Stem Cell Reports, Volume *7*

Supplemental Information

Integrated Genomic Analysis of Diverse Induced Pluripotent Stem Cells from the Progenitor Cell Biology Consortium

Nathan Salomonis, Phillip J. Dexheimer, Larsson Omberg, Robin Schroll, Stacy Bush, Jeffrey Huo, Lynn Schriml, Shannan Ho Sui, Mehdi Keddache, Christopher Mayhew, Shiva Kumar Shanmukhappa, James Wells, Kenneth Daily, Shane Hubler, Yuliang Wang, Elias Zambidis, Adam Margolin, Winston Hide, Antonis K. Hatzopoulos, Punam Malik, Jose A. Cancelas, Bruce J. Aronow, and Carolyn Lutzko

Supplemental Information

Study Design

The PCBC Core Standards Working Group identified the characteristics for the iPSC included in the study based on donor cell type, reprogramming vector and gene combinations. The Cincinnati Cell Characterization Core (C4) established standard protocols for thaw, adaptation to standardized feeder-free culture conditions, sample collection and analysis. Manufacturer lots were tested and controlled for standardization of key reagents. Lines were thawed directly into feeder-free mTeSR1 culture media on Matrigel (hereafter referred to as feeder-free) regardless of the conditions under which they were cultured or cryopreserved in the originating laboratory (syn2724705). Cell lines that did not viably recover using this strategy were subsequently re-thawed onto irradiated murine embryonic feeders (MEFs) with hES media using standard conditions (Thomson et al., 1998) or were re-requested to be sent as live cultures on MEFs from the originating laboratory. In either of these cases the cells were subsequently transitioned into the feeder-free conditions prior to downstream analysis.

Standardizing Metadata Fields, Terms, Collection and Confirmation

Metadata information was initially provided by the originating laboratory, and was subsequently augmented with *in vitro* genetic and experimental characterization data of the line, and resubmitted to the originating lab for confirmation. For example, sex was confirmed with karyotype results and lines submitted by a common donor were identified from SNP calls derived from genotyping arrays and RNA-Seq. Specifically an identity-by-descent analysis was performed in PLINK (Purcell et al., 2007) first using the subset of individuals for which genotyping had been performed and then verified with SNP calls based on RNA-Seq from a larger subset of samples which identified three more samples with common donors. The source SNP array PLINK input files and merged VCF genotype file are available (syn2391784).

iPSC Cell Culture, Flow Cytometry and Molecular Sample Collection

iPSC were cultured in mTeSR1 (Stem Cell Technologies, Vancouver, Canada) on Matrigel (Corning Inc., Corning, New York) coated 6-well dishes (Nunc, Waltham,

Massachusetts) and subcultured with dispase using protocols adapted from the manufacturer (syn2724700).

For analysis of cell surface markers, cells were harvested with Accutase (Innovative Cell Technologies, San Diego, California), blocked with 2% IVIG and 1% HSA in PBS, and stained with the specific antibodies for 30 minutes at +4°C. For intracellular marker analysis, cells were fixed and permeabilized with the Becton Dickinson Fix/Perm Kit (BD Biosciences, San Jose, California) prior to addition of antibodies. Samples were subjected to flow cytometry acquisition on a MACSQuant cytometer (Miltenyi Biotech, San Diego, California) and analyzed using FlowJo software (FlowJo, Ashland, Oregon).

Samples for mRNA and miRNA analysis were prepared by removing culture media from and adding 1mL Trizol Reagent (Ambion, Carlsbad, California) per well and incubating for 1 minute. Trizol was pipetted several times, transferred to RNAse-free tubes, and stored at -80°C until extraction. Samples were extracted using manufacturer's recommended protocols using chloroform:isoamyl alcohol (49:1) followed by ethanol precipitation and split into 2 aliquots. Half the sample was pelleted and retained for miRNA-Seq analysis without further preparation. The other half was subjected to mRNA purification using PureLink Spin Cartridges (Life Technologies, Carlsbad, California). Samples for DNA analysis were prepared by removing culture media from the plate, scraping the cell layer with a cell scraper, and collecting in DPBS. The cell suspension was centrifuged for 1 minute at 1000xg and excess DPBS was removed from the pellet. The pellet was stored at -80°C until DNA was isolated.

In vitro and In vivo Pluripotency Analysis

Cells were harvested for RNA, DNA, and flow cytometry as described above. Detailed protocols are available in the Synapse database (syn2512369). Cells were additionally differentiated in embryoid body (EB) cultures for 17 days. In brief, iPSC were disaggregated into clumps and cultured in suspension on non-adherent culture dishes. On day 7, the EB were transferred to gelatin coated tissue culture dishes where they adhered, and grew out from the EB On day 17, the cultures were harvested for DNA and RNA extraction and analysis (syn2512370).

Each line was also subjected to an *in vivo* teratoma pluripotency assay. In brief, 80-90% confluent plates were incubated in dispase for 2-3 minutes, washed with DMEM/F-12 media, and scraped to retain small clumps. The clumps were pelleted and

resupended in 30% Matrigel in mTeSR1 for injection into NOD.Cg-*Prkdc*^{scid}*Il2rg*^{tm1Wjl}/SzJ mice (NSG mice) from the Cincinnati Children's Hospital Medical Center (CCHMC) Comprehensive Mouse and Cancer Core. Tumors were harvested when they reached ~1cm^3 and were fixed in 4% paraformaldehyde. Tumors were paraffin embedded, sectioned, stained in hematoxylin-eosin and evaluated in the clinical pathology core at CCHMC using standard procedures (syn3103753).

Stained histological sections (syn2882776) and a table with the pathologist observations and interpretations of sections from every line are available (syn2882785) with an example in Figure S1.

Molecular Karyotypic Copy Number Variation Analysis of iPSC

CNV were classified as benign, non-benign or clinically significant by Board Certified Clinical Cytogeneticists using the Cincinnati Children's Hospital clinical genetics database. To determine differential gene expression compared to observed non-benign CNV occurrence, at least a 50% change in expression from the mean of all PSC was required. As two distinct Illumina genotyping arrays were used for these studies (syn1773109), for all described comparisons between PSC derived from the same donor, we required that the results were obtained from a single genotyping array platform.

Data Exploration and Distribution

To provide comprehensive data and evaluations of each line, all associated data has been deposited into the Synapse online data repository (syn1773109). This includes metadata, *in vitro* and *in vivo* differentiation, qPCR, RNA- and miRNA-seq, CNV and DNA methylation high-throughput sequence and processed data. In addition to data files, associated analysis code, analytical methods and provenance tracking for all associated files have been included. To aid in interactive analyses of this data, customized data exploration options have been integrated into Synapse to facilitate gene-level analysis, cluster analysis and ToppGene functional enrichment analysis.

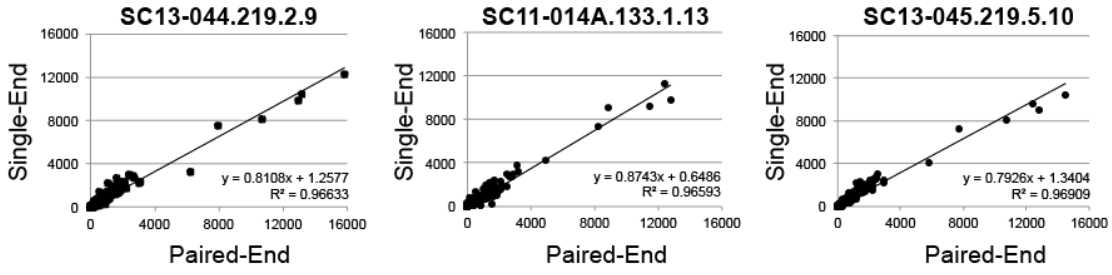
Genomic and Epigenetic Molecular Characterization

Details on the protocols and technologies employed are included in the supplementary

file, along with raw and processed data files in Synapse (syn1773109). To assay gene expression levels, total RNA was extracted and prepared with the Illumina TruSeq kit RNA V2. Single-end libraries were sequenced at a depth of between 10-30 million 50nt reads on an Illumina HiSeq 2000. miRNA expression levels were evaluated in a similar fashion, except that libraries were prepared with the Illumina TruSeq Small RNA kit and sequenced to only 1-4 million reads. Methylation was assessed with the Illumina HumanMethylation450 BeadChip with annotations provided by ENCODE (Encode Project Consortium, 2012). Two different assays were used for copy number variation (CNV) analysis: 21 cell lines were assayed with the Illumina CytoSNP-850K BeadChip, and 29 cell lines were assayed with the Illumina HD HumanOMNI-Quad BeadChip platform. For validation, 37 lines were analyzed using a TaqMan Low Density Array (TLDA) (Stem Cell Pluripotency Array, 4385344, Life Technologies) that evaluates a panel of stem cell and pluripotency marker genes (syn3107327).

Sensitivity Assessment of the PSC RNA-Seq

Prior to performing RNA-Seq on all PSC samples, we compared the use of deep (~50 million) sequenced paired-end (PE), stranded RNA-Seq (50nt reads) to that of more-shallow (~20 million) single-end (SE), non-stranded reads. Comparison of SE and PE for the same iPSC (n=2) and ESC (n=1), yielded an r^2 value of greater than 0.96 for all comparisons. For all of these comparisons, the number of PE gene measurements RPKM>1, was equivalent to the SE, lower depth samples, suggesting that these parameters are sufficiently to accurately quantify gene expression.



