ORIGINAL RESEARCH

Comparative Gene Expression Analysis of Mouse and Human Cardiac Maturation

CrossMark

Hideki Uosaki [1,*,a], Y-h Taguchi [2,b]

[1] *Division of Cardiology, The Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA*
[2] *Department of Physics, Chuo University, Tokyo 112-8551, Japan*

Abstract Understanding how human cardiomyocytes mature is crucial to realizing stem cell-based heart regeneration, modeling adult heart diseases, and facilitating drug discovery. However, it is not feasible to analyze human samples for maturation due to inaccessibility to samples while cardiomyocytes mature during fetal development and childhood, as well as difficulty in avoiding variations among individuals. Using model animals such as mice can be a useful strategy; nonetheless, it is not well-understood whether and to what degree gene expression profiles during maturation are shared between humans and mice. Therefore, we performed a **comparative gene expression analysis** of mice and human samples. First, we examined two distinct mice microarray platforms for shared gene expression profiles, aiming to increase reliability of the analysis. We identified a set of genes displaying progressive changes during maturation based on **principal component analysis**. Second, we demonstrated that the genes identified had a differential expression pattern between adult and earlier stages (*e.g.*, fetus) common in mice and humans. Our findings provide a foundation for further genetic studies of cardiomyocyte maturation.

Introduction

Pluripotent stem cells (PSCs) hold tremendous potential for regenerative medicine, disease modeling, and drug discovery in a broad spectrum of tissue and cell types, such as cardiomyocytes [1–4]. Recent advances in the field have rendered efficient and robust differentiation of cardiomyocytes from most of PSC lines [5–7]. Although the maturation of differentiated cardiomyocytes into the adult-like stage is essential to study adult-onset diseases *in vitro*, fully matured cardiomyocytes have never been obtained [8]. Moreover, there are no clear-cut and definitive markers available to evaluate cardiomyocyte maturation [8]. Therefore, a detailed understanding of the cardiac maturation process *in vivo* is a prerequisite for further development of methods to maturate PSC-derived cardiomyocytes *in vitro*.

Uosaki et al. examined the detailed process of mice cardiac maturation using meta-microarray analysis [9]. This and other studies demonstrated that the maturation of cardiomyocytes is

* Corresponding author.
  E-mail: huosaki1@jhmi.edu (Uosaki H).
[a] ORCID: 0000-0002-8964-8609.
[b] ORCID: 0000-0003-0867-8986.

a continuous process occurring during embryonic and postnatal development [9–12]. Because of limited human samples obtained during the early life (potentially collected from aborted fetus, babies that died from accidents or other medical reasons, and/or biopsies from transplanted hearts) and technical difficulty in repetitive sample collection from the same individual, it is difficult to dissect the progression in humans from individual variations, *e.g.*, by measuring gene expression. Therefore, studies of cardiac maturation rely heavily on model animals, *e.g.*, mice. Here, the key question remain to be addressed is whether and to what extent cardiac maturation progresses are similar in mice and humans.

Comparative gene expression analysis [13] is a useful strategy to evaluate consistency between species. It enables studying multiple human diseases in mice, which are hard to investigate directly in humans [14]. It can even help us to understand gene regulatory mechanisms in mammals using gene expression data from non-mammalian animals [15]. Moreover, it also helps in identifying highly-correlative expression profiles between putative orthologs across species [16].

In this study, we demonstrated the correlation of gene expression involved in cardiac maturation between mice and humans. We performed a meta-microarray analysis of data generated from mice samples ranging from the embryonic to the adult stages using two microarray platforms (Affymetrix Mouse Genome 430 2.0 Array, referred to as "mouse 430 2.0" hereafter and Mouse Gene 1.0 ST Array, referred as "mogene 1.0" hereafter) to collect a reliable set of genes correlating with the progression of cardiac maturation in mice. Subsequently, we evaluated whether highly-correlative expression profiles that were identified in the mice gene set exist in human samples.

