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

Quan H. Nguyen¹, Samuel W. Lukowski¹, Han Sheng Chiu¹, Anne Senabouth¹, Timothy J. C. Bruxner¹, Angelika N. Christ¹, Nathan J. Palpant¹, Joseph E. Powell¹,²*

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
transcriptional heterogeneity comprising the pluripotent state. In this study, we provide the largest dataset of single-cell transcriptional profiling of undifferentiated hiPSCs currently available, which cumulatively amount to 18,787 cells across five biological replicates. Our findings address the following hypotheses: (1) that transcriptional resolution at the single cell level reveals gene networks governing specific cell subpopulations, (2) transcripts can exhibit differences in gene expression heterogeneity between specific subpopulation of cells, and (3) pluripotent cells form distinct groups or subpopulations of cells based on biological processes or differentiation potential.

Results

Description of the parental hiPSC line, CRISPRi

WTC-CRISPRi hiPSCs (Mandegar et al., 2016) were chosen as the parental cell line for this study. These cells are genetically engineered with an inducible nuclease-dead Cas9 fused to a KRAB repression domain (Figure S1a). Transcriptional inhibition by gRNAs targeted to the transcriptional start site is doxycycline-dependent and can be designed to silence genes in an allele-specific manner. The versatility of this line provides a means to use this scRNA-seq data as a parental reference point for future studies aiming to assess the transcriptional basis of pluripotency at the single cell level. Cells were verified to have a normal 46 X,Y male karyotype by Giemsa banding analysis before analysis by scRNA-seq (Figure S1b).

Single-cell RNA Sequence data

After quality control of the sequencing data (Methods), we obtained 1,030,909,022 sequence reads for 20,482 cells from five hiPSC single cell samples (Table S1, Figure S2), with 63-71% confidently and uniquely mapped (mapping quality 255) to the human reference transcriptome hg19 (ENSEMBL, release 75). We sequenced 19,937 cells from four samples to an average depth of 44,506 reads-per-cell (rpc), while one sample consisting of 545 cells was sequenced to an average depth of 318,909 rpc. On average, 2,536 genes and 9,030 Unique Molecular Identifiers (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/](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
SOX2, NANOG and UTF1 were expressed across the subpopulations (Figure 2a-b, Tables S5 and S6) but showed differences in expression heterogeneity, suggesting differences in the pluripotent state across subpopulations (Table S5).

We sought to identify biological processes underlying classification of cell subpopulations by firstly performing a statistical analysis to identify significantly differentially expressed genes between subpopulations using a binomial test that accounts for both cell numbers and negative binomial distribution of a gene’s expression within a subpopulation (Methods, Figure 1c, Table S4). Differentially expressed genes with a fold-change significant at a Bonferroni-corrected p-value threshold (p < 3.1 x 10^{-7}) were evaluated for enrichment of functional pathways (Tables S7-S11).

Cells classified in subpopulations one and two, which comprise 96.1% of total cells analyzed, were distinguished from one another by significantly different expression levels of genes in alternate pathways controlling pluripotency and differentiation (Figure 3, Tables S7-S9). The Transcriptional Regulation of Pluripotent Stem Cells (TRPSC) pathway was consistently up-regulated in cells classified as subpopulation two compared to subpopulation one (Figure 3, Table S9 and S12). TRPSC is an auto-activation loop, which maintains expression of POU5F1, NANOG, and SOX2 at high levels. Complexes containing various combinations of these transcription factors (Lam et al., 2012) can activate the expression of genes whose products are associated with rapid cell proliferation, and also repress the expression of genes associated with cell differentiation (Forristal et al., 2010; Guenther, 2011) (Figure 3). In particular, POU5F1, NANOG, and SOX2 are more highly expressed in subpopulation two (Table S5), and the direction of differential expression of genes associated with cell proliferation and repression of cell differentiation (Forristal et al., 2010; Guenther, 2011) is consistent with subpopulation two containing cells that are more active in their self-renewal than cells in subpopulation one (Tables S5, S8 and S12).
Our unbiased differential expression analysis identified SALL4 (Spal-like transcription factor 4) as significantly higher in subpopulation two than in subpopulation one (Table S12, p-adjusted: 4.3x10^{-5}). SALL4 is one of the key transcription factors that participates in controlling transcriptional balance in pluripotent cells and suppressing differentiation (Miller et al., 2016) (Figure 3). Specifically, SALL4 activates transcription of POU5F1 and maintains pluripotency (Yang et al., 2010b). Another upregulated gene (p-adjusted: 7.0 x 10^{-5}) in subpopulation two, ZIC1 (Zic family member 1), was identified by GeneMANIA analysis to be related to SALL4 through shared protein domains (Figure S4). Both ZIC1 and SALL4 were predicted by the STRING database to interact with key pluripotency markers (Figure S4). Furthermore, ZIC1 and its paralog ZIC3, a key member in the TRPSC pathway (Figure 3), are involved in maintaining the undifferentiated state, for example in the case of neural precursor cells (Inoue et al., 2007). Moreover, we also identified another differentially expressed gene, NR6A1 (p-adjusted: 3.7 x 10^{-6}), which we predict is likely to participate in the TRPSC pathway since its paralog, NR5A1, is among the key members of this pathway (Figure S4). Based on these observations, we hypothesise that in subpopulation two the three differentially expressed (DE) genes, SALL4, NR5A1 and ZIC3, cooperate with key pluripotency transcription factors (POUSF1-OCT4, SOX2, and NANOG) to activate genes related to proliferation, but not genes involved in differentiation (Figure 3).

