Single-cell transcriptome sequencing of 18,787 human induced pluripotent stem cells identifies differentially primed subpopulations

Quan H. Nguyen[1*], Samuel W. Lukowski[1*], Han Sheng Chiu[1], Anne Senabouth[1], Timothy J. C. Bruxner[1], Angelika N. Christ[1], Nathan J. Palpant[1*], Joseph E. Powell[1,2*]

[1] Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland, 4072, Australia

[2] Queensland Brain Institute, University of Queensland, Brisbane, Queensland, 4072, Australia

[*] These authors contributed equally

Correspondence: Joseph E. Powell (j.powell@imb.uq.edu.au)

Summary

For pluripotent stem cells, transcriptional profiling is central to discovering the key genes and gene networks governing the undifferentiated state. However, the heterogeneity of cell states represented in pluripotent cultures have not been described at the transcriptional level. Since gene expression is highly heterogeneous between cells, single-cell RNA sequencing (scRNA-seq) can be used to increase our understanding of how individual pluripotent cells function. Here, we present the scRNA-seq results of 18,787 individual WTC CRISPRi human induced pluripotent stem cells. Four subpopulations were distinguishable on the basis of their pluripotent state including: quiescent (48.3%), proliferative (47.8%), early-primed for differentiation (2.8%) and late-primed for differentiation (1.1%). We identified novel genes and pathways defining each of the subpopulations and developed a multigenic prediction model to accurately classify single cells into subpopulations. This study provides a benchmark single cell dataset that expands our understanding of the cellular complexity of pluripotency.

Keywords

Single-cell RNA-seq, transcriptome, pluripotent stem cells, heterogeneity, gene expression, genomics, subpopulations, differentiation, pluripotency

Introduction

The transcriptome is a key determinant of the phenotype of a cell and regulates the identity and fate of individual cells. Much of what we know about the structure and function of the transcriptome comes from studies averaging measurements over large populations of cells, many of which are functionally heterogeneous. Such studies conceal the variability between cells and so prevent us from determining the nature of heterogeneity at the molecular level as a basis for understanding biological complexity. Cell-to-cell differences in any tissue or cell culture are an essential feature of their biological state and function.

In recent decades, the isolation of pluripotent stem cells, first in mouse followed by human (Evans and Kaufman, 1981; Thomson et al., 1998), and the more recent discovery of deriving pluripotent stem cells from somatic cell types (iPSCs) (Takahashi and Yamanaka, 2006), is a means to study lineage-specific mechanisms underlying development and disease to broaden our capacity for biological therapeutics (Palpant et al., 2017). Pluripotent stem cells are capable of unlimited self-renewal and can give rise to specialised cell types based on stepwise changes in the transcriptional networks that orchestrate complex fate choices from pluripotency into differentiated states.

In addition to individual published data, international consortia are banking human induced pluripotent stem cells (hiPSCs) and human embryonic stem cells (hESCs) and providing extensive phenotypic characterization of cell lines including transcriptional profiling, genome sequencing, and epigenetic analysis as data resources (Streeter et al., 2017; The Steering Committee of the International Stem Cell, 2005). These data provide a valuable reference point for functional genomics studies but continue to lack key insights into the heterogeneity of cell states that represent pluripotency.

While transcriptional profiling has been a central endpoint for analyzing pluripotency, the heterogeneity of cell states represented in pluripotent cultures has not been described at a global transcriptional level. Since each cell has a unique expression state comprising a particular collection of regulatory factors and target gene behavior, single-cell RNA Sequencing (scRNA-seq) can provide a transcriptome-level understanding of how individual cells function in pluripotency (Wen and Tang, 2016). These data can also reveal insights into the intrinsic

65 transcriptional heterogeneity comprising the pluripotent state. In this study, we provide the
66 largest dataset of single-cell transcriptional profiling of undifferentiated hiPSCs currently
67 available, which cumulatively amount to 18,787 cells across five biological replicates. Our
68 findings address the following hypotheses: (1) that transcriptional resolution at the single cell
69 level reveals gene networks governing specific cell subpopulations, (2) transcripts can exhibit
70 differences in gene expression heterogeneity between specific subpopulation of cells, and (3)
71 pluripotent cells form distinct groups or subpopulations of cells based on biological processes
72 or differentiation potential.
73
74 **Results**
75
76 ***Description of the parental hiPSC line, CRISPRi***
77 WTC-CRISPRi hiPSCs (Mandegar et al., 2016) were chosen as the parental cell line for this study.
78 These cells are genetically engineered with an inducible nuclease-dead Cas9 fused to a KRAB
79 repression domain **(Figure S1a)**. Transcriptional inhibition by gRNAs targeted to the
80 transcriptional start site is doxycycline-dependent and can be designed to silence genes in an
81 allele-specific manner. The versatility of this line provides a means to use this scRNA-seq data
82 as a parental reference point for future studies aiming to assess the transcriptional basis of
83 pluripotency at the single cell level. Cells were verified to have a normal 46 X,Y male karyotype
84 by Giemsa banding analysis before analysis by scRNA-seq (**Figure S1b**).
85
86 ***Single-cell RNA Sequence data***
87 After quality control of the sequencing data (Methods), we obtained 1,030,909,022 sequence
88 reads for 20,482 cells from five hiPSC single cell samples (**Table S1, Figure S2**), with 63-71%
89 confidently and uniquely mapped (mapping quality 255) to the human reference transcriptome
90 hg19 (ENSEMBL, release 75). We sequenced 19,937 cells from four samples to an average depth
91 of 44,506 reads-per-cell (rpc), while one sample consisting of 545 cells was sequenced to an
92 average depth of 318,909 rpc. On average, 2,536 genes and 9,030 Unique Molecular Identifiers
93 (UMIs) were detected per cell. We observed only a slight increase in the average number of

genes detected for cells sequenced at a greater depth (**Table S1, Figure S2f**) and no gain in the total number of genes detected for all cells in the whole sample, suggesting that an average of 44,506 rpc achieves close to sequencing saturation in our samples. Overall, we detected 16,064 unique genes, which were expressed in at least 1% of the total cells. We subsequently removed 1,738 cells due to a high percentage of expressed mitochondrial and/or ribosomal genes (**Methods, Table S2**), leaving a total of 18,787 high quality hiPSCs for further analysis. Following between-sample and between-cell normalisation, we observed no evidence for batch effects due to sample or sequencing run (**Figure 1a**, **Figure S3**).

Identification of four hiPSC subpopulations based on biological function

Using an unsupervised classification approach, we quantitatively assigned cells into clusters based on genome-wide transcriptome profiles **(Figure 1)**. This unbiased method identified four independent subpopulations of cells containing 48.3, 47.8, 2.8 and 1.1 percent of the 18,787 cells respectively. Importantly, after unsupervised clustering we also did not observe evidence for batch effects across any of the four cell populations identified (**Figure 1a**, **Table S3**, and interactive, gene-searchable figure at http://computationalgenomics.com.au/shiny/hipsc/), suggesting that the observed clustering was due to biological and not technical factors. By comparing gene expression between subpopulations, we identified four differentially expressed gene sets that distinguish each subpopulation from the remaining cells (**Figure 1c**, **Table S4**).