Sample ID	Sequencing Parameter	PSC	Number of Genes RPKM>1
SC13-045_PE.463.1.708	PE	iPSC	13,875
SC13-045.219.5.10	SE	iPSC	14,241
SC11-014_PE.457.1.703	PE	ESC	13,956
SC11-014A.133.1.13	SE	ESC	14,554
SC13-044_PE.457.1.705	PE	iPSC	13,774

SC13-044.219.2.9	SE	iPSC	14,213

RNA-Seq Data Processing

FASTQ files were aligned to the human genome build GRCh37 and University of California Santa Cruz (UCSC) transcriptome reference (Rosenbloom et al., 2014) using Tophat 2.0.9 (Kim et al., 2013). Processing scripts with exact parameters used and raw data are available and are linked together by provenance in Synapse (syn1773110). Gene-level Reads Per Kilobase per Million (RPKM) values and alternative splicing estimates were obtained using AltAnalyze (Emig et al., 2010). All samples were evaluated for a variety of quality control metrics including alignment percentage, proportion of exonic reads, and distribution of reads at the 5' and 3' ends of transcripts using the Cincinnati Children's Medical Center DNA sequencing core automated pipeline. Outlier samples with poor 5' to 3' ratios or other clear quality control issues (syn2332184) were flagged as FAIL and were not included in downstream covariate analyses. For AltAnalyze analysis, unique putative novel exons were determined from all Tophat junction alignments in AltAnalyze version 2.0.9 and analyzed for associated exon-read coverage using the BedTools function BAMtoBED, along with all AltAnalyze predicted exons (Ensembl 72 and UCSC annotated mRNAs). The resulting exon.bed and junction.bed files for AltAnalyze were used as input for downstream statistical and visualization analyses (clustering, PCA and network analysis) in AltAnalyze, using indicated stringency options for transcription, exon and reciprocal junction analyses (syn3105745). For splicing visualization, coverage plots were produced from the Broad's IGV Sashimi-Plot function (Robinson et al., 2011). Protein isoform, protein domain and miRNA functional prediction algorithms are described in detail at http://www.altanalyze.org . Gene-level Fragment per Kilobase per Million (FPKM) expression estimates were also obtained with Cufflinks2 (Trapnell et al., 2012) using corrections for sequence-specific bias and multi-mapped reads (syn2247799). The specific parameters for both Tophat and Cufflinks are stored in provenance records in Synapse (see for example syn2246887). Gene-level expression estimates based on the Transcript per Million (TPM) estimates were additionally calculated using the software eXpress (Roberts and Pachter, 2013) for redistribution and comparative analysis (syn3033755).

miRNA-Seq Data Processing

miRNA expression was quantified with mirExpress v2.1.4 (Wang et al., 2009) using the human miRBase 20.0 reference (syn2247097). The counts for each miRNA were further filtered and normalized. Specifically the 2306 annotated miRNAs were filtered down to 1302 miRNAs that had at least two reads aligned in more than 10% of the samples. Each sample was then normalized by dividing the read counts by the count of the 90th percentile miRNA in each sample followed by standardization by mean and standard deviation. The quality of the data was assessed by PCA analysis and hierarchical clustering. Samples were considered FAIL with low overall annotated miRNA read-depth prior to normalization (syn2701942). Some samples were re-run and concatenated but ultimately not included in our analyses as these samples ultimately were more correlated to each other. Differentially expressed miRNAs were assessed using a series of linear models where expression was a function of tissue of origin, gender and reprogramming vector. All p-values were assessed using an f-test and corrected using Benjamini-Hochberg false discovery rate correction.

DNA-Methylation Data Processing

DNA-methylation arrays were normalized with the minfi R package (Aryee et al., 2014) (syn2233188). Before processing, a single cell line was removed due to a grossly abnormal karyotype and 12 other samples were removed due to poor intensity. The 12 samples had log2 median intensities for both methylated and unmethylated probes below 10.5. The remaining samples were quantile normalized.

CNV Analysis

The Plug-in cnvPartition (v3.2.0) for GenomeStudio was used to identify CNVs from the SNP arrays. For this software, the default settings were used, with the exception of a minimum loss of heterozygosity (changed to 5 Mb) and minimum number of SNPs (changed to 10).

Statistical Analyses

For pairwise comparison group analyses, a moderated t-test p-value (Smyth, 2004) was calculated for all pairwise comparisons by a custom python script between all major comparable metadata variables (syn2246673). This script uses existing methods available in the software AltAnalyze. It excludes samples with abnormal karyotypes for

analysis, can consider both unique donors and non-unique donors, will perform miRNA target enrichment analysis and miRNA differentially expressed comparison analysis, compare methylation and expression profiles via a Pearson correlation for matching samples, and optionally filters genes based on a priori selected expressed genes. Genes with a moderated t-test $p<0.05$, following a Benjamini-Hochberg adjustment and fold change >1.5 were considered differentially expressed when all samples were considered. To ensure the detected differences were biologically significant, differentially expressed genes and miRNAs were required to be expressed at least 20% the mean expression in hESC derived embryoid bodies. To evaluate potential regulation by methylation, genes and miRNAs with anticorrelated (Pearson rho<-0.5) expression from comparison of the same cell lines were furthered evaluated. As a secondary filtering method, genes, exons, miRNAs and probes with a non-adjusted $p<0.05$ for unique donors only (substantially smaller dataset) were considered reliably differentially expressed. Percent spliced in (PSI) ratios for any reciprocally expressed exon-exon junctions or introns and junctions were obtained using AltAnalyze and used as input for this script. Enrichment analyses of miRNA targets from differentially expressed genes were performed using GO-Elite (Zambon et al., 2012). In addition, DESeq2 (Love et al., 2014) was used to perform a multivariate analysis from read counts (HTSeq, syn2822494)(Anders et al., 2015) for all analyzed RNA-Seq samples (syn2838880).

Additional Results

Molecular Karyotypic Copy Number Variation (CNV) Analysis of iPSC

The largest number of clinically significant CNV were observed on chromosomes X and 15. Of interest, 12 of 16 female iPSC (75%), and 1 of 3 (33%) female hESC had low levels of X-chromosome monosomy observed in <10% of cells, regardless of cell type of origin, vector type or reprogramming gene combinations. Though it did not reach statistical significance, CNV were observed at higher frequency in lines generated using integrating vectors (retroviral and lentiviral vectors) with 7 of 12 lines (58%) carrying clinically significant CNV compared to 13 of 32 lines (41%) generated using non-integrating vectors ($p=0.37$; Fisher's Exact Test).

Seven unique donors were used in the generation of multiple lines, with each set of lines reprogrammed from the same cell type of origin. iPSC from five of these donors

also used the same reprogramming vector and could therefore be used to identify reprogramming associated CNV (donors D001,2,3,4 and 10) (Table S6). For example, all three lines reprogrammed from donor D003 had the same 719kb mosaic duplication at 15q11.2 indicating that it is likely present in the donor's originating somatic cells (SC11-008, 9 and 10). In contrast, CNV were more variable in the lines generated from donors D001, 2, 3, 4 and 9. Of interest, the three lines from donor D002 had divergent CNV calls. One line (SC11-005) had duplications in both 20q11.21 and 6q21, another line (SC11-006) had only the duplication in 20q11.21, and the third line (SC11-007) had only the 6q21 duplication. This suggests that both duplications pre-existed in the original donor cells and were preserved in SC11-005, but one was lost in each of the other two lines. Overall, this indicates some level of instability in the lines from this donor.

References

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq-a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166-169.

Aryee, M.J., Jaffe, A.E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A.P., Hansen, K.D., and Irizarry, R.A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics *30*, 1363-1369.

Emig, D., Salomonis, N., Baumbach, J., Lengauer, T., Conklin, B.R., and Albrecht, M. (2010). AltAnalyze and DomainGraph: analyzing and visualizing exon expression data. Nucleic acids research *38*, W755-762.

Encode Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology *14*, R36.

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

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J.*, et al.* (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics *81*, 559-575.

Roberts, A., and Pachter, L. (2013). Streaming fragment assignment for real-time analysis of sequencing experiments. Nature methods *10*, 71-73.

Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nature biotechnology *29*, 24-26.

Rosenbloom, K.R., Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M.*, et al.* (2014). The UCSC Genome Browser database: 2015 update. Nucleic acids research.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J.*, et al.* (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology *25*, 1251-1255.

Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology *3*, Article3.

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

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols *7*, 562-578.

Whetzel, P.L., and Team, N. (2013). NCBO Technology: Powering semantically aware applications. Journal of biomedical semantics *4 Suppl 1*, S8.

Zambon, A.C., Gaj, S., Ho, I., Hanspers, K., Vranizan, K., Evelo, C.T., Conklin, B.R., Pico, A.R., and Salomonis, N. (2012). GO-Elite: a flexible solution for pathway and ontology over-representation. Bioinformatics *28*, 2209-2210.

