Supplementary Experimental procedures

Dataset pre-processing

All raw expression datasets were downloaded from Gene Expression Omnibus (GEO) or European Bioinformatics Institute (EBI). A list of the datasets is shown in **Table S1A**. The data were processed using Bioconductor packages in R. Affymetrix arrays were normalized using the RMA method, and genes with multiple probes were represented by the arithmetic mean value. Illumina arrays were similarly processed using the limma package in R, and normalized using quantile normalization. Since the downloaded samples came from multiple labs, batch effects were unavoidable. We utilized ComBat to correct for batch effects (Johnson et al., 2007), which were then subsequently used for all downstream analysis.

Different human datasets were processed slightly differently. Data from Gafni et al, Hanna et al., and Ware et al., were processed independently using the affy library in R then RMA normalized. For Theunissen et al., we incorporated additional data from primed hESCs and fibroblasts from the literature (GSE46397 [samples GSM1129746-8 and GSM1129752-4] and GSE48257 [samples GSM1173414-6]) to reduced false positives in the naive module. No batch normalization was needed. Similarly, for Takashima et al., we also incorporated additional data from primed hPSCs and fibroblasts from the EBI database (E-GEOD-42807 [samples GSM1050275-8] and E-GEOD-48275 [samples GSM1173885-9]). No batch normalization was needed. For Chan et al., RNA-seq reads were mapped to the hg19 genome using BWA-aligner (Li and Durbin, 2009), then aligned to splice junctions if the read did not map to the genome. Reads counts were normalized using the Reads Per Kilobsae Million (RPKM) metric (Mortazavi et al., 2008), then quantile normalized.

For pre-implantation datasets, each dataset was analyzed separately. For microarray datasets (Vassena et al., 2011; Xie et al., 2010), raw data was processed using Bioconductor packages in R and normalized using the RMA method. Microarray datasets were performed on whole embryos. Single-cell RNA-seq sequencing datasets (Deng et al., 2014; Tang et al., 2011; Xue et al., 2013; Yan et al., 2013) were processed as previously described (Xue et al., 2013). However, the Xue et al. dataset was excluded from downstream analysis due to lack of samples in the blastocyst stage, and the mouse preimplantation dataset by Deng et al. was also excluded due to the large variations in the blastocyst samples that produced weakly correlated blastocyst

modules. All processed data and scripts will be made available online at http://labs.genetics.ucla.edu/horvath/htdocs/CoexpressionNetwork/naive_pluripotency/

Weighted gene co-expression network analysis (WGCNA)

We analyzed all the datasets with weighted gene co-expression network analysis through the R package WGCNA (Zhang and Horvath, 2005). Analysis was carried out as previously described (Xue et al., 2013). Weighted gene co-expression networks were made for each pre-implantation datasets and stem cells datasets separately. To construct the networks, we firstly generated similarity matrices *S* based on Pearson correlations between all gene pairs. Their component S_{ij} was measured by the function, $S_{ij} = 0.5 + 0.5 \times cor(i,j)$., where *cor(i,j)* denotes Pearson correlation for expression profiles of gene *i* and *j*. Then we raised the correlation matrix to a user-defined power (varies for each dataset based on noise) in order to suppress noisy small correlations and then calculated the adjacency matrices. Based on the adjacency matrices, we calculated topological overlaps (TO), which reflect not only direct adjacency in gene pairs but also their interconnections. The degree of overlaps was used to build hierarchical trees with average linkage method. We divided the trees with the Dynamic Tree Cut algorithm to determine gene modules (Langfelder et al., 2008). Then we merged modules when their eigengenes (the first principal component of module expression pattern) show strong correlation. Based on the nature of different datasets, slightly different parameters were used. The detailed parameters for above calculations in each dataset are shown in **Table S1B**.

To construct consensus modules, we followed a similar logic as constructing normal WGCNA gene networks. We first built adjacency matrices for each dataset using a soft power threshold of 60, and used the adjacency matrices to calculate topological overlap matrices (TOM). We then determined the consensus TOM based on the 25% quantile of multiple TOMs. The consensus TO was used to build the hierarchical tree, and gene modules were calculated using the Dynamic Tree Cut algorithm. Finally, gene modules were merged when module eigengenes showed a correlation of 75% or better.