We initially examined transcript dynamics in these populations based on expression of known markers of pluripotency and lineage determination as previously described (Tsankov et al., 2015) (**Figure 2** and **Table S5**). Of the 18,787 cells, 99.8% expressed at least one of 19 pluripotency genes **(Table S6)**. Furthermore, genes with known roles in pluripotency had stronger expression across all subpopulations compared to genes involved in lineage determination (**Figure 2a-b, Tables S5 and S6**). For example, *POU5F1* (also known as *OCT4*), which encodes a transcription factor critically involved in the self-renewal of undifferentiated pluripotent stem cells was consistently expressed in 98.6% of cells comprising all four subpopulations (**Figure 2a-b**, **Tables S5 and S6**). Other known markers of pluripotency such as

123 *SOX2*, *NANOG* and *UTF1* were expressed across the subpopulations (**Figure 2a-b, Tables S5 and**
124 **S6**) but showed differences in expression heterogeneity, suggesting differences in the
125 pluripotent state across subpopulations (**Table S5**).
126
127 We sought to identify biological processes underlying classification of cell subpopulations by
128 firstly performing a statistical analysis to identify significantly differentially expressed genes
129 between subpopulations using a binomial test that accounts for both cell numbers and negative
130 binomial distribution of a gene's expression within a subpopulation (**Methods, Figure 1c, Table**
131 **S4)**. Differentially expressed genes with a fold-change significant at a Bonferroni-corrected *p*-
132 value threshold ($p < 3.1 \times 10^{-7}$) were evaluated for enrichment of functional pathways (**Tables**
133 **S7-S11**).
134
135 Cells classified in subpopulations one and two, which comprise 96.1% of total cells analyzed,
136 were distinguished from one another by significantly different expression levels of genes in
137 alternate pathways controlling pluripotency and differentiation (**Figure 3, Tables S7-S9**). The
138 Transcriptional Regulation of Pluripotent Stem Cells (TRPSC) pathway was consistently up-
139 regulated in cells classified as subpopulation two compared to subpopulation one (**Figure 3**,
140 **Table S9 and S12**). TRPSC is an auto-activation loop, which maintains expression of *POU5F1*,
141 *NANOG*, and *SOX2* at high levels. Complexes containing various combinations of these
142 transcription factors (Lam et al., 2012) can activate the expression of genes whose products are
143 associated with rapid cell proliferation, and also repress the expression of genes associated
144 with cell differentiation (Forristal et al., 2010; Guenther, 2011) (**Figure 3**). In particular, *POU5F1*,
145 *NANOG*, and *SOX2* are more highly expressed in subpopulation two (**Table S5**), and the
146 direction of differential expression of genes associated with cell proliferation and repression of
147 cell differentiation (Forristal et al., 2010; Guenther, 2011) is consistent with subpopulation two
148 containing cells that are more active in their self-renewal than cells in subpopulation one
149 (**Tables S5, S8 and S12**).
150

151 Our unbiased differential expression analysis identified *SALL4* (*Spal*-like transcription factor 4)
152 as significantly higher in subpopulation two than in subpopulation one (Table S12, *p-adjusted*:
153 $4.3x10^{-5}$). SALL4 is one of the key transcription factors that participates in controlling
154 transcriptional balance in pluripotent cells and suppressing differentiation (Miller et al., 2016)
155 (**Figure 3**). Specifically, SALL4 activates transcription of *POU5F1* and maintains pluripotency
156 (Yang et al., 2010b). Another upregulated gene (*p-adjusted*: 7.0×10^{-5}) in subpopulation two,
157 *ZIC1* (Zic family member 1), was identified by GeneMANIA analysis to be related to SALL4
158 through shared protein domains (**Figure S4**). Both ZIC1 and SALL4 were predicted by the STRING
159 database to interact with key pluripotency markers (**Figure S4**). Furthermore, ZIC1 and its
160 paralog ZIC3, a key member in the TRPSC pathway (**Figure 3**), are involved in maintaining the
161 undifferentiated state, for example in the case of neural precursor cells (Inoue et al., 2007).
162 Moreover, we also identified another differentially expressed gene, *NR6A1* (*p-adjusted*: $3.7 \times$
163 10^{-6}), which we predict is likely to participate in the TRPSC pathway since its paralog, *NR5A1*, is
164 among the key members of this pathway (**Figure S4)**. Based on these observations, we
165 hypothesise that in subpopulation two the three differentially expressed (DE) genes, *SALL4*,
166 *NR5A1* and *ZIC3*, cooperate with key pluripotency transcription factors (POU5F1-OCT4, SOX2,
167 and NANOG) to activate genes related to proliferation, but not genes involved in differentiation
168 (**Figure 3**).
169
170 Compared to subpopulations one and two, subpopulations three and four represent
171 pluripotent populations with significant down-regulation of key pluripotency network genes
172 (e.g. *NANOG*, *OTX2*, *SOX2* and *UTF1*) **(Figure 2a-b**). For subpopulation three, which comprises
173 2.8% of cells analyzed, Reactome pathway enrichment analysis of 2,534 DE genes between
174 subpopulations three and four showed the top pathways related to developmental signaling
175 and transcriptional regulation via chromatin modification (**Table S10**). Intracellular signaling
176 pathways that control cell proliferation, cell differentiation, and cell migration, such as EGFR,
177 PDGF, and NGF pathways (FDR $< 1.7 \times 10^{-6}$), were the top three most enriched pathways (**Table**
178 **S10**). Additionally, signalling pathways by FGFRs involved in differentiation were also
179 significantly enriched (FDR $< 3 \times 10^{-4}$). Comparing clusters three and one, signaling by TGF-beta,

180 and signaling by NODAL were in the top enriched pathways (FDR < 8 x 10^{-3}). Similarly, signaling
181 by NODAL (FDR < 0.04) (LeVincent et al., 2003) and pre-NOTCH processing (FDR < 0.04)
182 (Artavanis-Tsakonas et al., 1999), which are involved in cell fate decisions, were in the top
183 enriched pathways when comparing subpopulation three to subpopulations one and two (**Table**
184 **S10**). Thus, pluripotent cells in subpopulation three appear further advanced towards being
185 lineage primed compared to subpopulations one and two.
186
187 Pathway enrichment analysis by BiNGO in Cytoscape for all 1,706 DE genes in subpopulation
188 four vs. all other subpopulations (1.1% of analyzed cells) (**Table S11**) found the top enriched
189 pathways related to differentiation including genes involved in: gastrulation (FDR < 1.3 x 10^{-2})
190 and formation of primary germ layer (FDR < 1.4 x 10^{-2}); developmental process (FDR < 2.8 x 10^{-}
191 ^{3}); and cell differentiation (FDR < 1.2 x 10^{-2}); and more than 20 significantly (FDR < 5 x 10^{-2})
192 enriched pathways related to organogenesis (**Table S11**). Thus, although cells in subpopulation
193 four are still pluripotent, as indicated by the expression of pluripotent markers, they represent
194 cells at a late-primed state progressing toward differentiation.
195
196 Taken together, our transcriptional profiling of single cells revealed four subpopulations defined
197 by their pluripotency levels, cell proliferation, and potential for cell lineage commitment.
198 Subpopulation one pluripotent cells represent a quiescent pluripotent state, subpopulation two
199 represents proliferative pluripotent cells, subpopulation three as early-primed for
200 differentiation, and subpopulation four as late-primed for differentiation (**Figure 1b)**.
201
202 ***Cell classification can be predicted from transcriptome profiles***
203 Using the lists of differentially expressed genes, we built an unbiased predictor to identify the
204 pluripotency potential of a single cell. To avoid over-fitting the model due to co-expression of
205 genes, we used a variable selection regression model called LASSO (Tibshirani, 1996) to
206 estimate gene effects differentiating each cluster conditional upon the effects of other genes.
207 Using a 100-fold bootstrapping approach, we estimated the predictive accuracy of identifying a
208 cell in each of the four subpopulations (**Figure 4, Table S13**) (Tibshirani, 1996). To detect new

209 gene markers compared to the use of known pluripotency markers (**Table S5**), we applied
210 LASSO to selected sets of differentially expressed genes between one subpopulation compared
211 to the remaining three subpopulations. Consistently across four comparisons, we found that
212 our models based on the genes identified from our differential expression analysis had a higher
213 prediction accuracy, explained more deviance and performed with better sensitivity and
214 specificity (higher area under the curve - AUC) than those using known pluripotency and
215 differentiation markers (**Figure 3, Figure S5, Tables S13-S14**). We observed the highest
216 classification accuracy using genes identified using the LASSO model for cells in subpopulations
217 three and four than cells in subpopulations one and two, suggesting that these subpopulations
218 were more divergent from the remaining majority of the cell population. This observation
219 further supports the classification that subpopulations three and four are more primed to
220 differentiation than subpopulations one and two.
221
222 To confirm that genes selected by our LASSO analysis were also detected in other hiPSC lines,
223 we obtained open-access RNA-Seq transcript count data (tags per million - tpm) from the
224 Human Induced Pluripotent Stem Cells Initiative (HipSci) for 71 hiPSC lines derived from the skin
225 of normal individuals (Streeter et al., 2017). Consistently, we observed expression of LASSO
226 genes in 71 other independent hiPSC samples (**Figure S6**). Moreover, we observed high
227 correlation ($r > 0.85$) between the relative expression values among genes in our single-cell
228 dataset with those genes in the HipSci bulk RNA-seq dataset (**Figure S6c**). The high correlation
229 further confirms that the single-cell sequencing data accurately reflects the relative abundance
230 of transcripts.
231
232 ***Transcriptional heterogeneity revealed to be specific to cell subpopulations***
233 With the large scale dataset of 18,787 single cells and greater than 16,000 genes detected, we
234 were able to robustly analyze expression variation between different genes, different
235 subpopulations, and different cells (**Figure S7**). The inherently high heterogeneity of gene
236 expression in scRNA-seq data, especially for low abundant genes with a more frequent on-off
237 signal, may reduce the detection power of differential expression analysis between cells in

