Computational assignment Of cell-cycle stage from single-cell transcriptome data

Article *in* Methods · July 2015

DOI: 10.1016/j.ymeth.2015.06.021

8 authors, including:

Antonio Scialdone
Helmholtz Zentrum München
33 PUBLICATIONS **499** CITATIONS
SEE PROFILE

Luis R. Saraiva
Sidra Medical and Research Center
38 PUBLICATIONS **597** CITATIONS
SEE PROFILE

Oliver Stegle
European Molecular Biology Laboratory
306 PUBLICATIONS **5,135** CITATIONS
SEE PROFILE

Florian Buettner
EMBL-EBI
57 PUBLICATIONS **1,180** CITATIONS
SEE PROFILE

Computational assignment of cell-cycle stage from single-cell transcriptome data

Antonio Scialdone[1,2,*], Kedar N. Natarajan[1,2], Luis R. Saraiva[1,2], Valentina Proserpio[1,2], Sarah A.Teichmann[1,2], Oliver Stegle[2,*], John C. Marioni[1,2,*] and Florian Buettner[2,3,*]

[1] *Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK*

[2] *European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge, CB10 1SD, UK*

[3] *Institute of Computational Biology, Helmholtz Zentrum München, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany.*

[*] *To whom correspondence should be addressed: AS: as1@ebi.ac.uk, OS: stegle@ebi.ac.uk, JCM: marioni@ebi.ac.uk, FB: buettner@ebi.ac.uk*

Abstract

The transcriptome of single cells can reveal important information about cellular states and heterogeneity within populations of cells. Recently, single-cell RNA-sequencing has facilitated expression profiling of large numbers of single cells in parallel. To fully exploit these data, it is critical that suitable computational approaches are developed. One key challenge, especially pertinent when considering

dividing populations of cells, is to understand the cell-cycle stage of each captured cell. Here we describe and compare five established supervised machine learning methods and a custom-built predictor for allocating cells to their cell-cycle stage on the basis of their transcriptome. In particular, we assess the impact of different normalization strategies and the usage of prior knowledge on the predictive power of the classifiers. We tested the methods on previously published datasets and found that a PCA-based approach and the custom predictor performed best. Moreover, our analysis shows that the performance depends strongly on normalization and the usage of prior knowledge. Only by leveraging prior knowledge in form of cell-cycle annotated genes and by preprocessing the data using a rank-based normalization, is it possible to robustly capture the transcriptional cell-cycle signature across different cell types, organisms and experimental protocols.

1 Introduction

Recent technological advances have helped to establish single-cell RNA-sequencing (scRNA-seq) as a robust and routine assay, enabling the transcriptional profiling of thousands of cells to be processed in an unbiased manner [1, 2]. The application of scRNA-Seq to a wide range of different systems has already resulted in new insights in important areas such as embryogenesis [3] and tissue heterogeneity [4]. Indeed, scRNA-Seq enables detection and quantification of transcriptional changes at the level of single-cells, thereby unravelling dynamic aspects of the transcriptional heterogeneity between cells that is not accessible using bulk sequencing approaches. For example, scRNA-seq has helped to identify novel cell types [5] and to reveal dynamic changes of the transcriptome during temporal processes like cell differentiation [6].

Importantly, the state of each individual cell is reflected by a multitude of individual components, many of which are reflected by transcriptome signatures. A key component and major driver of transcriptional heterogeneity and cell decision processes is the cell cycle. Moreover, the cell cycle is known to be linked to fundamental biological processes, including cell differentiation [7] and oncogenesis [8, 9]. Consequently, accurately identifying the cell cycle stage of individual cells is needed

to fully understand a number of different biological problems.

So far, information about cell cycle stage has largely been obtained by experimental approaches. For instance, cells can be treated with chemicals to induce cell-cycle arrest in a specific phase [10]. Alternatively, cell sorting methods can be used to stratify cells by size (counterflow centrifugation elutriation [11]) or DNA content (e.g., Hoechst staining [12]), which facilitates enrichment of cells in different stages of the cell cycle. Alternatively, strategies based on genetic manipulation through insertion of fluorescent probes in genes that are differentially expressed in different cell-cycle stages (e.g., FUCCI technique [13]) can be employed. However, these approaches have major drawbacks as they can be very labour extensive and, due to their invasive nature, have the potential to disturb the biological system substantially (e.g., cell-cycle arrest can have a large impact on differentiation potential).

In the context of scRNA-Seq experiments, the transcriptome data itself provides informative cues about the cell cycle stage of individual cells [14, 15]. In particular, genome-wide transcriptome data provides information on the expression levels of informative cell-cycle marker genes, which have been carefully annotated in several systems and cell types (e.g., in human, yeast and *Arabidopsis* [16]). Consequently, we reasoned that these genes can be used to infer the cell cycle phase directly from the transcriptome. Such an approach would be complementary to experimental sorting procedures and could help reduce biases that might arise from more invasive experimental techniques. Moreover, the cell-cycle structure of unsorted populations of cells profiled by scRNA-seq could be investigated.

While strategies have been developed to remove cell cycle variation from scRNA-seq data without inferring the cell-cycle stage in order to improve the detection of sub-populations of cells [14], and computational analyses have been used to distinguish cycling from quiescent cells [17], the possibility of explicitly predicting the cell-cycle stage of cells from their transcriptome has not yet been explored. In this paper we analyze six supervised computational methods to predict G1, S or G2M phase given the transcriptome of a cell. We train each algorithm on a recently published scRNA-seq dataset where cell-cycle information is available from experimental annotation [14], and we assess their performance on scRNA-seq datasets generated

from a variety of cell types and organisms.

2 Material and Methods

We use a supervised machine learning approach to evaluate the ability of six algorithms to predict the unobserved cell cycle stage of a cell from its transcriptome profile. These include five established supervise machine learning approaches as well as a custom-built predictor. Each algorithm was trained on the same scRNA-seq dataset where the cell-cycle stage of each cell was known. Additionally, different sets of cell-cycle annotated genes were used to build each classifier. A schematic overview of our approach is shown in Figure 1. The six prediction algorithms' performance was measured using 10-fold cross-validation on the training dataset and a variety of independent datasets. The predictive power of all classifiers was quantified by calculating the F1-score (harmonic mean of recall and precision), which has been shown to be an effective summary statistic for multi-class classification [18]. In order to quantify how well the predictors perform across all cell-cycle phases, we also calculated the macro-averaged F1 score by taking the average of precision and recall over all cell cycle phases before computing the harmonic mean, so as to make it independent of the number of cells in each phase in the testing dataset.

2.1 Prediction algorithms

We compared a total of six classifiers including linear and non-linear predictors as well as one custom method specifically designed for in-silico cell-cycle allocation. Below, we provide brief details about the methods employed and their implementation.

2.1.1 Random forest

We used the scikit-learn implementation of random forests (ExtraTreesClassifier) [19] and trained 500 trees by minimizing the entropy in the leaves of the individual

randomized trees, constructed using a subset of all N features ($\sqrt{N}$).

2.1.2 Logistic regression and Lasso

Logistic regression was used, both without regularization and with an L1 penalty (lasso) [20]. The lasso penalty was determined using an internal 5-fold cross validation, maximizing the F1 score.

2.1.3 Support vector machines

We used support vector machines with an rbf kernel with feature selection [21]. Kernel parameters were determined using a cross-validated grid search. Due to the large number of variables, feature selection was performed based on a univariate feature ranking [22]. First, for each gene an ANOVA was performed and genes were ranked according to their F-statistic. Next, the best number of features was determined in an integrated cross-validated grid search.
Multi-class classification was performed via a one-vs-one scheme (scikit-learn implementation), allowing for multi-class classification based on standard binary SVMs.

2.1.4 PCA-based classification

Recently, we showed that the first principal component (PC) of a set of annotated cell cycle marker genes is sufficient for constructing a cell-cell covariance matrix, reflecting the cell cycle induced correlation among cells [14]. We therefore evaluated a Gaussian Naive Bayes classifier based on the first PC derived from the set of cell cycle markers. Furthermore, we explored the additional predictive power of higher order PCs (see supplementary figure B.6). Naive Bayes classifiers assign a probability to each data instance based on a set of features by assuming conditional independence of the features. The Gaussian Naive Bayesian classifier assumes a Gaussian distribution of each continuous feature (here PC1 and higher order PCs if applicable) with mean and variance specific to each class (here: cell-cycle phase). Mean and variance parameters are estimated using maximum likelihood as implemented in the scikit-learn framework [19].

2.1.5 Pairs

We developed a classification algorithm based on the idea of the relative expression of "marker pairs", which is also exploited in top scoring pairs classifier, developed for classifying cancer types based on microarray data [23, 24, 25]. The algorithm selects pairs of genes whose relative expression has a sign that changes with the cell-cycle phase. These pairs can then be used to quantify the evidence that a given cell is in G1, S or G2M phase, as described in the Appendix A. Since only the sign of the relative expression of gene pairs in the same cell is used, this method does not require any normalization for sequencing depth.