Compared to subpopulations one and two, subpopulations three and four represent pluripotent populations with significant down-regulation of key pluripotency network genes (e.g. NANOG, OTX2, SOX2 and UTF1) (Figure 2a-b). For subpopulation three, which comprises 2.8% of cells analyzed, Reactome pathway enrichment analysis of 2,534 DE genes between subpopulations three and four showed the top pathways related to developmental signaling and transcriptional regulation via chromatin modification (Table S10). Intracellular signaling pathways that control cell proliferation, cell differentiation, and cell migration, such as EGFR, PDGF, and NGF pathways (FDR < 1.7 x 10^{-6}), were the top three most enriched pathways (Table S10). Additionally, signalling pathways by FGFRs involved in differentiation were also significantly enriched (FDR < 3 x 10^{-4}). Comparing clusters three and one, signaling by TGF-beta,
and signaling by NODAL were in the top enriched pathways (FDR < 8 \times 10^{-3}). Similarly, signaling by NODAL (FDR < 0.04) (LeVincent et al., 2003) and pre-NOTCH processing (FDR < 0.04) (Artavanis-Tsakonas et al., 1999), which are involved in cell fate decisions, were in the top enriched pathways when comparing subpopulation three to subpopulations one and two (Table S10). Thus, pluripotent cells in subpopulation three appear further advanced towards being lineage primed compared to subpopulations one and two.

Pathway enrichment analysis by BiNGO in Cytoscape for all 1,706 DE genes in subpopulation four vs. all other subpopulations (1.1% of analyzed cells) (Table S11) found the top enriched pathways related to differentiation including genes involved in: gastrulation (FDR < 1.3 \times 10^{-2}) and formation of primary germ layer (FDR < 1.4 \times 10^{-2}); developmental process (FDR < 2.8 \times 10^{-3}); and cell differentiation (FDR < 1.2 \times 10^{-2}); and more than 20 significantly (FDR < 5 \times 10^{-2}) enriched pathways related to organogenesis (Table S11). Thus, although cells in subpopulation four are still pluripotent, as indicated by the expression of pluripotent markers, they represent cells at a late-primed state progressing toward differentiation.

Taken together, our transcriptional profiling of single cells revealed four subpopulations defined by their pluripotency levels, cell proliferation, and potential for cell lineage commitment. Subpopulation one pluripotent cells represent a quiescent pluripotent state, subpopulation two represents proliferative pluripotent cells, subpopulation three as early-primed for differentiation, and subpopulation four as late-primed for differentiation (Figure 1b).