238 different subpopulations (Shalek et al., 2013). We observed more variation for subpopulations
239 with smaller number of cells (**Figure S7a**), and also found more variation for genes with low
240 expression (**Figure S7b**). Tagwise dispersion, which is expression variability for a gene across all
241 cells in a subpopulation, decreased when average expression increased (**Figure S7b**). The
242 difference in the level of heterogeneity of gene expression for cells in a given subpopulation
243 compared to other subpopulations is an important indicator of the relative dynamic cellular
244 activity of the subpopulation. The red line in Figure S7b shows the median dispersion of all
245 genes across all cells within a subpopulation, thereby representing the average expression
246 heterogeneity of the subpopulation. We found the median dispersion was higher in
247 subpopulations three and four than in subpopulations one and two (**Figure S7b**). This was
248 consistent with the observation that subpopulations three and four were closer to a
249 differentiated state compared to cells in subpopulations one and two, which were more
250 pluripotent based on transcriptome analysis.
251
252 **Discussion**
253
254 While methods to dissect cell subpopulations at single cell resolution such as FACS and
255 immunohistochemistry have been available, a comprehensive profiling of transcriptional
256 state(s) defining functionally distinct cell subpopulations comprising a 'homogenous' hiPSC cell
257 line have not be described (Wilson et al., 2015). To address this, we generated and analyzed the
258 largest hiPSC single-cell transcriptomics dataset to date, from five biological replicates of an
259 engineered WTC-CRISPRi hiPSC line (Mandegar et al., 2016). The 18,787 high-quality
260 transcriptomes, collectively expressing 16,064 genes, provided strong statistical power for
261 unbiased decomposition of this hiPSC population. Using a conservative statistical threshold, we
262 identified hundreds to thousands of genes that are differentially expressed between cells
263 enabling us to functionally categorize four distinct subpopulations. To our knowledge, this
264 dataset provides the first demonstration that a pure hiPSC population comprises multiple
265 subpopulations distinguishable by single-cell transcriptomics profiling.
266

267 Comparison of transcriptomes between subpopulations revealed gene regulatory networks
268 controlling the identity and pluripotency differentiation potential of cell subpopulations. Across
269 five separate biological replicates, we consistently found the existence of two main
270 subpopulations including a pluripotent-quiescent and pluripotent-proliferative subpopulation,
271 accounting for 96.1% of all cells profiled. Differentially expressed genes between the two
272 subpopulations were enriched for a cell proliferation gene network coordinately regulated by
273 *SALL4, ZIC1, NR5A1, POU5F1, SOX2,* and *NANOG*. The separation of two major subpopulations
274 on the basis of cell proliferation states may in part be explained by evidence that
275 reprogramming is commonly a stochastic process dependent on cell-proliferation rate (Hanna
276 et al., 2009). It remains to be determined whether these subpopulations generally reflect a
277 common feature of pluripotency in hESC or hiPSC populations, are a specific variable of iPSC
278 reprogramming, and whether cells in a single population transition between quiescent and
279 proliferative states on the basis of population dynamics over time.
280
281 Furthermore, we detected two smaller subpopulations (2.8% and 1.1% of the total cells) with
282 transcriptional signatures of pluripotency but primed to differentiation based on enriched
283 signaling pathways and gene ontologies related to lineage specification. Interestingly, from
284 analysis of expression heterogeneity within and between subpopulations, we found higher
285 variability in these two subpopulations compared to the remaining cells. This observation is
286 consistent with recent single-cell studies showing that the transition from pluripotency to
287 lineage commitment phase is characterized by high gene expression variability (Semrau et al.,
288 2016) and by the gradual destabilization of the pluripotent stem cell networks (Bargaje et al.,
289 2017).
290
291 We developed an approach that can be widely applied to optimize prediction models based on
292 single-cell transcriptomics data to classify cells into subpopulations at a high accuracy.
293 Identifying cell types is often based on immunostaining, FACS, or targeted PCR quantification of
294 a small number of markers (Tsankov et al., 2015). Here, we constructed an unbiased
295 classification model based on differential gene expression selected by LASSO regression

296 optimization procedure, without requirement for prior knowledge. We identified combinations
297 of a large number of genes not previously reported as new predictors for pluripotency and
298 showed that prediction models from differentially expressed genes performed better than
299 models built from known markers. Further functional genomics assays are required to
300 determine the role these gene networks play in defining the characteristics of these cell sub-
301 populations. The result is consistent with multiple genetic loci contributing small individual
302 effects to polygenic traits (Yang et al., 2010a). Therefore, the results support the use of an
303 unbiased and genome-wide approach to developing gene prediction models, which can be
304 applied to classify cell types and discover novel markers for a phenotype.
305
306 Despite the large number of cells sequenced, this study was limited in that only 3' mRNA was
307 sequenced, and thus there remained variation between cell populations that could not be
308 taken into account. Nevertheless, our aim was to deconvolute a 'homogenous' hiPSC
309 population, and inclusion of transcriptional sequence data from other RNA species in the future
310 will likely improve our ability to further delineate subpopulations of cells. Furthermore, we
311 confirmed that the genes selected were expressed in 71 HipSci datasets (Streeter et al., 2017),
312 and that relative expression level among genes was consistent between scRNA and bulk RNA
313 sequencing.
314
315 The parental cell line selected for this study, WTC-CRISPRi hiPSCs (Mandegar et al., 2016), is an
316 important system for targeted transcription inhibition, and is a key feature for functional
317 genomics studies that build on this dataset to study the biology of pluripotency. The results of
318 this study provide a benchmark single-cell transcriptional dataset for the field to expand our
319 understanding of the gene networks underlying cell subpopulations in pluripotency. Future
320 work is required to expand this analysis to multiple hiPSC and hESC lines to identify common
321 features of single-cell subpopulations in pluripotency. This study also provides a reference
322 dataset for functional studies using WTC-CRISPRi hiPSCs as a platform for inhibiting expression
323 of novel candidate regulators of pluripotency or differentiation.
324

325 **Methods**

326 ***Cell culture***

327 Undifferentiated human induced pluripotent stem cells (hiPSC; WTC-wild type C) were provided
328 courtesy of Bruce Conklin (UCSD) as previously described (Mandegar et al., 2016). Cells were
329 maintained on Vitronectin XF (STEMCELL Technologies, cat. no. 07180) and cultured in mTeSR1
330 (STEMCELL Technologies, cat. no. 05850). Cytogenetic analysis by Giemsa banding showed a
331 normal 46, XY male karyotype. For scRNA-seq, samples one and two were harvested from a
332 single plate using Versene, split into two technical replicates resuspended in Dulbecco's PBS
333 dPBS) (Life Technologies,cat. no. 14190-144) with 0.04% bovine serum albumin (Sigma; , cat.
334 no. A9418-50G) and immediately transported for cell sorting. For samples 3-5 cells were
335 harvested from individual plates using 0.25% Trypsin (Life Technologies, cat. no. 15090-046) in
336 Versene, neutralized using 50% fetal bovine serum (HyClone, cat. no. SH30396.03) in
337 DMEM/F12 (Life Technologies, cat. no. 11320-033), centrifuged at 1200 rpm for 5 minutes and
338 re-suspended in dPBS + 0.04% BSA.

339

340 ***Cell sorting***

341 Viable cells were sorted on a BD Influx cell sorter (Becton-Dickinson) using Propidium Iodide
342 into Dulbecco's dPBS + 0.04% bovine serum albumin and retained on ice. Sorted cells were
343 counted and assessed for viability with Trypan Blue using a Countess automated counter
344 (Invitrogen), and then resuspended at a concentration of 800-1000 cells/µL (8×10^5-1×10^6
345 cells/mL). Final cell viability estimates ranged between 80-93%.

