Heliyon 6 (2020) e04035

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Profiling and quantification of pluripotency reprogramming reveal that WNT pathways and cell morphology have to be reprogramed extensively

Kejin Hu [a,*], Lara Ianov [b], David Crossman [c]

[a] Department of Biochemistry and Molecular Genetics, University of Alabama at Birmingham, Birmingham, AL, 35294, USA
[b] Civitan International Research Center, University of Alabama at Birmingham, Birmingham, AL, 35294, USA
[c] Heflin Center for Genomic Sciences, University of Alabama at Birmingham, Birmingham, AL, 35294, USA

ARTICLE INFO

Keywords:
Biological sciences
Cell biology
Systems biology
Mathematical biosciences
Bioinformatics
Transcriptomics
Molecular biology
Developmental biology
Regenerative medicine
Induced pluripotent stem cell
Cellular reprogramming
Quantification
Mathematical model
RNA-seq
Reprogramome
Human pluripotency
iPSC
Human fibroblasts
Transcriptional profiling

ABSTRACT

Pluripotent state can be established via reprogramming of somatic nuclei by factors within an oocyte or by ectopic expression of a few transgenes. Considered as being extensive and intensive, the full complement of genes to be reprogrammed, however, has never been defined, nor has the degree of reprogramming been determined quantitatively. Here, we propose a new concept of reprogramome, which is defined as the full complement of genes to be reprogrammed to the expression levels found in pluripotent stem cells (PSCs). This concept in combination with RNA-seq enables us to precisely profile reprogramome and sub-reprogramomes, and study the reprogramming process with the help of other available tools such as GO analyses. With reprogramming of human fibroblasts into PSCs as an example, we have defined the full complement of the human fibroblast-to-PSC reprogramome. Furthermore, our analyses of the reprogramome revealed that WNT pathways and genes with roles in cellular morphogenesis should be extensively and intensely reprogrammed for the establishment of pluripotency. We further developed a new mathematical model to quantitate the overall reprogramming, as well as reprogramming in a specific cellular feature such as WNT signaling pathways and genes regulating cellular morphogenesis. We anticipate that our concept and mathematical model may be applied to study and quantitate other reprogramming (pluripotency reprogramming from other somatic cells, and lineage reprogramming), as well as transcriptional and epigenetic differences between any two types of cells including cancer cells and their normal counterparts.

1. Introduction

An enucleated oocyte can reprogram an implanted somatic cell nucleus to pluripotent stem cells (PSCs) (Byrne et al., 2007; Markoulaki et al., 2008). Ectopic expression of a few transgenes can also induce pluripotent stem cells (iPSCs) from somatic cells, most commonly fibroblasts (Takahashi and Yamanaka, 2006; Yu et al., 2007). iPSC reprogramming from human fibroblasts is a prolonged and stochastic process with very low efficiency (Hu, 2014a; Takahashi et al., 2007; Yu et al., 2007). One reason for this inefficient conversion of cell fates is probably the great expanse of reprogramming required (Shao et al., 2016a,b). Although it is considered to be extensive as well as intensive, the degree of iPSC reprogramming has not been determined quantitatively. A method for the measurement of reprogramming expanse is yet to be developed. Here, we report a new concept, reprogramome, which provides a basis for measurement of reprogramming. Subsequently, we developed the related concepts of downreprogramome, upreprogramome, erasome, and activatome. Using these concepts, we have precisely defined the breadth of reprogramming required for the establishment of human pluripotency. We have additionally developed mathematical models for quantification of reprogramming intensity of each gene in reprogramming and the total expanse of reprogramming using a new reprogramming unit, log2-transformed fold changes (LFC). Using this new concept and means of quantification of reprogramming, we revealed that WNT pathways and genes involved in cellular morphogenesis should be reprogrammed extensively and intensely for a complete conversion of human fibroblasts into iPSCs, indicating the utility of our novel concepts and mathematical models.

* Corresponding author.
E-mail address: kejinhu@uab.edu (K. Hu).

https://doi.org/10.1016/j.heliyon.2020.e04035
Received 19 March 2020; Received in revised form 30 April 2020; Accepted 18 May 2020

2. Materials and methods

2.1. Cell lines and tissue culture

We used and reported to our sponsors two NIH-registered human embryonic stem cell (hESC) lines meeting federal and university regulations. We culture human embryonic stem cells (H1 and H9) and human iPSC lines in the chemically defined E8 media (Chen et al., 2011). RNA-seq data of four human iPSC lines, 3RIPSC3 (GEO#, GSM1632433), 3RIPSC4 (GSM1632434), JQ1IPSC5 (GSM1953940), JQ10IPSC (GSM2150917), were used to further define and test the reprogramome previously defined using RNA-seq data of hESCs. These four iPSC lines were previously established and characterized in the authors' laboratory. Their pluripoentcy has been verified by the five conventional surface markers, a set of pluripotency signature genes, PCA clustering, pluripotency morphology, growth properties, and teratoma test (Shao et al., 2016a,b). Human foreskin BJ fibroblasts (ATCC, CRL-2522) were cultured in fibroblast medium: DMEM, 10% heat-inactivated FBS, 0.1 mM 2-mercaptoethanol, 100 U ml^{-1} penicillin, 100 μg ml^{-1} streptomycin, 0.1 mM MEM NEAA and 4 ng ml^{-1} human bFGF.

2.2. RNA preparation

Cells were harvested with Trizol reagent and stored at -80 °C until use. Total RNA was extracted using the Direct-zol Miniprep kit (Zymo Research, R2052). The four RNA samples of fibroblasts for RNA-seq were harvested on different days at different passage number. The three ESC RNA samples for RNA-seq were from two different ESC lines. For the repeat RNA samples of H1, they are harvested from different passages on different days. RNA from the four human iPSC lines were prepared similarly, and has been described in detail previously (Shao et al., 2016a, b).

2.3. RNA-seq

mRNA-sequencing was carried out on the Illumina HiSeq2500 following the established protocols. RNA-seq library preparation was done using the Agilent SureSelect Stranded kit (Agilent, Santa Clara, CA) as per the manufacturer's instruction. The libraries were quantitated using qPCR in a Roche LightCycler 480 with the Kapa Biosystems kit for library quantitation (Kapa Biosystems, Woburn, MA) both immediately prior to and after library construction. We conducted paired end 50-bp sequencing for downstream analyses.

2.4. Bioinformatics

All samples contained a minimum of 28.1 million reads with an average number of 40.1 million reads across all biological replicates. The FASTQ files were uploaded to the UAB High Performance Computer cluster for bioinformatics analysis with the following custom pipeline built in the Snakemake workflow system (v5.2.2) (Koster and Rahmann, 2012): first, quality and control of the reads were assessed using FastQC, and trimming of the bases with quality scores of less than 20 was performed with Trim_Galore! (v0.4.5). All samples passed initial FASTQ QC, which included good quality scores through the read length and minimal adapter contamination. Following trimming, the transcripts were quasi-mapped and quantified with Salmon (Patro et al., 2017) (v0.12.0, with `–gencode` and `-k 21′ flags for index generation and `-l A, `–gcBias` and `–validateMappings` flags for quasi-mapping) to the hg38 human transcriptome from Gencode release 29. The average quasi-mapping rate was 88.8% and the logs of reports were summarized and visualized using MultiQC (Ewels et al., 2016) (v1.6). The quantification results were imported into a local RStudio session (R version 3.5.3) and the package "tximport" (Soneson et al., 2015) (v1.10.0) was utilized for gene-level summarization. Differential expression analysis was conducted with DESeq2 package (Love et al., 2014) (v1.22.1).

