

Computational methods for identifying cell type, signalling pathways and gene regulatory networks

Md Humayun Kabir

A thesis submitted in fulfilment of the requirements for the
degree of Doctor of Philosophy

School of Medicine
Western Sydney University
October 2018

Acknowledgements

I wish to express my sincere gratitude to my supervisor Dr. Michael O'Connor for his sincere support, direction and encouragement throughout the course of my research. His patience, flexibility and guidance throughout the time enabled me to complete my PhD smoothly.

I also wish to express my sincere gratitude to my co-supervisor Dr. Joshua W.K. Ho and for his invaluable expertise, sincere guidance and encouragement for completing this thesis.

I would like to express my gratitude to Dr. Djordje Djordjevic for his kind support and motivation during the initial stage of my PhD journey. Also I would like to thank all members of the Ho lab for their kind support to me. It is my privilege to thank all the previous and present regenerative medicine lab members for their friendship and co-operation.

I am grateful to the Australian government and Western Sydney University for funding the WSU PRA (International) scholarship during my PhD studies.

I would like to thank my teachers, relatives and friends in Bangladesh for their continuous support and encouragement from the other side of the world.

I am very much grateful to my wife and my little daughter for being by my side always and supporting me to complete this long PhD journey. I like to thank my younger sisters Jesmin and Mona for their endless support and making this all happen. Finally, I would like to thank my parents from the bottom of my heart for giving me life and then every opportunity I could ever hope for.

This thesis is dedicated to my grandparents.

Statement of Authentication

The work presented in this thesis is, to the best of my knowledge and belief, original except as acknowledged in the text. I hereby declare that I have not submitted this material, either in full or in part, for a degree at this or any other institution.

A handwritten signature in black ink, appearing to read 'H. Kabir' with a stylized flourish at the end.

Md Humayun Kabir

October 2018

Synopsis

Due to the advancement of genetic technologies it is now possible to accurately and simultaneously measure the expression levels of essentially all genes for species that have had their genomes sequenced. A detailed molecular understanding of these gene expression data will facilitate identification of new drug and cell therapy targets for disease treatment. There are increasing amounts of genome-wide public gene expression data that provide valuable information for the research community. The underlying hypothesis for this thesis is that development of new bioinformatics methods for both cell type and signal pathway characterization – that make use of public gene expression data – will facilitate application of stem cells and their differentiated derivatives. Each chapter in this thesis had specific aims. Chapter 1 presents a literature review that aimed to overview current bioinformatics tools applicable to robust definition of stem cell and differentiated cell identity. Chapter 2 aimed to develop a new tool (termed C3) for identifying unknown cell types based on their transcriptional profile. Chapter 3 aimed to characterize a new populations of purified cells (termed ROR1⁺ cells) obtained from differentiation of human pluripotent stem cells. Chapter 4 aimed to develop a new method (termed SPAGI) for predicting active signalling pathways for any cell type, based on the cell's transcriptome. Chapter 5 aimed to provide a large-scale prediction (or blueprint) of signal pathway-mediated regulation of lens epithelial cell gene expression. Lastly, the final chapter provides an overall reflection on the thesis combined with perspectives on future work made possible as a result of the thesis.

Manuscripts arising from this thesis

Published

1. Patricia Murphy*, Md Humayun Kabir*, Tarini Srivastava*, Michele E. Mason*, Chitra U. Dewi, Seakcheng Lim, Andrian Yang, Djordje Djordjevic, Murray C. Killingsworth, Joshua W. K. Ho, David G. Harman and Michael D. O'Connor, "**Light-focusing human micro-lenses generated from pluripotent stem cells model lens development and drug-induced cataract in vitro**", *Development* (2018), 145, dev155838. doi:10.1242/dev.155838, * co first authors

Author contributions

Conceptualization: M.D.O.; Methodology: P.M., M.H.K., A.Y., D.D., M.C.K., J.W.K.H., D.G.H., M.D.O.; Software: M.H.K., A.Y., D.D., J.W.K.H.; Validation: P.M., M.H.K., D.G.H., M.D.O.; Formal analysis: P.M., M.H.K., T.S., M.E.M., C.U.D., S.L., M.D.O.; Investigation: P.M., M.H.K., T.S., M.E.M., C.U.D., S.L., M.C.K., J.W.K.H., D.G.H., M.D.O.; Resources: M.D.O.; Data curation: J.W.K.H.; Writing - original draft: P.M., M.D.O.; Writing - review & editing: P.M., M.H.K., T.S., M.E.M., C.U.D., S.L., A.Y., D.D., M.C.K., J.W.K.H., D.G.H., M.D.O.; Visualization: P.M., M.H.K., T.S., M.E.M., C.U.D., S.L., J.W.K.H., M.D.O.; Supervision: M.D.O.; Project administration: M.D.O.; Funding acquisition: M.D.O.

2. Md Humayun Kabir, Djordje Djordjevic, Michael D. O'Connor, Joshua W. K. Ho, "**C3: An R package for cross-species compendium-based cell-type identification**", *Computational Biology and Chemistry* 77 (2018) 187-192. <https://doi.org/10.1016/j.compbiolchem.2018.10.003>

Author contributions

J.W.K.H. initiated the project; M.H.K. designed the method, implemented the package, performed evaluation and wrote the manuscript; D.D. contributed to method design and software testing; M.D.O.C and J.W.K.H. supervised the whole project and revised the manuscript. All authors read and approved the final manuscript.

3. Md Humayun Kabir, Ralph Patrick, Joshua W. K. Ho, Michael D. O'Connor, "**Identification of active signaling pathways by integrating gene expression and protein interaction data**", *BMC Systems Biology* 2018, 12(Suppl 9):120. <https://doi.org/10.1186/s12918-018-0655-x>

Author contributions

J.W.K.H and M.D.O.C conceived the SPAGI approach and supervised the project. M.H.K designed the method, implemented the R package, carried out the experiments and wrote the manuscript. R.P carried out critical evaluation and testing. All authors revised and approved the final version of the manuscript.

4. Md Humayun Kabir and Michael D. O'Connor, "**Stems cells, big data and compendium-based analyses for identifying cell types, signalling pathways and gene regulatory networks**", Biophysical Reviews (2019) 11:41–50
<https://doi.org/10.1007/s12551-018-0486-4>

Author contributions

M.H.K drafted the manuscript. M.H.K and M.D.O'C revised and approved the manuscript.

Under revision

5. Md Humayun Kabir, Patricia Murphy, Seakcheng Lim, Joshua W. K. Ho, Michael D. O'Connor, "**Large scale profiling of lens epithelial cell signalling pathways and target genes reveals regulatory networks for cataract-associated genes**", under revision in Experimental Eye Research (YEXER_2018_622)

Author contributions

M.D.O'C. conceived the project. J.W.K.H. and M.D.O'C supervised the project. M.H.K. performed the bioinformatics analysis. P.M. and S.L. performed the cell culture and Western blotting. All authors revised and approved the manuscript.

List of presentations

Oral presentation

1. Md Humayun Kabir, Djordje Djordjevic, Joshua W.K. Ho, Michael O'Connor, **Bioinformatics validation of human iPSC-derived ROR1+ cells as lens epithelial cells**, presented at ARVO-Asia 2017 conference

Fast forward presentation

1. Md Humayun Kabir, Djordje Djordjevic, Michael D. O'Connor, Joshua W. K. Ho, **C3: An R package for cross-species compendium-based cell-type identification**, presented at Sydney Bioinformatics Research Symposium (SBRS) 2017

Poster presentation

1. Md Humayun Kabir, Djordje Djordjevic, Michael D. O'Connor, Joshua W. K. Ho, **C3: An R package for cross-species compendium-based cell-type identification**, presented at Sydney Bioinformatics Research Symposium (SBRS) 2017
2. Md Humayun Kabir, Djordje Djordjevic, Michael D. O'Connor, Joshua W. K. Ho, **C3: An R package for cross-species compendium-based cell-type identification**, presented at Graduate Research School HDR showcase 2017 WSU, NSW, Australia

Table of Contents

Chapter 1: Stems cells, big data and compendium-based analyses for identifying cell types, signalling pathways and gene regulatory networks	1-12
- Introduction	3
- Stem cells enable molecular characterisation of human biology	3
- Molecular profiling using big data	4
- Current big data analysis tools	5
- Characterising pluripotency mechanisms using big data	5
- Big data repositories for defining cell identity	6
- Compendium-based methods for defining cell identity	6
- Compendium-based analyses for stem cell research	7
- Investigating extracellular regulation of cell behaviour	7
- Signal transduction pathways and target genes	8
- Bioinformatic determination of signal pathways	9
- Linking signal pathways and TG sets	9
- Defining disease mechanisms by integrating signal pathways and disease genes	9
- Conclusion	9
- References	10
 Chapter 2: C3: An R package for cross-species compendium-based cell-type identification	 13-21
1. Introduction	15
2. Methods	16
2.1. C3: a new R package for cross-species cell-type identification	16
2.2. The human and mouse gene expression compendia	16
2.3. Identification of specifically expressed genes in the query and compendium data	16
2.4. XGSA	17
2.5. Comparison with ExpressionBlast	17
3. Results	17
3.1. Evaluation of C3	17
3.2. Comparison with other similar software programs	19
4. Discussion	19
5. Conclusion	20
- References	20

Chapter 3: Light-focusing human micro-lenses generated from pluripotent stem cells model lens development and drug-induced cataract in vitro	22-49
- Introduction	24
- Results	24
- Discussion	28
- Materials and Methods	32
- References	34
- Supplementary material	36
 Chapter 4: Identification of active signaling pathways by integrating gene expression and protein interaction data	 50-71
- Background	52
- Methods	53
- Building background pathway data	53
- Housekeeping genes identification	55
- Potential signaling pathway identification	55
- Ranking of the potential signaling pathways	55
- Assessment of SPAGI false positive rate	55
- Results	56
- SPAGI analysis of tooth	56
- SPAGI analysis of lens gene expression data	57
- Comparison of SPAGI analysis on species-specific vs combined PPI data	58
- Analysis of the SPAGI false positive rate via random expression level assignment	59
- Analysis of the SPAGI false positive rate versus GO analysis	60
- Discussion	60
- Conclusions	60
- References	61
- Additional file	63
 Chapter 5: Large scale profiling of lens epithelial cell signalling and gene expression networks reveals regulatory pathways for known cataract genes	 72-116
1 Introduction	76
2 Materials and methods	78
2.1 Acquiring transcriptional and PPI datasets	78
2.2 Identification of house-keeping genes	78

2.3	Establishing a universe of known R/K/TR paths and pathways	79
2.4	Identification and ranking of LEC-specific paths and pathways	79
2.5	Identification of TG sets and cataract-associated genes for LEC-expressed TRs	80
2.6	Gene ontology and promoter analyses	80
2.7	Cell culture	80
2.8	Western blotting	81
3	Results and discussion	81
3.1	Lens signalling pathways defined through gene expression and PPIs	81
3.2	Mapping TGs, including cataract-associated genes, to the LEC blueprint	83
3.3	Ranking LEC pathways highlights pervasive as well as niche critical lens signalling pathways	84
3.4	Identification of overlapping and potentially niche roles for LEC signalling pathways	85
3.5	The LEC blueprint accurately predicts known roles for lens signalling pathways	87
3.6	The LEC blueprint accurately predicts known LEC transcriptional regulation events	87
3.7	Different lens TRs regulate specific subsets of cataract-associated genes	88
3.8	A new, large LEC gene regulatory network involving known critical lens regulators	89
3.9	A higher-level transcriptional network of 10 TRs predicted to control LECs	90
3.10	The LEC blueprint ascribes relevant functional roles to lens TRs and signalling pathways	90
3.11	Additional lens gene regulatory networks remain to be discovered	91
3.12	ELK1: confirmation the LEC transcriptional blueprint identifies new lens biology	92
4	Conclusion	93
-	References	94
Chapter 6: General discussion		117-123
-	Summary of outcomes from this thesis	118
-	Advances in cell type identification	118
-	Advances in identification of active signalling pathways	119
-	A more detailed molecular understanding of lens epithelial cell biology	121
-	Concluding remarks	122
-	References	169

Chapter 1

**Stems cells, big data and compendium-based
analyses for identifying cell types, signalling
pathways and gene regulatory networks**

The aim of this thesis was to develop and apply new methods to analyse gene expression data in order to identify uncharacterised cell types and to identify potential active signalling pathways en masse. The thesis is presented in manuscript format, with each publication status listed on the next page. The progression of manuscripts is arranged such that preceding chapters demonstrate the need, and provide background for, subsequent chapters.

The first manuscript is a review that describes the importance of bioinformatics analyses of large-scale transcriptomic and other data sets for progressing the field of stem cell biology. The manuscript is described in this chapter. This manuscript provides the rationale for why cell type characterisation and signalling pathway identification are important research areas.

The second manuscript describes a new method (termed C3) for Cross-species Compendium-based Cell-type identification from a gene expression profile. The C3 method was implemented by developing an open source R package. Extensive validation studies showed the method is applicable to cell type identification from a gene expression profile for a wide variety of species. One suitable application is the identification of purified but poorly characterised cell populations obtained from differentiating stem cells.

The third manuscript describes application of the C3 algorithm as part of a larger study to characterise ROR1⁺ cells derived from human pluripotent stem cells. The C3 method was applied to RNA-seq data obtained from ROR1⁺ cells, and showed the purified cells to be most similar to primary human lens epithelial cells. This finding was supported by additional bioinformatics studies including principal component analysis as well as extensive cell biology-based characterisation techniques.

The fourth manuscript describes a new method, termed SPAGI for Signal Pathway Analysis for Gene regulatory network Identification. The SPAGI method utilizes gene expression and protein-protein interaction data and is executed as an open source R package. It outputs a ranking of signal paths – each consisting of receptor(s), kinases, and transcriptional regulators – with paths grouped as receptor-defined pathways. The goal of the method was to identify active signalling pathways en masse from microarray or RNA-seq gene expression data. The method was validated using gene expression data sets from a variety of cell types.

The fifth manuscript provides an in-depth characterisation of a published newborn mouse lens epithelial cell dataset using the SPAGI method. The results of the SPAGI analysis were extended by comparison with lens epithelial cell target genes identified from sequencing data generated through the Fantom5 consortium. This analysis generated an interconnected, lens epithelial cell transcriptional blueprint of signalling pathways and associated target genes. Comparison of these target genes with the known cataract-associated genes identified 3 new gene regulatory networks and associated signal pathways predicted to control the networks.

The thesis is completed by a General Discussion that provides perspectives on future research areas that have arisen from the advances made during this thesis.



Stems cells, big data and compendium-based analyses for identifying cell types, signalling pathways and gene regulatory networks

Md Humayun Kabir^{1,2} · Michael D. O'Connor^{1,3} 

Received: 23 October 2018 / Accepted: 15 November 2018 / Published online: 25 January 2019
© International Union for Pure and Applied Biophysics (IUPAB) and Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract

Identification of new drug and cell therapy targets for disease treatment will be facilitated by a detailed molecular understanding of normal and disease development. Human pluripotent stem cells can provide a large in vitro source of human cell types and, in a growing number of instances, also three-dimensional multicellular tissues called organoids. The application of stem cell technology to discovery and development of new therapies will be aided by detailed molecular characterisation of cell identity, cell signalling pathways and target gene networks. Big data or ‘omics’ techniques—particularly transcriptomics and proteomics—facilitate cell and tissue characterisation using thousands to tens-of-thousands of genes or proteins. These gene and protein profiles are analysed using existing and/or emergent bioinformatics methods, including a growing number of methods that compare sample profiles against compendia of reference samples. This review assesses how compendium-based analyses can aid the application of stem cell technology for new therapy development. This includes via robust definition of differentiated stem cell identity, as well as elucidation of complex signalling pathways and target gene networks involved in normal and diseased states.

Keywords Pluripotent stem cell · Bioinformatics · Compendium · Signalling · Growth factor · Pathway · Gene regulatory network

Introduction

All somatic cells in a multicellular organism such as humans contain the same DNA. However, each normal distinct cell type within the organism only expresses a subset of the available genome required for proper functioning of that particular cell type (Ralston and Shaw 2008). Expression of particular sets of target genes (TGs) is regulated by a range of transcriptional regulators (TRs) including transcription factors and histone modifiers (Hoopes 2008; Ralston and Shaw 2008). Disease states typically involve acquisition of abnormal cellular transcriptional profiles that, in turn, alter cell phenotypes and function, for instance, during tumorigenesis.

This article is part of a Special Issue on ‘Big Data’ edited by Joshua WK Ho and Eleni Giannoulitou.

✉ Michael D. O'Connor
m.oconnor@westernsydney.edu.au

¹ School of Medicine, Western Sydney University, Campbelltown, NSW, Australia

² Department of Computer Science and Engineering, University of Rajshahi, Rajshahi, Bangladesh

³ Medical Sciences Research Group, Western Sydney University, Campbelltown, NSW, Australia

Maturation of cellular phenotype and function occurs through the interplay between environmental cues—sensed, for example, via growth factor receptors—and transcriptional changes that take place within the cell (Hoopes 2008; Ralston and Shaw 2008). For most cell type/external cue combinations, little molecular detail is known either of the molecular events that lead to transcriptional changes or the breadth of TGs changes that occur. Greater detail of these processes is recognised as a key frontier for the development of new therapies for a broad range of diseases (Berg 2016). Thus, there is a compelling need to identify TG sets that are regulated by particular signalling pathways and environmental factors, in order to better characterise the development and maintenance of cellular phenotypes, behaviours and biological processes. This information will also greatly facilitate improved understanding of how these events become dysregulated in ageing and disease.

Stem cells enable molecular characterisation of human biology

Historically, the inability to access large amounts of normal and diseased human tissue—particularly during the early

stages of disease initiation—significantly impeded efforts to define cell identity at a molecular level. The scarcity of human tissues has also hindered efforts to define how environmental cues alter cell biology and disease progression.

Significant genomic and functional similarities exist between human cells and tissues compared to those of other species. Consequently, many different animal models have been developed to try and progress investigation of normal and disease development. While valuable knowledge has been gained through decades of animal studies, the ability for animal models to specifically predict treatment responses in human patients is questionable (Shanks et al. 2009). This has led both academic researchers and the pharmaceutical industry to investigate human stem cells as an alternative source of information for both basic research and drug discovery (Cressey 2012; O'Connor 2013).

Human pluripotent stem (PS) cells offer a unique opportunity to rapidly progress our understanding of how environmental cues modulate signalling cascades and TG sets. This is due to key properties of human PS cells (O'Connor 2013; O'Connor et al. 2011a; Ungrin et al. 2007), including the ability to:

- 1) Self-renew (i.e., proliferate while retaining developmental potential), thereby enabling production of extremely large numbers of human cells in vitro
- 2) Differentiate into essentially any desired human cell type for research and clinical applications
- 3) Enable simple and highly targeted gene modification through technologies such as Crispr/Cas9
- 4) Obtain both normal and disease-specific human PS cells, either from donated IVF embryos (i.e., embryonic stem cells, or ES cells), by cell reprogramming (i.e., induced pluripotent stem cells) or by genome modification of these PS cell types
- 5) Directly model human biology without confounding species-specific differences that can arise through studies of animal models

As a result of these properties, use of human PS cell technology has become widespread. For example, in 2010 GE Healthcare announced the commercial availability of human ES cell-derived cardiomyocytes. These PS cell-derived cells provided a readily available and biologically relevant alternative to animal models and primary cells for cardiac drug discovery and toxicity testing.

Realising the full academic, industrial and clinical potential of human PS cells will require application of big data or 'omics' techniques to overcome major challenges that face the field. These challenges include (i) improving culture manipulations for optimal PS cell maintenance and directed differentiation, (ii) development of efficient cell purification strategies, and (iii)

establishment of robust characterisation assays for differentiated cell types.

Overcoming these challenges will require defining the similarities between differentiated cell types and desired primary cell types. This will include assessment of the developmental maturity of differentiated cells as relates to their phenotypes and functions, as well as the molecular events required to achieve and maintain cell phenotypes and functions. Doing so will provide both minimal characterisation criteria for reproducible production of desired differentiated cell types, and also a molecular framework for disease investigation and drug target discovery.

Molecular profiling using big data

Transcriptional changes that result from environmental cues occur via activation and/or repression of specific TG sets. Historically, investigations of signalling pathways and related TGs developed from the discovery of recombinant DNA technology (Cohen et al. 1973) and the ability to genetically modify mice and other organisms. Initial characterisation technologies for these studies included PCR, histology and electron microscopy. While these initial approaches yielded useful information, limited molecular detail of affected signalling pathways or TG sets was obtained.

The development of big data techniques for transcriptomics (from spotted arrays and microarrays to RNA-sequencing, also known as RNA-seq) (Bumgarner 2013) and proteomics (particularly mass spectrometry) (Han et al. 2008) enabled much higher resolution characterisation of the molecular changes that link environmental sensing, signal transduction and affected TG sets. Additionally, traditional immunoprecipitation techniques—that provide evidence of protein interactions through antibody-based protein capture—have been coupled with both microarray analysis and DNA sequencing. For example, chromatin immunoprecipitation (ChIP) techniques (termed ChIP-chip and ChIP-seq, respectively) enable interactions between proteins and DNA to be defined with high resolution of the chromosomal location (Furey 2012; Mardis 2007). Both ChIP-chip and ChIP-seq assays have been widely used with cell lines and animal tissue to determine the chromosomal location of post-translationally modified histones, histone variants, transcription factors and chromatin modifying enzymes (Bailey et al. 2013; Collas 2010).

Computational approaches have also been developed to investigate TG regulation by TRs. This has largely been driven by the capacity for genome-wide assessment of DNA-binding motifs within gene promoters, as a consequence of sequencing the human genome. Algorithms such as PASTAA (Roeder et al. 2009), Homer (Heinz et al. 2010), GeoSTAN (Zacher et al. 2017), iRegulon (Janky et al. 2014) and compendium-based approaches (Banks et al. 2016) are

examples of software that use different approaches to predict TG regulation by transcription factors. As these methods are evolved, the accuracy of TG predictions increases. Combinations of sequencing and computational-based approaches have also been developed for identification of TG regulation by TRs. For example, cap analysis of gene expression (CAGE) data generated through the Fantom5 consortium has provided sequencing data from the 5' region of mRNA transcripts (as opposed to traditional 3' sequencing approaches) for 975 human and 399 mouse cell samples (Andersson et al. 2014; Consortium et al. 2014). Computational analysis of this data has been used to predict TRs responsible for regulation of large sets of TGs across many cell types (Marbach et al. 2016).

Current big data analysis tools

The above technical and technological advances mean it is now possible to accurately and simultaneously measure the expression levels of essentially all genes for species that have had their genomes sequenced. It is also possible to begin interrogating the TRs involved in generating gene expression profiles, through ChIP-seq and or computational analyses. Alternatively, mass spectrometry enables simultaneous measurement of the levels of many thousands of proteins.

A variety of open source and proprietary software has been developed to analyse whole transcriptome expression data. For example, Gene Pattern (Broad Institute) (Reich et al. 2006) and GeneSpring (Agilent) for microarray data; limma for both microarray and RNA-seq gene expression data (Ritchie et al. 2015); and EdgeR (Robinson et al. 2010) for RNA-seq data. These different softwares enable identification of differentially expressed genes related to developmental and/or disease states. However, it should be noted that sequencing-based approaches, such as RNA-seq, tend to be better suited for identification of expressed vs. non-expressed genes, as opposed to identification of only differentially expressed genes. This is due to the digital nature of transcript detection by sequencing techniques, compared to the analogue nature of microarray based techniques (that typically rely on fluorescent-based methods for transcript detection, thereby making determination of absolute expression cut-off thresholds challenging).

Transcriptome analysis software can generate lists of expressed and/or differentially expressed genes from either new whole transcriptome data or reanalysis of published studies. These gene lists then provide insights into the signalling pathways and TGs involved in development or function of normal tissue, as well as pathways and TGs altered by disease states. A commonly used approach to investigate differentially expressed gene lists is identification of gene groupings via gene ontology (GO) analysis.

Various GO analysis software are available including the DAVID Gene Ontology Functional Annotation Clustering tool (Huang et al. 2009a, b), Enricher (Kuleshov et al. 2016), GO-Bayes (Zhang et al. 2010), Babelomics (Medina et al. 2010), etc. Alternatively, assessment of expressed growth factor signalling pathway members can be performed by comparison of gene lists against the Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway database (Kanehisa and Goto 2000).

Characterising pluripotency mechanisms using big data

Transcriptional, translational and ChIP profiling studies have been performed using cell lines and primary cells/tissue, and more recently using stem cells and their differentiated derivatives. For example, landmark studies have highlighted genes that are highly expressed across multiple human PS cell lines, thus identifying core transcriptional machinery consisting of the transcription factors OCT4/POU5F1, NANOG and SOX2 (Boyer et al. 2005; Cloonan et al. 2008; Hirst et al. 2007). These studies have also identified some TGs of these key pluripotency TRs. Additional studies have identified other human PS cell regulators including FOXD3, SALL4, Polycomb-group proteins, etc. (Lee et al. 2006; O'Connor et al. 2011b; Respuela et al. 2016).

Through comparison with mouse ES cell transcriptional data, these human studies provided a molecular framework for understanding the different culture requirements for PS cells obtained from different species. For instance, while mouse and human ES cells are both obtained from fertilised embryos, maintenance of mouse ES cells is LIF-dependent and FGF-independent. Conversely, human ES cells are LIF-independent and FGF-dependent. Transcriptional profiling studies have helped provide an explanation for these observations. The initially isolated mouse ES cell state is now recognised as a developmentally earlier state termed the 'naïve' pluripotency state. In contrast, the initially isolated human ES cell state is now termed the 'primed' pluripotency state that is analogous to pluripotent cells that can be isolated from the mouse epiblast. Naïve human ES cells can be transitioned between the naïve and primed pluripotency states (Chen et al. 2015; Duggal et al. 2015; Warrier et al. 2017), raising the possibility of obtaining naïve human ES cells directly from blastocysts (Van der Jeught et al. 2015). As naïve PS cells may enable better control of differentiated cell production, the transcriptomics studies described here provide evidence that big data might facilitate improvement and application of stem cell technology.

Big data repositories for defining cell identity

A major challenge for the stem cell field is the reliable production and characterisation of desired differentiated cell types. Cell-type identification via a whole transcriptome gene expression profile can provide a relatively rapid, broad and reasonably cost-effective approach. Accurate cell-type identification is needed to enable better manipulation of differentiated cells in culture (e.g., by identifying growth factor requirements), and also to provide a framework for understanding the molecular events that occur in a disease state.

Transcriptional and/or translational analyses typically involve characterisation of a control sample with or without comparison to treated sample(s) generated through chemical or genomic perturbations. Time-course components are also often included. The vast number of transcriptional and translational studies performed over the past 15 years has led to the establishment of large data repositories to facilitate public access to gene and protein expression data. Examples of public repositories for gene expression data include the Gene Expression Omnibus (GEO) that accepts data from any species (Barrett et al. 2013); human and mouse data available via the ENCODE consortium (Consortium TEP 2012; Consortium TME 2012); and human data available via GTEx (Consortium GT 2013). Protein data repositories include UniProt (Consortium TU 2007) and STRING (von Mering et al. 2003). These public gene and protein expression data repositories can provide compendia for more comprehensive/more robust cell-type identification for differentiated PS cell progeny.

Compendium-based methods for defining cell identity

Discovery of new biology by comparison of a test gene expression profile against a larger collection (i.e., compendium) of expression profiles has been used for almost two decades (Fig. 1a). However, compendium-based analyses have not yet been widely used by the stem cell field, despite the opportunity for robust cell type identification through compendia (Fig. 1a–c).

Two general approaches have been used for compendium-based cell-type identification: those that use a somewhat limited gene set as the query and those that use a larger expression profile as the query (DeFreitas et al. 2016). Compendium-based approaches can also be further divided into those that enable within-species comparisons and (less frequent) those that enable cross-species comparisons. For example, SPELL enables within-species identification (only for yeast) from a limited gene set against large gene expression microarray compendia (Hibbs et al. 2007). Alternatively, GEMINI uses a large transcriptome profile to query for similar profiles but

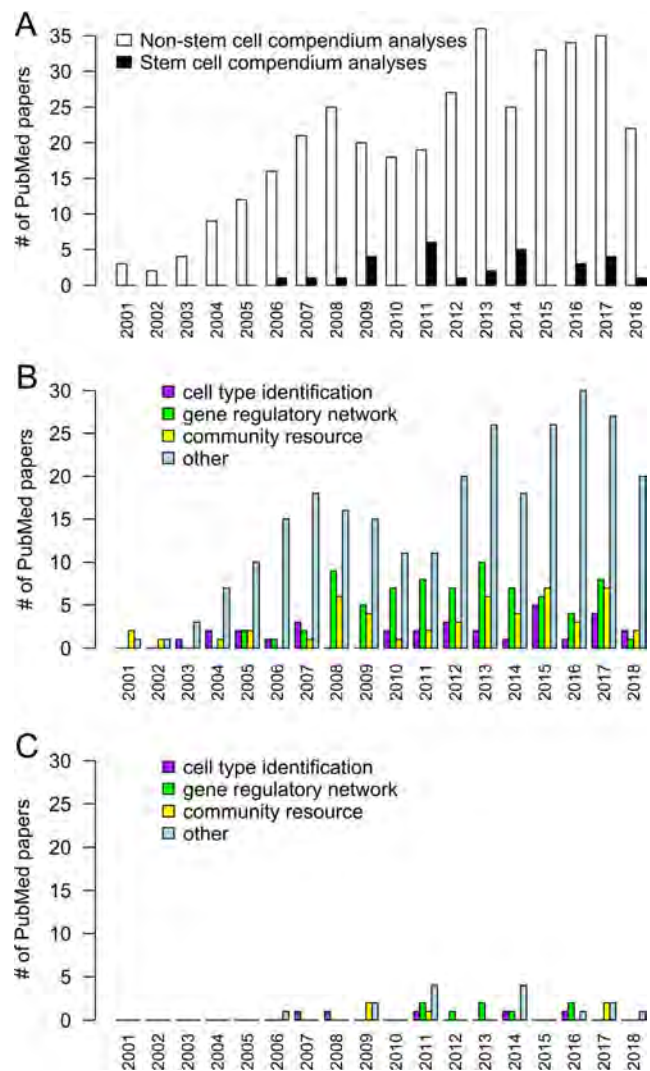


Fig. 1 Increases in the number of published articles making use of compendium-based analyses (as identified via PubMed searches). **a** The number of articles using compendium-based analyses for both non-stem cell types (white bars) and stem cell types (black bars). **b** An indication of the number of publications using particular compendium-based applications for analysis of non-stem cell types. **c** Publications using particular compendium-based applications for analysis of stem cell types

only within “The Cancer Genome Atlas” database (DeFreitas et al. 2016). GEMINI uses a principal component analysis to reduce the dimensionality of the query transcriptome, and then uses a distance function to search for the closest match within the compendium. It does not support cross-species comparisons.

A small number of compendium-based approaches that enable cross-species cell-type identification have recently been described. For example, the web server ProfileChaser mines only the curated GEO datasets for gene expression profiles that differentially regulate the same transcriptional programs as the query profiles (Engreitz et al. 2011). Another web server that matches query gene sets (to a maximum of 100 differentially expressed genes) by searching the GEO

database is ExpressionBlast. Required inputs are the limited query gene list together with their expression comparison values, a species type, a desired output species type and a distance metric (Euclidean/correlation/anti-correlation/anti-Euclidean). The algorithm then uses text analysis methods to perform similarity matching of the query gene set against GEO datasets. ExpressionBlast then outputs the relevant GEO datasets that similarly express the same genes as the query gene list (Zinman et al. 2013). The web server Cell Montage permits searching for similar gene expression profiles compared to a query gene profile (Fujibuchi et al. 2007). The method is platform specific (i.e., specific to similar microarray platforms) and also only allows users to query against GEO datasets that contain raw expression values.

Compendium-based analyses for stem cell research

A small number of groups have started utilising the compendium-based approach for stem cell research (Fig. 1c). For example, Germanguz et al. used a compendium-based approach consisting of 17 cell state-specific gene expression data (including PS cells) to identify genes that uniquely define cell states and developmental stages. They also identified core genes (including transcription factors) that can drive and maintain the cell states (Germanguz et al. 2016). StemCellNet is a web server for interactive network analysis and visualisation in the context of stem cell biology (Pinto et al. 2014). HAEMCODE is a repository of transcription factor binding maps for mouse blood cells generated by ChIP-seq (Ruau et al. 2013). Asp et al. generated a dataset of genome-wide locations for ten key histone marks and transcription factors. By using mouse myoblasts and terminally differentiated myotubes, they were able to discover key epigenetic changes underlying myogenesis (Asp et al. 2011). Hannah et al. described a ChIP-Seq compendium to discover transcriptional mechanisms operating in the haematopoietic system (Hannah et al. 2011). Sharov et al. identified a reliable set of direct TGs for Pou5f1, Sox2 and Nanog by utilising a compendium of published and new microarray data (Sharov et al. 2008). Hackney and Moore built a compendium of information and data derived from biological and molecular studies relating to haematopoietic stem cell regulation (Hackney and Moore 2005).

The above compendium-based stem cell studies tended to either compare multiple cell types or identify a specific cell type. These approaches are not optimised for identification of an unknown cell type. In comparison, a new open source R package developed by our group, termed C3, allows cross-species identification of any cell type. C3 uses a large transcriptomic profile rather than a limited gene list, and is

compatible with a wide variety of input compendia (Kabir et al. 2018a). The cross-species comparison enabled by C3 makes use of a recently developed cross-species gene set analysis method called XGSA (Djordjevic et al. 2016). C3 can identify unknown cell types for a wide variety of species by comparing gene expression profiles with a large compendium of public human and mouse gene expression datasets. This approach is suitable for identification of poorly characterised cell types obtained from stem cell differentiation strategies (Murphy et al. 2018). In this way, C3 fits well into the pipeline of cell analyses needed by the stem cell field (Fig. 2).

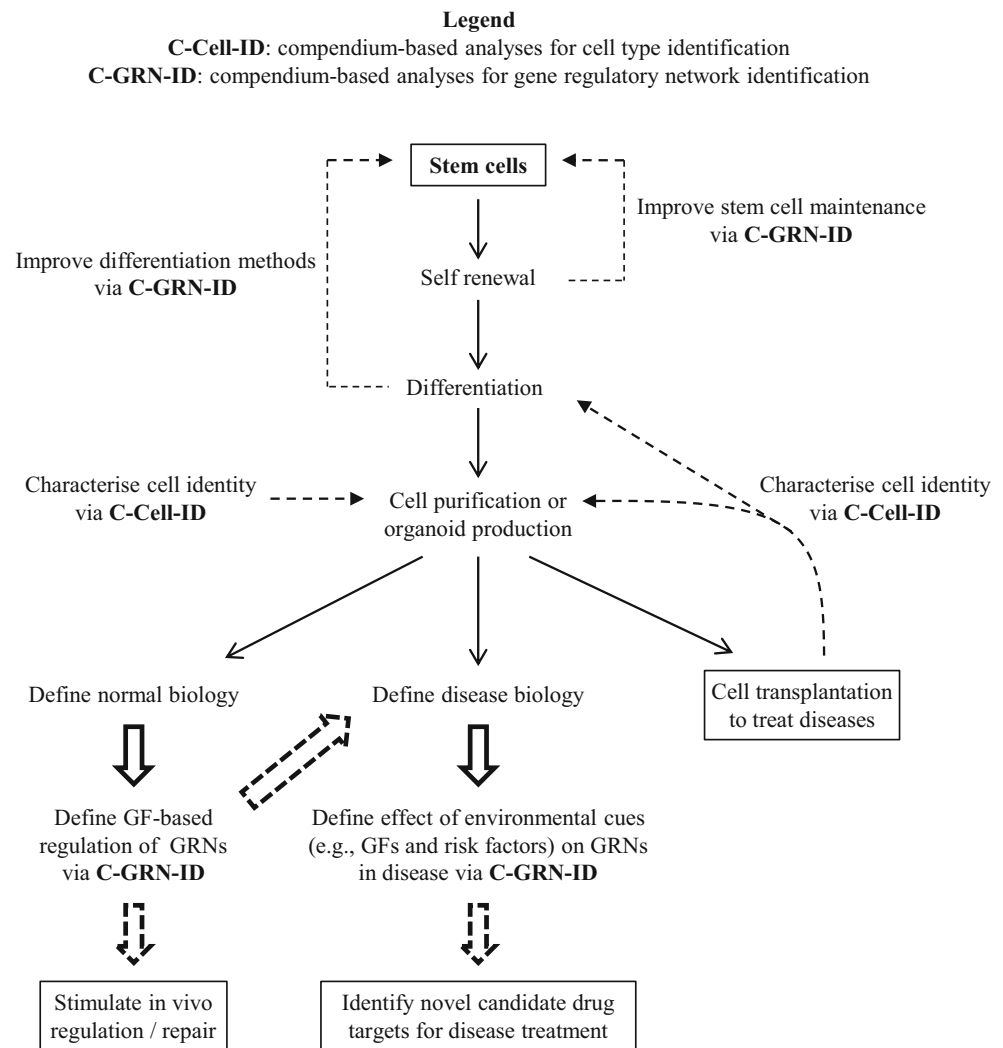
In addition to identification and characterisation of differentiated stem cell progeny, transcriptional profiles are also being used to guide stem cell differentiation strategies. For example, a recently published algorithm called MOGRIFY uses gene expression data to predict TRs responsible for generating cell type-specific transcriptional profiles (and thus cell-specific phenotypes and functions) (Rackham et al. 2016). These cell type-specific combinations of TRs can then be used to guide overexpression studies aimed at directly converting (i.e., trans-differentiating) one cell type into another.

Investigating extracellular regulation of cell behaviour

A second major challenge for the stem cell field, and disease research in general, is to define how extracellular signalling pathways regulate transcriptional events required for cell development, environmental sensing and disease progression (Berg 2016; Zhang and Mallick 2013). At the genome level, gene transcription is often activated or repressed by the action of transcription factors (also referred to as trans-regulatory factors) that bind to promoter regions generally upstream (i.e., 5') of a gene's transcription start site (termed cis-regulatory elements). The specific DNA sequences within the genome to which transcription factors bind are called DNA-binding motifs and are often described via position weight matrices (Babu et al. 2004; Boeva 2016; Spitz and Furlong 2012).

Transcriptional and translational profiles represent molecular snapshots that result from the combined action of an array of transcriptional, post-transcriptional and translational regulators, often under extracellular control via signalling pathways. Individual gene transcript abundance is largely determined by the net activity of the transcription factors bound to a gene's promoter (Beer and Tavazoie 2004; Chen and Rajewsky 2007; Kim and O'Shea 2008)—though other regulators of transcript abundance can also be involved such as transcriptional regulators acting at more distance (e.g., enhancer) sites and post-transcriptional regulators (such as micro-RNA). Overall, the ability of any particular transcription factor to activate or repress gene expression is dependent upon

Fig. 2 Schematic diagram showing how compendium-based analyses can be used to accelerate application of stem cell technology to identification and testing of new drug and cell-based therapies



the interplay between the intracellular context and regulatory cues received from the extracellular environment, for instance via growth factor signalling pathways.

Signal transduction pathways and target genes

As discussed above, a range of computational tools have been developed to elucidate gene regulatory networks by defining transcription factor/TG interactions (e.g., PASTAA, Homer, GeoSTAN, iRegulon, etc.). Sequencing approaches that target the 5' end of mRNA transcripts, such as CAGE, have also been developed. Significant recent progress has been made by applying these approaches within large, international collaborative efforts. For example, the Fantom5 consortium generated CAGE data across 975 human and 399 mouse samples, including primary cells, tissues and cancer cell lines (Andersson et al. 2014; Consortium et al. 2014). From these data, TG sets for transcription factors expressed by 394 human cell samples

have been defined via analysis of DNA-binding motifs within gene promoters and enhancers (Marbach et al. 2016).

While the above approaches have provided a wealth of information on transcription factor/TG interactions, there are relatively few open source or proprietary algorithms that exist for comprehensively linking signal pathways to TG sets. A typical signal transduction pathway for transmitting extracellular cues involves growth factors binding to specific cell surface receptors, subsequent modulation of intracellular kinase activities, and ultimately altered transcription factor activity and consequent changes in TG expression (Wang et al. 2011). The coordinated activity of different signalling pathways within and between multiple cell types is the basis of many important biological processes, such as development, tissue repair and immunity (Zhao and Li 2017; Zhao et al. 2008). Activation of different signalling pathways can lead to numerous physiological or cellular responses, such as cell proliferation, differentiation, metabolism and death—key processes relevant to stem cells and their progeny both in vitro and in vivo.

Bioinformatic determination of signal pathways

Various resources have been created to assist in defining signalling pathways. The collection of manually drawn pathway diagrams available via KEGG provides a starting point for understanding particular receptor-mediated signalling pathways. However, their use can be limiting when attempting to define cell type-specific signalling pathways. Conversely, the STRING database contains millions of known protein-protein interactions (PPIs); however, accessing cell type-specific subsets of these interactions can be challenging.

Several bioinformatics methods have been described that reconstruct known signalling pathways from PPI data, with or without inclusion of gene expression data (Bebek and Yang 2007; Gil et al. 2017; Ritz et al. 2016; Wang et al. 2011). CASCADE_SCAN uses a steepest descent method to build a specific pathway from a list of protein molecules (Wang et al. 2011). Pathlinker creates signal pathways by using input receptors and transcriptional regulators to interrogate PPI databases (Gil et al. 2017; Ritz et al. 2016). PathFinder uses characteristics of known signal pathways together with related association rules to find pathways from a receptor to a transcription factor in PPI networks (Bebek and Yang 2007). Gitter et al. proposed a method to handle the orientation problem (i.e., orienting protein interaction edges using directionless PPI data) in weighted protein interaction graphs (Gitter et al. 2011). Mei et al. proposed a multi-label, multi-instance transfer learning method to simultaneously reconstruct 27 human signalling pathways (Mei and Zhu 2015). Scott et al. proposed a method to reconstruct known signalling pathways by applying a colour coding algorithm (Scott et al. 2006). Tuncbag et al. formulated a forest approach (defined as a disjointed union of trees) to simultaneously reconstruct multiple pathways from biological networks that are altered in a particular condition (Tuncbag et al. 2013). Other methods identify known signalling pathways using gene expression datasets to calculate edge weights for PPI data (Liu and Zhao 2004; Steffen et al. 2002; Zhao and Li 2017; Zhao et al. 2008).

Linking signal pathways and TG sets

All the above methods for signal pathway analysis generate topological structures for known signalling pathways. One potential limitation is that most of the methods were assessed and applied only to yeast data, with few methods designed for complex mammalian data. Recent work from our group has demonstrated a novel approach—termed SPAGI (Signal Pathway Analysis for Gene regulatory network Identification)—that systematically identifies biologically relevant signalling pathways for mammalian cells (Kabir et al.

2018c). The SPAGI approach starts with a whole transcriptome expression profile and uses it to construct a comprehensive catalogue of signalling pathways from PPI data. Application of the SPAGI approach to mouse and human cell RNA-seq data, including from differentiated progeny of human PS cells, identified known critical signalling pathways relevant to the cell types used. Subsequent research using human lens epithelial cell gene expression data has coupled each of the SPAGI-generated receptor-defined paths to TG sets obtained from the Fantom5 consortium data (Kabir et al. 2018b). The resulting lens epithelial cell gene expression framework (or lens transcriptional blueprint) describes growth factor-mediated control of transcriptional programs important to lens epithelial cell biology. Initial validation studies have shown that known gene regulatory interactions were identified, and predicted new transcriptional regulators were validated via Western blotting. This approach directly addresses a major challenging in the stem cell and disease research fields, namely, the need for large-scale generation of discrete and testable molecular hypotheses that describe the influence of environmental factors during tissue development and disease progression (Fig. 2).

Defining disease mechanisms by integrating signal pathways and disease genes

A key motivation driving the establishment of integrated signalling pathways and TG networks is the need to better define disease processes to enable identification of novel drug targets (Butcher et al. 2004; Davidson et al. 2002). Information relating to genes and gene variants involved in disease phenotypes can be found within the Online Mendelian Inheritance in Man (OMIM) database (Hamosh et al. 2005). Tissue-specific disease gene databases also exist for numerous tissues including the kidney, heart, muscle, brain, lens, etc. By correlating the abovementioned lens transcriptional blueprint with the Cat-Map database of lens-related disease genes (Shiels et al. 2010), our group has been able to identify both known and novel gene regulation events and map them to growth factor signalling pathways (Kabir et al. 2018b). This approach can also be applied to other cell types, including differentiated stem cell derivatives, to define candidate drug targets—and therefore candidate novel therapeutics—for human diseases (as outlined in Fig. 2).