**Cell classification can be predicted from transcriptome profiles**

Using the lists of differentially expressed genes, we built an unbiased predictor to identify the pluripotency potential of a single cell. To avoid over-fitting the model due to co-expression of genes, we used a variable selection regression model called LASSO (Tibshirani, 1996) to estimate gene effects differentiating each cluster conditional upon the effects of other genes. Using a 100-fold bootstrapping approach, we estimated the predictive accuracy of identifying a cell in each of the four subpopulations (Figure 4, Table S13) (Tibshirani, 1996). To detect new
gene markers compared to the use of known pluripotency markers (Table S5), we applied LASSO to selected sets of differentially expressed genes between one subpopulation compared to the remaining three subpopulations. Consistently across four comparisons, we found that our models based on the genes identified from our differential expression analysis had a higher prediction accuracy, explained more deviance and performed with better sensitivity and specificity (higher area under the curve - AUC) than those using known pluripotency and differentiation markers (Figure 3, Figure S5, Tables S13-S14). We observed the highest classification accuracy using genes identified using the LASSO model for cells in subpopulations three and four than cells in subpopulations one and two, suggesting that these subpopulations were more divergent from the remaining majority of the cell population. This observation further supports the classification that subpopulations three and four are more primed to differentiation than subpopulations one and two.

To confirm that genes selected by our LASSO analysis were also detected in other hiPSC lines, we obtained open-access RNA-Seq transcript count data (tags per million - tpm) from the Human Induced Pluripotent Stem Cells Initiative (HipSci) for 71 hiPSC lines derived from the skin of normal individuals (Streeter et al., 2017). Consistently, we observed expression of LASSO genes in 71 other independent hiPSC samples (Figure S6). Moreover, we observed high correlation ($r > 0.85$) between the relative expression values among genes in our single-cell dataset with those genes in the HipSci bulk RNA-seq dataset (Figure S6c). The high correlation further confirms that the single-cell sequencing data accurately reflects the relative abundance of transcripts.

Transcriptional heterogeneity revealed to be specific to cell subpopulations

With the large scale dataset of 18,787 single cells and greater than 16,000 genes detected, we were able to robustly analyze expression variation between different genes, different subpopulations, and different cells (Figure S7). The inherently high heterogeneity of gene expression in scRNA-seq data, especially for low abundant genes with a more frequent on-off signal, may reduce the detection power of differential expression analysis between cells in
different subpopulations (Shalek et al., 2013). We observed more variation for subpopulations with smaller number of cells (Figure S7a), and also found more variation for genes with low expression (Figure S7b). Tagwise dispersion, which is expression variability for a gene across all cells in a subpopulation, decreased when average expression increased (Figure S7b). The difference in the level of heterogeneity of gene expression for cells in a given subpopulation compared to other subpopulations is an important indicator of the relative dynamic cellular activity of the subpopulation. The red line in Figure S7b shows the median dispersion of all genes across all cells within a subpopulation, thereby representing the average expression heterogeneity of the subpopulation. We found the median dispersion was higher in subpopulations three and four than in subpopulations one and two (Figure S7b). This was consistent with the observation that subpopulations three and four were closer to a differentiated state compared to cells in subpopulations one and two, which were more pluripotent based on transcriptome analysis.

Discussion

While methods to dissect cell subpopulations at single cell resolution such as FACS and immunohistochemistry have been available, a comprehensive profiling of transcriptional state(s) defining functionally distinct cell subpopulations comprising a ‘homogenous’ hiPSC cell line have not be described (Wilson et al., 2015). To address this, we generated and analyzed the largest hiPSC single-cell transcriptomics dataset to date, from five biological replicates of an engineered WTC-CRISPRi hiPSC line (Mandegar et al., 2016). The 18,787 high-quality transcriptomes, collectively expressing 16,064 genes, provided strong statistical power for unbiased decomposition of this hiPSC population. Using a conservative statistical threshold, we identified hundreds to thousands of genes that are differentially expressed between cells enabling us to functionally categorize four distinct subpopulations. To our knowledge, this dataset provides the first demonstration that a pure hiPSC population comprises multiple subpopulations distinguishable by single-cell transcriptomics profiling.
Comparison of transcriptomes between subpopulations revealed gene regulatory networks controlling the identity and pluripotency differentiation potential of cell subpopulations. Across five separate biological replicates, we consistently found the existence of two main subpopulations including a pluripotent-quiescent and pluripotent-proliferative subpopulation, accounting for 96.1% of all cells profiled. Differentially expressed genes between the two subpopulations were enriched for a cell proliferation gene network coordinately regulated by SALL4, ZIC1, NR5A1, POU5F1, SOX2, and NANOG. The separation of two major subpopulations on the basis of cell proliferation states may in part be explained by evidence that reprogramming is commonly a stochastic process dependent on cell-proliferation rate (Hanna et al., 2009). It remains to be determined whether these subpopulations generally reflect a common feature of pluripotency in hESC or hiPSC populations, are a specific variable of iPSC reprogramming, and whether cells in a single population transition between quiescent and proliferative states on the basis of population dynamics over time.