We prepared heat maps in RStudio using the package of *pheatmap* (Kolde, 2019); box plots with the package of ggplot2; ladder plots with the package of *plotrix*.

2.5. Read count cutoff of DESeq2 data for expressed gene

We used two types of cutoffs for DESeq2 read counts for genes considered as being expressed, mean read count and individual read count cutoff. In DESeq2 normalization, we previously used a mean normalized read counts of 50 for the cell type in question as a cutoff for a gene to be considered active (Shao et al., 2016a,b). This cutoff is supported by our current data (Table S1-3). To confirm our selection of 50 as the mean-read-count cutoff in our data used in this manuscript, we manually selected three groups of genes based on our experience and previous microarray data (Hu and Slukvin, 2012), pluripotency (26 genes, Table S1), fibroblast (21 genes, Table S2), and double negative genes (19 genes, Table S3), the last of which are known not to be expressed in both human fibroblasts and ESCs. For all these 66 genes except for POU5F1 (i.e., *OCT4)*, the mean normalized DESeq2 read counts range from 0 to 48.6 in the cell type in which they are traditionally treated as not expressed. Most of these read counts are below 30, and only 2 of those are over 40. However, *OCT4* has a mean DESeq2 read counts of 124.6 in human fibroblasts. This is because we used FGF2 in our culture of fibroblasts. FGF2 was reported to stimulate expression of *OCT4* in fibroblasts (Jez et al., 2014). The RNA-seq signals of OCT4 provide additional evidence that our RNA-seq is very sensitive and of high quality. In addition, many well-known pluripotency genes have read counts in the lower half of three-digit numbers, for examples, *DPPA2* (254), *GDF3* (272), *LEFTY2* (431), and *NODAL* (454). These read counts are in agreement with our previous microarray data, which displayed low levels of expression for these genes (Hu and Slukvin, 2012). As an autocrine factor regulated by OCT4 and SOX2 in human ESCs with a role in ESC self-renewal (Mayshar et al., 2008), *FGF4* is considered a gene characteristic of hPSCs based on a survey of 59 human ESC lines from 17 laboratories by The International Stem Cell Initiative (International Stem Cell et al., 2007) because its expression strongly correlates with that of *NANOG*. But FGF4 expression level is very low (International Stem Cell et al., 2007) serving a reference gene for the lower limit. Our data with *FGF4* are in agreement with that of The International Stem Cell Initiative, and the averaged normalized mean read counts for *FGF4* for human ESCs are 80.6 versus 0.3 for fibroblasts. Therefore, the mean DESeq2 read counts of 50 is a reasonable cutoff (for example, this cutoff retains *FGF4* as an expressed gene but *CD19* and *CGB7* as inactive genes in human ESCs) (see Tables S1, and S3). To be stricter in selecting reliable expressed genes, we further used an individual-read-count cutoff of 10. That is, we further excluded genes from the list obtained using the above criteria, for which the individual normalized read count is less than 10 for any of the repeat experiments.

2.6. Additional selection criteria

In addition to the read count cutoffs described above, we used other strict criteria to define the reliable reprogramomes. We use q values rather than p values. We used q values of <0.01 rather than <0.05. Furthermore, we included genes only with a least 2 fold of differences in expression levels rather than 1.5 fold as a cutoff.

3. Results

3.1. Definition of reprogramome

We define reprogramome as the subset of genes that will be reprogrammed so that one cell type can be converted into another one. A reprogramome generally includes two subgroups, downreprogramome and upreprogramome. Downreprogramome refers to the group of genes whose expression levels should be downregulated while

upreprogramome include the group of genes whose expression levels should be upregulated for a complete conversion of cell fates. A downreprogramome may include a subset of genes whose expression should be shut off completely, i.e., erasome. On the other hand, an upreprogramome may contain a subset of genes whose expression should be activated *de novo*, with a term of activatome in its own right. Reprogramome may be a concept of transcription or epigenetics, and therefore there are transcriptional reprogramome and epireprogramome, respectively. Transcriptional reprogramome is a special sub-transcriptome, while epireprogramome is a defined sub-epigenome. Reprogramome may characterize any conversion of cell fates including pluripotency and lineage reprogramming. As a proof of principle, below we will summarize our profiling of the human transcriptional reprogramome for fibroblast conversion to iPSCs.

In the case of human fibroblast reprograming to iPSCs, the transcriptional downreprogramome should be the group of genes that have higher expression levels in fibroblasts than in PSCs. This group of genes should be downregulated to the expression levels found in PSCs. On the other hand, the transcriptional upreprogramome is the group of genes that have higher expression levels in PSCs than in fibroblasts. These genes should be upregulated to the levels found in PSCs. Therefore, in order to define the downreprogramome and upreprogramome, we just need to define the group of genes with higher expression in fibroblasts (fibroblast-specific genes, or simply fibroblast genes hereafter, or downregulatome) and the other group of genes with higher expression in PSCs (PSC genes hereafter, or upregulatome). The sum of the fibroblast-specific and PSC-specific genes constitutes the entire reprogramome of fibroblast-to-iPSC reprogramming.

3.2. Extensive reprogramming revealed by reprogramome profiling

To this end, we sequenced RNA on human fibroblasts and PSCs. We used the NIH-registered human embryonic stem cell lines (ESCs), H1 and H9 because these are the widely used reference cell lines for PSCs (Thomson et al., 1998). Our RNA-seq is of high quality based on the quality control analyses and read counts for the signature genes of both cell types (see Methods, and Table S1-3). Using a set of strict criteria for selection (see Methods), we showed that the downregulatome contains 3,617 genes/transcripts (Figure 1A, Table S6), representing 26.4% of fibroblast transcriptome (Figure 1G, Tables S4 and S11). The upregulatome includes 4,190 genes/transcripts (Figure 1B, Table S7), equivalent to 30.6% of fibroblast transcriptome (Figure 1G, and Table S11) and representing 28.8% of the ESC transcriptome (Tables S5 and S11). Combining downregulatome (Table S6) and upregulatome (Table S7), the reprogramome contains 7,807 genes/transcripts (Table S10). This size of reprogramome is surprisingly large and is equivalent to 57% of the fibroblast transcriptome, and 53.6% of the ESC transcriptome. The actual reprogramome may be greater because our selection criteria may have excluded a subset of genes with very low expression levels, as well as the subset of genes whose differences in expression levels between the two types of cells are lower than 2 fold, the threshold we used.