Supplemental Figures

Figure S1, related to Figure 1: Standardized Preliminary Screening of iPSC lines. A) Normal female karyotype for SC11-010. B) Abnormal female karyotype for SC11-003 with 47,XX, del(8)p23,+12. C) Immunohistological staining with anti-OCT4 antibody of a teratoma with a poorly differentiated area that is confirmed to have undifferentiated cells by SC12-034. Histopathological analysis of Teratomas from iPSC line with representatives of D) mesoderm, endoderm and ectoderm. Immunohistological analysis of a teratoma from SC12-025 with E) OCT4, F) Alpha-feto protein, G) Neurofilament, and H) Muscle Specific Actin staining.

Figure S2, related to Figure 2: Copy Number Variation Analyses. A) Summary CNV by reprogramming vector (n=50). Bars in blue show non-benign CNV and in red show clinically significant CNV. Benign CNV are not shown. B) The mean of clinically significant CNV in hESC, iPSC generated with non-integrating or integrating vectors. None of the comparisons were significantly different by student's t-test: blastocyst vs integrating, p= 0.0891, integrating vs non-integrating, p=0.1734, blastocyst vs non-integrating, p=0.2405 (2-tailed, heteroscedastic). C) Non-Benign CNV partially observed in at least one of two iPSC from a single donor are shown.

Figure S3, related to Figure 3: Global Similarity of iPSC and hESC. A) Hierarchical clustering of gene expression differences present among hESC, iPSC and hESC derived EB by RNA-Seq. Genes with a 4 fold difference between at least 8 samples and correlated (rho>0.5 or rho<-0.5) with the expression of at least 10 other genes are shown. This filtering schema was used to enrich for differentially expressed genes with similar expression patterns that are shared across multiple lines. Relative expression calculated to the average of the EB and iPSC average expression for each gene. A singular value decomposition (SVD) analysis of the top three principal components for all RNA-Seq genes with a minimum RPKM of 5 and at least 500 reads/gene (n=5801) are displayed to the right of the heatmap. The same analysis workflow was run on B) DNA-methylation profiles with at least 3 samples with beta values less than 0.33 and at least 3 samples greater than 0.66 and C) microRNA-seq expression profiles with at least 50 reads per microRNA, to obtain correlated/anticorrelated probe or microRNA clusters. PCA plots were generated from the initial filtered sets, before correlated clusters were selected.

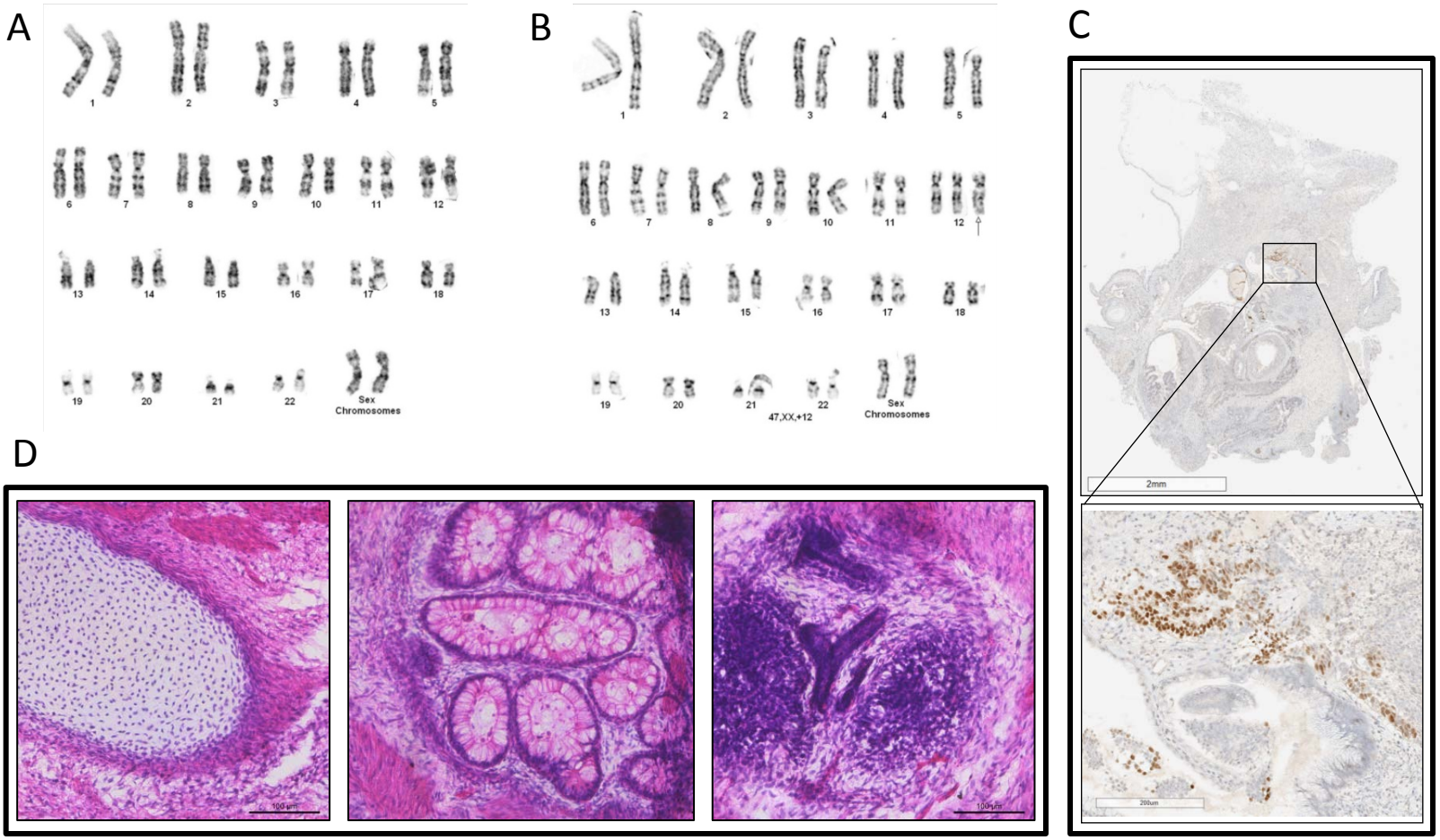
Detailed cell line data and expression values can be found in synapse: https://www.synapse.org/#!Synapse:syn1773109/files/

Figure S4, related to Figure 4: Global and Representative Reprogramming Specific Molecules. (A) Top-ranking DNA-methylation probes and corresponding genes for representative unique donor samples. Shapes are distinct by covariate. Anticorrelated gene expression and DNA-methylation are indicated by a red alpha and significant differentially expressed genes by an asterisk. B) The top ranking differentially expressed genes by pairwise comparison p-value for reprogramming associated variables extracted from the cell line metadata for representative unique donor samples. Measurements in red indicate the same parental genetic donor (D007). C) Correlated qPCR (TLDA) and RNA-Seq Gene Expression Profiles. Gene expression normalized to the mean of all evaluated pluripotent stem cells and single embryoid body are shown for the top 4 most correlated genes between TLDA and RNA-Seq.

Figure S5, related to Figure 5: Alternative Splicing Events. A) Hierarchical clustering of the top-ranking alternative splicing events for hESC versus EB. Values are reported as percent spliced in (PSI) from AltAnalyze normalized to the average of the hESC and EB average PSI. White indicates no PSI calculated due to insufficient detection of RNA isoforms for the indicated splicing events (drop-outs). B) Top reprogramming associated splicing events for representative unique donor samples.

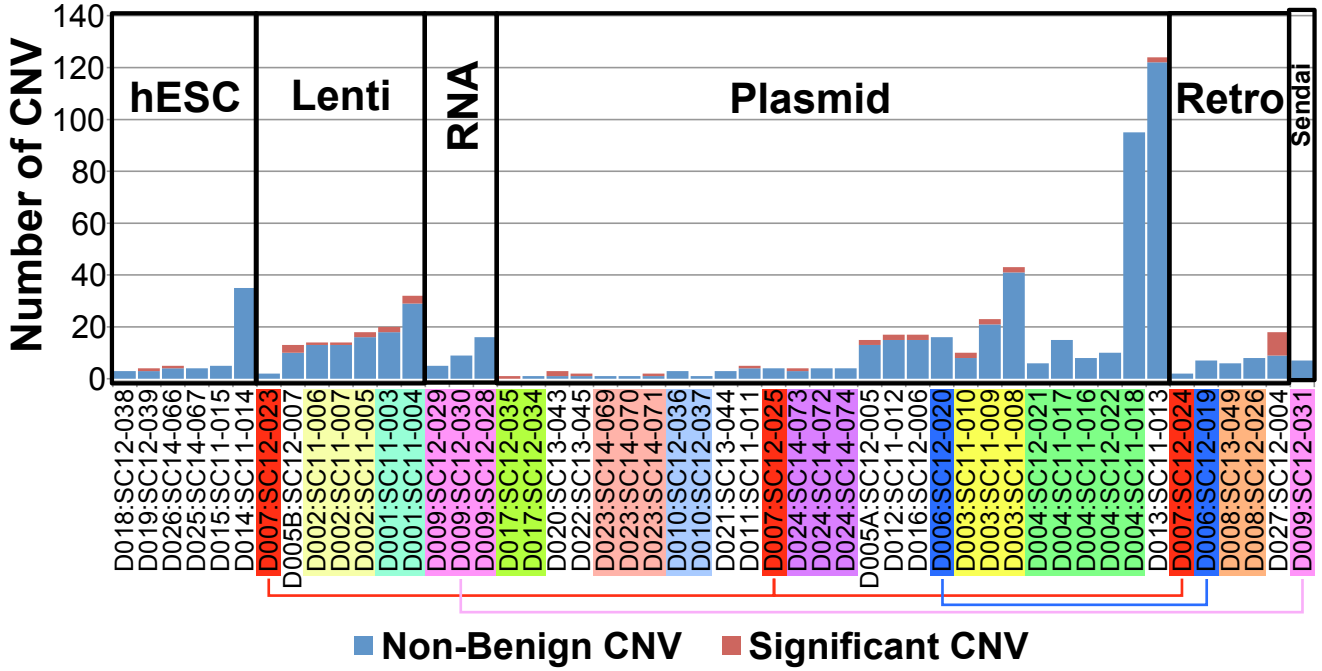
Figure S6, related to Figure 6: Sex-Associated Autosomal DNA-Methylation. A) Previously proposed X-chromosome inactivation specific Illumina 450k DNA-Methylation probes (n=3,279, Nazor et. al Cell Stem Cell 2012) and associated gene expression profiles. B) Unsupervised hierarchical clustering of the top differentially methylated autosomal methylation probes, filtering for any probe with at least one sample containing less than 0.3 beta and at least one other sample with greater than 0.6 beta (n=22,678). Probe cluster annotations are indicated below the heatmap. C) Biologically enriched (Benjamini-Hochberg adjusted $p<0.05$) categories from ToppGene corresponding to the 10 reported clusters from B) visualized in Cytoscape. D) DNA-methylation beta values for lineage directing transcription factors in high and low *XIST* female iPSC. E) Representative genes with expression anti-correlated to multiple measures of XCI in definitive endoderm (DE), mesoderm (Meso) and ectoderm (Ecto) directed differentiations of female PSC. F) Comparison of high and low *XIST* PSC derived teratomas based on the percent