Furthermore, we detected two smaller subpopulations (2.8% and 1.1% of the total cells) with transcriptional signatures of pluripotency but primed to differentiation based on enriched signaling pathways and gene ontologies related to lineage specification. Interestingly, from analysis of expression heterogeneity within and between subpopulations, we found higher variability in these two subpopulations compared to the remaining cells. This observation is consistent with recent single-cell studies showing that the transition from pluripotency to lineage commitment phase is characterized by high gene expression variability (Semrau et al., 2016) and by the gradual destabilization of the pluripotent stem cell networks (Bargaje et al., 2017).

We developed an approach that can be widely applied to optimize prediction models based on single-cell transcriptomics data to classify cells into subpopulations at a high accuracy. Identifying cell types is often based on immunostaining, FACS, or targeted PCR quantification of a small number of markers (Tsankov et al., 2015). Here, we constructed an unbiased classification model based on differential gene expression selected by LASSO regression...
optimization procedure, without requirement for prior knowledge. We identified combinations of a large number of genes not previously reported as new predictors for pluripotency and showed that prediction models from differentially expressed genes performed better than models built from known markers. Further functional genomics assays are required to determine the role these gene networks play in defining the characteristics of these cell sub-populations. The result is consistent with multiple genetic loci contributing small individual effects to polygenic traits (Yang et al., 2010a). Therefore, the results support the use of an unbiased and genome-wide approach to developing gene prediction models, which can be applied to classify cell types and discover novel markers for a phenotype.

Despite the large number of cells sequenced, this study was limited in that only 3’ mRNA was sequenced, and thus there remained variation between cell populations that could not be taken into account. Nevertheless, our aim was to deconvolute a ‘homogenous’ hiPSC population, and inclusion of transcriptional sequence data from other RNA species in the future will likely improve our ability to further delineate subpopulations of cells. Furthermore, we confirmed that the genes selected were expressed in 71 HipSci datasets (Streeter et al., 2017), and that relative expression level among genes was consistent between scRNA and bulk RNA sequencing.

The parental cell line selected for this study, WTC-CRISPRi hiPSCs (Mandegar et al., 2016), is an important system for targeted transcription inhibition, and is a key feature for functional genomics studies that build on this dataset to study the biology of pluripotency. The results of this study provide a benchmark single-cell transcriptional dataset for the field to expand our understanding of the gene networks underlying cell subpopulations in pluripotency. Future work is required to expand this analysis to multiple hiPSC and hESC lines to identify common features of single-cell subpopulations in pluripotency. This study also provides a reference dataset for functional studies using WTC-CRISPRi hiPSCs as a platform for inhibiting expression of novel candidate regulators of pluripotency or differentiation.
Methods

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

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

Generation of single cell GEMs and sequencing libraries
Single cell suspensions were loaded onto 10X Genomics Single Cell 3’ Chips along with the reverse transcription (RT) mastermix as per the manufacturer’s protocol for the Chromium Single Cell 3’ Library (10X Genomics; PN-120233), to generate single cell gel beads in emulsion (GEMs). Reverse transcription was performed using a C1000 Touch Thermal Cycler with a Deep Well Reaction Module (Bio-Rad) as follows: 55°C for 2h; 85°C for 5min; hold 4°C. cDNA was recovered and purified with DynaBeads MyOne Silane Beads (Thermo Fisher Scientific; Cat#
37002D) and SPRIselect beads (Beckman Coulter; Cat# B23318). Purified cDNA was amplified as 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. Amplified cDNA was purified using SPRIselect beads and sheared to approximately 200bp with a Covaris S2 instrument (Covaris) using the manufacturer’s recommended parameters. Sequencing libraries were generated with unique sample indices (SI) for each sample. Libraries for samples 1-3 and 4-5 were multiplexed respectively, and sequenced on an Illumina NextSeq500 (NextSeq control software v2.0.2/ Real Time Analysis v2.4.11) using a 150 cycle NextSeq500/550 High Output reagent Kit v2 (Illumina, FC-404-2002) in stand-alone mode as follows: 98bp (Read 1), 14bp (I7 Index), 8bp (I5 Index), and 10bp (Read 2).