3.3. iPSC reprogramming is intensive

Next, we investigated the intensity of reprogramming. To this end, we broke the differences in gene expression levels between fibroblasts and ESCs into four tiers: 1) expressed in one cell type only but not in the other one; and differences in gene expression levels between the two cell types is 2) greater than 10 fold changes (FC), 3) 5 to 10 FC, and 4) 2 to 5 FC (Figure 1C–F). The first tier includes activatome and erasome as defined above, representing the most radical reprogramming. The remaining three tiers are designated as dramatic, moderate, and mild reprogramming. Our data show that the activatome includes 1,788 genes/transcripts equivalent to 13.1% of fibroblast transcriptome (Figure 1D, F, G, and Tables S9 and S11), while the erasome contains 1,071 genes/transcripts, representing 7.8% of fibroblast transcriptome (Figure 1C, E, G,

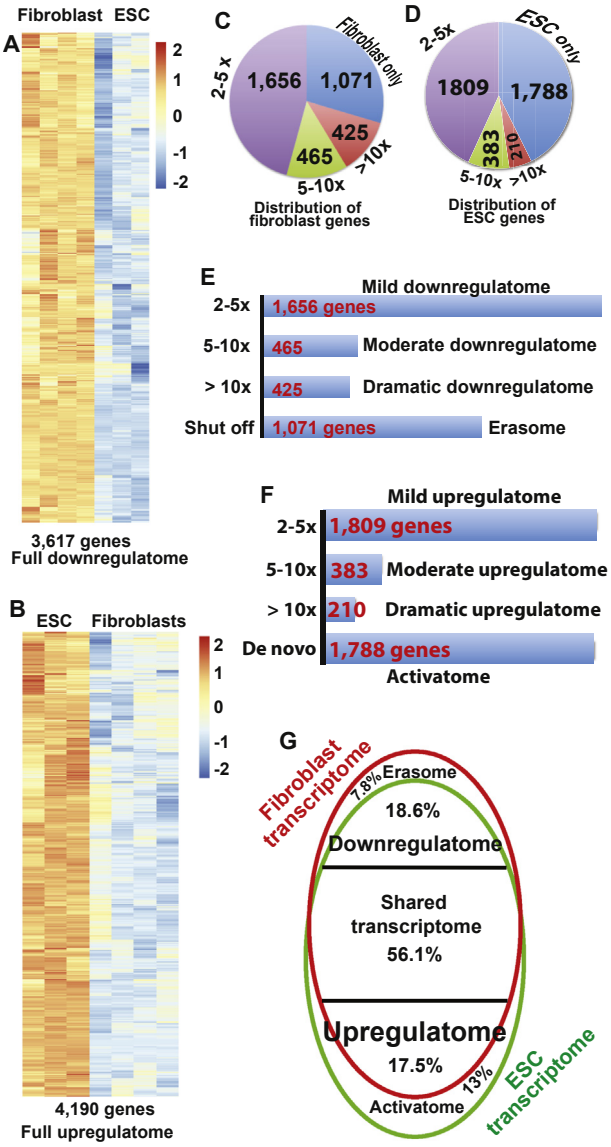

Figure 1. Profiling human fibroblast-to-iPSC reprogramome by RNA-seq. A, A heat map showing differences of gene expression levels with 2 fold or higher expression in human fibroblasts. Log2 scale. Fibroblast, n = 4; ESC, n = 3. q < 0.01. B,A heat map showing differences of gene expression levels with 2 fold or higher expression in human ESCs. Log2 scale. Fibroblast, n = 4; ESC, n = 3. q < 0.01. C, Distributions of differentially expressed genes into different levels of fold changes for fibroblast-enriched genes. D, Distributions of differentially expressed genes into different levels of fold changes for ESC genes. E, Numbers of genes in different groups of reprogramming levels for dowregulatome. F, Numbers of genes in different groups of reprogramming levels for upregulatome. G, Relative size to fibroblast transcriptome for the different sub-reprogramomes.

and Tables S8 and S11). Combining these two groups, the radical reprogramming tier includes 2,859 genes/transcripts, equivalent to 20.9% of the fibroblast transcriptome. There are 425 genes/transcripts in the category of dramatic downregulatome, and 210 genes/transcripts in dramatic upregulatome. The dramatic tier therefore includes 635 genes/transcripts, representing 4.6% of the fibroblast transcriptome. Thus, 3,494 genes/transcripts should be reprogrammed dramatically (10 fold above) or radically, equivalent to 24% of the ESC transcriptome and 25.5% of the fibroblast transcriptome. From these data, it is evident that pluripotency reprogramming is both extensive and intensive involving 57% the size of fibroblast transcriptome and a large activatome and erasome.

As expected, some well-known pluripotent genes (Hu and Slukvin, 2012) are among the activatome, for examples, *CLDN6*, *DPPA4*, *GDF3*, *L1TD1*, *LEFTY1*, *LEFTY2*, *LIN28A*, *LRRN1*, *NANOG*, *NODAL*, *PRDM14*, *SALL3*, *SOX2*, *TDGF1*, *TERT*, *ZFP42*, *ZIC5*, and *ZSCAN10* (Table S9). Unexpectedly, the master pluripotent gene, OCT4A (POU5F1), is not in the list of activatome, but is in the dramatic upreprogramome. This is because we used FGF2 in our culture of fibroblasts and FGF2 has been reported to stimulate *OCT4* expression (Jez et al., 2014). *OCT4* has a mean read counts of 124 in our fibroblasts, which is above the cutoff of 50. Some other established pluripotency genes are among the dramatic upregulatome, for examples, *DNMT3B*, *SALL2*, and *SALL4* (Table S7). *PODXL*, the gene encoding a carrier protein for the two widely used pluripotency surface markers TRA-1-60 and TRA-181 (Kang et al., 2016), is within the dramatic reprogramome. MYC, one of the original reprogramming factors (Hu, 2014a, 2014b; Takahashi and Yamanaka, 2006), is a member of the moderate upregulatome. Interestingly, *KLF2*, *KLF4* and *KLF5*, the well-known pluripotency genes in mouse (Jiang et al., 2008), are all among the human downregulatome (Table S6) rather than upregulatome, with *KLF2* unexpectedly in the erasome (Table S8). Although KLF4 is one of the four canonical reprogramming factors (Hu, 2014a, 2014b; Takahashi and Yamanaka, 2006) and plays a role in human pluripotency (Chan et al., 2009), it is widely expressed in various types of cells and tissues (Ghaleb and Yang, 2017). Of note, *KLF4* was first cloned from fibroblasts (Shields et al., 1996).

