Supplementary Information

Generation of mouse ES cell lines engineered for the forced induction of transcription factors

Lina S. Correa-Cerro[1], Yulan Piao[1], Alexei A. Sharov[1], Akira Nishiyama, Jean S. Cadet, Hong Yu, Lioudmila V. Sharova, Li Xin, Hien G. Hoang, Marshall Thomas, Yong Qian, Dawood B. Dudekula, Emily Meyers, Bernard Y. Binder, Gregory Mowrer, Uwem Bassey, Dan L. Longo, David Schlessinger, and Minoru S.H. Ko*

National Institute on Aging, National Institutes of Health, Baltimore, MD 21224, USA

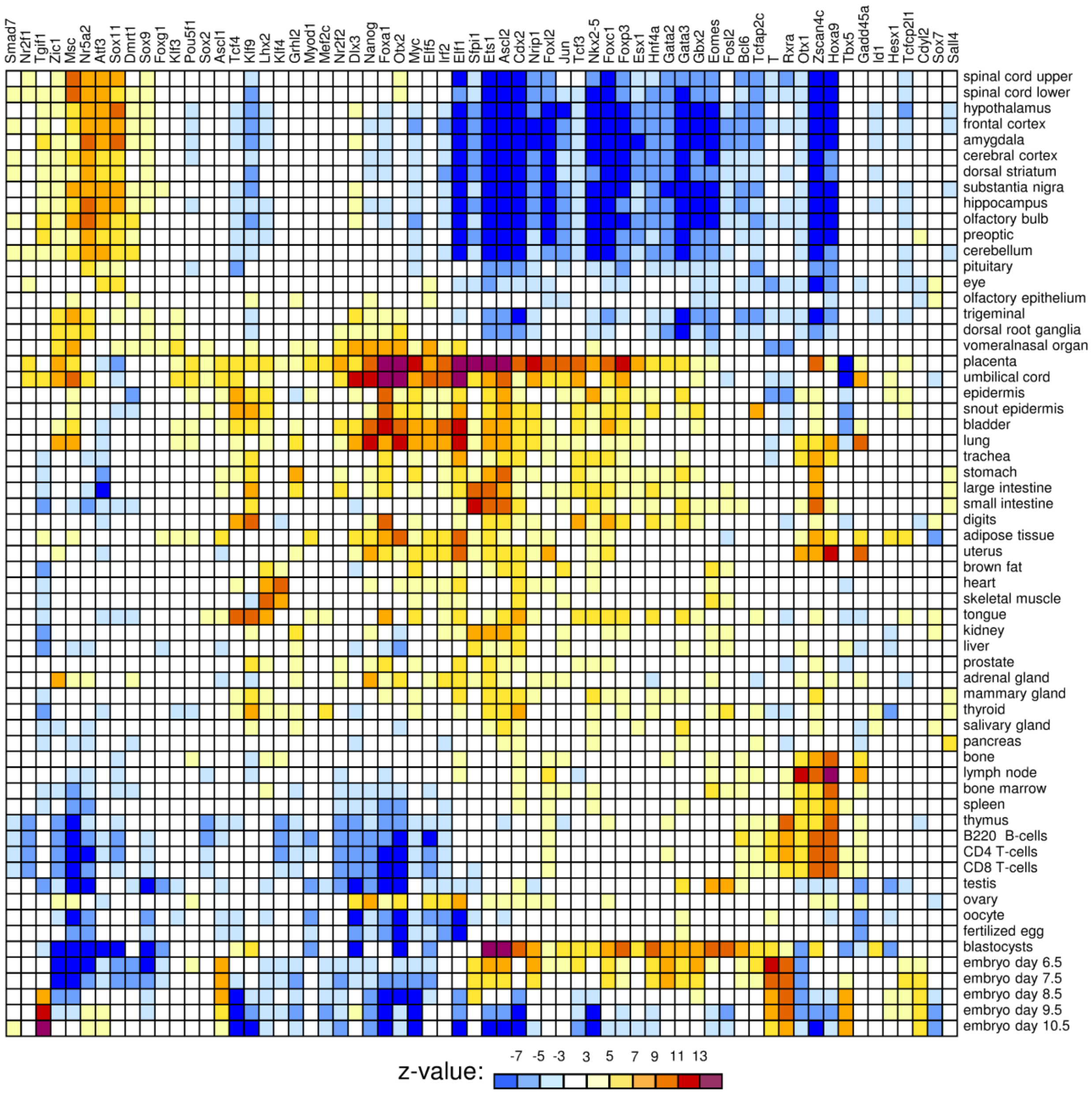
[1]These authors contributed equally to this work