Results

Performance comparison between frozen robust microarray analysis and microarray suite 5 method

In our previous paper [9], we employed the frozen robust microarray analysis (fRMA) [17] to analyze the gene expression profiles of more than 200 microarray datasets ranging from early embryonic to adult hearts. fRMA serves as a reliable platform to perform meta-microarray analysis [17]. Nonetheless, fRMA can only be applied to popular microarray platforms, such as mouse 430 2.0 and mogene 1.0, due to its requirement of preprocessed dataset. In addition, there is uncertainty on whether fRMA correctly performs batch effect extraction, although this is one of the primary reasons why fRMA is introduced. On the other hand, microarray suite 5 method (MAS5) is a method used for single-microarray preprocessing [18]. We hypothesized that MAS5 can replace fRMA for meta-microarray analysis.

To evaluate the performance of MAS5 for data preprocessing, we collected 646 microarray datasets (Table S1) and preprocessed them with MAS5 as well as fRMA. To allow comparison, MAS5-processed data was log2 transformed and scaled (mean = 0; standard deviation = 1). Signal intensities of all 45,101 probesets on mouse 430 2.0 platform were

well correlated between MAS5 and fRMA ($R = 0.90$; Pearson correlation) (**Figure 1**A). Although probes with medium signal intensities (6–12 in fRMA) showed better correlation, more variability was observed for probes with lower or higher signal intensities. To evaluate whether this variability would compromise the overall analysis, we conducted principal component analysis (PCA) for signal intensities of preprocessed data by fRMA (**Figure 1**B) and MAS5 (**Figure 1**C). The scatter plots of the first and second principal component (PC1 and PC2) values were almost identical. In addition, variable loadings for PC1 were well correlated between data preprocessed by fRMA and MAS5 ($R = 0.89$; Pearson correlation) (**Figure 1**D). These results suggest that MAS5 can replace fRMA for meta-microarray analysis. Therefore, data preprocessed by MAS5 were used for downstream analyses. As pointed out previously [9], PC1 represents the maturation process and PC2 seems to separate batch effects in either preprocessing method.

As PCA indicated a gradual maturation process in the heart [9], we next assessed how gene expression changes during the maturation process. To detect gross changes, we averaged the signal intensities of each probe at each developmental stage for ranking. **Figure 1**E depicts the distribution of the intensity ranks. As expected, the majority of probesets at the early embryonic and adult stages ranked either first or fifth, whereas more than one third of the probes at the late embryonic stage ranked third, suggesting that the expression of each gene changes gradually and unidirectionally. This finding is important when considering the limited datasets of human heart samples, which are mostly early-gestation fetal and adult samples, for comparative genomics.

Probe–gene conversion

To perform comparative gene expression analysis, it is necessary to convert probesets to genes. In mouse 430 2.0, there were more than 45,000 probesets for 20,736 genes. We used mouse 4302.db to annotate probesets to genes. As a result, 11,076 genes were annotated to single probesets, whereas the remaining genes were annotated to at least two probesets (**Figure 2**A). Seita et al. reported that identifying probes with the most dynamic ranges can be a good way to select probes [19]. However, such a method might be vulnerable to noise. Therefore, we decided to choose probes based on the interquartile ranges (IQRs) rather than the full dynamic ranges. For instance, *myomesin 2* (*Myom2*), encoding an M-protein that is expressed in mature cardiomyocytes [20], was annotated to 4 different probesets (**Figure 2**B). One probeset (1438372_at) showed a very small dynamic range, whereas the other three probesets displayed similar but distinct patterns, with the widest IQR observed for the 1457435_x_at probeset. Different from *Myom2*, *Slc2a1* that encodes glucose transporter 1 (Glut1) was annotated to 3 probesets (**Figure 2**C), which share similar IQRs. In contrast to mouse 430 2.0, more than 95% (19,925 out of 20,915 in total) of genes were annotated to a single probeset in mogene 1.0 when using mogene10sttranscriptcluster.db to annotate probesets to genes (**Figure 2**D). Therefore, for the mogene 1.0 data, we simply averaged the signal intensities from multiple probesets to obtain the expression level of a particular gene.