2.2 Selection of cell-cycle marker genes

To establish a set of cell cycle annotated genes, we combine all genes annotated to cell cycle in the Gene Ontology database (GO:0007049) [26] along with the 600 top-ranked genes from CycleBase [27, 16]. Furthermore, we construct an informative set of cell cycle marker genes, by excluding those genes whose variation was below the technical noise in the training dataset (see section 2.3). In addition, we demonstrate the benefits of using prior knowledge by evaluating the performance of a classifier based on the complete, unbiased set of expressed genes.

2.2.1 Data post-processing

To obtain gene expression values that are comparable across a wide range of protocols, we used a rank normalisation approach: for each cell we ranked the expression values (FPKM, RPKM or normalized with size factors as in the respective primary publication) of the set of genes used for training from lowest to highest. We then used these rank-normalized gene-expression values as input for all algorithms. In addition, we explored an alternative normalisation strategy where the data from each cell was normalized with the total number of reads mapped to the gene set used for prediction.

2.3 Training data set

We trained all classifiers on a recently published single-cell RNA-seq dataset comprised of 182 mouse embryonic stem cells (mESCs) with known cell-cycle phase [14]. In brief, Rex1-GFP-expressing mESCs (Rex1-GFP mESCs) were cultured using serum-free NDiff 227 medium (Stem Cells Inc.) supplemented with 2i inhibitors. Hoechst staining (Hoechst 33342; Invitrogen) was optimized for Rex1-GFP mESC, and cells were sorted using FACS for three different cell-cycle phases (G1, S and G2M phase). Next, single-cell RNA-seq was performed using the C1 Single Cell Auto Prep System (Fluidigm; 100-7000). We normalized the raw read counts using two different size factors derived from endogenous genes and ERCC spike-ins as proposed by Brenecke et al [28]. After normalization of both endogenous genes and ERCC spike-ins with their respective size factors and estimation of technical noise using ERCC spike-ins, we identified a set of 6,635 genes with variation above the technical background level (FDR <0.1) by following the approach proposed in [28]. To establish a set of informative cell-cycle marker genes, we determined the intersection of annotated cell cycle marker genes with the set of variable genes in the Hoechst-stained mESCs. We further reduced the set of informative cell cycle genes by determining the set of genes with variation above the technical background level for an additional single-cell RNA-seq dataset [12]. This resulted in a smaller set of 405 genes. The rank-normalised expression of these informative cell cycle markers for 182 cells constitutes the training data.

2.4 Datasets with cell cycle information

We tested the performance of all predictors on a variety of independent data-sets with known ground truth.

2.4.1 Mouse mESCs data (Quartz-Seq protocol)

We used the normalized gene expression data from the primary publication [12]. In brief, mESCs were FACS sorted into G1, S and G2M phases based on their Hoechst

33342-stained cell area. Next, seven S, eight G2M and 20 G1 cells were sequenced using the Quartz-seq protocol and gene expression was normalized to FPKM values. Due to the lack of spike-ins, we estimated the amount of technical (null) noise expected for genes with variable levels of expression using a log-linear fit between the expression mean and the squared coefficient of variation between cells [14, 28]. This approach resulted in a total of 5,546 highly variable genes.

2.4.2 Human leukemia cells (bulk)

We analysed data from bulk human myeloid leukemia cells [11]. Cells were assigned to cell-cycle stages (G1, S and G2M) using centrifugal elutriation and mRNA expression was quantified using RNA-seq.

2.4.3 Bulk mESCs

mESCs were stained with Hoechst 33342 and FACS sorted for cell cycle stages (G1, S and G2M). Approx 150,000-300,000 cells from an asynchronous population and from each cell cycle fractions (G1, S and G2M) were used for bulk mRNA sequencing, with libraries being generated using the Illumina TruSeq Stranded RNA Sample preparation kit. All libraries were prepared and sequenced using the Wellcome Trust Sanger Institute sample preparation pipeline. Sequencing quality control and data quality checks were performed by the Sanger Sequencing facility. Downstream data analysis (Alignment, Mapping and counting reads) was performed as described [14].

2.5 Datasets without cell cycle information

2.5.1 Liver cells

We tested the algorithms on two independently generated sets of individually sequenced liver cells, one previously published [29], one generated for this study (see Appendix A). Since most liver cells do not proliferate (see, e.g., [30]), they are expected to be in G1 phase.

Smart-Seq protocol. We used normalized gene expression data of five liver cells from the primary publication [29]. All cells were sequenced using the Smart-Seq

protocol.

C1 protocol. In addition we generated the individual transcriptomes of 70 liver cells using the Fluidigm C1 platform. In brief, a suspension of cells was prepared from the liver of a 14-week old B6CastF1 (C57Bl/6J mother x CAST/Ei father) female mouse and loaded onto a 10-17 μm C1 Single-Cell Auto Prep IFC (Fluidigm), and cell capture was performed according to the manufacturer's instructions (see Appendix A for the detailed protocol).
Paired-end reads were mapped simultaneously to the *M. musculus* genome (Ensembl version 38.75) and the ERCC sequences using GSNAP (version 2014-05-15) with default parameters. Htseq-count [31] was used to count the number of reads mapped to each gene (default options).

2.5.2 Blastomeres

We applied our algorithm to the transcriptomes of a total of 30 individual cells dissociated from early, mid and late 2-cell stage mouse embryos and sequenced using the Smart-seq protocol [29]. Most of these cells are expected to be in G2 phase, as blastomeres from 2-cell stage embryos have a very short G1 phase and spend more than half of their cell-cycle in G2 phase [32, 33].

2.5.3 T-cells

Single-cell RNAseq data - Finally, we applied our approach to 81 T-cells [34]. We used normalized gene expression data as in [14]. We evaluated our algorithm by comparing the fraction of cells assigned to individual cell-cycle phases in silico to the respective fractions obtained from flow cytometry analysis of Ruby-stained T-cells.
Flow cytometry analysis - Untouched Naive CD4+ cells were purified from Il13-eGFP homozygous spleens from six week old mice and stained with CellTrace Violet proliferation dye (Invitrogen). After 3.5 days of activation in standard condition for T_H2 polarisation (anti-CD28 ($4\mu g/ml$, eBioscience) and anti-CD3 ($1\ \mu g/ml$, eBioscience)) and IL-4 ($10\ ng/ml$, R&D Systems) cells were stained with Vybrant DyeCycle^{TM} Ruby (Invitrogen) stain to visualise DNA content and analysed on a

BD LSR II Fortessa.

3 Results and Discussion

3.1 Predictive power and generalizability

3.1.1 Single-cell data

First, we assessed the performance of the different prediction algorithms as well as all sets of marker genes using a cross-validation approach. In this "holdout" experiment a fraction of the training data is removed when fitting the model, which is then applied to the withheld data to assess its performance. This resulted in high precision and recall for all cell cycle phases (Fig. 2.a-c) for all classifiers and gene sets, indicating that all models fit the data well and are able to generalize well when applied to the same type of data (i.e., mESCs cultured in 2i+LIF).

Poor generalizability to independent test data for many methods but PCA and pairs method To assess how well the different approaches generalize to independent data sets, we tested the six predictors derived on an independent test set of 35 mESCs sequenced using a different protocol (Quartz-seq) and cultured in a different medium (serum; see Fig. 2.d-f). Two general features are shared by all predictors: all of them perform worst on S phase prediction (see also below), and their overall predictive power substantially increases when they are trained on cell-cycle annotated genes (Fig. 2.e-f), which indicates the importance of the inclusion of prior information. The best performance on the independent mESC test set was achieved by the PCA-based Naive Bayes Classifier and the custom predictor (the "Pairs" method) which had similar predictive power. As all methods yielded good performance on the cross-validation, this indicates that all methods but PCA and Pairs overfit to the training data without being able to generalize to cells from different conditions. For the large set of all variable genes, this over-fitting-effect is particularly strong and occurs for all methods, again reflecting the importance of using prior knowledge.

Alternative normalization strategy results in poor generalizability We also assessed the influence of the normalization step by training the predictors on the total read count normalised gene expression data. This resulted in a notable decrease in performance and highlights the importance of robust normalisation strategies which hold for different experimental protocols (Supplementary Fig. B.3).

Feature importance In order to assess the relevance of individual genes for classification, we analysed the loadings on PC1 for the PCA-based method and assigned a score to the pairs of genes for the pairs method (following the approach introduced in [23]; see Appendix B.3 and figures B.4 and B.5).
While the majority of the most relevant genes are well known markers for specific cell-cycle phases (e.g., Plk1, Aurka, etc.), we could identify several genes that were not previously annotated to any particular stage of the cell cycle but were among the most important for classification. For example, Tmem2 and Tex14, which have the highest negative loadings on PC1 (i.e., the two strongest G1 markers) were not annotated with a specific peak time or phenotype in Cyclebase.