*Correspondence and requests for materials should be addressed to M.S.H.K. (kom@mail.nih.gov)

Project Website : http://esbank.nia.nih.gov/

Microarray Data : GEO/NCBI (http://www.ncbi.nlm.nih.gov/geo; accession number GSE31381)
Microarray Data : NIA Array Analysis software (http://lgsun.grc.nia.nih.gov/ANOVA)

- Supplementary Figure S1
- Supplementary Table S1
- Supplementary Table S2 (separate file)
- Supplementary Methods



Supplementary Figure S1. Correlation of gene expression response to the induction of TFs with tissue-specific gene expression from the GNF ver. 2 database (Su et al. 2002).

Supplementary Table S1. qPCR primers

Gene symbol	Forward	Reverse
Aff1	GAGGCATTTCCCGAGAAGGCTC	TGGTCTGGATCCGACTTGATAGCTC
Ankrd22	GGATGCTCCTTAATGCTGGCGTAG	ATGGGCTTCCAGAAGCAGAGGG
Arnt2	AGATGGCGTCAGACATACCAGGATC	TTTACTGGGACCTTCACCATCTTCG
Ash2l	AATGGTTCACCGCTGACACCTTTG	GCACACATTGCAATGGAAGCTGTAG
Atxn1	TGGCCGTGATACAGTTTGCTGTTG	GGATGACCAGCCCTGTCCAAATAC
Batf3	GAAGAAGCAGACCCAGAAGGCTGAC	TGCGCAGCACAGAGTTCTCCTG
Bcl6	CCGTGAGCAGTTTAGAGCCCATAAG	CCCTCAGGGCTGATTTCAGGATC
Cdyl	TGGTTCGAATCAAGGAGCTGGC	TCATATTGCAGCGCACCAGGG
Cdyl2	AACGACCAGCTTGGAGAGCAGG	TCACAGGGCTGTGCAGGTTCAG
Ctbp2	AAGCAGCAGCCACTGAGATCCG	TGACCAAGGAGCTGAAGTCACGAAG
Dedd2	TCACGACCTCCTGCCACATCTG	TGGAAGAAGAGCTGGGATTGCC
Dmrt1	GAAGTGCAGCCTGATTGCGGAG	CTTTTGACCAGGAGCTCGGCTG
Dppa3	TGTCGGTGCTGAAAGACCCTATAGC	CACTGTCCCGTTCAAACTCATTTCC