Conclusion

Stem cells provide an opportunity to examine normal and disease human biology on a scale not possible with primary cells and tissues. Realising these opportunities requires overcoming specific challenges relating to determination of cell

type identity, and definition of how environmental cues including growth factor signalling pathways regulate gene transcription involved in tissue development, repair/regeneration and disease. Compendium-based analyses hold promise for rapid and robust identification of first-reported differentiated stem cell types, as well as batch-produced cells for industry or cell therapy applications. Bioinformatic methods that generate comprehensive and integrated combinations of signalling pathways and gene regulatory networks are starting to provide specific molecular disease hypotheses that can be investigated using human PS cell-derived cell types. Thus compendium-based big data approaches to stem cell research present significant opportunities for the development of novel cell and drug therapies.

Author contributions M.H.K drafted the manuscript. M.H.K and M.D.O'C revised and approved the manuscript.

Funding M.H.K was supported by WSU Postgraduate Research Awards. M.D.O'C was supported by The Medical Advances Without Animals Trust.

Compliance with ethical standards

Conflict of interest Md Humayun Kabir declares that he has no conflict of interest. Michael D. O'Connor declares that he has no conflict of interest.

Ethical approval This article does not contain any studies with human or animal subjects performed by any of the authors.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Andersson R et al (2014) An atlas of active enhancers across human cell types and tissues. *Nature* 507:455–461. <https://doi.org/10.1038/nature12787>
- Asp P et al (2011) Genome-wide remodeling of the epigenetic landscape during myogenic differentiation. *Proc Natl Acad Sci U S A* 108: E149–E158. <https://doi.org/10.1073/pnas.1102223108>
- Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA (2004) Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol* 14:283–291. <https://doi.org/10.1016/j.sbi.2004.05.004>
- Bailey T et al (2013) Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput Biol* 9:e1003326. <https://doi.org/10.1371/journal.pcbi.1003326>
- Banks CJ, Joshi A, Michoel T (2016) Functional transcription factor target discovery via compendia of binding and expression profiles. *Sci Rep* 6:20649. <https://doi.org/10.1038/srep20649>
- Barrett T et al (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 41:D991–D995. <https://doi.org/10.1093/nar/gks1193>
- Bebek G, Yang J (2007) PathFinder: mining signal transduction pathway segments from protein-protein interaction networks. *BMC Bioinformatics* 8:335. <https://doi.org/10.1186/1471-2105-8-335>
- Beer MA, Tavazoie S (2004) Predicting gene expression from sequence. *Cell* 117:185–198
- Berg J (2016) Gene-environment interplay. *Science* 354:15. <https://doi.org/10.1126/science.aal0219>
- Boeva V (2016) Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic. *Cells Front Genet* 7:24. <https://doi.org/10.3389/fgene.2016.00024>
- Boyer LA et al (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* 122:947–956. <https://doi.org/10.1016/j.cell.2005.08.020>
- Bumgarner R (2013) Overview of DNA microarrays: types, applications, and their future. *Curr Protoc Mol Biol Chapter 22:Unit 22.21*. <https://doi.org/10.1002/0471142727.mb2201s101>
- Butcher EC, Berg EL, Kunkel EJ (2004) Systems biology in drug discovery. *Nat Biotechnol* 22:1253–1259. <https://doi.org/10.1038/nbt1017>
- Chen K, Rajewsky N (2007) The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet* 8:93–103. <https://doi.org/10.1038/nrg1990>
- Chen H et al (2015) Reinforcement of STAT3 activity reprogrammes human embryonic stem cells to naive-like pluripotency. *Nat Commun* 6:7095. <https://doi.org/10.1038/ncomms8095>
- Cloonan N et al (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* 5:613–619. <https://doi.org/10.1038/nmeth.1223>
- Cohen SN, Chang AC, Boyer HW, Helling RB (1973) Construction of biologically functional bacterial plasmids in vitro. *Proc Natl Acad Sci U S A* 70:3240–3244
- Collas P (2010) The current state of chromatin immunoprecipitation. *Mol Biotechnol* 45:87–100. <https://doi.org/10.1007/s12033-009-9239-8>
- Consortium F et al (2014) A promoter-level mammalian expression atlas. *Nature* 507:462–470. <https://doi.org/10.1038/nature13182>
- Consortium GT (2013) The genotype-tissue expression (GTEx) project. *Nat Genet* 45:580–585. <https://doi.org/10.1038/ng.2653>
- Consortium TEP (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74. <https://doi.org/10.1038/nature11247>
- Consortium TME (2012) An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol* 13:418. <https://doi.org/10.1186/gb-2012-13-8-418>
- Consortium TU (2007) The universal protein resource (UniProt). *Nucleic Acids Res* 35:D193–D197. <https://doi.org/10.1093/nar/gkl929>
- Cressey D (2012) Stem cells take root in drug development. *Nat News*
- Davidson EH et al (2002) A genomic regulatory network for development. *Science* 295:1669–1678. <https://doi.org/10.1126/science.1069883>
- DeFreitas T, Saddiki H, Flaherty P (2016) GEMINI: a computationally-efficient search engine for large gene expression datasets. *BMC Bioinf* 17:102. <https://doi.org/10.1186/s12859-016-0934-8>
- Djordjevic D, Kusumi K, Ho JW (2016) XGSA: a statistical method for cross-species gene set analysis. *Bioinformatics* 32:i620–i628. <https://doi.org/10.1093/bioinformatics/btw428>
- Duggal G et al (2015) Alternative routes to induce naive pluripotency in human embryonic stem cells. *Stem Cells* 33:2686–2698. <https://doi.org/10.1002/stem.2071>
- Engreitz JM, Chen R, Morgan AA, Dudley JT, Mallewar R, Butte AJ (2011) ProfileChaser: searching microarray repositories based on genome-wide patterns of differential expression. *Bioinformatics* 27:3317–3318. <https://doi.org/10.1093/bioinformatics/btr548>
- Fujibuchi W, Kiseleva L, Taniguchi T, Harada H, Horton P (2007) CellMontage: similar expression profile search server. *Bioinformatics* 23:3103–3104. <https://doi.org/10.1093/bioinformatics/btm462>

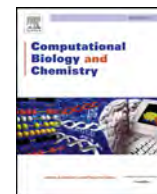
- Furey TS (2012) ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat Rev Genet* 13:840–852. <https://doi.org/10.1038/nrg3306>
- Germanguz I, Listgarten J, Cinkompumin J, Solomon A, Gaeta X, Lowry WE (2016) Identifying gene expression modules that define human cell fates. *Stem Cell Res* 16:712–724. <https://doi.org/10.1016/j.scr.2016.04.008>
- Gil DP, Law JN, Murali TM (2017) The PathLinker app: connect the dots in protein interaction networks. *F1000Res* 6:58. <https://doi.org/10.12688/f1000research.9909.1>
- Gitter A, Klein-Seetharaman J, Gupta A, Bar-Joseph Z (2011) Discovering pathways by orienting edges in protein interaction networks. *Nucleic Acids Res* 39:e22. <https://doi.org/10.1093/nar/gkq1207>
- Hackney JA, Moore KA (2005) A functional genomics approach to hematopoietic stem cell regulation. *Methods Mol Med* 105:439–452
- Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 33:D514–D517. <https://doi.org/10.1093/nar/gki033>
- Han X, Aslanian A, Yates JR 3rd (2008) Mass spectrometry for proteomics. *Curr Opin Chem Biol* 12:483–490. <https://doi.org/10.1016/j.cbpa.2008.07.024>
- Hannah R, Joshi A, Wilson NK, Kinston S, Gottgens B (2011) A compendium of genome-wide hematopoietic transcription factor maps supports the identification of gene regulatory control mechanisms. *Exp Hematol* 39:531–541. <https://doi.org/10.1016/j.exphem.2011.02.009>
- Heinz S et al (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 38:576–589. <https://doi.org/10.1016/j.molcel.2010.05.004>
- Hibbs MA, Hess DC, Myers CL, Huttenhower C, Li K, Troyanskaya OG (2007) Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics* 23:2692–2699. <https://doi.org/10.1093/bioinformatics/btm403>
- Hirst M et al (2007) LongSAGE profiling of nine human embryonic stem cell lines. *Genome Biol* 8:R113. <https://doi.org/10.1186/gb-2007-8-6-r113>
- Hoopes L (2008) Introduction to the gene expression and regulation topic room. *Nat Educ* 1(1)
- Huang DW, Sherman BT, Lempicki RA (2009a) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37:1–13. <https://doi.org/10.1093/nar/gkn923>
- Huang DW, Sherman BT, Lempicki RA (2009b) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4:44–57. <https://doi.org/10.1038/nprot.2008.211>
- Janky R et al (2014) iRegulon: from a gene list to a gene regulatory network using large motif and track collections. *PLoS Comput Biol* 10:e1003731. <https://doi.org/10.1371/journal.pcbi.1003731>
- Kabir MH, Djordjevic D, O'Connor MD, Ho JWK (2018a) C3: an R package for cross-species compendium-based cell-type identification. *Comput Biol Chem* 77:187–192
- Kabir MH, Murphy P, Lim S, Ho JWK, O'Connor MD (2018b) Large scale profiling of lens epithelial cell signalling pathways and target genes reveals regulatory networks for cataract-associated genes. *Exp Eye Res* (under review)
- Kabir MH, Patrick R, Ho JWK, O'Connor MD (2018c) Identification of active signaling pathways by integrating gene expression and protein interaction data. *BMC Syst Biol* in press
- Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28:27–30
- Kim HD, O'Shea EK (2008) A quantitative model of transcription factor-activated gene expression. *Nat Struct Mol Biol* 15:1192–1198. <https://doi.org/10.1038/nsmb.1500>
- Kuleshov MV et al (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 44:W90–W97. <https://doi.org/10.1093/nar/gkw377>
- Lee TI et al (2006) Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* 125:301–313. <https://doi.org/10.1016/j.cell.2006.02.043>
- Liu Y, Zhao H (2004) A computational approach for ordering signal transduction pathway components from genomics and proteomics. *Data BMC Bioinf* 5:158. <https://doi.org/10.1186/1471-2105-5-158>
- Marbach D, Lamparter D, Quon G, Kellis M, Kutalik Z, Bergmann S (2016) Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat Methods* 13:366–370. <https://doi.org/10.1038/nmeth.3799>
- Mardis ER (2007) ChIP-seq: welcome to the new frontier. *Nat Methods* 4:613–614. <https://doi.org/10.1038/nmeth0807-613>
- Medina I et al (2010) Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic Acids Res* 38:W210–W213. <https://doi.org/10.1093/nar/gkq388>
- Mei S, Zhu H (2015) Multi-label multi-instance transfer learning for simultaneous reconstruction and cross-talk modeling of multiple human signaling pathways. *BMC Bioinf* 16:417. <https://doi.org/10.1186/s12859-015-0841-4>
- Murphy P et al (2018) Light-focusing human micro-lenses generated from pluripotent stem cells model lens development and drug-induced cataract in vitro. *Development* 145. <https://doi.org/10.1242/dev.155838>
- O'Connor MD (2013) The 3R principle: advancing clinical application of human pluripotent stem cells. *Stem Cell Res Ther* 4:21. <https://doi.org/10.1186/scrt169>
- O'Connor MD, Kadel MD, Eaves CJ (2011a) Functional assays for human embryonic stem cell pluripotency. *Methods Mol Biol* 690:67–80. https://doi.org/10.1007/978-1-60761-962-8_4
- O'Connor MD et al (2011b) Retinoblastoma-binding proteins 4 and 9 are important for human pluripotent stem cell maintenance. *Exp Hematol* 39:866–879 e861. <https://doi.org/10.1016/j.exphem.2011.05.008>
- Pinto JP, Reddy Kalathur RK, Machado RS, Xavier JM, Braganca J, Futschik ME (2014) StemCellNet: an interactive platform for network-oriented investigations in stem cell biology. *Nucleic Acids Res* 42:W154–W160. <https://doi.org/10.1093/nar/gku455>
- Rackham OJ et al (2016) A predictive computational framework for direct reprogramming between human cell types. *Nat Genet* 48:331–335. <https://doi.org/10.1038/ng.3487>
- Ralston A, Shaw K (2008) Gene expression regulates cell differentiation. *Nat Educ* 1(1)
- Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP (2006) GenePattern 2.0. *Nat Genet* 38:500–501. <https://doi.org/10.1038/ng0506-500>
- Respuela P, Nikolic M, Tan M, Frommolt P, Zhao Y, Wysocka J, Rada-Iglesias A (2016) Foxd3 promotes exit from naive pluripotency through enhancer decommisioning and inhibits germline specification cell. *Stem Cell* 18:118–133. <https://doi.org/10.1016/j.stem.2015.09.010>
- Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43:e47. <https://doi.org/10.1093/nar/gkv007>
- Ritz A et al (2016) Pathways on demand: automated reconstruction of human signaling networks. *NPJ Syst Biol Appl* 2:16002. <https://doi.org/10.1038/npjsba.2016.2>
- Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene

- expression data. *Bioinformatics* 26:139–140. <https://doi.org/10.1093/bioinformatics/btp616>
- Roider HG, Manke T, O'Keeffe S, Vingron M, Haas SA (2009) PASTAA: identifying transcription factors associated with sets of co-regulated genes. *Bioinformatics* 25:435–442. <https://doi.org/10.1093/bioinformatics/btn627>
- Ruau D et al (2013) Building an ENCODE-style data compendium on a shoestring. *Nat Methods* 10:926. <https://doi.org/10.1038/nmeth.2643>
- Scott J, Ideker T, Karp RM, Sharan R (2006) Efficient algorithms for detecting signaling pathways in protein interaction networks. *J Comput Biol* 13:133–144
- Shanks N, Greek R, Greek J (2009) Are animal models predictive for humans? *Philos Ethics Humanit Med* 4:2. <https://doi.org/10.1186/1747-5341-4-2>
- Sharov AA et al (2008) Identification of Pou5f1, Sox2, and Nanog downstream target genes with statistical confidence by applying a novel algorithm to time course microarray and genome-wide chromatin immunoprecipitation data. *BMC Genomics* 9:269. <https://doi.org/10.1186/1471-2164-9-269>
- Shiels A, Bennett TM, Hejtmancik JF (2010) Cat-Map: putting cataract on the map. *Mol Vis* 16:2007–2015
- Spitz F, Furlong EE (2012) Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* 13:613–626. <https://doi.org/10.1038/nrg3207>
- Steffen M, Petti A, Aach J, D'Haeseleer P, Church G (2002) Automated modelling of signal transduction networks. *BMC Bioinf* 3:34
- Tuncbag N et al (2013) Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem. *J Comput Biol* 20:124–136. <https://doi.org/10.1089/cmb.2012.0092>
- Ungrin M, O'Connor M, Eaves C, Zandstra PW (2007) Phenotypic analysis of human embryonic stem cells. *Curr Protoc Stem Cell Biol* Chapter 1:Unit 1B 3. <https://doi.org/10.1002/9780470151808.sc01b03s2>
- Van der Jeught M et al (2015) Application of small molecules favoring naive pluripotency during human embryonic stem cell derivation. *Cell Reprogram* 17:170–180. <https://doi.org/10.1089/cell.2014.0085>
- von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* 31:258–261
- Wang K et al (2011) CASCADE_SCAN: mining signal transduction network from high-throughput data based on steepest descent method. *BMC Bioinf* 12:164. <https://doi.org/10.1186/1471-2105-12-164>
- Warrier S et al (2017) Direct comparison of distinct naive pluripotent states in human embryonic stem cells. *Nat Commun* 8:15055. <https://doi.org/10.1038/ncomms15055>
- Zacher B, Michel M, Schwalb B, Cramer P, Tresch A, Gagneur J (2017) Accurate promoter and enhancer identification in 127 ENCODE and roadmap epigenomics cell types and tissues by GenoSTAN. *PLoS One* 12:e0169249. <https://doi.org/10.1371/journal.pone.0169249>
- Zhang L, Mallick BK (2013) Inferring gene networks from discrete expression data. *Biostatistics* 14:708–722. <https://doi.org/10.1093/biostatistics/kxt021>
- Zhang S, Cao J, Kong YM, Scheuermann RH (2010) GO-Bayes: Gene Ontology-based overrepresentation analysis using a Bayesian approach. *Bioinformatics* 26:905–911. <https://doi.org/10.1093/bioinformatics/btq059>
- Zhao XM, Li S (2017) HISP: a hybrid intelligent approach for identifying directed signaling pathways. *J Mol Cell Biol* 9:453–462. <https://doi.org/10.1093/jmcb/mjx054>
- Zhao XM, Wang RS, Chen L, Aihara K (2008) Uncovering signal transduction networks from high-throughput data by integer linear programming. *Nucleic Acids Res* 36:e48. <https://doi.org/10.1093/nar/gkn145>
- Zinman GE, Naiman S, Kanfi Y, Cohen H, Bar-Joseph Z (2013) ExpressionBlast: mining large, unstructured expression databases. *Nat Methods* 10:925–926. <https://doi.org/10.1038/nmeth.2630>

Chapter 2

C3: An R package for cross-species compendium-based cell-type identification

Identification of cell-type of a biological sample based on its gene expression profile is an important research question when investigating novel cell populations resulting from differentiation of pluripotent stem cells, or after isolation of a cell population in a non-model organism. Application of a cell type compendium-based method may be particularly useful if the compendium consists of large collection of available cell transcriptome data for human and mouse organisms. Here we have developed a procedure associated with an open source R package, known as C3, to identify the cell type of a gene expression profile and tested it against a variety of cell samples.



C3: An R package for cross-species compendium-based cell-type identification



Md Humayun Kabir^{a,b,c}, Djordje Djordjevic^{b,d}, Michael D. O'Connor^{a,e}, Joshua W.K. Ho^{b,d,f,*}

^a School of Medicine, Western Sydney University, Campbelltown, NSW, Australia

^b Victor Chang Cardiac Research Institute, Darlinghurst, NSW, Australia

^c Department of Computer Science and Engineering, University of Rajshahi, Bangladesh

^d St. Vincent's Clinical School, University of New South Wales, Sydney, NSW, Australia

^e Medical Sciences Research Group, Western Sydney University, Campbelltown, NSW, Australia

^f School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Pokfulam, Hong Kong SAR, China

ARTICLE INFO

Keywords:

Bioinformatics
Transcriptomics
Cell type identification
Cross-species
Gene set analysis

ABSTRACT

Cell type identification from an unknown sample can often be done by comparing its gene expression profile against a gene expression database containing profiles of a large number of cell-types. This type of compendium-based cell-type identification strategy is particularly successful for human and mouse samples because a large volume of data exists for these organisms. However, such rich data repositories often do not exist for most non-model organisms. This makes transcriptome-based sample classification in these species challenging. We propose to overcome this challenge by performing a *cross-species* compendium comparison. The key is to utilise a recently published cross-species gene set analysis (XGSA) framework to correct for biases that may arise due to potentially complex homologous gene mapping between two species. The framework is implemented as an open source R package called C3. We have evaluated the performance of C3 using a variety of public data in NCBI Gene Expression Omnibus. We also compared the functionality and performance of C3 against some similar gene expression profile matching tools. Our evaluation shows that C3 is a simple and effective method for cell type identification. C3 is available at <https://github.com/VCCRI/C3>.

1. Introduction

The key question we seek to address in this article is *how can we identify the cell-type of a biological sample given its gene expression profile?* This question commonly arises when investigating a novel cell population resulting from differentiation of pluripotent stem cells or isolation of a cell population in a non-model organism. The most popular bioinformatics approach is a compendium-based identification approach, in which the unknown sample's gene expression profile is used as a query profile against a large gene expression compendium consisting of many cell types. A number of tools have been developed to perform such a task, such as GEMINI (DeFreitas et al., 2016), ProfileChaser (Engreitz et al., 2011), ExpressionBlast (Zinman et al., 2013) and CellMortage (Fujibuchi et al., 2007). All these tools work in a similar fashion: match the query gene expression profile or a gene set against a database of gene expression profiles to identify its best matches. Importantly, most of these tools implicitly assume there is a one-to-one correspondence between genes in the query sample and the

compendium sample, which can be violated when comparing data from different species. Beyond supporting filtering for genes with one-to-one homology mapping across species, none of the current tools effectively handle a cross-species query in a statistically rigorous fashion.

Therefore, when using currently available tools it is important to always use a database of the same species as the query sample. This is often practically impossible because most publicly available data sets are only available for a small number of species. For example, one of the largest public gene expression repositories - the NCBI Gene Expression Omnibus (GEO) - contained more than 57,000 GEO series (GSE) generated by microarrays or RNA-Seq (as of March 2017) (Barrett et al., 2013). Collectively, these data are a valuable resource for researchers to discover new biological insights. Nonetheless, most of these GSE data sets were generated from just two species: *Homo sapiens* (human) and *Mus musculus* (mouse). In fact, around two thirds of these GSE data sets are derived from human or mouse samples (Fig. 1). The other third come from more than 1300 species, with only 33 species having over 100 GSE (Fig. 1). In other words, while it is possible to curate a useful

* Corresponding author at: School of Biomedical Sciences, The University of Hong Kong, L4 Laboratory Block, 21 Sassoon Road, Pokfulam, Hong Kong SAR, China.
E-mail address: jwkho@hku.hk (J.W.K. Ho).

<https://doi.org/10.1016/j.compbiolchem.2018.10.003>

Received 8 March 2018; Received in revised form 29 August 2018; Accepted 4 October 2018

Available online 09 October 2018

1476-9271/ © 2018 Elsevier Ltd. All rights reserved.

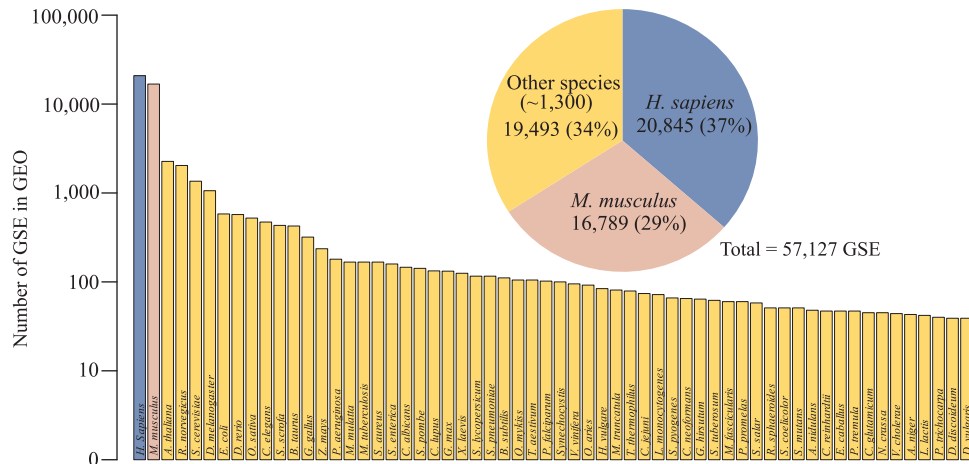


Fig. 1. Summary of GSE based on species in NCBI GEO.

The pie chart shows the total number of GSE for *H. sapiens* (blue), *M. musculus* (pink) and all other species (orange). The bar plot shows the top 60 species according to the number of GSE in NCBI GEO.

gene expression compendium for human and mouse, it is practically impossible for other species, especially non-model organisms.

We propose to alleviate this lack of species-specific compendia by performing a *cross-species* cell identification, where a query profile is matched against a database of samples which come from different organisms. A key challenge to implementing such a cross-species analysis scheme is that many pairs of species, especially those that are evolutionary distant, can have complex “many-to-many” homologous gene relationships. Failure to properly account for the homology gene mapping can lead to statistical biases (Djordjevic et al., 2016).

In this article, we present a new open source R package – C3 – that implements this cross-species compendium-based cell type identification approach using a recently developed cross-species gene set analysis method called XGSA (Djordjevic et al., 2016). XGSA has been shown to reduce the false positive bias while still maintain good statistical power for gene sets affected by highly complex homology structures. Using C3, we can harness the large collection of human and mouse public data as a resource to identify unknown cell types for a wide variety of species. We demonstrate the effectiveness of C3 using a large collection of GEO data. We also compare its performance with other similar tools.

2. Methods

2.1. C3: a new R package for cross-species cell-type identification

C3 is an open source R package for identifying an unknown cell-type from its gene expression profile based on a large compendium of gene expression data that can be derived from different species. A key aspect of this approach is that it is most useful when the compendium represents many different tissue or cell types, preferably from a well-studied organism such as human or mouse. Examples of public data sources that can be used to form this kind of compendium include ENCODE (The ENCODE Project Consortium, 2012; The Mouse ENCODE Consortium, 2012) and GTEx (The GTEx Consortium, 2013). The full description of the method implemented in C3 is described in detail in the rest of this section, but an overview of the framework can be found in Fig. 2. Briefly, C3 first identifies genes considered to be specifically highly-expressed genes in the query and the compendium profiles, by removing genes ubiquitously highly-expressed across these expression profiles. Next, C3 performs XGSA between the query gene set and each of the compendium gene sets to account for “many-to-many” gene relationships, and thereby determine which compendium gene sets are statistically enriched in the query gene set. A *p*-value is reported for each compendium sample. The cell-types of the most highly ranked

compendium gene sets (according to *p*-value) are then used to predict the cell-type of the query profile. C3 is available at <https://github.com/VCCRI/C3>.

2.2. The human and mouse gene expression compendia

For both mouse and human, we constructed a large compendium of tissue-specific genes using RNA data from the ENCODE project. ENCODE gene expression data, summarised as FPKMs, were obtained for human (hg19; 144 tissues or cell lines) (The ENCODE Project Consortium, 2012) and for mouse (mm9; 94 tissues or cell types) (The Mouse ENCODE Consortium, 2012). Most tissues or cell types in the ENCODE data set are represented by more than one replicate. We combined replicates of the same tissue or cell type by calculating the mean expression value for each gene. The GTEx gene expression data, summarised as median of TPM values, of 53 different tissues were downloaded from GTEx web portal (<https://www.gtexportal.org/home/datasets>). When a compendium was constructed from multiple data sources, we only considered genes that were common among all data sets.

2.3. Identification of specifically expressed genes in the query and compendium data

Using the compendium data, for each sample in the compendium we identified sets of highly-expressed genes that are specific to each sample using two parameters: *n* – the number of highly expressed genes to consider for marker gene status; *t* – the proportion of samples a marker gene can appear in before it is discarded as non-unique/non-specific. Using these two parameters we could identify then remove genes that are consistently highly expressed (within the top *n* highly expressed genes in each sample) in more than *t* × 100% of samples. The goal of this step is to remove ubiquitously expressed genes such as house-keeping genes. The remaining gene sets is enriched for cell-type specific genes. To identify the highly-expressed specific genes within the query data set, first we calculated the mean expression value of the replicates for each gene and then identified the top *n* highly expressed genes. We then removed the ubiquitously expressed genes identified by the compendium from the top *n* expressed genes. When the query sample species is different from the species used to create the compendium, we use XGSA to identify the homologs of the set of ubiquitously expressed genes for the query cell species. We then remove this set of gene homologs from the query cell top expressed genes.

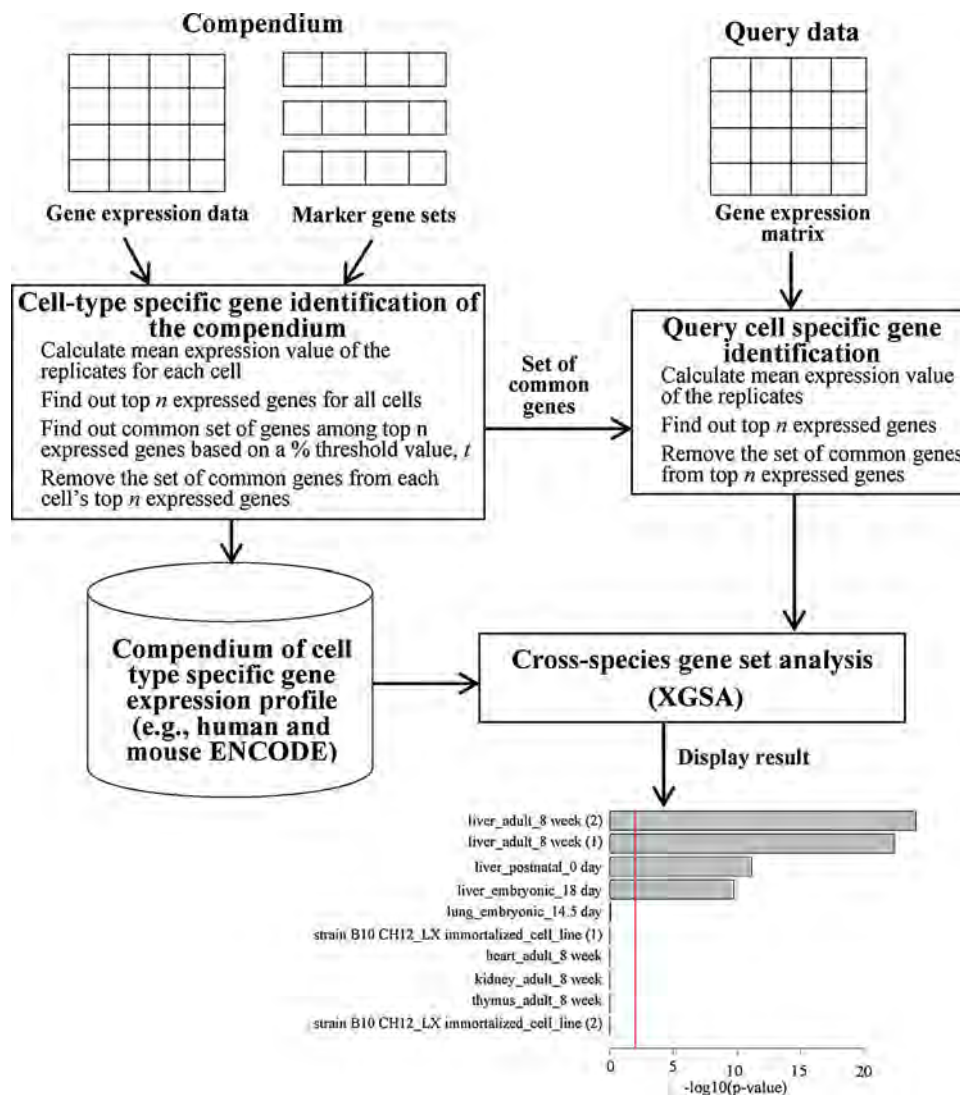


Fig. 2. Overall workflow diagram of C3.

2.4. XGSA

To provide the required input for XGSA, all genes names are first converted to ENSEMBL gene IDs. XGSA then applies a simple statistical method that computes a conservative p -value based on Fisher's Exact test. This approach takes into account the homology gene mapping structure between two cross-species gene sets (Djordjevic et al., 2016). If the two compared gene sets are from the same gene sets, the resulting p -value is identical to that of a standard gene set test based on a Fisher's Exact test. The package then performs Benjamini-Hochberg multiple testing corrections on the raw p -values, and reports and visualises the $-\log_{10}$ of the corrected p -values. Accordingly, the lowest p -value indicates the best similarity between the query sample and the compendium sample. However, if more similar cell samples of the query exist in the compendium we can go with a number of top most matches.

2.5. Comparison with ExpressionBlast

For the comparison with ExpressionBlast, we used brain, kidney and liver sample data sets from the *R. norvegicus* species (Fushan et al., 2015). We identified the specific highly expressed genes for each of the sample tissue types using our C3 package by setting parameter values as $n = 1000$ and $t = 0.10$. Among these specific highly expressed genes, we have selected the top 100 expressed genes based on their expression

values. We used this set of highly-expressed tissue specific genes with \log_2 expression values as the input to the ExpressionBlast web tool. In this way we tested each of the three tissue types against both the human and mouse organisms using ExpressionBlast.

3. Results

3.1. Evaluation of C3

To evaluate the performance of C3, we collected gene expression profiles from four GEO data series (GSE43013 (Fushan et al., 2015), GSE74754 (Mayrhofer et al., 2017), GSE78770 (Morey et al., 2016), and GSE53393 (Chapalamadugu et al., 2014)), which collectively contain data from 13 different species (*B. taurus*, *C. familiaris*, *C. porcellus*, *E. caballus*, *E. europaeus*, *F. catus*, *M. musculus*, *O. cuniculus*, *R. norvegicus*, *S. scrofa*, *D. rerio*, *T. truncatus*, and *M. mulatta*) across five different tissue types (brain, kidney, liver, blood, and skeletal muscle). We tested whether C3 could correctly identify the cell type of the samples when compared against a human compendium or a mouse compendium constructed from ENCODE data (The ENCODE Project Consortium, 2012; The Mouse ENCODE Consortium, 2012). Initially, we tested two combinations of parameters in C3 (n and t). The summary result is shown in Fig. 3 and the detailed results are shown in the Supplementary materials [see Supplementary Tables 1–6]. Overall,

a

	Sample name	n=500, t=0.05		n=1000, t=0.10	
		Human	Mouse	Human	Mouse
Data set 1 (GSE43013)	<i>B. taurus</i> brain	1	1	1	1
	<i>C. familiaris</i> brain	1	1	1	1
	<i>C. porcellus</i> brain	1	1	1	1
	<i>E. caballus</i> brain	1	1	1	1
	<i>E. europaeus</i> brain	1	1	1	1
	<i>F. catus</i> brain	1	1	1	1
	<i>M. musculus</i> brain	1	1	1	1
	<i>O. cuniculus</i> brain	1	1	1	1
	<i>R. norvegicus</i> brain	1	1	1	1
	<i>S. scrofa</i> brain	1	1	1	1
	<i>B. taurus</i> kidney	2	1	2	1
	<i>C. familiaris</i> kidney	3	1	2	1
	<i>C. porcellus</i> kidney	3	1	2	1
	<i>E. caballus</i> kidney	2	1	2	1
	<i>E. europaeus</i> kidney	3	1	2	1
	<i>F. catus</i> kidney	6	1	2	1
	<i>M. musculus</i> kidney	2	1	2	1
	<i>O. cuniculus</i> kidney	2	1	2	1
	<i>R. norvegicus</i> kidney	2	1	2	1
	<i>S. scrofa</i> kidney	5	1	3	1
	<i>B. taurus</i> liver	1	1	1	1
	<i>C. familiaris</i> liver	1	1	1	1
	<i>C. porcellus</i> liver	1	1	1	1
	<i>E. caballus</i> liver	1	1	1	1
	<i>E. europaeus</i> liver	1	1	1	1
	<i>F. catus</i> liver	1	1	1	1
	<i>M. musculus</i> liver	1	1	1	1
	<i>O. cuniculus</i> liver	1	1	1	1
	<i>R. norvegicus</i> liver	1	1	1	1
	<i>S. scrofa</i> liver	1	1	1	1
Data set 2 (GSE74754)	<i>D. rerio</i> brain (control)	1	1	1	1
	<i>D. rerio</i> brain (tumour)	1	1	1	1
Data set 3 (GSE78770)	<i>T. truncatus</i> blood (hua)	1	1	1	1
	<i>T. truncatus</i> blood (kai)	1	1	1	1
	<i>T. truncatus</i> blood (keo)	1	1	1	1
	<i>T. truncatus</i> blood (pele)	1	1	1	1
Data set 4 (GSE53393)	<i>M. mulatta</i> skeletal muscle (early BPA)	1	1	2	1
	<i>M. mulatta</i> skeletal muscle (early control)	1	1	2	1
	<i>M. mulatta</i> skeletal muscle (late BPA)	2	1	2	1
	<i>M. mulatta</i> skeletal muscle (late control)	2	1	2	1

b

	Sample name	Parameters settings	
		n=500, t=0.05	n=1000, t=0.10
Data set 1 (GSE43013)	<i>R. norvegicus</i> kidney	1	1
	<i>B. taurus</i> liver	1	1
Data set 2 (GSE74754)	<i>D. rerio</i> brain (control)	1	1
Data set 3 (GSE78770)	<i>T. truncatus</i> blood (hua)	1	1
Data set 4 (GSE53393)	<i>M. mulatta</i> skeletal muscle (early BPA)	1	1

barring a few exceptions which will be discussed below, C3 was able to consistently identify the correct or the most closely related cell type across all species (Fig. 3a).

To investigate the robustness of C3, we investigated the impact of using a different high quality gene expression compendium. Using a compendium of human tissue from GTEx (The GTEx Consortium, 2013), we repeated the C3 analysis on a subset of representative samples (from different tissues and organisms) from the four data sets. The summary results are shown in Fig. 3b and the detailed results can be found in Supplementary Table 6. The result confirm that C3 is robust in producing accurate tissue/cell-type prediction when using a different high quality compendium, Viewed together, the ENCODE and GTEx analyses demonstrate that either combination of *n* and *t* enable accurate cell/tissue identification when comparing the mouse and human

Fig. 3. Evaluation of C3. Gene expression profiles of tissues from 13 different organisms were selected from four GEO data sets. These profiles were used to evaluate whether C3 could correctly identify its cell type of the sample when compared against a human ENCODE compendium (Human) or a mouse ENCODE compendium (Mouse) (a). The five different samples from different GSE IDs also been tested with the GTEx human compendium (b). *n*: top number of highly expressed genes; *t*: cut-off threshold value; 1 = Statistically significant and in top position; 2 = Statistically significant but in top 2-3rd position; 3 = Statistically significant but in top 4-5 th position; 4 = Not statistically significant but in top position; 5 = Not statistically significant but in top 2-5 th position; 6 = Not statistically significant and not in 2-5 th position

compendium data within each parameter set, with *n* = 1000 and *t* = 0.1 being preferable.

Notably, further assessing the *n* and *t* parameter settings (by changing them to: *n* = 300, *t* = 0.05; *n* = 500, *t* = 0.04; *n* = 700, *t* = 0.03; etc.) gave consistent cell type identifications using the ENCODE compendium (data not shown), with, the best results obtained using *n* = 1000 and *t* = 0.10. Similarly, when we tested the five different query samples with the GTEx compendium (Fig. 3b), the parameter values of *n* = 1000 and *t* = 0.10 gave accurate results for all cases. Thus values of 1000 and 0.10 for the parameters *n* and *t* (respectively) appear optimal for accurate identification of query samples regardless of the source of the compendium data.

Examining the cell type identifications for each of the datasets in more detail, GSE43013 (Fushan et al., 2015) contains a gene expression

data set from three different tissue types (brain, kidney and liver) in 33 mammalian species, among which 10 have homology mapping information available via ENSEMBL. C3 was able to correctly identify the cell types in all the brain and liver samples across all 10 species. For the kidney data, C3 correctly identified the cell type when compared against the mouse compendium across 10 species, but was much less effective when compared against the human compendium. Interestingly, this comparison against the human compendium resulted in most of the kidney gene sets being identified as liver samples ahead of the human kidney samples. Nonetheless, comparison of the mouse and human predictions for GSE43013 allowed correct assignment of kidney as the most likely cell type for the kidney samples.

We also tested three more GSE datasets that contained data from 3 additional species; *D. rerio* (GSE74754; brain) (Mayrhofer et al., 2017), *T. truncatus* (GSE78770; blood) (Morey et al., 2016), and *M. mulatta* (GSE53393; skeletal muscle) (Chapalamadugu et al., 2014). Through these analysis C3 correctly identified the cell types of *D. rerio* brain and *T. truncatus* blood. The *M. mulatta* skeletal muscle samples were correctly identified by C3 when they compared to the mouse compendium but were not as effectively identified using the human compendium (top hit was heart/tongue sample) (Fig. 3a). Nevertheless, comparison of the mouse and human predictions for GSE53393 allowed correct assignment of skeletal muscle as the most likely cell type for these samples.

Overall, a total of 160 C3 analyses were performed (80 against the ENCODE mouse compendium and 80 against the ENCODE human compendium) using two combinations of n/t parameters (i.e., 500/0.05 and 1000/0.1). Notably, all the cell type identity predictions made by C3 using the mouse compendium were correct for at least one of the parameter combinations (i.e., typically at least 1000/0.1 if not also 500/0.05). For comparison against the human compendium: correct predictions were made for 67.5% of the queries, and for a further 25% of the queries the correct prediction was ranked second or third by C3 (i.e., the correct prediction was in the top 3 positions 92.5% of the time using the human compendium). Only 1 out of the 80 predictions made by C3 using the human compendium (0.625%; *F. cattus*, kidney) did not include the correct identification in the top 5 predictions. Notably, only two cell types were not predicted correctly by the human compendium (i.e., as the top prediction): kidney and skeletal muscle. These tissues are both highly vascularised, and this may be a confounding factor when comparing against human samples. However, as shown in Fig. 3, all the kidney and skeletal muscle datasets were correctly identified when compared against the mouse compendium. Thus, as mentioned above, correct cell type identification is achieved by comparing the predictions from both the human and mouse compendia.

To investigate the impact of data normalisation, we repeated the C3 analysis after performing quantile normalisation. In particular, we performed quantile normalisation of the ENCODE compendium data sets and then tested the selected query samples for both parameter settings. The summary and detailed results are given in Supplementary Tables 3–5. From the results we can see that the results are same as those above. This result supports that the robustness of C3 against data normalisation.

3.2. Comparison with other similar software programs

A comparison of the features of C3 and other similar methods is illustrated in Table 1. The other four similar methods are primarily web-based with only GEMINI offering a Python command-line version. GEMINI lacks the ability to perform cross-species cell type identification. It uses level 3 gene expression datasets from The Cancer Genome Atlas (TCGA) project (The Cancer Genome Atlas Network, 2012). Also, CellMontage can compare only the expression data from similar microarray platforms. As a result, neither GEMINI nor CellMontage could be included in our comparative analysis. ProfileChaser supports cross-species analyses using NCBI HomoloGene for only 6 species, and uses

only the set of genes that have one-to-one human homology mapping. However, ProfileChaser searches only the curated GEO DataSets (GDS) (i.e., supporting only 1815 GDS) for similar biological conditions based on differential gene expression from reduced set of gene expression features. Consequently, we were unable to meaningfully include this tool in our comparative analysis.

The only C3 alternative we are aware of that can compare a transcriptomic profile to a compendium of data across species in order to identify an unknown cell type is ExpressionBlast (Zinman et al., 2013). ExpressionBlast is a web-based tool that takes a maximum of one hundred differentially expressed genes with their expression values, and compares it to microarray data from 8 different species on GEO. For cross-species comparisons, ExpressionBlast uses homologous gene groups from InParanoid and handles multiple homologs using the closest expression value of the input gene. In contrast, C3 is an open source R package that takes gene expression profiles as input. C3 leverages XGSA to perform cross-species analysis between any of species in the growing list of species in Ensembl Compara (currently 85 species).

To compare the performance of ExpressionBlast with C3, we analysed the brain, kidney and liver sample data from *R. norvegicus* (GSE43013) (Fushan et al., 2015) using both methods, as the rat is one of the eight species supported by ExpressionBlast. For C3, we tested against the human and mouse compendiums with parameter values $n = 1000$ and $t = 0.10$. For ExpressionBlast, we input the 100 highly expressed tissue specific genes with their $\log_2(FPKM + 1)$ expression values. The summary results for C3 and ExpressionBlast are shown in Table 2, and the detailed results are presented in Supplementary Table 2 (for C3) and Supplementary Fig. 1 (for ExpressionBlast). From the comparative test results, it is clear that C3 can identify cell types at least as accurately as ExpressionBlast. However, C3 has markedly greater flexibility than ExpressionBlast in that: i) it can handle the whole query gene expression profile; ii) it can be applied to data from a wide range of organisms; and iii) its R package enables it to be easily incorporated into any analytical pipeline.

4. Discussion

This work highlights the utility of cross-species analysis in cell-type identification using a gene expression compendium-based approach. This is particularly important when considering that the majority (two thirds) of transcriptomic data in the GEO database is from human and mouse, with the remaining third of data shared between over 1000 organisms (Fig. 1), most of which have very scant genomic resources. Our aim with C3 was to leverage the many published data sets from the well characterised human and mouse organisms to identify an unknown cell type from a potentially poorly characterised organism, or a new/previously undescribed cell type (e.g., cells obtained from differentiated stem cell cultures). We demonstrated that the accuracy of C3 cell type predictions is independent of the data source used (i.e., C3 provides accurate predictions using both ENCODE and GTEx data). Moreover, C3 outperforms the few other methods available for cross-species cell type identification. Accurate C3 predictions can be obtained using a range of values for the parameters n and t , however, values of 1000 and 0.10 (respectively) are optimal.