3.1.2 Bulk data

We also applied each approach to predict cell-cycle stage from bulk RNA-seq datasets where all cells had been cell-cycle staged. These datasets are fundamentally different from single-cell transcriptomics, and most of the predictors we tested are only able to distinguish G1 phase from G2M phase (Figure 3). Indeed, correct allocation of cells in the S-phase is challenging, mainly because of less specific transcriptional patterns (see Supplementary Figure B.2). Nevertheless, the PCA-based and the pairs predictor correctly allocate all samples to their true cell-cycle phase, including those in the S-phase.
Consequently, we concluded that of the six predictors evaluated and the gene sets used, the PCA-based and the pairs approaches trained on cell-cycle annotated genes had the best overall performance and, importantly, had the strongest ability to distinguish S-phase cells .

3.2 Application to datasets without ground truth

We next applied the PCA-based approach to a variety of datasets including liver cells, T-cells and blastomeres to predict their cell-cycle phases (see Appendix B for the application of the pairs method to the same datasets). For all data sets, scatter plots with G1 and G2M score of individual cells with decision boundaries are shown in Figure 4.

3.2.1 Liver cells

As expected, given the non-proliferative state of most liver cells [30], all profiled liver cells were allocated with a high degree of confidence to G1 phase (Figure 4; blue colour). This applied to cells profiled in two independent laboratories, suggesting that the PCA-based predictor is relatively robust to technical biases that may arise during sample preparation.

3.2.2 Blastomeres

The cell cycle of blastomeres from 2-cell stage embryos takes about $\sim 20h$ to complete, with the S-phase starting $\sim 1h$ after the first mitosis [32] and lasting approximately $\sim 6h$. G2 phase is very long and its length varies between ~ 12 and ~ 16 hours [33]. Apart from a few cells allocated to G1, most of the blastomeres analysed were predicted to be in S phase by the PCA-based method (see figure 4), and in G2M phase by the pairs method (see figure B.1.A), which agrees better with our prior expectations. Despite the difference in allocation probabilities, the G2M scores from the two methods have a high rank correlation (Figure B.8)), suggesting that in the PCA-based approach, due to the weak signal for the S phase, the probability for a cell being in S is less reliable than the G1/G2M probability and can also result in a badly calibrated score, in particular for cell types other than those in the training set. The pairs method is less affected by this issue, possibly due to the higher robustness of the signal captured by the relative rankings of pairs of genes [23, 24].

3.2.3 T-cells

This dataset includes T_H2 polarised cells at different stages of differentiation, which can be in any of the three cell-cycle phases as confirmed by our flow cytometry analysis of a set of Ruby-stained cells (51%, 14% and 35% allocated to G1, S and G2M phase respectively, see figure 4). By analysing the scRNA-seq dataset with the PCA-based method, we found 65.4%, 27.1% and 7.4% in G1, S and G2M phase respectively (similar percentages are obtained with the pairs method, see Appendix B.1). The method successfully predicts the cycling nature of these cells, with a relevant proportion of cells allocated to S and G2M phase. The difference with the flow cytometry analysis can be explained by, e.g., poor resolution of S in the flow cytometry and possible biases in the capture and processing of single cells for RNA-seq, which can affect the relative percentages of cells in the different phases.

4 Conclusions

We evaluated six computational methods for predicting the cell-cycle stage of single cells from their transcriptome. To train the algorithms, we used a scRNA-seq dataset where cells had been previously sorted by their cell-cycle phase, and exploited available databases of genes known to be involved in cell cycle progression (GO and cyclebase [16]) to optimise the set of predictors. We quantified the predictive power of all methods for a wide range of single cells as well as bulk data from different organisms (mouse and human) and show that a parameter-free PCA-based approach and the custom predictor (the "Pairs" method) performed best and correctly allocated cells from all data sets to their cell-cycle phase.

In order for our method to be broadly applicable to a wide range of experimental protocols, we used a rank-based approach to normalize our data, resulting in a good performance on a large variety of data-set. We also explored total count normalisation where each cell was normalised by the total number of reads mapped to the genes used as input for the predictors. This resulted in a notable decrease in performance for all methods, which may be explained by two factors. First, the cell-

cycle signature is considerably weaker without rank normalisation (Supplementary Figure B.3c and B.2). Second, rank-based normalization robustly preserves the cell-cycle related information across different experimental protocols and, particularly in combination with PCA, results in a highly regularized cell-cycle allocator with good generalizability across data sets. In contrast, more sophisticated approaches such as random forest or SVM in combination with total count normalised data, easily overfit to a specific data-set.

Similarly, the strong regularization enforced by the PCA explains the good performance of the PCA-based predictor on all datasets. All predictors achieve very high F1-scores in the cross-validation, which indicates that the predictors do not overfit within the training set and would generalize well to similar data-sets generated using the same cell type and experimental protocol. However, in order to be useful in practice, it is crucial that a predictor will also generalize well to different cell types, experimental conditions and sequencing techniques. Here we show that only the PCA-based predictor and the pairs method achieve a strong enough regularization to robustly capture a generalizable cell-cycle signature in the transcriptome. Interestingly, the signal captured by the pairs method based on the relative rankings of pairs of genes is probably the most robust and generalizable (see also [23, 24]), as it is shown by the analysis of the 2-cell stage blastomeres.
Between the three phases, the S phase proved to be the most challenging to identify. This can be explained by the least specific transcriptional signature of S-phase markers at the single-cell level (Supplementary Fig. B.2) along with the poor resolution of S phase in flow cytometry data affecting both training and testing datasets.

While the dataset we used for training only provided information on cell-cycle phase, without being able to monitor the progress within a given stage, the methods we tested assign a continuous score to each cell that can potentially provide information at higher temporal resolutions. For instance, while most of 2-cell stage blastomeres are allocated to G2M by the pairs method (see figure B.1.A), the average G2M score is lowest for early blastomeres and highest for late blastomeres,

possibly reflecting the progress of cells within the G2M stage. However, more data is needed to assess how valid this is across different cell types.

In silico allocation of cells to specific cell-cycle phases can be important for a range of applications. By integrating the knowledge of cell transcriptome with cell-cycle phase it will be possible to reveal interactions between cell cycle and other cellular processes. Furthermore, cell-cycle allocation can be crucial for the correct interpretation of single-cell data, since many genes have been shown to correlate with cell cycle, and these correlations can mask the existence of cell sub-populations [14] especially in rapidly cycling cells (e.g., cancer or stem cells), where a greater fraction of variability is attributable to cell cycle.

Accession codes - mESc and liver data have been deposited at ArrayExpress: XXX. The code for the implementation of the six cell-cycle predictors will be made available on GitHub.

Acknowledgements - We acknowledge support of the European Research Council (Starting Grant no. 260507 thSWITCH to S.A.T), the Sanger-EBI Single Cell Centre (K.N.N. & A.S.) and the UK Medical Research Council (biostatistics career development fellowship to F.B.).

Figures

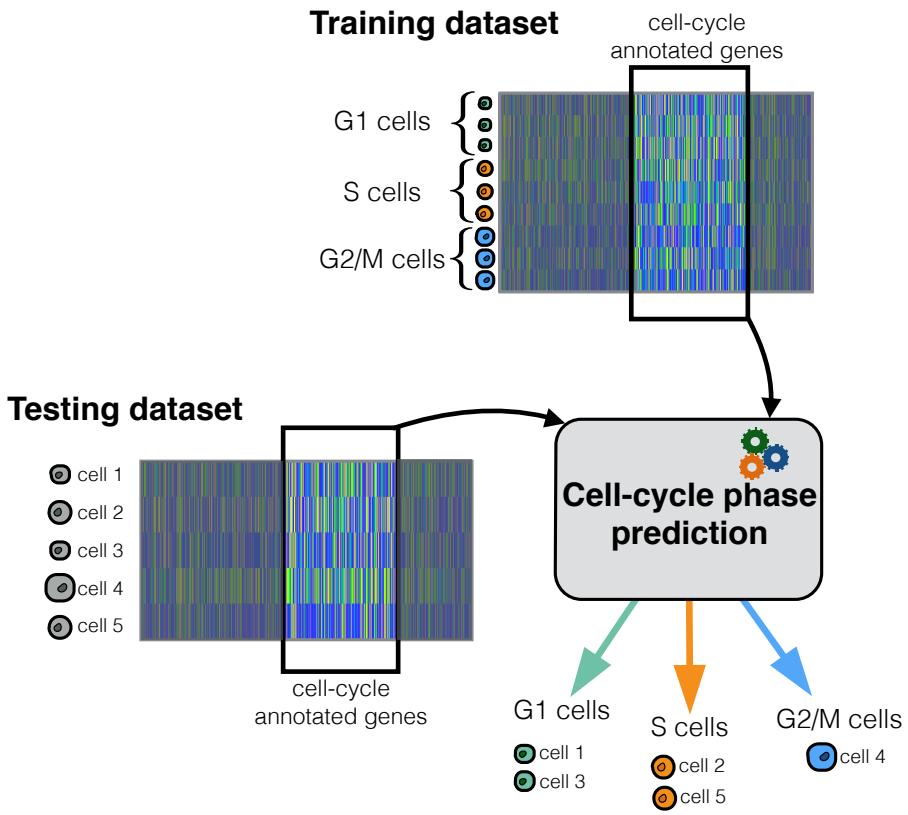

Figure 1: Overview of our approach. The transcriptional profiles of individual cells are taken as input and information on cell cycle markers is extracted (left). The expression profiles of these genes in a training dataset are then used to train an algorithm (top) that can predict the cell cycle stage of individual cells in independent datasets.

