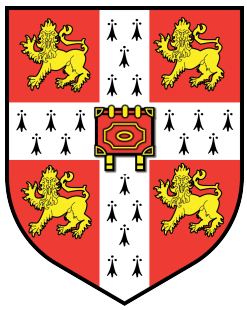
Single cell mRNA-sequencing of mESCs reveals cell-to-cell variation in pluripotency and cell cycle genes



Aleksandra Anna Kołodziejczyk

Trinity College
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy
May 2016

Summary

Cell culture conditions for embryonic stem cells are important for their selfrenewal capacity and for them to maintain pluripotency. Depending on the media that cells are cultured in, they exhibit different morphology and gene expression patterns. It was shown that ES cells cultured in 2i versus serum results in cells with more homogeneous morphology and more uniform Nanog expression.
I analysed the transcriptomes of over 700 individual mESCs cultured in three conditions (serum, 2i and alternative 2i) using full-transcript single cell RNAsequencing to understand the causes of culture medium-dependent differences in gene expression variability. I aimed to quantify and dissect the cell-to-cell variation in the three conditions in an unbiased way by high-throughput single cell mRNAsequencing and statistical data analysis in a way that was not possible before.
Firstly, I found that global levels of intercellular heterogeneity in gene expression are indistinguishable between conditions. At the same time, specific groups of genes (pluripotency genes in serum, cell cycle genes in 2i) do differ in their noise levels across culture conditions. The heterogeneity of pluripotency genes in the serumcultured mES cells is a consequence of subpopulations of cells that are differentiating away from the pluripotent state. In 2i and a2i-cultured cells, the transcriptomic heterogeneity originated in gene expression signatures of different cell cycle stages.
Secondly, I showed that the transcriptomic signatures of cells grown in the three media are distinct, with cells grown in 2i medium being most similar to the blastocyst cells of the early embryo.
Additionally, I found that differences in cell cycle genes' noise profiles correlate with proliferation rate, where slowly-cycling cells have broader, more noisy expression profiles and clearer separation between cells in G1/S and G2/M phases.
Moreover, I observed a previously described but poorly understood 2C-like population in 2i-cultured cells. I characterized this population in detail and compared it to in vivo data from early stages of mouse embryo development to determine whether it truly is equivalent to the embryonic 2-cell stage. I observed that these cells globally are more transcriptionally similar to blastocyst cells than cells from the 2-cell stage of the embryo.
Finally, I investigated the pluripotency gene regulatory network by analyzing correlations between transcription factors and chromatin-associated genes in the mouse ES cell data. I found two major clusters: pluripotency factors and differentiation regulators. In the pluripotency cluster, I identified new putative pluripotency regulators (Ptma, Zfp640, Zfp710). I validated these by knockdown with CRISPR repression technology, and demonstrated that even partial depletion of these genes causes a shift towards a more differentiated state.
Single cell RNA sequencing allowed me to look at cell populations and genes in the dataset to unravel cell identities and genes that regulate processes in these cells. This work highlights the power of single cell sequencing whilst providing data and analytical approaches that will be a useful resource for further study.

Declaration

The work presented in this dissertation was carried out at the MRC Laboratory of Molecular Biology, the EMBL European Bioinformatics Institute and the Wellcome Trust Sanger Institute between October 2012 and May 2016. This thesis is the result of my own work except when explicitly stated in the main text and is unlike any work I have previously submitted for any other qualification. This thesis does not exceed the word limit of 60,000 words required by the University of Cambridge School of Biological Sciences.

Aleksandra A. Kołodziejczyk
May 2016

Acknowledgements

Although the last four years were challenging, overall my time in Cambridge has made me really happy both personally and professionally and I have achieved more than I hoped for when I started my PhD. There are many people who made this happen and I feel I should thank.

First and foremost, I am deeply grateful to my supervisor Sarah Teichmann for guidance, support, patience, and trusting me. I was really lucky to be able to work in the exciting and emerging field of single cell genomics, which together with Sarah's optimism and insights allowed me to make new discoveries. Naturally, having a supervisor who taught me how to do scientific research both in and outside of the lab will impact my whole future career.

