Stem Cell Reports, Volume *3*

Supplemental Information

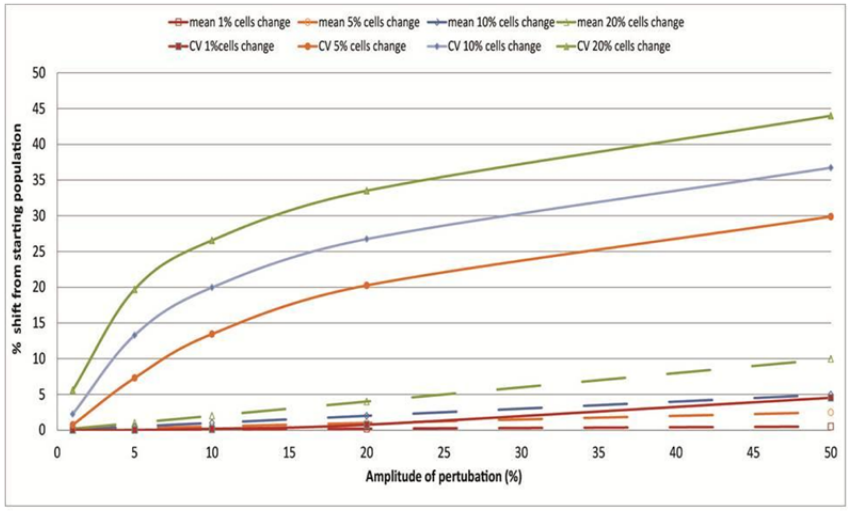
Gene Expression Variability as a Unifying Element of the Pluripotency Network

Elizabeth A. Mason, Jessica C. Mar, Andrew L. Laslett, Martin F. Pera, John Quackenbush, Ernst Wolvetang, and Christine A. Wells

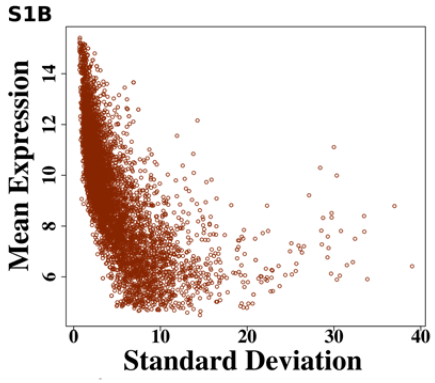
Supplemental Data

Figure S1:

S1A



S1B



S1C

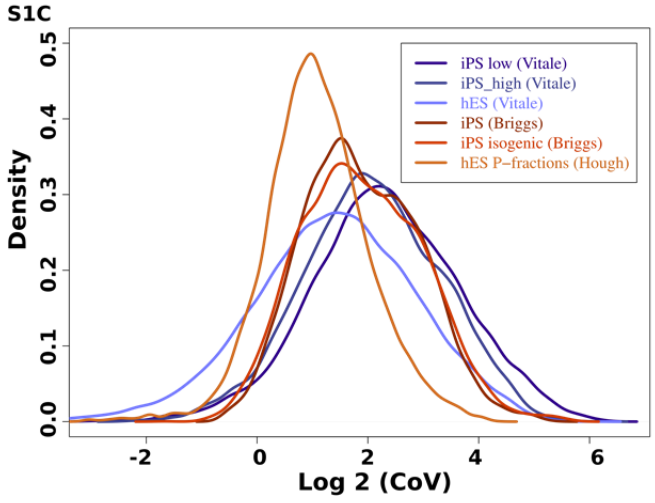


Figure S1 Legends:

S1A. *Mean changes in a small percentage of the population have little effect on the population mean, but are reflected in large changes to the population variance.* Y-axis shows the proportional shift as % deviation from the original population value. X-axis shows the amplitude of perturbation imposed on the cell population. Legend: Solid line and filled symbols for CoV values, hatched lines and open symbols for Mean values. Red lines: 1% of the cells changing; Orange lines: 5% of the cells changing; Blue lines 10% of the cells changing. Green lines 20% of the cells changing.