Elf5	TGACAGGATGACGTACGAGAAGCTG	TTGTACACTAACCTCCGGTCAACCC
Ell2	GGAGGAATCCCGTAATCGAAGCAC	CGAGATGGCTTGAGGAGCTTTACG
Ets1	AGCCGACTCTCACCATCATCAAGAC	TGCTCGGAGTTAACAGCGGGAC
Etv1	AGCTCATACTCCGAAACCTGACCG	ACATAGGACGCCCTTCCCTTGG
Etv5	CTGAGCCGCTCTCTCCGCTATTAC	AAGAGTGCATCCGGGTCACACAC
Fbxo15	AAACCAGCACAGCGAGAAGCG	AATGCACACCAAGGTCACCGC
Fgfbp1	AGGATCCAGATGTGCTCAACCAGAG	TGTCGCCTGTAACATGTTGAGGAAG
Fhl2	TGAGGAGTGTGGAACACCCATCG	GAAGCAGCCTTCATGCCAGTGC
Fosl2	AAGACCATCGGTACCACCGTGG	TGTTTCTCTCCCTCCGGATTCG
Foxa1	CAACGACTGGAACAGCTACTACGCG	GCCGGAGTTCATGTTGCTGACAG
Foxc1	GCAGCCCAAGGACATGGTGAAG	TGTCCGGGGCATTCTGGATG
Foxg1	CCCTCAACAAGTGCTTCGTGAAGG	CGCGCTTAAAGGCCAGCTTG
Foxl2	Mm00843544_s1; TaqMan probe from ABI	
Foxn3	GTCCGGCCGTTACCAATCACTC	GTTGTGGTCCTCCTTCGGATCG
Foxp3	GTTCGCCTACTTCAGAAACCACCC	TCTCCACTCGCACAAAGCACTTG
Gata2	CCCCTAAGCAGAGAAGCAAGGCTC	ATTGCACAGGTAGTGGCCCGTG
Gbx2	ATGCGGAAGACGGCAAAGCC	CCACCTTTGACTCGTCTTTCCCTTG
Grhl2	GCAAAGCAAGTGACAGCCAAGAAG	CTCAAATTGATCTGGGCTTCACTGG
Hesx1	TCAGCTCCGGGAAAGCAAGC	TGAAGTCTCACTGGGAAGATCTGGG
Hmga2	TCCACATCAGCCCAGGGACAAC	TGGGTCTTCCTCTGGGTCTCTTAGG
Hnf4a	AGGTCAAGCTACGAGGACAGCAGC	CGAATGTCGCCATTGATCCCAG
Hoxa2	TTTATCAATAGCCAGCCGTCGCTC	CGAGTGTGAAAGCGTCGAGG
Hoxa9	AAACAACCCAGCGAAGGCGC	AGTTGGCAGCCGGGTTATTGG
Hsf2bp	AGAAACTGCACAGGCAGACAGTGG	TGTTTTGCCTCATTCAGCTGCTG
Id3	AGGAGCCTCTTAGCCTCTTGGACG	TAAGCTGAGTGCCTCGCGGG
Inppl1	AGAGAGCCAGACCCACCAGATGAC	AGGGGCAGAAATGCTGGTAGAGC
Irf2	TGGCTGGAGGAGCAGATAAATTCC	TCCTTTTCCACGTCCCATCCG