**Bioinformatics mapping of reads to original transcripts and cells**

Processing of the sequencing data into transcript count tables was performed using the Cell Ranger Single Cell Software Suite 1.2.0 by 10X Genomics (http://10xgenomics.com/). Raw base call files from the NextSeq500 sequencer were demultiplexed, using the `cellranger mkfastq` pipeline, into sample-specific FASTQ files. These FASTQ files were then processed with the `cellranger count` pipeline where each sample was processed independently. First, `cellranger count` used STAR (Dobin et al., 2013) to align cDNA reads to the hg19 human reference transcriptome, and aligned reads were filtered for valid cell barcodes and unique molecular identifiers (UMI). Observed cell barcodes were retained if they were 1-Hamming-distance away from an entry in a whitelist of known barcodes. UMIs were retained if they were not homopolymers and had a quality score > 10 (90% base accuracy). `cellranger count` corrected mismatched barcodes if the base mismatch was due to sequencing error, determined by the quality of the mismatched base pair and the overall distribution of barcode counts. A UMI was corrected to another, more prolific UMI if it was 1-Hamming-distance away and it shared the same cell barcode and gene. `cellranger count` examined the distribution of UMI counts for each unique cell barcode in the sample and selected cell barcodes with UMI counts that fell within the 99th percentile of the range defined by the estimated cell count value. The default estimated cell count value of 3000 was used for this experiment. Counts that fell within an order of magnitude of the 99th percentile were also retained. The resulting analysis files for
each sample were then aggregated using the cellranger aggr pipeline, which performed a between-sample normalization step and merged all 5 samples into one. Post-aggregation, the count data was processed and analyzed using a comprehensive pipeline assembled and optimized in-house as described below.

Preprocessing

To preprocess the mapped data, we constructed a cell quality matrix based on the following data types: library size (total mapped reads), total number of genes detected, percent of reads mapped to mitochondrial genes, and percent of reads mapped to ribosomal genes. Cells that had any of the 4 parameter measurements higher than 3x median absolute deviation (MAD) of all cells were considered outliers and removed from subsequent analysis (Table S2). In addition, we applied two thresholds to remove cells with mitochondrial reads above 20% or ribosomal reads above 50% (Table S2). To exclude genes that were potentially detected from random noise, we removed genes that were detected in fewer than 1% of all cells. Before normalization, abundantly expressed ribosomal protein genes and mitochondrial genes were discarded to minimize the influence of those genes in driving clustering and differential expression analysis.

Data normalization

Two levels of normalization were performed to reduce possible systematic bias between samples and between cells. To reduce potential confounding effects caused by differences in sequencing depths between five samples, a subsampling process (Zheng et al., 2017) was used to scale the mean mapped reads (MMR) per cell of all samples down to the level of the sample with the lowest MMR. For each sample, a binomial sampling process randomly selected reads and UMIs for each gene in a cell at a sample-specific subsampling rate. The subsampling rate for each sample was determined using the ratios of expected total reads (given the expected mean reads per cell (minimum MMR of all samples), the known number of cells, and the fraction of mapped reads to total reads) to the original total mapped reads (equation 1). Following
resampling, the MMRs for the five samples were scaled, while the expression data distribution
for genes in all cells of the sample was maintained.

\[
\text{Rate}_i = \min(\text{MMR}_j) \times N_j \times (\text{ReadFraction}_i) / \text{Total\_mapped\_reads}_i
\]  

(1)

where: \(\min(\text{MMR}_j)\) is the minimum MMR of all samples to be merged; \(N_j\) is the number of cells
in sample \(j\); \(\text{ReadFraction}_i\) is the ratio of confidently mapped reads in a cell to the total number
of reads detected for that cell in sample \(i\); \(\text{Total\_mapped\_reads}_i\) is the total number reads that
share the same cell barcode. For each gene in each cell, the sampling process of reads was
performed using the function: \(\text{rbinom(length(reads), reads, subsample\_rate)}\) in R. This process
is more robust than standard scaling options because it takes into account unique read
information associated to mapped genes and cells.