346

347 ***Generation of single cell GEMs and sequencing libraries***

348 Single cell suspensions were loaded onto 10X Genomics Single Cell 3' Chips along with the
349 reverse transcription (RT) mastermix as per the manufacturer's protocol for the Chromium
350 Single Cell 3' Library (10X Genomics; PN-120233), to generate single cell gel beads in emulsion
351 (GEMs). Reverse transcription was performed using a C1000 Touch Thermal Cycler with a Deep
352 Well Reaction Module (Bio-Rad) as follows: 55°C for 2h; 85°C for 5min; hold 4°C. cDNA was
353 recovered and purified with DynaBeads MyOne Silane Beads (Thermo Fisher Scientific; Cat#

354 37002D) and SPRIselect beads (Beckman Coulter; Cat# B23318). Purified cDNA was amplified as
355 follows: 98°C for 3min; 12x(98°C for 15s, 67°C for 20s, 72°C for 60s); 72°C for 60s; hold 4°C.
356 Amplified cDNA was purified using SPRIselect beads and sheared to approximately 200bp with a
357 Covaris S2 instrument (Covaris) using the manufacturer's recommended parameters.
358 Sequencing libraries were generated with unique sample indices (SI) for each sample. Libraries
359 for samples 1-3 and 4-5 were multiplexed respectively, and sequenced on an Illumina
360 NextSeq500 (NextSeq control software v2.0.2/ Real Time Analysis v2.4.11) using a 150 cycle
361 NextSeq500/550 High Output reagent Kit v2 (Illumina, FC-404-2002) in stand-alone mode as
362 follows: 98bp (Read 1), 14bp (I7 Index), 8bp (I5 Index), and 10bp (Read 2).
363
364 ***Bioinformatics mapping of reads to original transcripts and cells***
365 Processing of the sequencing data into transcript count tables was performed using the Cell
366 Ranger Single Cell Software Suite 1.2.0 by 10X Genomics (http://10xgenomics.com/). Raw base
367 call files from the NextSeq500 sequencer were demultiplexed, using the *cellranger mkfastq*
368 pipeline, into sample-specific FASTQ files. These FASTQ files were then processed with the
369 *cellranger count* pipeline where each sample was processed independently. First, *cellranger*
370 *count* used STAR (Dobin et al., 2013) to align cDNA reads to the hg19 human reference
371 transcriptome, and aligned reads were filtered for valid cell barcodes and unique molecular
372 identifiers (UMI). Observed cell barcodes were retained if they were 1-Hamming-distance away
373 from an entry in a whitelist of known barcodes. UMIs were retained if they were not
374 homopolymers and had a quality score > 10 (90% base accuracy). *cellranger count* corrected
375 mismatched barcodes if the base mismatch was due to sequencing error, determined by the
376 quality of the mismatched base pair and the overall distribution of barcode counts. A UMI was
377 corrected to another, more prolific UMI if it was 1-Hamming-distance away and it shared the
378 same cell barcode and gene. *cellranger count* examined the distribution of UMI counts for each
379 unique cell barcode in the sample and selected cell barcodes with UMI counts that fell within
380 the 99th percentile of the range defined by the estimated cell count value. The default
381 estimated cell count value of 3000 was used for this experiment. Counts that fell within an
382 order of magnitude of the 99th percentile were also retained. The resulting analysis files for

383 each sample were then aggregated using the *cellranger aggr* pipeline, which performed a
384 between-sample normalization step and merged all 5 samples into one. Post-aggregation, the
385 count data was processed and analyzed using a comprehensive pipeline assembled and
386 optimized in-house as described below.
387
388 ***Preprocessing***
389 To preprocess the mapped data, we constructed a cell quality matrix based on the following
390 data types: library size (total mapped reads), total number of genes detected, percent of reads
391 mapped to mitochondrial genes, and percent of reads mapped to ribosomal genes. Cells that
392 had any of the 4 parameter measurements higher than 3x median absolute deviation (MAD) of
393 all cells were considered outliers and removed from subsequent analysis (**Table S2**). In addition,
394 we applied two thresholds to remove cells with mitochondrial reads above 20% or ribosomal
395 reads above 50% (**Table S2**). To exclude genes that were potentially detected from random
396 noise, we removed genes that were detected in fewer than 1% of all cells. Before
397 normalization, abundantly expressed ribosomal protein genes and mitochondrial genes were
398 discarded to minimize the influence of those genes in driving clustering and differential
399 expression analysis.
400
401 ***Data normalization***
402 Two levels of normalization were performed to reduce possible systematic bias between
403 samples and between cells. To reduce potential confounding effects caused by differences in
404 sequencing depths between five samples, a subsampling process (Zheng et al., 2017) was used
405 to scale the mean mapped reads (MMR) per cell of all samples down to the level of the sample
406 with the lowest MMR. For each sample, a binomial sampling process randomly selected reads
407 and UMIs for each gene in a cell at a sample-specific subsampling rate. The subsampling rate for
408 each sample was determined using the ratios of expected total reads (given the expected mean
409 reads per cell (minimum MMR of all samples), the known number of cells, and the fraction of
410 mapped reads to total reads) to the original total mapped reads (equation 1). Following

411 resampling, the MMRs for the five samples were scaled, while the expression data distribution
412 for genes in all cells of the sample was maintained.

413 $$Rate_i = min(MMR_j)*N_i*(ReadFraction_i)/Total_mapped_reads_i \quad (1)$$

414 where: $min(MMR_j)$ is the minimum MMR of all samples to be merged; N_j is the number of cells
415 in sample *j*; $ReadFraction_i$ is the ratio of confidently mapped reads in a cell to the total number
416 of reads detected for that cell in sample *i*; $Total_mapped_reads_i$ is the total number reads that
417 share the same cell barcode. For each gene in each cell, the sampling process of reads was
418 performed using the function: rbinom(length(reads), reads, subsample_rate) in R. This process
419 is more robust than standard scaling options because it takes into account unique read
420 information associated to mapped genes and cells.
421
422 To reduce cell-specific systematic bias, possibly caused by technical variation such as cDNA
423 synthesis, PCR amplification efficiency and sequencing depth for each cell, expression values for
424 all genes in a cell were scaled based on an estimated cell-specific size factor. Before
425 normalization, counts were log_2-transformed (by log_2(count+1)) to stabilize variance due to the
426 large range of count values (spanning 6 orders of magnitude, **Figure S1e**). To estimate the
427 scaling size factor for each cell, a deconvolution method (Lun et al., 2016) was applied for
428 summation of gene expression in groups of cells. This summation approach reduced the
429 number of stochastic zero expression of genes that are lowly expressed (higher dropout rates),
430 or genes that are turned on/off in different subpopulations of cells.

431 $$E(V_{ik}) = \lambda_{i0}\sum_{j\in S_k} \theta_j t_j^{-1} \quad (2)$$

432 Where S_k is a pool of cells, V_{ik} is the sum of adjusted expression value ($Z_{ij} = \theta_j*\lambda_{i0}$, where λ_{i0} is
433 the expected transcript count and θ_j is the cell specific bias) across all cells in pool V_k for gene *i*,
434 $\theta_j t_j^{-1}$ is the cell-specific scale factor for cell *j* (where t_j is the constant adjustment factor for
435 cell *j*).
436 The estimated size factor of a gene in in pool S_k, named as $E(R_{ik})$, is the ratio between the
437 estimated V_{ik} and the average Z_{ij} across all cells in the population. $E(R_{ik}) \approx \sum_{S_k} \theta_j t_j^{-1} C-1$ where C
438 $= N-1\sum_{S_0} \theta_j t_j^{-1}$, where *N* is the number of cells, S_0 represents all cells and is a constant for the
439 whole population and thus can be set to unity or ignored. The cell pools were sampled using a