positive quantification of muscle specific actin (MSA) or neurofilament (NF) staining.


A
B
C
D
E
F
G
H
Sex Chromosomes
47,XX,+12
2mm
100 μm
200um
300um

Supplemental Figure 2

A

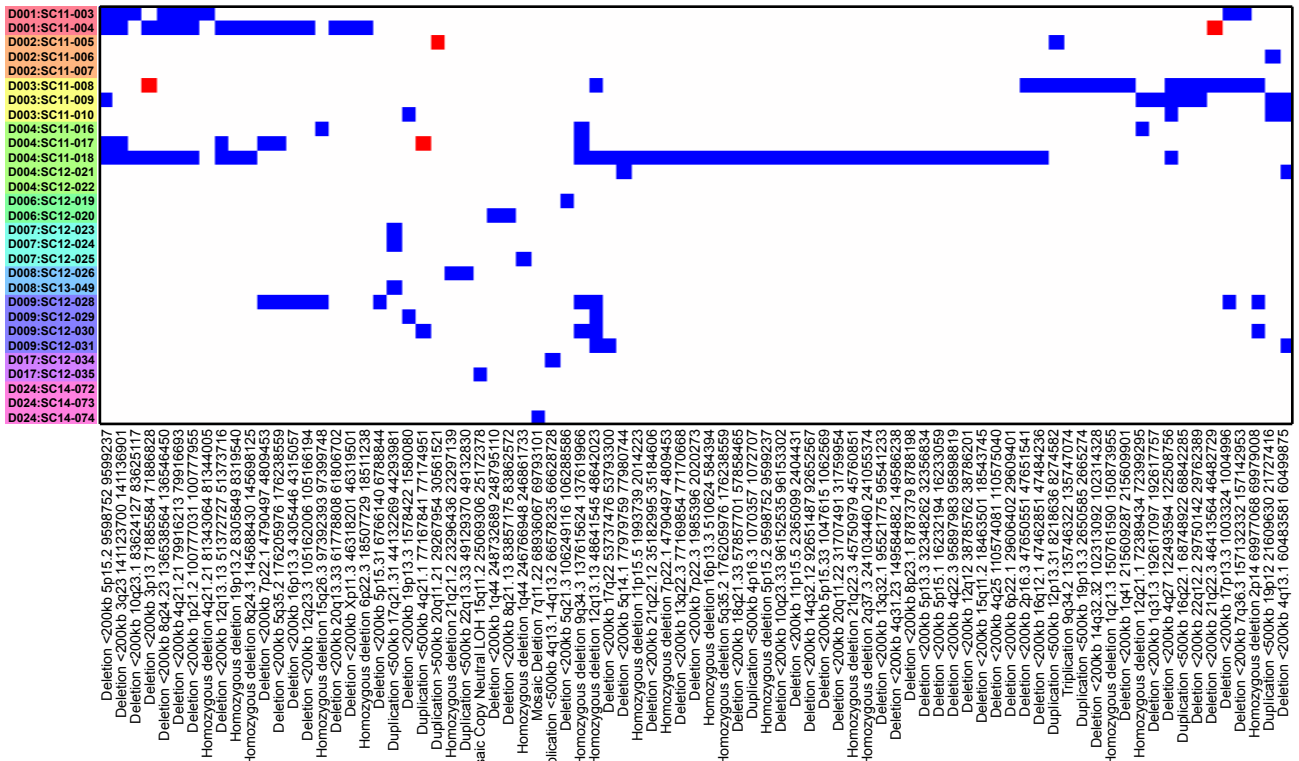


B

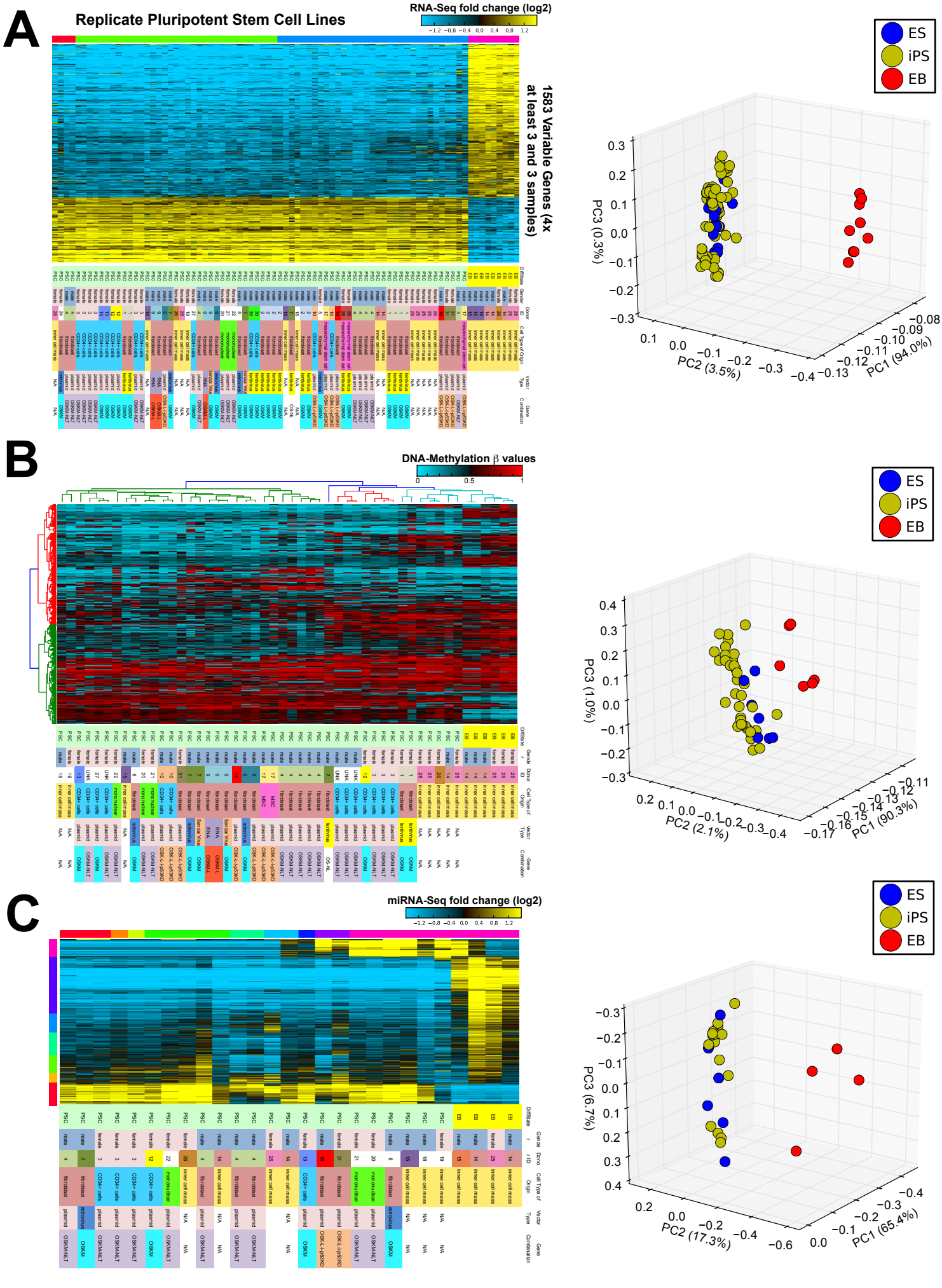
Summary of CNV by vector type.

