DAVID. For each set of differential expression analysis, genes that were expressed (FPKM ≥ 3) in at least one individual under comparison were used as the background. These include 9919 genes for ISCH vs. NF comparison, 10,462 genes for DCM vs. NF comparison, and 10,190 for ISCH vs. DCM comparison.

3.3.1. ISCH vs. NF comparison

For genes that had higher expression levels in ISCH, they were enriched and only enriched for CARDIOVASCULAR (p-value = 0.0028) in disease class and ECM-receptor interaction pathway (p-value = 0.000152) in KEGG. For Gene Ontology (GO), these genes were

Table 1
Summary of differentially expressed (DE) genes for comparisons of ISCH vs. NF, DCM vs. NF, and ISCH vs. DCM.

Comparison	Pair	No. of DE genes	No. of overlapping DE genes	No. of union DE genes
ISCH vs. NF	234 vs. 1207	522	70	983
	234 vs. 1256	418		
	234 vs. D111	492		
DCM vs. NF	333 vs. 1207	491	12	1109
	333 vs. 1256	482		
	333 vs. D111	343		
	X2182 vs. 1207	361		
	X2182 vs. 1256	393		
	X2182 vs. D111	491		
ISCH vs. DCM	234 vs. 333	484	129	825
	234 vs. X2182	492		

significantly enriched for extracellular matrix formation processes (p-value $< 10^{-25}$) (Supplementary Figs. 2(A)–(C)). ECM-receptor interaction and their formation related processes had been shown to play critical roles in ischemic heart remodeling [24]. For genes that had higher expression levels in the NF individuals, they were enriched for CARDIOVASCULAR (p-value = 0.011) and RENAL (p-value = 0.0024) in disease class, but no significant enrichment was found in KEGG. For GO, these genes were significantly enriched for terms related to system development (p-value $= 8.68 \times 10^{-8}$) and organ development (p-value $= 9.34 \times 10^{-6}$) (Supplementary Fig. 2(D)).

3.3.2. DCM vs. NF comparison

For genes that had higher expression levels in DCM, they were enriched for CARDIOVASCULAR (p-value = 0.00022) in disease class. They were also enriched for focal adhesion pathway (p-value = 0.029) in KEGG. For GO, these genes were significantly enriched for terms related to plasma membrane (p-value $< 10^{-10}$) and extracellular region (p-value $< 10^{-6}$) (Supplementary Figs. 2(E)–(G)). For genes that had higher expression levels in NF, no significant enrichment was found in disease class and KEGG, but for GO, these genes were significantly enriched for terms related to extracellular matrix (Supplementary Fig. 2(H)).

3.3.3. ISCH vs. DCM comparison

Genes that had higher expression levels in ISCH were enriched for CARDIOVASCULAR (p-value = 0.0015) and IMMUNE (p-value = 0.0001) in disease class. In KEGG, only the ECM-receptor interaction pathway was significantly enriched (p-value $= 1.85 \times 10^{-6}$). For GO, terms related to extracellular matrix formation processes (p-value $< 10^{-28}$), response to external stimulus (p-value $= 2.06 \times 10^{-14}$) and inflammatory response (p-value $= 6.99 \times 10^{-12}$) were significantly enriched (Supplementary Figs. 2(I)–(K)). Extracellular matrix formation process was again found highly enriched, indicating that extracellular matrix related genes were not only differentially expressed in ISCH vs. NF, but also differentially expressed in subtypes of heart failure. No significant enrichment was found for genes that had higher expression levels in DCM.

3.4. Overrepresented transcription factor binding sites in heart failure

To investigate whether there is a discrete set of cardiac transcription factors potentially driving the observed gene expression changes in heart failure cases relative to controls, we examined whether certain transcription factor binding sites are over- or under-represented among those that showed altered gene expression patterns in heart failure using a computational approach developed by Hannenhalli et al. [3]. Specifically, we determined the enrichment of TRANSFAC motifs by counting the frequency with which a given binding site was present in the promoters of differentially expressed genes relative to the frequency in the reference set of genes (i.e., genes expressed in the heart). We performed analysis separately for genes that were up-regulated or down-regulated in heart failure cases.