440 sliding window on a list of cells ranked by library size for each cell. Four sliding windows with
441 20, 40, 60, and 80 cells were independently applied and results were combined to generate a
442 linear system that can be decomposed by QR decomposition to estimate $\theta_j t_j^{-1}$ size factor for
443 each of the cell. The normalized counts are the results of taking the raw counts divided by cell-
444 specific normalized size factors.
445
446 ***Analyzing transcriptional heterogeneity in 18,787 single cells***
447 To assess transcriptional heterogeneity among cells and genes, we first removed potential
448 variation due to technical sources by the subsampling process and the cell-specific
449 normalization as described above. Depending on experimental designs, an additional step using
450 a generalized linear model (GLM) to regress out other potential confounding factors can be
451 included. After reducing technical variation via normalization, we calculated the coefficient of
452 variation and expression dispersion of each gene across all cells. For cell-to-cell variation, we
453 first performed principal component analysis (PCA) on general cell data, which included percent
454 counts of the top 100 genes, total number of genes, percent of mitochondrial and ribosomal
455 genes. To investigate variation between genes, the distribution of dispersion across a range of
456 expression values was calculated (equation 3). This approach is useful because technical
457 variation often appears greater in lowly expressed genes than in more abundant genes (Shalek
458 et al., 2013). Denoting x_i as the vector of expression values (in cpm) for gene *i* across all cells,
459 we use the following formula to compute coefficient of variation:
460 $$\text{Coefficient of variation} = \text{sqrt}(\text{var}(x_i))/\text{mean}(x_i) \quad (3)$$
461 We estimated the BCV (biological coefficient of variation) with an empirical Bayesian approach
462 to estimate dispersion between genes and between samples (McCarthy et al., 2012). Common
463 dispersion (shared dispersion value of all genes), trended dispersion (mean dispersion trend for
464 lowly expressed genes to abundant genes), and gene-specific dispersion was estimated to
465 reflect variation of all genes across the whole population (**Figure S7**).
466
467 ***Dimensionality reduction***

468 After merging five samples, preprocessing, and normalizing the dataset was scaled to *z*-
469 distribution and PCA was performed for dimensionality reduction using the prcomp function in
470 R (McCarthy et al., 2016). To assess PCA results, we examined the top genes that were most
471 correlated to PC1 and PC2, and the distribution of cells and percent variance explained by the
472 top five PCs. Importantly, the optimal number of PCs explaining the most variance in the
473 dataset was determined using a Scree test calculated by the fa.parallel function in the psych
474 package. The fa.parallel was run based on expression data for the top variable genes.
475
476 Cells are represented using t-SNE (t-distributed Stochastic Neighbor Embedding) and diffusion
477 map (van der Maaten and Hinton, 2008). We applied the RtSNE package v0.1.3 on the
478 normalized expression data (16,064 genes x 18,787 cells) to calculate a three-dimensional t-SNE
479 projection dataset (a 16,064 cells x three t-SNE dimensions), which was then combined with
480 other data types to display cells on two- and three-dimensional t-SNE plots.
481
482 ***Clustering***
483 We developed an unsupervised clustering method, *i.e.* without taking into account any
484 predetermined parameters. Using the cell-PCA eigenvector matrix, agglomerative hierarchical
485 clustering using Ward's minimum distance option was applied to construct a dendrogram tree,
486 which contains all cells grouped into multiple layers of branches of similar cells, based on
487 transcriptome profiles. To determine the number of subpopulations, branches of the
488 dendrogram were pruned by a Dynamic Tree Cut method, which does not employ a constant
489 (supervised) height cutoff, but dynamically performs top-down iterative decomposition and
490 combination of clusters from larger to smaller neighbouring clusters until the number of
491 clusters becomes stable (Langfelder et al., 2008). Cluster information was then represented in t-
492 SNE graphs.
493
494 ***Differential expression analysis***
495 To select genes that distinguish subpopulations, we performed pairwise differential expression
496 analysis between cells in pairs or groups of clusters by fitting a general linear model and using a

497 negative binomial test as described in the DESeq package (Anders and Huber, 2010). Each cell
498 was considered as one biological replicate in each cluster. We found that the shrinkage
499 estimation of dispersion approach used in DESeq produced stable estimation of scale factors for
500 genes and cells between clusters and was more conservative in detecting differentially
501 expressed genes, especially when comparing subpopulations with a larger number of cells, such
502 as subpopulation one and two, to subpopulations with small cell numbers, such as three and
503 four. Specifically, DESeq detected fewer DE genes that expressed highly in a small proportion of
504 cells in a subpopulation while remaining cells in that subpopulation had zero or very low
505 expression. Significantly differentiated genes were those with Bonferroni adjusted *P*-values less
506 than 5% ($p < 3.1 \times 10^{-7}$).
507
508 ***LASSO regression analysis***
509 To develop predictive models based on single cell transcriptomics data, we applied Least
510 absolute shrinkage and selection operator (LASSO) procedure, to choose gene predictors for
511 classifying cells into one of the four subpopulations. Briefly, penalized logistic regression was
512 applied to fit a predictor matrix containing expression values of top differentiated genes in all
513 cells (or a subsample of a randomly selected 10 percent of the total cells) and a response vector
514 assigning cells into one of the subpopulations (dichotomous variable). The LASSO procedure
515 optimizes the combination set of coefficients for all predictors in a way that the residual sum of
516 squares is smallest for a given lambda value (Friedman et al., 2010). In other words, the LASSO
517 procedure identified an optimal combination of genes (predictors) and fitted a logistic
518 regression model, in which expression values of the selected genes were predictors and the
519 binary labels of cells were response variables. The fitted model could either explain the highest
520 deviance (compared to the full model) or classify cells to subpopulations with the lowest 10-
521 fold classification error. The glmnet R package was applied to select top genes that contributed
522 to classifying cells into each subpopulation (Tibshirani et al., 2012). The LASSO model was
523 trained using one subsampled dataset and then evaluated in a new, non-overlapping
524 subsampled dataset. Prediction accuracy was estimated by applying the trained model, with
525 only selected predictors and their corresponding coefficients, on a new set of randomly

526 sampled cells that were not used in the model training dataset. Bootstrap was used to calculate
527 classification accuracy for 100 iterations.
528
529 ***Pathway and gene functional analysis***
530 To functionally characterize the four subpopulations, we performed a network analysis using
531 significant DE genes between cells within a subpopulation and the remaining cells, or between
532 cells in pairs of subpopulations. We used Cytoscape to apply three main programs: GeneMANIA
533 (Warde-Farley et al., 2010), with a comprehensive background database containing 269
534 networks and 14.3 million interactions, the Reactome functional interaction network analysis, a
535 reliably curated protein functional network (Wu et al., 2010), and the STRING protein-protein
536 interaction database (Szklarczyk et al., 2015).
537
538 **Author contributions**
539 J.E.P and N.J.P designed the study, acquired funding and led analysis. S.W.L. , H.S.C. , T.J.C.B.
540 and A.N.C. performed experiments. Q.H.N, S.W.L., A.S performed analysis. All authors wrote
541 and edited the manuscript.
542
543 **Acknowledgments**
544 Sequencing was performed by the Institute for Molecular Bioscience Sequencing Facility at the
545 University of Queensland. This work was supported by the Australian National Health and
546 Medical Research Council (NHMRC) grant APP1083405 and APP1107599.

547 **References**

548
549 Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol.
550 *11*.
551
552 Artavanis-Tsakonas, S., Rand, M.D., and Lake, R.J. (1999). Notch Signaling: Cell Fate Control and Signal
553 Integration in Development. Science *284*, 770-776.
554
555 Bargaje, R., Trachana, K., Shelton, M.N., McGinnis, C.S., Zhou, J.X., Chadick, C., Cook, S., Cavanaugh, C.,
556 Huang, S., and Hood, L. (2017). Cell population structure prior to bifurcation predicts efficiency of
557 directed differentiation in human induced pluripotent cells. Proc. Natl. Acad. Sci. U. S. A. *114*, 2271–
558 2276.
559
560 Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and
561 Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.