3.4. Extensive reprogramming in WNT pathway

To understand the unique features that should be established during reprogramming, we further conducted gene ontology (GO) analyses with the activatome using the PANTHER platform (Mi et al., 2013). Out of the 1,549 uniquely mapped genes, 1,405 genes fall into the unclassified group. Nevertheless, 275 genes in the activatome can be assigned to at least one PANTHER pathway. Out of the 163 pathways available in PANTHER databases, 100 are represented in the activatome, and 76 pathways are over-represented (Table S12). Among them, 29 pathways have a p value less than 0.05 and 19 pathways have FDR less than 0.05. Figure S1A shows the top 20 pathways over-represented by activatome in terms of p values. Of note, 47 genes in WNT signaling pathway are in the group of activatome ($p = 2.58 \times 10^{-5}$) (Figure 2A, Table S14). Other interesting pathways include cadherin signaling (33 genes, $p = 1.3 \times 10^{-6}$), FGF signaling (18 genes, $p = 0.01$), and VEGF signaling (12 genes, $p = 0.01$). Considering that a large number of genes in WNT pathway are represented by the activatome, we

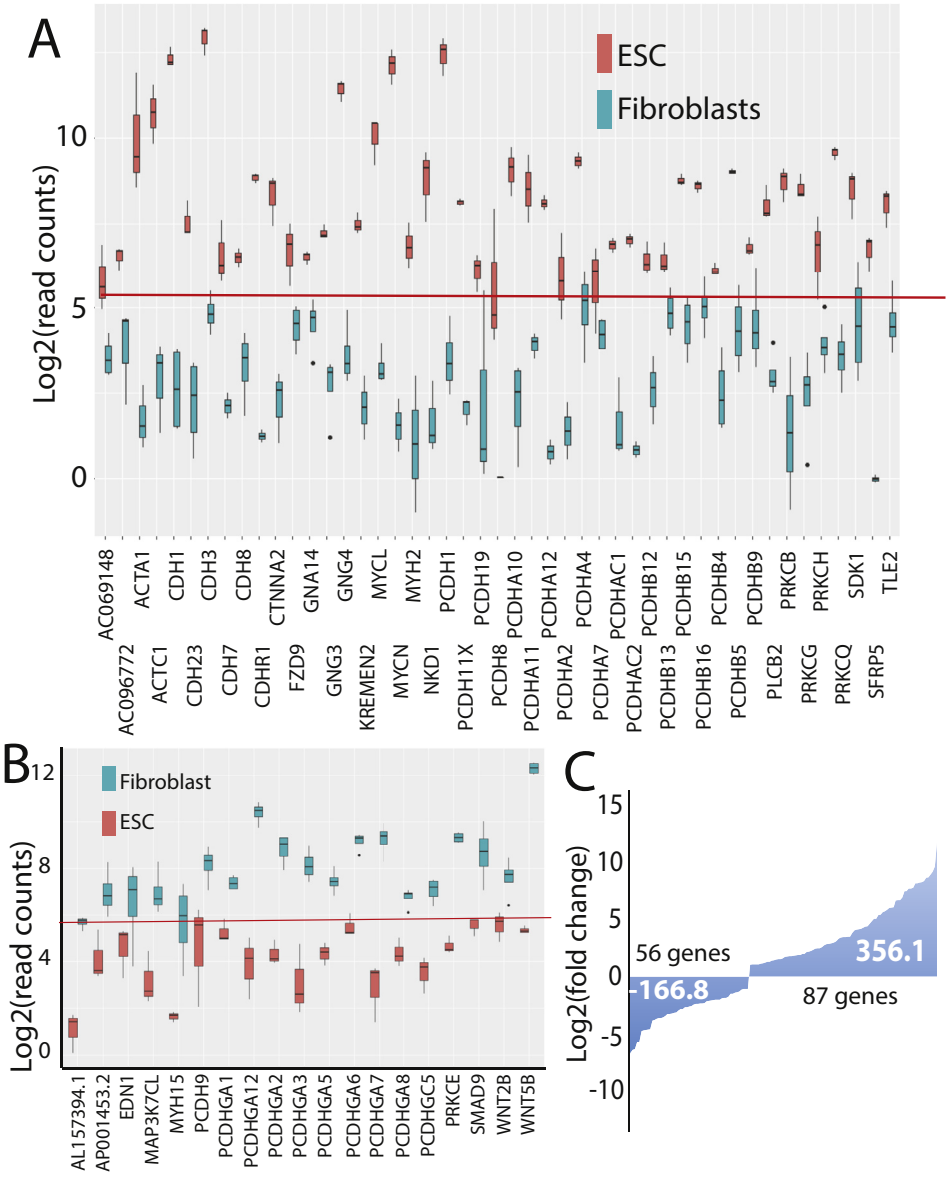

Figure 2. Profiling of WNT sub-reprogramome, and quantification of reprogramming in WNT pathways. A, Box plots showing WNT genes that have to be activated de novo for iPSC generation. Red lines in A and B mark the threshold for expressed genes. B, Box plots showing WNT genes that have to be shut off for iPSC generation. C, Total reprogramming for genes in WNT pathways. Unit for y-axis is LFC. The number within the waterfall plot is the reprogramming values in LFC calculated using our mathematical model. Numbers of WNT genes to be down- and up-reprogrammed are indicated along the waterfall plots.

further analyzed the WNT-pathway genes in the upreprogramome. Surprisingly, the upreprogramome includes 87 WNT-pathway genes ($p = 5.62 \times 10^{-4}$) (Figure 2C). This prompted our further pathway analyses with erasome and found that it contains 19 WNT-pathway genes (Figure 2B, and Table S15). We then analyzed the full downreprogramome and found that it contains 56 WNT-pathway genes (Figure 2C). In summary, 143 Wnt genes have to be reprogrammed (Table S13). Therefore, genes in WNT pathways should be intensely and extensively reprogrammed (also see quantification below).

3.5. A mathematic model for quantification of reprogramming

Pluripotency reprogramming is intensive and extensive, but there is no specific method to quantitate reprogramming. After establishment of the reprogramome concept that allows for profiling of reprogramming genes, we further reasoned that the total expanse of pluripotency reprogramming could be measured by total numbers of genes to be reprogrammed along with the degree of reprogramming for each gene. The degree of reprogramming for each gene is reflected by its fold change (FC) in transcription. To distinguish down-from up-reprogramming, we propose to use the log2-transformed fold change (LFC). For gene $_i$, the log2-transformed fold change to achieve the complete reprogramming is $G_i = \log2(FC_i)$ (Figure 3A). We assume that under an ideal condition (for example, reprogramming that happens in a fertilized egg, or reprogramming in a reconstructed egg with a transferred somatic nucleus into a mature oocyte (Hu, 2019)), every gene has the same reprogramming constant. Given a reprogramming constant of α, the amount (or intensity) of reprogramming for gene i is: $R_i = \alpha G_i$ (Figure 3A). The total reprogramming (reprogramming expanse) for the set of genes that should be upregulated is $R_{up} = \alpha \sum G_i$ (Figure 3B). The total reprogramming for the set of genes that should be downregulated can be calculated similarly, but R_{down} is a negative value. The reprogramming constant α can be arbitrarily set as 1 for clarity, and the formulas become $R_{up} = \sum G_i$ and $R_{down} = \sum G_i$ (Figure 3C). The total amount of reprogramming (total reprogramming expanse) would be: $R = R_{up} + |R_{down}|$ (Figure 3D). Based on this model, we have calculated the reprogramming expanse of human fibroblast reprogramming into pluripotency. The R_{down} is -11,096.4 LFC; R_{up} is 14,936.4 LFC, and the total reprogramming expanse R is 26,032.8 LFC (Table S11, and Figure 3E). We also calculated the amount of reprogramming for the erasome and activatome to be -5,244.6 LFC and 10,190 LFC, respectively (Table S11). R_{down} is 74.3% of R_{up}, while the amount of reprogramming for erasome is 51.5% that of activatome. These data indicate that upreprogramming is more dramatic than downreprogramming.