For the comparison of ISCH vs. NF, the binding sites of NKX2-5, MAZ, and MZF1 were over-represented in down-regulated genes in all three pairs. Further examination of the gene expression of these transcription factors suggests that gene that encodes NKX2-5 was differentially expressed between the ISCH subject 234 and the NF subject 1256 (p-value = 0.0062), with subject 234 showing lower expression level than subject 1256, which is consistent with the fact that the target genes of NKX2-5 were down-regulated in ISCH. The expression level of subject 234 was also lower than the other two NF subjects, 1207 and D111, but the gene expression difference was not statistically significant (Supplementary Table 5). We did not find evidence of differential expression in genes that encode MAZ and MZF1. Several factors may affect the lack of differential expression in transcription factor genes. First, transcription factor genes are generally expressed at very low levels, and this affects the statistical power in detecting their differential expression. Second, transcription factors are often regulated at post-translational levels and are not expected to exhibit differential expression [25]. Third, a large fraction of the observed differences in target gene expression is likely due to genotype differences in *cis*-elements and not differential expression of the transcription factor regulators. Fourth, it is entirely possible that certain transcription factors can both activate certain genes and repress others, depending on the genomic context and also whether the binding site is polymorphic between cases and controls and the effect of the corresponding polymorphisms. We performed similar enrichment analysis for up-regulated genes in the ISCH vs. NF comparison and both up- and down-regulated genes for the DCM vs. NF comparison, but did not identify transcription factor motifs that were significantly enriched in all pairs.

3.5. Overlap of differentially expressed genes with GWAS loci for cardiometabolic traits

We queried our differentially expressed genes against published GWAS loci for cardiometabolic traits as derived from the NHGRI GWAS catalog (Supplementary Table 6). Compared to genes that were not differentially expressed, we found a statistically significant overlap with GWAS loci for cardiometabolic traits for the ISCH vs. NF comparison (p-value = 0.00029), and the DCM vs. NF comparison (p-value = 0.0012). The overlap was less significant for the ISCH vs. DCM comparison (p-value = 0.0093) (Table 2). Our results suggest that the identified differentially expressed genes might play a specific role for cardiometabolic disease.

3.6. RNA-Seq selected genes classify heart failure status in samples with microarray data

Genes that were globally differentially expressed were used as feature vectors to classify the 313 individuals with microarray data, which include 95 with ISCH, 82 with DCM, and 136 with NF. In the classification of the 231 ISCH/NF individuals, the K-means clustering algorithm using the RNA-Seq determined feature vector with 70 genes correctly classified 216 of the individuals, yielding a misclassification rate of 6.5% (Fig. 1(A)). In the classification of the 218 DCM/NF individuals, the RNA-Seq determined feature vector with 12 genes was used in the K-means clustering algorithm and this led to the correct classification of 194 individuals, yielding a misclassification rate of 11.0% (Fig. 1(B)). Notably, all six individuals who had both RNA-Seq and microarray data were correctly classified into the right group. Our results suggest that by using feature vectors determined from a training set with six individuals only, we were able to correctly classify nearly 200 individuals in the testing dataset into the correct clinical phenotype category. It is remarkable to achieve such high classification accuracy with datasets of extremely small training/testing ratio (4:231 for ISCH/NF and 5:218 for DCM/NF) (Table 3A).

As a comparison, we repeated the classification analysis based on feature vectors determined from the microarray data. We focused on the six individuals that had both RNA-Seq and microarray data, but the genes were selected by mean fold change of expression levels based on the microarray data. We used a different method to select signature genes because there is only a single intensity measure per probe, which disallows pairwise comparisons due to the lack of variance estimate for gene expression. For the ISCH/NF classification, we used the top 70 genes that had the largest fold change on gene expression as the feature vector. The clustering algorithm with this feature vector gave a misclassification rate of 12%, which is about twice as high as the misclassification rate when the RNA-Seq determined feature vector was used (Supplementary Fig. 3). Similarly, for the DCM/NF classification, the clustering algorithm with microarray determined feature vector gave a misclassification rate of 46%, which is about four times as high as that for RNA-Seq determined feature vector (Supplementary Fig. 4).

Table 2
Overlap of differentially expressed (DE) genes with GWAS loci for cardiometabolic traits.

Comparison	Category	Total	Overlap with GWAS loci	% Overlap	P-value
ISCH vs. NF	DE genes	983	72	7.32	0.0013
	Non-DE genes	8936	408	4.57	
DCM vs. NF	DE genes	1109	78	7.03	0.0041
	Non-DE genes	9353	438	4.68	
ISCH vs. DCM	DE genes	825	57	6.90	0.017
	Non-DE genes	9365	448	4.78	

The reduced misclassification rates using RNA-Seq determined feature vectors indicate the much higher accuracy of RNA-Seq in quantifying gene expression levels than microarray (Table 3B).