562
563 Evans, M.J., and Kaufman, M.H. (1981). Establishment in culture of pluripotential cells from mouse
564 embryos. Nature *292*, 154-156.
565
566 Forristal, C.E., Wright, K.L., Hanley, N.A., Oreffo, R.O., and Houghton, F.D. (2010). Hypoxia inducible
567 factors regulate pluripotency and proliferation in human embryonic stem cells cultured at reduced
568 oxygen tensions. Reproduction *139*, 85-97.
569
570 Friedman, J.H., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models
571 via Coordinate Descent. J Stat Softw *33*, 22.
572
573 Guenther, M.G. (2011). Transcriptional control of embryonic and induced pluripotent stem cells.
574 Epigenomics *3*, 323-343.
575
576 Hanna, J., Saha, K., Pando, B., van Zon, J., Lengner, C.J., Creyghton, M.P., van Oudenaarden, A., and
577 Jaenisch, R. (2009). Direct cell reprogramming is a stochastic process amenable to acceleration. Nature
578 *462*, 595-601.
579
580 Inoue, T., Ota, M., Ogawa, M., Mikoshiba, K., and Aruga, J. (2007). Zic1 and Zic3 regulate medial
581 forebrain development through expansion of neuronal progenitors. J. Neurosci. *27*, 5461-5473.
582
583 Lam, C.S., Mistri, T.K., Foo, Y.H., Sudhaharan, T., Gan, H.T., Rodda, D., Lim, L.H., Chou, C., Robson, P.,
584 Wohland, T.*, et al.* (2012). DNA-dependent Oct4-Sox2 interaction and diffusion properties characteristic
585 of the pluripotent cell state revealed by fluorescence spectroscopy. Biochem. J. *448*, 21-33.
586
587 Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the
588 Dynamic Tree Cut package for R. Bioinformatics *24*, 719-720.
589
590 LeVincent, S.D., Dunn, N.R., Hayashi, S., Norris, D.P., and Robertson, E.J. (2003). Cell fate decisions within
591 the mouse organizer are governed by graded Nodal signals. Genes Dev. *17*, 1646-1662.
592
593 Lun, A.T.L., Bach, K., and Marioni, J.C. (2016). Pooling across cells to normalize single-cell RNA
594 sequencing data with many zero counts. Genome Biol. *17*, 75.
595
596 Mandegar, Mohammad A., Huebsch, N., Frolov, Ekaterina B., Shin, E., Truong, A., Olvera, Michael P.,
597 Chan, Amanda H., Miyaoka, Y., Holmes, K., Spencer, C.I.*, et al.* (2016). CRISPR Interference Efficiently
598 Induces Specific and Reversible Gene Silencing in Human iPSCs. Cell Stem Cell *18*, 541-553.
599
600 McCarthy, D.J., Campbell, K.R., Lun, A.T.L., and Wills, Q.F. (2016). scater: pre-processing, quality control,
601 normalisation and visualisation of single-cell RNA-seq data in R. bioRxiv.
602
603 McCarthy, D.J., Chen, Y., and Smyth, G.K. (2012). Differential expression analysis of multifactor RNA-seq
604 experiments with respect to biological variation. Nucleic Acids Res. *40*, 4288-4297.
605
606 Miller, A., Ralser, M., Kloet, S.L., Loos, R., Nishinakamura, R., Bertone, P., Vermeulen, M., and Hendrich,
607 B. (2016). Sall4 controls differentiation of pluripotent cells independently of the Nucleosome
608 Remodelling and Deacetylation (NuRD) complex. Development *143*, 3074-3078.
609

610 Palpant, N.J., Pabon, L., Friedman, C.E., Roberts, M., Hadland, B., Zaunbrecher, R.J., Bernstein, I., Zheng,
611 Y., and Murry, C.E. (2017). Generating high-purity cardiac and endothelial derivatives from patterned
612 mesoderm using human pluripotent stem cells. Nat. Protocols *12*, 15-31.
613
614 Semrau, S., Goldmann, J., Soumillon, M., Mikkelsen, T.S., Jaenisch, R., and van Oudenaarden, A. (2016).
615 Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic
616 stem cells. bioRxiv.
617
618 Shalek, A.K., Satija, R., Adiconis, X., Gertner, R.S., Gaublomme, J.T., and Raychowdhury, R. (2013). Single-
619 cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature *498*.
620
621 Streeter, I., Harrison, P.W., Faulconbridge, A., Flicek, P., Parkinson, H., and Clarke, L. (2017). The human-
622 induced pluripotent stem cell initiative—data resources for cellular genetics. Nucleic Acids Res. *45*,
623 D691-D697.
624
625 Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth,
626 A., Santos, A., Tsafou, K.P.*, et al.* (2015). STRING v10: protein–protein interaction networks, integrated
627 over the tree of life. Nucleic Acids Res. *43*, D447-D452.
628
629 Takahashi, K., and Yamanaka, S. (2006). Induction of Pluripotent Stem Cells from Mouse Embryonic and
630 Adult Fibroblast Cultures by Defined Factors. Cell *126*, 663-676.
631
632 The Steering Committee of the International Stem Cell, I. (2005). The International Stem Cell Initiative:
633 toward benchmarks for human embryonic stem cell research. Nat Biotech *23*, 795-797.
634
635 Thomson, J.A., Itskovitz-Eldor, J., Shapiro, S.S., Waknitz, M.A., Swiergiel, J.J., Marshall, V.S., and Jones,
636 J.M. (1998). Embryonic Stem Cell Lines Derived from Human Blastocysts. Science *282*, 1145-1147.
637
638 Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical
639 Society. Series B (Methodological) *58*, 267-288.
640
641 Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., and Tibshirani, R.J. (2012). Strong
642 rules for discarding predictors in lasso-type problems. Journal of the Royal Statistical Society. Series B,
643 Statistical methodology *74*, 245-266.
644
645 Tsankov, A.M., Akopian, V., Pop, R., Chetty, S., Gifford, C.A., Daheron, L., Tsankova, N.M., and Meissner,
646 A. (2015). A qPCR ScoreCard quantifies the differentiation potential of human pluripotent stem cells. Nat
647 Biotech *33*, 1182-1192.
648
649 van der Maaten, L., and Hinton, G.E. (2008). Visualizing data using t-SNE. J. Mach. Learn.Research *9*,
650 2579–2605.
651
652 Warde-Farley, D., Donaldson, S.L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C.,
653 Kazi, F., Lopes, C.T.*, et al.* (2010). The GeneMANIA prediction server: biological network integration for
654 gene prioritization and predicting gene function. Nucleic Acids Res. *38*, W214-220.
655
656 Wen, L., and Tang, F. (2016). Single-cell sequencing in stem cell biology. Genome Biol. *17*, 71.
657

658 Wilson, N.K., Kent, D.G., Buettner, F., Shehata, M., Macaulay, I.C., Calero-Nieto, F.J., Sanchez Castillo, M.,
659 Oedekoven, C.A., Diamanti, E., Schulte, R.*, et al.* (2015). Combined Single-Cell Functional and Gene
660 Expression Analysis Resolves Heterogeneity within Stem Cell Populations. Cell Stem Cell *16*, 712-724.
661
662 Wu, G., Feng, X., and Stein, L. (2010). A human functional protein interaction network and its application
663 to cancer data analysis. Genome Biol. *11*, R53.
664
665 Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C.,
666 Martin, N.G., Montgomery, G.W.*, et al.* (2010a). Common SNPs explain a large proportion of the
667 heritability for human height. Nat. Genet. *42*, 565-569.
668
669 Yang, J., Gao, C., Chai, L., and Ma, Y. (2010b). A novel SALL4/OCT4 transcriptional feedback network for
670 pluripotency of embryonic stem cells. PLoS ONE *5*, e10766.
671
672 Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D.,
673 McDermott, G.P., Zhu, J.*, et al.* (2017). Massively parallel digital transcriptional profiling of single cells.
674 Nat Commu *8*, 14049.
675
676