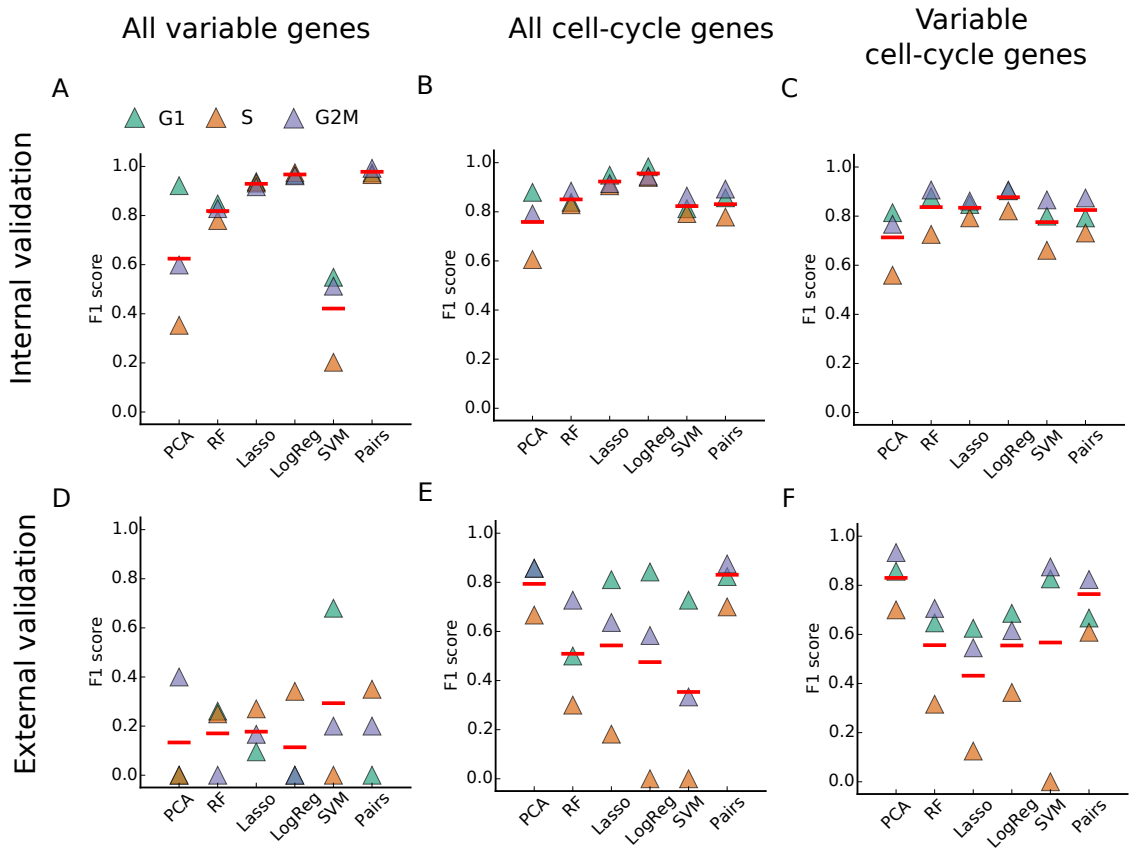

Figure 2: Validation on data with known cell-cycle phase. a-c, F1 scores from internal cross validation for different gene sets; F1 score for G1 phase is shown in green, for S-phase in orange and for G2M phase in blue. Red lines represent the macro-averaged F1 score. A, all variable genes, B, all annotated cell-cycle genes, C, all variable cell-cycle genes. D-F, F1 scores on independent test set. D, all variable genes, E, all annotated cell-cycle genes, F, all variable cell-cycle genes.

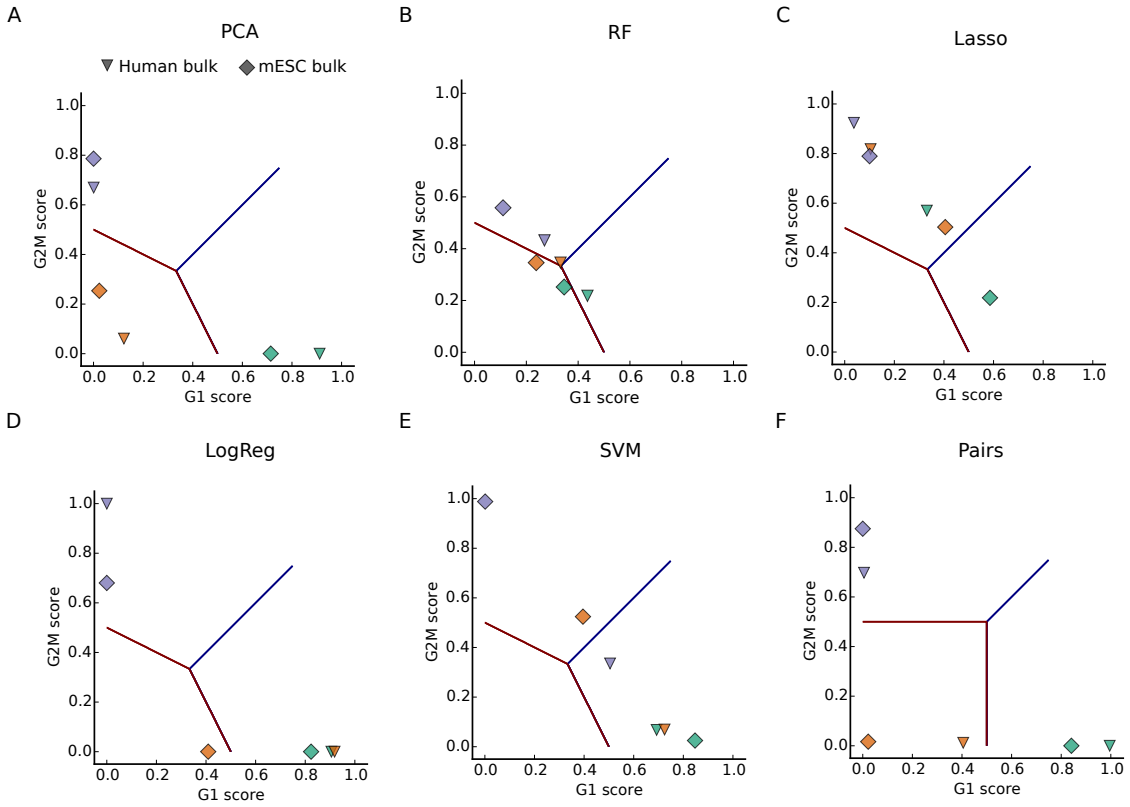

Figure 3: Application of the different prediction algorithms to bulk data with known cell-cycle stage for all six predictors. Bulk samples from mESCs are shown as diamonds, bulk samples from human myeloid leukemia cells are shown as triangles. Colours indicate true cell-cycle phase as in figure 2: G1 phase is shown in green, S-phase in orange and G2M phase in blue A, PCA-based method, B, random forest, C, Lasso, D, logistic regression, E, SVM, F, pairs. All predictors but the pairs method were trained on the informative set of annotated cell cycle genes, the pairs predictor was trained on all annotated cell cycle genes.