Module preservation analysis

To test whether gene modules we found were conserved in other datasets, we assessed their preservation degrees with following two methods. We firstly examined overlaps between modules derived from different datasets and estimated their statistical significance from the hypergeometric distribution. We secondly utilized a composite preservation statistics proposed

by WGCNA so as to calculate Zsummary scores. Module preservation analysis uses a permutation test to define a test statistic that summarizes the evidence that the network connectivity and density of the module is preserved in an independent dataset (Langfelder et al., 2011). The Zsummary scores correspond to degrees of how much their composition and structure are preserved in the network of control samples. As shown in previous study, reference modules with Zsummary scores > 10 are highly preserved in the other (test) dataset. In contrast, Zsummary scores < 2 generally represent not to be preserved. Zsummary scores, which are above 2 and below 10, indicate the modules are moderately or weakly preserved.

Hub gene analysis

Eigengenes summarize the expression profiles for modules, and show strong correlation with a number of genes in their modules. Therefore in order to choose intramodular hub genes, we calculated kME = cor(E^q,*x(i)*), where E^q denotes eigengene of module *q* and *x(i)* denotes expression pattern of gene *i*. Since we use signed networks here, we expect that module genes have significant positive MM values. The advantage of using a correlation to quantify module membership is that the corresponding statistical significance (p-values) can be easily computed. In this analysis, we considered genes as hubs when they have a high kME score. We regarded hub genes as those with kME score greater than 0.80 ($p=10^{-22}$).

Gene ontology analysis

We performed gene ontology analysis using top hub genes (kME>0.8) of respective modules using the web-based gene enrichment analysis tool DAVID (Huang da et al., 2009). The consensus module gene ontology was performed on all 308 genes.

Supplementary References

Deng, Q., Ramskold, D., Reinius, B., and Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science *343*, 193-196.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols *4*, 44-57.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118-127.

Langfelder, P., Luo, R., Oldham, M.C., and Horvath, S. (2011). Is my network module preserved and reproducible? PLoS computational biology *7*, e1001057.

Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics *24*, 719-720.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

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

Tang, F., Barbacioru, C., Nordman, E., Bao, S., Lee, C., Wang, X., Tuch, B.B., Heard, E., Lao, K., and Surani, M.A. (2011). Deterministic and stochastic allele specific gene expression in single mouse blastomeres. PloS one *6*, e21208.

Vassena, R., Boue, S., Gonzalez-Roca, E., Aran, B., Auer, H., Veiga, A., and Izpisua Belmonte, J.C. (2011). Waves of early transcriptional activation and pluripotency program initiation during human preimplantation development. Development *138*, 3699-3709.

Xie, D., Chen, C.C., Ptaszek, L.M., Xiao, S., Cao, X., Fang, F., Ng, H.H., Lewin, H.A., Cowan, C., and Zhong, S. (2010). Rewirable gene regulatory networks in the preimplantation embryonic development of three mammalian species. Genome research *20*, 804-815.

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C.Y., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y.E*., et al.* (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature *500*, 593-597.

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J*., et al.* (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nature structural & molecular biology *20*, 1131-1139.

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. Statistical applications in genetics and molecular biology *4*, Article17.

Supplemental Figure and Table Legends

Figure S1. Reproducible and robust gene co-expression networks in mouse 2i/LIF, serum/LIF, and primed datasets

A-B) Hierarchical cluster tree of genes (black lines, n>12,000) forming co-expression modules identified using WGCNA. Co-expression modules correspond to branches and are labeled by arbitrary colors as indicated by the first color band underneath the tree. Y-axis denotes distance based on topology overlap (TO), a measure of gene interconnectedness. "Deep" branches indicate highly discrete modules. Remaining color bands indicate genes that are correlated (red) or anti-correlated (blue) to the 2i/LIF, serum/LIF, or primed pluripotent states. Modules that have significant enrichment (p<0.01) of highly correlated genes were assigned with the corresponding biological name for remaining figures (e.g. 2i/LIF, serum/LIF, or primed). Samples hybridized to the Affymetrix Mouse Gene 1.0 ST Array (A) or Illumina WG-6 expression beadchip (B) were analyzed separately.

C) Heatmap representing significance of gene overlaps between 2i/LIF, serum/LIF, and primed modules in Illumina (y-axis) or Affymetrix (x-axis) datasets (these modules were as defined in **Figures S1A and S1B**). Values in cells indicate number of overlapped genes and P-values as determined by the hypergeometric test. Color legend represents –log10 p-value.