Secondly, I would like to thank John Marioni for his advice and discussions throughout our fruitful collaboration. I extend my gratitude to my collaborators: especially to Jong Kyoung Kim, for his invaluable help with bioinformatics and statistics; and to Jason Tsang, Xuefei Gao and Pentao Liu for all of the discussions about the biology of stem cells and for their help with the experimental parts of my project.

I would like to thank all current and previous members of the Teichmann lab at LMB, EBI and Sanger. I have to mention Tomislav Ilicic for assistance with various computational aspects of the project and his friendship, and Xiuwei Zhang for patience, teaching me basics of R and all Ragusa chocolate she brought from Switzerland. I am grateful to Tina Perica, Valentina Proserpio, Liora Haim-Vilmovsky and Bidesh Mahata for warm welcome and showing me around when we were still at the LMB. Furthermore, I am thankful to Kedar Natarajan for our discussions about stem cells, to Alex Tuck for sharing his data with me and to Johan Henriksson for turning my dataset into a database.

This thesis would not read well without the editing assistance and the 'unPolishing' of my English by Mike Stubbington.

This PhD would not have been possible without the support, love and hours on Skype with mom, dad and Paweł.

I would like to thank my friends: Łukasz Kopeć, Filip Szczypiński, Monika Folkierska-Żukowska, Julia Majewska and Kasia Wojtczak for their support, for visiting me, for telling me I do not have nine lives, for Cake/Vodka Mondays, and for words like "doktormatka". My friends from Trinity College: Rebecca Berrens, Ana Casanova, Annette LaRocco, Sun Lee, Tilman Flock, James Kane, Andreas Jakowetz, Matthew Dunstan, Janina Voigt and Tobias Schmidutz, who were always there for me and who helped to provide much needed fun in Cambridge and during our trips in Spain, Germany and Poland.

My life in Cambridge has been enriched by activities outside of my doctoral research, especially with my friends from CU Polish Society and Federation of Polish Societies in UK: Michał Włodarski, Dominika Kampa, Dominika Wolańska, Kasia Doniec, Tomek Cebo, Ola Pędraszewska, Marta Tondera, Czarek Łastowski and Kasia Rachuta. Thanks for the conferences, cultural events and fun we had organizing and running the society. Special thanks go to Tamás Sztanka-Tóth, for our Polish-Hungarian friendship.

I also need to thank Victor Sourjik (and all Sourjik lab members) and Dmitry Veprintsev, who inspired and encouraged me. Without them on my path I probably would not have attempted to pursue a PhD in the first place.

Finally, I would like to thank the BBSRC, Abcam and Wellcome Trust for funding to support my research.

Summary

Cell culture conditions for embryonic stem cells are important for their self-renewal capacity and for them to maintain pluripotency. Depending on the media that cells are cultured in, they exhibit different morphology and gene expression patterns. It was shown that ES cells cultured in 2i *versus* serum results in cells with more homogeneous morphology and more uniform *Nanog* expression.

I analysed the transcriptomes of over 700 individual mESCs cultured in three conditions (serum, 2i and alternative 2i) using full-transcript single cell RNA-sequencing to understand the causes of culture medium-dependent differences in gene expression variability. I aimed to quantify and dissect the cell-to-cell variation in the three conditions in an unbiased way by high-throughput single cell mRNA-sequencing and statistical data analysis in a way that was not possible before.

Firstly, I found that global levels of intercellular heterogeneity in gene expression are indistinguishable between conditions. At the same time, specific groups of genes (pluripotency genes in serum, cell cycle genes in 2i) do differ in their noise levels across culture conditions. The heterogeneity of pluripotency genes in the serum-cultured mES cells is a consequence of subpopulations of cells that are differentiating away from the pluripotent state. In 2i and a2i-cultured cells, the transcriptomic heterogeneity originated in gene expression signatures of different cell cycle stages.

Secondly, I showed that the transcriptomic signatures of cells grown in the three media are distinct, with cells grown in 2i medium being most similar to the blastocyst cells of the early embryo.