S1B**.** *Genes with low mean expression tend to show increased standard deviation*.
We have displayed the standard deviation as a function of mean expression for all expressed genes in the iPS unrelated (Briggs) population. Y axis displays mean expression and X axis displays standard deviation of expression. Genes with a low mean expression tend to display a higher standard deviation, perhaps due to a small proportion of cells in the population expressing the gene at a detectable level. Genes with a high mean expression do not tend to contribute to the standard deviation disproportionately.

S1C. *There are no significant differences in gene expression variability between phenotypes*. Density plots of gene expression variance were computed using a Gaussian kernel density estimator for the coefficient of variation (*R* statistical software) for all detected genes in each dataset. Y-axes display the density of log2(expression) and the Y-axes display the log2(CoV) of gene expression. Datasets were independently normalised using quantile normalisation (*lumi* Bioconductor package for *R*). Distributions were not statistically different (Levene's test; *lawstat* CRAN package for *R*) between phenotypes.

Figure S2.

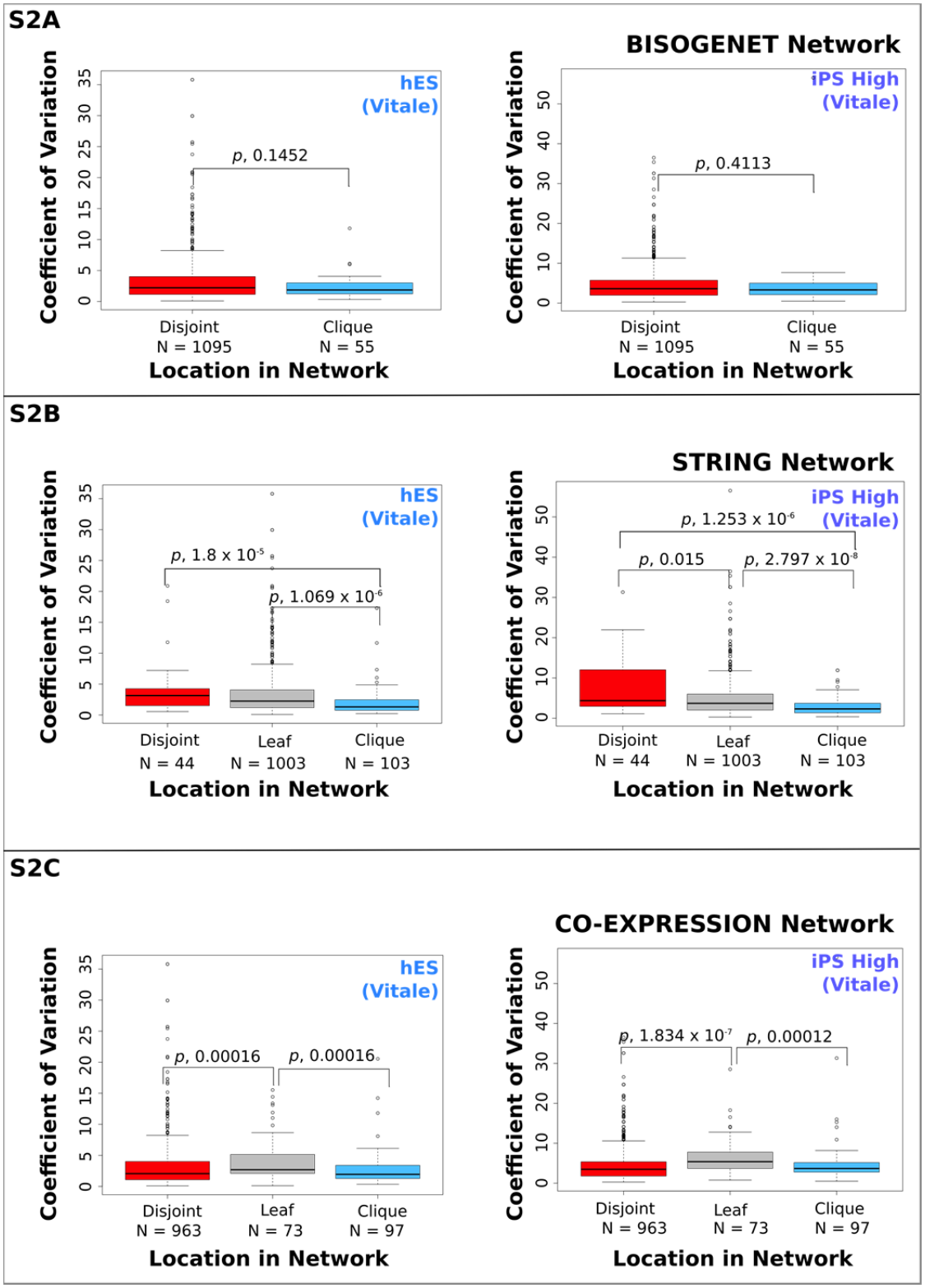

S2A
BISOGENET Network
hES
(Vitale)
iPS High
(Vitale)
Coefficient of Variation
p, 0.1452
p, 0.4113
Disjoint
N = 1095
Clique
N = 55
Location in Network
S2B
STRING Network
p, 1.8 x 10-5
p, 1.069 x 10-6
p, 1.253 x 10-6
p, 0.015
p, 2.797 x 10-8
Disjoint
N = 44
Leaf
N = 1003
Clique
N = 103
S2C
CO-EXPRESSION Network
p, 0.00016
p, 0.00016
p, 1.834 x 10-7
p, 0.00012
Disjoint
N = 963
Leaf
N = 73
Clique
N = 97

Figure S2 Legend:

Figure S3 displays CoV profiles for each region of the 3 networks generated: Protein-Protein (S2A and S2B) and co-expression (S2C) in 2 cell phenotypes (iPS and hES) from an independent dataset (Vitale). X-axis describes the network regions and Y-axis describes the coefficient of variation. P-values assess significant differences in gene expression variability between each network region (*p*, 0.05, *Wilcoxon rank sum*).

Figure S3.

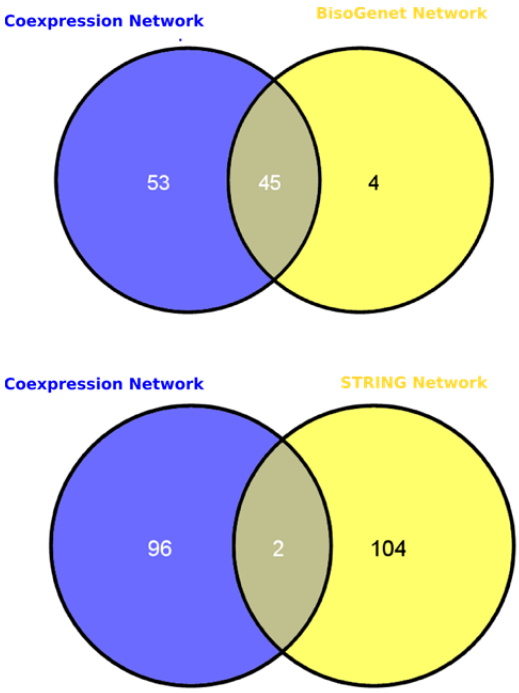


Figure S3 Legend: *Elements are shared between network cliques*. Venn diagrams in Figure S3C display the overlap in membership between the co-expression network clique, with the BisoGenet and STRING network cliques.