Vector	Sig CNV	N	Mean Significant CNV	Standard Error	Range
Blastocyst	2	6	0.33	0.210819	0-1
Non-Integrating	21	28	0.75	0.73983	0-2
Integrating	21	12	1.75	0.15299	0-9

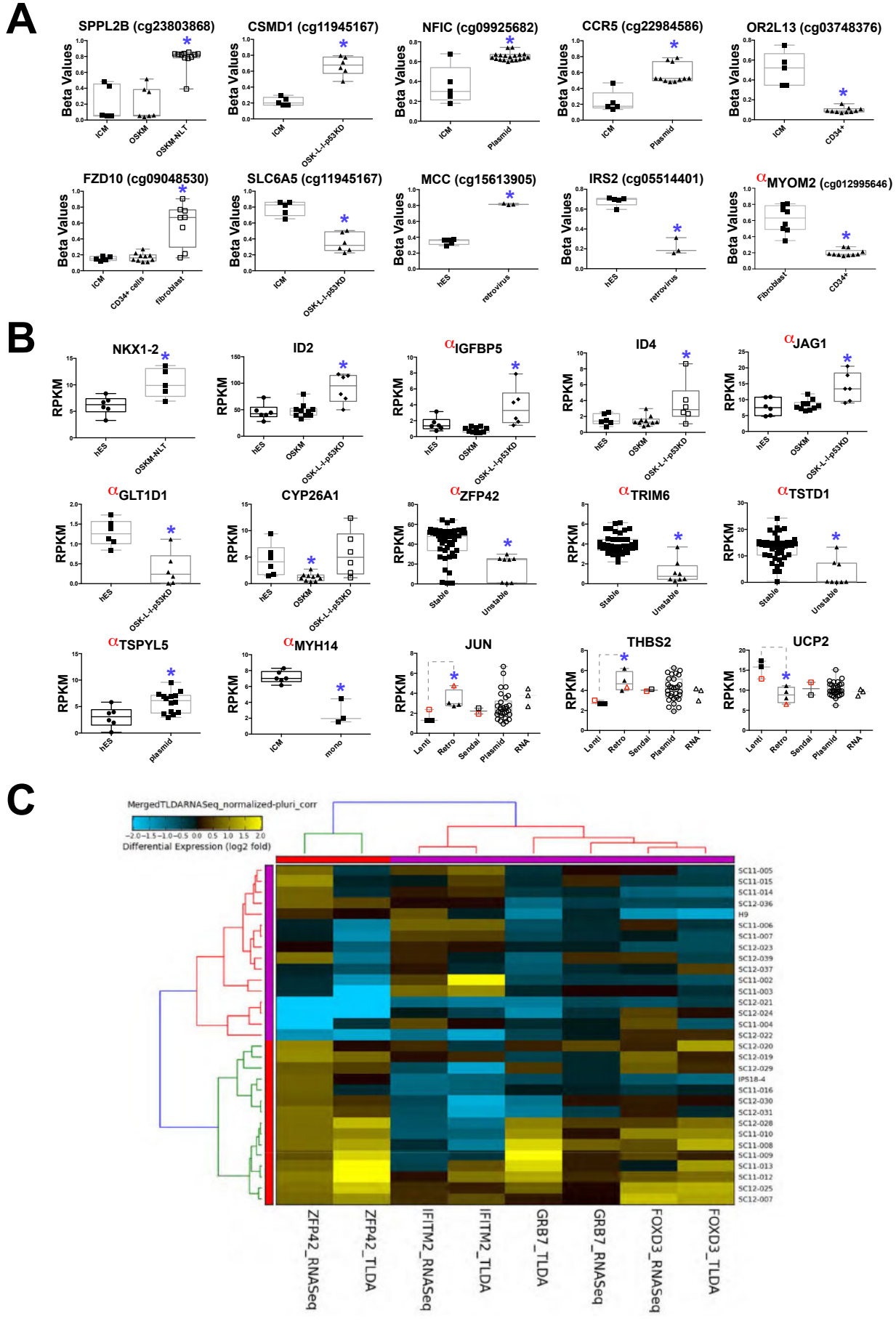
C



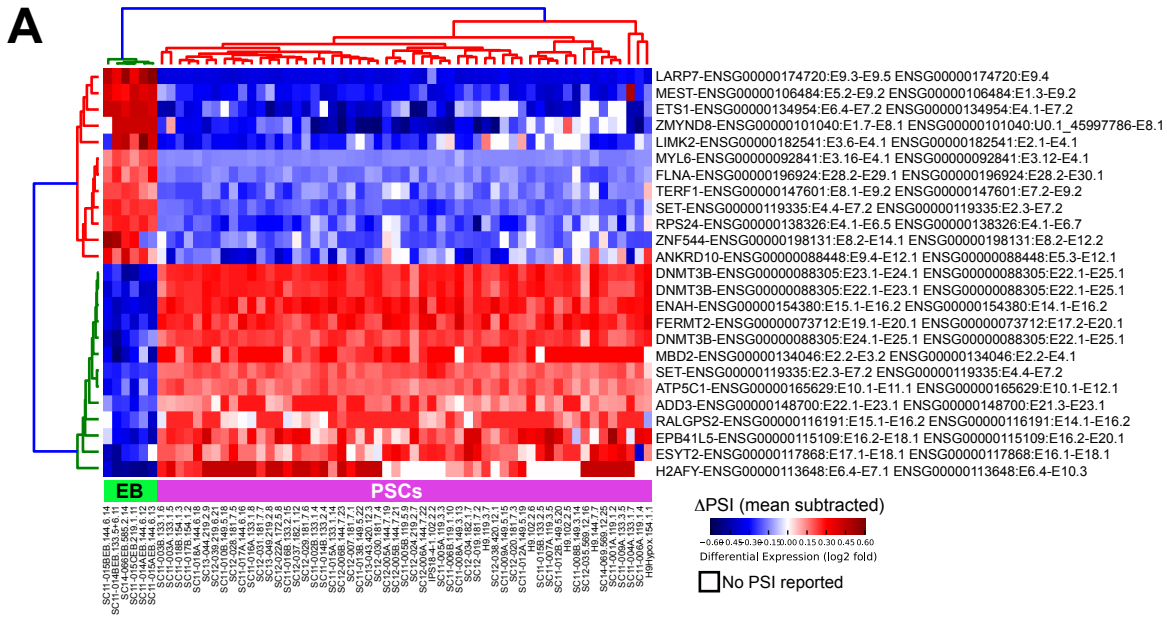
Supplemental Figure 3

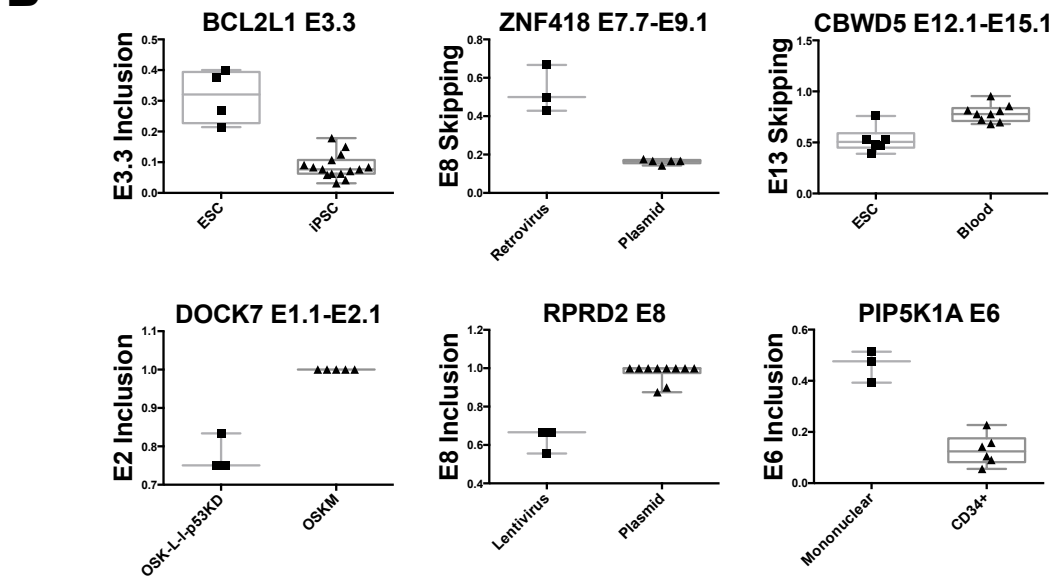


Supplemental Figure 4

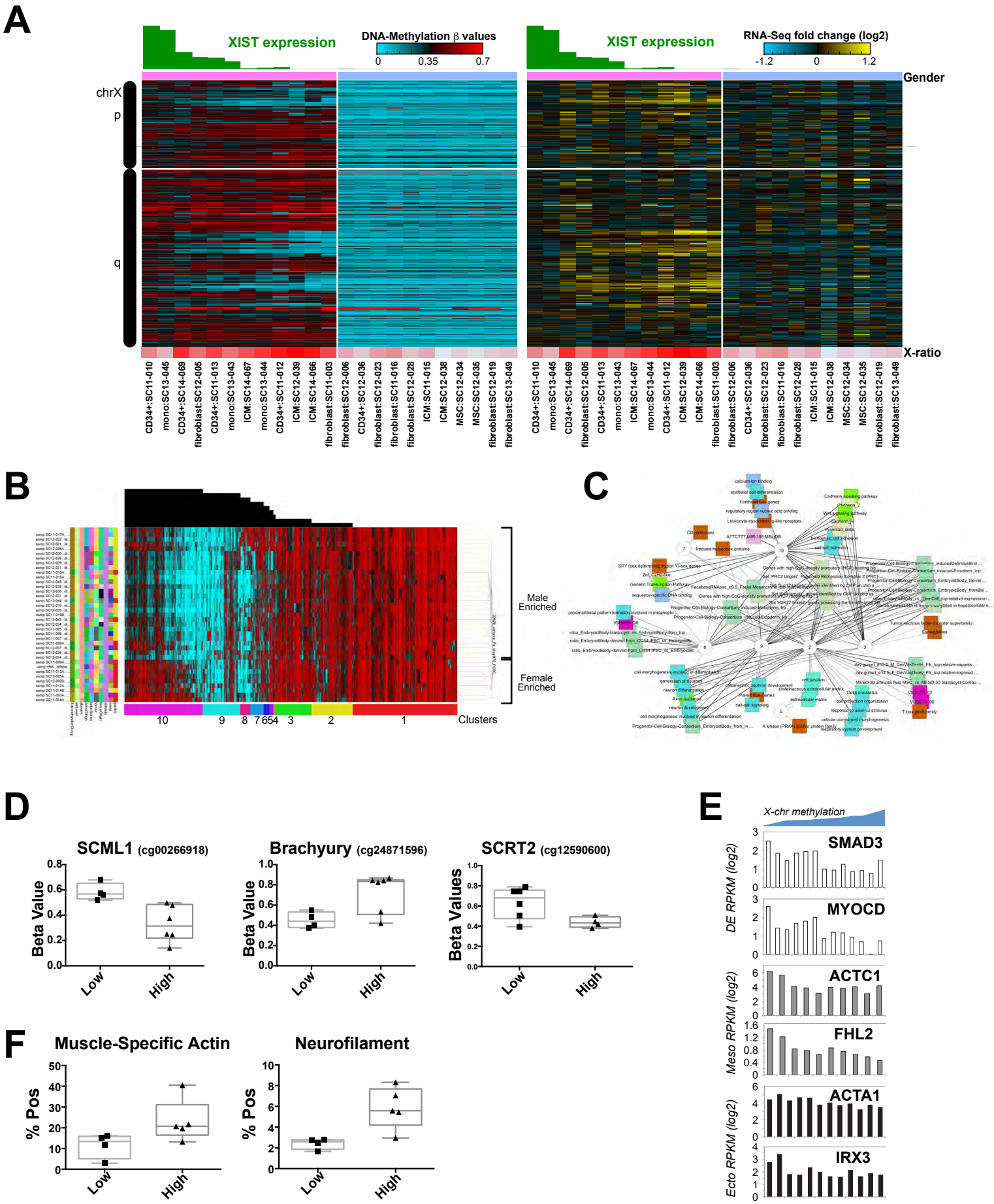


Supplemental Figure 5





Supplemental Figure 6



Supplementary Table 1: Cell line nomenclature, contributors and references (if previously published).

PCBC Cell Line Name	C4 Cell Line ID	Originating Lab ID	Principle Investigator	Other Significant Contributor	PMID or other reference info
PCBC01hsi2011070101	SC11-002	CHOP_WT1.1	Weiss, Gadue, & French	Jason A Mills	PMCID: 3778548
PCBC01hsi2011070102	SC11-003	CHOP_WT1.2	Weiss, Gadue, & French	Jason A Mills	PMCID: 3778548
PCBC01hsi2011070103	SC11-004	CHOP_WT1.3	Weiss, Gadue, & French	Jason A Mills	PMCID: 3778548
PCBC01hsi2011070104	SC11-005	CHOP_WT2.1	Weiss, Gadue, & French	Jason A Mills	PMCID: 3778548
PCBC01hsi2011070105	SC11-006	CHOP_WT2.2	Weiss, Gadue, & French	Jason A Mills	PMCID: 3778548
PCBC01hsi2011070106	SC11-007	CHOP_WT2.3	Weiss, Gadue, & French	Jason A Mills	PMCID: 3778548
PCBC03hsi2011080401	SC11-008	CBiPS-6.2	Elias Zambidis	N/A	PMCID: PMC3072973
PCBC03hsi2011080402	SC11-009	CBiPS-19.11	Elias Zambidis	N/A	PMCID: PMC3072973
PCBC03hsi2011080403	SC11-010	CBiPS-6.13	Elias Zambidis	N/A	PMCID: PMC3072973
PCBC03hsi2011080404	SC11-011	CBiPS-E5C3	Elias Zambidis	Tea Soon Park	PMCID: PMC3414503
PCBC03hsi2011080405	SC11-012	CBiPS-E12C1	Elias Zambidis	Tea Soon Park	PMCID: PMC3414503
PCBC03hsi2011080406	SC11-013	CBiPS-E17C6	Elias Zambidis	Tea Soon Park	PMCID: PMC3414503
PCBC02hse2011100705	SC11-014	WA01	James Thomson	N/A	9804556
PCBC02hse2011100706	SC11-015	WA24	James Thomson	N/A	9804556
PCBC02hsi2011100701	SC11-016	DF19-9-7T/DF19.7	James Thomson	Junying Yu	19325077
PCBC02hsi2011100703	SC11-017	DF4-3-7T.A/DF4.7	James Thomson	Junying Yu	19325077
PCBC02hsi2011100702	SC11-018	DF6-9-9T.B/DF6.9	James Thomson	Junying Yu	19325077
PCBC15hsi2011102602	SC12-003	virWTb	Bruce Conklin	Shiro Baba	PMID: 24509632, PMC4063274
PCBC15hsi2011102603	SC12-004	virWTa	Bruce Conklin	Shiro Baba	PMID: 24509632, PMC4063274
PCBC15hsi2012040401	SC12-005	epiWTb	Bruce Conklin	Yohei Hayashi	PMID: 24509632, PMC4063274
PCBC15hsi2012040402	SC12-006	epiWTc	Bruce Conklin	Yohei Hayashi	PMID: 24509632, PMC4063274