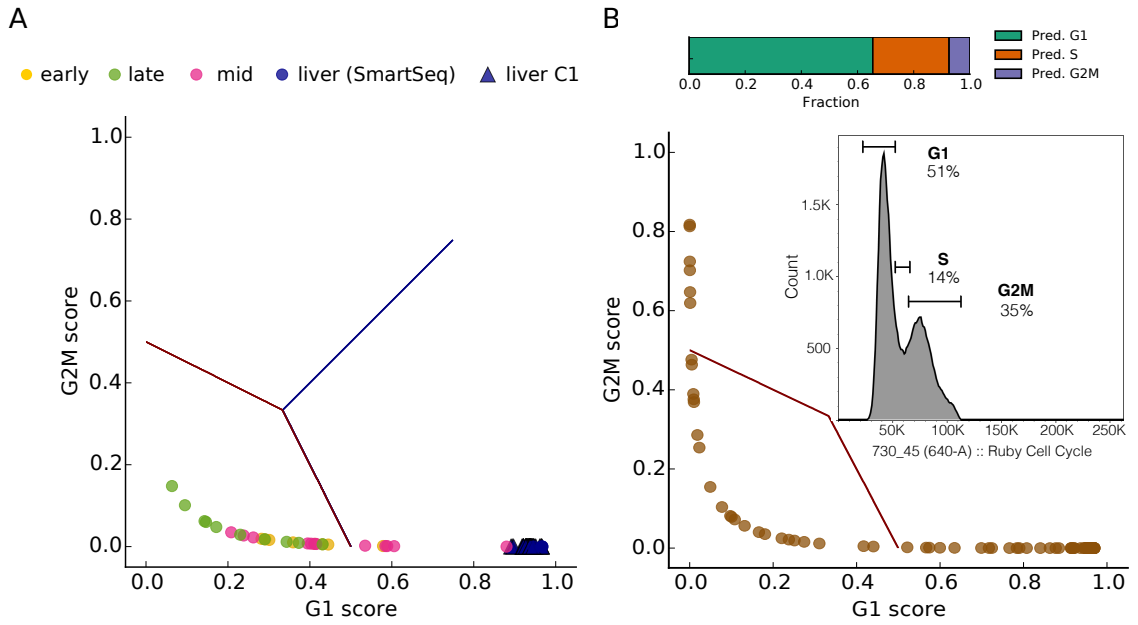

Figure 4: Application of the PCA-based approach to data without known cell-cycle stage. A, scatter plot of predicted G1 score and G2M score for single cells from the early, mid and late blastomere (yellow, pink, green circles) as well as individual liver cells from two different studies (dark blue circles and triangles). B, Scatter predicted G1 score and G2M score for the single T-cells. Top, bar plot showing relative fraction of cells predicted to be in G1, S and G2M phase. Inset, density plot of Ruby staining showing the relative fractions of cells in G1, S and G2M phase.

References

[1] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, John J Trombetta, David A Weitz, Joshua R Sanes, Alex K Shalek, Aviv Regev, and Steven A McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–14, May 2015.

[2] Oliver Stegle, Sarah A Teichmann, and John C Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet*, 16(3):133–45, Mar 2015.

[3] Liying Yan, Mingyu Yang, Hongshan Guo, Lu Yang, Jun Wu, Rong Li, Ping Liu, Ying Lian, Xiaoying Zheng, Jie Yan, Jin Huang, Ming Li, Xinglong Wu, Lu Wen, Kaiqin Lao, Ruiqiang Li, Jie Qiao, and Fuchou Tang. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nature structural & molecular biology*, 20(9):1131–9, September 2013.

[4] Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos Tanay, and Ido Amit. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science (New York, N.Y.)*, 343(6172):776–9, February 2014.

[5] Amit Zeisel, Ana B Muñoz Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, Charlotte Rolny, Gonçalo Castelo-Branco, Jens Hjerling-Leffler, and Sten Linnarsson. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, Feb 2015.

[6] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed

by pseudotemporal ordering of single cells. *Nat Biotechnol*, 32(4):381–6, Apr 2014.

[7] Siim Pauklin and Ludovic Vallier. The Cell-Cycle State of Stem Cells Determines Cell Fate Propensity. *Cell*, 155(1):135–147, September 2013.

[8] Michael B Kastan and Jiri Bartek. Cell-cycle checkpoints and cancer. *Nature*, 432:316–323, 2004.

[9] Ziv Bar-Joseph, Zahava Siegfried, Michael Brandeis, Benedikt Brors, Yong Lu, Roland Eils, Brian D Dynlacht, and Itamar Simon. Genome-wide transcriptional analysis of the human cell cycle identifies genes differentially regulated in normal and cancer cells. *Proc Natl Acad Sci U S A*, 105(3):955–960, January 2008.

[10] Lyubomir T Vassilev. Cell Cycle Synchronization at the G 2 /M Phase Border by Reversible Inhibition of CDK1. *Cell Cycle*, 5(22):2555–2556, October 2014.

[11] Tony Ly, Yasmeen Ahmad, Adam Shlien, Dominique Soroka, Allie Mills, Michael J Emanuele, Michael R Stratton, and Angus I Lamond. A proteomic chronology of gene expression through the cell cycle in human myeloid leukemia cells. *eLife*, 3:e01630, January 2014.

[12] Yohei Sasagawa, Itoshi Nikaido, Tetsutaro Hayashi, Hiroki Danno, Kenichiro D Uno, Takeshi Imai, and Hiroki R Ueda. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome biology*, 14(4):R31, January 2013.

[13] Asako Sakaue-Sawano, Hiroshi Kurokawa, Toshifumi Morimura, Aki Hanyu, Hiroshi Hama, Hatsuki Osawa, Saori Kashiwagi, Kiyoko Fukami, Takaki Miyata, Hiroyuki Miyoshi, Takeshi Imamura, Masaharu Ogawa, Hisao Masai, and Atsushi Miyawaki. Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. *Cell*, 132(3):487–98, February 2008.

[14] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and

Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, 33(2):155–160, January 2015.

[15] Iftach Nachman, Aviv Regev, and Sharad Ramanathan. Dissecting timing variability in yeast meiosis. *Cell*, 131(3):544–56, Nov 2007.

[16] Alberto Santos, Rasmus Wernersson, and Lars Juhl Jensen. Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. *Nucleic acids research*, 43(Database issue):D1140–4, January 2015.

[17] Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, David N Louis, Orit Rozenblatt-Rosen, Mario L Suvà, Aviv Regev, and Bradley E Bernstein. Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–401, Jun 2014.

[18] Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani. *The elements of statistical learning*, volume 2. Springer, 2009.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[20] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[21] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[22] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[23] Donald Geman, Christian d'Avignon, Daniel Q Naiman, and Raimond L Winslow. Classifying gene expression profiles from pairwise mrna comparisons. *Stat Appl Genet Mol Biol*, 3:Article19, 2004.

[24] Aik Choon Tan, Daniel Q Naiman, Lei Xu, Raimond L Winslow, and Donald Geman. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, 21(20):3896–904, Oct 2005.

[25] Bahman Afsari, Elana J Fertig, Donald Geman, and Luigi Marchionni. switchbox: an r package for k-top scoring pairs classifier development. *Bioinformatics*, 31(2):273–4, Jan 2015.

[26] Gene Ontology Consortium et al. The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(suppl 1):D258–D261, 2004.

[27] Nicholas Paul Gauthier, Lars Juhl Jensen, Rasmus Wernersson, Søren Brunak, and Thomas S Jensen. Cyclebase. org: version 2.0, an updated comprehensive, multi-species repository of cell cycle experiments and derived analysis results. *Nucleic acids research*, 38(suppl 1):D699–D702, 2010.

[28] Philip Brennecke, Simon Anders, Jong Kyoung Kim, Aleksandra A Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah A Teichmann, John C Marioni, and Marcus G Heisler. Accounting for technical noise in single-cell rna-seq experiments. *Nat Methods*, 10(11):1093–5, Nov 2013.

[29] Qiaolin Deng, Daniel Ramsköld, Björn Reinius, and Rickard Sandberg. Single-cell rna-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343(6167):193–6, Jan 2014.

[30] Agnes Klochendler, Noa Weinberg-Corem, Maya Moran, Avital Swisa, Nathalie Pochet, Virginia Savova, Jonas Vikeså, Yves Van de Peer, Michael Brandeis, Aviv Regev, Finn Cilius Nielsen, Yuval Dor, and Amir Eden. A transgenic mouse marking live replicating cells reveals in vivo transcriptional program of proliferation. *Dev Cell*, 23(4):681–90, Oct 2012.

[31] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. Htseq–a python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–9, Jan 2015.

[32] Maria A Ciemerych and Peter Sicinski. Cell cycle in mouse development. *Oncogene*, 24(17):2877–98, Apr 2005.

[33] Jérôme Artus and Michel Cohen-Tannoudji. Cell cycle regulation during early mouse embryogenesis. *Mol Cell Endocrinol*, 282(1-2):78–86, Jan 2008.

[34] Bidesh Mahata, Xiuwei Zhang, Aleksandra A Kolodziejczyk, Valentina Proserpio, Liora Haim-Vilmovsky, Angela E Taylor, Daniel Hebenstreit, Felix A Dingler, Victoria Moignard, Berthold Göttgens, Wiebke Arlt, Andrew N J McKenzie, and Sarah A Teichmann. Single-cell rna sequencing reveals t helper cells synthesizing steroids de novo to contribute to immune homeostasis. *Cell Rep*, 7(4):1130–42, May 2014.

Appendices

A Supplementary Methods

A.1 Detailed protocol for liver cells

Liver cell isolation protocol - 14 week old B6CastF1 (C57Bl/6J mother x CAST/Ei father) female mouse were anesthetized and perfused first with PBS, and then with Liver Perfusion Medium (Gibco, Life Technologies). The whole liver was dissected and transferred to $10mL$ of Liver Digest Medium (Gibco, Life Technologies). The liver was mechanically dissociated, and incubated at 37°C for 15-20 minutes with continuous shaking (225 rpm). Remaining tissue aggregates were gently triturated until dispersed, and the resulting liver cell suspension centrifuged for five minutes at 1000 rpm (at room temperature). The supernatant was carefully removed and the cells re-suspend in 10 mL of ice-cold Hepatocyte Wash Medium (Gibco, Life Technologies). The centrifugation and re-suspension steps were repeated twice additional times, and in the final re-suspension step, Hepatozyme (Gibco, Life Technologies) media was used. Finally the cells were filtered using a CellTrics 30 μM mesh filter (Partec), and kept on ice until loaded into the C1 IFC unit. All experiments involving mice were approved by the local ethical review committee, and a certificate of designation from the UK Home Office (the national authority for animal experimentation) was obtained.