Table S1: Gene lists for the full co-expression network, clique and disjoint regions

Table S2: Table of significantly enriched terms in the disjoint region of the co-expression network

Table S3: K-means clustering gene lists

Table S4: List of K-means clusters of the PluriNet genes across sub-cellular fractions

Table S5: Gene lists for 3 co-expression PluriNet networks

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Microarray datasets:

All microarray data was generated on the Illumina HT-12 platform, and raw data was summarized using Bead Studio (Illumina, Inc). Background correction (*affy*) and quantile normalization was performed using *R* statistical software Bioconductor package *lumi (Du et al., 2008)*. We tested the distribution of variability in each phenotype and found no significant differences (Supplementary information 1C). All downstream analyses were performed using quantile normalised data with background correction, and only probes passing the Illumina detection threshold were included in the analysis. A probe was considered detected if its p-value was ≤ 0.01 in at least 75% of individuals in the same phenotype. We had previously tested the impact of 5 different normalisation strategies on the genome wide variance distribution, and showed that Quantile normalization offered the least pertubation of variance patterns seen in the raw data(Mar et al., 2011a).

The Illumina probe (ILMN_1659013) assigned to Nanog maps to a retrotransposed variant (NanogP1), which may be under different regulatory control to the canonical transcript. The probe mapping to the canonical transcript (ILMN_3307710) was not represented in the datasets we selected (surveyed using the Illumina HT12-V3 chips), so Nanog was excluded entirely from our analysis.

<u>Isogenic and unrelated iPS cell phenotypes (Briggs et al., 2012)</u>

The full iPSC (induced pluripotent stem cell) experimental series (GEO accession number GSE42956) assessed the derivation of *bona-fide* iPS cells from patients with Down's syndrome and healthy controls. All iPS cells were generated from fibroblasts using non-viral episomal reprogramming, and FACS sorted on TRA160 expression prior to profiling. 6 iPSCs lines from the same donor formed the isogenic iPS cell population (iPS_isogenic)(Briggs). This population was used to assess changes in CoV independent of genetic background. The unrelated iPS population (iPS_unrelated)(Briggs) encompassed all 18 iPSC samples derived from 3 different donors, thus representing a total population with mixed genetic background.

<u>Human embryonic stem cells with varying pluripotency potential (Hough et al., 2009)</u>

The hESC experimental series (GEO accession number GSE13201), surveyed four different fractions (P4, P5, P6, P7) of HES2 cells that had been FACS sorted based on two surface markers (GCTM2 and CD9) whose expression was highly correlated with self-renewal. These fractions were concordant with the

architecture of a hESC colony, such that the cells from the P4 fraction had the lowest proportion of self-renewing cells (defined as the least pluripotent) and generally located in the middle of the colony, whereas cells from the P7 fraction were found on the edge of the colony and had the largest number of self-renewing cells (defined as the most pluripotent phenotype). Where samples from all fractions were combined to produce the full colony, the population was named hES_all_P_fractions (Hough).

Phenotypic variance in induced pluripotent stem cells (Vitale et al., 2012)

The full experimental series available in Array Express (ID E-MTAB-1040) compared human ESC (Mel1) with completely reprogrammed iPSC grouped by high or low expression of the pluripotency cell surface marker SSEA4. The data in this study represented a subset of cell types representing 9 control iPSC (grouped as iPS_high and iPS_low) and 3 hESC samples from the larger dataset.

Simulating gene expression changes in the cell population:

We used *python* programming language to model a matrix of 10^7 cells, reflecting the size of a typical cell population in culture. A 1D array fitting a normal distribution was simulated using the range of expression values typically seen in the linear range of a microarray experiment (5000-50000 FU). The mean, median, standard deviation, and co-variance were calculated, and normality was tested based on D'Agostino's K-squared test. Randomized 'pooled' samples (representing a summary of 10^6 entries, or 1 'pool') were taken from the original array and the mean and CoV of these pooled samples were exported to a table (n=100 pools). Increasing percentages (we selected 1, 5, 10 and 20%) of entries in the original array were perturbed, and the degree of perturbation was also scaled (we selected 5 -50% in increasing increments of 5%), prior to resampling randomized pooled samples for each perturbation, as described above. The proportional deviation from the original population values were recorded, and were visualised in a line graph where N= 100 for either the CoV or the mean at each point.