Jarid1a	CAACTTTGCCGAAGCGGTGAAC	GTGAAAAGACACAATGGCGCCTG
Jarid2	GAAGCAGAAGTCTTGCCGTGGG	CGTGTTTGCCAGACACTTTGCC
Jmjd2c	GCTCCTTCAGCAGAGACACATTTCC	TGGATGACTTCTCCCTCCGCAG
Jun	AAAGGAAGCTGGAGCGGATCG	TTCCCTGAGCATGTTGGCCG
Klf3	TTTGATCCAGTCCCTGTCAAGCAG	GCAACGGTGTGGAGTAAATGACCC
Klf9	GGGGAAACACGCCTCCGAAAAG	TTTCCCCAGTGTGGGTCCGGTAGTG
Lass2	ATTCTGCGTATGGCCCACAAGTTC	CAGTCTCCTCCCCCTCTGAACTCTC
Lhx2	TACTACAACGGCGTGGGCACTGTGC	TGCGCATGCGCTTTGTCTTTTGG
Mbd3	GAATAAGAGTCGCCAGCGTGTGC	TGGATGCAGTCTGCCGTACAGG
Meis2	TCTTCGCCAAGCAGGTTCGC	ACAGCTAATGTACCGGTGGCAGAAG
Mettl5	TGGATGGATTCGAAAAGCCCAAG	AACCGCTTTGTTTTCAATGTCATCG
Mkrn1	AGTCCATACGGCGTAGTGTGCAAG	AGGGATGGTTTTGCACTCAGATCAG
Nkx2-5	ACCCAGCCAAAGACCCTCGG	GACAGGTACCGCTGTTGCTTG
Nr2f1	CCTCAAGAAGTGCCTCAAAGTGGG	TGGATTGGGCTGGGTTGGAG
Nsbp1	CGAATGCAACATGGAAAATGCTG	CTGCTGCCACTGCTTCTTTCTTTTC
Nupr1	ACCCTTCCCAGCAACCTCTAAACC	CAGCAGCTTCTCTCTTGGTCCGAC
Ostf1	TGGAAGGGTTATGCAGACATTGTCC	TCCAAGGCCAGCTTCTTCTCATTG
Otx1	GGTGGCACTCAAGATCAACCTGC	TTCCATTCCCGCTCTGCTGC
Pdlim1	AGCTGCCCATCTGTGACAAATGTG	TTCAGGGTGGCGATGGTGATC
Prickle1	AGAGTATGCATGGGTCCCACCG	GAACCTTTTCCTCTGGCAAGCATG
Rest	GAAACACCTGAGAAACCATTTCCCC	TGAATGAGTCCGCATGTGTCGC
Sap30	GGAGACTCGCCTGTTCAGGACATC	GGTCTGGTTGGAAGCTTGAAGTGTC
Sfrs6	AAGCCATAGGCGCTCCTACTCTGG	CTGCGACTCCTACTCCGAGACCTTC
Sirt3	ATGCACGGTCTGTCGAAGGTCC	TTCACAACGCCAGTACAGACAGGG
Six1	ACTGCTTTAAGGAGAAGTCTCGGGG	ATTGTTTTCGGTGTTCTCCCTTTCC
Smad6	CCTATTCTCGGCTGTCTCCTCCTG	TTGGTGGCCTCGGTTTCAGTG
Sox11	AAGAAGTGCGCCAAGCTCAAGG	TCATCGTCGTCGTCCAGGAAGAC
Sox15	ACCCAAGGGAGCAGAGGCTTTG	AGGGGAGAAAGAGGGTCTTAGCTCC
Sox7	ATGAGAGGAAACGTCTGGCAGTGC	GTGTCAGCGCCTTCCATGACTTTC
Stra13	AGCTCATGGCGGAGTTCCTGAG	CCACAACATCCAGGTCTTCTGCC
Sub1	TGTCAGTGTTCGGGACTTCAAAGG	CCTTCAGCTGGCTCCATTGTTCC
Tbx3	CCCTTCCACCTCCAACAACACG	GTAAGGAAACAGGCTCCCGAAAGG
Tbx5	TTTGCACCCACGTCTTCCCG	CCCGAAAGCCTTTGGCGAAG
Tcfap2c	CGCACTTGCTCCTACACGATCAGAC	TCACTGGGGTTCATGACCACTCC
Tcfcp2l1	ACAGAAGCAGGATGACAGTGGGG	TCCAGGGTAGTCAGCTCTTCCAGG
Tcl1	CCAACCGCCTGTGGATCTGG	CCTGGCGCAAGATCACCTGG
Tgif1	AAGAGAAAGCACTGCTGTCCCAGC	TCTCAGCATGTCAGGAAGGAGCC
Tgm2	GCCACTTCATCCTGCTCTACAATGC	TATTCCCGTCGCTCCTCCTCTG
Trpv2	CCAAATCGGTTTGACCGTGACC	CTCTAGCAGTCCAGTCAGCTCCTCG
Txlng	TGAAATTGGCACAATGGAAGAAGC	TTCCTGCTGCACTCTGAATCTTGC
Ugp2	GTGAATTCCCTACAGTGCCCTTGG	GGTCCAGTTCCAGCATATCGGG
Zfp57	TGAGGACGTGGCAGTGTCTTTCAC	CCCTGTGCAACTGGAGGACTTCTC
Zic1	CTGGCTGCGGCAAGGTTTTC	CTCGCACTTGAAGGGCTTCTCC
Zmat4	GCACAGCTGATATCCGAGTCCCAG	GGTGAAGCATGTAATACAGCCGGAC