Recently we have used a preliminary version of C3 to discover the identity of a novel PAX7+ cell population in lizard *Anole carolinensis* (Palade et al., 2018). Without C3, definitively demonstrating the identity of a novel cell-type in a relatively under-studied organism such as *A. carolinensis* is very challenging as there are very few *Anole*-specific transcriptomic data available for this organism. The PAX7+ cells analysed by C3 are implicated in the regeneration that occurs after amputation of the tail in *A. carolinensis*. Using a preliminary version of C3, we were able to identify from both the human and mouse ENCODE compendia that the PAX7+ cell-type resembles muscle satellite cells. This information has allowed us to further investigate the cellular and molecular basis of lizard muscle regeneration. As another real-life

Table 1
Comparison of software features of C3 and other similar methods.

	C3	ExpressionBlast	ProfileChaser	GEMINI	CellMontage
Cross-species method	Ensembl BioMart portal, complete homology structure using XGSA	Inparanoid, handles multiple orthologues using closest value of input gene	One-to-one human homology	Not supported	Not mentioned
How many species	As many as ENSEMBL mapping	8	6	–	–
Input	Gene expression matrix	Max 100 differentially expressed genes with expression values	Gene expression matrix	Gene expression matrix	Gene expression matrix with raw expression values
User interface	R command line	Web	Web	Web and Python command-line	Web
Availability	Open source	Free	Free	Free	Free
Application	General	General	Specific to GDS	Level 3 gene expression from TCGA project	Specific to similar microarray platforms
Dependency	Previously made compendium	Differentially expressed genes	Reduced set of gene expression features	Reduced dimension of expression profile	UniGene names for gene ids

Table 2
Comparison of cross-species cell type identification using C3 and ExpressionBlast.

	Identified cell type by C3	Identified cell type by ExpressionBlast
<i>R. norvegicus</i> brain with Human compendium	brain	other than brain (no brain sample among top 5)
<i>R. norvegicus</i> brain with Mouse compendium	brain	brain
<i>R. norvegicus</i> kidney with Human compendium	liver at top position and then kidney	liver (no kidney sample among top 5)
<i>R. norvegicus</i> kidney with Mouse compendium	kidney	kidney
<i>R. norvegicus</i> liver with Human compendium	liver	liver
<i>R. norvegicus</i> liver with Mouse compendium	liver	liver

application, we used the C3 method to demonstrate that an ROR1 + cell population derived from human pluripotent stem cells is similar to lens epithelial cells in both human and mouse (Murphy et al., 2018). Both of these examples highlight the power of C3 in determining or confirming the identity of a cell type using a compendium of gene expression profiles from different species, including poorly characterised species.

C3 can only correctly identify the cell type of an unknown transcriptomic profile if a similar cell type is represented in the compendium. With this in mind, the quality, variety and size of the compendium is paramount and future work should investigate larger compendiums such as based on ARCHS4 (Alexander Lachmann et al., 2017), as well as domain specific compendiums such as for identifying cancer subtypes.

5. Conclusion

Overall, we have demonstrated that C3 can prioritise identification of the correct corresponding cell type as the most significant hit. We believe C3 should facilitate rapid cell type identification for less-well characterised species, or for poorly characterised cell types obtained from stem cell differentiation strategies.

Competing interests

The authors declare no competing financial interests.

Authors' contributions

J.W.K.H. initiated the project; M.H.K. designed the method, implemented the package, performed evaluation and wrote the manuscript; D.D. contributed to method design and software testing; M.D.O'C and J.W.K.H. supervised the whole project and revised the manuscript.

All authors read and approved the final manuscript.

Acknowledgements

M.H.K. is supported by a UWS Postgraduate Research Award (International). J.W.K.H. is supported by a Career Development Fellowship by the National Health and Medical Research Council (1105271) and a Future Leader Fellowship by the National Heart Foundation of Australia (100848).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <https://doi.org/10.1016/j.compbiolchem.2018.10.003>.

References

- Alexander Lachmann, D.T., Keenan, Alexandra B., Jagodnik, Kathleen M., Lee, Hyojin J., Wang, Lily, Silverstein, Moshe C., Ma'ayan, Avi, 2017. 'Massive mining of publicly available RNA-seq data from human and mouse'. *Nat. Commun.* 18 (9), 1366.
- Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., et al., 2013. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* 41, D991–5.
- Chapalamadugu, K.C., Vandevoort, C.A., Settles, M.L., Robison, B.D., Murdoch, G.K., 2014. Maternal bisphenol A exposure impacts the fetal heart transcriptome. *PLoS One* 9, e89096.
- DeFreitas, T., Saddiki, H., Flaherty, P., 2016. GEMINI: a computationally-efficient search engine for large gene expression datasets. *BMC Bioinf.* 17, 102.
- Djordjevic, D., Kusumi, K., Ho, J.W., 2016. XGSA: a statistical method for cross-species gene set analysis. *Bioinformatics* 32, i620–i628.
- Engreitz, J.M., Chen, R., Morgan, A.A., Dudley, J.T., Mallewar, R., Butte, A.J., 2011. ProfileChaser: searching microarray repositories based on genome-wide patterns of differential expression. *Bioinformatics* 27, 3317–3318.
- Fujibuchi, W., Kiseleva, L., Taniguchi, T., Harada, H., Horton, P., 2007. CellMontage: similar expression profile search server. *Bioinformatics* 23, 3103–3104.
- Fushan, A.A., Turanov, A.A., Lee, S.G., Kim, E.B., Lobanov, A.V., Yim, S.H., et al., 2015. Gene expression defines natural changes in mammalian lifespan. *Aging Cell* 14, 352–365.
- Mayrhofer, M., Gourain, V., Reischl, M., Affaticati, P., Jenett, A., Joly, J.S., et al., 2017. A novel brain tumour model in zebrafish reveals the role of YAP activation in MAPK- and PI3K-induced malignant growth. *Dis. Model. Mech.* 10, 15–28.
- Morey, J.S., Neely, M.G., Lunardi, D., Anderson, P.E., Schwacke, L.H., Campbell, M., et al., 2016. RNA-seq analysis of seasonal and individual variation in blood transcriptomes of healthy managed bottlenose dolphins. *BMC Genomics* 17, 720.
- Murphy, P., Kabir, M.H., Srivastava, T., Mason, M.E., Dewi, C.U., Lim, S., et al., 2018. Light-focusing human micro-lenses generated from pluripotent stem cells model lens development and drug-induced cataract in vitro. *Development* 145.
- Palade, J., Djordjevic, D., Hutchins, E.D., George, R.M., Cornelius, J.A., Rawls, A., et al., 2018. Identification of satellite cells from anole lizard skeletal muscle and demonstration of expanded musculoskeletal potential. *Dev. Biol.* 433, 344–356.
- The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70.
- The ENCODE Project Consortium, 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74.
- The GTEx Consortium, 2013. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* 45, 580–585.
- The Mouse ENCODE Consortium, 2012. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol.* 13, 418.
- Zinman, G.E., Naiman, S., Kanfi, Y., Cohen, H., Bar-Joseph, Z., 2013. ExpressionBlast: mining large, unstructured expression databases. *Nat. Methods* 10, 925–926.

The functionality and performances of C3 were compared with some other gene expression profile matching tools. The evaluation shows that C3 is a simple and effective method for cell type identification.

Chapter 3

Light-focusing human micro-lenses generated from pluripotent stem cells model lens development and drug-induced cataract in vitro

The C3 method described in the previous chapter was developed as part of a larger study to characterise ROR1⁺ cells derived from human pluripotent stem cells. The C3 method was applied to RNA-seq data obtained from these ROR1⁺ cells. This analysis showed the purified ROR1⁺ cells are most similar to primary human lens epithelial cells. This finding result was supported by further bioinformatics studies including principal component analysis as well as extensive cell biology-based characterisation techniques.

HUMAN DEVELOPMENT

RESEARCH ARTICLE

Light-focusing human micro-lenses generated from pluripotent stem cells model lens development and drug-induced cataract *in vitro*

Patricia Murphy^{1,2,*}, Md Humayun Kabir^{1,2,3,*}, Tarini Srivastava^{1,2,*}, Michele E. Mason^{1,2,*}, Chitra U. Dewi^{1,2}, Seakcheng Lim^{1,2}, Andrian Yang^{3,4}, Djordje Djordjevic^{3,4}, Murray C. Killingsworth⁵, Joshua W. K. Ho^{3,4}, David G. Harman^{1,2} and Michael D. O'Connor^{1,2,‡}

ABSTRACT

Cataracts cause vision loss and blindness by impairing the ability of the ocular lens to focus light onto the retina. Various cataract risk factors have been identified, including drug treatments, age, smoking and diabetes. However, the molecular events responsible for these different forms of cataract are ill-defined, and the advent of modern cataract surgery in the 1960s virtually eliminated access to human lenses for research. Here, we demonstrate large-scale production of light-focusing human micro-lenses from spheroidal masses of human lens epithelial cells purified from differentiating pluripotent stem cells. The purified lens cells and micro-lenses display similar morphology, cellular arrangement, mRNA expression and protein expression to human lens cells and lenses. Exposing the micro-lenses to the emergent cystic fibrosis drug Vx-770 reduces micro-lens transparency and focusing ability. These human micro-lenses provide a powerful and large-scale platform for defining molecular disease mechanisms caused by cataract risk factors, for anti-cataract drug screening and for clinically relevant toxicity assays.

KEY WORDS: Lens development, Stem cell, Organoid, Focus, Cataract, Vx-770

INTRODUCTION

During embryogenesis, the ocular lens arises from the lens placode in the surface ectoderm opposite the optic cup (Mann, 1964; Tholozan and Quinlan, 2007). Although the exact process can differ between vertebrate species, key lens features shared by vertebrates include an anterior lens epithelial cell (LEC) monolayer expressing α -crystallins overlying a mass of lens fibre cells expressing α -, β - and γ -crystallins (Thomson and Augusteyn, 1985). In mammals, invagination of the lens placode is followed by formation of the lens vesicle – a spherical LEC monolayer surrounding an acellular

lumen. Differentiation of the posterior LECs into lens fibre cells fills the lens vesicle lumen to establish the basic lens architecture. For decades these features have provided a framework for *in vitro* lens and cataract studies using explanted primary rat LECs. For example, our group reported *in vitro* regeneration of light-focusing rat lenses from paired rat LEC monolayers arranged to mimic lens vesicles (O'Connor and McAvoy, 2007). The size, cellular arrangement and protein expression within these *in vitro* regenerated rat lenses closely resembled newborn rat lenses. Continued culture of these regenerated rat lenses resulted in formation of a human-like cataract, as seen by reduced light transmission and reduced focusing ability.

To improve the suitability of *in vitro* lens regeneration for targeted and large-scale cataract studies, we investigated human pluripotent stem cells (hPSCs) as a source of LECs. A handful of studies have differentiated hPSCs to relatively impure populations of lens cells or 'lentoids' – small aggregates of randomly organised LECs and lens fibre cells (Fu et al., 2017; Li et al., 2016; Yang et al., 2010). Limitations with these approaches include the presence of contaminating non-lens cells, the spontaneous and random nature of lentoid production, and the production of only tens-to-hundreds (Fu et al., 2017; Li et al., 2016) or thousands (Yang et al., 2010) of lentoids. Although one report describes limited magnification ability of the lentoids (Fu et al., 2017), none of the published methods have been shown to produce biconvex lentoids that focus light to a point – the fundamental functional requirement of the lens – due to abnormal attachment of the lentoids to culture surfaces and/or other cell types.

Here, we describe a simple and efficient system for production of 10^6 – 10^8 purified LECs from hPSCs, and the subsequent controlled, robust and reproducible production of 10^3 – 10^5 light-focusing human micro-lenses. These micro-lenses possess anatomical and molecular features of primary human lenses, and exposing the micro-lenses to the cystic fibrosis drug Vx-770 decreases their ability to transmit and focus light. This platform provides a robust and accessible human system for modelling lens and cataract development, anti-cataract drug screening, and drug toxicity studies.

RESULTS

Characterisation of ROR1 as a LEC marker

We hypothesised that the impurity of LECs generated from PSCs via published methods, together with suboptimal culture conditions for these LECs, leads to uncontrolled lentoid production, uncontrolled lentoid shape, random detachment and loss of lentoids from the culture, and the inability to focus light. By modifying (Fig. 1A) an elegant three-stage growth factor treatment for lens cell differentiation (Yang et al., 2010), we

¹School of Medicine, Western Sydney University, Campbelltown, NSW 2560, Australia. ²Medical Sciences Research Group, Western Sydney University, Campbelltown, NSW 2560, Australia. ³Victor Chang Cardiac Research Institute, Darlinghurst, NSW 2010, Australia. ⁴St Vincent's Clinical School, University of New South Wales, Sydney, NSW 2010, Australia. ⁵Electron Microscopy Laboratory, NSW Health Pathology and Correlative Microscopy Facility, Ingham Institute, Liverpool, NSW 2170, Australia.

*These authors contributed equally to this work

‡Author for correspondence (m.oconnor@westernsydney.edu.au)

DOI: 10.1242/dev.155838

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution and reproduction in any medium provided that the original work is properly attributed.

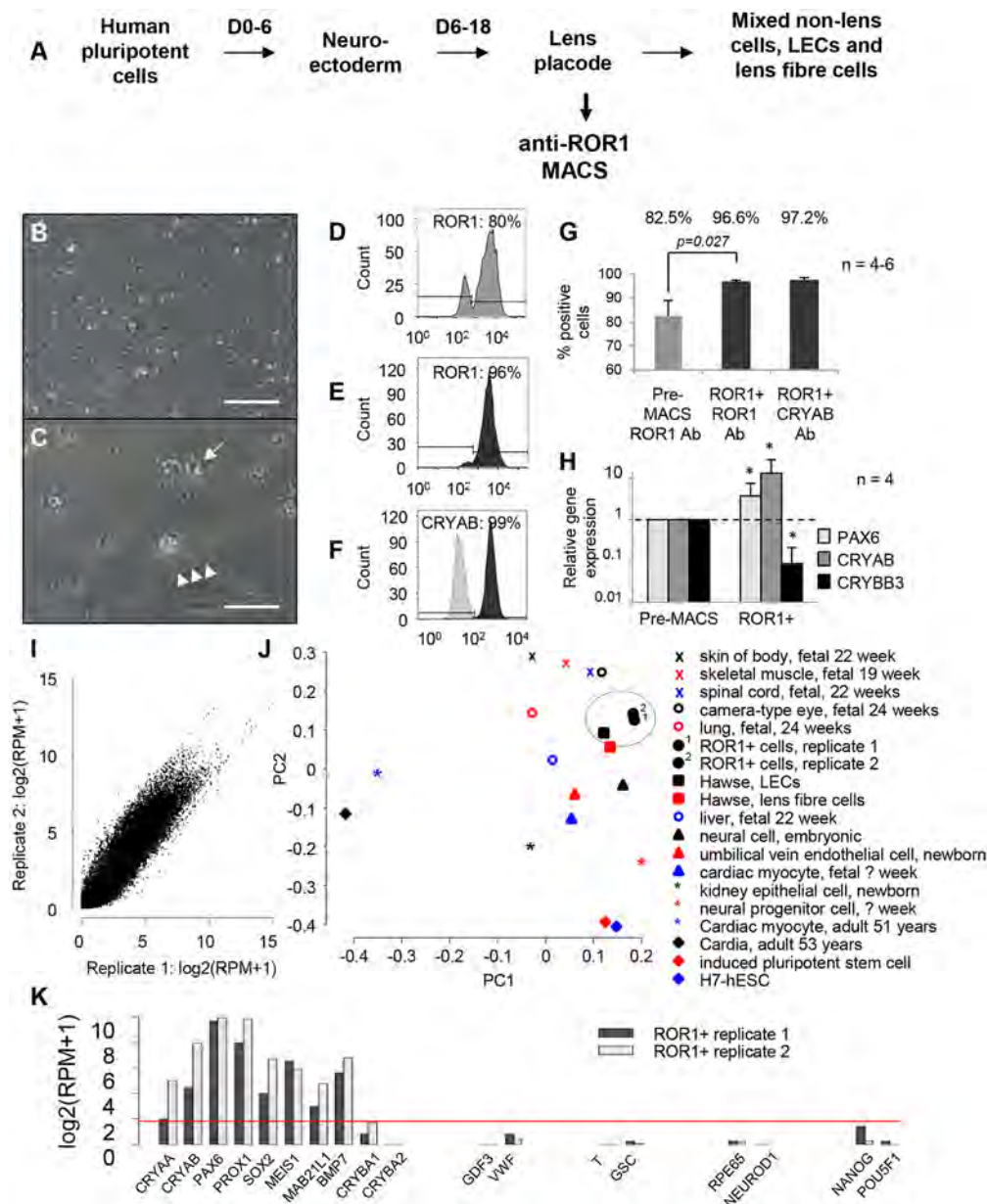


Fig. 1. Identification and characterisation of ROR1 as a LEC marker. (A) Schematic diagram showing the three-stage lens differentiation protocol, with modification to enable ROR1-based purification of LECs. (B,C) ROR1⁺ cells cultured at high cell densities showed uniform polygonal morphologies that formed tightly packed monolayers (B). When cultured at low cell densities or passaged in medium containing only FGF2 (C), ROR1⁺ cells became large and vacuolated (arrow) with stress fibres (arrowheads; cells shown 18 days after plating; $n=3$). Scale bars: 100 μm. (D-G) Flow cytometry data showing expression levels of: ROR1 prior to (D) and after (E) ROR1-based purification; CRYAB after ROR1-based purification (F); and average expression levels before and after purification (G). (H) Relative mRNA transcript expression levels for PAX6, CRYAB and the lens fibre-specific gene *CRYBB3* after ROR1⁺ cell separation ($*P<0.05$). (I) Pearson correlation showing high similarity (>0.96) between RNA-seq libraries generated from two independent ROR1⁺ cell samples. (J) Principal component analysis shows the ROR1⁺ RNA-seq transcriptomes are most similar to primary human LECs (circled). (K) Representative data from the ROR1⁺ RNA-seq libraries shows key genes required by LECs are expressed (*CRYAA*, *CRYAB*, *PAX6*, *PROX1*, *SOX2*, *MEIS1*, *MAB21L1*, *BMP7*). In contrast, genes expressed by lens fibre cells (*CRYBA1*, *CRYBA2*) or various endodermal cells (*GDF3*, *VWF*), mesodermal cells (*T*, *GSC*), non-lens ectodermal cells (*RPE65*, *NEUROD1*) and pluripotent cells (*NANOG*, *POU5F1*) are not expressed. Data shown in B,C and D-H are representative of 50 and four (respectively) independent differentiation experiments using four different hPSC lines; data are mean \pm s.e.m. in G,H.

increased lentoid production, lentoid retention, and expression of LEC and lens fibre cell genes (Fig. S1). Nevertheless, heterogeneous cell morphologies were still obtained, lentoid production was still uncontrolled, lentoids still detached and were lost, and the lentoids did not focus light when assessed via light microscopy. As an alternative approach, analysis of published lens microarray data (Hawse et al., 2005) identified the receptor tyrosine

kinase-like orphan receptor 1 (ROR1) as a potential LEC purification antigen (Fig. S2). *In situ* hybridisation showed ROR1 is highly expressed by mouse LECs at embryonic day 14, and PCR showed ROR1 transcript expression at a similar stage of the three-stage lens differentiation protocol.

Magnetic-activated cell sorting (MACS) using an anti-ROR1 antibody during stage 2 of the lens differentiation protocol (Fig. 1A)

consistently produced a homogeneous population of polygonal cells (supplementary material File S2; Fig. 1B). However, this LEC-like morphology changed to large and vacuolated when the ROR1⁺ cells were cultured at low density or passaged in medium containing only FGF2 (Fig. 1C). These polygonal (LEC-like) and vacuolated (abnormal) cell morphologies are highly similar to primary human foetal LECs either when first explanted or when passaged (Ringens et al., 1982). Flow cytometry revealed that 95–100% of the captured cells were ROR1⁺, and over 99% expressed CRYAB (crystallin α B) (Fig. 1D–G). Comparing the ROR1⁺ cells with the starting population revealed that their transcriptional profile is consistent with purification of LECs. This included a threefold increase in PAX6, a 10-fold increase in CRYAB and a 10-fold decrease in CRYBB3 (crystallin β B3) expression (Fig. 1H). Comparing whole-transcriptome RNA-seq profiles from different ROR1⁺ cell batches showed highly similar expression patterns (Pearson correlation >0.96; Fig. 1I), indicating high reproducibility of the cell separation. Principal component analysis found the ROR1⁺ RNA-seq libraries to be most closely related to primary adult human LECs (Fig. 1J). This was further reinforced by gene set analysis where comparison of the ROR1⁺ RNA-seq data against a compendium of 145 human and 95 mouse gene expression datasets revealed the ROR1⁺ transcriptomes to be most similar to human LECs (Hawse et al., 2005) and mouse LECs (Hoang et al., 2014) (false discovery rates <9.6 $\times 10^{-3}$ and <5.88 $\times 10^{-8}$, respectively). Over 90 transcripts indicative of LECs (Lachke et al., 2012) were reproducibly expressed in the ROR1⁺ RNA-seq libraries, whereas genes associated with various endodermal, mesodermal, non-lens ectodermal or pluripotent cells were not (Fig. 1K), thus supporting the high purity and LEC nature of the ROR1⁺ cells. Consistent with this, transplantation of ROR1⁺ cells into immunocompromised mice showed no teratoma formation unless hPSCs were deliberately transplanted with the ROR1⁺ cells (Fig. S2).

Combinatorial growth factor screening for ROR1⁺ proliferation

To avoid the large vacuolated phenotype seen with initial passaging of the ROR1⁺ cells, a combinatorial growth factor screen was undertaken to test six signalling pathways (nine growth factors) whose receptors are expressed by LECs (Fig. 2A). As FGF signalling is crucial for lens development (Lovicu et al., 2011; Wu et al., 2014) FGF2 was included in the basal medium at 10 ng/ml (TM32; Fig. 2A) – a concentration known to stimulate rat LEC proliferation and migration but not differentiation to fibre cells. All combinations of the remaining five test pathways were assayed on ROR1⁺ cells derived from four hPSC lines. Imaging Hoechst-stained nuclei revealed that media containing insulin and IGF1 (insulin-like growth factor 1) greatly increased ROR1⁺ cell yield (Fig. 2B–I). Mass spectrometry revealed ROR1⁺ cells cultured in these media expressed α - but not β -crystallin proteins (Fig. S3). The high CRYAA sequence coverage obtained indicates that it is one of the most abundant proteins expressed by the cultured ROR1⁺ cells. In contrast, media that contained BMPs had lower cell yield (Fig. 2H,I), and ROR1⁺ cells cultured in these media expressed lens fibre cell crystallin proteins, including CRYBB1, CRYBB2 and CRYBB3 (Fig. S3). As TM17 was among the best-performing proliferation media for ROR1⁺ cells derived from all four PSC lines in both low and high cell-seeding conditions, it was used as the routine ROR1⁺ maintenance medium. TM17 supported ROR1⁺ cell freeze/thawing with retention of α -crystallin protein expression

and cell morphology (Fig. 2J,K and Table S1). After ~2 weeks of high density culture in TM17, or after exposure to media containing FGF2 and BMP4/7, random lentoid production and lens fibre cell crystallin expression was seen (Fig. 2L,M and Table S2).

Large-scale production of light-focusing human micro-lenses

For controlled and large-scale production of *in vitro* lenses suitable for drug-screening, ROR1⁺ cells underwent forced aggregation to generate small (~100 μ m diameter) LEC aggregates similar to the LEC mass seen during zebrafish lens development. This approach is capable of generating 1200 spherical aggregates per well of a 24-well plate (Fig. S3). These aggregates were embedded in agarose to minimise attachment to each other or the culture dish, and then maintained for up to 6 weeks in stage 3 lens differentiation medium (Yang et al., 2010) on top of the agarose. The cultured aggregates were imaged at various times using phase microscopy (their small size precluding non-phase imaging). Initially, these spherical aggregates transmitted less light than the surrounding culture medium due to an underlying opacity throughout the aggregates (Fig. 3A). However, as culture progressed, this opacity gradually reduced in both size and intensity, such that by ~2 weeks of culture the aggregates transmitted similar levels of light to the surrounding culture medium (Fig. 3C,E,H,Q). Concomitant with this increase in light transmittance was a striking increase in light-focusing ability. At the beginning of the culture, the cell aggregates displayed very little focusing ability, with the light intensity at the maximal focal point only equivalent to the light intensity of the surrounding culture medium (Fig. 3B). However, as the culture progressed and the light transmittance increased, so too did the light intensity at the focal point below the cell aggregates (Fig. 3D,F). Detailed characterisation of this light-focusing property immediately after aggregation (Fig. 3G,I,K,M,O,Q) compared with after ~3 weeks of culture (Fig. 3H,J,L,N,P,Q) revealed the cell aggregates developed a remarkable capacity to focus light. In some experiments, some micro-lenses were seen to have clusters of non-transparent cells located adjacent to the micro-lens periphery, likely due to damage during the embedding process (Fig. S4); these clusters tended not to prohibit assessment of light transmittance or focusing ability. Taken together, these data demonstrate that the initial shape and internal composition of the ROR1⁺ cell aggregates is insufficient for either transparency or significant focusing ability, and that gross morphological changes observed over a number of weeks are associated with the development of both maximal light transmittance (relative to the culture medium alone) and maximal light-focusing ability.

Micro-lens functions are associated with lens fibre crystallin expression

To determine whether the observed changes in light transmittance and focusing ability were associated with changes in crystallin expression, PCR, immunofluorescence and mass spectrometry were used to assess the expression of α -, β - and γ -crystallins at various times during the aggregate culture. The PCR analyses (Fig. 4A) revealed the opposite trend from that seen with ROR1 cell purification (Fig. 1H): i.e. culture of the ROR1⁺ cell aggregates resulted in decreased relative expression of PAX6 and CRYAB mRNA but increased expression of CRYBB3 mRNA. The immunofluorescence analyses showed homogenous expression of CRYAA protein at day 14 (Fig. 4C), whereas expression of

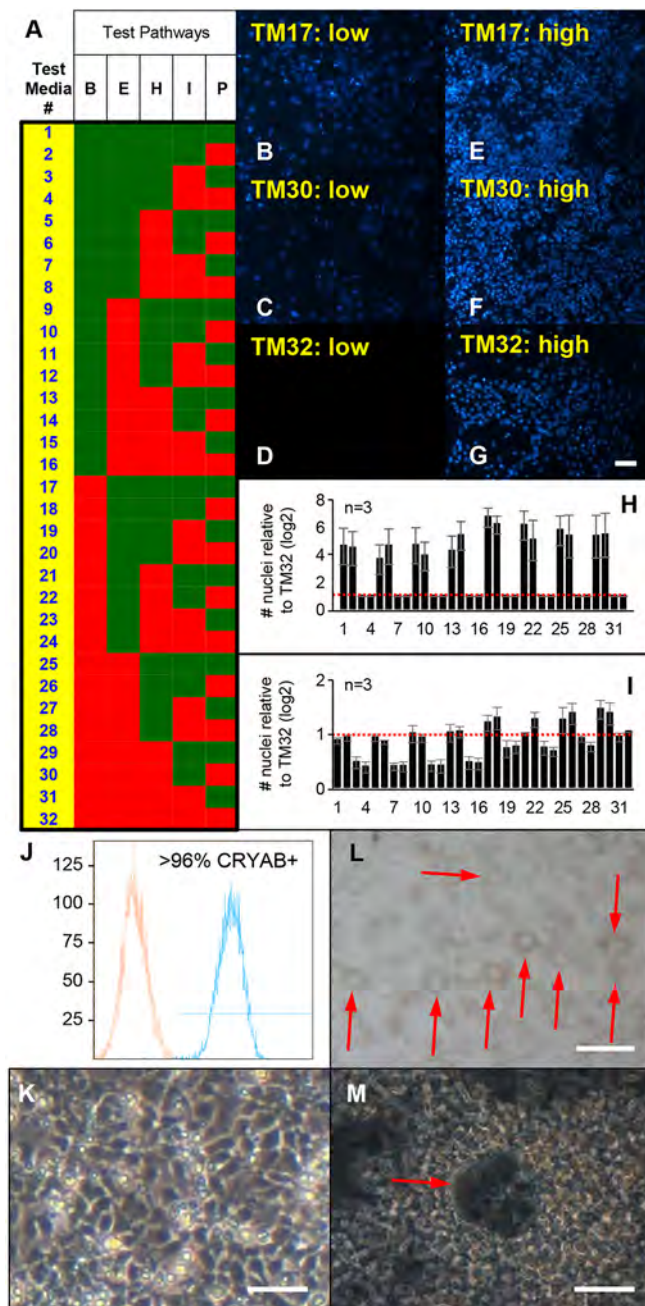


Fig. 2. Combinatorial growth factor screening identified media for ROR1⁺ cell expansion and differentiation. (A) Schematic diagram showing composition of the test media. FGF2 was included in the basal medium (TM32), with all combinations of the five other test pathways (eight growth factors) tested as shown [B, BMP4, BMP7; E, EGF, TGF α , H, HGF; I, insulin, IGF1; P, PDGF-AA; green square represents factor(s) present; red square represents factor(s) absent]. (B-I) Data from Hoechst-stained ROR1⁺ cells cultured in TM17 (B,E), TM30 (C,F) and TM32 (D,G) after seeding at low (B-D) and high (E-G) cell density, as well as average Hoechst-stained nuclei counts for all media (H,I). These data reveal that TM17 promoted expansion of ROR1⁺ cell cultures while maintaining expression of α - but not β -crystallins (see supplementary material Fig. S3). Scale bar: 20 μ m. (J,K) Flow cytometry and light microscopy data show ROR1⁺ cells expanded, frozen, thawed and cultured for 6 days in TM17 retain high levels of CRYAB expression (J) with expected morphology (K) but without detectable expression of β -crystallins (see Table S1 and Fig. S4). Scale bar: 40 μ m. (L,M) Light microscopy images show spontaneous production of lentoid-like structures after being expanded, frozen and thawed in TM17, then cultured in stage 2 lens differentiation medium. Cells in these cultures expressed α - and β -crystallins (see Table S2 and Fig. S3). Scale bars: 200 μ m in L; 40 μ m in M. The data shown in B-I are each representative of three independent differentiation experiments; data are mean \pm s.e.m. in H,I.

large nuclei at the periphery of the aggregates (Fig. 5A). At this stage, the bulk of the aggregates consisted of relatively small cells with large nuclei and numerous organelles (Fig. S6A). The nuclei in these bulk cells were typically rod-shaped, with prominent nucleoli and darker nuclear substance compared with the surrounding cytoplasm (Fig. 5B) – similar to the nuclear morphology seen during the early stages of lens fibre cell differentiation *in vivo* (Kuwabara and Imaizumi, 1974; Vrensen et al., 1991).

Later in culture, LEC-like cells could be found at the periphery of the transparent and focusing micro-lenses (Fig. 5C). In larger aggregates (~200 μ m and more diameter) multi-layering of the LECs could be seen that was not apparent in smaller micro-lenses (Fig. S7A). In these later cultures, the bulk of the micro-lenses of all sizes were composed of large cells with varied cross-sectional sizes and relatively homogenous cytoplasm (Fig. S6B,C). These lens fibre-like cells were typically joined by complex membrane interdigitations (Fig. 5D); enlarged and degenerating organelles such as mitochondria could also be found (Fig. 5E) similar to those seen in lens fibre cells *in vivo* (Vrensen et al., 1991). The nuclei within some of these lens fibre-like cells were large and circular-profiled with spoke-like nucleoli (Fig. 5F). In other cells, the nuclear membrane was only recognisable as a chain of vesicles, with some of the nuclear substance appearing to be indistinguishable from the cell cytoplasm (Fig. 5G) – these nuclear morphologies being similar to those seen *in vivo* within terminally differentiating fibre cells of the late bow zone (Kuwabara and Imaizumi, 1974). In some instances, secondary lens fibre-like cells could be seen at the periphery of the aggregates overlying the bulk cells and adjacent to LEC-like cells (Fig. S6D).

Micro-lenses show evidence of lens capsule formation

As a first step towards investigating production of lens capsule-related material within the ROR1 system, the ROR1⁺ RNA-seq libraries were examined. This analysis revealed ROR1⁺ cells express a range of integrins, collagens and laminins that are known to be required for normal lens development *in vivo* (Table S4). Subsequent immunofluorescence experiments localised laminin and collagen IV expression within the micro-lenses to peripherally located LEC-like cells, which appeared multi-layered in larger micro-lenses (Fig. S8). Electron microscopy revealed the presence of a thin, lens capsule-like material around the

both β - and γ -crystallins was sometimes variable (Fig. 4E,G). Immunofluorescence at day 24 showed homogenous expression of CRYAA (Fig. 4I) and more homogenous expression of β - and γ -crystallins (Fig. 4K,M). Expression of lens fibre cell crystallins was similarly supported by mass spectrometry analyses that routinely identified predominantly α - and β -crystallin proteins among the most abundant proteins expressed by the light-focusing micro-lenses (Table S3 and Fig. S5).

Micro-lens functions are associated with lens fibre cell maturation

As specific ultrastructural changes have also been associated with lens fibre cell development, electron microscopy was performed to characterise the cellular organisation within the developing micro-lenses. Early in culture, these analyses revealed LEC-like cells with

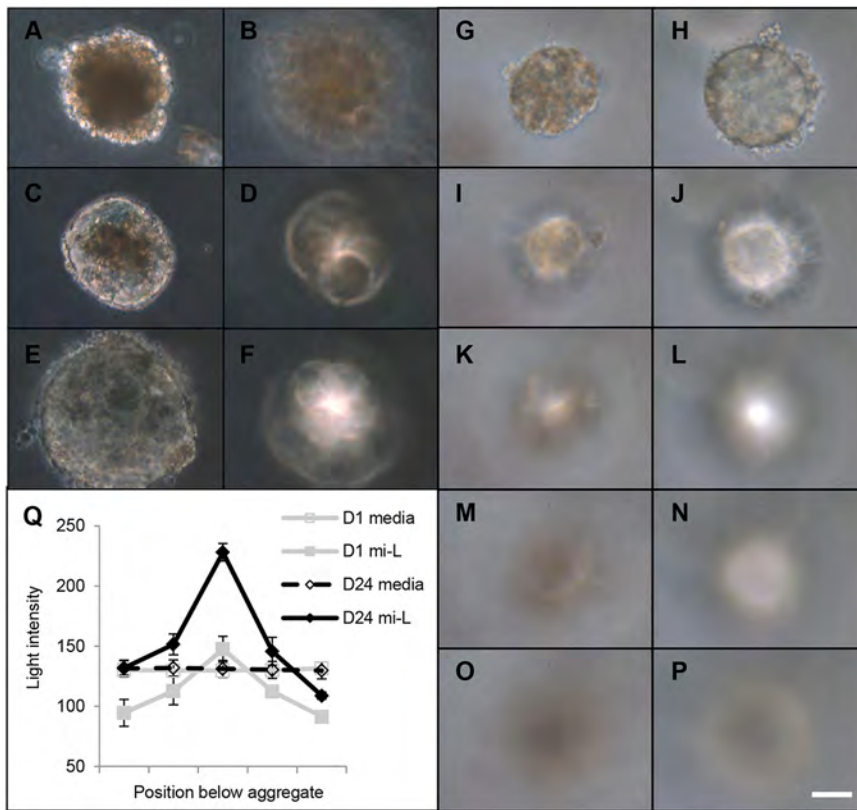


Fig. 3. ROR1⁺ cell aggregation leads to transparent and light-focusing micro-lenses. (A–Q) Light microscopy data from ROR1⁺-cell aggregates and the maximal focal point below them. After 3 days of culture, the aggregates transmitted less light than the surrounding culture medium (A) and did not focus light (B). As culture progressed, the aggregates transmitted more light (C, day 7; E, day 14) and began focusing light (D, day 7; F, day 14). More-detailed characterisation of a single aggregate shows it had limited transparency (G) and focusing ability (I,K,M,O) on day 3 of culture, but by day 27 it transmitted the same amount of light as the surrounding culture medium (H) and had developed significant focusing ability (J,L,N,P). Quantification of the light transmittance and focusing ability confirms these findings (Q). Scale bar: 40 μ m. The images are representative of five micro-lenses from two biological replicates; data in Q are mean \pm s.e.m.

micro-lenses that appeared thicker in micro-lenses that had been cultured for longer (Fig. S7A–D).

Vx-770 induces cataract in micro-lenses

The underlying reason for generating functional human lenses *in vitro* was to provide a source of functional human lens tissue for investigating cataract risk factors. To determine whether the micro-lens system might be suitable for investigation of clinically relevant cataracts, the micro-lenses were exposed to Vx-770 – a potentiator of activity for the cystic fibrosis transmembrane conductance regulator (CFTR) protein. This emerging cystic fibrosis drug is reported to have caused cataracts in rats (McColley, 2016). Cataracts have also been reported in children and adolescents receiving Vx-770 (Talamo Guevara and McColley, 2017), and clinical trials are yet to discount an association with its use and cataract formation in paediatric patients (Dryden et al., 2016; McColley, 2016; Van Goor et al., 2009). The concentrations used to test the effect of Vx-770 on ROR1⁺ aggregates (i.e. up to 2000 ng/ml) covered the plasma concentration range reported for children treated with Vx-770 (Davies et al., 2016). When included from the start of culture, the ROR1⁺ cell-aggregates exposed to vehicle or 200 ng/ml Vx-770 transmitted similar levels of light compared with the culture medium (Fig. 6A,B). In contrast, the aggregates exposed to 500 ng/ml and higher were less transparent, transmitting significantly less light than the culture medium and the vehicle- and 200 ng/ml-treated aggregates (Fig. 6C,D). Continued culture of these treated aggregates resulted in focusing ability developing in both the vehicle- and 200 ng/ml-treated aggregates (Fig. 6E,F,H), but not the aggregates treated with 500 ng/ml Vx-770 or more (Fig. 6G,H).

To test whether Vx-770 would affect micro-lens function after focusing ability had developed, aggregates were first allowed to develop focusing ability and only then were they exposed to

Vx-770. Interestingly, comparison of micro-lens transparency before and after treatment revealed no measurable decrease in light transmittance with any of the treatments (Fig. 6I–O). When comparing micro-lens focusing ability before and after treatment, neither control-treatment nor 200 ng/ml Vx-770 decreased the focusing ability (Fig. 6P,Q,S,T,V). In contrast, micro-lenses treated with 2000 ng/ml Vx-770 showed a large and significant reduction in focusing ability (Fig. 6R,U,V). Notably, these Vx-770-induced effects occurred regardless of the micro-lens size (i.e. from 80 μ m to 200 μ m in diameter).

DISCUSSION

ROR1⁺ cells closely resemble human LECs

The inability to reliably access large amounts of functional human lens tissue has hampered cataract research for decades. Until now, no effective conditions have been identified for simple, robust and large-scale generation of either purified LEC populations or light-focusing lenses, from PSCs of any species. Previous reports of hPSC-based LEC models have been limited by the presence of contaminating non-lens cells, in some cases up to ~60% non-CRYAA-expressing cells (Yang et al., 2010). Other limitations include the requirement for either: successive rounds of minimally scalable manual purification of LEC progenitor cells that display spontaneous lentoid body formation (Fu et al., 2017; Li et al., 2016); or complex five-colour flow cytometry that requires simultaneous positive and negative selection to obtain small numbers of partially purified LEC progenitors that undergo spontaneous lentoid production (Mengarelli and Barberi, 2013). In contrast, the extensive characterisation data shown here demonstrate simple, robust and large-scale MACS-based purification of ROR1⁺ human LECs. This semi-automated process is capable of generating 10⁶ to 10⁸ ROR1-purified human LECs (from 1 \times 35 mm dish to 6 \times T175

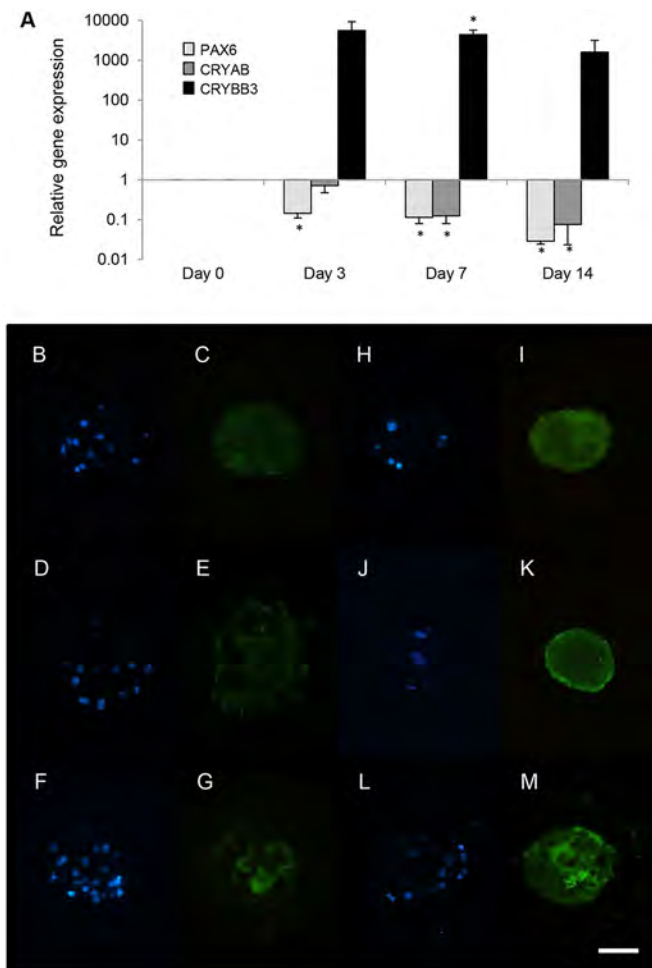


Fig. 4. Aggregation of ROR1⁺ cells induces lens fibre cell crystallin expression. (A) Real-time PCR analysis of aggregated ROR1⁺ cells results in decreased relative expression of PAX6 and CRYAB, and increased expression of CRYBB3 (**P*<0.01; data obtained from four biological replicates and presented as mean±s.e.m.). (B–M) Immunofluorescence analysis shows that after 14 days of culture, αA-crystallin (C) was expressed uniformly throughout the bulk of the micro-lenses, whereas β-crystallin (E) and γ-crystallin (G) were not. After 24 days of culture, αA-crystallin (I), β-crystallin (K) and γ-crystallin (M) were all expressed relatively uniformly throughout the bulk of the micro-lens. The location of DAPI-stained nuclei within the day 14 (B,D,F) and day 24 (H,J,L) aggregates are shown. Scale bar: 40 μm. Each image is representative of five micro-lenses from two biological replicates.

flasks, respectively). Morphological, transcriptomic and proteomic analyses of these ROR1⁺ cells demonstrate them to be most similar to human LECs, thereby providing a simple and large-scale source of purified human LECs for lens and cataract research.

ROR1⁺ cells for investigation of posterior capsule opacification

The defined, proliferation-inducing culture conditions identified here provide an extended period of time for investigation of factors that affect human LEC biology compared with previously reported PSC-based lentoid systems. The identification of insulin and IGF1 as significant inducers of ROR1⁺ cell proliferation, and BMPs as inducers of β-crystallin expression, is consistent with known roles for these factors in lens cells from other species (Lovicu et al., 2011). A variety of chick- and rat-based studies have shown that insulin and IGF1 can induce a proliferative response in LECs, and BMP signalling can potentiate aspects of lens fibre cell differentiation (which has also

been shown for insulin/IGF1 in some circumstances). The observation that ROR1⁺ cells plated at low density change into large, vacuolated cells with stress fibres suggests they might be suited to investigating posterior capsule opacification (PCO) – the most common complication arising from cataract surgery. The simplicity and scalable nature of the ROR1⁺ cells may provide advantages over existing primary human LEC models of PCO (Wormstone and Eldred, 2016). Ongoing work is aimed at further elucidating how TGFβ signalling integrates with signalling via FGF, insulin and IGF1, BMP and other pathways in ROR1⁺ cells, for comparison with what is currently known of the molecular development of PCO.