Population variance analyses:

We examined the average gene expression variance distributions for each population across the three data sets which were processed as described above, and log(2) transformed. As a measure of variance we used the coefficient of variation (CoV), computed for each gene by dividing the standard deviation of its expression measures across a sample population by its average expression. This provides a snapshot of expression variability for each gene across a population of cells. Basing our analysis on CoV protects against detecting patterns in variability influenced by trends in absolute expression alone. Log transformation protects highly up-regulated genes from contributing to CoV disproportionately, and

thus provides an additional variance stabilizing measure. Box-plots were generated from average and CoV values of all probes. Data were considered to be outliers when falling greater than 1.5 times the inter-quartile range and are indicated by open circles. Density plots of gene expression variance were computed using a Gaussian kernel density estimator for the coefficient of variation in *R* statistical software.

Constructing a co-expression network from known pathways, enriched in the pluripotent phenotype: Pathway-based significance between fibroblast and iPSC phenotypes in the Briggs et al. (2012) dataset was determined using the *attract* algorithm (Mar et al., 2011b; Mar et al., 2011c). All pathways in KEGG were assessed, and the *PluriNet* originally described by Muller et al. (2007) was assessed individually against all pathways in KEGG(Franz-Josef Muller, 2008) (Kanehisa et al., 2002). Gene sets were identified for the synexpression groups of *PluriNet* and ECMR-interaction (Extracellular Matrix Receptor) pathways. Correlated partners of the synexpression groups were computed at a Pearson coefficient cut-off of +0.9. The list of probes representing the *PluriNet* and ECMR- interaction pathways and their correlated partners of expression was mapped from probe to official gene symbol level (for a full description of methods see Supplementary Information 3: Mapping) using *python*. Correlated partners of expression of the synexpression groups identified in *PluriNet* and the ECMR-interaction pathways were generated using the *attract* algorithm. The Pearson R correlation threshold was set at above or equal to +0.9. A single list of genes was generated which comprised members of the ECMR-interaction and *PluriNet* pathways, and their correlated partners of expression. Those gene pairs with a Pearson R value equal to or above +0.995 and below -0.995 were selected as network nodes. The network was visualized using a force directed spring embedded layout in *Cytoscape,* where the correlation coefficient between the pair of genes represents an edge weight (Shannon et al., 2003). Associated with an edge was either positive (Pearson R >= 0.995; green) or negative (Pearson R <= -0.995; red) correlation in gene expression.

Constructing protein-protein interaction networks in Cytoscape:
Cytoscape plug-ins (STRING.db and BisoGenet) were used to construct edges representing protein-protein interactions. This produced 2 different protein-protein interaction networks with node colour and shape. BisoGenet (Martin et al., 2010) is a Cytoscape plugin which integrates data from well-known interaction databases including DIP, BIOGRID, HPRD, BIND, MINT and INTAC. STRING.db (Francheschini et al., 2012) is a database which provides known and predicted (scored) associations between proteins,

which results in comprehensive protein networks covering >1100 organisms. We imported our co-expression node list into STRING.db to form a medium-stringency network for *Homo sapiens*. Details provided in Supplementary Information S3B.

Network analyses:

Network architecture:

The larger network was divided into 3 regions based with different connectivity:

1. Clique: Nodes that form part of the densely connected network core. Characterized by blue circles.
2. Leaf: Nodes peripherally connected to the main network hub. Characterized by grey triangles.
3. Disjoint: Nodes that were disconnected from the main network. Characterized by red squares.

Supplementary Information S2A contains gene lists for each region

The force directed spring embedded algorithm pushes nodes with a higher degree toward the centre (clique region), and nodes with a reduced degree further away.

CoV profiles:

Box-plots were generated from the CoV values for each group, in the iPS_unrelated (Briggs) and the hES_P_fractions (Hough) datasets. A *Wilcoxon rank sum test* assessed whether the differences between the distributions were statistically significant.