To reduce cell-specific systematic bias, possibly caused by technical variation such as cDNA
synthesis, PCR amplification efficiency and sequencing depth for each cell, expression values for
all genes in a cell were scaled based on an estimated cell-specific size factor. Before
normalization, counts were log\(_2\)-transformed (by log\(_2(\text{count}+1))\) to stabilize variance due to the
large range of count values (spanning 6 orders of magnitude, Figure S1e). To estimate the
scaling size factor for each cell, a deconvolution method (Lun et al., 2016) was applied for
summation of gene expression in groups of cells. This summation approach reduced the
number of stochastic zero expression of genes that are lowly expressed (higher dropout rates),
or genes that are turned on/off in different subpopulations of cells.

\[
E(V_{ik}) = \lambda_0 \sum_{j \in S_k} \theta_j t_j^{-1}
\]  

(2)

Where \(S_k\) is a pool of cells, \(V_{ik}\) is the sum of adjusted expression value (\(Z_{ij} = \theta_j \times \lambda_{i0}\), where \(\lambda_{i0}\) is
the expected transcript count and \(\theta_j\) is the cell specific bias) across all cells in pool \(V_k\) for gene \(i\),
\(\theta_j t_j^{-1}\) is the cell-specific scale factor for cell \(j\) (where \(t_j\) is the constant adjustment factor for
the whole population and thus can be set to unity or ignored. The cell pools were sampled using a
sliding window on a list of cells ranked by library size for each cell. Four sliding windows with
20, 40, 60, and 80 cells were independently applied and results were combined to generate a
linear system that can be decomposed by QR decomposition to estimate $\theta_j t_j^{-1}$ size factor for
each of the cell. The normalized counts are the results of taking the raw counts divided by cell-
specific normalized size factors.

Analyzing transcriptional heterogeneity in 18,787 single cells

To assess transcriptional heterogeneity among cells and genes, we first removed potential
variation due to technical sources by the subsampling process and the cell-specific
normalization as described above. Depending on experimental designs, an additional step using
a generalized linear model (GLM) to regress out other potential confounding factors can be
included. After reducing technical variation via normalization, we calculated the coefficient of
variation and expression dispersion of each gene across all cells. For cell-to-cell variation, we
first performed principal component analysis (PCA) on general cell data, which included percent
counts of the top 100 genes, total number of genes, percent of mitochondrial and ribosomal
genes. To investigate variation between genes, the distribution of dispersion across a range of
expression values was calculated (equation 3). This approach is useful because technical
variation often appears greater in lowly expressed genes than in more abundant genes (Shalek
et al., 2013). Denoting $x_i$ as the vector of expression values (in cpm) for gene $i$ across all cells,
we use the following formula to compute coefficient of variation:

$$\text{Coefficient of variation} = \frac{\sqrt{\text{var}(x_i)}}{\text{mean}(x_i)}$$

We estimated the BCV (biological coefficient of variation) with an empirical Bayesian approach
to estimate dispersion between genes and between samples (McCarthy et al., 2012). Common
dispersion (shared dispersion value of all genes), trended dispersion (mean dispersion trend for
lowly expressed genes to abundant genes), and gene-specific dispersion was estimated to
reflect variation of all genes across the whole population (Figure S7).

Dimensionality reduction
After merging five samples, preprocessing, and normalizing the dataset was scaled to z-distribution and PCA was performed for dimensionality reduction using the `prcomp` function in R (McCarthy et al., 2016). To assess PCA results, we examined the top genes that were most correlated to PC1 and PC2, and the distribution of cells and percent variance explained by the top five PCs. Importantly, the optimal number of PCs explaining the most variance in the dataset was determined using a Scree test calculated by the `fa.parallel` function in the `psych` package. The `fa.parallel` was run based on expression data for the top variable genes.

Cells are represented using t-SNE (t-distributed Stochastic Neighbor Embedding) and diffusion map (van der Maaten and Hinton, 2008). We applied the `Rtsne` package v0.1.3 on the normalized expression data (16,064 genes x 18,787 cells) to calculate a three-dimensional t-SNE projection dataset (a 16,064 cells x three t-SNE dimensions), which was then combined with other data types to display cells on two- and three-dimensional t-SNE plots.