Functional human organoids from non-mammalian developmental templates

Previous work from our group demonstrated that physiologically sized, transparent and light-focusing rat lenses could be generated *in vitro* by mimicking aspects of the lens vesicle stage of mammalian lens development (O'Connor and McAvoy, 2007). These paired explant-derived *in vitro* lenses contained LEC monolayers and bulk compartments of lens fibre cells undergoing terminal differentiation. The desire to generate much smaller yet still functional human micro-lenses (suitable for developmental biology and drug screening applications) led to the hypothesis that partially mimicking aspects of teleost lens development using ROR1⁺ LECs might produce transparent and light-focusing lenses. Zebrafish lens development was chosen as a template as these teleosts have small lenses, and because anatomical features of their lens development have been well described (Greiling and Clark, 2012).

In zebrafish, cells of the surface ectoderm delaminate to form a lens cell mass, rather than a lens vesicle, that measures ~80 μm in diameter during primary lens fibre cell differentiation. Differentiation of the cells within the lens cell mass forms the primary fibre cells, while at the same time the lens epithelium forms from cells at the periphery of the lens placode. To approximate these events *in vitro*, ROR1⁺ cells underwent forced aggregation to generate spheroidal masses of LECs. The aggregates were then embedded in agarose to minimise attachment to each other or the culture surface, before being asymmetrically exposed to medium containing FGF2 and Wnt3a (Yang et al., 2010) to mimic growth factor delivery to the lens. This approach generated both transparent and light-focusing human micro-lenses, and is the first demonstration of its kind. This approach also provides significantly greater control over the size, timing and location of micro-lens generation compared with existing methods of hPSC-based lentoid generation. Notably, no spontaneous loss of micro-lenses occurs, and the process can generate ~2.5×10³ to ~2.5×10⁵ light-focusing micro-lenses (from 1×35 mm dish to 6×T175 flasks, respectively). The appearance of LEC multi-layering in the larger micro-lenses (~200 μm diameter or more, or ~2.5× the diameter of equivalently staged zebrafish lens masses) suggests there is an upper size limit to the utility of this approach. Nevertheless, the establishment of more anatomically correct functional lens tissue below this size limit (achieved by varying the input cell number during ROR1⁺ cell aggregation) suggests mimicking non-human anatomical templates may be a useful approach for generating small, rudimentary functional tissues for developmental biology and drug screening applications.

Lens fibre cell differentiation in micro-lenses mimics events seen *in vivo*

The gradual appearance of transparency and light-focusing within the micro-lenses indicates these functional properties developed as a result of specific cellular and molecular changes that occurred

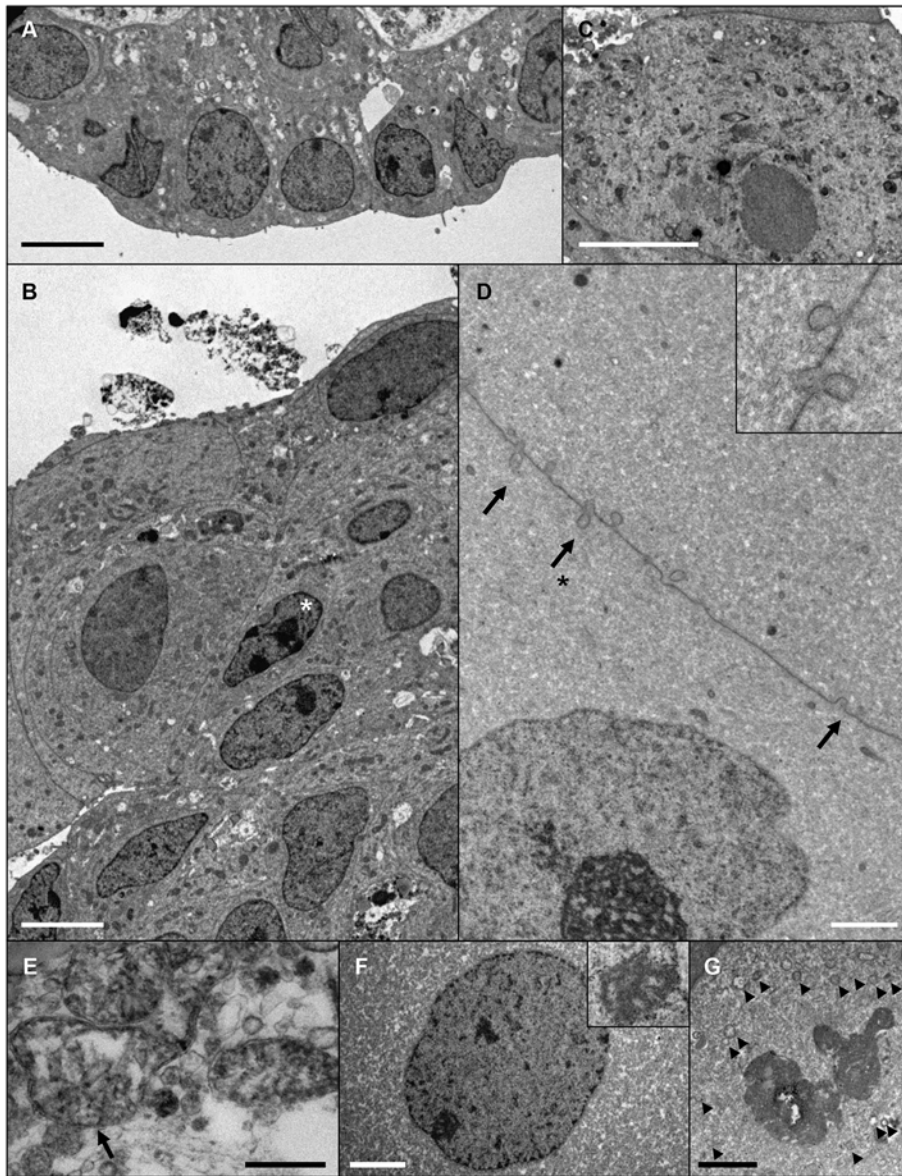


Fig. 5. Evidence of progressive lens fibre cell differentiation in ROR1⁺ cell aggregates. Electron microscopy data from cultured aggregates. (A,B) A micro-lens cultured for 14 days shows a monolayer of LEC-like cells at the periphery of the tissue (A), and cells with small, rod-shaped nuclei (asterisk) and numerous organelles within the bulk of the tissue (B). (C) LEC-like cell with numerous organelles present at the periphery of a micro-lens after 24 days of culture. (D-G) Ultrastructural indicators of lens fibre cell differentiation within a micro-lens cultured for 42 days. (D) Ball-and-socket type membrane interdigitations (arrows) between adjacent lens fibre-like cells (inset shows a higher magnification of the region indicated with an arrow and asterisk). (E) A swollen mitochondria (arrow). (F) An enlarged nuclei with spoke-like nucleolus (inset). (G) A degraded nuclei with nuclear membrane visible as a series of vesicles (arrowheads). Scale bars: 5 µm in A-C; 2 µm in D, F, G; 0.5 µm in E. Images are representative of seven micro-lenses obtained from two biological replicates.

within the ROR1⁺ cell aggregates over a period of weeks. Thus, neither the initial shape nor the internal composition of the freshly aggregated ROR1⁺ cells is the main determinant of either transparency or focusing ability.

In vivo, the establishment of lens transparency and focusing ability is thought to result from a combination of specific events that occur in parallel. Maintenance of an anterior LEC monolayer occurs while cells at the edges of this monolayer differentiate into primary and then secondary lens fibre cells. At the same time, lens fibre cell production is associated with: cell elongation (the increasing cell size helping to provide the required lens shape); cell alignment (to minimise extracellular light scatter); expression and accumulation of α -, β - and γ -crystallins (to provide the required refractive index); loss of organelles, including the nucleus (to remove intracellular light-scattering particles); and accumulation of complex membrane interdigitations (to maintain maximal alignment of the fibre cells and thus minimise the possibility of light-scattering due to intercellular spaces).

Loss of the opaque phenotype initially seen in the aggregated ROR1⁺ cells, and concomitant development of transparency and

light focusing, appear to recapitulate key aspects of the above lens-development processes. Regions of LEC-like cells were present at the periphery of the aggregates. Within the bulk of the aggregates, lens fibre-like cells became larger and their cross-sectional profiles were varied, as seen in mouse, bovine and chick primary lens fibre cells (al-Ghoul and Costello, 1997; Shestopalov and Bassnett, 2000; Taylor et al., 1996). The cytoplasm of these fibre-like cells became more homogenous as β - and γ -crystallins were expressed, with some evidence suggestive of rudimentary secondary fibre cell production. Organelles in these lens fibre-like cells showed evidence of being degraded, including progressive appearance of classic nuclear degradation morphologies indicative of terminal lens fibre cell denucleation (i.e. rod-shaped nuclei early in culture; later in culture spoke-like nucleoli and evidence of nuclear membrane breakdown) (Kuwabara and Imaizumi, 1974; Vrensen et al., 1991). The lens fibre-like cells also accumulated complex membrane interdigitations that are characteristic of the lens, such as 'ball-and-socket' type junctions, a feature not described before in other PSC-derived lentoids.

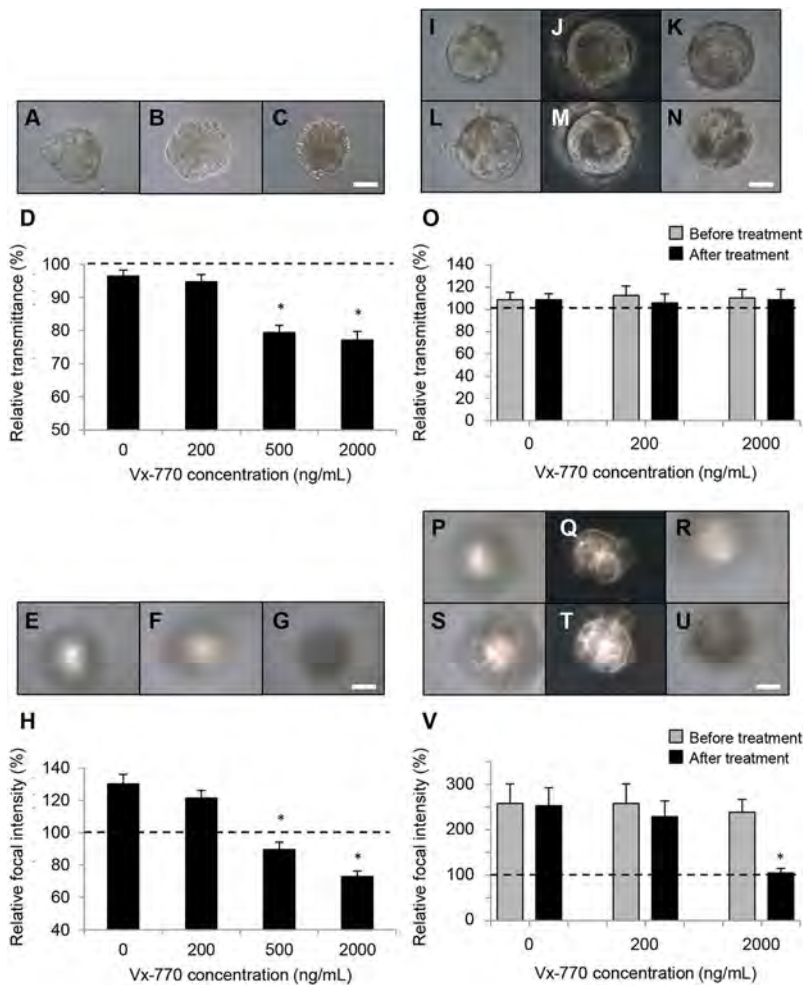


Fig. 6. The CFTR potentiator Vx-770 inhibits light focusing in ROR1 micro-lenses. (A-D) Light microscopy data showing ROR1⁺ micro-lenses treated with DMSO-only (A) or 200 ng/ml Vx-770 (B) transmitted light at similar levels to the culture medium after 24 days of culture, whereas a micro-lens treated with 2000 ng/ml Vx-770 transmitted less light (C). Scale bar: 40 μ m. Quantitative data are shown in D. (E-H) Light microscopy data showing ROR1 micro-lenses treated with DMSO-only (E) or 200 ng/ml Vx-770 (F) had developed focusing ability after 24 days in culture, whereas micro-lenses treated with 2000 ng/ml Vx-770 had not (G). Scale bar: 40 μ m. Quantitative data are shown in H. (I-O) Light microscopy data showing micro-lenses treated after they had developed focusing ability. Micro-lenses treated with DMSO-only (I,L), 200 ng/ml Vx-770 (J,M) or 2000 ng/ml Vx-770 (K,N) all transmitted similar levels of light after 7 days of treatment (L-N) compared with before treatment (I-K). Scale bar: 40 μ m. Quantitative data are shown in O. (P-V) Light microscopy data showing micro-lenses treated after having developed focusing ability. Micro-lenses treated with DMSO-only (S) and 200 ng/ml Vx-770 (T) retained focusing ability after 7 days of treatment compared with before treatment (P,Q, respectively). A micro-lens treated with 2000 ng/ml Vx-770 focused light prior to treatment (R) but did not after 7 days of treatment (U). Scale bar: 40 μ m. Quantitative data are shown in V. In D,O,H,V, * $P < 1 \times 10^{-4}$. The data shown (mean \pm s.e.m.) were each obtained from measurements of 15 micro-lenses from three biological replicates.

Production of the lens capsule and expression of associated integrins are also crucial elements of lens development; loss of capsule components or related integrins leads to lens malformations (Walker and Menko, 2009; Wederell and de Jongh, 2006). The data presented here shows ROR1⁺ LECs express transcripts for key extracellular matrix molecules and lens-related integrins. Additionally, LEC-like cells within the light-focusing micro-lenses expressed laminin and collagen IV. Thin, nascent lens capsule-like material could also be found, of a similar thickness to that seen with published stem cell-derived lentoids (Fu et al., 2017; Li et al., 2016; Yang et al., 2010). These observations – when combined with evidence of LEC maintenance, terminal lens fibre cell differentiation and well-contained lens substance in the majority of micro-lenses – suggest that integrin-related signalling pathways are being sufficiently stimulated for development and maintenance of transparency and focusing. The role of this lens capsule-like material in development of the ROR1⁺ micro-lenses could be further tested, e.g. by generating micro-lenses from hPSCs that have been derived from individuals known to develop lens capsule-related cataract, such as those with Alport syndrome (Chen et al., 2015; Song et al., 2011).

Towards a systems biology blueprint for development of lens function

Given that key features of normal lens development were observed as the micro-lenses become transparent and focused light, these micro-lenses may be useful for defining and testing an emerging

systems biology blueprint for development of lens function. Namely, how signalling via defined growth factors leads to integration of multiple intracellular signalling cascades that activate progressive transcriptional changes (and other changes – membrane dynamics, protein packing, etc.) that lead to establishment of a functional three-dimensional tissue. Decades of important research have identified growth factors (including FGFs, insulin and IGF, BMPs, Wnt) and related kinases (Lovicu et al., 2011), as well as transcription factors (PAX6, FOXE3, etc.) and target genes (Cvekl and Zhang, 2017) required for normal lens development – and how some of these elements interact. Recent studies have added to this body of knowledge, with this new information yet to be fully incorporated with prior knowledge (Anand and Lachke, 2017). The defined ROR1⁺ cell type and associated culture media of the micro-lens system suggests it may be possible to integrate this new information to create and test a comprehensive molecular blueprint for development of lens transparency and focusing, e.g. via growth factor variation, time-course transcriptional profiling, as well as gain- and/or loss-of-function studies. Interesting studies could include investigation of lens capsule production, which appears less extensive in all of the PSC-based lentoid systems compared with the normal lens. Similarly, variability in the timing of denucleation between fibre-like cells has been described in each of the PSC-based lentoid systems. These observations could be due to influences such as the limited growth factor set used to induce lens fibre cell differentiation (i.e. FGF2 and Wnt3a), and/or the relatively large space around the

lentoids (compared with the lens *in vivo*) that could alter the concentration, and therefore effectiveness, of autocrine and paracrine factors. The ROR1⁺ micro-lens system has the potential to further interrogate these issues and provide human-specific information related to development of lens function. A human-specific lens development blueprint could also have relevance for understanding how changes resulting from congenital mutation, environmental insults and ageing lead in isolation or combination to presbyopia and cataracts.

Investigating human cataracts using ROR1⁺ cells and micro-lenses

The clinical relevance of the micro-lens system is supported by the finding that focusing ability was decreased by treatment with high concentrations of Vx-770 (at the upper-range of minimum circulating plasma concentrations detected in paediatric patients treated with Vx-770). This loss of focusing ability occurred regardless of the timing of Vx-770 treatment (i.e. with treatment before or after development of micro-lens focusing) and independently of micro-lens diameter.

Nuclear and posterior subcapsular cataracts have been noted in rat pups treated with Vx-770, and cataracts have also been noted in child and adolescent cystic fibrosis patients treated with Vx-770. Eye examinations are therefore recommended for children being treated with Vx-770 (Talamo Guevara and McColley, 2017). Little-to-no detail is available on the histology or mechanism of Vx-770-induced cataract in either rats or humans. Thus, the ROR1⁺ micro-lens system appears to be a relevant and useful human model of lens function for future investigations into the molecular mechanisms of Vx-770-induced cataract (e.g. assessing protein aggregation, changes to membrane properties, vacuolisation, cell death, etc.). The findings that neither transparency nor focusing ability developed when Vx-770 was included from the start of culture suggests that LECs are affected by high Vx-770 concentrations. Whether the lens fibre-like cells are also affected needs to be determined – a possibility based on the loss of focusing observed when Vx-770 treatment was applied after focusing had developed.

The Vx-770 data also suggest that the micro-lens system may be a relevant model for investigating other known cataract risk factors, the mechanisms of action of which are yet to be fully defined (e.g. genetics, age, diabetes, smoking, UV light, radiation, drugs, etc.) (Robman and Taylor, 2005; Shiels and Hejtmancik, 2017). Recent studies have identified small molecules that can reverse some forms of cataract that arise due to protein aggregation (Makley et al., 2015; Quinlan, 2015; Zhao et al., 2015), though their efficacy in individuals with cataracts remains to be demonstrated. Although protein aggregates have been identified in cataractous human lenses, the common and unique molecular events initiated by different cataract risk factors are currently unclear. Moreover, other particles that appear distinct from protein aggregates have been identified in human lenses that may account for the light-scattering associated with some forms of cataract (Costello et al., 2010; Gilliland et al., 2001). Thus the micro-lens system holds potential for elucidating cataract mechanisms resulting from individual risk factors, and for identifying additional candidate anti-cataract therapeutic targets.

Summary

The cellular, molecular and functional features of human ROR1⁺ LECs and micro-lenses suggest they share sufficient similarities with human LECs and lenses to provide a useful *in vitro* tool with which to investigate lens and cataract development. In addition, the simplicity, scalability and defined nature of these systems represent

a significant advance over existing hPSC-based approaches. The ROR1⁺ LECs and micro-lenses will enable: functional genomic studies with relevance to developmental biology; investigation of PCO and a wide variety of primary cataract risk factors; clinical toxicity assays; as well as targeted and/or high-throughput anti-cataract drug screening. The capacity to generate disease-specific hPSCs suggests the micro-lenses are also a likely platform for investigating a wide range of poorly understood whole-body syndromes that include cataract as a symptom. For all of these studies the micro-lenses provide a large-scale, predictable, robust and highly purified human system with two reliable and fundamentally appropriate functional assays: the ability to quantify effects on lens transparency and on focusing ability.

MATERIALS AND METHODS

Pluripotent cell culture

Human pluripotent cells were obtained as follows: embryonic stem cells were provided by A. Nagy (CA1 line) (International Stem Cell Initiative et al., 2007) and the StemCore facility (MEL1 line), University of Queensland, Australia; induced pluripotent stem cells hiPSC-TT and hiPSC-LacZ were obtained from E. Stanley and A. Elefanti, Murdoch Children's Research Institute (Melbourne, Australia). Approval for use of these cells was obtained from the Western Sydney University Human Research Ethics Committee (Australia). Pluripotent cells were cultured in mTeSR1 (StemCell Technologies) on plates coated with Matrigel (Corning), and passaged as clumps using 1 mg/ml dispase as previously described (O'Connor et al., 2008a). For differentiation experiments, pluripotent cells were plated as single cells on Matrigel-coated dishes and cultured in mTeSR1 until confluent, after which the cells were exposed to the stage 1 lens differentiation medium.

Lens differentiation and ROR1⁺ cell separation/culture

A three-stage differentiation protocol was used to generate heterogeneous cultures containing lens cells (Yang et al., 2010). Growth factors were sourced from Miltenyi Biotec and Peprotech, and the base medium for each stage was DMEM:F12 (Thermo Fisher Scientific). Initial modification of this protocol involved increasing the concentration of noggin to 500 ng/ml and including 10 nM SB431542 in the stage 1 medium, followed by reducing the concentration of FGF2 to 10 ng/ml in stage 3. For purification of ROR1⁺ cells via magnetic cell separation, single-cell suspensions were obtained using TrypLE (Thermo Fisher Scientific) during stage 2 of the lens differentiation protocol. The cells were then incubated with a biotinylated anti-human ROR1 antibody (BioScientific; AF2000) and labelled cells purified using anti-biotin microbeads and an autoMACS cell separator (Miltenyi Biotec). Purified ROR1⁺ cells were plated on Matrigel-coated dishes in M199 medium (Thermo Fisher Scientific) containing 10 ng/ml of FGF2 or in test media (TM) consisting of M199 and combinations of the following growth factors: BMP4/BMP7 (20 ng/ml each); EGF/TGF α (5 ng/ml each); HGF (10 ng/ml); IGF1/insulin (10 ng/ml and 10 μ g/ml); and PDGF-AA (10 ng/ml). All four human pluripotent stem cell lines (two embryonic and two induced pluripotent) tested behaved similarly.

Micro-lens formation, culture and focal point image analysis

ROR1⁺ cells were aggregated via centrifugation at 300 *g* for 5 min using AggreWell plates (StemCell Technologies; 1200 micro-wells per 24-well plate well). Aggregates were cultured in the plates for 1 to 2 days before being collected and embedded in 0.25% agarose (Amresco) in M199, and then cultured for up to 42 days in stage 3 medium described above. At various times during culture, the developing micro-lenses were assessed for light transmission and focusing ability via light microscopy, using an adaptation of our published method for assessment of *in vitro*-regenerated rat lenses (O'Connor and McAvoy, 2007; O'Connor et al., 2008b). Briefly, using an inverted microscope, individual micro-lenses were brought into focus and an image taken. The objective lens was then lowered until a focal point was reached, at which location another image was taken and the distance travelled to this point recorded. The objective lens was lowered

again an equivalent distance and a third image taken. The objective lens was then raised and two more images taken, half-way between the first and second images and the other half-way between the second and third images. Measures of transmitted light and focal points were obtained by quantifying the grey-level approximately within the central quarter diameter of each micro-lens using ImageJ. Measurements were compared using the Student's *t*-test and are shown as mean \pm s.e.m.

Flow cytometry

Single cell suspensions of differentiated cells were obtained using TrypLE and stained as previously described (O'Connor et al., 2008a,b; Ungrin et al., 2007). Primary antibodies used included anti-human ROR1 (BioScientific) and anti-human CRYAB (ENZO Life Sciences; ADI-SPA-223); the secondary antibody used was an AlexaFluor-488 anti-IgG antibody (Thermo Fisher Scientific; A11001). Labelled cells were analysed using a MACSquant cell analyser (Miltenyi Biotec) and data analysed using the Student's *t*-test (mean \pm s.e.m.).

RNA-seq, PCR and *in situ* hybridisation

RNA from ROR1⁺ cells was collected immediately after cell separation and purified using an Isolate II RNA purification kit (Bioline) as per the manufacturer's instructions. The quality of purified RNA samples was assessed using a Bioanalyser (Agilent) before RNA-seq libraries were prepared using a TruSeq Stranded Total RNA Sample Preparation Kit (Illumina). Samples were sequenced on an Illumina HiSeq 2500 instrument using 2 \times 100 paired end reads, and the data have been deposited in GEO (<http://www.ncbi.nlm.nih.gov/geo/>) under accession number GSE94296. Reads were analysed using FastQC to assess the quality of the data and processed using the Falco framework (Yang et al., 2016), with HISAT2 (Kim et al., 2015) as the aligner and featureCounts (Liao et al., 2014) as the read quantification tool (extra arguments for featureCounts: -s 2 -t exon -g gene_name -primary). The hg19 genome reference and GTF annotation used for building the alignment index and quantification, respectively, were obtained from GENCODE (GRCh37.p13) (Harrow et al., 2012). Gene expression was normalised as reads per million reads (RPM) per sample. For principal component analysis of cell-type specific genes, the ROR1⁺ RNA-seq data was compared against a set of published gene expression data that consisted of adult human LECs and lens fibre cells (Hawse et al., 2005), as well as various ENCODE samples, including foetal lens tissue, pluripotent stem cells, and other foetal and adult tissues. To minimise the batch effect when comparing gene expression profiles from different data sets, expression values were discretised to a value of 1 for the 1000 most highly expressed genes in each cell type, or 0 otherwise. Genes expressed in >20% of the datasets were removed before computing a pairwise dissimilarity matrix using the binary distance function in R. The dissimilarity matrix was then used as the input to the principal component analysis (PCoA) function implemented in an R package CIDR (Lin et al., 2017). Results were visualised using the first two principal components (PC1 and PC2). For gene-set analysis against published gene expression data, a large compendium of tissue-specific gene expression was generated consisting of 144 human and 94 mouse ENCODE datasets, as well as published lens transcriptomic data sets for human (Hawse et al., 2005) and mouse (Hoang et al., 2014; Khan et al., 2015; Lachke et al., 2012) (i.e. 145 human and 95 mouse cell/tissue types total). Where more than one replicate was available for any tissue or cell type, the mean expression value for each gene was calculated. To ensure a consistent gene universe across this large compendium, non-uniquously mapped gene symbols were removed. To generate a compendium of cell/tissue-specific marker genes, the top 3000 highest expressed genes in each tissue were filtered to only keep genes that were highly expressed in 5% or fewer tissue/cell types. The highest-expressed ROR1⁺ cell-specific genes were determined using the same approach, and then compared against the compendium of cell/tissue-specific marker gene set using Fisher's exact tests and Benjamini-Hochberg correction on the resulting *P*-values (Djordjevic et al., 2016). For semi-quantitative real-time PCR analysis, cDNA was synthesised from >100 ng purified RNA using Bioscript (Bioline) and a Mastercycler (Stratagene). Semi-quantitative real-time PCR was performed using Go-Flexi Taq and SYBR Green (Bioline) and an MX3005P real-time PCR machine (Agilent

Technologies); PCR primer sequences used are shown in Extended Data Fig. 1. Data were analysed using Student's *t*-test (mean \pm s.e.m.). *In situ* hybridisation analysis was performed using the mouse embryo *in situ* hybridisation resource at www.genepaint.org.

Teratoma assay

All experiments involving animals were approved by the Animal Research Ethics Committee at Western Sydney University. Assessment of the teratoma-forming ability of purified, ROR1⁺ cells was undertaken as previously described (O'Connor et al., 2011). Grafts containing single cell suspensions of 10⁶ cultured ROR1⁺ cells were transplanted in 100 μ l of \sim 10 mg/ml Matrigel under the back flank of 12-week-old NOD/SCID mice; grafts were randomly assigned amongst littermates, with the number of mice used minimised by transplanting multiple grafts in each animal. Control grafts contained, in addition to the ROR1⁺ cells, single-cell suspensions of up to 5 \times 10⁵ undifferentiated pluripotent cells. Mice were housed for up to 12 weeks post-transplantation, euthanized with CO₂, and grafts were fixed in 10% neutral buffered formalin and assessed without blinding.

Immunofluorescence staining

Cultured micro-lenses were fixed at room temperature without removal from the surrounding agarose which was \sim 2 mm thick. Fixation was performed with 10% neutral buffered formalin for 1 h for laminin and collagen IV detection (though section shrinkage was noted) or 24 h for crystallin detection (leakage of crystallins into the surrounding agarose was noted at lower fixation times). The fixed agarose samples containing micro-lenses were washed three times with phosphate-buffered saline. Samples were dehydrated in a Microm STP-120 Tissue Processor (Thermo Fisher Scientific) in 50%, 70% and 80% ethanol (each for 60 min), followed by 90% and 100% ethanol (each 2 \times 90 min), xylene (3 \times 90 min) and paraffin at 60°C (1 \times 60 min and 1 \times 90). Samples were embedded in paraffin and 5 μ m sections cut. Immunofluorescent staining was performed as previously described (O'Connor and McAvoy, 2007; O'Connor et al., 2008b) using the following anti-human primary antibodies at \sim 4 μ g/ml (Santa Cruz Biotechnology): anti-CRYAA (sc-22743); anti- β -crystallin (sc-22745); and anti- γ -crystallin (sc-22746). Control primary antibody staining was performed using rabbit IgG (Innovative Research; 121266101). Secondary antibody staining was performed using an Alexafluor-488 anti-rabbit IgG antibody (Thermo Fisher Scientific; A11078). Nuclei were counterstained with 1 μ g/ml Hoechst or DAPI (Thermo Fisher Scientific). Images were photographed using a CKX-41 microscope (Olympus) and digital camera with QCapture 6 software (QImaging); images are shown with no digital manipulation.

Mass spectrometry

For detection of only the most-abundant proteins, cultured ROR1⁺ cells or whole micro-lenses were collected directly from culture in 15 μ l of 0.5% RapiGest SF (Waters) in 50 mM NH₄HCO₃. Samples were homogenised on ice for 5 min before reduction with 100 μ l of 5 mM dithiothreitol (Cabriochem) in 50 mM NH₄HCO₃ for 1 h at 60°C, then alkylated with 100 μ l of 15 mM iodoacetamide (Merck) in 50 mM NH₄HCO₃ for 1 h at room temperature. Samples were proteolysed overnight at 37°C with 10 ng/ml of trypsin (Promega) in 75 mM NH₄HCO₃. Peptides were purified by solid phase extraction using Waters Oasis HLB cartridges (30 mg, 1 ml). Pre-cleaning with 1 ml acetonitrile (ACN) was followed by conditioning with 1 ml 0.1% trifluoroacetic acid (TFA). Samples were acidified with 250 μ l of 0.4% aqueous TFA prior to loading. Samples were washed consecutively with 1 ml of 0.1% TFA to remove salts, 1 ml of ultrapure H₂O to remove aqueous soluble contaminants and TFA, then peptides were eluted into low-binding microcentrifuge tubes using 500 μ l of 70% aqueous ACN. Solvents were removed using rotational vacuum concentrator for 2–3 h. Dried peptide samples were treated with 15 μ l of 0.1% aqueous formic acid and rested for 30 min. Samples were triturated prior to centrifugation at 14,000 *g* for 10 min. Supernatants containing peptides were analysed by LC-MS/MS using a nanoAcquity UPLC and Xevo QToF mass spectrometer (Waters); 3 μ l of sample were loaded at 3 μ l/min onto a C18 Symmetry trapping column of dimensions 180 μ m \times 0 mm (Waters) and desalted at this flow rate for 5 min using 1% ACN in water with 0.1% formic acid. Peptides

were washed off the trap at 400 nl/min onto a C18 BEH analytical column (Waters) packed with 1.7 µm particles of pore size 13 nm of dimensions 100 µm×100 mm, using a ramped method from 1% to 85% ACN (with 0.1% formic acid) over 37 min. Eluting peptides were identified by MS/MS using a Xevo QToF mass spectrometer (Waters), fitted with a nanospray source with an emitter tip tapered to 10 µm at 2300 V in positive ion mode. Data-dependent acquisition was performed with continuous scanning for 2⁺ to 4⁺ charged peptides, an intensity of >50 counts and a maximum of three ions in any given 3 s scan (precursor peptides were excluded for 30 s). The MS/MS data files were analysed using Mascot Daemon and queried against the SwissProt database using Homo sapiens-specific searches. Variable modifications of carbamidomethyl (C), deamidated (NQ), oxidation (M) and propionamide (C) were used with peptide and MS/MS mass tolerances of 0.05 Da. Only peptide hits with *P*<0.05 were reported. Peptides identified by Mascot were further validated by manual inspection of the MS/MS spectra for the peptide to ensure the b- and y-ion series are sufficiently extensive for an accurate identification. Percolator-based decoy searches (Käll et al., 2007) were also performed on the samples, and these revealed false discovery rates of 0%.

Electron microscopy

Micro-lenses were fixed for 1 h at room temperature in 2.5% glutaraldehyde in 0.1 M phosphate buffer, pH 7.4. A 3 mm biopsy punch and tweezers were used to isolate micro-lenses from the agarose gel. Samples were then fixed in 2.5% glutaraldehyde for a further 48–72 h at 4°C, after which they were washed four times at hourly intervals with 5 ml phosphate buffer at 4°C. Samples were then transferred to 0.1 M sodium cacodylate buffer (pH 7.4) for 2 h, post-fixed in 2% OsO₄ for 4 h then rinsed in cacodylate buffer. Samples were then stained with 1% tannic acid for 30 min at room temp and rinsed in cacodylate buffer. This was followed by rinsing in 2% sodium acetate, en bloc staining with 2% uranyl acetate for 1 h, dehydration in a series of graded alcohols and dry acetone, infiltration with Spurr's resin diluted in acetone, and polymerisation in 100% standard hardness Spurr's resin at 70°C. The embedded micro-lenses were sectioned at 90 nm using a Powertome ultramicrotome (RMC Boeckeler) and imaged with a Morgagni 268D transmission electron microscope (FEI) at 80 kV.

Acknowledgements

We thank E. Stanley and A. Elefanty for providing the human induced pluripotent stem cells, and J. M. Polo for critical review of the manuscript.

Competing interests

The authors declare no competing or financial interests.

Author contributions

Conceptualization: M.D.O.; Methodology: P.M., M.H.K., A.Y., D.D., M.C.K., J.W.K.H., D.G.H., M.D.O.; Software: M.H.K., A.Y., D.D., J.W.K.H.; Validation: P.M., M.H.K., D.G.H., M.D.O.; Formal analysis: P.M., M.H.K., T.S., M.E.M., C.U.D., S.L., M.D.O.; Investigation: P.M., M.H.K., T.S., M.E.M., C.U.D., S.L., M.C.K., J.W.K.H., D.G.H., M.D.O.; Resources: M.D.O.; Data curation: J.W.K.H.; Writing - original draft: P.M., M.D.O.; Writing - review & editing: P.M., M.H.K., T.S., M.E.M., C.U.D., S.L., A.Y., D.D., M.C.K., J.W.K.H., D.G.H., M.D.O.; Visualization: P.M., M.H.K., T.S., M.E.M., C.U.D., S.L., J.W.K.H., M.D.O.; Supervision: M.D.O.; Project administration: M.D.O.; Funding acquisition: M.D.O.

Funding

This work was supported by The Medical Advances Without Animals Trust (MAWA) and the Rebecca L. Cooper Medical Research Foundation. J.W.K.H. is supported by a Career Development Fellowship from the National Health and Medical Research Council and the National Heart Foundation of Australia. Deposited in PMC for immediate release.

Data availability

The RNA-seq data have been deposited in GEO under accession number GSE94296.

Supplementary information

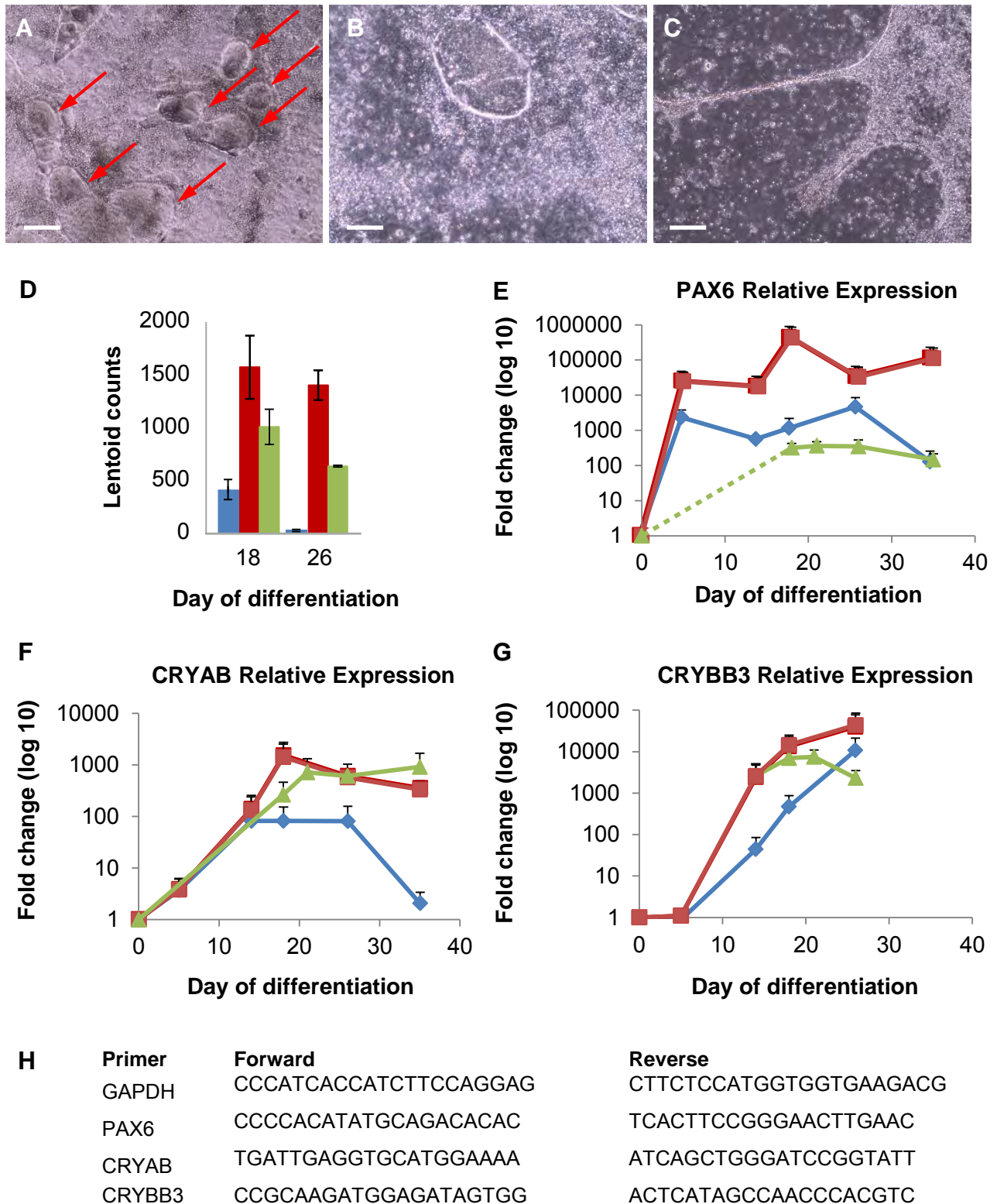
Supplementary information available online at <http://dev.biologists.org/lookup/doi/10.1242/dev.155838.supplemental>

References

- Al-Ghoul, K. J. and Costello, M. J. (1997). Light microscopic variation of fiber cell size, shape and ordering in the equatorial plane of bovine and human lenses. *Mol. Vis.* **3**, 2.
- Anand, D. and Lachke, S. A. (2017). Systems biology of lens development: a paradigm for disease gene discovery in the eye. *Exp. Eye Res.* **156**, 22–33.
- Chen, W., Huang, J., Yu, X., Lin, X. and Dai, Y. (2015). Generation of induced pluripotent stem cells from renal tubular cells of a patient with Alport syndrome. *Int. J. Nephrol. Renovasc. Dis.* **8**, 101–109.
- Costello, M. J., Johnsen, S., Metlapally, S., Gilliland, K. O., Frame, L. and Balasubramanian, D. (2010). Multilamellar spherical particles as potential sources of excessive light scattering in human age-related nuclear cataracts. *Exp. Eye Res.* **91**, 881–889.
- Cvekl, A. and Zhang, X. (2017). Signaling and gene regulatory networks in mammalian lens development. *Trends Genet.* **33**, 677–702.
- Davies, J. C., Cunningham, S., Harris, W. T., Lapey, A., Regelman, W. E., Sawicki, G. S., Southern, K. W., Robertson, S., Green, Y., Cooke, J. et al. (2016). Safety, pharmacokinetics, and pharmacodynamics of ivacaftor in patients aged 2–5 years with cystic fibrosis and a CFTR gating mutation (KIWI): an open-label, single-arm study. *Lancet Respir. Med.* **4**, 107–115.
- Djordjevic, D., Kusumi, K. and Ho, J. W. (2016). Xgsa: a statistical method for cross-species gene set analysis. *Bioinformatics* **32**, 1620–1628.
- Dryden, C., Wilkinson, J., Young, D., Brooker, R. J. and Scottish Paediatric Cystic Fibrosis Managed Clinical Network (Spcfmcn). (2016). The impact of 12 months treatment with ivacaftor on Scottish paediatric patients with cystic fibrosis with the G551d mutation: a review. *Arch. Dis. Child.* **103**, 68–70.
- Fu, Q., Qin, Z., Jin, X., Zhang, L., Chen, Z., He, J., Ji, J. and Yao, K. (2017). Generation of functional lentoid bodies from human induced pluripotent stem cells derived from urinary cells. *Invest. Ophthalmol. Vis. Sci.* **58**, 517–527.
- Gilliland, K. O., Freel, C. D., Lane, C. W., Fowler, W. C. and Costello, M. J. (2001). Multilamellar bodies as potential scattering particles in human age-related nuclear cataracts. *Mol. Vis.* **7**, 120–130.
- Greiling, T. M. S. and Clark, J. I. (2012). New insights into the mechanism of lens development using zebra fish. *Int. Rev. Cell. Mol. Biol.* **296**, 1–61.
- Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S. et al. (2012). Gencode: the reference human genome annotation for the encode project. *Genome Res.* **22**, 1760–1774.
- Hawse, J. R., Deamicis-Tress, C., Cowell, T. L. and Kantorow, M. (2005). Identification of global gene expression differences between human lens epithelial and cortical fiber cells reveals specific genes and their associated pathways important for specialized lens cell functions. *Mol. Vis.* **11**, 274–283.
- Hoang, T. V., Kumar, P. K., Sutharzan, S., Tsonis, P. A., Liang, C. and Robinson, M. L. (2014). Comparative transcriptome analysis of epithelial and fiber cells in newborn mouse lenses with RNA sequencing. *Mol. Vis.* **20**, 1491–1517.
- International Stem Cell Initiative, Adewumi, O., Aflatoonian, B., Ahrlund-Richter, L., Amit, M., Andrews, P. W., Beighton, G., Bello, P. A., Benvenisty, N., Berry, L. S., Bevan, S. et al. (2007). Characterization of human embryonic stem cell lines by the international stem cell initiative. *Nat. Biotechnol.* **25**, 803–816.
- Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. and Maccoss, M. J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**, 923–925.
- Khan, S. Y., Hackett, S. F., Lee, M. C., Pourmand, N., Talbot, C. C., Jr. and Riazuddin, S. A. (2015). Transcriptome profiling of developing murine lens through RNA sequencing. *Invest. Ophthalmol. Vis. Sci.* **56**, 4919–4926.
- Kim, D., Langmead, B. and Salzberg, S. L. (2015). Hisat: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360.
- Kuwabara, T. and Imaizumi, M. (1974). Denucleation process of the lens. *Invest. Ophthalmol.* **13**, 973–981.
- Lachke, S. A., Ho, J. W. K., Kryukov, G. V., O'Connell, D. J., Aboukhalil, A., Bulky, M. L., Park, P. J. and Maas, R. L. (2012). Isyte: integrated systems tool for eye gene discovery. *Invest. Ophthalmol. Vis. Sci.* **53**, 1617–1627.
- Li, D., Qiu, X., Yang, J., Liu, T., Luo, Y. and Lu, Y. (2016). Generation of human lens epithelial-like cells from patient-specific induced pluripotent stem cells. *J. Cell. Physiol.* **231**, 2555–2562.
- Liao, Y., Smyth, G. K. and Shi, W. (2014). Featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930.
- Lin, P., Troup, M., Ho, J. W. K. (2017). CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.* **18**, 59.
- Lovicu, F. J., Mcavoy, J. W. and de longh, R. U. (2011). Understanding the role of growth factors in embryonic development: insights from the lens. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **366**, 1204–1218.
- Makley, L. N., Mcmenimen, K. A., Devree, B. T., Goldman, J. W., Mcglasson, B. N., Rajagopal, P., Dunyak, B. M., Mcquade, T. J., Thompson, A. D., Sunahara, R. et al. (2015). Pharmacological chaperone for alpha-crystallin partially restores transparency in cataract models. *Science* **350**, 674–677.
- Mann, I. (1964). *The Development Of The Human Eye*. New York: Grune & Stratton, Incorporated.