To evaluate the confidence of our classification results, we performed random sampling. For the classification of the ISCH/NF individuals, we randomly selected 70 genes from the union of the 9919 expressed genes obtained from RNA-Seq and used this list of genes as the feature vector for classification. We repeated this process 100 times and obtained an empirical distribution of the misclassification rate, which ranges from 8.2% to 49%. Similar analysis was carried out for the classification of the DCM/NF individuals, and the range of the misclassification rate was 16% to 49%. Results from these analyses suggest that the low misclassification rates observed in our original analysis are unlikely due to random variation.

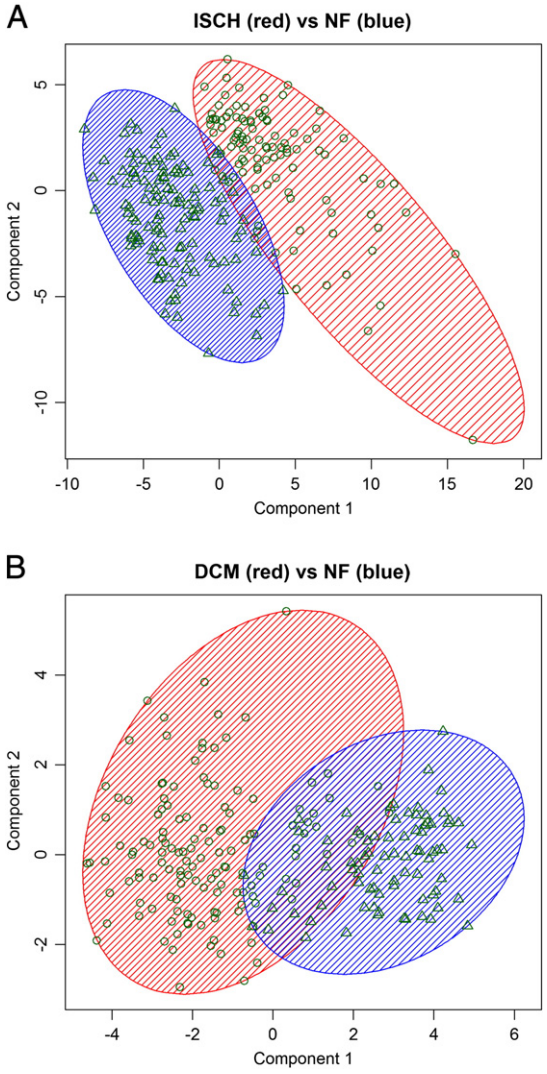

Fig. 1. K-means clustering results based on RNA-Seq determined feature vectors. (A) Clustering results for the 231 ISCH/NF individuals. (B) Clustering results for the 218 DCM/NF individuals.

We also attempted to classify the ISCH/DCM individuals. However, with the 129 RNA-Seq selected differentially expressed genes serving as the feature vector, the accuracy for classification was only slightly better than random (misclassification rate was 45%, Supplementary Fig. 5(A)). The misclassification rate based on genes selected from microarray based on mean fold change of expression was also 45% (Supplementary Fig. 5(B)). In the case of heart failure subtype classification, the differentially expressed genes could not distinguish between ISCH and DCM in an independent dataset. This failure of classification might be due to several reasons: 1) all of the ISCH and DCM individuals had end-stage heart failure, and this had obviated their initial differences [26]; 2) the relative small sample size of the training dataset; and 3) gene expression levels in the testing dataset were not accurately measured by microarray. A recent study also reported the difficulty of discriminating between cardiomyopathies of different causes [17].

4. Discussion

Heart failure results from abnormalities in multiple biological processes that contribute to cardiac dysfunction. In this study, we tested the hypothesis that a small set of genes with distinct expression patterns between failing and non-failing hearts can accurately classify disease status for complex diseases such as heart failure. Using RNA-Seq data on six individuals, we identified genes that were differentially expressed between ISCH and NF, DCM and NF, and ISCH and DCM individuals. A remarkable finding of our study is that using the gene signatures identified from this small RNA-Seq dataset, we were able to classify a much larger set of 313 individuals with failing or non-failing hearts, and the misclassification rates for the classification of ISCH/NF and DCM/NF individuals were one to three times lower than those obtained from microarray data. Such remarkable results are likely due to the highly accurate gene expression measurements obtained from RNA-Seq and careful selection of feature vectors in classification.