**Clustering**

We developed an unsupervised clustering method, *i.e.* without taking into account any predetermined parameters. Using the cell-PCA eigenvector matrix, agglomerative hierarchical clustering using Ward’s minimum distance option was applied to construct a dendrogram tree, which contains all cells grouped into multiple layers of branches of similar cells, based on transcriptome profiles. To determine the number of subpopulations, branches of the dendrogram were pruned by a Dynamic Tree Cut method, which does not employ a constant (supervised) height cutoff, but dynamically performs top-down iterative decomposition and combination of clusters from larger to smaller neighbouring clusters until the number of clusters becomes stable (Langfelder et al., 2008). Cluster information was then represented in t-SNE graphs.

**Differential expression analysis**

To select genes that distinguish subpopulations, we performed pairwise differential expression analysis between cells in pairs or groups of clusters by fitting a general linear model and using a
negative binomial test as described in the DESeq package (Anders and Huber, 2010). Each cell was considered as one biological replicate in each cluster. We found that the shrinkage estimation of dispersion approach used in DESeq produced stable estimation of scale factors for genes and cells between clusters and was more conservative in detecting differentially expressed genes, especially when comparing subpopulations with a larger number of cells, such as subpopulation one and two, to subpopulations with small cell numbers, such as three and four. Specifically, DESeq detected fewer DE genes that expressed highly in a small proportion of cells in a subpopulation while remaining cells in that subpopulation had zero or very low expression. Significantly differentiated genes were those with Bonferroni adjusted P-values less than 5% ($p < 3.1 \times 10^{-7}$).

**LASSO regression analysis**

To develop predictive models based on single cell transcriptomics data, we applied Least absolute shrinkage and selection operator (LASSO) procedure, to choose gene predictors for classifying cells into one of the four subpopulations. Briefly, penalized logistic regression was applied to fit a predictor matrix containing expression values of top differentiated genes in all cells (or a subsample of a randomly selected 10 percent of the total cells) and a response vector assigning cells into one of the subpopulations (dichotomous variable). The LASSO procedure optimizes the combination set of coefficients for all predictors in a way that the residual sum of squares is smallest for a given lambda value (Friedman et al., 2010). In other words, the LASSO procedure identified an optimal combination of genes (predictors) and fitted a logistic regression model, in which expression values of the selected genes were predictors and the binary labels of cells were response variables. The fitted model could either explain the highest deviance (compared to the full model) or classify cells to subpopulations with the lowest 10-fold classification error. The glmnet R package was applied to select top genes that contributed to classifying cells into each subpopulation (Tibshirani et al., 2012). The LASSO model was trained using one subsampled dataset and then evaluated in a new, non-overlapping subsampled dataset. Prediction accuracy was estimated by applying the trained model, with only selected predictors and their corresponding coefficients, on a new set of randomly
sampled cells that were not used in the model training dataset. Bootstrap was used to calculate classification accuracy for 100 iterations.

**Pathway and gene functional analysis**

To functionally characterize the four subpopulations, we performed a network analysis using significant DE genes between cells within a subpopulation and the remaining cells, or between cells in pairs of subpopulations. We used Cytoscape to apply three main programs: GeneMANIA (Warde-Farley et al., 2010), with a comprehensive background database containing 269 networks and 14.3 million interactions, the Reactome functional interaction network analysis, a reliably curated protein functional network (Wu et al., 2010), and the STRING protein-protein interaction database (Szklarczyk et al., 2015).

**Author contributions**

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. and A.N.C. performed experiments. Q.H.N, S.W.L., A.S performed analysis. All authors wrote and edited the manuscript.

**Acknowledgments**

Sequencing was performed by the Institute for Molecular Bioscience Sequencing Facility at the University of Queensland. This work was supported by the Australian National Health and Medical Research Council (NHMRC) grant APP1083405 and APP1107599.

**References**

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. 11.

Artavanis-Tsakonas, S., Rand, M.D., and Lake, R.J. (1999). Notch Signaling: Cell Fate Control and Signal Integration in Development. Science 284, 770-776.