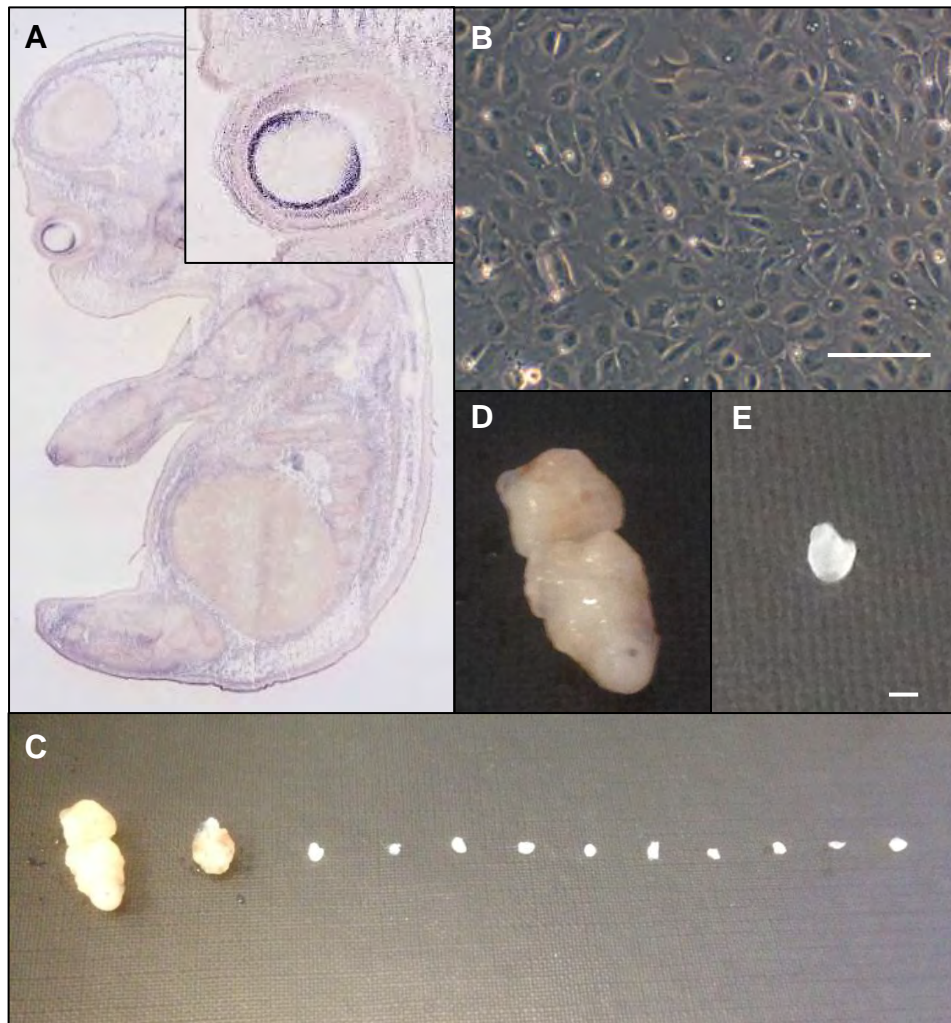
- McColley, S. A.** (2016). A safety evaluation of ivacaftor for the treatment of cystic fibrosis. *Expert Opin. Drug Saf.* **15**, 709-715.
- Mengarelli, I. and Barberi, T.** (2013). Derivation of multiple cranial tissues and isolation of lens epithelium-like cells from human embryonic stem cells. *Stem Cells Transl. Med.* **2**, 94-106.
- O'Connor, M. D. and McAvoy, J. W.** (2007). In vitro generation of functional lens-like structures with relevance to age-related nuclear cataract. *Invest. Ophthalmol. Vis. Sci.* **48**, 1245-1252.
- O'Connor, M. D., Kardel, M. D., Iosfina, I., Youssef, D., Lu, M., Li, M. M., Vercauteren, S., Nagy, A. and Eaves, C. J.** (2008a). Alkaline phosphatase-positive colony formation is a sensitive, specific, and quantitative indicator of undifferentiated human embryonic stem cells. *Stem Cells* **26**, 1109-1116.
- O'Connor, M. D., Wederell, E. D., de longh, R., Lovicu, F. J. and McAvoy, J. W.** (2008b). Generation of transparency and cellular organization in lens explants. *Exp. Eye Res.* **86**, 734-745.
- O'Connor, M. D., Kardel, M. D. and Eaves, C. J.** (2011). Functional assays for human embryonic stem cell pluripotency. *Methods Mol. Biol.* **690**, 67-80.
- Quinlan, R. A.** (2015). Drug discovery. A new dawn for cataracts. *Science* **350**, 636-637.
- Ringens, P., Mungyer, G., Jap, P., Ramaekers, F., Hoenders, H. and Bloemendal, H.** (1982). Human lens epithelium in tissue culture: biochemical and morphological aspects. *Exp. Eye Res.* **35**, 313-324.
- Robman, L. and Taylor, H.** (2005). External factors in the development of cataract. *Eye (Lond)* **19**, 1074-1082.
- Shestopalov, V. I. and Bassnett, S.** (2000). Three-dimensional organization of primary lens fiber cells. *Invest. Ophthalmol. Vis. Sci.* **41**, 859-863.
- Shiels, A. and Hejtmancik, J. F.** (2017). Mutations and mechanisms in congenital and age-related cataracts. *Exp. Eye Res.* **156**, 95-102.
- Song, B., Niclis, J. C., Alikhan, M. A., Sakkal, S., Sylvain, A., Kerr, P. G., Laslett, A. L., Bernard, C. A. and Ricardo, S. D.** (2011). Generation of induced pluripotent stem cells from human kidney mesangial cells. *J. Am. Soc. Nephrol.* **22**, 1213-1220.
- Talamo Guevara, M. and McColley, S. A.** (2017). The safety of lumacaftor and ivacaftor for the treatment of cystic fibrosis. *Expert Opin. Drug Saf.* **16**, 1305-1311.
- Taylor, V. L., Al-Ghoul, K. J., Lane, C. W., Davis, V. A., Kuszak, J. R. and Costello, M. J.** (1996). Morphology of the normal human lens. *Invest. Ophthalmol. Vis. Sci.* **37**, 1396-1410.
- Tholozan, F. M. and Quinlan, R. A.** (2007). Lens cells: more than meets the eye. *Int. J. Biochem. Cell Biol.* **39**, 1754-1759.
- Thomson, J. A. and Augusteyn, R. C.** (1985). Ontogeny of human lens crystallins. *Exp. Eye Res.* **40**, 393-410.
- Ungrin, M., O'Connor, M. D., Eaves, C. J. and Zandstra, P. W.** (2007). Phenotypic analysis of human embryonic stem cells. *Curr. Protoc. Stem Cell Biol.* **2**, 3.1-3.25.
- Van Goor, F., Hadida, S., Grootenhuys, P. D., Burton, B., Cao, D., Neuberger, T., Turnbull, A., Singh, A., Joubbran, J., Hazlewood, A. et al.** (2009). Rescue of Cf airway epithelial cell function in vitro by a Cftr potentiator, Vx-770. *Proc. Natl. Acad. Sci. USA* **106**, 18825-18830.
- Vrensen, G. F., Graw, J. and De Wolf, A.** (1991). Nuclear breakdown during terminal differentiation of primary lens fibres in mice: a transmission electron microscopic study. *Exp. Eye Res.* **52**, 647-659.
- Walker, J. and Menko, A. S.** (2009). Integrins in lens development and disease. *Exp. Eye Res.* **88**, 216-225.
- Wederell, E. D. and de longh, R. U.** (2006). Extracellular matrix and integrin signaling in lens development and cataract. *Semin. Cell Dev. Biol.* **17**, 759-776.
- Wormstone, I. M. and Eldred, J. A.** (2016). Experimental models for posterior capsule opacification research. *Exp. Eye Res.* **142**, 2-12.
- Wu, W., Tholozan, F. M., Goldberg, M. W., Bowen, L., Wu, J. and Quinlan, R. A.** (2014). A gradient of matrix-bound Fgf-2 and perlecan is available to lens epithelial cells. *Exp. Eye Res.* **120**, 10-14.
- Yang, C., Yang, Y., Brennan, L., Bouhassira, E. E., Kantorow, M. and Cvekl, A.** (2010). Efficient generation of lens progenitor cells and lentoid bodies from human embryonic stem cells in chemically defined conditions. *FASEB J.* **24**, 3274-3283.
- Yang, A., Troup, M., Lin, P. and Ho, J. W.** (2016). Falco: a quick and flexible single-cell Rna-seq processing framework on the cloud. *Bioinformatics* **33**, 767-769.
- Zhao, L., Chen, X. J., Zhu, J., Xi, Y. B., Yang, X., Hu, L. D., Ouyang, H., Patel, S. H., Jin, X., Lin, D. et al.** (2015). Lanosterol reverses protein aggregation in cataracts. *Nature* **523**, 607-611.

Supplementary material Fig. S1



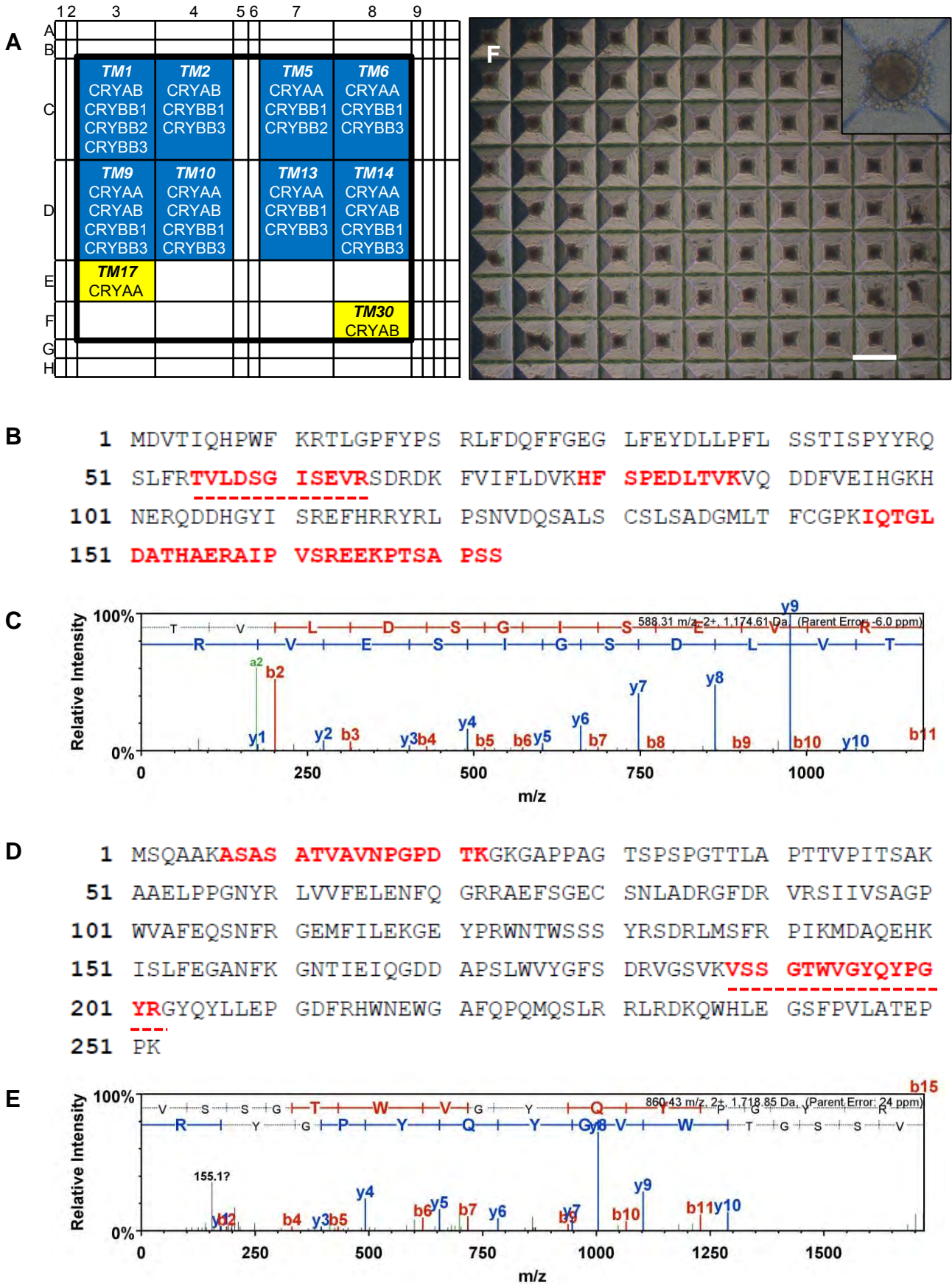
Supplementary Figure S1. Modification of the 3-stage lens differentiation protocol increased lentoids and lens gene expression but did not produce pure LEC populations. (A-C) Images of lentoids (red arrows) generated after modification of the published 3-stage lens differentiation protocol (Yang et al., 2010). Increasing the concentration of Noggin from 100 to 500 ng/mL and including 10 nM SB431542 (Activin/BMP/TGF- β pathway inhibitor) in the Stage 1 medium produced large numbers of lentoids (A). Some lentoids had light-refractive borders (B), however they did not focus light. As described in the published protocol, the lentoids detached from the culture surface and were lost when changing the culture medium (C). Scale bars, 500 μ m, 100 μ m and 500 μ m, respectively. (D-G) Time-course comparison of lentoid production and lens-related gene expression between the published 3-stage lens differentiation protocol (blue) and two modified protocols (red, green; n = 3). Increasing the concentration of Noggin to 500 ng/mL and including 10 nM SB431542 in the Stage 1 medium (red) increased: the number of lentoids produced per 35mm dish and the time they were retained in the culture (D); PAX6, CRYAB and CRYBB3 expression as detected by semi-quantitative real-time PCR (E-G). Subsequently reducing the concentration of FGF2 in Stage 3 from 100 ng/mL to 10 ng/mL (green) in an attempt to maintain LECs decreased lentoid production (D) as well as PAX6 and CRYBB3 mRNA expression, while maintaining CRYAB expression (E-G: green). Despite this, heterogeneous morphologies and random lentoid production still occurred (A-C). (H) PCR primers used for analysis of differentiating cells. The data shown in A-G were obtained from 3 independent differentiation experiments.

Supplementary material Fig. S2



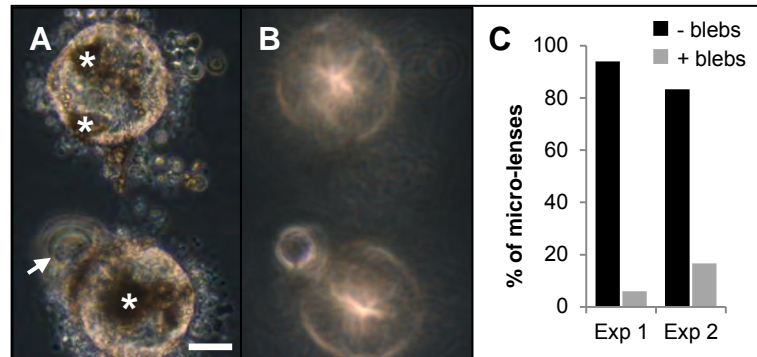
Supplementary Figure S2. ROR1 expression in embryos and ROR1 teratoma assay data. (A) In situ hybridisation data (www.genepaint.org) showing ROR1 transcript expression is predominantly expressed by LECs at embryonic day 14. (B) ROR1⁺ cells plated at high cell densities after purification showed uniform polygonal morphologies (cells shown 2 days after plating). Scale bar: 100 μ m. (C) Twelve grafts, each containing 10^6 ROR1⁺ cells, show that teratoma formation only occurred when undifferentiated pluripotent cells were deliberately included with the ROR1⁺ cells in control grafts (i.e., the 1st and 2nd grafts on the left were seeded with 500,000 and 50,000 disaggregated pluripotent stem cells, respectively, equivalent to \sim 5,000 and 500 colony-forming cells). (D) Higher magnification of the left-hand control graft shown in (C). (E) Higher magnification of a graft that received only ROR1⁺ cells. Control grafts were collected 6 weeks after transplantation; all other grafts were collected 12 weeks after transplantation. Scale bar: 1 mm.

Supplementary material Fig. S3



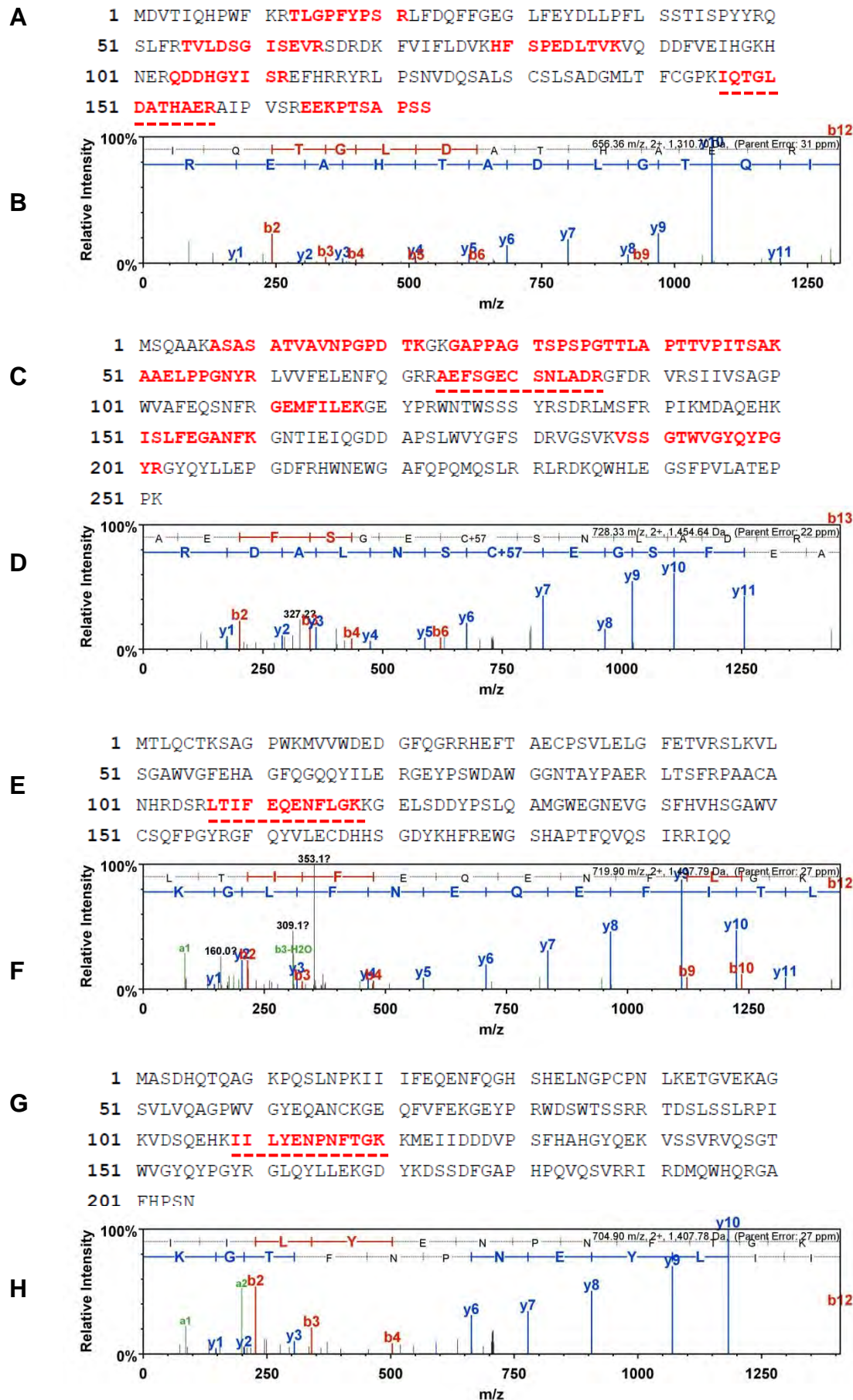
Supplementary Figure S3. Mass spectrometry analysis of ROR1+ cells cultured in TM17 express LEC but not lens fibre cell crystallins. (A) Schematic diagram summarising the crystallin proteins detected in ROR1+ cells cultured in the test media. (B) MS/MS analysis revealed 28% CRYAA protein sequence coverage obtained from ROR1+ cells expanded, frozen, thawed and then cultured for 6 days all in TM17. (C) Raw mass spectrometry data showing identification of the CRYAA peptide underlined in (B). (D) MS/MS analysis revealed 12% CRYBB1 protein sequence coverage obtained from ROR1+ cells expanded, frozen and thawed in TM17 and then cultured for 6 days in Stage 2 lens differentiation medium. (E) Raw mass spectrometry data showing identification of the CRYBB1 peptide underlined in (D). These data are representative of data obtained from 3 independent differentiation experiments. (F) Light microscopy image showing relatively homogenous and large-scale production of ROR1+ cell aggregates. Scale bar: 400 μm .

Supplementary material Fig. S4



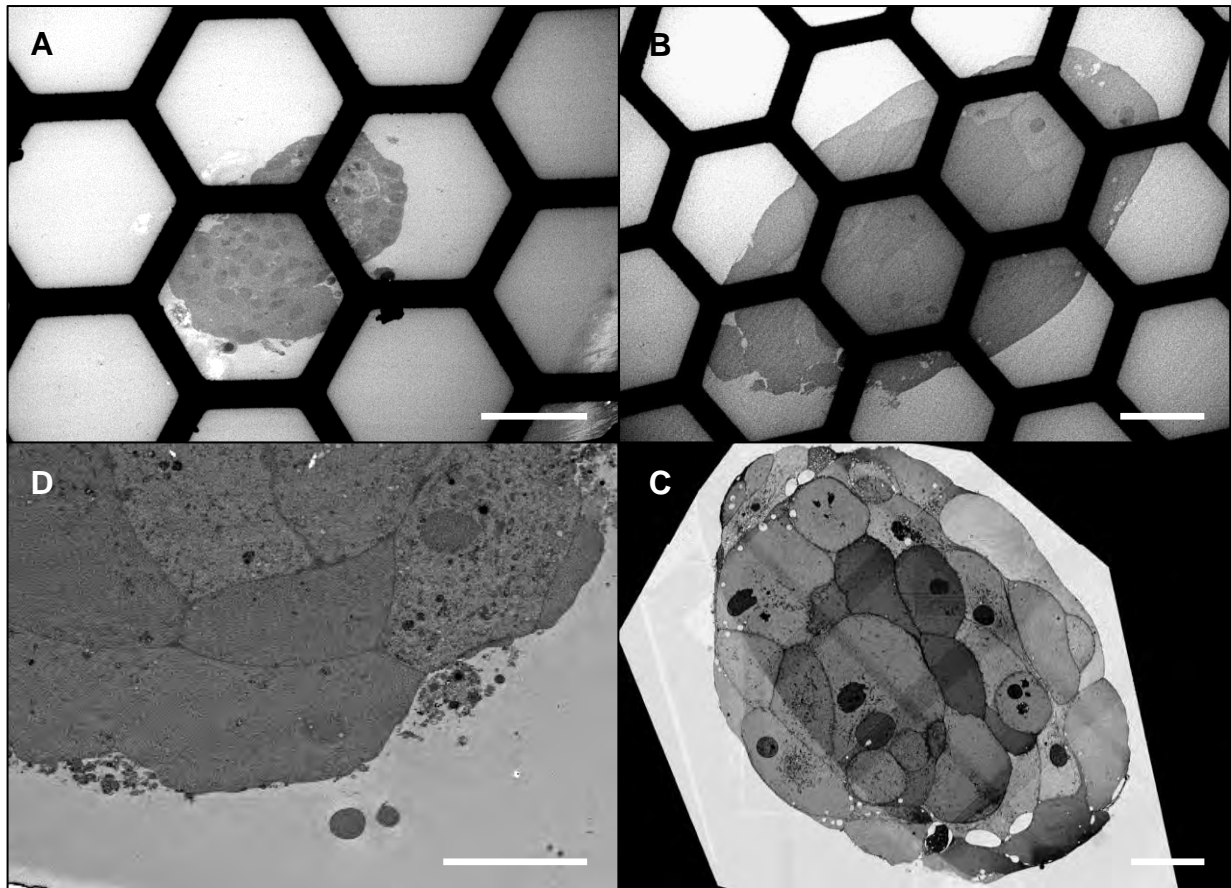
Supplementary Figure S4. Occasional micro-lens features. (A-C) Light microscopy data showing clusters of non-transparent cells (A, asterisks), adjacent to the periphery of two micro-lenses, that did not preclude assessment of light-transmitting regions (A) or focusing ability (B). The presence of 'bleb'-like structures on some aggregates (A, arrow) had little effect on focusing (B) and were typically relatively infrequent (C). Scale bar: 40 μ m.

Supplementary material Fig. S5



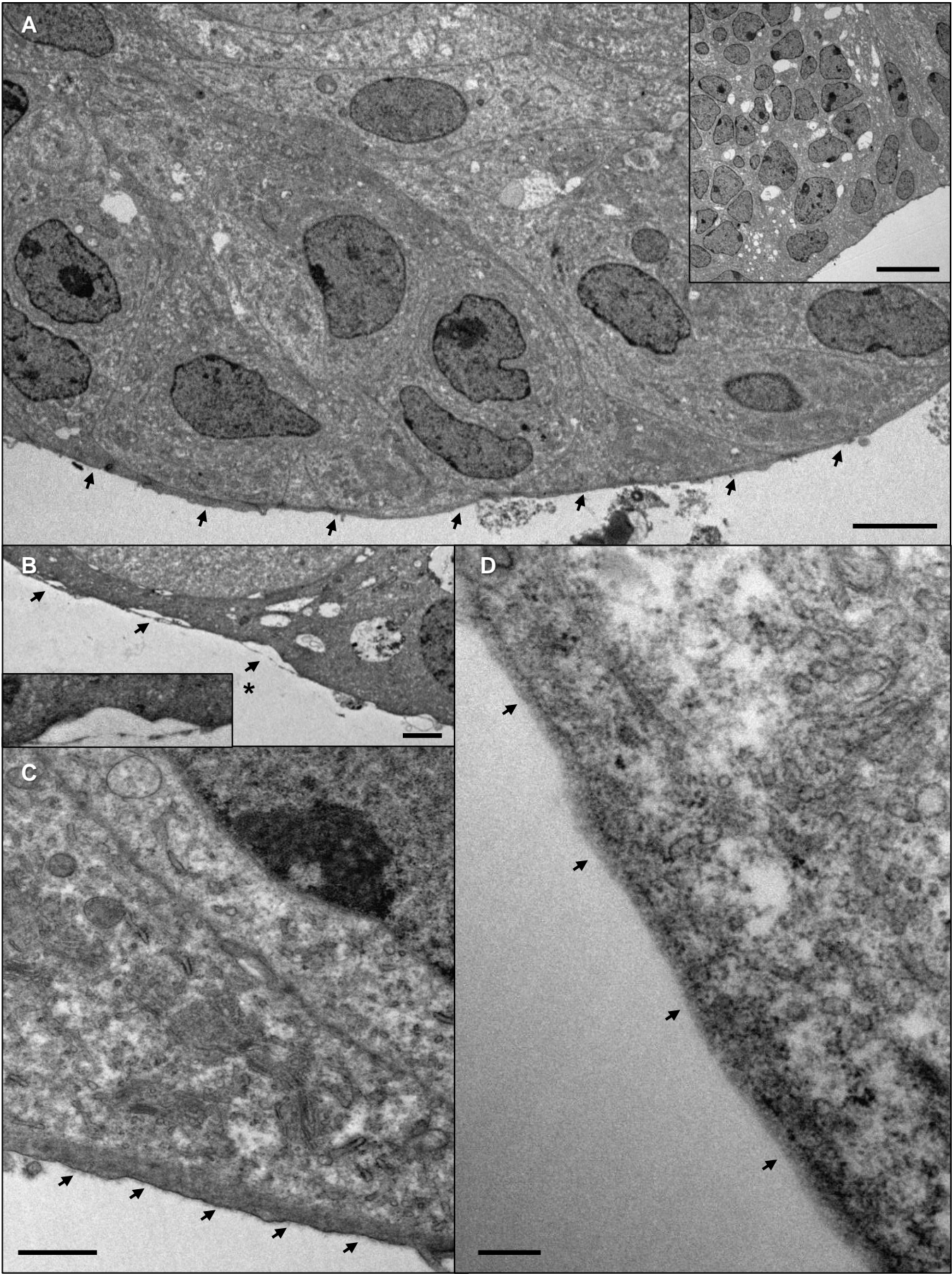
Supplementary Figure S5. Mass spectrometry analysis of LEC and lens fibre cell crystallins in micro-lenses. (A-H) MS/MS analysis of micro-lenses cultured for 27 days in Stage 3 lens differentiation medium revealed 35% protein coverage of CRYAA (A), 38% coverage of CRYBB1 (C), 6% coverage of CRYBA4 (E) and 5% coverage of CRYBB2 (G). Example raw data peptide identifications are shown for the underlined sequences in CRYAA (A, B), CRYBB1 (C, D), CRYBA4 (E, F) and CRYBB2 (G, H). These data, obtained from 10 pooled micro-lenses, are representative of data obtained from 2 independent micro-lens experiments.

Supplementary material Fig. S6



Supplementary material Figure S6. Cellular organisation in ROR1+ cell aggregates. (A-C) Electron microscopy data from an aggregate cultured for 14 days (A) shows the bulk of the tissue consisted of small cells. Scale bar: 50 μm . (B, C) Images of aggregates cultured for 42 days show the bulk of the tissues consist of larger cells with fewer organelles. Scale bars: 50 μm (B) and 10 μm (C, composite image). (D) Lens fibre-like cell adjacent to a LEC-like cell in an aggregate cultured for 24 days, suggestive of rudimentary secondary lens fibre cell differentiation. Scale bar: 10 μm . Images representative of 6 micro-lenses from 2 biological replicates.

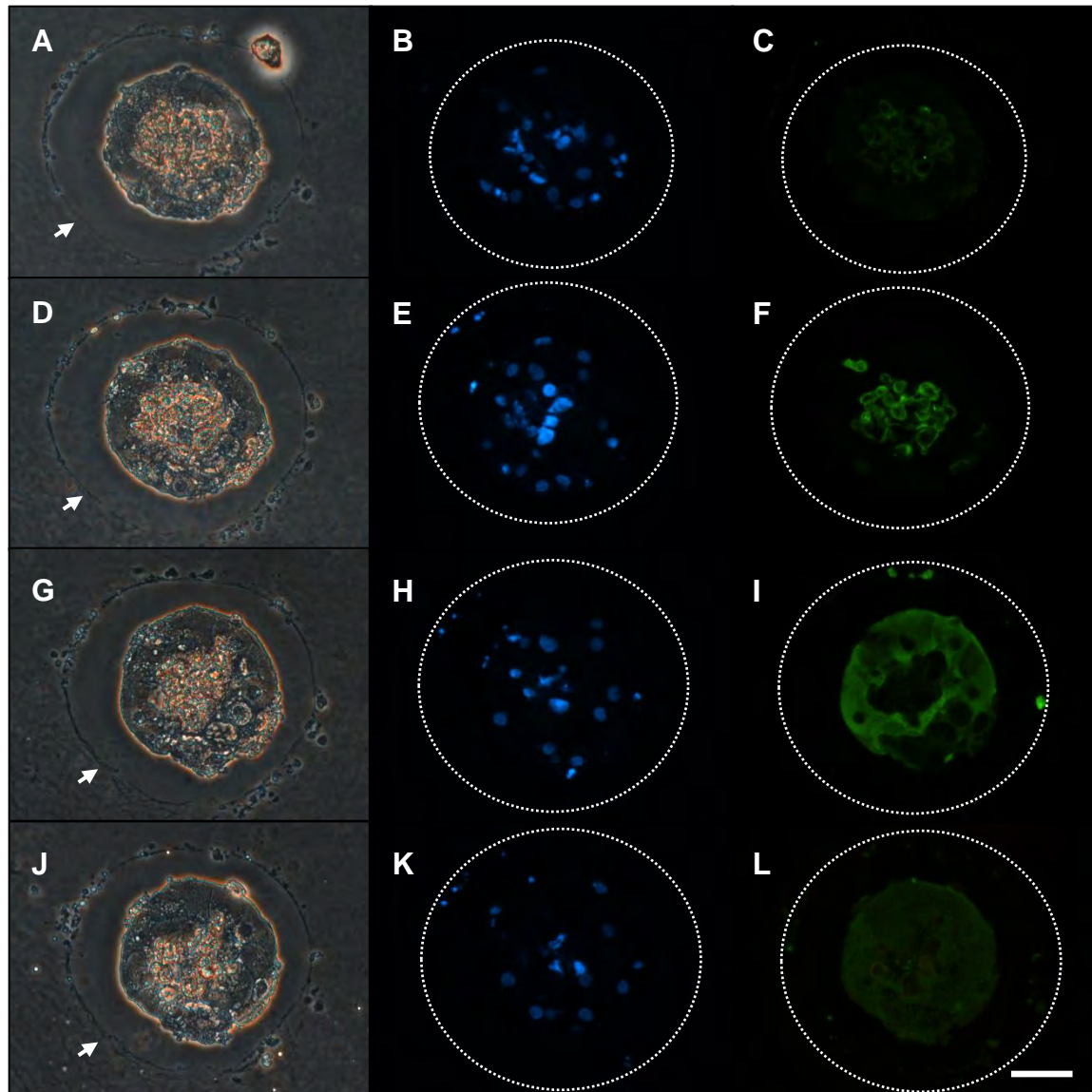
Supplementary material Fig. S7



250 nm

Supplementary material Figure S7. Lens capsule surrounding ROR1+ cell aggregates. (A-C) Electron microscopy data from a large aggregate (diameter >200 μm) cultured for 42 days. (A) Multilayering of peripheral LEC-like cells adjacent to lens capsule-like material (arrows). Scale bar: 5 μm . Inset: a small area of multilayering adjacent to the peripheral LEC-like cells. Scale bar: 10 μm . (B, C) Lens capsule-like material adjacent to lens fibre-like cells. (B, inset: higher magnification of region indicated by asterisk). Scale bars: 1 μm . (D) Lens capsule-like material (indicated by arrows) adjacent to LEC-like cells. Scale bar: 250 nm.

Supplementary material Fig. S8



Supplementary material Figure S8. Lens capsule components expressed by LEC-like cells in ROR1+ cell aggregates. (A, D, G, J) Consecutive peripheral sections of a large aggregate (diameter $\sim 200\ \mu\text{m}$) cultured for 24 days (A is the outermost section). The short fixation time (60 min) permitted detection of laminin and collagen IV (C, F), however, it also resulted in significant shrinkage of the sections that was not observed with longer (24 hour) fixation times. The surrounding agarose used to embed the aggregates during culture is indicated by arrows. (B, E, H, K) DAPI staining of the same sections shown in A, D, G, and J shows the location of nuclei within the fixed micro-lenses. (C, F, I, L) The same sections shown in A, D, G and J after immunofluorescence using anti-laminin (C), anti-collagen IV (F), anti- γ -crystallin (I) and anti- α A-crystallin (L) antibodies. Dotted white lines estimate the original micro-lens boundary. Scale bar: $40\ \mu\text{m}$. Data representative of 7 micro-lenses from 2 biological replicates.

Supplementary Table S1 and S2. Mass spectrometry analysis of ROR1+ cells cultured in TM17 express LEC but not lens fibre cell crystallins. A list of proteins identified from ROR1+ cells expanded, frozen, thawed and then re-cultured for 6 days in TM17 reveals expression of a- but not b-crystallins. (B) A list of proteins identified from ROR1+ cells expanded, frozen and thawed in TM17 and then cultured for 6 days in Stage 2 lens differentiation medium. These data show expression of a variety of lens fibre cell-specific crystallin proteins.

[Click here to Download Table S1-S2](#)

Supplementary Table S3. Mass spectrometry analysis of micro-lenses derived from ROR1+ cells. A list of proteins identified from ROR1+ cell-derived micro-lenses cultured in Stage 3 lens differentiation medium for 27 days shows expression of a-crystallin as well as a variety of lens fibre cell-specific b-crystallin proteins. These data, obtained from 10 pooled micro-lenses, are representative of data obtained from 2 independent micro-lens experiments.

[Click here to Download Table S3](#)

Supplementary Table S4. Expression of capsule components and integrins by ROR1+ cells. A list of the most highly expressed integrin, collagen and laminin mRNA transcripts detected in the ROR1+ RNA-seq libraries.

[Click here to Download Table S4](#)

The C3 method was applied to build two different compendia consisting of large number of cell sample data from human and mouse organisms and then compared the ROR1⁺ cells data with the compendia to characterize it. Without C3 it was very hard to characterize the ROR1⁺ cells as lens epithelial cells because there were almost no or very few previous cell populations like ROR1⁺ cells derived from stem cells. Although the C3 provides an efficient approach to identify the cell type of a gene expression profile a method for identifying active signal pathways for the profile is required to provide more information for clinical applications.

Chapter 4

Identification of active signalling pathways by integrating gene expression and protein interaction data

Signaling pathways are the key biological mechanisms that transduce extracellular signals to affect transcription factor-mediated gene regulation within cells. We have developed a new method – named SPAGI (Signal Pathway Analysis for Gene regulator network Identification) – associated with an R package that can simultaneously predict the set of active signaling pathways in a cell, together with their pathway structure. This is done by integrating protein-protein interaction network and gene expression data. The method was validated using gene expression data sets from a variety of cell types.

RESEARCH

Open Access



Identification of active signaling pathways by integrating gene expression and protein interaction data

Md Humayun Kabir^{1,2,3}, Ralph Patrick^{2,4,5}, Joshua W. K. Ho^{2,4,6*} and Michael D. O'Connor^{1,7*}

From 29th International Conference on Genome Informatics
Yunnan, China. 3-5 December 2018

Abstract

Background: Signaling pathways are the key biological mechanisms that transduce extracellular signals to affect transcription factor mediated gene regulation within cells. A number of computational methods have been developed to identify the topological structure of a specific signaling pathway using protein-protein interaction data, but they are not designed for identifying active signaling pathways in an unbiased manner. On the other hand, there are statistical methods based on gene sets or pathway data that can prioritize likely active signaling pathways, but they do not make full use of active pathway structure that link receptor, kinases and downstream transcription factors.

Results: Here, we present a method to simultaneously predict the set of active signaling pathways, together with their pathway structure, by integrating protein-protein interaction network and gene expression data. We evaluated the capacity for our method to predict active signaling pathways for dental epithelial cells, ocular lens epithelial cells, human pluripotent stem cell-derived lens epithelial cells, and lens fiber cells. This analysis showed our approach could identify all the known active pathways that are associated with tooth formation and lens development.

Conclusions: The results suggest that SPAGI can be a useful approach to identify the potential active signaling pathways given a gene expression profile. Our method is implemented as an open source R package, available via <https://github.com/VCCRI/SPAGI/>.

Keywords: Signaling pathway, Gene expression, Protein-protein interaction, Dental epithelial cells, Lens epithelial cells, Lens fiber cells, Pluripotent stem cells, ROR1⁺ cells

Background

A key role cell signaling (also known as signal transduction) plays within biological systems is to relay extracellular signals in order to regulate intracellular gene expression. The signal transduction process is typically initiated by the binding of a ligand to a membrane-bound receptor, which triggers a cascade of intercellular signaling activities through multiple kinases - ultimately impacting on how transcription factors regulate downstream gene expression [1]. The coordinated activity of different signaling pathways within

and between multiple cell types is the basis of many important biological processes, such as development, tissue repair and immunity [2, 3].

Activation of different signaling pathways can lead to numerous physiological or cellular responses, such as cell proliferation, death, differentiation, and metabolism [4, 5]. Any interruption that occurs within these extra-/intra-cellular communication chains can cause diseases including developmental disorders and cancers [6–9]. Conversely, a clear understanding of the activity of, and interaction between, signaling pathways can help to design rational disease treatment and tissue regeneration strategies [10]. It is therefore important to understand the signaling pathways that are activated in a cell, in

* Correspondence: jwkho@hku.hk; m.oconnor@westernsydney.edu.au

²Victor Chang Cardiac Research Institute, Darlinghurst, NSW, Australia

¹School of Medicine, Western Sydney University, Campbelltown, NSW, Australia

Full list of author information is available at the end of the article



© The Author(s). 2018 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

order to provide a framework for understanding critical pathways affected by disease.

In principle, it should be possible to identify the important signaling pathways of a cell by using gene expression and protein-protein interaction (PPI) data sets. Extensive, publically-available PPI data provide an opportunity to establish a general signaling pathway blueprint, to which cell type-specific gene expression data can be mapped so as to refine the general signaling pathway blueprint into a cell-type specific blueprint. In this way it should be possible to construct a set of cell-type specific active signaling pathways for any cell that summarizes the information flow from a receptor (R) to kinases (Ks), then to transcription factors (TFs).

PPI data is a direct source of information about the structure of signaling pathways [3, 11]. A number of PPI databases are available for human and model organisms such as STRING [12]. A number of bioinformatics methods have been proposed for the reconstruction of known signaling pathways by using PPI data. For example, *CASCADE_SCAN* generates a specific pathway for a list of protein molecules using a steepest descent method. That is, the method takes the input proteins and then finds their interaction partners iteratively based on some evidences (i.e., high scored interactions) [1]. On the other hand, *Pathlinker* reconstructs the known signaling pathways by taking a subnetwork of PPI that consists of the Rs and TFs of interest [13]. The *PathLinker App* is a software tool of the *Pathlinker* method implemented as a *Cytoscape app* [14]. *PathFinder* identifies signaling pathways from a specific R protein to a TF protein in PPI networks by extracting the characteristics of known signal transduction pathways and their functional annotations in the form of association rules [15].

A number of methods use PPI data alone to infer signaling pathway structure. Gitter et al. proposed a method to handle the orientation problem in weighted protein interaction graphs as an optimization problem and present three approximation algorithms based on either weighted Boolean satisfiability solvers or probabilistic assignments [16]. Mei et al. proposed a multi-label multi-instance transfer learning method to simultaneously reconstruct 27 human known signaling pathways, and model their cross-talk [17]. Scott et al. proposed a method to reconstruct the known signaling pathways efficiently in protein interaction networks by assigning well-founded reliability scores to PPI data and by applying a color coding algorithm [18].

There are also methods that combine PPI and genetic interaction data to identify signaling pathway structure. The activity pathway network (APN) approach utilizes high-throughput genetic interaction data and applies the Bayesian learning method to identify detailed structure of known signaling pathways [19]. Another method utilizes

the same approach to restructure the pathway by also combining PPI data with genetic interaction data [20].