D) Heatmap of module preservation scores between two datasets. On the top panel are module preservation scores of the Affymetrix 2i/LIF, serum/LIF, and primed modules (reference) in the Illumina dataset (test). The bottom panel shows the opposite analysis, that is, module preservation scores of the Illumina (reference) 2i/LIF, serum/LIF, and primed modules in the Affymetrix dataset (test). Generally, $Z_{summary}$ score less than 2 indicates the reference module is not preserved in the test dataset, $Z_{summary}$ greater than 10 is strongly preserved, and $Z_{summary}$ scores between 2 and 10 are weakly or moderately preserved. Because Affymetrix appeared to have more robust preservation, this dataset is used as reference for all downstream analysis.

Figure S2. Comparisons of naive hESCs generated from five different methods

A) A pairwise comparison of naive and primed modules between multiple datasets (indicated on the x and y axis). Heatmap shows the significance of gene overlaps between naive networks (left panel) and primed networks (right panel). Values in the cells indicate the number of overlapped

genes and *P*-values calculated from the hypergeometric distribution. Color legend represents –log10 *p*-value.

B) Hierarchical cluster tree of genes (black lines, *n*=12,055) forming consensus modules identified using WGCNA. Consensus modules correspond to branches and are labeled by arbitrary colors as indicated by the first color band underneath the tree. Grey color is reserved for genes that were unassigned, (i.e. not part of a consensus network across five datasets). Remaining color bands indicate genes that are highly correlated to naive pluripotent state (red) or primed pluripotent state (blue). White bands represent lack of correlation to either naive or primed states.

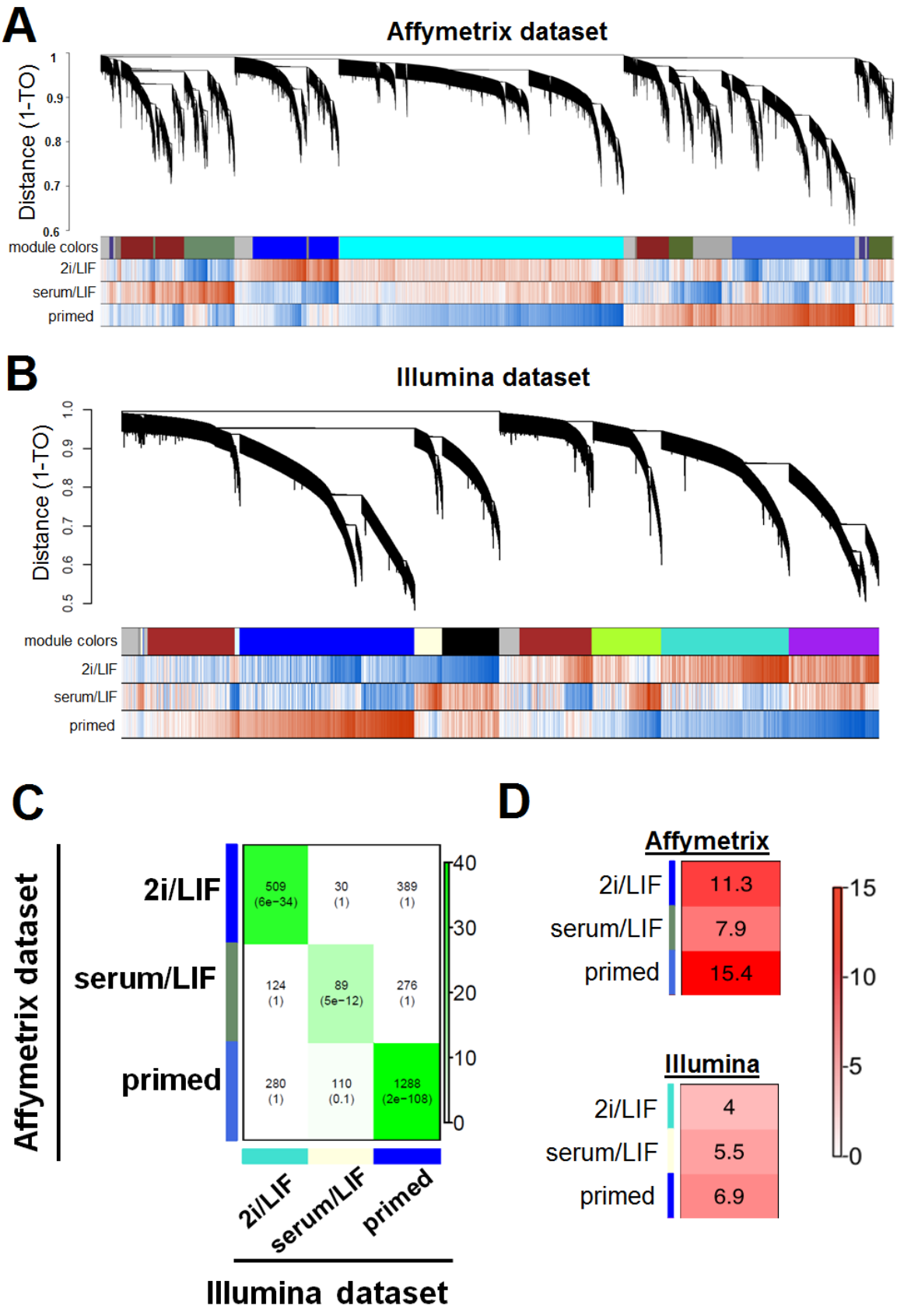
C) Bar plot of enriched gene ontology terms for all genes within the consensus naive module (*n*=317). Top terms for each category were reported. X-axis represents –log10 transformed Benjamani-Hochberg adjusted *p*-values.