The unbiased RNA-Seq approach as we employed in this study identified genes that were differentially expressed between individuals with heart failure and those with non-failing hearts. Typical differential expression analysis involves group-wise comparison, i.e., comparing gene expression levels between two groups (each with multiple biological replicates), and searching for genes with different mean expression levels. Instead of searching for such overall differentially expressed genes between heart failure cases and controls, we did pairwise comparisons between every two individuals that have different disease status. We then took the intersection of the identified genes and used them as feature vectors for downstream clustering of microarray data. By so doing, we achieved a high accuracy with extremely small training: testing ratio (4:231 for ISCH/NF and 5:218 for DCM/NF); in other words, using ~2% of the samples, we correctly classified disease status for the remaining ~98% of the samples.

The advantage of the pairwise comparison lies in its ability to identify sets of genes that were differentially expressed in all pairs, i.e., globally differentially expressed, and this minimizes the contribution of less informative genes in the classification. To demonstrate this point, we compared the misclassification rates from pairwise comparisons with those obtained from group-wise comparisons (Supplementary Table 7). As

Table 3
Misclassification rates using K-means clustering for RNA-Seq determined feature vectors (A), and microarray determined feature vectors (B).

Groups	No. genes in feature vector	No. of individuals in group 1	No. of individuals in group 2	Total	No. of misclassified individuals	Misclassification rate
(A)						
ISCH + NF	70	95	136	231	15	6.49%
DCM + NF	12	82	136	218	24	11.01%
ISCH + DCM	129	95	82	177	80	45.20%
(B)						
ISCH + NF	70	95	136	231	29	12.55%
DCM + NF	12	82	136	218	101	46.33%
ISCH + DCM	129	95	82	177	79	44.63%

expected, the group-wise comparisons identified more differentially expressed genes; for the ISCH vs. NF comparison, 50 more differentially expressed genes were identified but the misclassification rate increased from 6.5% to 13.4%, which almost doubled the misclassification rate obtained from pairwise comparison; for the DCM vs. NF comparison, the number of differentially expressed genes increased more than four times and the misclassification rate reduced slightly from 11% to 7.8%; for the ISCH vs. DCM comparison, the number of differentially expressed genes was almost doubled, but the misclassification rate was only 1% lower than that from pairwise comparison. Taken together, these results indicate that pairwise comparisons significantly reduced the number of signature genes but achieved similar level or even better classification accuracy than group-wise comparisons.

In our comparison with microarray, we have used the same number of individuals ($n = 6$) in the training set as RNA-Seq. Since the cost of microarray is lower than RNA-Seq on a per-sample basis, it is of interest to compare the performance of RNA-Seq with microarray when larger number of training individuals is used in microarray data analysis. We randomly selected half of the microarray samples for the training set and used the other half for testing. The misclassification rates for each comparison were shown in Supplementary Table 8. The number of signature genes identified from this analysis was much larger than that from the pairwise RNA-Seq analysis (~8 times more for the ISCH vs. NF comparison and ~48 times more for the DCM vs. NF comparison). Although using much larger number of genes in the feature vector, the misclassification rate for the ISCH vs. NF comparison was still higher than that from RNA-seq: 7.9% vs. 6.5%; for the DCM vs. NF comparison, microarray had slightly lower misclassification rate: 7.4% vs. 11.0%. However, it is worth noting that the cost of generating microarray data for 165 subjects is much higher than that for six RNA-Seq samples. Therefore, the 3.6% reduction in misclassification rate represents a modest improvement.

Since we only had one ISCH patient, to assess the robustness of our conclusion, we analyzed two recently sequenced ISCH subjects from our ongoing MAGNet study. We repeated our analyses using these two new subjects, and obtained misclassification rates of 6.5% and 7.8%, respectively, which are comparable to the 6.5% misclassification rate obtained from the original ISCH subject. Although preliminary, this result suggests that our conclusion is robust to the choice of different ISCH patients.

Although RNA-Seq has demonstrated its superior power in studying the complexity of eukaryotic transcriptomes, this approach is still relatively new for cardiovascular genomics research. Here we report an RNA-Seq study of advanced heart failure together with a microarray study of the same population. Most studies reporting gene expression variations in heart failure have focused on small numbers of samples with advanced heart failure. Because of the relative small sample size, the resulting genes have frequently failed to be replicated. Our study represents the largest heart failure transcriptomic study reported to date. An important implication of our findings is the identification of myocardial genes associated with heart failure in humans.