Supplementary Table S2. Response of gene expression 48hr after the induction of transcription factors in mouse ES cells. (Tab-delimited text file)

Supplementary Methods

Normalization of microarray data and detection of outliers

Two methods of array hybridizations were used in this study: (1) RNA extracted from cells with induced transcription factors (TFs) (cultured in Dox- conditions) and from controlled cells (cultured in Dox+ conditions) were Cy-3 labeled and all hybridized on separate arrays together with reference RNA labeled with Cy5; and (2) RNA extracted from cells with induced TFs (Dox-) were labeled with Cy3 and hybridized together with RNA from control cells (Dox+) which were labeled with Cy5. The second method does not use reference RNA. Data processing depended on the method of hybridization. Potential Cy3/Cy5 bias in microarrays with the hybridization of Dox- vs. Dox+ samples was removed by normalization to the median logratio of gene expression change in all TF-manipulation experiments.

Microarrays with Cy3-labeled sample RNA vs. Cy5-labeled reference RNA were processed as follows:

Step 1: Apply fixed cutoff=10 to all data (both Cy3 and Cy5): if(x<10) then x=10;

Step 2: Log-transform data (Log10);

Step 3: Adjust to reference RNA (Cy5): Cy3(adjusted) = Cy3 – Cy5 + average(Cy5);

Step 4: Remove outliers. Each TF is now characterized by 4 values (4 arrays): control cells (Dox+) in 2 replications and cells with induced TFs (Dox-) in 2 replications, all are log-transformed (log10). To check if any of these values is an outlier, we first we estimate the average square difference (ASD) between replication pairs:
ASD = SUM[(x(i,1) – x(i,2))2+(y(i,1) – y(i,2))2]/N/2
where x(i,j) and y(i,j) are the logintensities of Dox+ and Dox- cells, respectively, with induced i-th TF and replication j. Then we estimate the z-value for each TF: z = (x(i,1) – x(i,2))/sqrt(ASD). If abs(z) > 4 then one of the Dox+ values is an outlier. The value that is farthe away from the median Dox+ value is considered an outlier and it is replaced with the value from another replication. Similarly, we remove outliers for Dox- samples: estimate z = (y(i,1) – y(i,2))/sqrt(ASD), where y(i,j) = logintensity of Dox- cells with induced TF=i and replication=j. If abs(z) > 4 then one of the Dox- values is an outlier. The value that is farther away from the average Dox+ value for the same clone is considered an outlier and it is replaced with the value from another replication. If Dox- replications deviate from the average Dox+ value in different directions, then we assume no change in the expression of this gene.

Step 5: Adjust for gene expression variability in various transgenic clones. The main idea is that if the ES clone has an aberrant expression of gene i in Dox+ (i.e., without TF induction) and in Dox- (after TF induction) the expression returns closer to normal, then this change is not viewed as an effect of TF induction. First we estimate the significance of deviation of average expression in Dox+ from the median:

z = (average(Dox+) – median(Dox+))/SD,

where SD is standard deviation for average(Dox+) values estimated for all manipulated TFs.
If abs(z) > 2 then do the following:

```
if(average(Dox+) > median && average(Dox-) < average(Dox+)){
    average(Dox+) = median+2*SD;
```

```
        if(average(Dox-) > median+2*SD){
            average(Dox-) = $median+2*SD;
        }
    }else if(average(Dox+) < median && average(Dox-) > average(Dox+)){
        average(Dox+) = median-2*SD;
        if(average(Dox-) < median-2*SD){
            average(Dox-) = $median-2*SD;
        }
    }
```

where “median” = median expression of gene for all cell lines in Dox+ conditions.
This method of data processing was used to re-analyze microarrays from our previous study (Nishiyama et al. 2009), as well as for 7 new TFs (Ctbp2, Etv5, Jarid2, Jmjd2c, Mettl5, Tbx3, Tcl1). Statistics of data correction: 0.223% values were outliers, 0.930% values were adjusted for clone variability.