PCBC15hsi2012062201	SC12-007	virWTb	Bruce Conklin	Shiro Baba	PMID: 24509632, PMC4063274
PCBC05hsi2012061401	SC12-019	HFF12	Beverly Torok-Storb	Aravind Ramakrishnan	N/A
PCBC05hsi2012061402	SC12-020	niPSC	Beverly Torok-Storb	Aravind Ramakrishnan	N/A

PCBC02hsi2011100704	SC12-021	mND1	James Thomson	Guokai Chen	21478862
PCBC02hsi2012082101	SC12-022	MIRJT7-mND2-0-WB0119	James Thomson	Guokai Chen	21478862
PCBC16hsi2011111501	SC12-023	lenti-8.4.1	Jonathan Slack	Lucas V. Greder, James Dutton	23197849
PCBC16hsi2011111502	SC12-024	retro-20.1	Jonathan Slack	Lucas V. Greder, James Dutton	N/A
PCBC16hsi2011111503	SC12-025	Sendai-9-1	Jonathan Slack	Lucas V. Greder, James Dutton	23326500, 24485793
PCBC16hsi2011081101	SC12-026	kyba029	Michael Kyba	Abhijit Dandapat, Jakub Tolar	N/A
PCBC08hsi2012082303	SC12-027	BJ Epi 5	George Daley, Thorsten Schlaeger	Alexander DeVine	N/A
PCBC08hsi2012082304	SC12-028	BJ RiPS 1	George Daley, Thorsten Schlaeger	Andrew ttenger	N/A
PCBC08hsi2012082305	SC12-029	BJ RiPS 2	George Daley, Thorsten Schlaeger	Andrew ttenger	N/A
PCBC08hsi2012082306	SC12-030	BJ RiPS 3	George Daley, Thorsten Schlaeger	Andrew ttenger	N/A
PCBC08hsi2012082310	SC12-031	BJ Sendai 1	George Daley, Thorsten Schlaeger	Kelly Fitzgerald	N/A
PCBC08hsi2012082311	SC12-032	BJ Sendai 2	George Daley, Thorsten Schlaeger	Kelly Fitzgerald	N/A
PCBC08hsi2012082312	SC12-033	BJ Sendai 3	George Daley, Thorsten Schlaeger	Kelly Fitzgerald	N/A
PCBC08hsi2012082315	SC12-034	haMSC 17	George Daley, Thorsten Schlaeger	Fauza, Alexander DeVine	N/A
PCBC08hsi2012082316	SC12-035	haMSC 18	George Daley, Thorsten Schlaeger	Fauza, Alexander DeVine	N/A
PCBC08hsi2012082318	SC12-036	CD34+ 1	George Daley, Thorsten Schlaeger	Colin Sieff, Kelly Fitzgerald	N/A
PCBC08hsi2012082319	SC12-037	CD34+ 2	George Daley, Thorsten Schlaeger	Colin Sieff, Kelly Fitzgerald	N/A
PCBC08hse2012100901	SC12-038	CHB4	George Daley	Paul Lerou	18223642
PCBC08hse2012100902	SC12-039	CHB8	George Daley	Paul Lerou	18223642
PCBC10hsi2012051001	SC12-040	BJ iPSC	John Cooke	Sheena Abraham, Eduard Yakubov	N/A
PCBC02hsi2012090602	SC13-043	IISH1i-BM1	Igor Slukvin	Kejin Hu	21296996
PCBC02hsi2012090601	SC13-044	IISH2i-BM9	Igor Slukvin	Kejin Hu	21296996
PCBC02hsi2012090603	SC13-045	IISH3i-CB6	Igor Slukvin	Kejin Hu	21296996
PCBC16hsi2013040201	SC13-049	029 iPS clone 4	Michael Kyba	Abhijit Dandapat, Jakub Tolar	N/A
PCBC03hsi2013090602	SC13-059	E20C2	Elias Zambidis	Tea Soon Park	PMCID: PMC3414503