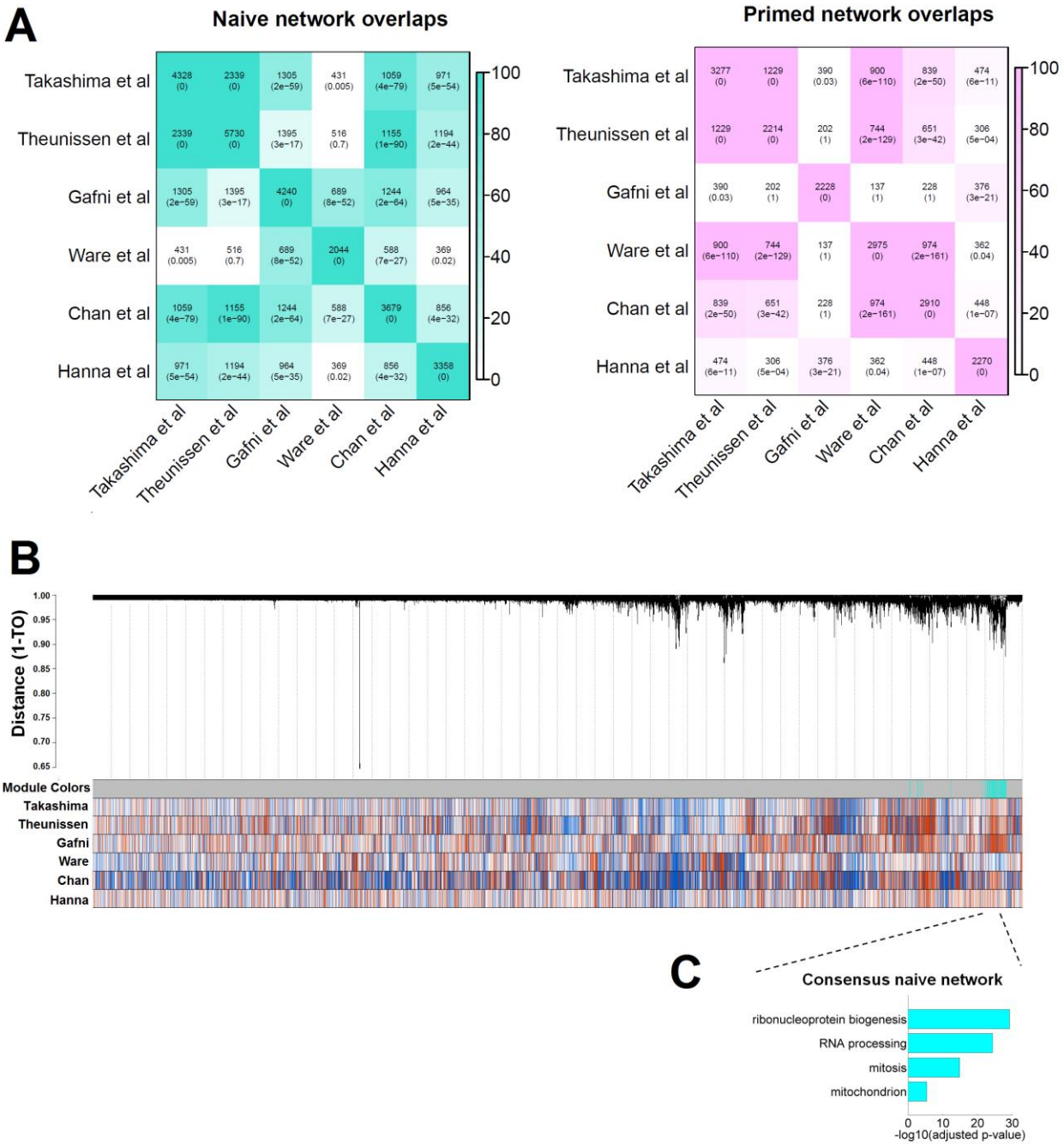
Table S1. WGCNA of mouse and human naive and primed datasets