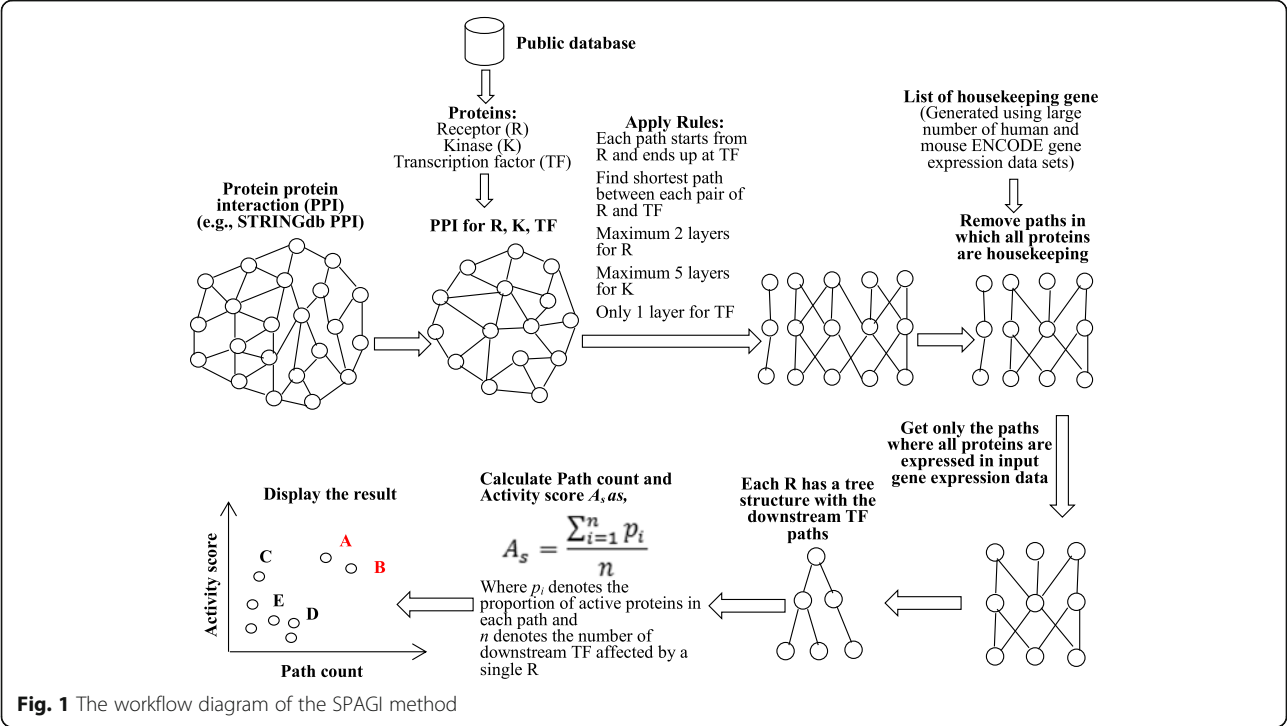
A number of computational methods utilize PPI data along with gene expression data to uncover known signaling pathways [2, 3, 21, 22]. In these methods the gene expression data sets are usually used to calculate the edge weight by gene expression correlation for the network. One approach utilizes PPI and gene expression data sets and applies integer linear programming to get an optimal subnetwork from the PPI network starting from membrane proteins and ending at transcription factors [3]. A recently published method called *HISP* uses the same approach, but in addition applies genetic algorithms with operations including selection, crossover, and mutation to select the candidate topologies of resultant signaling pathways and uses gene knockout data to get directionality of the signaling pathways [2]. *Netsearch* determines networks by integrating protein-protein interaction data with microarray expression data by extracting subnetworks of the protein interaction dataset whose members have the most correlated expression profiles [22]. It generates a specific pathway based on the input proteins (R and TF) and the PPI networks. Another method highlights the order of signaling pathway components, assuming all the components on the pathways are known [21]. It constructs a score function based on the correlations between each gene pair to determine the final signal transduction network.

All of the above methods aim to restructure the topologies of known signaling pathways. However, to our knowledge, no open-source methods have been reported that simultaneously and comprehensively identify the set of active signaling pathways *and* the likely pathway structures for a gene expression profile (i.e., all R, K and effector TF paths for each identified pathway). Additionally, most of the above methods were evaluated and applied to yeast PPI data, with only a few methods designed specifically to deal with the significantly greater complexity of mammalian data. Here we propose an approach to systematically identify the set of active receptor-mediated signaling pathways within any given cell, by combining PPI and gene expression data. This method is implemented as an open source packaging using the 'R' programming language. This open source software is called SPAGI (Signaling Pathway Analysis for putative Gene regulatory network Identification), and is available via <https://github.com/VCCRI/SPAGI/>.

Methods

Building background pathway data

The overall workflow of the SPAGI approach is approach is depicted in Fig. 1. First we collected the known R, K and TF signaling molecules (2134 genes/proteins in total) from public data sets [23–25]. The list of R proteins was



collected from a curated database of the Fantom5 project [24]. The list of K proteins was collected from the Uniprot curated database [23]. The list of TF proteins was obtained from a database of sequence-specific DNA-binding TFs identified by gene ontology (GO) based annotation [25]. Next we separately used both the mouse and human PPI data from *STRING* database (version 10) [26] to obtain all currently known PPIs for the 2134 known R/K/TF signaling molecules - while keeping the human and mouse separate. Please note that we have considered here all the physical and other inferred (e.g., co-expression) interactions when defining PPIs to maximize our ability to detect the full network structure. The confidence (*combined_score*) values assigned to interactions within *STRING* range from 0 to 999. We selected PPIs defined by *STRING* as ‘high confidence’ (i.e. *confidence_score* ≥ 700) to further maximise our

ability to construct networks representative of true biological pathways. This thresholding yielded 16,550 and 19,502 PPIs for mouse and human respectively. After obtaining these highly scored PPIs both for the human and mouse organisms we have merged all the PPIs by assuming that the molecules have one-to-one homology mapping between the organisms. Note that after filtering and considering the presence of bi-directional interactions within *STRING* (e.g., R to K and K to R), the set of all known R/K/TF interactions involves 39,004 PPIs in human and 33,100 PPIs in mouse (with 27,790 PPIs common to both). We then took the union of all PPIs and have assigned the larger score value of a PPI if it is present in both organisms. The merged PPI network has 44,314 edges (See Table 1 for details). From the combined high scored PPIs, we collected only the PPIs for the signaling pathways that have

Table 1 SPAGI pathway path background data summary

	For Mouse	For Human
# R, K, TF	2134	2134
# R/K, K/K, K/TF interactions (<i>combined_score</i> > 0)	234,603	249,571
# high-confidence (<i>combined_score</i> ≥ 700) R/K, K/K, K/TF interactions (assuming bi-directional interaction)	33,100	39,004
# common interaction		27,790
# combined unique interaction		44,314
# high-confidence complete R/K/TF paths		102,842
# high-confidence complete R/K/TF paths without housekeeping gene paths (# of R-defined pathways)		89,161 (548)

interactions able to make full paths from R to K to TF. This process included interactions from:

- (a) R (not directly connected with K) to R (directly connected with K)
- (b) R to K
- (c) K to K and
- (d) K to TF

Finally, we collected all the filtered PPIs from the above step, keeping their associated PPI *combined_score* value for each of the interactions. Note that for clarity, the word ‘path’ is defined as a single R/K/TF prediction, whereas the word ‘pathway’ is defined as the collection of paths that all start from the same R (i.e., all paths defined by a single R constitutes a pathway).

To make the signaling pathway paths, we first made a directed weighted graph from the PPI data using the *igraph* R package. As *igraph* considers the weight of the interaction as a cost (i.e., higher weight means it needs more effort to travel), we have modified the PPI *combined_score* value as weight by calculating *1000-combined_score*. After assigning the *combined_score* as weight we collected the reachable shortest paths from each R protein to each TF protein by utilizing the *shortest_path* function of the package. The *shortest_path* function uses the Dijkstra’s graph algorithm for the weighted directed graph. We have collected all the complete paths (a path is being called complete if it starts from a R protein and ends up to a TF protein) that have a length from 3 to 7, allowing for at most 2 layers for RP, 5 layers for KN and 1 layer for TF. To identify cell type-specific paths, we then filtered out the complete paths where all factors were designated as housekeeping genes (see the next section for how the list of housekeeping genes was generated). As a result of these steps, the final collection of complete paths consists only of those that are not designated as housekeeping paths. These paths are used as background pathway path data for our method.

Housekeeping genes identification

We collected the published RNA-seq gene expression data sets for different cells and tissues both for mouse and human from the ENCODE project [27, 28], and processed them separately. We examined the expression distribution pattern of these data sets and found that on average the $\log_2(FPKM + 1) = 1.5$ value could be used as the expression cut-off for the data sets. Using this cut-off we identified the expressed genes for all the cells and tissues. We then designated a gene as a housekeeping gene if it was found to be expressed in at least 75% of the total number of cells and tissues for that particular organism. This approach was used to identify both the mouse and human housekeeping genes. These 2 lists

of housekeeping genes were then combined to generate a unique list of housekeeping genes, assuming one-to-one homology mapping between human and mouse genes. This combined list of unique housekeeping genes was used as background data.

Potential signaling pathway identification

The background signaling pathway path data was used to identify the potential signaling pathways for a particular gene expression data set. As input we took the gene expression data matrix of \log_2 transformation of *RPKM/FPKM/CPM* values, an expression cut-off threshold to identify the expressed genes, and a high expression threshold (generally an expression value greater than the expression value of the peak of distribution) to calculate the activity score of the pathways.

Processing steps

- (1) From the gene expression data set, first we calculated the average expression value of the replicates and then identified the expressed genes by using the cut-off threshold described above.
- (2) From the background path data we obtained only those paths for which all the protein factors are expressed according to the input gene expression data. This set of paths is treated as potential signaling pathway paths for the gene expression data set.

Ranking of the potential signaling pathways

For each potential signaling pathway, we first calculated the proportion of active molecules (defined as highly expressed genes based on the above high expression threshold) for each path. We then summed all the proportions of all the paths for the pathway and divided the total proportion value by the total number of paths of the pathway. This final value was termed the *Activity score* (A_s) for a pathway and mathematically can be written as:

$$A_s = \frac{\sum_{i=1}^n p_i}{n}$$

Where p_i denotes the proportion of active molecules in each path and n denotes the number of downstream TFs for the pathway. Next we plotted the values of n and A_s to display the results of top ranked active signaling pathways in the upper positions.

Assessment of SPAGI false positive rate

The SPAGI false positive rate was obtained by randomly assigning gene expression data and then re-performed the SPAGI analyses. The number of highly ranked active pathways for each sample was then counted. The false

positive rate for highly ranked pathways was obtained by dividing the number of highly ranked pathways obtained from the randomly assigned data by the number of highly ranked pathways obtained from the original sample. GO analysis was also performed on the randomly assigned gene expression data used to determine the SPAGI false discovery rate. Each GO analysis was performed separately using the online version of Enrichr for biological processes [29]. Results were filtered to retain only the significant terms and for signaling GO terms using the raw p -value. The false positive rate for the GO analysis was calculated by dividing the number of highly ranked pathways obtained via the randomly assigned data by the number of highly ranked pathways obtained from the original data.

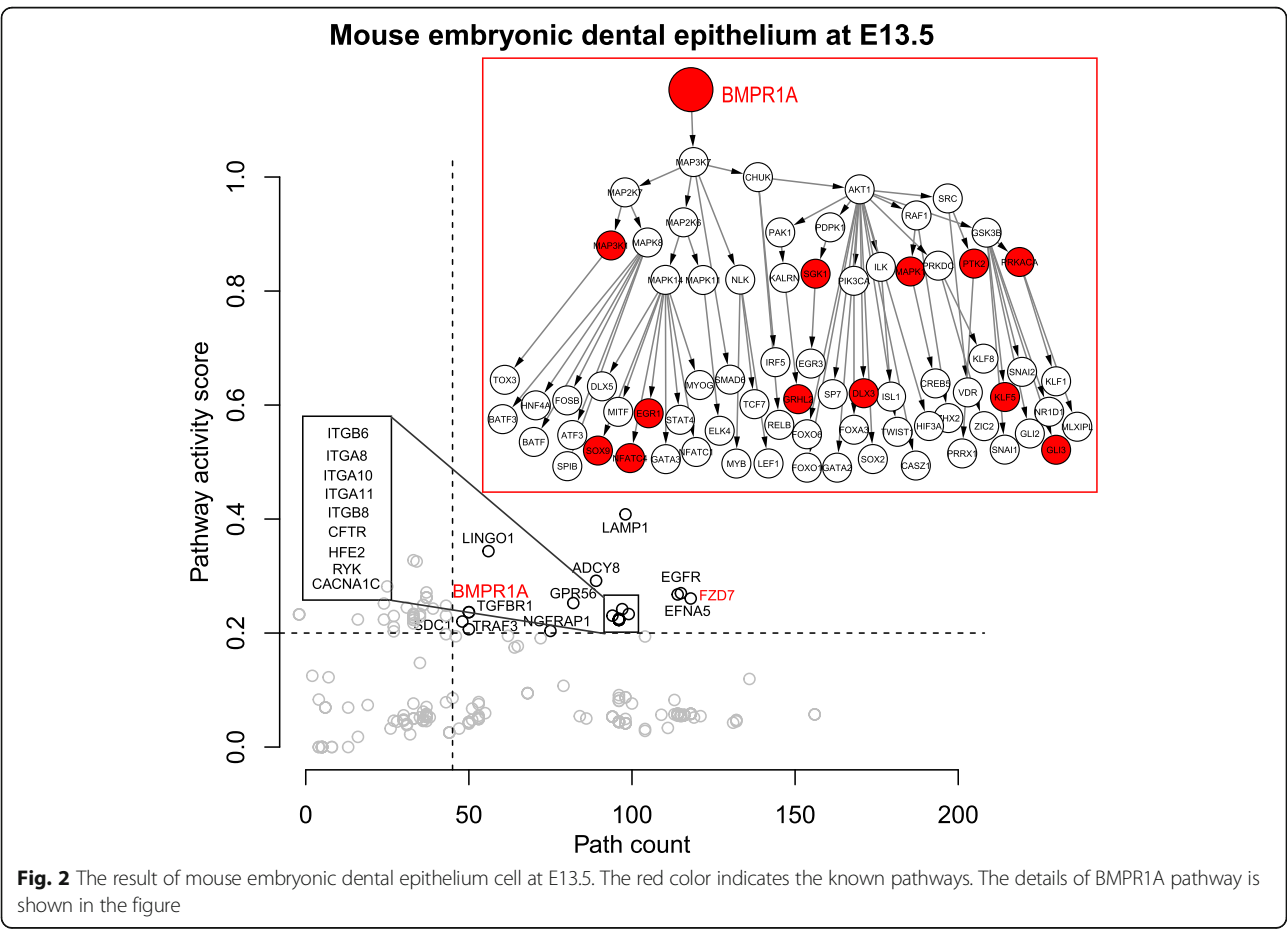
Results

The ability of SPAGI to identify known, critical, tissue-specific signaling pathways was tested using four cell types obtained from three different gene expression data sets (two are RNA-seq and one is microarray). These four cell types were chosen as there is an extensive body of literature for them that has already identified critical pathways, thus enabling biological validation

of the SPAGI output. The first data set used is from mouse dental epithelial cells at the development stage E13.5 ($n = 3$) [30]. The remaining two data sets were from the ocular lens: one is a newborn mouse lens data set that consists of gene expression profiles from lens epithelial cells (LECs; $n = 3$) and lens fiber (LF) cells ($n = 3$) [31]; the other data set is from human pluripotent stem cell-derived ROR1⁺ LEC-like cells ($n = 2$) [32].

SPAGI analysis of tooth

Published data have shown that BMP and WNT (through FZD receptors) signaling pathways are important for embryonic mouse tooth development [30]. Loss of function of BMPR1A in dental epithelial cells reduces WNT expression and prevents tooth formation [30]. To test whether SPAGI can identify BMP and WNT/FZD pathways from published dental epithelial cell gene expression data, we applied the SPAGI method to gene expression data from embryonic development stage E13.5. After all filtering, we have obtained 14,657 specific paths (i.e., 14.25% of total paths) for the dental epithelial cell. This analysis revealed SPAGI identified both the BMPR1A and FZD7 receptor-mediated pathways (Fig. 2), together with a range of other pathways.

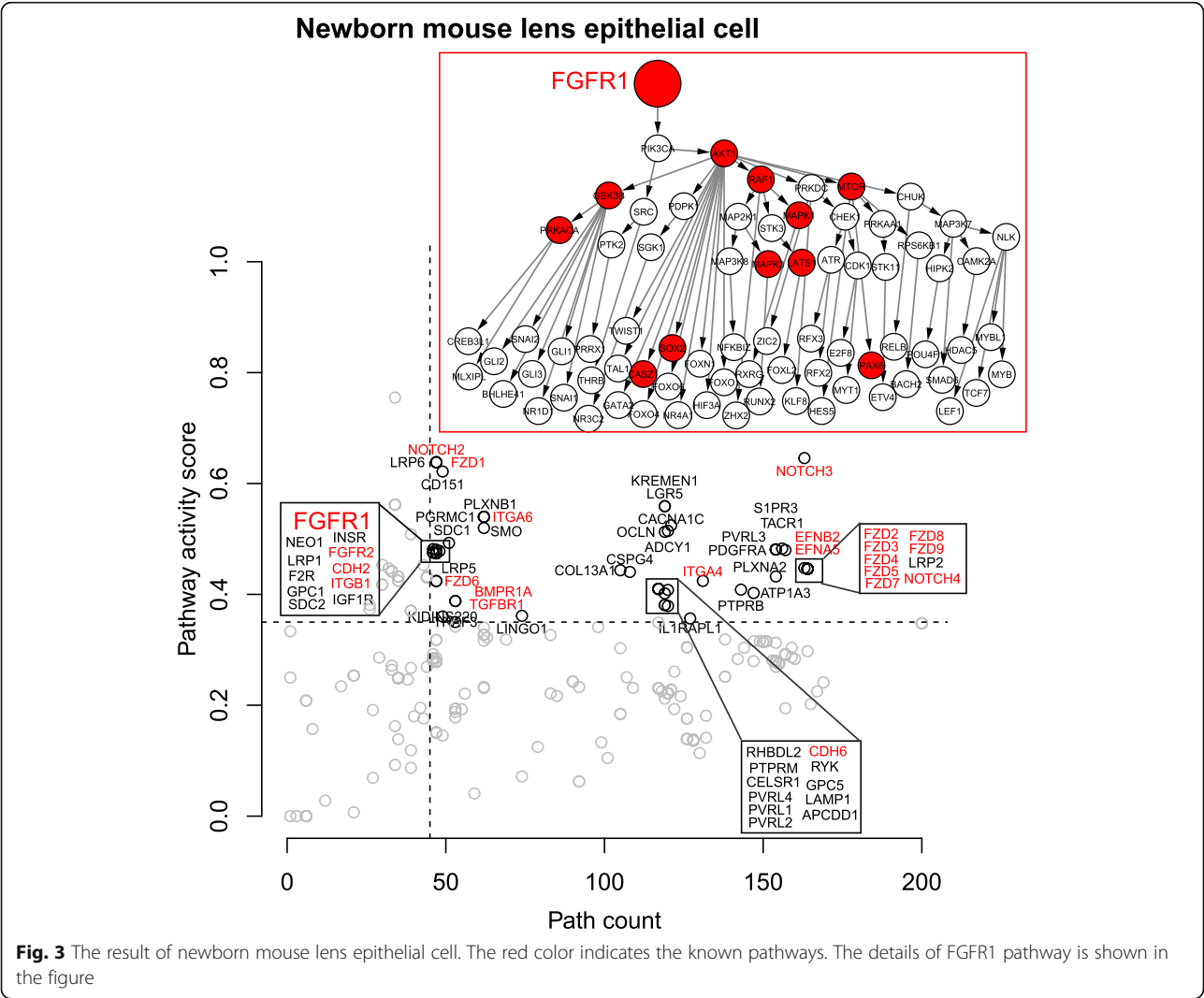


SPAGI analysis of lens gene expression data

A large number of studies over decades have described the requirement for different signaling pathways during lens development. As summarized in a recent review by Cvekl and Zhang [33], while critical lens pathways have been broadly identified, the precise R/K/TF signaling paths utilized within each pathway are not fully understood, nor are the path nodes (typically Ks and TFs) where different signaling pathways intersect. An accurate method for comprehensively identifying R/K/TF paths that operate within lens (and other) tissue is therefore needed. For example, the FGF pathway induces the pre-placodal region required for lens formation, as well as subsequent proliferation of LECs and differentiation of LECs to LF cells. The BMP pathway is also involved in pre-placodal induction, invagination of the lens placode, LEC proliferation and survival, and LF cell differentiation. The FZD pathway works as an inhibitor at the pre-placodal region, and in LEC adhesion, integrity and

polarity. NOTCH signaling controls lens growth and acts as a differentiator for both LECs and LF cells. Signaling through different integrins is required early in lens differentiation, and for cell adhesion, lens capsule assembly and normal development of both LECs and LFs. Cadherins are required for appropriate polarity, adhesion and survival of LECs, and for LF cell elongation. EPHs and Ephrins are involved in cell adhesion and polarity, and LF cell elongation and alignment. The TGFβ pathway acts as an inhibitory signal in the pre-placodal region for proper lens growth, and is implicated in lens diseases such as anterior subcapsular cataract and posterior capsule opacification. Critically, how molecular integration of all these pathways occurs during lens development or formation of different cataract subtypes is currently unclear [34].

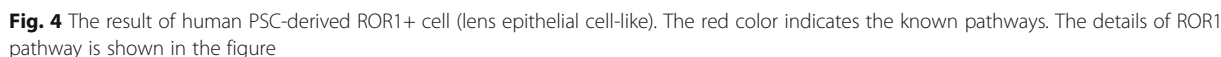
Analysis of published mouse LEC gene expression data [31] using our SPAGI method identified all of the pathways mentioned above (Fig. 3). After all filtering this

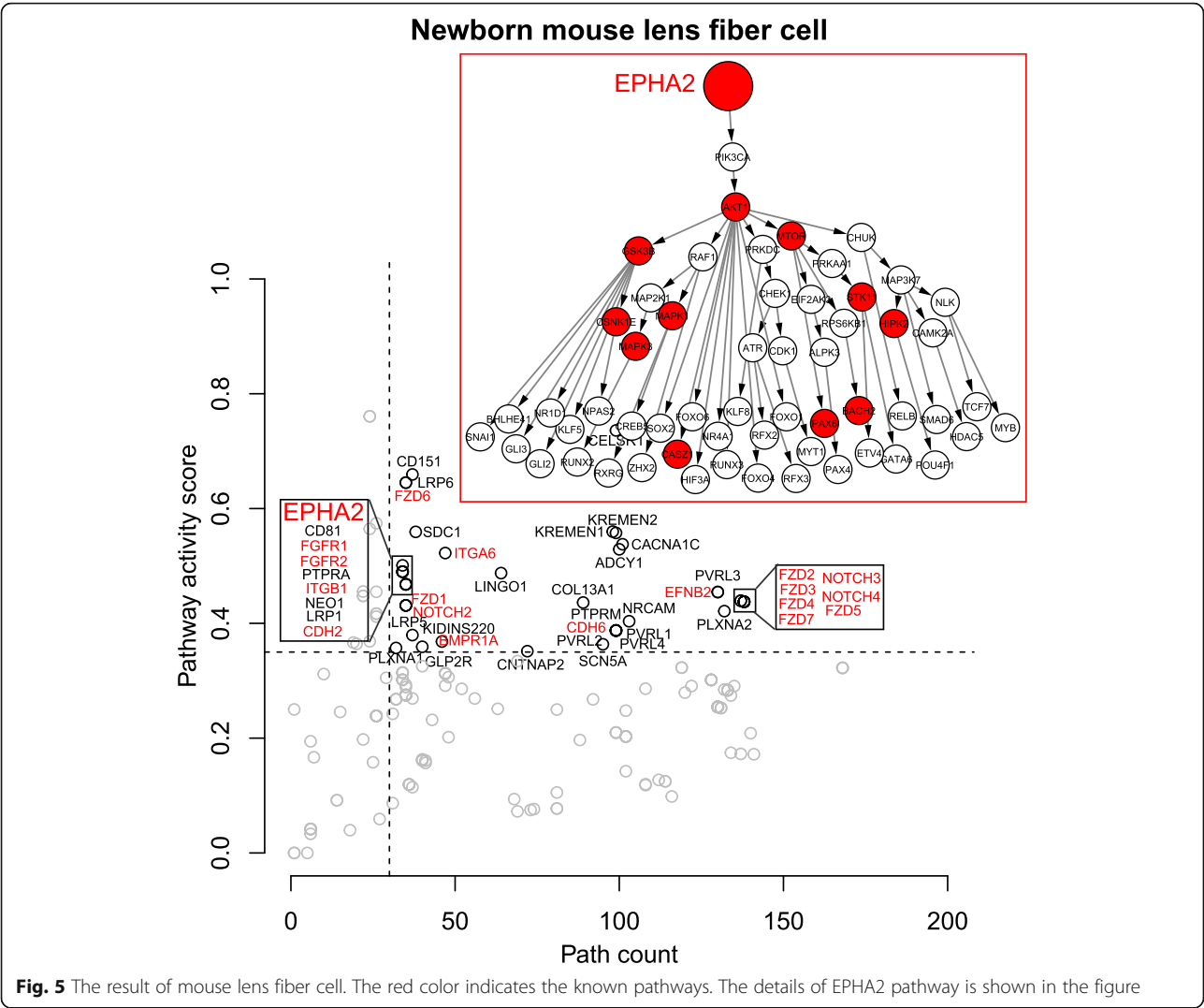


Analysis of published mouse LF cell data shows that, as expected, LF cell signaling pathways are very similar to LECs (Fig. 5). We have obtained 13,790 specific paths (i.e., 13.41% of total paths) after all filtering for mouse LF cell. Differences in the ranking of particular pathways provide indications of how these pathways are integrated in the transition from LECs to LF cells (e.g., EPHA2 in Figs 3, 4 and 5). Overall, these results show that the SPAGI R package can accurately identify and rank known, critically-important signaling pathways from the gene expression profiles of different cell and tissue types. As shown in Figs. 2 to 5, the SPAGI approach identifies each specific R, K and TF within each path

Comparison of SPAGI analysis on species-specific vs combined PPI data

To determine the breadth of PPI data coverage within the mouse and human STRING datasets, we also performed the SPAGI analysis separately for the human and mouse query data – i.e., mouse samples compared only against the mouse STRING data and the human sample compared only against the human STRING data (see Additional File 1: Figures S1–4). From these results we see that a number of known pathways were identified for each sample. However, pathways known from previous studies to be





important for particular cell types were not identified using this species-specific STRING analysis approach. For example, the ROR1 pathway was not identified in the human ROR1⁺ cell sample, despite ROR1 being critical for capturing this population of human LECs. Similarly, the EPHA2 pathway was not identified within the mouse LF cell samples, despite this being a key pathway that leads to disease (i.e., cataract) if disrupted [38]. Thus, combining the mouse and human PPI datasets prior to SPAGI analysis led to more biologically-relevant results for the query samples than obtained when using the human and mouse PPI data separately.

Analysis of the SPAGI false positive rate via random expression level assignment

We investigated the false positive rate of the SPAGI method by randomly assigning gene expression data using both the mouse dental epithelial cell and mouse

LEC gene expression data sets. First, we have randomly assigned the gene names amongst the gene expression values for each sample and then re-performed the SPAGI analyses as done for the original data. We then counted the number of highly ranked active pathways for each sample, and looked for identification of known pathways within the high ranked active pathways to investigate the SPAGI false discovery rate for known pathways from the randomly assigned expression data. Next we calculated the false positive rate for each sample utilizing the number of high ranked pathways of randomly assigned expression sample by dividing the number of high ranked pathways of original sample. We repeated this analysis 10 times for each sample and calculated the average false positive rate for each sample. The average false positive rate for mouse dental epithelial cell is 0.128 and for mouse LEC gene expression data is 0.022 (see Additional File 1: Tables S1 and S2).

Analysis of the SPAGI false positive rate versus GO analysis

We also compared the performance of the SPAGI method with GO analysis method. The GO analysis was performed based on the unique set of molecules (i.e., Rs, Ks and TFs) from the original mouse dental epithelial cell and mouse LEC data. For comparison, we also performed GO analysis based on the same random assignment used to determine the SPAGI false discovery rate described above. Each GO analysis was performed separately using the online version of Enrichr [36], and captured all the results associated with biological process. These results were filtered to retain only the significant terms based on raw *p*-value and for signaling GO terms. Finally we searched for known pathways for each sample. Additional File 1: Table S3 shows the comparison results of the original cell samples, and Additional File 1: Tables S4 and S5 show the comparison results obtained using the randomly assigned cell samples. The results show that both the SPAGI and GO methods can identify almost all the known pathways for the original sample data, although the GO method did not identify the Cadherins pathways in the mouse LECs data. However, the results of the randomly assigned gene expression data showed that the false identification rate of known pathways by SPAGI was much smaller (0–0.2) than for the GO analysis method (0.4–1) (Additional File 1: Table S6).

Discussion

In this manuscript we described a new bioinformatics method, SPAGI, that can simultaneously and comprehensively discover the set of active signaling pathways and their putative defined path structures. Our evaluation demonstrates that the SPAGI method can accurately identify known and biologically-relevant signaling pathways from multiple gene expression data sets across different tissue types, while providing specific detail of the molecular cascades involved in these pathways. The SPAGI method therefore provides capabilities not available with other current open-source software. While some pathway analysis software is commercially available (e.g., IPA), SPAGI provides a free and open-source approach that can routinely provide updated data through updates to the STRING database.

In addition to validation of the SPAGI method by comparison against known biology, the SPAGI approach was also validated by assessment of its false positive rate - both on its own and in comparison to the false positive rate obtained via GO analysis. The SPAGI approach identified few pathways when using randomly assigned gene expression data

for the mouse dental epithelial cells (0.128) and mouse LECs (0.022). Moreover, the results of the randomly assign gene expression data showed the false positive rate was smaller for the SPAGI method (0–0.2) than the false positive rate obtained via the GO analysis method (0.4–1). These data provide strong support for SPAGI being both more sensitive and more specific than pathway identification via GO analysis alone.

To assess whether the SPAGI method is best applied to species-specific PPI data or combined/multi-species PPI data, we performed SPAGI analysis on both single species and combined species PPI data. While large numbers of pathways were identified via the single species analyses, some biologically-relevant pathways were not identified. This included the ROR1 receptor-mediated pathway not being identified via the human PPI data, and the EPHA2 pathway not being identified in the mouse LF cell data. As both these pathways appear to be important in their respective cell types [32, 38], SPAGI is currently best performed (i.e., identifies the largest number of biologically-relevant pathways) using the combined species PPI data.

It should be noted that as currently applied, the SPAGI method detects receptor-mediated signaling pathways. Modification of the SPAGI approach could be used to identify other cellular control mechanisms involving PPIs independent of TFs. Also, at this stage it is not clear whether the other pathways highly ranked by the activity score are truly active, as protein expression and protein activation state (e.g., via phosphorylation) within a tissue cannot be determined from gene expression data. Nonetheless, the breadth of data provided by SPAGI can provide specific testable hypotheses for cell biologists to guide functional genomic studies to identify critical regulators involved in health and disease. As such, more studies are required to investigate these pathways.

Conclusions

The SPAGI method represents a new, interesting and open-source method to comprehensively identify important receptor-mediated signaling pathways from a gene expression data set. We have applied our method to four different gene expression data sets from three different cell types and shown that the SPAGI method correctly identified all the known signaling pathways for the cells, with low false discovery rate and lower false discovery than using GO analysis alone. Our results suggest that SPAGI can be a useful approach to identify the potential active signaling pathways given a gene expression profile.

Additional file

Additional File 1: Figure S1. The result of mouse embryonic dental epithelium cell at E13.5 with only the mouse PPI background pathway data. **Figure S2.** The result of newborn mouse lens epithelium cell with only the mouse PPI background pathway data. **Figure S3.** The result of human PSC-derived ROR1+ cell (lens epithelium cell-like) with only the human PPI background pathway data. **Figure S4.** The result of newborn mouse lens fiber cell with only the mouse PPI background pathway data. **Table S1.** False positive rate calculation for SPAGI method of randomly assigns gene expression values of new born mouse lens epithelial cell and mouse tooth epithelial cell at embryonic day E13.5. **Table S2.** SPAGI test result for randomly assign gene expression values of new born mouse lens epithelial cell and mouse tooth epithelial cell at embryonic day E13.5. **Table S3.** Identification of known pathways by SPAGI and GO analysis methods. **Table S4.** Summary of known pathways identification by SPAGI and GO methods for randomly assigns genes of mouse lens epithelial cell. **Table S5.** Summary of known pathways identification by SPAGI and GO methods of randomly assigns genes for mouse tooth epithelial cell. **Table S6.** False positive rate of SPAGI and GO analysis method for known pathways. (PDF 658 kb)

Abbreviations

APN: Activity pathway network; K: Kinase; LEC: Lens epithelial cell; LF: Lens fiber; PPI: Protein-protein interaction; R: Receptor; SPAGI: Signaling pathway analysis for putative gene regulatory network identification; TF: Transcription factor

Acknowledgements

We thank members of the Ho Laboratory for their insightful discussion.

Funding

M.H.K was supported by a UWS Postgraduate Research Award (International). M.D.O'C was supported by The Medical Advances Without Animals Trust. J.W.K.H is supported by a Career Development Fellowship by the National Health and Medical Research Council (1105271) and a Future Leader Fellowship from the National Heart Foundation of Australia (100848). Publication of this article was sponsored by the Future Leader Fellowship from the National Heart Foundation of Australia (100848).

Availability of data and materials

The R package and the associated example data to execute it are available via <https://github.com/VCCRI/SPAGI/>.

About this supplement

This article has been published as part of *BMC Systems Biology Volume 12 Supplement 9, 2018: Proceedings of the 29th International Conference on Genome Informatics (GIW 2018): systems biology*. The full contents of the supplement are available online at <https://bmcsystbiol.biomedcentral.com/articles/supplements/volume-12-supplement-9>.

Authors' contributions

J.W.K.H and M.D.O'C conceived the SPAGI approach and supervised the project. M.H.K designed the method, implemented the R package, carried out the experiments and wrote the manuscript. R.P carried out critical evaluation and testing. All authors revised and approved the final version of the manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing financial interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹School of Medicine, Western Sydney University, Campbelltown, NSW, Australia. ²Victor Chang Cardiac Research Institute, Darlinghurst, NSW, Australia. ³Department of Computer Science and Engineering, University of Rajshahi, Rajshahi, Bangladesh. ⁴St. Vincent's Clinical School, University of New South Wales, Sydney, NSW, Australia. ⁵Stem Cells Australia, Melbourne Brain Centre, University of Melbourne, Parkville, VIC 3010, Australia. ⁶School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Pokfulam, Hong Kong, SAR, China. ⁷Molecular Medicine Research Group, Western Sydney University, Campbelltown, NSW, Australia.

Published: 31 December 2018

References

- Wang K, Hu F, Xu K, Cheng H, Jiang M, Feng R, et al. CASCADE_SCAN: mining signal transduction network from high-throughput data based on steepest descent method. *BMC Bioinformatics*. 2011;12:164.
- Zhao XM, Li S. HISP: a hybrid intelligent approach for identifying directed signaling pathways. *J Mol Cell Biol*. 2017;9(6):453–62.
- Zhao XM, Wang RS, Chen L, Aihara K. Uncovering signal transduction networks from high-throughput data by integer linear programming. *Nucleic Acids Res*. 2008;36(9):e48.
- Hunter T. Signaling—2000 and beyond. *Cell*. 2000;100(1):113–27.
- Takahashi A, Ohtani N, Hara E. Irreversibility of cellular senescence: dual roles of p16INK4a/Rb-pathway in cell cycle control. *Cell Div*. 2007;2:10.
- Jiang R. Walking on multiple disease-gene networks to prioritize candidate genes. *J Mol Cell Biol*. 2015;7(3):214–30.
- Liu KQ, Liu ZP, Hao JK, Chen L, Zhao XM. Identifying dysregulated pathways in cancers from pathway interaction networks. *BMC Bioinformatics*. 2012;13:126.
- Wang YCH, Pan Z, Ren J, Liu Z, Xue Y. Reconfiguring phosphorylation signaling by genetic polymorphisms affects cancer susceptibility. *J Mol Cell Biol*. 2015;7(3):187–202.
- Zhang W, Zeng T, Liu X, Chen L. Diagnosing phenotypes of single-sample individuals by edge biomarkers. *J Mol Cell Biol*. 2015;7(3):231–41.
- Smith RJ Jr, Koobatian MT, Shahini A, Swartz DD, Andreadis ST. Capture of endothelial cells under flow using immobilized vascular endothelial growth factor. *Biomaterials*. 2015;51:303–12.
- Basha O, Flom D, Barshir R, Smoly I, Tirman S, Yeger-Lotem E. MyProteinNet: build up-to-date protein interaction networks for organisms, tissues and user-defined contexts. *Nucleic Acids Res*. 2015;43(W1):W258–63.
- von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res*. 2003;31(1):258–61.
- Ritz A, Poirel CL, Tegge AN, Sharp N, Simmons K, Powell A, et al. Pathways on demand: automated reconstruction of human signaling networks. *NPJ Syst Biol Appl*. 2016;2:16002.
- Gil DP, Law JN, Murali TM. The PathLinker app: Connect the dots in protein interaction networks. *F1000Res*. 2017;6:58.
- Bebek G, Yang J. PathFinder: mining signal transduction pathway segments from protein-protein interaction networks. *BMC Bioinformatics*. 2007;8:335.
- Gitter A, Klein-Seetharaman J, Gupta A, Bar-Joseph Z. Discovering pathways by orienting edges in protein interaction networks. *Nucleic Acids Res*. 2011;39(4):e22.
- Mei S, Zhu H. Multi-label multi-instance transfer learning for simultaneous reconstruction and cross-talk modeling of multiple human signaling pathways. *BMC Bioinformatics*. 2015;16:417.
- Jacob Scott TI, Richard M. Karp and Roded Sharan. Efficient algorithms for detecting signaling pathways in protein interaction networks. *J Comput Biol*. 2006;13(2):133–44.
- Battle A, Jonikas MC, Walter P, Weissman JS, Koller D. Automated identification of pathways from quantitative genetic interaction data. *Mol Syst Biol*. 2010;6:379.
- Fu C, Deng S, Jin G, Wang X, Yu ZG. Bayesian network model for identification of pathways by integrating protein interaction with genetic interaction data. *BMC Syst Biol*. 2017;11(Suppl 4):81.
- Liu Y, Zhao H. A computational approach for ordering signal transduction pathway components from genomics and proteomics data. *BMC Bioinformatics*. 2004;5:158.
- Steffen M, Petti A, Aach J, D'Haeseleer P, Church G. Automated modelling of signal transduction networks. *BMC Bioinformatics*. 2002;3:34.

23. UniProt - Swiss-Prot Protein Knowledgebase 2017 [Human and mouse protein kinases: classification and index]. Available from: <http://www.uniprot.org/docs/pkinfam>. Accessed on 10 January 2018.
24. Ramilowski JA, Goldberg T, Harshbarger J, Kloppmann E, Lizio M, Satagopam VP, et al. A draft network of ligand-receptor-mediated multicellular signalling in human. *Nat Commun.* 2015;6:7866.
25. Tripathi S, Christie KR, Balakrishnan R, Huntley R, Hill DP, Thommesen L, et al. Gene Ontology annotation of sequence-specific DNA binding transcription factors: setting the stage for a large-scale curation effort. *Database (Oxford).* 2013;2013:bat062.
26. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 2015;43(Database issue):D447–52.
27. Consortium TEP. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012;489(7414):57–74.
28. Consortium TME, Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, et al. An encyclopedia of mouse DNA elements (mouse ENCODE). *Genome Biol.* 2012;13(8):418.
29. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 2016;44(W1):W90–7.
30. O'Connell DJ, Ho JW, Mammoto T, Turbe-Doan A, O'Connell JT, Haseley PS, et al. A Wnt-bmp feedback circuit controls intertissue signaling dynamics in tooth organogenesis. *Sci Signal.* 2012;5(206):ra4.
31. Hoang TV, Kumar PK, Sutharzan S, Tsonis PA, Liang C, Robinson ML. Comparative transcriptome analysis of epithelial and fiber cells in newborn mouse lenses with RNA sequencing. *Mol Vis.* 2014;20:1491–517.
32. Murphy P, Kabir MH, Srivastava T, Mason ME, Dewi CU, Lim S, et al. Light-focusing human micro-lenses generated from pluripotent stem cells model lens development and drug-induced cataract in vitro. *Development.* 2018;145(1):dev155838.
33. Cvekl A, Zhang X. Signaling and gene regulatory networks in mammalian Lens development. *Trends Genet.* 2017;33(10):677–702.
34. Lovicu FJ, McAvoy JW, de longh RU. Understanding the role of growth factors in embryonic development: insights from the lens. *Philos Trans R Soc Lond Ser B Biol Sci.* 2011;366(1568):1204–18.
35. Mishra SK, Funair L, Cressley A, Gittes GK, Burns RC. High-affinity Dkk1 receptor Kremen1 is internalized by clathrin-mediated endocytosis. *PLoS One.* 2012;7(12):e52190.
36. Niehrs C. Function and biological roles of the Dickkopf family of Wnt modulators. *Oncogene.* 2006;25(57):7469–81.
37. Lachke SA, Higgins AW, Inagaki M, Saadi I, Xi Q, Long M, et al. The cell adhesion gene PVRL3 is associated with congenital ocular defects. *Hum Genet.* 2012;131(2):235–50.
38. Shiels A, Bennett TM, Knopf HL, Maraini G, Li A, Jiao X, et al. The EPHA2 gene is associated with cataracts linked to chromosome 1p. *Mol Vis.* 2008;14:2042–55.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



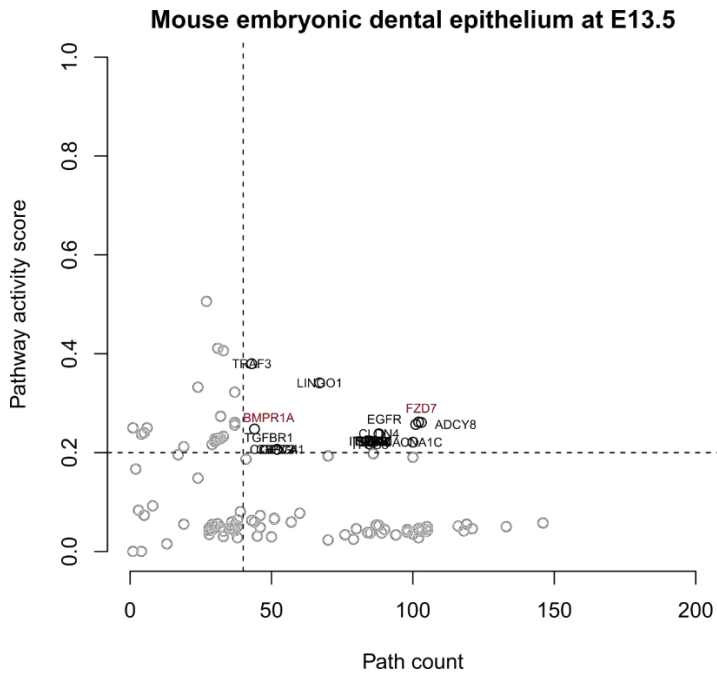


Figure S1. The result of mouse embryonic dental epithelium cell at E13.5 with only the mouse PPI background pathway data.

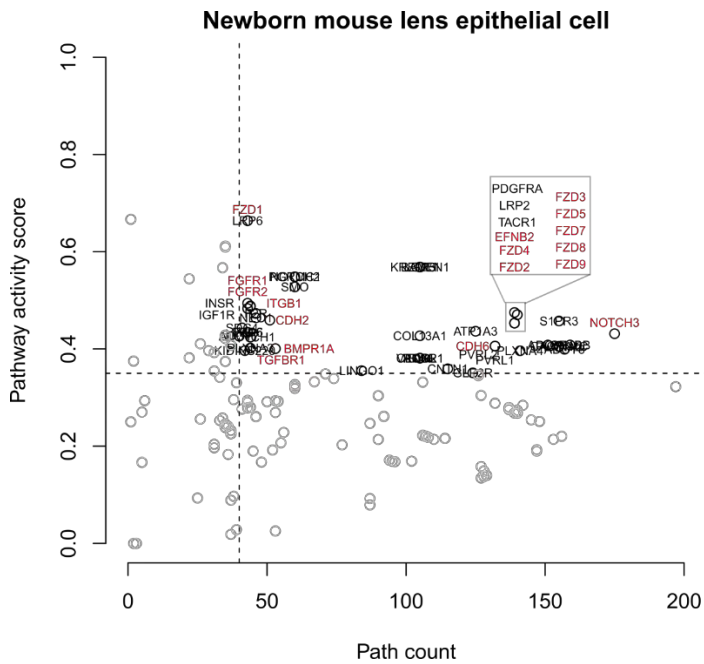


Figure S2. The result of newborn mouse lens epithelium cell with only the mouse PPI background pathway data.