Bargaje, R., Trachana, K., Shelton, M.N., McGinnis, C.S., Zhou, J.X., Chadick, C., Cook, S., Cavanaugh, C., Huang, S., and Hood, L. (2017). Cell population structure prior to bifurcation predicts efficiency of directed differentiation in human induced pluripotent cells. Proc. Natl. Acad. Sci. U. S. A. 114, 2271–2276.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15-21.
Evans, M.J., and Kaufman, M.H. (1981). Establishment in culture of pluripotential cells from mouse embryos. Nature 292, 154-156.

Forristal, C.E., Wright, K.L., Hanley, N.A., Oreffo, R.O., and Houghton, F.D. (2010). Hypoxia inducible factors regulate pluripotency and proliferation in human embryonic stem cells cultured at reduced oxygen tensions. Reproduction 139, 85-97.

Friedman, J.H., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 33, 22.

Guenther, M.G. (2011). Transcriptional control of embryonic and induced pluripotent stem cells. Epigenomics 3, 323-343.

Hanna, J., Saha, K., Pando, B., van Zon, J., Lengner, C.J., Creyghton, M.P., van Oudenaarden, A., and Jaenisch, R. (2009). Direct cell reprogramming is a stochastic process amenable to acceleration. Nature 462, 595-601.

Inoue, T., Ota, M., Ogawa, M., Mikoshiba, K., and Aruga, J. (2007). Zic1 and Zic3 regulate medial forebrain development through expansion of neuronal progenitors. J. Neurosci. 27, 5461-5473.

Lam, C.S., Mistri, T.K., Foo, Y.H., Sudhaharan, T., Gan, H.T., Rodda, D., Lim, L.H., Chou, C., Robson, P., Wohland, T., et al. (2012). DNA-dependent Oct4-Sox2 interaction and diffusion properties characteristic of the pluripotent cell state revealed by fluorescence spectroscopy. Biochem. J. 448, 21-33.

Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics 24, 719-720.
Palpant, N.J., Pabon, L., Friedman, C.E., Roberts, M., Hadland, B., Zaunbrecher, R.J., Bernstein, I., Zheng, Y., and Murry, C.E. (2017). Generating high-purity cardiac and endothelial derivatives from patterned mesoderm using human pluripotent stem cells. Nat. Protocols 12, 15-31.

Semrau, S., Goldmann, J., Soumillon, M., Mikkelsen, T.S., Jaenisch, R., and van Oudenaarden, A. (2016). Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells. bioRxiv.

Shalek, A.K., Satija, R., Adiconis, X., Gertner, R.S., Gaublomme, J.T., and Raychowdhury, R. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature 498.

The Steering Committee of the International Stem Cell, I. (2005). The International Stem Cell Initiative: toward benchmarks for human embryonic stem cell research. Nat Biotech 23, 795-797.

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.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58, 267-288.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., and Tibshirani, R.J. (2012). Strong rules for discarding predictors in lasso-type problems. Journal of the Royal Statistical Society. Series B, Statistical methodology 74, 245-266.

Tsankov, A.M., Akopian, V., Pop, R., Chetty, S., Gifford, C.A., Daheron, L., Tsankova, N.M., and Meissner, A. (2015). A qPCR ScoreCard quantifies the differentiation potential of human pluripotent stem cells. Nat Biotech 33, 1182-1192.

van der Maaten, L., and Hinton, G.E. (2008). Visualizing data using t-SNE. J. Mach. Learn.Research 9, 2579–2605.

Wen, L., and Tang, F. (2016). Single-cell sequencing in stem cell biology. Genome Biol. 17, 71.
Wilson, N.K., Kent, D.G., Buettner, F., Shehata, M., Macaulay, I.C., Calero-Nieto, F.J., Sanchez Castillo, M., Oedekoven, C.A., Diamanti, E., Schulte, R., et al. (2015). Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations. Cell Stem Cell 16, 712-724.

Wu, G., Feng, X., and Stein, L. (2010). A human functional protein interaction network and its application to cancer data analysis. Genome Biol. 11, R53.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010a). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42, 565-569.

Yang, J., Gao, C., Chai, L., and Ma, Y. (2010b). A novel SALL4/OCT4 transcriptional feedback network for pluripotency of embryonic stem cells. PLoS ONE 5, e10766.

Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat Commu 8, 14049.
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).