Liver cell capture and library preparation for mouse cells using the Fluidigm C1 system - 5000 liver cells were loaded onto a 10-17 μm C1 Single-Cell Auto Prep IFC (Fluidigm), and cell capture was performed according to the manufacturer's instructions. The capture efficiency was inspected using a microscope, and there were single cells in 70 wells, cell debris in four wells, and more than one cell (or one cell plus debris) in 22 wells. The data from wells that contained more than one cell or debris was subsequently removed from analysis. Upon capture, reverse transcription and cDNA pre-amplification were performed in the 10-17 μm C1 Single-Cell Auto Prep IFC using the SMARTer PCR cDNA Synthesis kit (Clon-

tech) and the Advantage 2 PCR kit (Clontech). 1 μl of the ERCC Spike-In Control Mix (Ambion) in a 1:400 dilution in C1 Loading Reagent was added to the lysis mix for the liver cells. cDNA was harvested, quantified with the Bioanalyzer DNA High-Sensitivity kit (Agilent Technologies), and Nextera libraries prepared using the Nextera XT DNA Sample Preparation Kit and the Nextera Index Kit (Illumina) by following the instructions in the Fluidigm manual ("Using the C1 Single-Cell Auto Prep System to Generate mRNA from Single Cells and Libraries for Sequencing"). Libraries were pooled, and paired-end 100-bp sequencing was performed on one flow-cell (two lanes) of an Illumina HiSeq 2500.

A.2 The pairs method

We describe below the selection of marker pairs and the computation of the score for the G1 phase with the pairs method. An analogous procedure can be followed for the other two phases.

Selection of marker pairs - The training dataset was used to select pairs of genes whose relative expression levels differ in G1 phase compared to cells in S and G2M. In other words, we found pairs of genes, g_1 and g_2, such that:

$$\begin{aligned} g_1 - g_2 > 0 \quad &\text{in at least a fraction } f \text{ of G1 cells} \\ g_1 - g_2 < 0 \quad &\text{in at least a fraction } f \text{ of S and G2M cells} \end{aligned} \tag{1}$$

So in a marker pair, the first gene, g_1, is more highly expressed than g_2 in G1 cells, whereas the opposite happens in S and G2M cells. See Figure A.1 for an example of a G1 marker pair. These marker pairs can be used as indicators of the cell-cycle phase, as their relative expression level has a specific behaviour that changes with the phase.

Higher values of the fraction f result in the selection of fewer, more specific marker pairs, whereas lower values of f increase the number of marker pairs and decrease their overall specificity. We found that a good trade-off is reached with $f = 50\%$, around which the best performance is achieved as measured by 10-fold cross validation (see figure A.2).

Computing the score - Once the marker pairs for G1 have been selected from

the training dataset, they can be used to compute a "G1 score" for a given single cell, which is calculated as follows:

1. The number N_{G1} of "hits" in the list of G1 marker pairs is counted, i.e., the number of marker pairs where the first gene is expressed at a level higher than the second.

2. The probability distribution $\mathcal{R}(N)$ of the number of "hits" with randomised lists of marker pairs is obtained.

3. The G1 score is defined as the probability of getting a number of hits lower than N_{G1} with a randomised list of markers:

$$S_{G1} = \sum_{N \leq N_{G1}} \mathcal{R}(N) \tag{2}$$

Cell-cycle phase allocation - The scores S_{G1}, S_S, S_{G2M} for each of the three phases can be calculated following the procedure described above. Cells can be allocated to the phase that corresponds to the highest score.
However, we found that the method performs better when the allocation is carried out only on the basis of the G1 and G2M scores (see also appendix B.2). Therefore, if either S_{G1} or S_{G2M} is greater than 0.5, cells are allocated to the phase with the highest score among G1 and G2M. Conversely, if both $S_{G1}, S_{G2M} < 0.5$, the cell is allocated to the S phase.

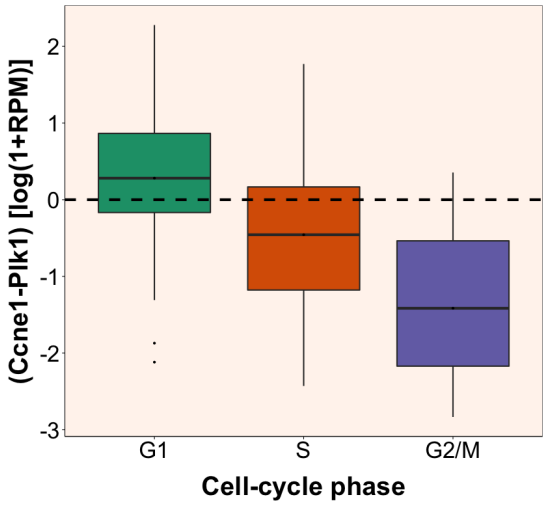

Figure A.1: Example of a G1 marker pair ($Ccne1$ and $Plk1$ genes).

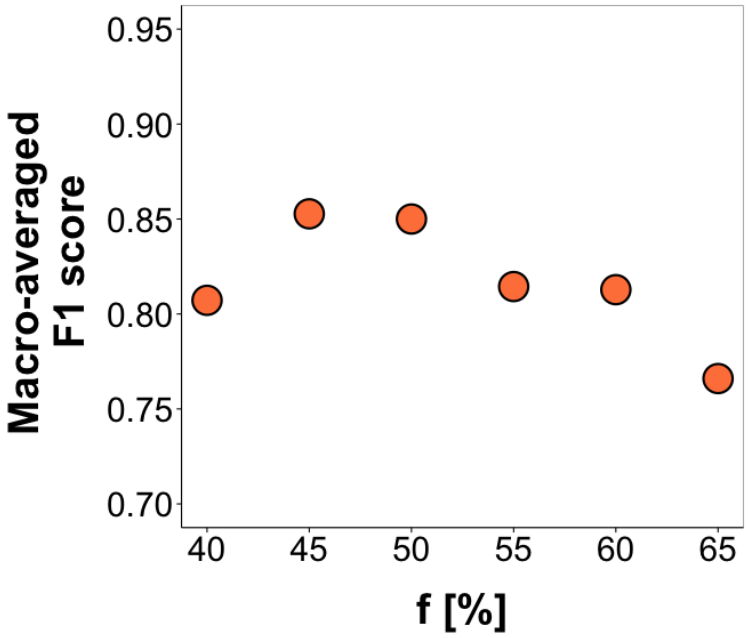

Figure A.2: The macro-averaged F1 score of the pairs method computed from internal 10-fold cross validation at different values of the parameter f.

B Supplementary Results

B.1 Allocation of data with no ground truth by using the pairs method

We used the pairs method to predict the cell-cycle phase of datasets with no ground truth, which we already discussed in the main text where we showed the results with the PCA-based method (fig. 4). The list of cell-cycle annotated genes was used for training.

Blastomeres - As expected, the vast majority of them is allocated to G2M (fig. B.1.A). Interestingly, the difference between early (yellow circles) and late (green circles) is also detected, as the average G2M score is higher for the late blastomeres compared to early and mid blastomeres.

Liver cells - Most of them predicted to be in G1, consistently with the expectation (fig. B.1.B).

T-cells - A percentage of $\sim$ 40% of cells are allocated to S and G2M phase (fig. B.1.C), which is consistent with the known proliferative state of these cells and with the results we obtained from the flow cytometry analysis of Ruby-stained cells (fig. 4.B).

B.2 The average expression levels of cell-cycle markers in different phases

We considered the cell-cycle markers listed in Cyclebase and checked their average expression levels in the cells in the different phases in our training dataset. A quantile normalization was carried out before calculating the averages.

Fig. B.2 shows the average normalized expression of G1, S and G2M markers in G1, S and G2M cells. While G1 and G2M markers peak respectively in G1 and G2M phases, the S markers are expressed approximately at the same level in G1 and S and do not have a clear peak in the S phase. This could explain why, in general, S phase cells are more difficult to identify from their transcriptome than cells in G1

and G2M.