Figures

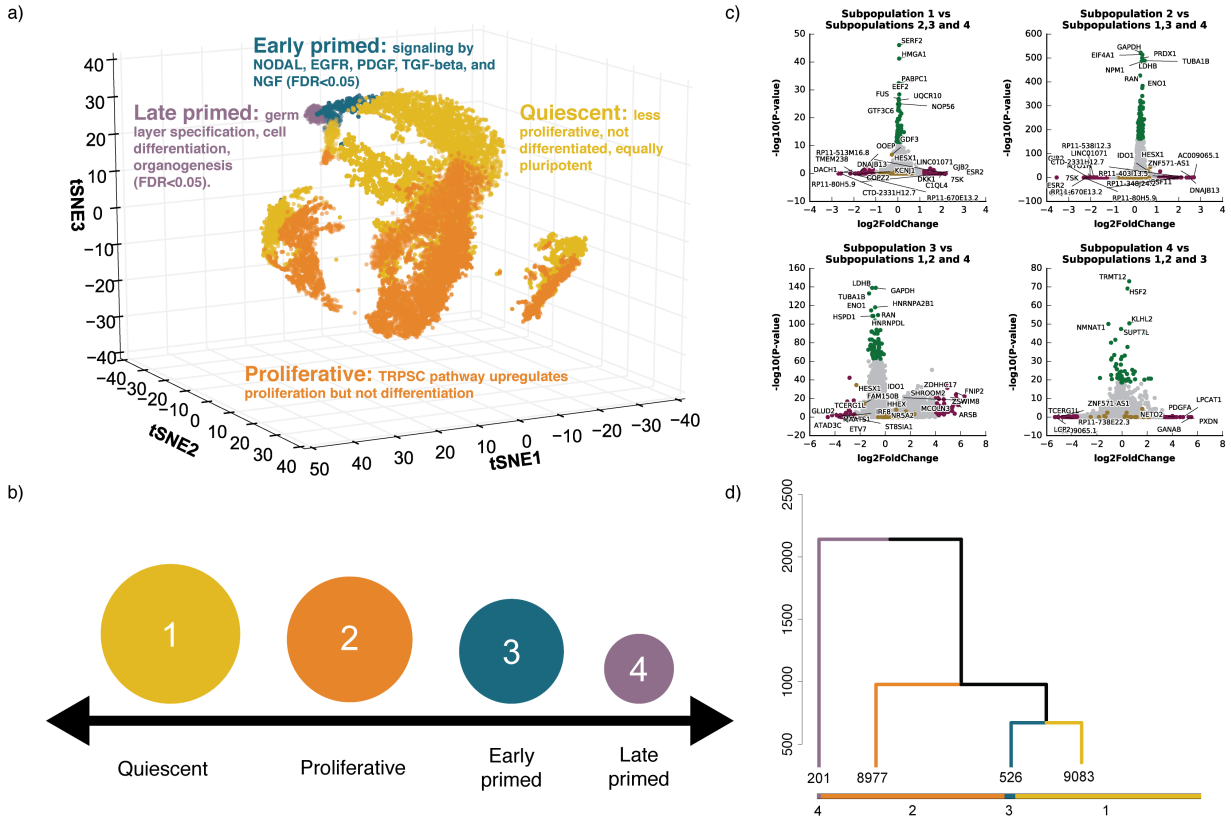

Figure 1. Identification of four cell subpopulations from 18,787 hiPSC cells, sequenced from five biological replicates. (a) Three-dimensional t-SNE distribution of cells based on gene expression value. Each point represent a single cell in three-dimensional space. A t-SNE transformation of the data was used for positioning cells, while four cell subpopulation labels (marked by different colors) represent results from clustering, and are independent of t-SNE data transformation (see http://computationalgenomics.com.au/shiny/hipsc/ for interactive, searchable figure). Pathway analysis based on differential expression identified functional properties that distinguish each subpopulation. **(b)** Four pluripotent subpopulations functionally separated from a homogeneous hiPSC population. **(c)** The top significantly differentially expressed genes of cells in a subpopulation compared to cells in the remaining three subpopulations. Genes denoted with orange points are known naive and primed markers. Genes represented with blue and purple points are those in the top 0.5% highest logFC or -log(*P-value*) respectively. **(d)** Unsupervised clustering of all cells into four subpopulations. The dendrogram tree displays distance and agglomerative clustering of the cells. Each branch represents one subpopulation. The clustering is based on a dynamic tree cut that performs a bottom-up merging of similar branches. The number of cells in each of the four subpopulations are given below branches.

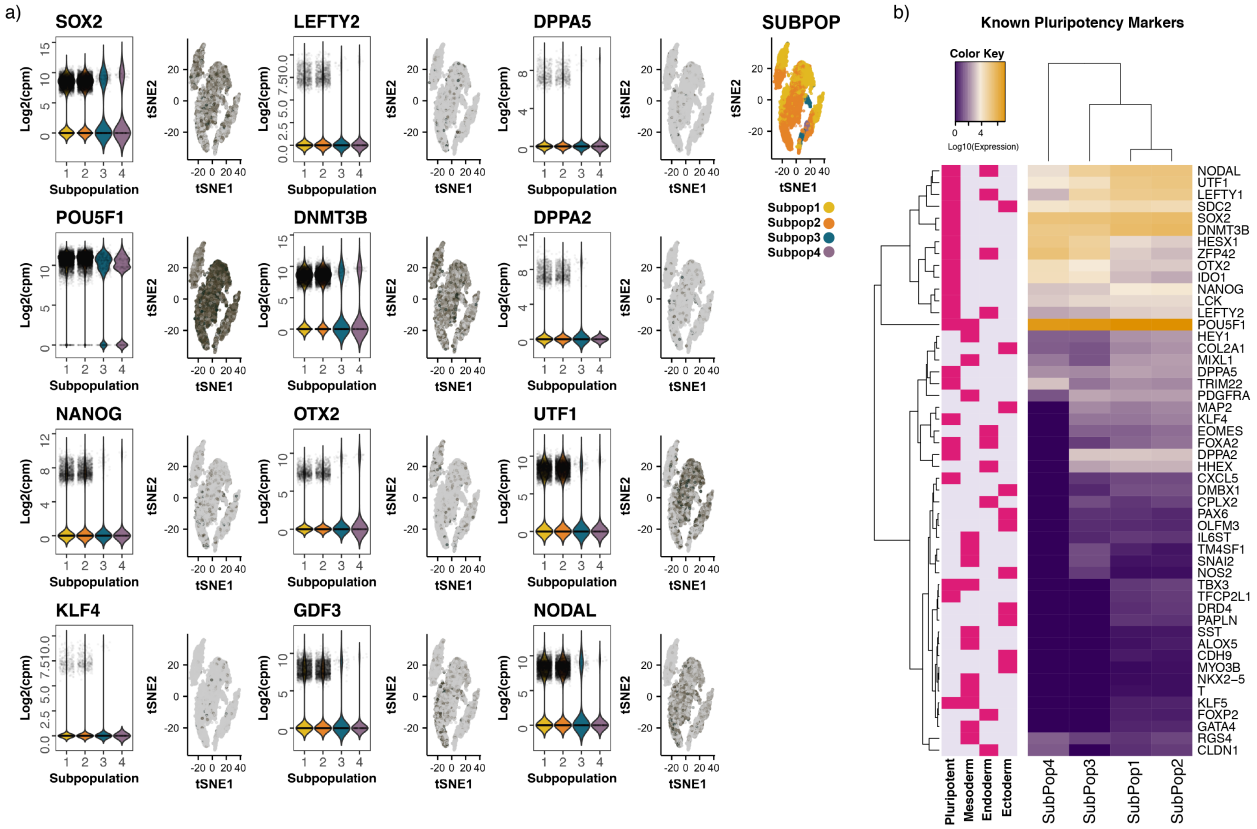

Figure 2. Expression levels of known pluripotency and lineage primed markers. (a) Violin and jitter figures for expression of top pluripotency markers and expression of the selected genes represented by t-SNE plots. Each point represents a single cell. Color gradient in the t-SNE plot represents relative expression level of the gene in a cell compared to in other cells (light = low; dark = high). **(b)** Heatmap of the mean expression of known markers within each cluster. The left-hand panel shows the classifications of genes into pluripotency and lineage-primed markers.