PCBC03hsi2013090603	SC13-060	E24C2	Elias Zambidis	Tea Soon Park	PMCID: PMC3414503
PCBC03hsi2013090604	SC13-061	E17C1	Elias Zambidis	Tea Soon Park	PMCID: PMC3414503
PCBC03hsi2013090605	SC13-062	E32C9	Elias Zambidis	Tea Soon Park	PMCID: PMC3414503
PCBC03hsi2013090606	SC13-063	E7C1	Elias Zambidis	Tea Soon Park	PMCID: PMC3414503
PCBC03hsi2013090607	SC13-064	E7C9	Elias Zambidis	Tea Soon Park	PMCID: PMC3414503
PCBC03hsi2013090608	SC13-065	E7C12	Elias Zambidis	Tea Soon Park	PMCID: PMC3414503
PCBC02hse2014030501	SC14-066	WA07	James Thomson	N/A	9804556
PCBC02hse2014030502	SC14-067	WA09	James Thomson	N/A	9804556
PCBC03hsi2014031101	SC14-069	4F CB-iPSC-MSC, LZ6-1	Elias Zambidis	Ludovic Zimmerlin	N/A
PCBC03hsi2014031102	SC14-070	4F CB-iPSC-MSC, LZ6-2	Elias Zambidis	Ludovic Zimmerlin	N/A
PCBC03hsi2014031103	SC14-071	4F CB-iPSC-MSC, LZ6-12	Elias Zambidis	Ludovic Zimmerlin	N/A
PCBC03hsi2014031104	SC14-072	4F CB-iPSC-MSC, LZ6+2	Elias Zambidis	Ludovic Zimmerlin	N/A
PCBC03hsi2014031105	SC14-073	4F CB-iPSC-MSC, LZ6+3	Elias Zambidis	Ludovic Zimmerlin	N/A
PCBC03hsi2014031106	SC14-074	4F CB-iPSC-MSC, LZ6+10	Elias Zambidis	Ludovic Zimmerlin	N/A

Supplementary Table 2: List of Assays Used to Characterize iPSC

Preliminary Assays
Colony and cellular morphology
Sterility
Mycoplasma
Flow cytometry (SSEA-1,SSEA-4, TRA-1-61, TRA-1-80, CD9, OCT-4)
Karyotype
Comprehensive Assays
RNA-seq
mi-RNA seq
dna methylation (450K CpG)
copy number variation (SNP)
Stem Cell Gene RT-PCR Expression Panel TLDA - 92 genes
Teratoma with Histopathological Analysis

!

Supplementary Table 3: Criteria for a line to be designated as stable.

Assay	Result	
Morphology	Normal	
Sterility (14 days)	No growth	
Mycoplasma	No mycoplasma	
Karyotype	20 of 20 [46,XY or XX]	
Flow Cytometry Analysis Self-Renewal Markers	SSEA-4	>95%
	TRA-1-60	>85%
	TRA-1-81	>85%
	OCT-4	>90%
	CD9	>90%
Flow Cytometry Analysis Differentiation Marker	SSEA-1	<5%

Supplementary Table 4: Summary of rationale for unstable lines determination and/or line not eligible for complete analysis.

Assay	Number of lines (%)
Non-Sterile	0
Mycoplasma Contamination	6 (9%)
Flow cytometry profile and/or differentiated morphology	6 (9%)
Abnormal Karyotype	7 (11%)
Quarantine	4 (6%)

Supplementary Table 5: Summary of karyotype abnormalities in iPSC lines.

Cell Line	Karyotype	Additional Information for Complex Karyotypes	Sex	Cell of Origin	Vector Type	Genes
SC11-002	47, XX, +12 [2] / 46, XX [18]		Female	fibroblast	lenti (Cre-excised)	OSKM
SC11-003	47, XX, +12 [15] / 47, XX, idem, del 8p23 [2] / 46, XX [2]	2 cells normal, 1 cell trisomy 5 (non-clonal), 17 cells trisomy 12, of those 17, 2 cells have a deletion at 8p23	Female	fibroblast	lenti (Cre-excised)	OSKM
SC12-003	46, XX [49]/ 47, XX +12 [1]		Female	fibroblast	retro	OSKM
SC12-004	46, XY, add 1q21	All 20 cells have material of unknown origin added to 1q21, creating a functional monosomy for distal 1q and partial trisomy for the region of genome added to 1q.	Male	fibroblast	retro	OSKM
SC12-026	47, XYY [2] / 46, XY [18]		Male	fibroblast	retro	OSKM
SC12-040	65-71,XXX,+add(X)(q28)[9],add(1)(q32)x2,-2[6],add(2)(q25)[3],add(3)(p13),add(3)(q21),-4[4],add(4)(21)[6],+6[7],-7[4],-9[8],add(9)(q22)[3],+10[4],+10[2],+12[2],-13,add(13)(p11.1),-14[3],-15[9],-15[3],-16[4],+17[7],add(17)(p11.2)[8],add(18)(q22)[8],20[3],+21[5],add(21)(22)[5],i(21)(q10)x2,+22[6],add(22)(11.2)[6],+mar1[8],+mar2[3],+1-2mar[9][cp10]	10 of 10 cells	Female	fibroblast	mRNA	OSKM
SC12-033	48, XY, +12, +20[2]/46, XY[37]		Male	BJ fibroblast	Sendai Vector	OSKM

Supplementary Table 6: Comparison of the CNV detected among the iPSC lines generated from common donors.

Unique Donor	Sex	Cell of origin	Vector	Genes	iPSC Line	Clinically Significant CNV
D001	Female	fibroblast	lenti	OSKM	SC11-003	low level mosaic monosomy for Xp22.33-Xq28
						interstitial duplication of 1.3Mb of 7q11.22
					SC11-004	interstitial duplication of 1.2Mb from 5q34
						interstitial duplication of 1.3Mb from 3q26.31
						low level mosaic monosomy for Xp22.33-Xq28
D002	Male	fibroblast	lenti	OSKM	SC11-005	1.3 Mb duplication at 20q11.21
						941Kb duplication at 6q21
					SC11-006	2.15 Mb Mosaic Duplication at 20q11.21
					SC11-007	1.69 Mb bp duplication at 6q21
D003	Female	UCB CD34+	plasmid	OSKM - NLT	SC11-008	719Kb Mosaic Duplication at 15q11.2
						Mosaic monosomy at Xp22.33-Xq28
					SC11-009	719Kb Mosaic Duplication at 15q11.2
						Mosaic monosomy at Xp22.33-Xq28
					SC11-010	719Kb Mosaic Duplication at 15q11.2
						Mosaic monosomy at Xp22.33-Xq28
D004	Male	fibroblast	plasmid	OSKM - NLT	SC11-016	No clinically significant chr imbalances
					SC11-017	No clinically significant chr imbalances
					SC11-018	No clinically significant chr imbalances
					SC12-021	No clinically significant chr imbalances
					SC12-022	No clinically significant chr imbalances
D007	Male	fibroblast	lenti	OS-NL	SC12-023	No clinically significant chr imbalances
			Retro	OSKM	SC12-024	No clinically significant chr imbalances
			plasmid	OSKM	SC12-025	No clinically significant chr imbalances
D009	Male	fibroblast	mRNA	OSKM-L	SC12-028	No clinically significant chr imbalances
			mRNA	OSKM-L	SC12-029	No clinically significant chr imbalances
			RNA	OSKM-L	SC12-030	No clinically significant chr imbalances
			Sendai Virus	OSKM	SC12-031	No clinically significant chr imbalances
D010	Male	BM CD34+	Episomal	OSK-L-I-p53KD	SC12-036	No clinically significant chr imbalances
				OSK-L-I-p53KD	SC12-037	No clinically significant chr imbalances

Supplememtary Table 7: Immunostaining analysis of teratomas generated from PSC lines with differentiated morphology in culture and independent control lines generated from the same donors.