RNA-Seq is a recently developed approach for transcriptome profiling that uses deep-sequencing technologies. Studies using this approach have already altered our view of the extent and complexity of eukaryotic transcriptomes. As shown by our results, RNA-Seq provides a far more precise measurement of levels of gene expression than microarray. In this study, we focused on gene expression quantification. However, using RNA-Seq, we can also quantify gene expression at the isoform level [27]. Additionally, we can examine differences of alternative splicing between two conditions, and integrate with DNA sequence data to examine allelic imbalance and RNA editing. In contrast to gene expression quantification, these analyses require a much higher sequencing depth to yield reliable results [28]. We will explore these various aspects of transcriptomic variations as we generate RNA-Seq data with higher sequencing depths.

In conclusion, we have utilized the RNA-Seq technology to identify genes with distinct expression patterns between failing and non-failing hearts. Our study demonstrates how knowledge gained from a small set of samples with accurately measured gene expressions using RNA-Seq and creative selection of classifier genes can be leveraged as a complementary strategy to discern the genetics of complex diseases. We note that analysis methods for RNA-Seq data are continuing to evolve. Additional studies employing improved analytical methods hold the potential to reveal a more complete picture of the genetic architecture of heart failure.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2014.12.002.

Conflict of interest

The authors declare that they have no conflict of interest.

Authors' contributions

M.L., Y.L. and T.C. conceived and designed the study. K.B., T.C., E.A., W.H.T., and C.M. collected the data. Y.L., M.M., S.H., Y.H. and M.L. analyzed the data. M.L. and Y.L. wrote the paper. J.B. performed the microarray and RNA-Seq experiments. K.B. and T.C. critically reviewed the manuscript. All authors read and approved the final manuscript.

Acknowledgments

This project was supported by the U.S. National Institutes of Health: R01 HL105993 to K.B.M., R01HL089847 to K.B.M. and T.P.C., R01HL088577 to T.P.C., R01GM108600, R01HL113147, R01HG006465, R01GM097505 to M.L., and R01HL103931 to W.H.T.

References

[1] C.C. Liew, V.J. Dzau, Molecular genetics and genomics of heart failure, Nat. Rev. Genet. 5 (11) (2004) 811–825.
[2] M.M. Kittleson, S.Q. Ye, R.A. Irizarry, K.M. Minhas, G. Edness, J.V. Conte, G. Parmigiani, L.W. Miller, Y. Chen, J.L. Hall, et al., Identification of a gene expression profile that differentiates between ischemic and nonischemic cardiomyopathy, Circulation 110 (22) (2004) 3444–3451.
[3] S. Hannenhalli, M.E. Putt, J.M. Gilmore, J. Wang, M.S. Parmacek, J.A. Epstein, E.E. Morrisey, K.B. Margulies, T.P. Cappola, Transcriptional genomics associates FOX