A) Contains the list of samples used for WGCNA analysis in both Affymetrix and Illumina platforms.

B) Details of naive and primed hESC culture conditions in multiple datasets.

C) WGCNA parameters used for generating modules in various datasets.


A
Affymetrix dataset
Distance (1-TO)
1
0.9
0.8
0.7
0.6
module colors
2i/LIF
serum/LIF
primed
B
Illumina dataset
Distance (1-TO)
1.0
0.9
0.8
0.7
0.6
0.5
module colors
2i/LIF
serum/LIF
primed
C
Affymetrix dataset
2i/LIF
serum/LIF
primed
509 (6e-34)
30 (1)
389 (1)
124 (1)
89 (5e-12)
276 (1)
280 (1)
110 (0.1)
1288 (2e-108)
40
30
20
10
0
2i/LIF
serum/LIF
primed
Illumina dataset
D
Affymetrix
2i/LIF
serum/LIF
primed
11.3
7.9
15.4
Illumina
2i/LIF
serum/LIF
primed
4
5.5
6.9
15
10
5
0


A
Naive network overlaps
Takashima et al
Theunissen et al
Gafni et al
Ware et al
Chan et al
Hanna et al
4328 (0)
2339 (0)
1305 (2e−59)
431 (0.005)
1059 (4e−79)
971 (5e−54)
2339 (0)
5730 (0)
1395 (3e−17)
516 (0.7)
1155 (1e−90)
1194 (2e−44)
1305 (2e−59)
1395 (3e−17)
4240 (0)
689 (8e−52)
1244 (2e−64)
964 (5e−35)
431 (0.005)
516 (0.7)
689 (8e−52)
2044 (0)
588 (7e−27)
369 (0.02)
1059 (4e−79)
1155 (1e−90)
1244 (2e−64)
588 (7e−27)
3679 (0)
856 (4e−32)
971 (5e−54)
1194 (2e−44)
964 (5e−35)
369 (0.02)
856 (4e−32)
3358 (0)
100
80
60
40
20
0
Takashima et al
Theunissen et al
Gafni et al
Ware et al
Chan et al
Hanna et al
Primed network overlaps
Takashima et al
Theunissen et al
Gafni et al
Ware et al
Chan et al
Hanna et al
3277 (0)
1229 (0)
390 (0.03)
900 (6e−110)
839 (2e−50)
474 (6e−11)
1229 (0)
2214 (0)
202 (1)
744 (2e−129)
651 (3e−42)
306 (5e−04)
390 (0.03)
202 (1)
2228 (0)
137 (1)
228 (1)
376 (3e−21)
900 (6e−110)
744 (2e−129)
137 (1)
2975 (0)
974 (2e−161)
362 (0.04)
839 (2e−50)
651 (3e−42)
228 (1)
974 (2e−161)
2910 (0)
448 (1e−07)
474 (6e−11)
306 (5e−04)
376 (3e−21)
362 (0.04)
448 (1e−07)
2270 (0)
100
80
60
40
20
0
Takashima et al
Theunissen et al
Gafni et al
Ware et al
Chan et al
Hanna et al
B
Distance (1-TO)
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
Module Colors
Takashima
Theunissen
Gafni
Ware
Chan
Hanna
C
Consensus naive network
ribonucleoprotein biogenesis
RNA processing
mitosis
mitochondrion
0
10
20
30
-log10(adjusted p-value)