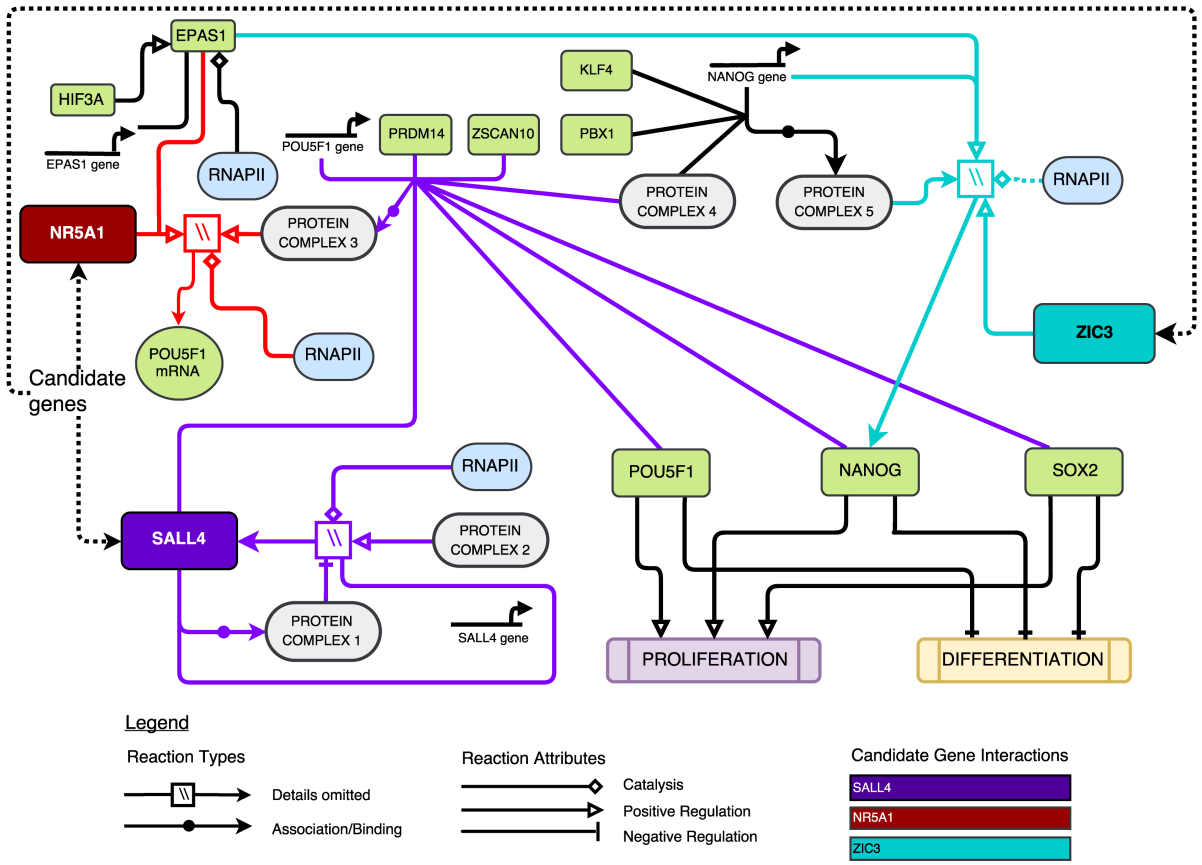

Figure 3. Network analysis of 49 differentially expressed genes between subpopulations one and two. Reactome pathway enrichment analysis was applied for 49 DE genes. The enriched Reactome pathway is 'transcriptional regulation of pluripotent stem cells', and its highly enriched child-pathway 'POU5F1 (OCT4), SOX2, NANOG activate genes related to proliferation'. Purple denotes effects of the SALL4 gene. Lines with an arrow indicate activation and non-arrowed lines indicate suppression. Protein Complex 1 - SALL4:SALL4; Protein Complex 2 - POU5F1:STAT3; Protein Complex 3 - POU5F1:SOX2:NANOG:ZSCAN10:PRDM14:SMAD2:SALL4 :POU5F1; Protein Complex 4 - SMAD4:p-SMAD2:p-SMAD2; Protein Complex 5 - POU5F1:SOX2: NANOG:KLF4:PBX1:SMAD2:NANOG.

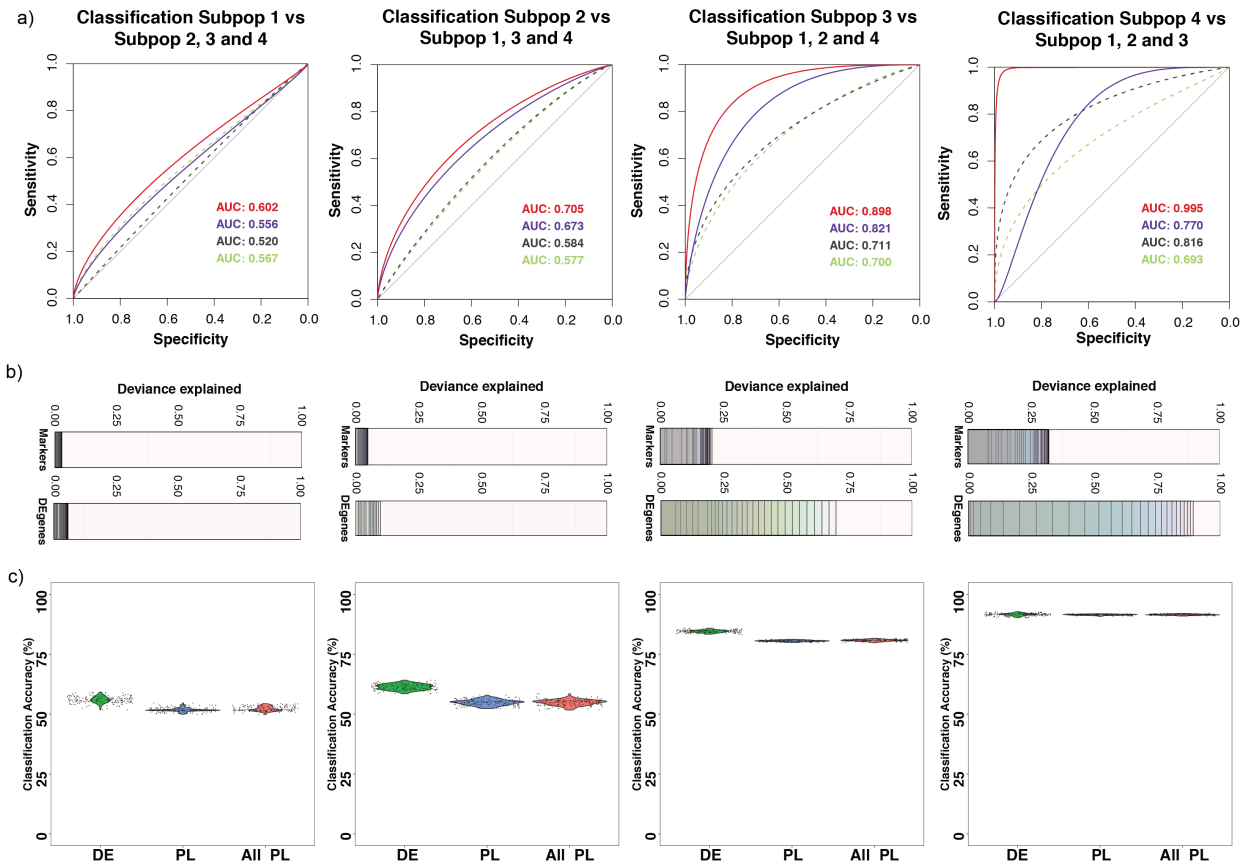

Figure 4. Selection of significant gene predictors for classifying each cluster using LASSO regression. (a) For each cluster, LASSO model was run using a set of differentially expressed (DE) genes and another set of known markers. Dashed lines are ROC (Receiver Operating Characteristic) curves for models using known markers. Continuous lines are for models using differentially expressed genes. The text shows corresponding AUC (Area Under the Curve) values for ROC curves. For each case (known markers or DE genes), a model with the lowest AUC and another model with the highest AUC are given. Lower AUC values (and ROC curves) in the prediction models using known markers suggested that the models using DE genes performed better in sensitivity and specificity. **(b)** Each deviance plot (bottom panel) shows the deviance explained (x-axis) by a set of gene predictors (numbers of genes is shown as vertical lines and varies from 1 to maximum value as the total number of gene input or to the minimum number of genes that can explain most of the deviance). The remaining space between the last gene and 1.0 border represents deviance not explained by the genes in the model. **(c)** Classification accuracy calculated by a Bootstrap method using all known markers (both pluripotent markers and primed lineage markers) or markers from our differentially expressed gene list is shown. Expression of LASSO selected genes for subpopulation one and subpopulation two is shown in Figure S5. The X-axis labels are for three cases: using LASSO selected differentially expressed genes (DE); LASSO selected pluripotency/lineage-primed markers (PL); and all pluripotency/lineage-primed markers (All PL)