Additionally, I found that differences in cell cycle genes' noise profiles correlate with proliferation rate, where slowly-cycling cells have broader, more noisy expression profiles and clearer separation between cells in G1/S and G2/M phases.

Moreover, I observed a previously described but poorly understood 2C-like population in 2i-cultured cells. I characterized this population in detail and compared it to *in vivo* data from early stages of mouse embryo development to determine whether it truly is equivalent to the embryonic 2-cell stage. I observed that these cells globally are more transcriptionally similar to blastocyst cells than cells from the 2-cell stage of the embryo.

Finally, I investigated the pluripotency gene regulatory network by analysing correlations between transcription factors and chromatin-associated genes in the mouse ES cell data. I found two major clusters: pluripotency factors and differentiation regulators. In the pluripotency cluster, I identified new putative pluripotency regulators (*Ptma*, *Zfp640*, *Zfp710*). I validated these by knockdown with CRISPR repression technology, and demonstrated that even partial depletion of these genes causes a shift towards a more differentiated state.

Single cell RNA sequencing allowed me to look at cell populations and genes in the dataset to unravel cell identities and genes that regulate processes in these cells. This work highlights the power of single cell sequencing whilst providing data and analytical approaches that will be a useful resource for further study.

Publications

1. Kolodziejczyk AA, Kim JK, Tsang JCH, Ilicic T, Henriksson J, Natarajan KN, Tuck AC, Gao X, Bühler M, Liu P, Marioni JC, Teichmann SA (2015) Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. Cell Stem Cell 17 (4), 471-485.

2. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA (2015) The Technology and Biology of Single-Cell RNA Sequencing. Molecular cell 58 (4), 610-620.

3. Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy D, Marioni JC, Teichmann SA. (2016) Classification of low quality cells from single cell RNA-seq data. Genome Biology 17 (1), 1.

4. Kim JK, Kolodziejczyk AA, Ilicic T, Teichmann SA, Marioni JC (2015) Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. Nat. Commun 6, 8687.

5. Tsang JCH, Yu Y, Burke S, Buettner F, Wang C, Kolodziejczyk AA, Teichmann SA, Lu L, Liu L. (2015) Single-cell transcriptomic reconstruction reveals cell cycle and multi-lineage differentiation defects in Bcl11a-deficient hematopoietic stem cells. Genome Biology 16 (1), 1-16.

6. Mahata B, Zhang X, Kolodziejczyk AA, Proserpio V, Haim-Vilmovsky L, Taylor AE, Hebenstreit D, Dingler FA, Moignard V, Göttgens B, Arlt W, McKenzie ANJ, Teichmann SA (2014) Single-cell RNA sequencing reveals T helper cells synthesizing steroids de novo to contribute to immune homeostasis. Cell Rep. 22;7(4):1130-42.

7. Brennecke P, Anders S, Kim JK, Kolodziejczyk AA, Zhang X, Proserpio V, Baying B, Benes V, Teichmann SA, Marioni JC, Heisler MG (2013) Accounting for technical noise in single-cell RNA-seq experiments. Nature Methods 10: 1093–1095.

Contributions

This thesis is the result of my own work except:

- CV^2 plots showing technical noise on Figure 1.8 were generated by Dr Jong Kyoung Kim.
- Microscopy pictures on Figure 3.5 were taken from Roeder and Radtke, 2009.
- Cell culture was done in collaboration with Dr Jason Cheuk-Ho Tsang and Dr Xuefei Gao.
- NPC differentiation dataset was kindly shared with us by Dr Alex Tuck.
- mRNA sequencing libraries of bulk controls (not CRISPR experiment) were done by Wellcome Trust Sanger Institure sample preparation pipeline team.
- dCas9-Krab and hyPBase plasmids as well as PB-gRNA-BsaI plasmid backbone are kind gift of Dr Xuefei Gao.
- Gene expression heterogeneity measurement using DM was developed by Dr Jong Kyoung Kim.
- Plots showing DM level on Figure 3.8 and Figure 3.9 were generated by Dr Jong Kyoung Kim.
- Single cell sequencing data batch correction was done by Dr Jong Kyoung Kim.

Table of contents