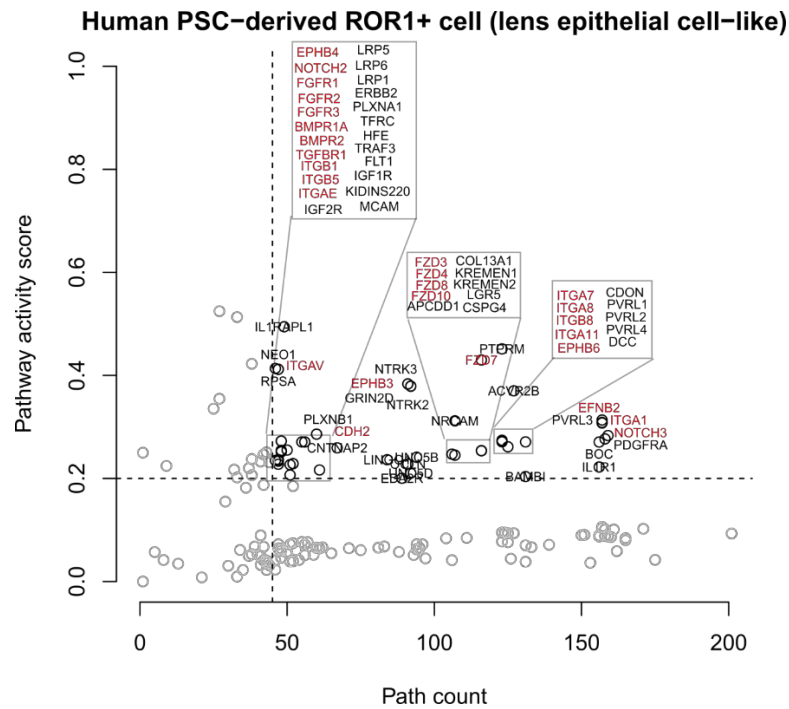


Figure S3. The result of human PSC-derived ROR1+ cell (lens epithelium cell-like) with only the human PPI background pathway data.

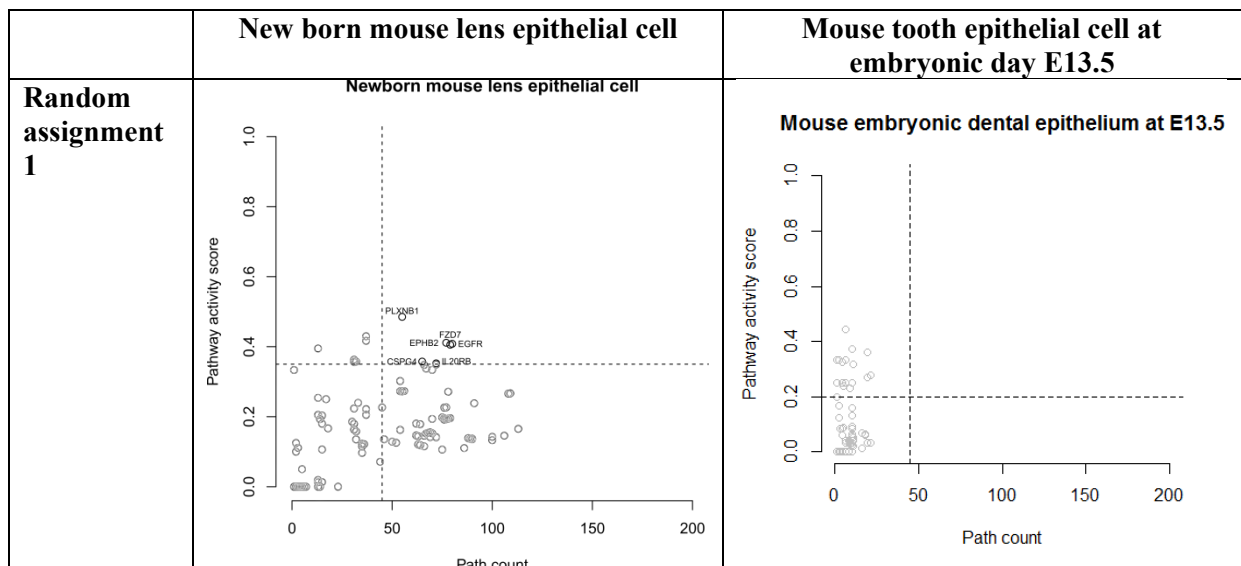


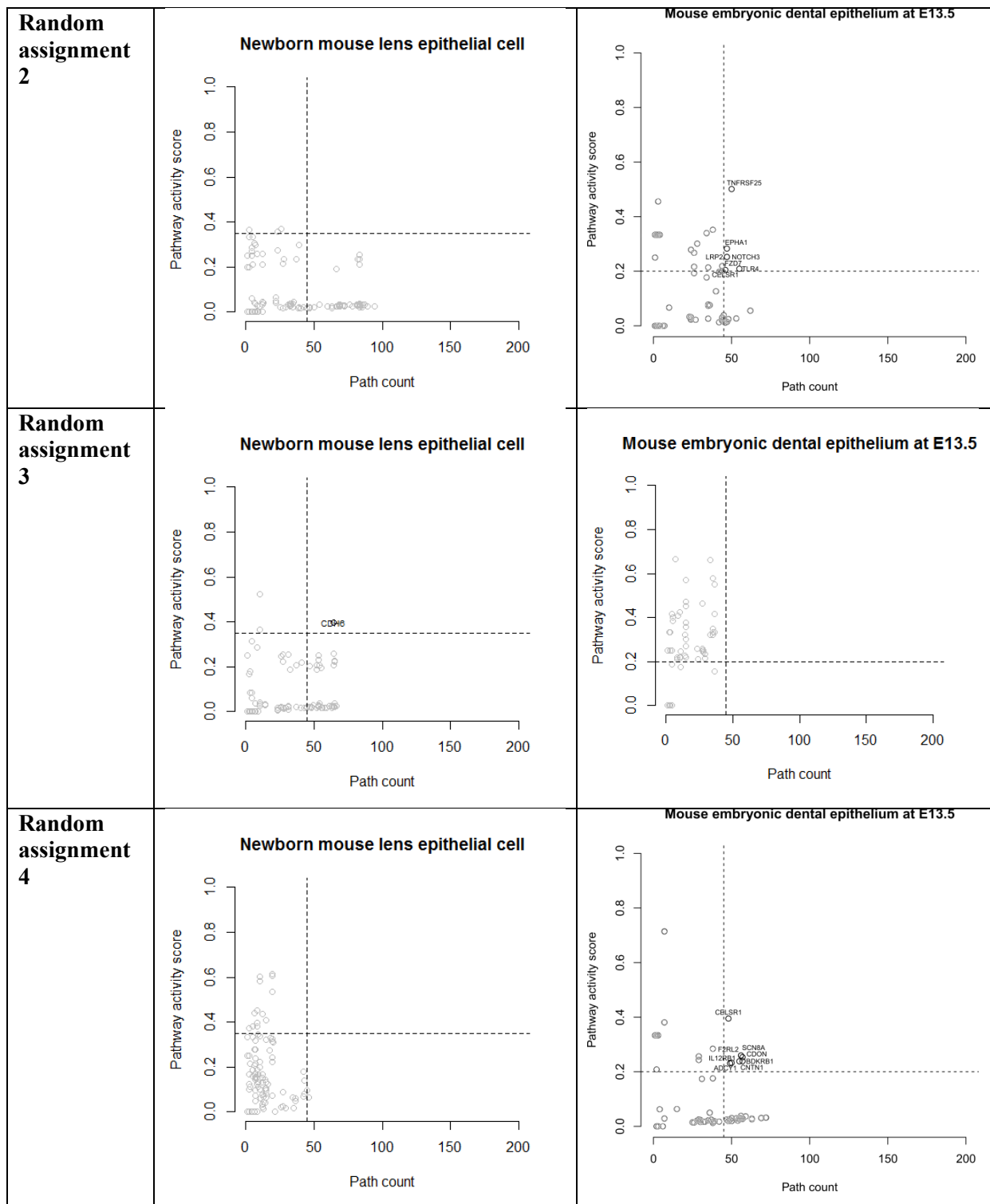
Figure S4. The result of newborn mouse lens fiber cell with only the mouse PPI background pathway data.

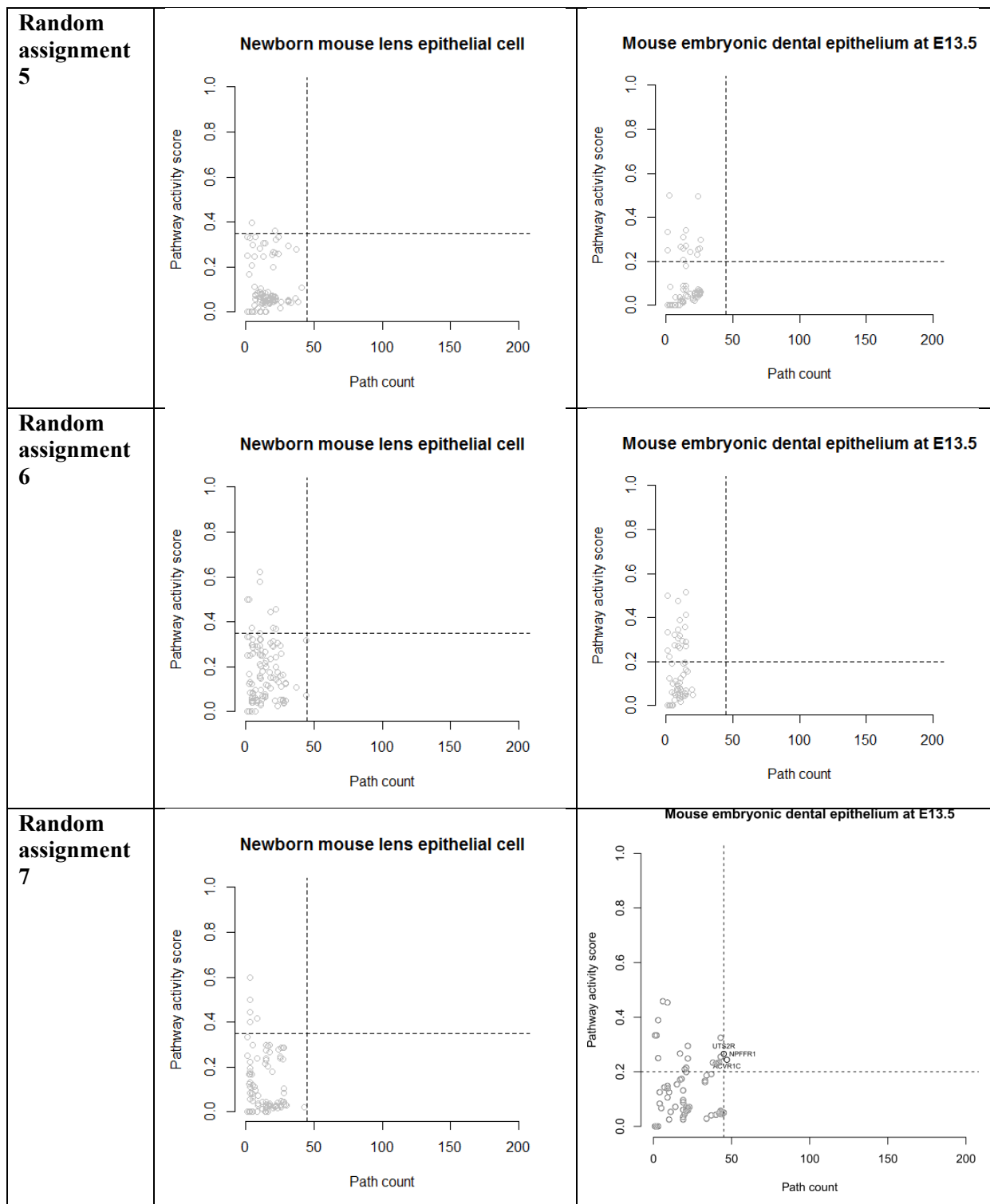
Table S1. False positive rate calculation for SPAGI method of randomly assigns gene expression values of new born mouse lens epithelial cell and mouse tooth epithelial cell at embryonic day E13.5

	New born mouse LEC		Mouse tooth epi at E13.5	
	# of high ranked pathways	False positive rate	# of high ranked pathways	False positive rate
With original expression	66		21	
Random assignment 1	6	0.09	0	0
Random assignment 2	0	0	7	0.33
Random assignment 3	1	0.02	0	0
Random assignment 4	0	0	8	0.38
Random assignment 5	0	0	0	0
Random assignment 6	0	0	0	0
Random assignment 7	0	0	3	0.14
Random assignment 8	7	0.11	9	0.43
Random assignment 9	0	0	0	0
Random assignment 10	0	0	0	0
	Total false positive rate	0.22	Total false positive rate	1.28
	Average false positive rate	0.022	Average false positive rate	0.128

Table S2. SPAGI test result for randomly assign gene expression values of new born mouse lens epithelial cell and mouse tooth epithelial cell at embryonic day E13.5







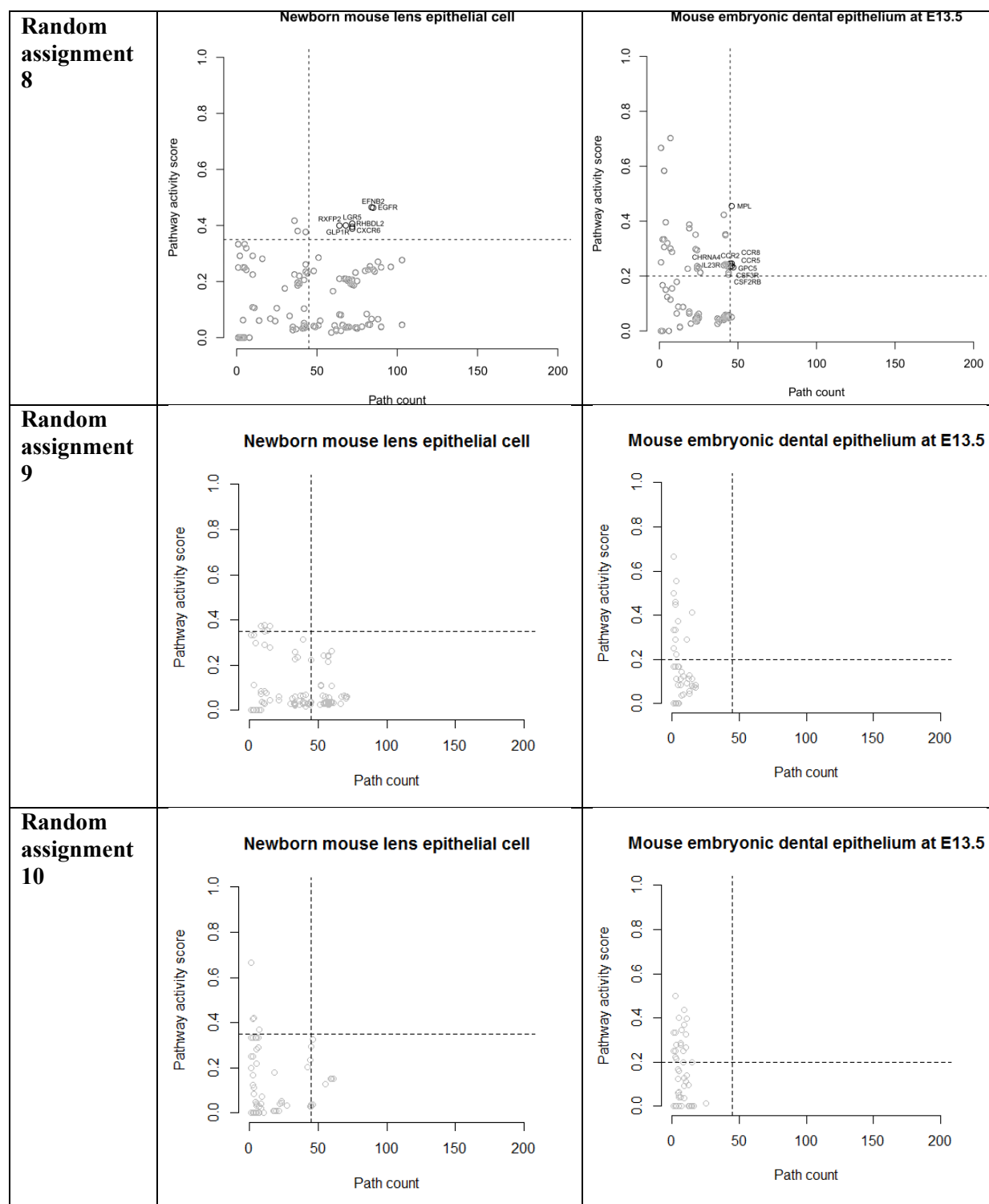


Table S3. Identification of known pathways by SPAGI and GO analysis methods

	BMP	Cadherins	EPH/Ephrin	FGF	Integrins	Notch	TGFB	Wnt
Known for lens epithelial cell	Y	Y	Y	Y	Y	Y	Y	Y
Identified by SPAGI	Y	Y	Y	Y	Y	Y	Y	Y
Identified by GO	Y	N	Y	Y	Y	Y	Y	Y
Known for tooth epithelial cell	Y							Y
Identified by SPAGI	Y							Y
Identified by GO	Y							Y

Table S4. Summary of known pathways identification by SPAGI and GO methods for randomly assigns genes of mouse lens epithelial cell

	BMP		Cadherins		EPH/Ephrin		FGF		Integrins		Notch		TGFB		Wnt	
	SPAGI	GO	SPAGI	GO	SPAGI	GO	SPAGI	GO	SPAGI	GO	SPAGI	GO	SPAGI	GO	SPAGI	GO
Random assignment 1	0	1	0	0	1	1	0	0	0	1	0	1	0	1	1	1
Random assignment 2	0	1	0	0	0	1	0	0	0	1	0	1	0	1	0	1
Random assignment 3	0	1	1	0	0	1	0	0	0	1	0	1	0	1	0	1
Random assignment 4	0	1	0	0	0	1	0	1	0	1	0	1	0	1	0	1
Random assignment 5	0	1	0	0	0	1	0	1	0	1	0	1	0	1	0	1
Random assignment 6	0	1	0	0	0	1	0	1	0	1	0	1	0	1	0	1
Random assignment 7	0	0	0	0	0	1	0	0	0	1	0	1	0	1	0	1
Random assignment 8	0	1	0	0	1	1	0	1	0	1	0	1	0	1	0	1
Random assignment 9	0	1	0	0	0	1	0	0	0	1	0	1	0	1	0	1
Random assignment 10	0	1	0	0	0	1	0	0	0	1	0	1	0	1	0	1
Total	0	9	1	0	2	10	0	4	0	10	0	10	0	10	1	10

Table S5. Summary of known pathways identification by SPAGI and GO methods of randomly assigns genes for mouse tooth epithelial cell

	BMP		Wnt (Fzd)	
	SPAGI	GO	SPAGI	GO
Random assignment 1	0	0	0	1
Random assignment 2	0	1	1	1
Random assignment 3	0	1	0	1
Random assignment 4	0	1	0	1
Random assignment 5	0	1	0	1
Random assignment 6	0	0	0	1
Random assignment 7	0	1	0	1
Random assignment 8	0	0	0	1
Random assignment 9	0	0	0	0
Random assignment 10	0	1	0	1
Total	0	6	1	9

Table S6. False positive rate of SPAGI and GO analysis method for known pathways

		BMP	Cadherins	EPH/Ephrin	FGF	Integrins	Notch	TGFB	Wnt
Lens epithelial cell	SPAGI	0	0.1	0.2	0	0	0	0	0.1
	GO	0.9	0	1	0.4	1	1	1	1
Tooth epithelial cell	SPAGI	0							0.1
	GO	0.6							0.9

The SPAGI is an efficient approach to identify active signalling pathways en masse from microarray or RNA-seq gene expression data. It outputs a ranking of signal paths – each consisting of receptor(s), kinases, and transcriptional regulators – with paths grouped as receptor-defined pathways. This result provides detail information of signal pathways for clinical applications.

Chapter 5

**Large scale profiling of lens epithelial cell signalling
and gene expression networks reveals regulatory
pathways for known cataract genes**

An important application of systems biology is illumination of how inputs from large numbers of signaling pathways are integrated, in order to precisely regulate specific gene expression networks involved in tissue development, repair, regeneration or disease. Here we provide an in-depth characterisation of a published newborn mouse lens epithelial cell dataset using the SPAGI method. The results generated by the SPAGI analysis were extended by comparison with published lens epithelial cell target genes, and also cataract associated genes, to generate a transcriptional blueprint for lens epithelial cells.

Large scale profiling of lens epithelial cell signalling pathways and target genes reveals regulatory networks for cataract-associated genes

Md Humayun Kabir^{1,2,3}, Patricia Murphy¹, Seakcheng Lim¹, Joshua W. K. Ho^{2,4,5}, Michael D. O'Connor^{1,6,*}

¹*School of Medicine, Western Sydney University, Campbelltown, NSW, Australia*

²*Victor Chang Cardiac Research Institute, Darlinghurst, NSW, Australia*

³*Department of Computer Science and Engineering, University of Rajshahi, Bangladesh*

⁴*St. Vincent's Clinical School, University of New South Wales, Sydney, NSW, Australia*

⁵*School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Pokfulam, Hong Kong SAR, China*

⁶*Medical Sciences Research Group, Western Sydney University, Campbelltown, NSW, Australia*

*Corresponding author.

E-mail: m.oconnor@westernsydney.edu.au

Keywords

signalling pathway; target gene, lens epithelial cell; lens fiber; pluripotent stem cell; gene regulatory network; cataract; micro-lens

Abstract

A key frontier in systems biology is elucidation of how inputs from large numbers of signalling pathways are integrated, in order to precisely regulate specific gene expression networks involved in tissue development, repair, regeneration or disease. The relatively simple anatomy of the ocular lens - consisting of an anterior epithelial monolayer from which differentiate the lens fibre cells that form the bulk of the tissue - makes the lens a valuable model for investigating growth factor-mediated transcriptional events. To catalogue the breadth of signalling paths that could potentially operate in the lens, we first analysed public gene expression data to identify all lens-expressed receptors, kinases and transcription factors. We subsequently used experimentally-validated, one-to-one protein-protein interaction data to build lens signalling paths from receptors to transcriptional regulators. The pathways were extended to include publically-available target gene sets associated with lens-expressed transcriptional regulators. Cataract-associated genes were then mapped onto the target gene sets, based on the hypothesis that identifying the gene regulatory networks for cataract-associated genes will identify important lens transcriptional regulation events. This approach identified a large number of interconnected signalling pathways and associated gene regulatory networks. One network identified in this way consisted of 64 transcriptional regulators controlling expression of 63 cataract-associated genes. This network included a large number of known lens transcription factors and also known lens gene regulatory events, with known lens-related signalling pathways predicted to regulate different aspects of the network. Assessment of the degree of inter-regulation between the 64 transcriptional regulators identified a higher-level network centred on Pax6, Sp1, Ets1, Creb3l1, Klf4, Egr1 and E2f4 in lens epithelial cells and Atf4, Creb4 and Rxrg in lens fibre cells. The utility of the integrated signalling pathways and gene regulatory networks identified here was confirmed by the demonstration that ELK1 is expressed and phosphorylated in ROR1⁺ human lens epithelial cells. Thus the integrated signalling pathway and gene regulatory networks compiled here provide a powerful new predictive blueprint for hypothesis-driven investigation of the molecular mechanisms controlling lens and cataract formation.

1. Introduction

Understanding how inputs from large numbers of extrinsic and intrinsic signalling pathways are integrated - to precisely regulate the expression of genes required for tissue specification, tissue repair and regeneration, or disease - is a major frontier in systems biology. The ocular lens has for over a century proven to be a valuable model for investigating factors that control tissue development [1-3], regeneration [4-6] and disease [7, 8]. This is due in part to the lens' relative anatomical simplicity in comparison to other tissues, namely, an anterior lens epithelial cell (LEC) monolayer, a mass of lens fibre (LF) cells and an overlying basement membrane (lens capsule). The lens also provides easily identifiable and quantifiable functional readouts of impaired development or disease, specifically transparency and focusing ability.

A large number of elegant studies over recent decades have begun to define important lens signalling pathways and gene regulatory networks. This includes studies of: receptors (Rs) for various growth factors (GFs) (reviewed by Lovicu et al [9]) and cell attachment/cell motility proteins (e.g., Eph/Efn [10, 11], integrins [12], cadherins [13], connexins [14]; kinases (Ks) [9, 15]; transcriptional regulators (TRs; such as transcription factors, histone modifiers, etc.) and target genes (TGs) as reviewed by Zhang and Cvekl, 2017 [16]. For example, an increasing anterior-to-posterior gradient [17] of fibroblast growth factor (FGF) signalling is thought to be a key driver of LF cell differentiation: FGFs, acting via FGF Rs, activate Ks such as PI3K and MAPKs leading to regulation of TRs such as PAX6 and subsequent alteration of gene expression (e.g., crystallin expression) [18-20]. Similarly, bone morphogenic proteins (BMPs), acting at least via type I Bmp Rs, activate MAPKs leading to regulation of TRs such as Smad proteins and subsequent alteration of gene expression [21].

Transcriptional profiling studies of whole lenses or micro-dissected mouse and human LECs and LF cells have catalogued genes expressed by these different lens cell types. Most recently, RNA-seq profiling of micro-dissected E14.5 to P0.5 mouse lenses revealed cell-type specific gene expression patterns consistent with known lens biology (e.g., crystallin accumulation, autophagy, etc.). That study provided significant new and valuable information on non-signalling pathways implicated in lens biology but for which more detailed information is needed (e.g., Mtor, ubiquitination,

sumoylation), as well as a strong indication that little-known lens TRs need to be further investigated (e.g., Sp1, E2f1, E2f4, Ets1, Elk1).

The accumulated knowledge from the above and other lens studies provides a strong foundation for understanding how signalling through Rs, Ks and TRs might regulate the TGs required for lens formation, growth and maintenance. However, a major current challenge is to determine how the activity of individual pathways and gene regulatory networks is integrated to achieve normal lens development. In turn, this knowledge will provide a detailed molecular framework for understanding how disease (e.g., cataract) - caused by genetic or environmental factors - perturbs signalling pathways and gene regulatory networks, and potentially how disease activates lens protection mechanisms.

Defining how risk factors initiate disease by adversely modifying existing cell-cell interactions and growth factor expression cascades is recognized as a crucial frontier for development of improved disease treatments [22]. In this context the lens has broader relevance to human health and disease, as signalling pathways that operate in the lens also operate in many other cell types, including other ectodermal derivatives such as retinal and neural cells [23, 24]. An increased understanding of how lens signalling pathways, gene regulatory networks and cataract-associated genes are integrated will therefore define molecular mechanisms (potentially including protective mechanisms) that have more broad relevance to normal and disease development in other tissues. Progression of these studies will be aided by establishment of a predictive model – or transcriptional blueprint – that comprehensively integrates lens TRs and their TGs (including disease-associated genes) with the interconnected regulatory signalling pathways that control tissue development, growth, repair and regeneration.

To beginning addressing these issues, we analysed publically-available lens RNA-seq gene expression data and known protein-protein interactions (PPIs) to construct lens signalling pathways consisting of Rs, Ks and TRs. Correlating these R/K/TR paths with publically-available LEC TG sets [25] enabled establishment of a LEC transcriptional blueprint that predicts how expression of cataract-associated genes is regulated.

Assessment of 3 new gene regulatory networks that arose empirically from the LEC blueprint showed they involve critical lens TRs and known lens regulatory events. Moreover, these new

networks were predicted by the LEC blueprint to be regulated by important lens signalling pathways, and gene ontology (GO) analyses not only supported these predictions but also highlighted previously unknown roles for poorly understood LEC TRs. Further validation of the predictive capacity of the LEC transcriptional blueprint is provided through demonstration that the poorly understood TR, ELK1, is both expressed and phosphorylated in human pluripotent stem cell-derived ROR1⁺ LECs.

The LEC transcriptional blueprint and associated signalling and gene regulatory networks described here provide new insights into how the molecular circuitry required for normal lens function is integrated and regulated. Importantly, the LEC blueprint and networks shown here offer a large and diverse array of discrete and testable molecular hypotheses for use in defining the molecular mechanisms of lens and cataract formation.

2. Materials and methods

2.1 Acquiring transcriptional and PPI datasets

All datasets used in this study were downloaded from the following public repositories: Rs were acquired from the Fantom5 project [26]; Ks were collected from Uniprot; TRs were obtained from a database of sequence-specific DNA-binding proteins identified by GO-based annotation [27]; experimentally-determined PPIs were obtained from the STRING database (version 10) [ref]; mouse and human RNA-seq gene expression datasets from various cells and tissues were collected from the ENCODE project [28, 29] with additional LEC expression profiling data [30, 31] obtained from the Gene Expression Omnibus; and LEC TG data was collected from <http://regulatorycircuits.org/> [25].

2.2 Identification of house-keeping genes

Lists of known Rs, Ks and TRs were obtained as described above. To identify ‘housekeeping’ Rs, Ks and TRs common to many cell types, 144 human and 94 mouse cell and tissue RNA-seq gene expression datasets were obtained from ENCODE and grouped by species. These two groups of datasets were processed separately, with genes being designated as ‘housekeeping’ genes if expressed in at least 75% of all cell and/or tissue datasets within a species data-grouping. A combined list of

unique housekeeping genes was then generated by assuming one-to-one homology mapping between human and mouse genes.

2.3 Establishing a universe of known R/K/TR paths and pathways

Using the recently generated SPAGI algorithm [32] we identified a universe of all currently known, one-to-one PPIs for R, K and TR proteins expressed in any cell type. This process was performed separately for both mouse and human Rs, Ks and TRs. Scores were obtained for each PPI via STRING [33], with both directions of PPI being assessed. Scores of 700 or above were deemed as significant and these PPIs were kept. Where more than one PPI score was found between two proteins, the highest score was kept. The resulting mouse and human PPIs were compared, and where any particular PPI was identified in both species the larger PPI score was kept. Paths were then generated through the SPAGI method [32] by supplying possible R and TR combinations to the Dijkstra graph algorithm, together with all 1-to-1 PPIs identified by SPAGI. The highest scoring path for any R and TR combination was then obtained. Path lengths were permitted to range from 3 to 7 PPIs, consisting of: up to one R/R PPI; one R/K PPI; up to 4 consecutive K/K PPIs; and one K/TR PPI. The resulting collection of completed R/K/TR paths was sub-divided according to the R at the start of the path, with each R-defined collection of paths termed a pathway. The subset of paths consisting solely of housekeeping Rs, Ks and TRs was also identified using the list of housekeeping genes (described in per 2.2).

2.4 Identification and ranking of LEC-specific paths and pathways

To obtain all currently possible R/K/TR paths for LECs, RNA-seq data from p0 mouse LECs was used. Lists of LEC-expressed and LF cell-expressed genes were obtained using an expression cut-off threshold of $\log_2(RPKM+1)=3$ based on the expression distribution profile shown in Supplementary Figure S1. From the list of LEC-expressed genes, a sub-list of LEC-expressed Rs, Ks and TRs was created and used to search the above universe of R/K/TR paths (see 2.3) in order to identify: i) all possible LEC R/K/TR paths; ii) the subset of paths that contain only housekeeping Rs, Ks and TRs; and iii) the complementary subset of 'LEC-specific' paths that excluded any paths composed solely of

housekeeping Rs, Ks and TRs. From the resulting collection of lens-specific paths, each R-defined pathway was assigned an ‘activity score’ based on the expression levels of the Rs/Ks/TRs contained within the pathway (for details see Kabir et al. 2018) [32]. All lens-specific pathways were then ranked by plotting their activity scores against the total number of TRs contained with each pathway.

2.5 Identification of TG sets and cataract-associated genes for LEC-expressed TRs

A list of non-house-keeping LEC TRs was obtained from the LEC-specific paths (see 2.4). Publically-available LEC TG sets (see 2.1) were obtained for these non-house-keeping LEC TR, with different thresholds used depending on the application for which the TG sets were used (as described in the Results section). Comparison against the Cat-Map [7] database enabled the frequency of cataract-associated genes within these LEC TG sets to be calculated, both for the TG of each individual TR and also for the combined set of TGs for all the TRs.

2.6 Gene ontology and promoter analyses

GO analysis was performed using the David Gene Ontology Functional Annotation Clustering tool (<https://david.ncifcrf.gov/home.jsp>) [34, 35]. Promoter analyses were performed using the PASTAA web server [36] to identify transcription factor binding sites within +/- 400 base pairs of the transcription start sites of lens-expressed genes.

2.7 Cell culture

CA1 human pluripotent stem (PS) cells [37] were provided by A. Nagy and used as per approval from the Western Sydney University Human Research Ethics Committee. The PS cells were passaged as aggregates using 1 mg/mL dispase [38] before being plated in mTeSR1 (StemCell Technologies, Canada) onto Matrigel-coated plates (Corning, Australia). Populations of ROR1-expressing LECs were obtained by differentiating the human PS cells as described by Murphy et al [39].

2.8 Western blotting

Cultures of human PS cell-derived LECs were harvested for protein using a total protein lysis buffer (25 mM Tris, 150 mM NaCl, 1 mM EDTA, 1% Triton-X 100, 1 mM Na Vanadate, 1 mM PMSF, 5 µg Aprotinin, X Protease Inhibitor, pH 7.4). Protein concentration was determined using the EZQ Protein Quantification kit (Thermo Fisher Scientific): 25 µg protein per lane was separated via SDS-PAGE and transferred onto 0.2 µm PVDF membrane (Merck, Massachusetts, USA) using 120V for 1 hour at 4°C. Membranes were incubated with 2 µg anti-ELK1 antibody (Abcam, Cambridge, UK) or 2 µg anti-phospho-ELK1 antibody (Abcam) at 4 °C overnight. Membranes were then probed with horseradish peroxidase (HRP)-conjugated secondary antibodies for 1 hour at room temperature, before being visualised using a Luminata Crescendo Western HRP Substrate (Merck).

3. Results and discussion

3.1 Lens signalling pathways defined through gene expression and PPIs

Establishment of a testable transcriptional blueprint for growth factor-mediated control of LEC gene expression requires detailed and high-confidence knowledge of PPIs for lens signalling pathway molecules. Toward this end we first generated lists of all known mouse and human Rs, Ks and TRs (expressed in any tissue) and then identified all currently known one-to-one PPIs between these 2,137 signal pathway proteins. This approach identified 44,672 one-to-one PPIs: 19,104 identified via the mouse STRING data, and 25,568 via the human STRING data (Table 1). This included 16,452 PPIs common to both the mouse and human data, and 28,220 PPIs unique to either the mouse or human data. These 44,672 one-to-one PPIs were then used to identify a universe of currently possible complete R to K to TR paths (74,856). After removal of paths that involve only house-keeping proteins, 63,629 individual paths remained constituting 464 R-defined pathways that are representative of all currently known signalling pathways that could operate in any known cell type.

We next used high-confidence lens transcriptional profiles to determine the set of R/K/TR paths that could operate in LECs. Published mouse LEC RNA-seq data [30] was analysed to determine the Rs, Ks, and TRs expressed at postnatal day 0 (534 in total: 253 Rs, 126 Ks, 155 TRs). This included

known lens Rs (e.g., Fgfr1-4, Bmprs, etc.), Ks (e.g., Mapk1, Pik3ca, Pik3cg, etc.) and TRs (e.g., Pax6, Prox1, Oct1, etc.).

By comparison with the universe of possible signalling pathways constructed above, the LEC-expressed signalling molecules were shown to form 30,201 individual paths (253 pathways). Paths involving only house-keeping genes were then removed, leaving 20,245 lens-specific paths that could be condensed into 253 R-defined pathways (Table 2) involving a total of 155 TRs. Similar analysis of the LF cell gene expression data revealed 12,250 LF cell paths (186 pathways) involving 144 TRs (Table 3). This comprehensive database of lens signalling pathways enables assessment of lens signalling cascades via both ‘top-down’ (i.e., from R to TRs) or ‘bottom-up’ (i.e., from TR to Rs) analyses (Fig. 1), with the ability to identify Ks and TRs that are common to or distinct between different lens signalling pathways.

Table 1: Summary of the known signal pathway universe identified via SPAGI

	Mouse	Human
# R, K, TF	2,137	2,137
# R/K, K/K, K/TF interactions (known PPI score > 0)	11,353	18,603
# high-confidence (score >= 700) R/K, K/K, K/TF interactions	19,104	25,568
(assuming bi-directional interaction)		
# common interaction		16,452
# combined unique interaction		28,220
# high-confidence complete R/K/TF paths		74,856
# high-confidence complete R/K/TF paths without housekeeping gene paths		63,629
(# pathways)		(464)

Table 2: Summary of SPAGI-identified signalling pathways in mouse LECs

# LEC expressed all paths	30,201
(# pathways)	(253)
# LEC expressed all paths TFs	155
# LEC expressed specific paths (i.e., without housekeeping gene paths)	20,245
(# pathways)	(253)
# LEC expressed specific paths TFs	155

Table 3: Summary of SPAGI-identified signalling pathways in mouse LFs

# LF expressed all paths	19,321
(# pathways)	(186)
# LF expressed all paths TRs	144
# LF expressed specific paths (i.e., without housekeeping gene paths)	12,250
(# pathways)	(186)
# LF expressed specific paths TRs	144

3.2 Mapping TGs, including cataract-associated genes, to the LEC blueprint

To expand the above collection of 253 high-confidence, experimentally-determined, R-defined lens signalling pathways into a comprehensive transcriptional blueprint of lens biology, we next investigated publically-available LEC TG sets generated by Marbach et al [25] from the Fantom5 consortium cap analysis of gene expression (CAGE) data. These LEC TG sets map the TRs that are active in human LECs to gene transcripts (obtained via the Fantom5 consortium) [40, 41] that are expressed by human LECs. At present, only LEC TG data is available as LF cell data was not generated by the Fantom5 consortium [40, 41].

As the LEC TG data can be accessed based on a threshold value, we first examined how varying the magnitude of this threshold impacted on the size of the TG list obtained for LECs. As shown in Fig. 1C, increasing the threshold value reduced the total number of TGs obtained for the 155 LEC-expressed TRs. To determine the biological relevance of the TGs obtained, we assessed how many known cataract-associated genes were captured by these different TG access thresholds. Comparison of 311 cataract-associated genes obtained from the Cat-Map database [7] with the TG sets obtained via different threshold values revealed a gradual reduction in cataract genes captured as the threshold value was increased: from 248 at a threshold value of 0, to 11 at a threshold value of 0.5 (Fig. 1C and D). At the same time, the frequency of cataract-associated genes within the TG sets increased from ~1.6% (the background frequency in the genome) at a threshold value of 0 to almost 50% at a threshold value of 0.5 (Fig. 1D).

3.3 Ranking LEC pathways highlights pervasive as well as niche critical lens signalling pathways

To provide a large number of LEC TGs for subsequent analyses, while including known cataract-associated genes at a higher frequency than they are present in the genome, the LEC TG threshold value of 0.1 was used. This process captured 1,390 LEC TGs. Comparison of these TGs against both mouse [8, 30] and human [31] LEC gene expression data (to confirm expression of these TGs in other LEC gene expression datasets) revealed 63 cataract-associated genes (at a frequency of ~4.8%) that are regulated by one or more of 64 LEC-expressed TRs. We next ranked the 253 LEC-specific signalling pathways by plotting their ‘activity score’ (a measure derived from the proportion of highly expressed genes within any given path) [32] against the total number of TRs regulated by each pathway (Fig. 2). This analysis revealed a trend in that pathways known to be important in LEC biology tended to: i) have relatively high activity scores (i.e., pathway members tended to be highly expressed); ii) regulate large numbers of LEC-expressed TRs; and iii) regulate large numbers of known cataract-associated genes.

Signalling pathways identified through this analysis included pathways known to be involved in lens development and/or growth (for extensive reviews of these pathways as relates to lens biology see Zhang and Cvekl 2017 and Lovicu 2011) [9, 16]. Some examples of the LEC pathways identified

here include: the Fgf pathway (including Fgfr1 to Fgfr4) involved in LEC proliferation and differentiation to LF cells; the Bmp pathway involved in LEC proliferation and survival, and LF cell differentiation; the Wnt pathway (via Fzds) that regulates LEC adhesion, integrity and polarity; the Notch pathway that controls lens growth through LEC formation and LF cell differentiation; integrins required for early lens differentiation, LEC/LF cell adhesion and normal development, as well as lens capsule assembly; cadherins required for LEC polarity/adhesion/survival, and LF cell elongation; Ephs and Efns involved in cell adhesion, cell polarity and LF cell elongation/alignment; and the Tgf β pathway that provides an inhibitory signal in the pre-placodal region for proper lens growth, and is implicated in anterior subcapsular cataract and posterior capsule opacification.

Interestingly, the above trend (for critical lens pathways to have high activity scores, to regulate large number of TRs, and to regulate large numbers of cataract-associated genes) was not an invariable rule. For instance, normal levels of Lrp6 [42], Smo [43], Itgb1 [44-46] and Itga6 [47, 48] are required for normal lens development. While these pathways have high activity scores, they are shown here to involve relatively few LEC TRs and to regulate few known cataract-associated genes (Fig. 3). These data are consistent with the idea that while a few signalling pathways may control large parts of lens behaviour, additional pathways likely perform critical niche functions that may not always be predicted based on expression levels or breadth of transcriptional footprint. Thus, the LEC transcriptional blueprint enables a new way to visualise molecular hypotheses that describe pervasive and niche critical contributions to lens biology.

3.4 Identification of overlapping and potentially niche roles for LEC signalling pathways

The above LEC transcriptional blueprint allows assessment of the specific molecular nodes that are either common to multiple pathways or unique to individual pathways at the level of Rs, Ks, TRs and TGs. To begin examining the interconnectivity between different LEC signalling pathways, comparison of the Fgfr and Pdgfr pathways was performed.

Fgfr pathways were chosen as Fgf signalling is critically-required for lens development [9, 19]. Analysis of the LEC transcriptional blueprint revealed that, individually, the Fgfr1 to Fgfr4 paths are quite highly expressed (Fig. 2). These Fgfr pathways individually regulate different numbers of TGs

and cataract-associated genes. With a LEC TG threshold of 0.1, the Fgfrs collectively are seen to regulate 142 LEC TRs (including Atf, Ets, Etv, Jun, Lef1, Myb, Myc, Mycn, Pax6, Pou2f1, Rara, Smad7) and 63 cataract-associated genes.

In comparison, in the normal lens Pdgfr signalling is generally considered to regulate LEC proliferation, though in vitro it can also act as a potentiator of LF cell differentiation [9]. Analysis of the LEC blueprint showed that Pdgfr pathway members are quite highly expressed, and that with a LEC TG threshold of 0.1 the Pdgfr paths regulate 123 LEC TRs including known TRs (e.g., Atf, Ets, Etv, Jun, Lef1, Myb, Myc, Mycn, Pax6, Rara, Smad1, Smad, Yap) and relatively unknown lens TRs (e.g., Elk1). Additionally, with a LEC TG threshold of 0.1 the Pdgfr pathway was also seen to regulate all 63 cataract-associated genes.

Further comparison of the combined Pdgfr and combined Fgfr paths indicated that 49 Ks and 139 TRs are common to both pathways; this represents 92.7% of the Pdgfr TRs and 97.9% of the Fgfr TRs. For the remaining 7.3% of Pdgfr TRs - i.e., those not shared with Fgfr pathways (e.g., Elk1, Ets1, Smad1, Smad3, Stat5a, Stat5b, Yap1) - GO analysis showed some of their TGs are involved in glycoprotein biosynthesis (but not lens fibre development), including Gcnt2, Pomgnt1, Vcan, each of which is a cataract associated gene. Conversely, GO analysis showed that some of the TGs for the 2.1% of Fgfr TRs not regulated by Pdgfr paths are involved in lens fibre development (but not glycoprotein biosynthesis), including Epha2 and Bfsp2, both of which are cataract-associated genes.

Overall, these data are consistent with Fgfr signalling, but not Pdgfr signalling, being a driver of LF cell differentiation, and with reports that show manipulation of either pathway can lead to loss of normal lens biology [9]. Given the large overlap in TG regulation by Pdgfr and Fgfr paths, it is possible that both pathways provide redundancy during lens development and growth. There may also be niche functions of these pathways resulting from the small subsets of genes regulated by the few TRs unique to both the Fgfr and Pdgfr pathways. More broadly, the Pdgfr data suggest that important in vivo roles may exist for a wider range of lens signalling pathways than is currently understood.

3.5 The LEC blueprint accurately predicts known roles for lens signalling pathways

To assess how accurately the LEC blueprint identifies functional roles for lens signalling pathways, the TGs for all TRs within particular pathways were grouped and analysed via GO. Encouragingly, this analysis identified GO categories that match known or suspected roles for these pathways in lens development (Table 4), indicating the LEC blueprint accurately reflects lens biology.