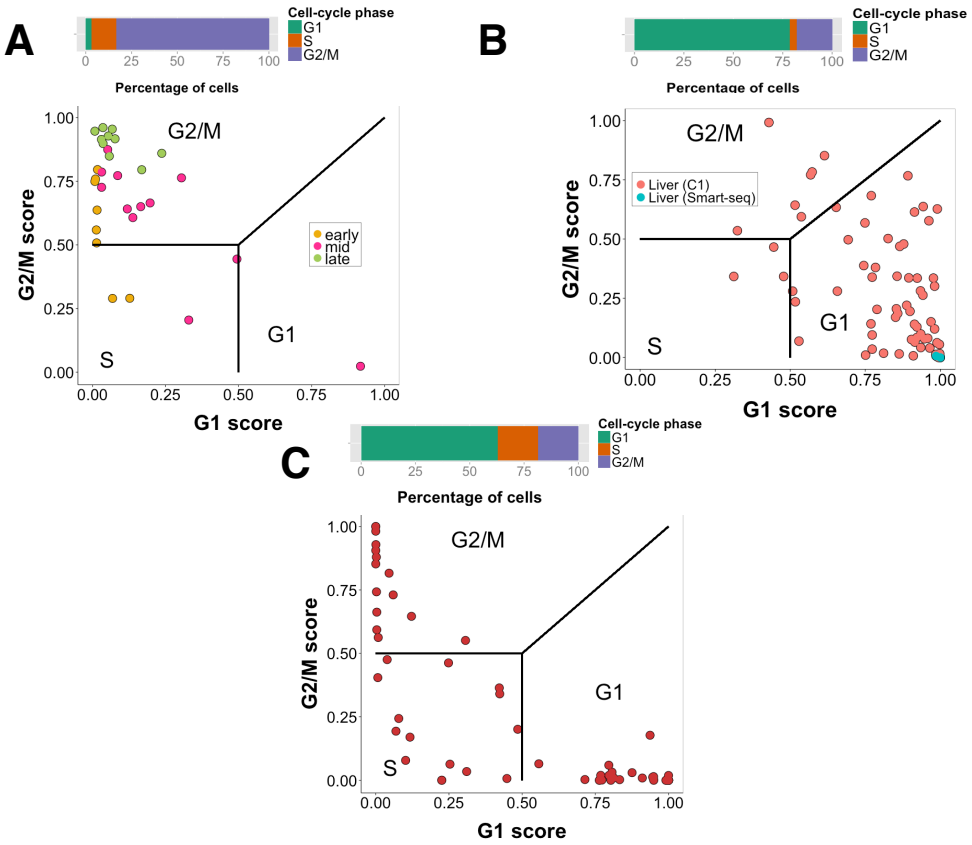

Figure B.1: Prediction of cell-cycle phases in single-cell datasets with no ground truth with the pairs method. The G1 and G2M scores are plotted. Continuous black lines mark the decision boundaries for the different phases. Bars on top show the percentages of cells in each phase. **Panel A -** Early, mid and late 2-cell blastomeres [29]. **Panel B -** Two independent sets of liver cells processed with Smart-seq [29] and C1 protocol. **Panel C -** Allocation of T-cells with the percentages of cells predicted in G1, S and G2M.

B.3 Feature importance

We used the loadings of PC1 as a measure for feature relevance in the PCA-based method. In figure B.4, the top 40 genes with the largest loadings are shown.

In the pairs method, we evaluated the relevance of each gene pair with a score com-

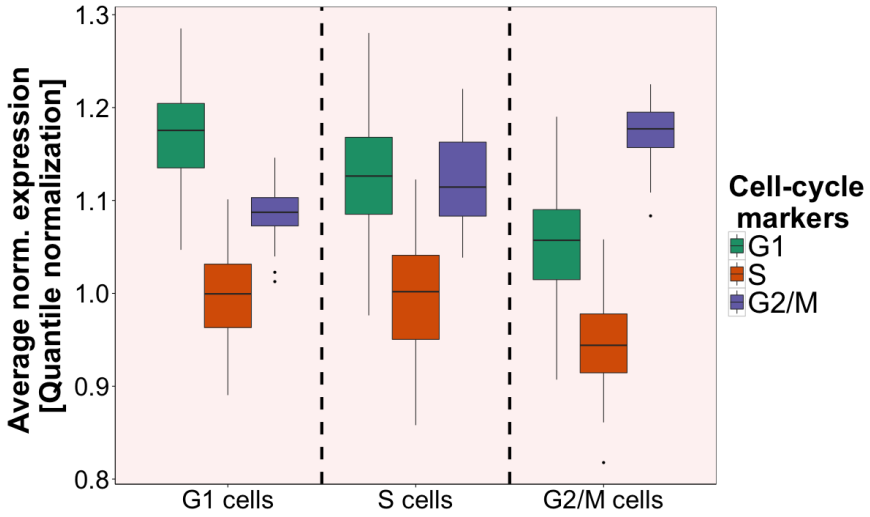

Figure B.2: Average normalized expression levels of G1, S and G2M markers in G1, S and G2M cells in the training dataset.

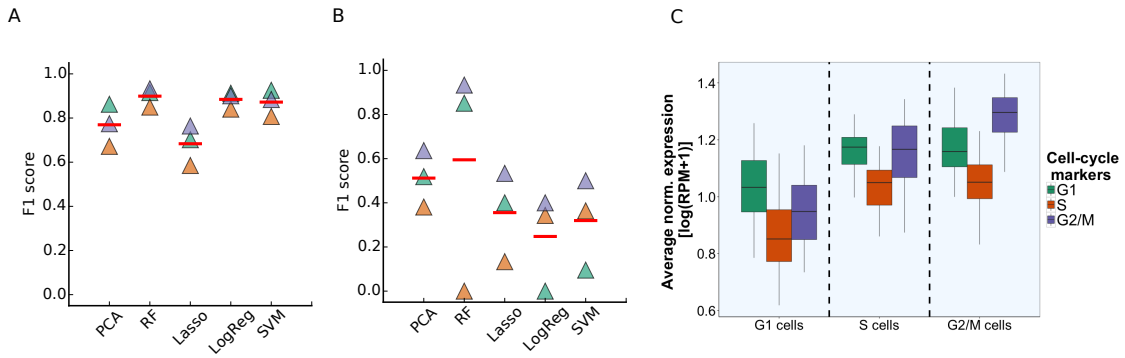

Figure B.3: Predictive power on the independent test set decreases when using a different normalisation strategy. A, cross-validated F1 score based on the set of informative/variable annotated cell cycle genes. B, F1 score on independent test data. C, Average normalized expression levels of G1, S and G2M markers in the training dataset. The gene cell cycle signature is considerably less pronounced compared to rank normalization (Suppl. Fig B.2).

puted as follows: for each given pair of genes g_i and g_j, we calculated the quantities $p_{ij}(C) = Prob(g_i > g_j)$, i.e., the fraction of samples in the training data set annotated to cell-cycle phase $C = \{G1, S, G2M\}$ where the expression level of gene

g_i is higher than that of gene g_j. The score of a marker pair (g_i, g_j) for a phase C_1 is defined as $\Delta_{ij} = |p_{ij}(C_1) - Mean(p_{ij}(C_2), p_{ij}(C_3))|$. Figure B.5 shows the ten disjoint pairs (i.e., pairs involving different genes, see [24]) with the highest scores in each cell-cycle phase.

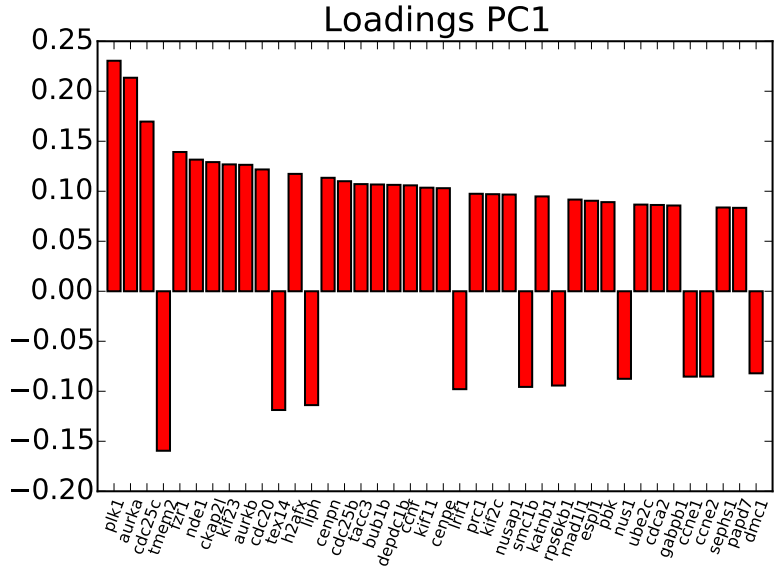

Figure B.4: Most relevant genes for the PCA-based method. As the loadings are derived from PC1 only, genes with positive loadings can be interpreeted as G2/M markers, genes with negative loadings can be interpreted as G1 (or S) markers.