Since WNT pathways are intensely and extensively reprogrammed, we also quantitate the amount of reprogramming in WNT pathways. The WNT downreprogramming $R_{WNT\text{-}down}$ is -166.8 LFC while the $R_{WNT\text{-}up}$ is 356.1 LFC (Figure 2C). These results indicate that the components of WNT pathways are 2.1 times more upreprogrammed than downreprogrammed. The total WNT reprogramming R_{WNT} is 522.9 LFC, representing 2% of the overall reprogramming (Table S11).

3.6. Profiling and quantification of reprogramming in cell morphorgenesis

To demonstrate further the utility of our quantification method for reprogramming, we analyzed genes involved in cellular morphogenesis and its regulation. Conversion of fibroblasts into iPSCs involves dramatic changes in cell morphology and establishes a unique cellular colony characteristic of PSC culture. In fact, high quality iPS cell lines can be established by selecting colonies with the characteristic cell and colony morphology without a reporter (Hu et al., 2011; Shao et al., 2016a,b), and automatic imaging system can be used for identification of high quality iPSC colonies (Tokunaga et al., 2014). GO analyses of reprogramome reveal that 22 and 23 such GO terms are associated with the downreprogramome and upreprogramome, respectively (Tables S16 and 17). Figure 4B, D shows the top 10 GO terms under the category of cellular morphogenesis for upregulatome and downregulatome, respectively. There are 269 genes with roles in cellular morphogenesis that should be downregulated at least 2 fold; and 252 such genes are in the upregulatome (Figures 4A, C, E). Thus, a total of 521 genes with roles in cellular morphogenesis should be reprogrammed for pluripotency establishment, representing 3.8% of fibroblast transcriptome (Tables S11 and S18–20). $R_{morph\text{-}down}$ is -859.5 LFC while $R_{morph\text{-}up}$ is 1,133 LFC, and the total reprogramming in cellular morphogenesis R_{morph} is 1,993 LFC, representing 7.7% of the overall reprogramming (Figure 4E, and Table S11). These two data indicate that genes in cellular morphogenesis should be reprogrammed more in intensity than extensiveness (7.7% vs 3.8%). That is, the average reprogramming of each genes in cellular morphogenesis (3.8 LFC, equivalent to 13.9 FC for each gene) is higher than that for the entire reprogramome (3.3 LFC, equivalent to 9.9 FC for each gene). Although there are more genes in cellular morphogenesis that should be downregulated, the upreprogramming for such genes are more pronounced since $R_{morp\text{-}down}$ is only 75.9% of $R_{morph\text{-}up}$. In addition, the erasome includes 56 genes with roles in cellular morphogenesis while

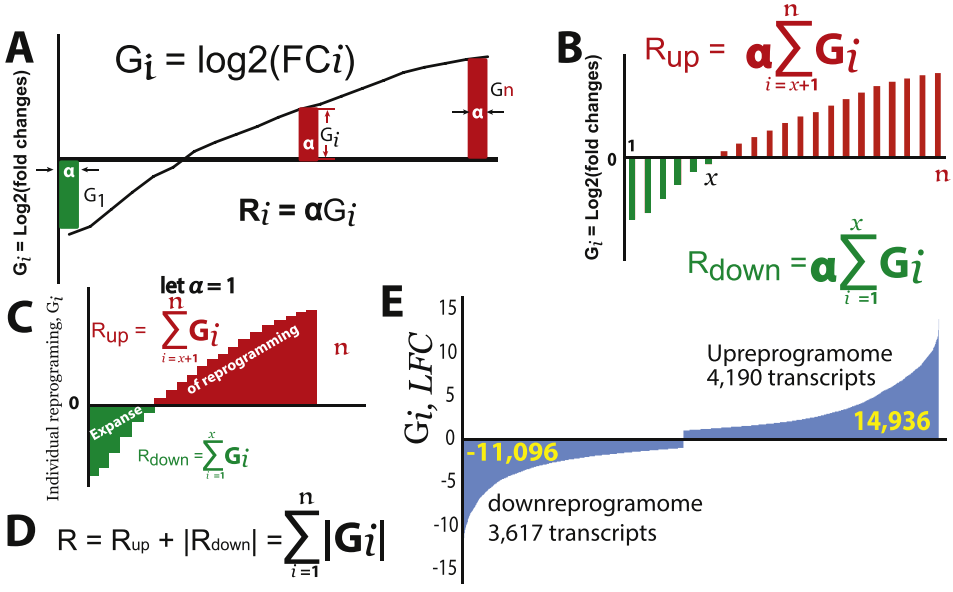

Figure 3. A mathematical model for quantification of reprogramming under an ideal condition (e.g., in a fertilized egg or a re-constructed egg with a somatic nucleus transferred into an enucleated oocyte). A, A mathematical model for calculating reprogramming amount of an individual gene. B, Mathematical models for calculating total reprogramming for downreprogramming (green) and upreprogramming (red). C, A mathematical model for calculating the total reprogramming when an arbitrary reprogramming constant a is set to be 1. The x-axis is number of genes (from 1 to n) ordered based on log2(FC) values from low to high. D, The formula for calculating the total reprogramming. E, Waterfall plots showing amount of reprogramming of human fibroblast reprogramming into pluripotency.