Microarrays without reference RNA (where RNA from cells cultured in Dox+ and Dox- conditions were labeled with Cy5 and Cy3, respectively, and hybridized on the same array) may potentially show more variability because data cannot be globally normalized using a reference as a yardstick. But these data can be normalized by the quantile method which is usually applied to single-dye microarrays. To take advantage of the competitive hybridization of RNA from cells in Dox- vs. Dox+ conditions here we also used a direct Cy3/Cy5 logratio as an alternative method. Finally, we compared the logratio from direct Cy3/Cy5 comparison and from quantile-normalized data, and then selected the value that was closer to zero. This is a conservative approach based on the assumption that true changes in gene expression should be detectable with both methods. Another specific feature of these data is a potential dye-related bias. To remove the bias, we used the median logratio from quantile-normalized data (Cy3/Cy5) and added this value to all Cy5 log-intensities. Below is the detailed description of the procedure:

Step 1: Log-transform all data (Log10)

Step 2: Use quantile normalization separately to Cy3 and Cy5 channels of each microarray. To make normalization smooth we approximate the cumulative probability distribution with piece-linear functions with 20 quantile nodes.

Step 3: Dye bias adjustment: estimate median logratio from quantile-normalized data (Cy3/Cy5) and add this value to all Cy5 log-intensities.

Step 4: Apply fixed cutoff=1 to all log-transformed data (it is equivalent to cutoff=10 for not log-transformed values).

Step 5: Remove outliers. Each TF is now characterized by 4 quantile-normalized values from 2 arrays: y1 = Dox- (Cy3) and x1 = Dox+ (Cy5) for the first replication and y2 = Dox- (Cy3) and x2 = Dox+ (Cy5) for the second replication, all are log-transformed (log10). First, to check if any of Dox+ values is an outlier we estimate z-value: z = (x1-x2)/sqrt(ASD), where ASD = average square difference (ASD) between replication pairs:
ASD = SUM[(x1 – x2)2+(y1 – y2)2]/N/2.
If abs(z) > 4 then one of the Dox+ values is an outlier. The Dox+ value that is farther away from the median Dox+ value for all TFs is considered an outlier and it is replaced with the value from another replication together with corresponding Dox- value. Because samples from Dox+ and Dox- conditions are hybridized to the same array, any problem with one color channel (Cy3 or Cy5) interferes with another color channel. Thus, we always replace both values for the outlier feature in the array. Second, to check if any of the Dox- values is an outlier, we estimate z = (y1-y2)/sqrt(ASD). If abs(z) > 4, then we consider

the array with a higher logratio as the outlier, and if logratios have different signs, then both replications are discarded:

```
If ((y1-x1)*(y2-x2) < 0){
        Discard both replications. Logratio = 0.
}else if (abs(y1-x1) > abs(y2-x2)){
        Values (x1,y1) are replaced by (x2,y2). Logratio = y2 – x2.
}else if (abs(y2-x2) > abs(y1-x1)){
        Values (x2,y2) are replaced by (x1,y1). Logratio = y1 – x1.
}
```

Discarding the higher absolute logratio value represents our conservative approach, according to which the change of gene expression after manipulation of TFs should be reproducible.

Step 6: Adjustment for gene expression variability in various transgenic clones (justification see above). First we estimate the significance of deviation of average expression in Dox+ from the median:

z = (average(Dox+) – median)/SD,

where SD is standard deviation for average(Dox+) values estimated for all manipulated TFs, and "median" is the median expression for all cell lines in Dox+ conditions.

```
If abs(z) > 2 then:
        if(average(Dox+) > median && average(Dox-) < average(Dox+)){
                average(Dox+) = median+2*SD;
                if(average(Dox-) > median+2*SD){
                        average(Dox-) = $median+2*SD;
                }
        }else if(average(Dox+) < median && average(Dox-) > average(Dox+)){
                average(Dox+) = median-2*SD;
                if(average(Dox-) < median-2*SD){
                        average(Dox-) = $median-2*SD;
                }
        }
```

Step 7: Combine logratio estimated from quantile-normalized data and non-normalized data for each array. From two logratios (normalized and not normalized) we select the one that has smaller absolute values. If logrations have different signs, then the response of a gene is set to zero.

Statistics of data correction: 0.263% values were outliers, 0.842% values were adjusted for clone variability.