transcription factors with human heart failure, Circulation 114 (12) (2006) 1269–1276.
[4] E.O. Weinberg, M. Mirotsou, J. Gannon, V.J. Dzau, R.T. Lee, R.E. Pratt, Sex dependence and temporal dependence of the left ventricular genomic response to pressure overload, Physiol. Genomics 12 (2) (2003) 113–127.
[5] Z. Tang, B.S. McGowan, S.A. Huber, C.F. McTiernan, S. Addya, S. Surrey, T. Kubota, P. Fortina, Y. Higuchi, M.A. Diamond, et al., Gene expression profiling during the transition to failure in TNF-alpha over-expressing mice demonstrates the development of autoimmune myocarditis, J. Mol. Cell. Cardiol. 36 (4) (2004) 515–530.
[6] P. Vanburen, J. Ma, S. Chao, E. Mueller, D.J. Schneider, C.C. Liew, Blood gene expression signatures associate with heart failure outcomes, Physiol. Genomics 43 (8) (2011) 392–397.
[7] A.R. Whitney, M. Diehn, S.J. Popper, A.A. Alizadeh, J.C. Boldrick, D.A. Relman, P.O. Brown, Individuality and variation in gene expression patterns in human blood, Proc. Natl. Acad. Sci. U. S. A. 100 (4) (2003) 1896–1901.
[8] F.L. Tan, C.S. Moravec, J. Li, C. Apperson-Hansen, P.M. McCarthy, J.B. Young, M. Bond, The gene expression fingerprint of human heart failure, Proc. Natl. Acad. Sci. U. S. A. 99 (17) (2002) 11387–11392.
[9] K.B. Margulies, S. Matiwala, C. Cornejo, H. Olsen, W.A. Craven, D. Bednarik, Mixed messages: transcription patterns in failing and recovering human myocardium, Circ. Res. 96 (5) (2005) 592–599.
[10] D.R. Bentley, S. Balasubramanian, H.P. Swerdlow, G.P. Smith, J. Milton, C.G. Brown, K.P. Hall, D.J. Evers, C.L. Barnes, H.R. Bignell, et al., Accurate whole human genome sequencing using reversible terminator chemistry, Nature 456 (7218) (2008) 53–59.
[11] D. Field, G. Garrity, T. Gray, N. Morrison, J. Selengut, P. Sterk, T. Tatusova, N. Thomson, M.J. Allen, S.V. Angiuoli, et al., The minimum information about a genome sequence (MIGS) specification, Nat. Biotechnol. 26 (5) (2008) 541–547.
[12] A. Mortazavi, B.A. Williams, K. McCue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq, Nat. Methods 5 (7) (2008) 621–628.
[13] Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics, Nat. Rev. Genet. 10 (1) (2009) 57–63.
[14] C. Trapnell, L. Pachter, S.L. Salzberg, TopHat: discovering splice junctions with RNA-Seq, Bioinformatics 25 (9) (2009) 1105–1111.
[15] A. Roberts, H. Pimentel, C. Trapnell, L. Pachter, Identification of novel transcripts in annotated genomes using RNA-Seq, Bioinformatics 27 (17) (2011) 2325–2329.
[16] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D.R. Kelley, H. Pimentel, S.L. Salzberg, J.L. Rinn, L. Pachter, Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks, Nat. Protoc. 7 (3) (2012) 562–578.
[17] K.C. Yang, K.A. Yamada, A.Y. Patel, V.K. Topkara, I. George, F.H. Cheema, G.A. Ewald, D.L. Mann, J.M. Nerbonne, Deep RNA sequencing reveals dynamic regulation of myocardial noncoding RNA in failing human heart and remodeling with mechanical circulatory support, Circulation 129 (9) (2014) 1009–1021.
[18] W. Huang da, B.T. Sherman, Q. Tan, J. Kir, D. Liu, D. Bryant, Y. Guo, R. Stephens, M.W. Baseler, H.C. Lane, et al., DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists, Nucleic Acids Res. 35 (Web Server issue) (2007) W169–W175.
[19] W. Huang da, B.T. Sherman, R.A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, Nat. Protoc. 4 (1) (2009) 44–57.
[20] V. Matys, O.V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, et al., TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes, Nucleic Acids Res. 34 (Database issue) (2006) D108–D110.
[21] R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, et al., Bioconductor: open software development for computational biology and bioinformatics, Genome Biol. 5 (10) (2004) R80.
[22] W.E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods, Biostatistics 8 (1) (2007) 118–127.
[23] C. Chen, K. Grennan, J. Badner, D. Zhang, E. Gershon, L. Jin, C. Liu, Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods, PLoS One 6 (2) (2011) e17238.
[24] G.L. Gallagher, C.J. Jackson, S.N. Hunyor, Myocardial extracellular matrix remodeling in ischemic heart failure, Front. Biosci. 12 (2007) 1410–1419.
[25] L.J. Everett, S.T. Jensen, S. Hannenhalli, Transcriptional regulation via TF-modifying enzymes: an integrative model-based analysis, Nucleic Acids Res. 39 (12) (2011) e78.
[26] M. Steenman, Y.W. Chen, M. Le Cunff, G. Lamirault, A. Varro, E. Hoffman, J.J. Leger, Transcriptomal analysis of failing and nonfailing human hearts, Physiol. Genomics 12 (2) (2003) 97–112.
[27] Y. Hu, Y. Liu, X. Mao, C. Jia, J.F. Ferguson, C. Xue, M.P. Reilly, H. Li, M. Li, PennSeq: accurate isoform-specific gene expression quantification in RNA-Seq by modeling non-uniform read distribution, Nucleic Acids Res. 42 (3) (2014) e20.
[28] Y. Liu, J.F. Ferguson, C. Xue, I.M. Silverman, B. Gregory, M.P. Reilly, M. Li, Evaluating the impact of sequencing depth on transcriptome profiling in human adipose, PLoS One 8 (6) (2013) e66883.