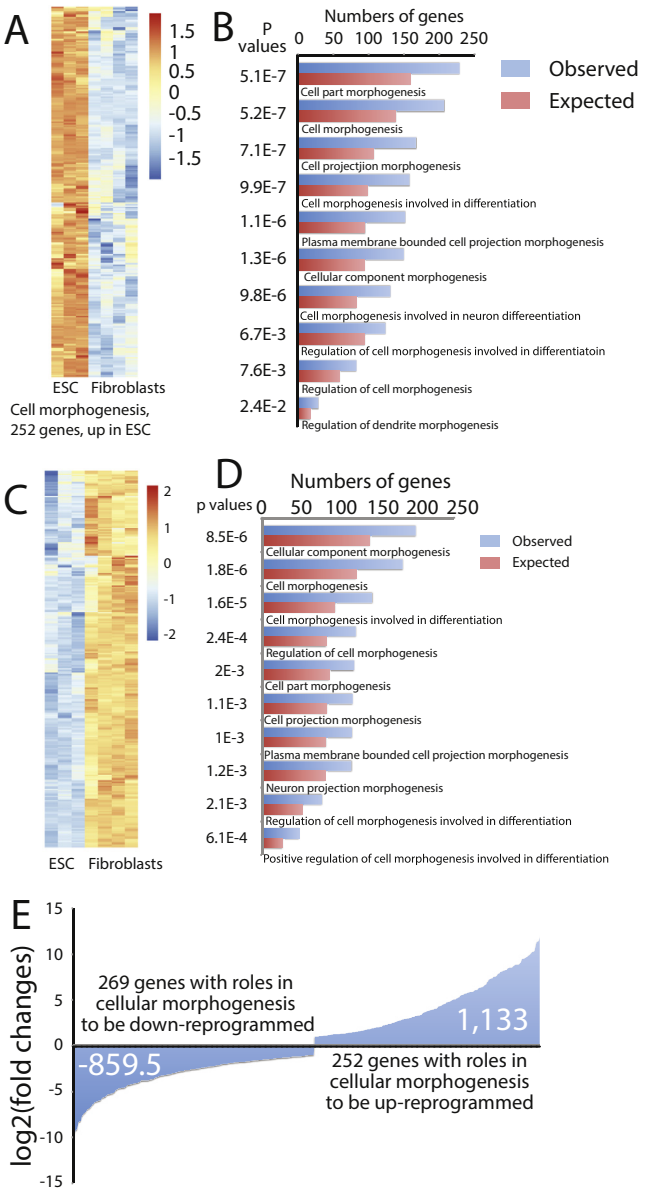

Figure 4. Gene sets in cellular morphogenesis that needs to be reprogrammed. A, A heat map showing 252 genes regulating cellular morphogenesis that have to be upreprogrammed. B, Top GO terms under the category of cellular morphogenesis for upreprogramome. C, A heat map showing 269 genes regulating cellular morphogenesis that have to be downreprogrammed. D, Top GO terms under the category of cellular morphogenesis for downreprogramome. E, Quantification of reprogramming for genes with roles in cellular morphogenesis. Statistical overrepresentation tests of GO terms for an input gene list were conducted using the "statistical overrepresentation test" tool associated with the PANTHER GO database.

the activatome contains 130 genes with such roles (Tables S18–20 and Figure S2), indicating that upreprogramming plays a more critical role in the establishment of cell and colony morphology of iPSCs. In sum, there is intensive and extensive reprogramming in genes with roles in cellular morphogenesis.

3.7. The established iPSC lines define a very similar reprogramome

Next, we asked how reliable our defined reprogramome is. To answer this, we used the established iPSC lines considering that they represent a very different source. We exploited the RNA-seq data of the four human iPSC lines our laboratory has established, characterized and published before (Shao et al., 2016a,b). Using the same criteria, we defined a downreprogramome of 3,497 genes (Figure 5A, Supplementary Table 21), and an upreprogramome of 3,885 genes (Figure 5D, Supplementary Table 22). Most of the member genes are shared with those in the reprogramome defined using the hESC transcriptional data (2,987 and 3,434 genes, respectively). However, there are 625 genes unique to the hESC-defined downreprogramome, and 751 genes unique to the hESC-defined upreprogramome. We then examined these two sets of genes missing from the HiPSC-defined reprogramome. Interestingly, very few (27 genes) of the 625 have higher mean read counts in HiPSCs, and only 7 of those are higher by 2 fold (Supplementary Table 23). We then used less stringent criteria to define the fibroblast-enriched genes ($p < 0.05$ rather than $q < 0.01$; 1.5 fold of differences rather than 2), and found that 479 out of the 625 genes are significantly enriched in fibroblasts compared to HiPSCs with additional two at the 1.3× level. This biased result indicates that almost all members in the hESC-defined downreprogramome are in fact members of the HiPSC-defined downreprogramome when less stringent criteria are applied for a small fraction of the members. The same is true for the upreprogramome. Only 14 out of the 751 genes unique to the hESC-defined upreprogramome have higher average read counts for fibroblasts (from 1.06× to 2.2×), but none of those differences is statistically significant (Supplementary Table 24). Impressively, 600 out of the 751 genes were expressed significantly higher in HiPSCs than in the starting fibroblasts ($p < 0.05$, 591 genes at >1.5×, and 9 genes at >1.34×), indicating that almost all members of the hESC-defined upreprogramome are members of the the HiPSC-defined one. The expression levels of both set of missing genes are very close to that of hESCs but lie in between that of hESCs and the starting fibroblasts (Figure 5B, E, Supplementary Figures 3A,C), indicating that the differences are largely because of incomplete reprogramming although the reprogramming is significant enough to cluster these two sets of missing genes in the HiPSC-defined reprogramome to that of hESCs (Figure 5C, F, Supplementary 3B,D). In conclusion, the hESC-defined reprogramome may serve as a bench mark for successful reprogramming especially for the core sets of 2,987 and 3,434 genes (Supplementary Tables 25 and 26, respoectively), and can identify minor incomplete reprogramming including the fibroblast transcriptional memory and insufficient establishment of the pluripotency transcriptional features.

4. Discussion

In this report, we have developed a new concept, reprogramome. This is analogous to interferome, exome, transcriptome, epigenome, reactome, proteome, and kinome. This novel concept allows for fine profiling of genes to be reprogrammed. Using the NIH-registered and widely used hESC lines of H1 and H9 as the reference transcriptomes, we defined the fibroblast-to-iPSC reprogramome. Impressively, the majority of the members (6,421 genes) consists a more reliable core fibroblast-to-iPSCs reprogramome as supported further by data from the established iPSC lines. The majority of the remaining member genes may represent a pool of genes that retain some degree of transcriptional memory of fibroblasts, or that cannot be completely upreprogrammed to the full pluripotent state. However, there may be a small set of unreliable member genes in the current reprogramome due to the limited cell lines used and other technical and experimental factors. More extensive investigations may further refine the reprogramome.

Further GO analyses, in combination with reprogramome profiling will provide many more insights into reprogramming. This is important because it is difficult to study the molecular mechanism because of very low efficiency of pluripotency reprogramming using the current protocols (Hu, 2014a, 2014b). With less than 1% of cells going to the pluripotent state, the signals we gain from the reprogramming population are mostly noise. For example, using our concept of reprogramome we were able to reveal that WNT pathways have to be extensively and intensively reprogrammed, and that the reprogramming is complicated