Cell Line	PSC culture morphology	Donor ID	Cell Type	MSA	NF	AFP	OCT4	histopathologic analysis summary	Teratoma interpretation
SC12-021	Spontaneous differentiation	4	iPSC	+++	+	+++	-	Well differentiated pluripotent teratoma	Within the cut section of the tumor, there is benign well differentiated teratoma composed of solid and cystic areas and contain tissue types arising from all three primordial germ cell layers. Solid areas are composed of mixed population of moderately cellular neural tissue with glial cells and smaller primitive neuroepithelial cells in a background of neurofibrillary material interspersed by cartilage. In few places, solid areas are contain numerous columnar cells that frequently form glands and acini consistent with sebaceous glands. Cysts are lined by variety of cells ranging from cuboidal to simple columnar with multifocal cilia. There are many goblet cells are present in some cysts. Some cysts are lined by columnar epithelium with occasional cilia and interspersed with goblet cells (putative respiratory epithelium). Some cysts contain columnar epithelium that frequently folds into villi like structures with few goblet cells (putative intestine)
SC12-022	Undifferentiated	4	iPSC	++	++	+++	-	Well differentiated pluripotent teratoma	Within the cut section of the tumor, there is benign well differentiated teratoma composed of solid and cystic areas and contain tissue types arising from all three primordial germ cell layers. Solid areas are composed primarily of moderately cellular neural tissue with glial cells and smaller primitive neuroepithelial cells in a background of neurofibrillary material. Additional solid tissue include cartilaginous tissue. Cysts are lined by a variety of epithelial cells, ranging from squamous to cuboidal to simple and pseudostratified columnar with multifocal cilia. Few to many goblet cells are present in some cysts. Some cysts contain eosinophilic to amphophilic amorphous to fibrillar material.
SC12-023	Spontaneous differentiation	7	iPSC	+++	+	+	-	Well differentiated pluripotent teratoma	Within the cut section of the tumor, there is benign well differentiated teratoma composed of solid and cystic areas and contain tissue types arising from all three primordial germ cell layers. Solid areas are composed of mixed population of moderately cellular neural tissue with glial cells and smaller primitive neuroepithelial cells in a background of neurofibrillary material. Cysts are lined by variety of cells ranging from cuboidal to simple columnar with multifocal cilia. There are many goblet cells are present in some cysts. These cysts are vary from being respiratory epithelium to digestive tract epithelium.
SC12-025	Undifferentiated	7	iPSC	++	+	++	+ (a few small positivite columnar epithelial cells)	Well differentiated pluripotent teratoma	Within the cut section of the tumor, there is benign well differentiated teratoma composed of solid and cystic areas and contain tissue types arising from all three primordial germ cell layers. Solid areas are composed of mixed population of moderately cellular neural tissue with glial cells and smaller primitive neuroepithelial cells in a background of neurofibrillary material admixed with cartilage. Cysts are lined by variety of cells ranging from cuboidal to simple columnar with multifocal cilia. There are many goblet cells are present in some cysts. These cysts represent variety of putative differentiated tissues ranging from intestine, hair follicle, skin and respiratory epithelium. Frequently these cysts appear mixture of differentiated tissues containing eosinophilic to amphophilic amorphous to fibrillar material and occasionally keratin in lumen.
SC14-082	Spontaneous differentiation	unique	iPSC	ND	ND	ND	ND	Well differentiated pluripotent teratoma	Within the cut section of the tumor, there is benign well differentiated teratoma composed of solid and cystic areas and contain tissue types arising from all three primordial germ cell layers. Solid areas are composed primarily of moderately cellular neural tissue with glial cells and smaller primitive neuroepithelial cells in a background of neurofibrillary material. Additional solid tissue include cartilage, bone, hair, teeth and structures resembling acini glands. Cysts are lined by a variety of epithelial cells, ranging from squamous to cuboidal to simple and pseudostratified columnar with multifocal cilia. Few to many goblet cells are present in some cysts. Some cysts contain eosinophilic to amphophilic amorphous to fibrillar material.

Supplememtary Table 8: Immunostaining of teratomas with histopathologically identified undifferentiated regions and control teratomas generated from the same PSC line

Cell Line	PSC culture morphology	Donor ID	Cell Type	MSA	NF	AFP	OCT4	Teratoma histopathologic analysis	Teratoma Interpretation
SC12-034	Undifferentiated	17	iPSC	+++	++	++	++ (a few small strong OCT4 areas)	Poorly differentiated areas within one teratoma	Within the cut section of the tumor, there is benign well differentiated teratoma composed of solid and cystic areas and contain tissue types arising from all three primordial germ cell layers. Solid areas are composed of mixed population of moderately cellular neural tissue with glial cells and smaller primitive neuroepithelial cells in a background of neurofibrillary material. Multifocally, there are numerous capillaries frequently filled with erythrocytes. There are numerous cysts are lined by variety of cells ranging from simple columnar to stratified squamous epithelium. These cysts are vary from being digestive tract epithelium to skin.
SC12-034	Undifferentiated	17	iPSC	++	+	++	-	Well differentiated pluripotent teratoma	Within the cut section of the tumor, there is a benign cystic spaces that are lined by variety of cells ranging from cuboidal to simple squamous epithelium. These cystic spaces occasionally have mucin like material in the lumen. Majority of the solid space is composed of neuronal tissue and loosely arranged mesenchymal tissue. Part of the solid tissue components include primitive tooth like structures, cartilage and transitional epithelium.
SC11-013	Undifferentiated	13	iPSC	+	+	++	-	Well differentiated pluripotent teratoma	Within the cut section of the tumor, there is benign well differentiated teratoma composed of solid and cystic areas and contain tissue types arising from all three primordial germ cell layers. Solid areas are composed primarily of moderately cellular neural tissue with glial cells and smaller primitive neuroepithelial cells in a background of neurofibrillary material. Additional solid tissue include cartilaginous tissue. Cysts are lined predominantly by cuboidal to simple columnar with multifocal cilia. There are few goblet cells are present in some cysts.
SC11-013	Undifferentiated	13	iPSC	+	+	+	-	Poorly differentiated teratoma	Within the cut section of the tumor, there is benign poorly differentiated teratoma composed of solid (predominant) and cystic areas. There is no identifiable mesodermal tissue. Solid areas are composed primarily of moderately cellular neural tissue with primitive neuroepithelial cells with frequent resetting and pigmentation resembling developing eye. Cysts are lined by a cuboidal epithelial cells.
SC11-014	Undifferentiated	14	ESC	++	+	++	-	Well differentiated pluripotent teratoma	Within the cut section of the tumor, there is benign well differentiated teratoma composed of solid and cystic areas and contain tissue types arising from all three primordial germ cell layers. Solid areas are composed of mixed population of moderately cellular neural tissue with glial cells and smaller primitive neuroepithelial cells in a background of neurofibrillary material. Additional solid areas include cartilage & muscle tissue. There are numerous cysts are lined by variety of cells ranging from simple columnar to stratified squamous epithelium. These cysts are vary from being respiratory tract epithelium to skin
SC11-014	Undifferentiated	14	ESC	+++	+	+++	-	Poorly differentiated teratoma	Within the cut section of the tumor, there is benign poorly differentiated teratoma composed of solid and cystic areas and contain tissue types arising from all three primordial germ cell layers. Solid areas are composed primarily of moderately cellular neural tissue with neurons, glial cells and smaller, primitive neuroepithelial cells in a background of neurofibrillary material. Additional solid tissue include cartilaginous tissue. Cysts are lined by a variety of epithelial cells, ranging from squamous to cuboidal to simple and pseudostratified columnar with multifocal cilia. Few to many goblet cells are present in some cysts. Some cysts contain eosinophilic to amphophilic amorphous to fibrillar material.

Table S9: Summary of files and accession numbers available in Synapse. Enter the accession number into the search field at www.synapse.org to access the resource.

Resource	Accession #
Study Homepage	syn1773109
Cell Line Descriptions	
Cell line Metadata	syn2767694
Sample/Assay Metadata Folder	syn2247883
Experimental Protocols	
Top-level folder	syn2512369
Cell thaw/plate protocol	syn2724705
PSC culture protocol	syn2724700
Embryoid Body differentiation protocol	syn2512370
Preliminary Screening	
Teratoma Reports	syn2882776
Teratoma Report Spreadsheet	syn2882785
Karyotype full reports	syn2679104
Data (Raw and Normalized, potential additions after publication)	
Top-level folder	syn1773110
RNA-seq raw data	syn2247098
Exon and junction bed files (RNA-seq)	syn2246520
Gene expression normalized data	syn3034437
Alternative splicing PSI data	syn3091916
miRNA-seq raw data	syn2247097
microRNA normalized data	syn2701942
DNA methylation raw data	syn2653626
DNA methylation normalized data	syn2233188
Taqman Low-Density Array (TLDA) data	syn3107327
SNP array clinical reports	syn2679103
Compiled CNV data and Excel graphs	syn3105726
Analysis Scripts and Results (frozen at publication)	
Scripts Top-level Folder	syn2246673
Methylation Normalization script	syn2677441
Covariate Analyses Top Level	syn3094629
Gene Expression Covariate Results	syn3106206
Alternative Splicing Covariate Results	syn3106266
DNA Methylation Covariate Results	syn3106255
microRNA Covariate Results	syn3106244
Alternative Splicing hESC vs EB results	syn3106284
Ancestry Analysis	syn3107098
AltAnalyze Sample Group Predictions	syn3107554
Other Documents	
Manuscript Homepage	syn2731183
Suppl - Xchr_methylation-RNASeq_anticorrelated.xlsx	syn3107536
Suppl - XchrNazor_methylation-RNASeq.xlsx	syn3107535