Constructing networks which represent pluripotent and transitioning cell populations

The PluriNet pathway was identified as significant in the *attract* analysis, and was decomposed into distinct modes of expression variability. We used agglomerative hierarchical clustering with average linkage to cluster the log2-transformed CoV data and used the Gap statistic with 1000 bootstrap samples to determine the number of appropriate variance clusters. A unique list of probes with a 1:1 mapping to official gene symbol represents all genes in these variance clusters, and there are 60, 97 and 39 genes associated with each cluster respectively, totalling 196 unique genes. (Supplementary Information S4)

The sub-fractions were grouped as follows:

Network 1: P4 & P5 microarray data

Network 2: P5 & P6 microarray data

Network 3: P6 & P7 microarray data

For each group we selected the full list of 196 probes and performed a pair-wise Pearson correlation of gene expression was performed using *R* statistical software. The gene pairs with a Pearson R value equal to or above +0.9 and below -0.9 were selected as network nodes, with the correlation between them representing an edge. The networks were visualized using a force directed spring embedded lay out in *Cytoscape* (Shannon et al., 2003), where the correlation coefficient between the pair of genes represents an edge weight. Genes were represented as circular nodes, and their pair-wise correlation of expression represented as an edge. Associated with an edge was either positive (Pearson R >= 0.9; green) or negative (Pearson R <= -0.9; red) correlation in gene expression, corresponding to the Pearson R coefficient.

S5. SUPPLEMENTARY REFERENCES

Briggs, J.A., Sun, J., Shepherd, J., Ovchinnikov, D.A., Chung, T.L., Nayler, S.P., Kao, L.P., Morrow, C.A., Thakar, N.Y., Soo, S.Y., *et al.* (2012). Integration-Free Induced Pluripotent Stem Cells Model Genetic and Neural Developmental Features of Down Syndrome Etiology. Stem Cells.

Du, P., Kibbe, W.A., and Lin, S.M. (2008). lumi: a pipeline for processing Illumina microarray. Bioinformatics *24*, 1547-1548.

Francheschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., *et al.* (2012). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Research *41*, 1-8.

Franz-Josef Muller, L.C.L., Dennis Kostka, Igor Ulitsky, Roy Williams, Christiana Lu, In-Hyun Park, Mahendra S. Rao, Ron Shamir, Phillip H. Schwartz, Nils O. Schmidt, Jeanne F. Loring (2008). Regulatory networks define phenotypic classes of human stem cell lines. Nature *455*, 5.

Hough, S.R., Laslett, A.L., Grimmond, S.B., Kolle, G., and Pera, M.F. (2009). A continuum of cell states spans pluripotency and lineage commitment in human embryonic stem cells. PLoS One *4*, e7708.

Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A. (2002). The KEGG databases at GenomeNet. Nucleic Acids Res *30*, 42-46.

Mar, J.C., Matigian, N.A., Mackay-Sim, A., Mellick, G.D., Sue, C.M., Silburn, P.A., McGrath, J.J., Quackenbush, J., and Wells, C.A. (2011a). Variance of gene expression identifies altered network constraints in neurological disease. PLoS Genet *7*, e1002207.

Mar, J.C., Matigian, N.A., Quackenbush, J., and Wells, C.A. (2011b). attract: A method for identifying core pathways that define cellular phenotypes. PLoS One *6*, e25445.

Mar, J.C., Wells, C.A., and Quackenbush, J. (2011c). Defining an informativeness metric for clustering gene expression data. Bioinformatics *27*, 1094-1100.

Martin, A., Ochagavia, M.E., Rabasa, L.C., Miranda, J., Fernandez-de-Cossio, J., and Bringas, R. (2010). BisoGenet: a new tool for gene network building, visualization and analysis. BMC Bioinformatics *11*, 91.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res *13*, 2498-2504.

Vitale, A.M., Matigian, N.A., Ravishankar, S., Bellette, B., Wood, S.A., Wolvetang, E.J., and Mackay-Sim, A. (2012). Variability in the generation of induced pluripotent stem cells: importance for disease modeling. Stem Cells Transl Med *1*, 641-650.