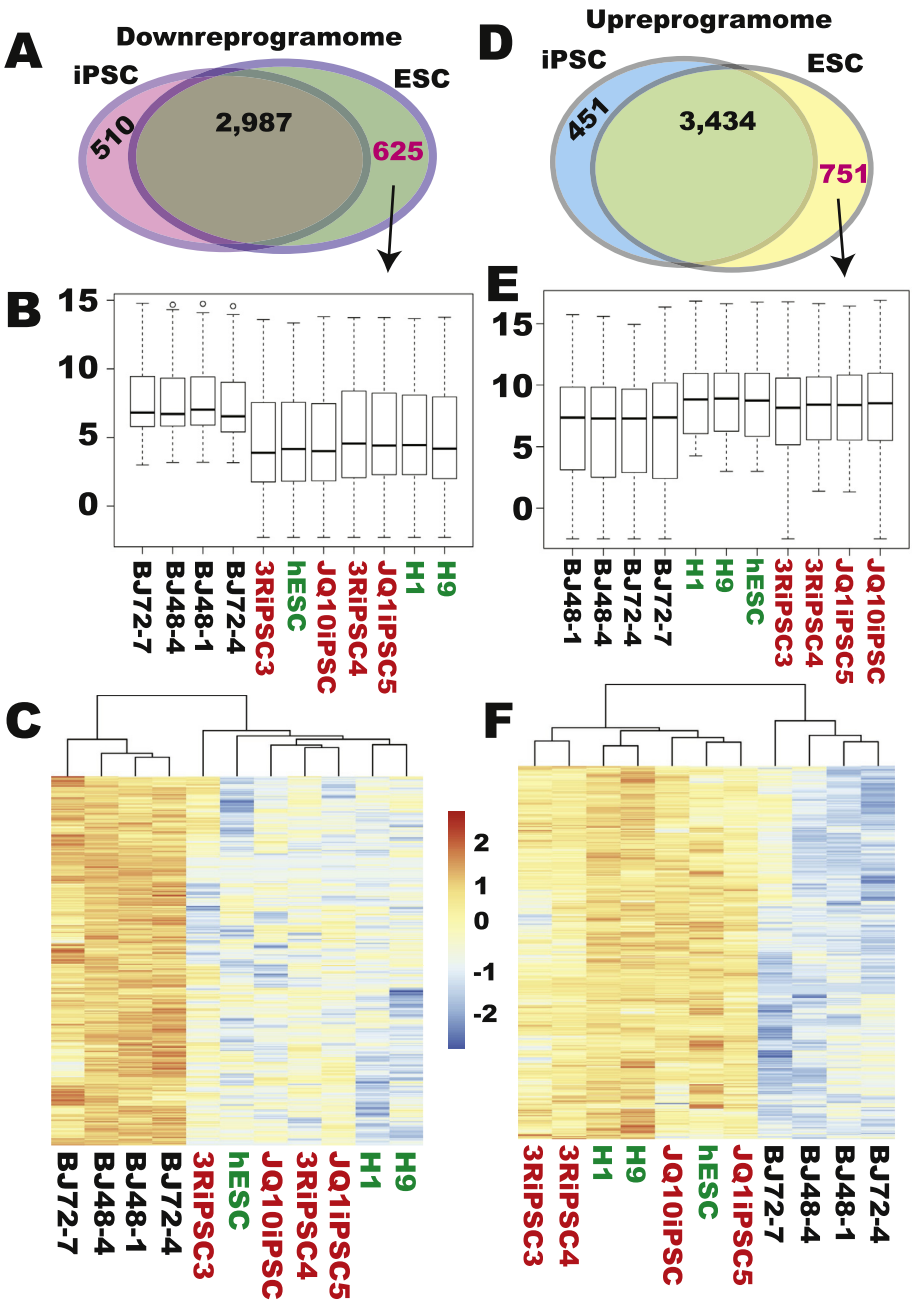

Figure 5. The reprogramome defined by the transcriptional data of hESCs is largely conserved in the established HiPSCs. A, Venn diagram showing number of genes shared by and unique to the two downreprogramomes as defined based on the RNA-seq data of hESCs and HiPSCs. B, Box plots showing that 481 out of the 625 missing genes in the HiPSC-defined downreprogramome are still statistically enriched in fibroblasts when compared with that of HiPSCs. C, A heat map for the 481 genes in B showing their similar expression in iPSCs to those of hESCs, and dissimilar to those of fibroblasts. D, Venn diagram showing numbers of genes shared by and unique to the two upreprogramomes as defined based on the RNA-seq data of hESCs and HiPSCs. E, Box plots showing that 600 out of the 751 genes unique to the hESC-defined upreprogramome are still expressed statistically higer in HiPSCs than in fibroblasts. F, A heat map for data in E showing similar expression levels for the 600 genes in HiPSCs to those in hESCs, but dissimilar to those in fibroblasts. Human ESC lines are highlighted in green, and iPSC lines in red in panels B, C, E, and F.

although upreprogramming is predominant. This is not surprising because WNT pathways regulate various cellular and developmental processes. WNT pathways are complicated. There are canonical and at least two non-canonical WNT pathways, and these are interconnected. In the human genome there are 19 WNT ligands, more than 15 WNT receptors or co-receptors, and many downstream effectors (Niehrs, 2012). WNT roles in reprogramming and pluripotency are poorly understood and warrant further investigation.

Another case study using reprogramome concept is the genes with roles in cellular morphogenesis and its regulation. We revealed that as high as 7.7% reprogramming involves genes that control and/or regulate cellular morphogenesis. This is also not surprising because iPSC generation involves dramatic changes in cellular morphology, and requires the establishment of strong cell-cell interaction among iPSCs and formation of a characteristic pluripotency colony. Our results are in agreements with a report that 1,454 genes were related to unusual colony morphology of human PSCs (Kato et al., 2016). The reprogramome for genes in morphogenesis should be much greater because we focused on cellular morphogenesis and excluded genes for morphogenesis of tissues and organ in our current analyses.

With the concept of reprogramome, here we further developed a mathematic model to quantitate reprogramming. We noticed that other methods such as PCA and t-SNE may be able to measure the transcriptional differences between the starting cells of reprograming and the endpoint cells, but our calculation focus on the degree of reprogramming/changes while PCA and t-SNE deal mainly with relationship by means of visualization after complicated dimension reduction of high-dimension complex data. Indeed, our models allows for easy estimation of reprogramming amounts in the unit of LCF for the entire reprogramome, different subreprogramome, specific pathways (e.g. WNT pathway), and cellular features (e.g., cellular morphogenesis) as demonstrated here.

Our concept and methods can be applied to pluripotency reprogramming from other starting cell types, as well as reprogramming to

various lineages such as neural and cardiac reprogramming. Furthermore, our concepts and methods can be applied to study the epigenetic changes required for a complete conversion of cell fates, i.e., epi-reprogramome. Of note, the same concept and methods may be applied to study the differences in transcription and epigenetics between any two types of cells including differences between cancer and their corresponding normal cells.

The concept of reprogramome further allows the author to evaluate the reprogramming legitimacy of transcriptional response of a gene to the conventional Yamanaka reprogramming factors (Hu, 2020). Without clear consideration of reprogramming legitimacy, the previous studies on transcriptional responses of genes to reprogramming factors were compromised by significant noises due to low efficiency (<1%), slow kinetics (>10 days), and the stochastic natures of iPSC reprogramming. With the new concepts of reprogramome and reprograming legitimacy, the author's analyses explain well both the potency and limitations of Yamanaka reprogramming.

Significance statement

The significance of this essay is at least two fold. First, we report a new concept of reprogramome, which is similar to transcriptome, kinome, exome, or interferome. Related concepts about the sub-groups of reprogramome were also proposed. The concept of reprogramome allows for fine mapping and analyses of genes that have to be reprogrammed for a complete conversion of cell fates from one type of cells to another. Second, we report new mathematical models for quantification of reprogramming. Our concepts and mathematical models will have impact beyond cellular reprogramming since they can be used to profile and quantitate transcriptional or epigenetical differences between any two types of cells.

Declarations

Author contribution statement

K. Hu: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

L. Ianov and D. Crossman: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

K. Hu was supported by the National Institutes of Health (1R01GM127411) and American Heart Association (17GRNT33670780).

Competing interest statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2020.e04035.