B.4 Higher principal components do not improve predictive power

In addition to allocating cells based on the first PC only, we also assessed the predictive power of a classifier based on more PCs. In order to establish the relevant number of PCs, a scree plot can be generated and the number of PCs can be chosen analysing the gap between the (normalized) eigenvalues B.6. As there is a large gap between the first and the second eigenvalue, we show detailed results based on the first PC only in the main text. This is also illustrated in figure B.6 where the first PC is most informative, both for the training and the test data. We also computed

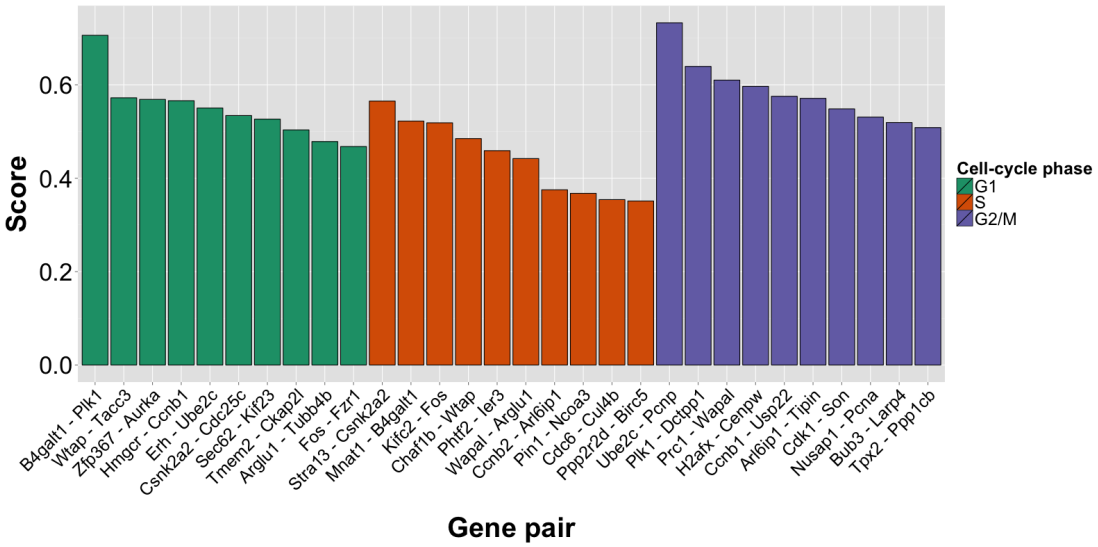

Figure B.5: Ten disjoint gene pairs with the highest score in each of the three phases. the cross-validated F1-score on the training data as well as the F1-score on the independent test data for up to five PCs. While the cross-validated F1-score varied little with the inclusion of more PCs, the performance on the test-set degraded in comparison to using only one PC, indicating that while the first PC captures generalizable cell-cycle effects, higher PCs capture more data-set specific properties. Similarly, we assessed how the performance of the other classifiers (random forest, SVM and logistic regression/lasso) would be affected when including the first 5 PCs as features and found no improvement in the predictive power (data not shown).

B.5 Training on the combined C1 and Quartz-Seq data

We reasoned that by increasing the size of the training data set we would be able to train a classifier with better generalizability. However, the drawback of using both data-sets for training is that no independent single-cell data with known cell-cycle stage set is available for external validation. Therefore, we evaluated the classifiers based on the combined C1 and Quartz-Seq data using 10-fold cross validation and observed similar performance as for using the C1 data for training only (fig. B.7 A). We then used the classifier trained on the combined mESC data to predict the cell-cycle stage for the same single-cell and bulk data-sets as described in the main text. This again yielded a similar performance as after training on the C1 data only (fig.

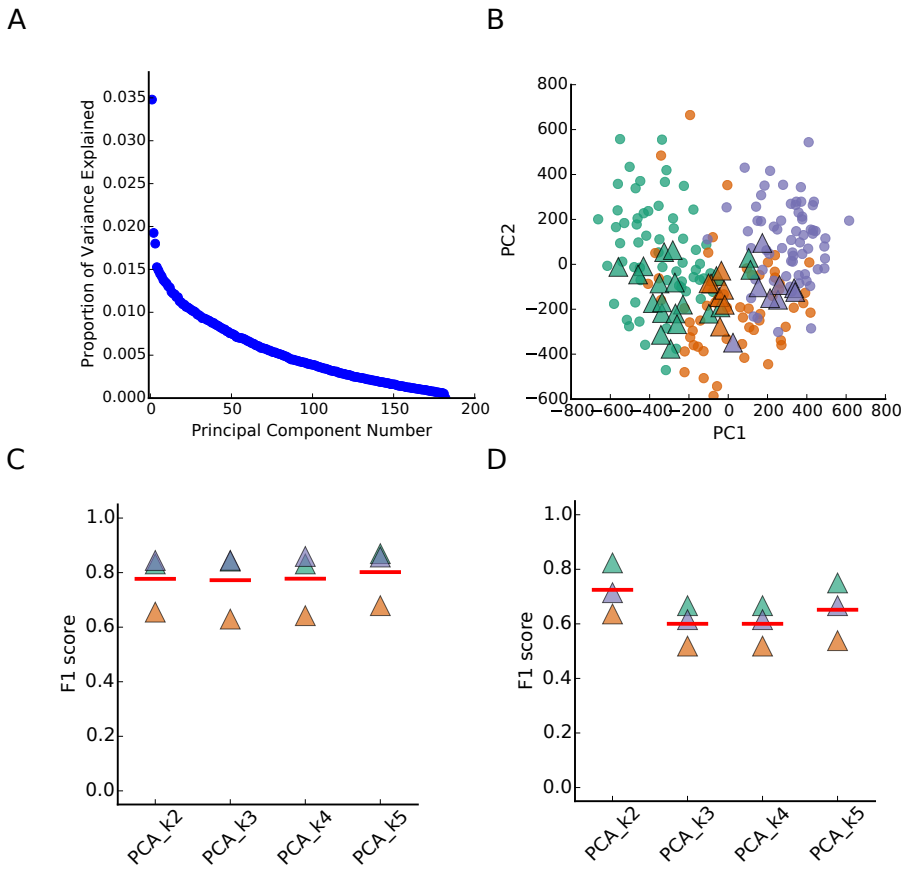

Figure B.6: A, scree plot for the informative set of annotated cell cycle genes. B, PCA of the training data for the informative set of annotated cell cycle genes (circles). The test data, shown in triangles, is projected into the PCA. C, Cross-validated F1 score for increasing number of PCs. D, F1 score on the test data for an increasing number of PCs.

B.7 B-C). For the PCA-based method we fitted the Gaussian Naive Bayes classifier on the first two PCs as some of the variability in PC1 was caused by the different media (and thus differences in pluripotency), all other classifiers were trained as described in the Methods section.

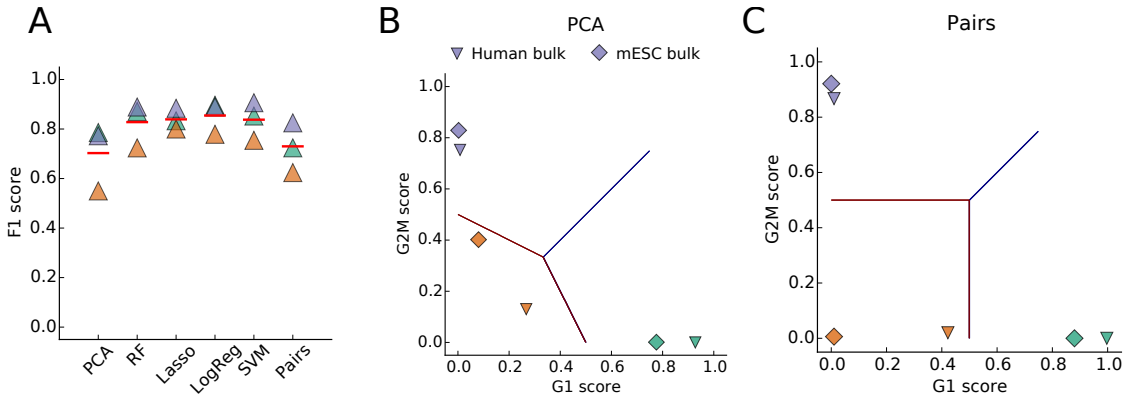

Figure B.7: A, F1 scores from internal cross validation for different gene sets; F1 score for G1 phase is shown in green, for S-phase in orange and for G2M phase in blue. Red lines represent the macro-averaged F1 score. Training was performed based on the set of variable cell-cycle genes for all methods but the pairs method, where the full list of cell-cycle annotated genes was used. B-C, Application to bulk data of the PCA-based (B) and the pairs method (C). Bulk samples from mESCs are shown as diamonds, bulk samples from human myeloid leukemia cells are shown as triangles. Colours indicate true cell-cycle phase as in figure 2: G1 phase is shown in green, S-phase in orange and G2M phase in blue.

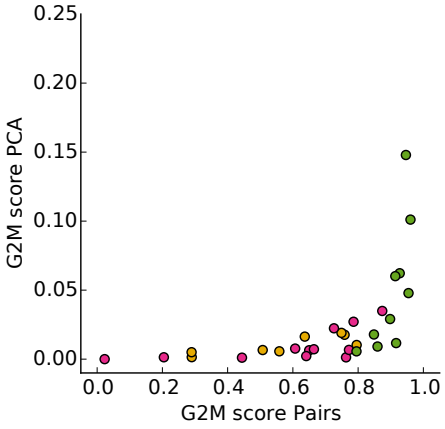

Figure B.8: Scatter plot for G2M score on the blastomeres data from PCA-based method and the pairs method. While absolute probabilities differ, the rank correlation was very high (0.81).