Acknowledgements

We appreciate the administrative support by Dr. Craig M. Powell, and technical help in RNA-sequencing by Dr. Michael Crowley. We also thank our colleagues at UAB for their critical reading of this manuscript.

References

Byrne, J.A., Pedersen, D.A., Clepper, L.L., Nelson, M., Sanger, W.G., Gokhale, S., Mitalipov, S.M., 2007. Producing primate embryonic stem cells by somatic cell nuclear transfer. Nature 450 (7169), 497–502.

Chan, K.K., Zhang, J., Chia, N.Y., Chan, Y.S., Sim, H.S., Tan, K.S., Choo, A.B., 2009. KLF4 and PBX1 directly regulate NANOG expression in human embryonic stem cells. Stem Cell. 27 (9), 2114–2125.

Chen, G., Gulbranson, D.R., Hou, Z., Bolin, J.M., Ruotti, V., Probasco, M.D., Thomson, J.A., 2011. Chemically defined conditions for human iPSC derivation and culture. Nat. Methods 8 (5), 424–429.

Ewels, P., Magnusson, M., Lundin, S., Kaller, M., 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32 (19), 3047–3048.

Ghaleb, A.M., Yang, V.W., 2017. Kruppel-like factor 4 (KLF4): what we currently know. Gene 611, 27–37.

Hu, K., 2014a. All roads lead to induced pluripotent stem cells: the technologies of iPSC generation. Stem Cell. Dev. 23 (12), 1285–1300.

Hu, K., 2014b. Vectorology and factor delivery in induced pluripotent stem cell reprogramming. Stem Cell. Dev. 23 (12), 1301–1315.

Hu, K., 2019. On mammalian totipotency: what is the molecular underpinning for the totipotency of zygote? Stem Cell. Dev. 28 (14), 897–906.

Hu, K., 2020. A PIANO (proper, insufficient, aberrant, and NO reprogramming) response to the Yamanaka factors in the initial stages of human iPSC reprogramming. Int. J. Mol. Sci. 21 (9).

Hu, K., Slukvin, I., 2012. Induction of pluripotent stem cells from umbilical cord blood. In: Meyers, R.A. (Ed.), Reviews in Cell Biology and Molecular Medicine. Wiley-VCH Verlag GmbH & Co. KGaA, pp. 1–25.

Hu, K., Yu, J., Suknuntha, K., Tian, S., Montgomery, K., Choi, K.D., Slukvin II., 2011. Efficient generation of transgene-free induced pluripotent stem cells from normal and neoplastic bone marrow and cord blood mononuclear cells. Blood 117 (14), e109–119.

International Stem Cell, I., Adewumi, O., Aflatoonian, B., Ahrlund-Richter, L., Amit, M., Andrews, P.W., et al., 2007. Characterization of human embryonic stem cell lines by the International Stem Cell Initiative. Nat. Biotechnol. 25 (7), 803–816.

Jez, M., Ambady, S., Kashpur, O., Grella, A., Malcuit, C., Vilner, L., Dominko, T., 2014. Expression and differentiation between OCT4A and its Pseudogenes in human ESCs and differentiated adult somatic cells. PloS One 9 (2), e89546.

Jiang, J., Chan, Y.S., Loh, Y.H., Cai, J., Tong, G.Q., Lim, C.A., Ng, H.H., 2008. A core Klf circuitry regulates self-renewal of embryonic stem cells. Nat. Cell Biol. 10 (3), 353–360.

Kang, L., Yao, C., Khodadadi-Jamayran, A., Xu, W., Zhang, R., Banerjee, N.S., Hu, K., 2016. The universal 3D3 antibody of human PODXL is pluripotent cytotoxic, and identifies a residual population after extended differentiation of pluripotent stem cells. Stem Cell. Dev. 25 (7), 556–568.

Kato, R., Matsumoto, M., Sasaki, H., Joto, R., Okada, M., Ikeda, Y., Furue, M.K., 2016. Parametric analysis of colony morphology of non-labelled live human pluripotent stem cells for cell quality control. Sci. Rep. 6, 34009.

Kolde, R., 2019. Pheatmap: Pretty Heatmaps. Retrieved from. https://cran.r-project.org/web/packages/pheatmap/pheatmap.pdf.

Koster, J., Rahmann, S., 2012. Snakemake–a scalable bioinformatics workflow engine. Bioinformatics 28 (19), 2520–2522.

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15 (12), 550.

Markoulaki, S., Meissner, A., Jaenisch, R., 2008. Somatic cell nuclear transfer and derivation of embryonic stem cells in the mouse. Methods 45 (2), 101–114.

Mayshar, Y., Rom, E., Chumakov, I., Kronman, A., Yayon, A., Benvenisty, N., 2008. Fibroblast growth factor 4 and its novel splice isoform have opposing effects on the maintenance of human embryonic stem cell self-renewal. Stem Cell. 26 (3), 767–774.

Mi, H., Muruganujan, A., Casagrande, J.T., Thomas, P.D., 2013. Large-scale gene function analysis with the PANTHER classification system. Nat. Protoc. 8 (8), 1551–1566.

Niehrs, C., 2012. The complex world of WNT receptor signalling. Nat. Rev. Mol. Cell Biol. 13 (12), 767–779.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., Kingsford, C., 2017. Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods 14 (4), 417–419.

Shao, Z., Yao, C., Khodadadi-Jamayran, A., Xu, W., Townes, T.M., Crowley, M.R., Hu, K., 2016a. Reprogramming by de-bookmarking the somatic transcriptional program through targeting of BET bromodomains. Cell Rep. 16 (12), 3138–3145.

Shao, Z., Zhang, R., Khodadadi-Jamayran, A., Chen, B., Crowley, M.R., Festok, M.A., Hu, K., 2016b. The acetyllysine reader BRD3R promotes human nuclear reprogramming and regulates mitosis. Nat. Commun. 7, 10869.

Shields, J.M., Christy, R.J., Yang, V.W., 1996. Identification and characterization of a gene encoding a gut-enriched Kruppel-like factor expressed during growth arrest. J. Biol. Chem. 271 (33), 20009–20017.

Soneson, C., Love, M.I., Robinson, M.D., 2015. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res 4, 1521.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., Yamanaka, S., 2007. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell 131 (5), 861–872.

Takahashi, K., Yamanaka, S., 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126 (4), 663–676.

Thomson, J.A., Itskovitz-Eldor, J., Shapiro, S.S., Waknitz, M.A., Swiergiel, J.J., Marshall, V.S., Jones, J.M., 1998. Embryonic stem cell lines derived from human blastocysts. Science 282 (5391), 1145–1147.

Tokunaga, K., Saitoh, N., Goldberg, I.G., Sakamoto, C., Yasuda, Y., Yoshida, Y., Nakao, M., 2014. Computational image analysis of colony and nuclear morphology to evaluate human induced pluripotent stem cells. Sci. Rep. 4, 6996.

Yu, J., Vodyanik, M.A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J.L., Tian, S., Thomson, J.A., 2007. Induced pluripotent stem cell lines derived from human somatic cells. Science 318 (5858), 1917–1920.