Table 4: GO analysis of TGs of pathways known to be involved in LECs.

Pathway (# paths)	Example GO categories for pathway TGs (raw $p < 0.05$; Benjamini $p < 0.05$)	# cataract genes
FGF (320)	GO:0040036 regulation of FGFR signalling	62
BMP (72)	GO:0071772 response to BMP	22
WNT (846)	GO:0016055 Wnt signalling pathway	61
PDGF (186)	GO: 0002009: morphogenesis of an epithelium	63
NOTCH (309)	GO:0008360 regulation of cell shape	56
EPH (1337)	GO:0070307 lens fiber cell development	63
EFN (855)	GO:1905114 R signaling pathway in cell-cell signaling	63
CDH (180)	GO:0043010 camera-type eye development	57
ITG (1728)	GO:001654 eye development	63

3.6 The LEC blueprint accurately predicts known LEC transcriptional regulation events

As the LEC blueprint shows lens signalling pathways regulate both multiple TRs and multiple cataract-associated genes (Fig. 2), the relationship between TRs and cataract-associated genes was further investigated. Using the 0.1 LEC TG threshold, 87 LEC TRs were captured (Fig. 3A) including

extensively studied lens TRs (such as Pax6 [49], Ctf [50, 51], Atf4 [52], Rxra [53], etc.) and other TRs with currently little or no known role in lens biology (such as Sp1 [54], Elk1, etc.).

A wide range was seen in the number of TGs regulated by each TR using the 0.1 LEC TG threshold, from 1 TG (e.g., EP300) to 79 TGs for ELK1 and 667 TGs for SP1. Analysis of cataract-associated genes within the TG sets showed that TRs with poorly understood current roles in lens development regulate varying numbers of cataract-associated genes. For example 35 cataract-associated genes are regulated by SP1 and 3 by ELK1, whereas 23 TRs do not regulate any cataract-associated genes (Fig. 3B).

Interestingly, for the TRs that regulate at least one cataract-associated gene, the frequency of cataract-associated genes per hundred TGs is inversely related to the total number of TGs (Fig. 3C). This suggests that some TRs may have quite niche functions in lens biology. For example, with the 0.1 LEC TG threshold, SP1 regulates the largest number of TGs but has the lowest frequency of cataract-associated genes (~4.9%, though this is still higher than the genome frequency of ~1.6% for cataract-associated genes). In contrast, EP300 is shown to regulate only a single TG, the cataract-associated gene CRYBB3 - indicating that EP300 has the highest frequency of cataract-associated genes (100%) at the 0.1 LEC TG threshold. This data is consistent with the published 100-fold loss of Crybb3 expression seen in CBP/Ep300 knockout mice [55]. Moreover, mutation or loss of Ep300 is associated with cataract in both mice and humans as seen in Rubinstein-Taybi syndrome [56]. These data suggest that the number of TGs a TR regulates does not necessarily indicate the importance of that TR to the lens. This is reinforced by the finding that 33 of the 63 cataract-associated genes are shown to be regulated by only 1 TR with the 0.1 LEC TG threshold.

3.7 Different lens TRs regulate specific subsets of cataract-associated genes

Examination of the classes of cataract-associated genes regulated by each LEC TR (Fig. 3D) showed that: 33 TRs regulate crystallin genes associated with cataract; 38 TRs regulate TRs that are themselves associated with cataract; 3 TRs regulate extracellular matrix proteins associated with cataract; 11 TRs regulate membrane proteins associated with cataract; and 38 TRs regulate other types of proteins associated with cataract. Individual cataract-associated genes varied in how many TRs

were involved in their regulation. For example, the 0.1 LEC TG threshold showed that the cataract-associated gene PAX6 was regulated by the most TRs (17) whereas 21 cataract-associated genes were regulated by only 1 TR each (Fig. 4A) - again suggesting some TRs have niche functions compared to other lens TRs.

As the LEC transcriptional blueprint shows that each LEC TR is regulated by more than one pathway, each cataract-associated gene is regulated by more than one pathway (Fig. 4B). Importantly, the top-down and bottom-up data analysis approaches provided by the LEC blueprint enable identification of the specific Rs, Ks and TRs that regulate specific (e.g., cataract-associated) LEC genes, as well as identification of key interconnections between signalling pathways via particular Ks or TRs.

3.8 A new, large LEC gene regulatory network involving known critical lens regulators

As normal lens development requires precise transcriptional regulation of genes associated with cataract formation, we examined whether any gene regulatory network is formed by the 64 LEC TRs and 63 cataract-associated genes captured via the 0.1 human LEC TG threshold. To assist interpretation of the resulting interconnected LEC gene regulatory network, genes were colour-coded based on whether their expression is higher in LECs, higher in LF cells, or similar in LECs and LF cells. Additionally, to identify networks that were common to lens biology of multiple species, the 0.1 LEC TGs were compared against both mouse [8, 30] and human [31] LEC gene expression data.

The resulting LEC gene regulatory network for 63 cataract-associated genes (Fig. 5A) involves many known lens TRs including Pax6, Sox2, Prox1, Maf, Myb, Atf4/Creb2, Pou2f1/Oct1, Smads, Otx2, Myc and Ep300. Within the network Sp1, Pax6 and Sox2 are identified as key nodes: Sp1 is shown to regulate key lens TRs (e.g., Prox1, Pax6, Sox2) [57, 58] as well as crystallins and other genes; Pax6 is shown to be regulated by 17 different TRs; and Sox2 is regulated by 14 TRs.

The network also identifies known gene regulatory interactions, including: regulation of Pax6 by Oct1 (Oct1^{-/-}, Sox2^{+/-} mice lack both Pax6 expression and lens placode induction [59]); regulation of Pax6 by ETS TRs (there are ETS binding sites in the Pax6 promoter [16]); and EP300 regulation of Crybb3 (Wolf et al show a 100-fold decrease in Crybb3 in CBP/p300^{-/-} mice [55]).

3.9 *A higher-level transcriptional network of 10 TRs predicted to control LECs*

The LEC TG data provides TG sets for all the 64 TRs shown in the network in Fig. 5A, thus enabling GO analyses to infer functional roles not just for subsets of TGs but also for the TR and signalling pathways that regulates these TGs.

To begin this process, we first examined whether any hierarchy existed among the 64 TRs in the large LEC cataract-associated gene network shown in Fig. 5A - i.e., whether a subset of the 64 TRs controlled expression of any of the LEC TRs and cataract-associated genes. This analysis revealed 27 TRs controlled expression of 9 of the 64 TRs and 51 of the 63 cataract-associated genes (Fig. 5B). Strikingly, Pax6 is at the centre of this higher-level network where it is regulated by 17 TRs: 3 TRs more highly expressed in LECs (Sp1, E2f4 and Ets1), 12 TRs similarly expressed in LECs and LF cells (including Elk1), and 3 TRs more highly expressed in LF cells (Atf4, Creb3 and Rxrg).

The graphical presentation of this higher-order network was then altered to show only the relative fold change in gene expression between LECs and LF cells for the 27 TRs (Fig.5C). As a result, the role of 10 TRs was highlighted, suggesting that: i) Pax6, Etx1, Creb3L1, Klf4, Sp1, Egr1 and E2f4 are important for regulating expression of the LEC network controlling expression of cataract-associated genes shown in Figure 5A; and ii) Atf4, Creb4 and Rxrg might play important roles in regulating this same network during differentiation to LF cells.

3.10 *The LEC blueprint ascribes relevant functional roles to lens TRs and signalling pathways*

Characterisation of the TGs for the 10 TRs involved in the higher-level network shown in Fig. 5C was performed using GO analyses. These GO analyses were coupled with bottom-up analyses using the SPAGI signalling pathway data in an attempt to identify the signalling pathways that regulate each of the 10 TRs. The GO categories shown in Figures 6 and 7 (and Supplementary Table S1) indicate that each of the 10 TRs in the higher-level gene regulatory network appear to perform specific functions within LECs that match known LEC biology (e.g., morphogenesis of a polarised epithelium). The TGs and cataract-associated genes regulated by this 10 TR higher-level network are shown in Supplementary Figures S2 and S3.

Further analysis of the TG GO categories for the 10 TRs shown in Figures 6 and 7 (and Supplementary Table S2) suggests that each of the TRs is regulated by different combinations of signalling pathways. Encouragingly, the GO-based pathway predictions support both the SPAGI-generated lens signalling pathway predictions and known lens biology, such as Wnt, Pdgf, Fgf, Egf, and Notch. Additionally, the mutual identification of Tgf β signalling by both the GO analysis and SPAGI signalling pathway analysis indicates underlying LEC networks that may be involved in development of posterior capsule opacification, and warrants more detailed investigation in future studies.

Noticeably smaller p-values were found for the GO terms arising from the TGs of the 7 TRs more highly expressed in LECs. The larger p-values associated with the 3 TRs more highly expressed in LF cells indicates CAGE data for LF cells should be obtained to better define the signalling and transcriptional networks that regulate LF cell differentiation. However, the current GO and SPAGI pathway analyses for these 3 TRs both highlight the role of Fgf and Wnt signalling in LF cell differentiation (Fig. 7 and Supplementary Table S2) - consistent with known roles for these pathways in LF cell production both in vivo and in vitro.

Taken together, the GO and SPAGI analyses identify a novel lens regulatory circuit centred around Pax6, Ets1, Creb3L1, Klf4, Sp1, Egr1 and E2f4 as important for establishing and/or maintaining the LEC phenotype, as well as Atf4, Creb4 and Rxrg as playing roles in driving or facilitating LEC differentiation to LF cells). These data also support the LEC transcriptional blueprint as a new and valuable resource for targeted molecular hypotheses to better define lens and cataract formation.

3.11 Additional lens gene regulatory networks remain to be discovered

It is worth noting that only ~20% (63 of 311) of the known cataract-associated genes are involved in the networks shown in Fig. 5A, and fewer in Fig. 5B and C. Clearly then, while the new LEC transcriptional networks shown here fit lens biology and are supported by the literature, they are not the only ones that operate in the lens. Lowering the LEC TG threshold to 0 increased the total number of TGs to 13,801 and the number of cataract-associated genes to 248 (after mapping against both mouse [8, 30] and human [31] LEC gene expression data to broadly confirm lens expression of the

TGs). Possible reasons why the remaining 63 (i.e., 311-248) cataract-associated genes may not be captured include species-specific differences, differences in the developmental timing of gene expression, expression in other tissues that affect lens transparency, and/ restricted expression to LF cells (that cannot be fully assessed due to the current lack of LF cell CAGE data). Future investigation of all 248 cataract-associated genes not directly addressed in the present study - and their accompanying signalling pathways - using the LEC transcriptional blueprint will further expand our understanding of the molecular circuitry controlling lens and cataract formation.

3.12 *ELK1: confirmation the LEC transcriptional blueprint identifies new lens biology*

The 3 novel LEC networks described in Figure 5 contain a large number of well-characterised LEC TRs as well as known transcriptional regulatory events. However, little is known of the roles played in the lens by some of the other TRs. For example, Elk1 was shown to regulate LEC TGs (Fig. 3A), including Pax6 (Fig. 5). In support of this finding, a role for Elk1 in the mouse lens has recently been suggested through promoter analyses of micro-dissected embryonic mouse LECs and LF cells [60]. To assess whether ELK1 might also be involved in human lenses we performed promoter analyses of published adult human lens gene expression data [31]. These analysis identified ELK1 binding sites within +/- 400 base pairs of the transcriptional start site for a large number of genes expressed by both LECs and LF cells (Fig. 8A). This is consistent with the gene expression data shown ELK1 is not differentially expressed between LECs and LF cells.

To assess whether ELK1 protein is expressed by human lens cells, we performed PCR and Western blot analysis using a recently described population of human (ROR1⁺) LECs obtained from pluripotent stem cells. The PCR analysis (not shown) demonstrated expression of *ELK1* mRNA, and the Western blot analysis confirmed expression of both ELK1 protein and phosphorylated ELK1 protein in the human LECs (Fig. 8B). These data suggest that more detailed investigation of lens signalling pathways and TGs related to Elk1 in mouse and/or human lens biology is warranted, and that ROR1⁺ human LECs and micro-lenses present a useful human system for these studies.

The data also demonstrate that the LEC transcriptional blueprint can be used to identify new biology relevant to both mouse and human lenses. Examples of additional important questions that can be isolated as testable molecular hypotheses using the LEC blueprint include:

- How is expression of β - and γ -crystallins regulated by Ets TRs, as indicated by Figure 5?
- Is BMP signalling involved in the Smad-mediated regulation of crystallin gene expression shown in Figure 3D?
- What are the molecular mechanisms by which Atf and Creb family members regulate lens development, given that Atf4 and Creb3 are shown to regulate Pax6 (Fig. 5A-C) and Creb3 is shown to regulate Maf, EphA2, Sox2 (Fig. 5A) and Creb3l1 is shown to regulate Cryba4 (Fig. 5A)?
- Do Sp1 and Rxrg interact to control expression of the cataract-associated gene Gfer in the lens (Fig. 5), given published reports of Sp1/Rxr interactions in non-lens cells [61, 62]?

4. Conclusion

The LEC transcriptional blueprint presented here (i.e., LEC signalling pathways and associated gene regulatory networks) is defined by known lens mRNA expression levels [30], experimentally-validated PPIs [33], empirically-determined transcript initiation sites [25], and known cataract-associated genes [7]. The LEC blueprint provides a comprehensive, integrated, multi-pathway framework for describing R-mediated transcriptional control of LEC behaviour. The R-mediated transcriptional networks presented here and derived from the LEC blueprint are consistent with known LEC biology accumulated over decades, as well as emerging molecular detail of critical lens TRs and cataract-associated genes. Demonstration that ELK1 is both expressed and phosphorylated in human ROR1 LECs - as predicted by the LEC blueprint - shows both that the blueprint enables hypothesis-driven interrogation of lens biology and that it can accurately predict new lens biology. The LEC blueprint thus provides a new and powerful tool for defining the molecular control of lens formation and growth, including investigation of niche functions specific to particular signalling pathways and/or TRs. The specificity of the molecular hypotheses within the LEC blueprint suggests it may also be applicable to investigation of the molecular events that occur during the initial stages of primary human cataract formation. As many of the signalling pathways and TGs contained within the

LEC transcriptional blueprint are involved in regulation of other tissues, particularly eye and neural tissues, further investigation of the LEC blueprint will provide a more detailed understanding of molecular events broadly relevant to human health.

Contributors

M.D.O'C. conceived the project. J.W.K.H. and M.D.O'C supervised the project. M.H.K. performed the bioinformatics analysis. P.M. and S.L. performed the cell culture and Western blotting. All authors revised and approved the manuscript.

Acknowledgements

M.H.K., P.M., and S.L. were supported by WSU Postgraduate Research Awards. M.D.O'C was supported by The Medical Advances Without Animals Trust. J.W.K.H is supported by a Career Development Fellowship from the National Health and Medical Research Council (1105271) and a Future Leader Fellowship from the National Heart Foundation of Australia (100848).

References

1. Coulombre JL, Coulombre AJ: Lens Development: Fiber Elongation and Lens Orientation. *Science* 1963, 142:1489-1490.
2. Spemann H: Anatomischen Anzeiger. 1901, 15:16-79.
3. H S: Neue Tatsachen zum Linsenproblemen. *Zool. Anz.* 1907, 31:379-386.
4. Tsonis PA, Del Rio-Tsonis K: Lens and retina regeneration: transdifferentiation, stem cells and clinical applications. *Exp Eye Res* 2004, 78:161-172.
5. Henry JJ, Hamilton PW: Diverse Evolutionary Origins and Mechanisms of Lens Regeneration. *Mol Biol Evol* 2018, 35:1563-1575.
6. Gwon AE, Gruber LJ, Mundwiler KE: A histologic study of lens regeneration in aphakic rabbits. *Invest Ophthalmol Vis Sci* 1990, 31:540-547.
7. Shiels A, Bennett TM, Hejtmancik JF: Cat-Map: putting cataract on the map. *Mol Vis* 2010, 16:2007-2015.

8. Kakrana A, Yang A, Anand D, Djordjevic D, Ramachandruni D, Singh A, Huang H, Ho JWK, Lachke SA: iSyTE 2.0: a database for expression-based gene discovery in the eye. *Nucleic Acids Res* 2018, 46:D875-D885.
9. Lovicu FJ, McAvoy JW, de Iongh RU: Understanding the role of growth factors in embryonic development: insights from the lens. *Philos Trans R Soc Lond B Biol Sci* 2011, 366:1204-1218.
10. Shiels A, Bennett TM, Knopf HL, Maraini G, Li A, Jiao X, Hejtmancik JF: The EPHA2 gene is associated with cataracts linked to chromosome 1p. *Mol Vis* 2008, 14:2042-2055.
11. Cooper MA, Son AI, Komlos D, Sun Y, Kleiman NJ, Zhou R: Loss of ephrin-A5 function disrupts lens fiber cell packing and leads to cataract. *Proc Natl Acad Sci U S A* 2008, 105:16620-16625.
12. Walker J, Menko AS: Integrins in lens development and disease. *Exp Eye Res* 2009, 88:216-225.
13. Logan CM, Rajakaruna S, Bowen C, Radice GL, Robinson ML, Menko AS: N-cadherin regulates signaling mechanisms required for lens fiber cell elongation and lens morphogenesis. *Dev Biol* 2017, 428:118-134.
14. Beyer EC, Berthoud VM: Connexin hemichannels in the lens. *Front Physiol* 2014, 5:20.
15. Gong X, Wang X, Han J, Niesman I, Huang Q, Horwitz J: Development of cataractous macrophthalmia in mice expressing an active MEK1 in the lens. *Invest Ophthalmol Vis Sci* 2001, 42:539-548.
16. Cvekl A, Zhang X: Signaling and Gene Regulatory Networks in Mammalian Lens Development. *Trends Genet* 2017, 33:677-702.
17. Schulz MW, Chamberlain CG, de Iongh RU, McAvoy JW: Acidic and basic FGF in ocular media and lens: implications for lens polarity and growth patterns. *Development* 1993, 118:117-126.
18. Cvekl A, Zhao Y, McGreal R, Xie Q, Gu X, Zheng D: Evolutionary Origins of Pax6 Control of Crystallin Genes. *Genome Biol Evol* 2017, 9:2075-2092.
19. Robinson ML: An essential role for FGF receptor signaling in lens development. *Semin Cell Dev Biol* 2006, 17:726-740.
20. Faber SC, Dimanlig P, Makarenkova HP, Shirke S, Ko K, Lang RA: Fgf receptor signaling plays a role in lens induction. *Development* 2001, 128:4425-4438.

21. Rajagopal R, Huang J, Dattilo LK, Kaartinen V, Mishina Y, Deng CX, Umans L, Zwijsen A, Roberts AB, Beebe DC: The type I BMP receptors, *Bmpr1a* and *Acvr1*, activate multiple signaling pathways to regulate lens formation. *Dev Biol* 2009, 335:305-316.
22. Berg J: Gene-environment interplay. *Science* 2016, 354:15.
23. Tanihara H IM, Honda Y: Growth factors and their receptors in the retina and pigment epithelium. *Prog Retin Eye Res* 1997, 16(2):271-301.
24. Frederikse P, Kasinathan C: Lens Biology is a Dimension of Neurobiology. *Neurochem Res* 2017, 42:933-942.
25. Marbach D, Lamparter D, Quon G, Kellis M, Kutalik Z, Bergmann S: Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat Methods* 2016, 13:366-370.
26. Ramilowski JA, Goldberg T, Harshbarger J, Kloppmann E, Lizio M, Satagopam VP, Itoh M, Kawaji H, Carninci P, Rost B, Forrest AR: A draft network of ligand-receptor-mediated multicellular signalling in human. *Nat Commun* 2015, 6:7866.
27. Tripathi S, Christie KR, Balakrishnan R, Huntley R, Hill DP, Thommesen L, Blake JA, Kuiper M, Laegreid A: Gene Ontology annotation of sequence-specific DNA binding transcription factors: setting the stage for a large-scale curation effort. *Database (Oxford)* 2013, 2013:bat062.
28. Consortium TEP: An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012, 489:57-74.
29. Consortium TME, Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, Groudine M, Bender M, Kaul R, et al: An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol* 2012, 13:418.
30. Hoang TV, Kumar PK, Sutharzan S, Tsonis PA, Liang C, Robinson ML: Comparative transcriptome analysis of epithelial and fiber cells in newborn mouse lenses with RNA sequencing. *Mol Vis* 2014, 20:1491-1517.
31. Hawse JR, DeAmicis-Tress C, Cowell TL, Kantorow M: Identification of global gene expression differences between human lens epithelial and cortical fiber cells reveals specific genes and their associated pathways important for specialized lens cell functions. *Mol Vis* 2005, 11:274-283.

32. Md Humayun Kabir RP, Joshua W. K. Ho, Michael D. O'Connor: Identification of active signaling pathways by integrating gene expression and protein interaction data. *BMC Systems Biology* 2018:In Press.
33. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, et al: STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 2015, 43:D447-452.
34. Huang da W, Sherman BT, Lempicki RA: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009, 4:44-57.
35. Huang da W, Sherman BT, Lempicki RA: Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 2009, 37:1-13.
36. Roider HG, Manke T, O'Keefe S, Vingron M, Haas SA: PASTAA: identifying transcription factors associated with sets of co-regulated genes. *Bioinformatics* 2009, 25:435-442.
37. International Stem Cell I, Adewumi O, Aflatoonian B, Ahrlund-Richter L, Amit M, Andrews PW, Beighton G, Bello PA, Benvenisty N, Berry LS, et al: Characterization of human embryonic stem cell lines by the International Stem Cell Initiative. *Nat Biotechnol* 2007, 25:803-816.
38. O'Connor MD, Kardel MD, Iosfina I, Youssef D, Lu M, Li MM, Vercauteren S, Nagy A, Eaves CJ: Alkaline phosphatase-positive colony formation is a sensitive, specific, and quantitative indicator of undifferentiated human embryonic stem cells. *Stem Cells* 2008, 26:1109-1116.
39. Murphy P, Kabir MH, Srivastava T, Mason ME, Dewi CU, Lim S, Yang A, Djordjevic D, Killingsworth MC, Ho JWK, et al: Light-focusing human micro-lenses generated from pluripotent stem cells model lens development and drug-induced cataract in vitro. *Development* 2018, 145.
40. Consortium F, the RP, Clst, Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Haberle V, Lassmann T, et al: A promoter-level mammalian expression atlas. *Nature* 2014, 507:462-470.
41. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al: An atlas of active enhancers across human cell types and tissues. *Nature* 2014, 507:455-461.

42. Stump RJ, Ang S, Chen Y, von Bahr T, Lovicu FJ, Pinson K, de Iongh RU, Yamaguchi TP, Sassoon DA, McAvoy JW: A role for Wnt/beta-catenin signaling in lens epithelial differentiation. *Dev Biol* 2003, 259:48-61.
43. Choi JJ, Ting CT, Trogrlic L, Milevski SV, Familiar M, Martinez G, de Iongh RU: A role for smoothensin during murine lens and cornea development. *PLoS One* 2014, 9:e108037.
44. Wang Y, Terrell AM, Riggio BA, Anand D, Lachke SA, Duncan MK: beta1-Integrin Deletion From the Lens Activates Cellular Stress Responses Leading to Apoptosis and Fibrosis. *Invest Ophthalmol Vis Sci* 2017, 58:3896-3922.
45. Scheiblin DA, Gao J, Caplan JL, Simirskii VN, Czymmek KJ, Mathias RT, Duncan MK: Beta-1 integrin is important for the structural maintenance and homeostasis of differentiating fiber cells. *Int J Biochem Cell Biol* 2014, 50:132-145.
46. Pathania M, Wang Y, Simirskii VN, Duncan MK: beta1-integrin controls cell fate specification in early lens development. *Differentiation* 2016, 92:133-147.
47. Walker JL, Zhang L, Menko AS: A signaling role for the uncleaved form of alpha 6 integrin in differentiating lens fiber cells. *Dev Biol* 2002, 251:195-205.
48. Walker JL, Zhang L, Zhou J, Woolkalis MJ, Menko AS: Role for alpha 6 integrin during lens development: Evidence for signaling through IGF-1R and ERK. *Dev Dyn* 2002, 223:273-284.
49. Walther C, Gruss P: Pax-6, a murine paired box gene, is expressed in the developing CNS. *Development* 1991, 113:1435-1449.
50. Burke LJ, Hollemann T, Pieler T, Renkawitz R: Molecular cloning and expression of the chromatin insulator protein CTCF in *Xenopus laevis*. *Mech Dev* 2002, 113:95-98.
51. Li T, Lu Z, Lu L: Regulation of eye development by transcription control of CCCTC binding factor (CTCF). *J Biol Chem* 2004, 279:27575-27583.
52. Tanaka T, Tsujimura T, Takeda K, Sugihara A, Maekawa A, Terada N, Yoshida N, Akira S: Targeted disruption of ATF4 discloses its essential role in the formation of eye lens fibres. *Genes Cells* 1998, 3:801-810.

53. Kastner P, Grondona JM, Mark M, Gansmuller A, LeMeur M, Decimo D, Vonesch JL, Dolle P, Chambon P: Genetic analysis of RXR alpha developmental function: convergence of RXR and RAR signaling pathways in heart and eye morphogenesis. *Cell* 1994, 78:987-1003.
54. Liu X, Zhou P, Fan F, Li D, Wu J, Lu Y, Luo Y: CpG site methylation in CRYAA promoter affect transcription factor Sp1 binding in human lens epithelial cells. *BMC Ophthalmol* 2016, 16:141.
55. Wolf L, Harrison W, Huang J, Xie Q, Xiao N, Sun J, Kong L, Lachke SA, Kuracha MR, Govindarajan V, et al: Histone posttranslational modifications and cell fate determination: lens induction requires the lysine acetyltransferases CBP and p300. *Nucleic Acids Res* 2013, 41:10199-10214.
56. CA S: *Rubinstein-Taybi Syndrome*. University of Washington, Seattle; 2002.
57. Muta M, Kamachi Y, Yoshimoto A, Higashi Y, Kondoh H: Distinct roles of SOX2, Pax6 and Maf transcription factors in the regulation of lens-specific delta1-crystallin enhancer. *Genes Cells* 2002, 7:791-805.
58. Duncan MK, Xie L, David LL, Robinson ML, Taube JR, Cui W, Reneker LW: Ectopic Pax6 expression disturbs lens fiber cell differentiation. *Invest Ophthalmol Vis Sci* 2004, 45:3589-3598.
59. Donner AL, Episkopou V, Maas RL: Sox2 and Pou2f1 interact to control lens and olfactory placode development. *Dev Biol* 2007, 303:784-799.
60. Zhao Y, Zheng D, Cvekl A: A comprehensive spatial-temporal transcriptomic analysis of differentiating nascent mouse lens epithelial and fiber cells. *Exp Eye Res* 2018, 175:56-72.
61. Shimada J, Suzuki Y, Kim SJ, Wang PC, Matsumura M, Kojima S: Transactivation via RAR/RXR-Sp1 interaction: characterization of binding between Sp1 and GC box motif. *Mol Endocrinol* 2001, 15:1677-1692.
62. Ohoka Y, Yokota-Nakatsuma A, Maeda N, Takeuchi H, Iwata M: Retinoic acid and GM-CSF coordinately induce retinal dehydrogenase 2 (RALDH2) expression through cooperation between the RAR/RXR complex and Sp1 in dendritic cells. *PLoS One* 2014, 9:e96512.

Fig. 1. SPAGI-generated LEC signalling paths and LEC TGs. A. Example of a ‘top-down’ analysis that identifies signalling cascades relating to a single R. B. Example of a ‘bottom-up’ analysis that identifies different signalling pathways that regulate a particular TR. C. Increasing the LEC TG threshold decreased both the number of known cataract-associated genes obtained (y-axis) and the total number of TGs obtained (numbers above open circles). D. Increasing the LEC TG threshold increased the frequency of known cataract associated genes captured (y-axis; absolute numbers of known cataract-associated genes captured at each threshold is indicated by the numbers above the open circles).

Fig. 2: Relationship between signalling pathway activity scores, TR numbers per pathway and number of cataract-associated genes per pathway. A. LEC pathways identified by the SPAGI pipeline and then mapped against the LEC TG data (0.1 threshold). A selection of signalling pathways known to be important for LEC biology are circled in red. Label colours indicate the number of cataract-associated genes regulated by the pathway (i.e., green: 0 to 20; blue: 21 to 40; orange: 41 to 60; dark red: ≥ 61).

Fig. 3. Assessment of 64 LEC TRs and their TGs. A. Bar-plot showing the number of TGs regulated by each of the 64 LEC TRs as shown by the 0.1 LEC TG threshold. B. The number of cataract-associated genes regulated by each of the 64 LEC TRs. C. The proportion of cataract-associated genes relative to the total number of TGs for each of the 64 LEC TRs. D. Classes of cataract-associated genes regulated by LEC TRs.

Fig. 4. The number of regulatory TRs and pathways for 63 cataract-associated genes. A. Bar-plot showing PAX6 is seen to be regulated by the largest number of TRs in the 0.1 LEC TG threshold data, whereas 21 cataract-associated genes are regulated by only 1 TR. B. Bar-plot (with the same x-axis arrangement as A) showing the number of pathways that regulate each of the 63 cataract-associated genes shown in A.

Fig. 5. Gene regulatory networks present in LECs. A. The gene regulatory network consisting of 64 LEC TRs and the 63 cataract-associated genes obtained using the 0.1 LEC TG threshold (yellow circles indicate known lens TRs; blue indicates higher expression in LECs; red indicates higher expression in LF cells; grey indicates similar expression in LECs and LF cells; black border indicates a known cataract-associated gene). B. A network of 27 TRs that regulate 9 of the 64 LEC TRs and 53 of the 63 cataract-associated genes shown in (A). C. A reduced version of the network shown in (B) obtained by showing the gene expression differences between LECs and LF cells (large expression difference is represented by large circles, small expression difference is represented by no or small circles).

Fig. 6. Gene regulatory network involving 7 TRs with higher expression in LECs. Mapping of GO terms - that arose from GO analysis of TGs regulated by the 7 TRs indicated - shows that particular GO terms are both unique to each TR and indicative of LEC functions. GO terms related to particular signalling pathways were also identified that correlate with the signalling pathways identified by the SPAGI analysis. (Legend for text in red boxes: *italicized text* for GO terms specific to that TR; black text for GO terms with Benjamini $p < 0.05$; grey text for GO terms with raw $p < 0.05$; **bold text** indicates where GO analysis matches SPAGI predictions)

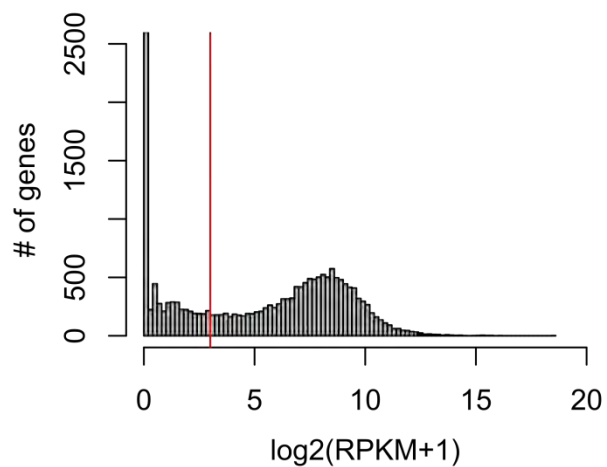
Fig. 7. Gene regulatory network involving 3 TRs with higher expression in LF cells. Mapping of GO terms - that arose from GO analysis of TGs regulated by the 3 TRs indicated - shows a GO term that is unique to ATF4 and a non-unique GO term for RXRG and CREB3. GO terms related to particular signalling pathways were also identified that correlate with the signalling pathways identified by the SPAGI analysis. (Legend for text in red boxes: *italicized text* for GO terms specific to that TR; black text for GO terms with Benjamini $p < 0.05$; grey text for GO terms with raw $p < 0.05$; **bold text** indicates where GO analysis matches SPAGI predictions)

Fig. 8. Analysis of ELK1 DNA-binding motifs and ELK1 protein expression in human lens cells. A. PASTAA-based promoter analyses show that ELK1 binding sites are present within +/-400 base

pairs of transcription start sites in human LEC and LF cell genes. B. Western blot analysis shows ELK1 (n = 2) and phospho-ELK1 (n = 2) protein are expressed in human ROR1⁺ LECs.

Supplementary Figure S1

A Lens epithelial cell data distribution



B Lens fiber cell data distribution

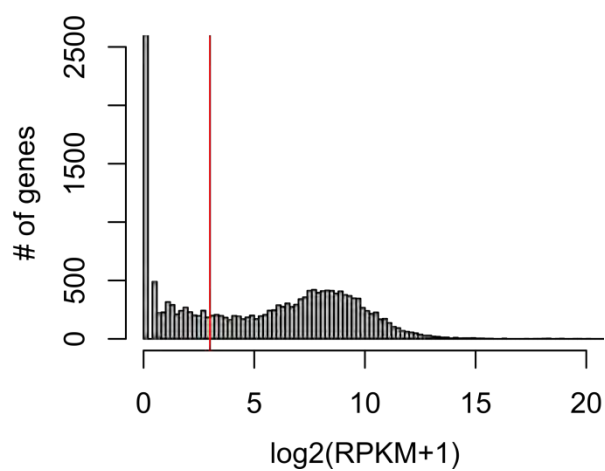


Figure 1

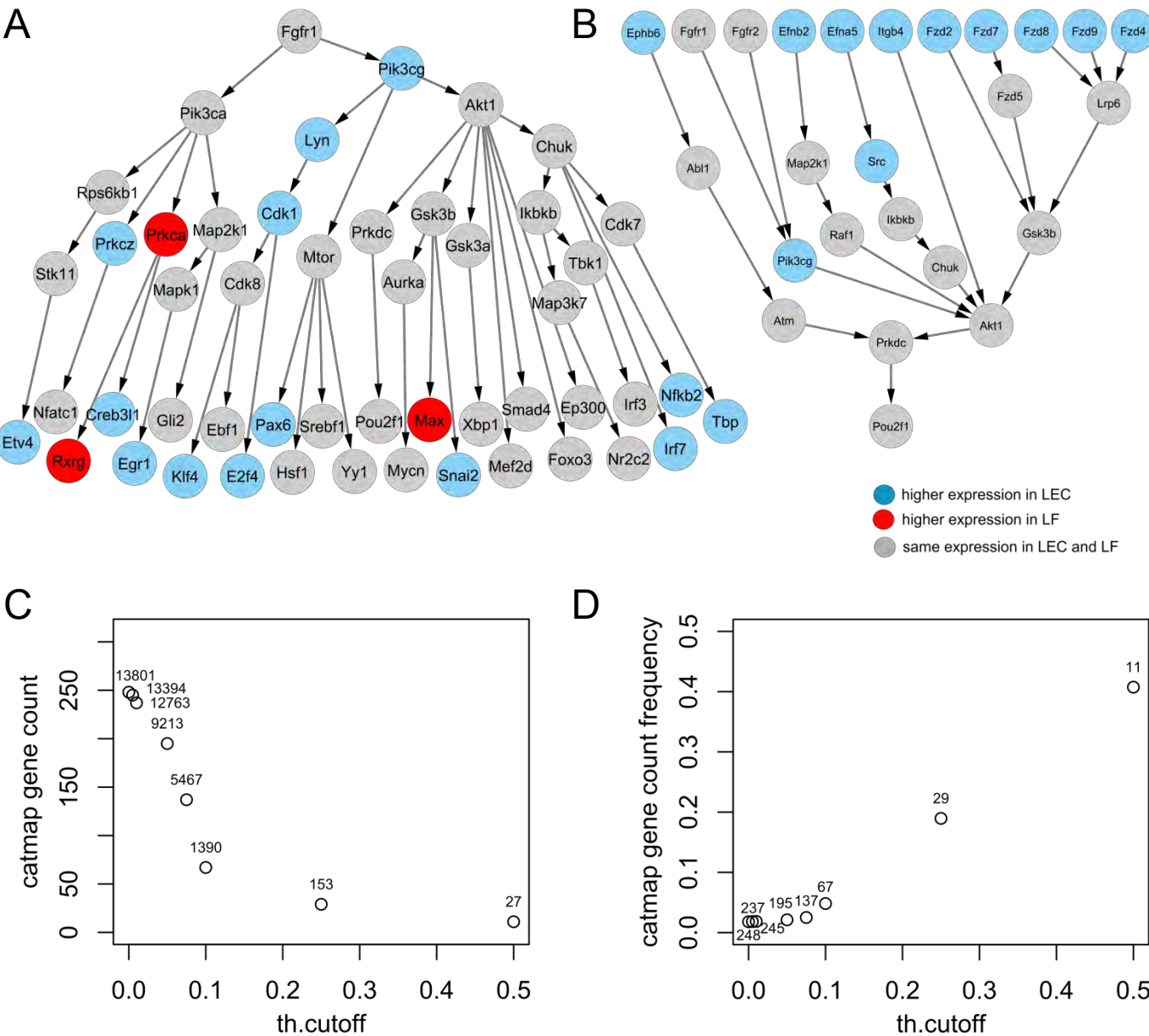


Figure 2

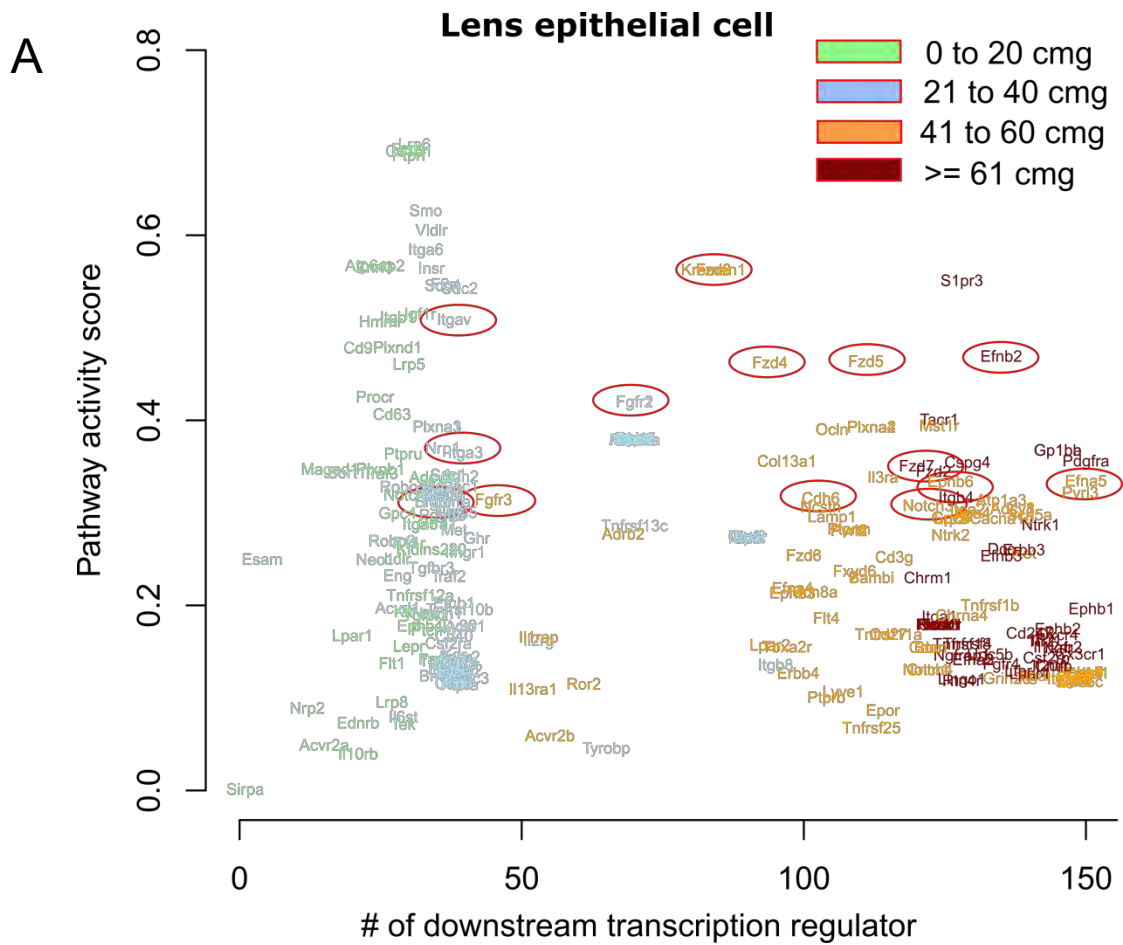


Figure 3

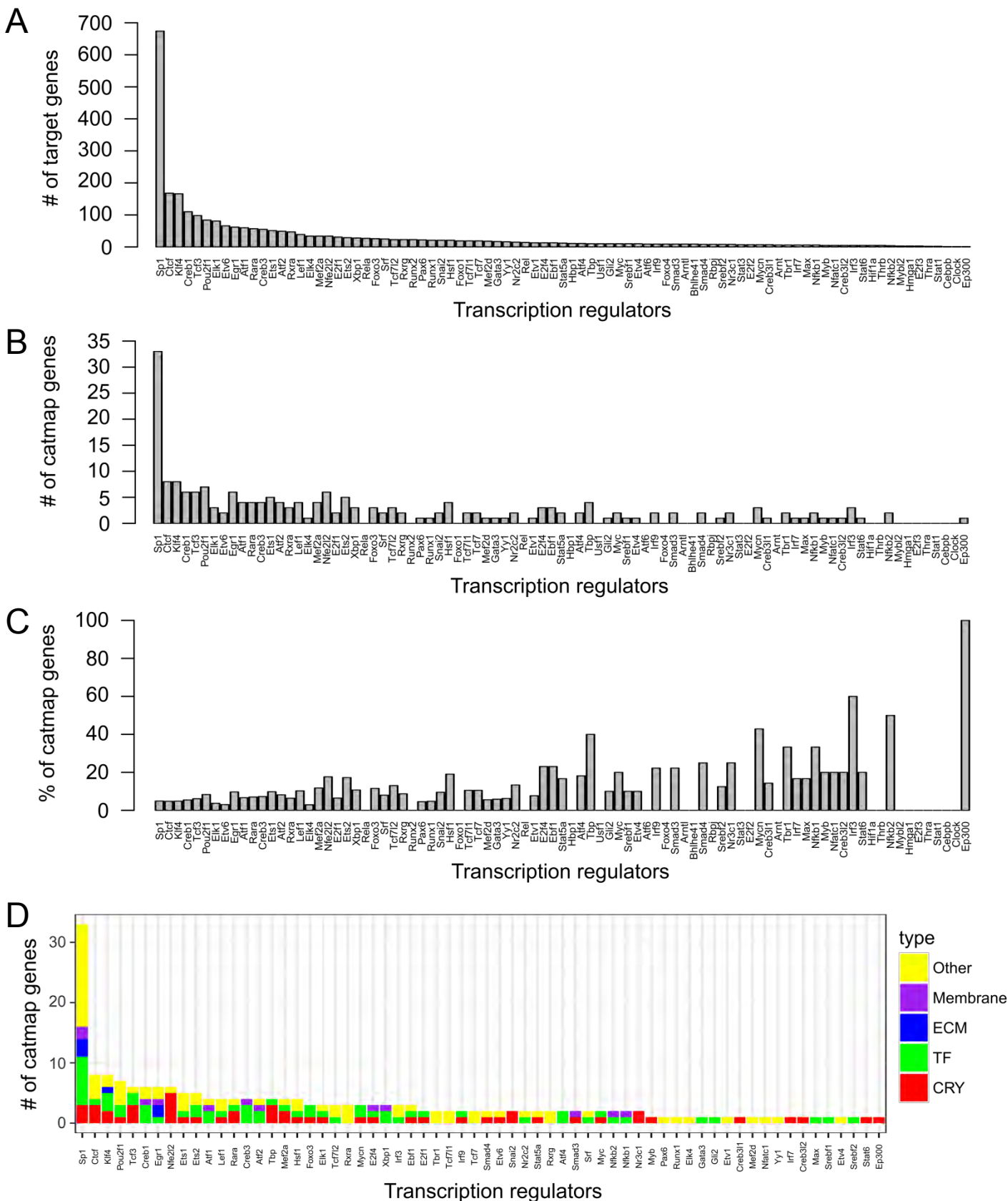


Figure 4

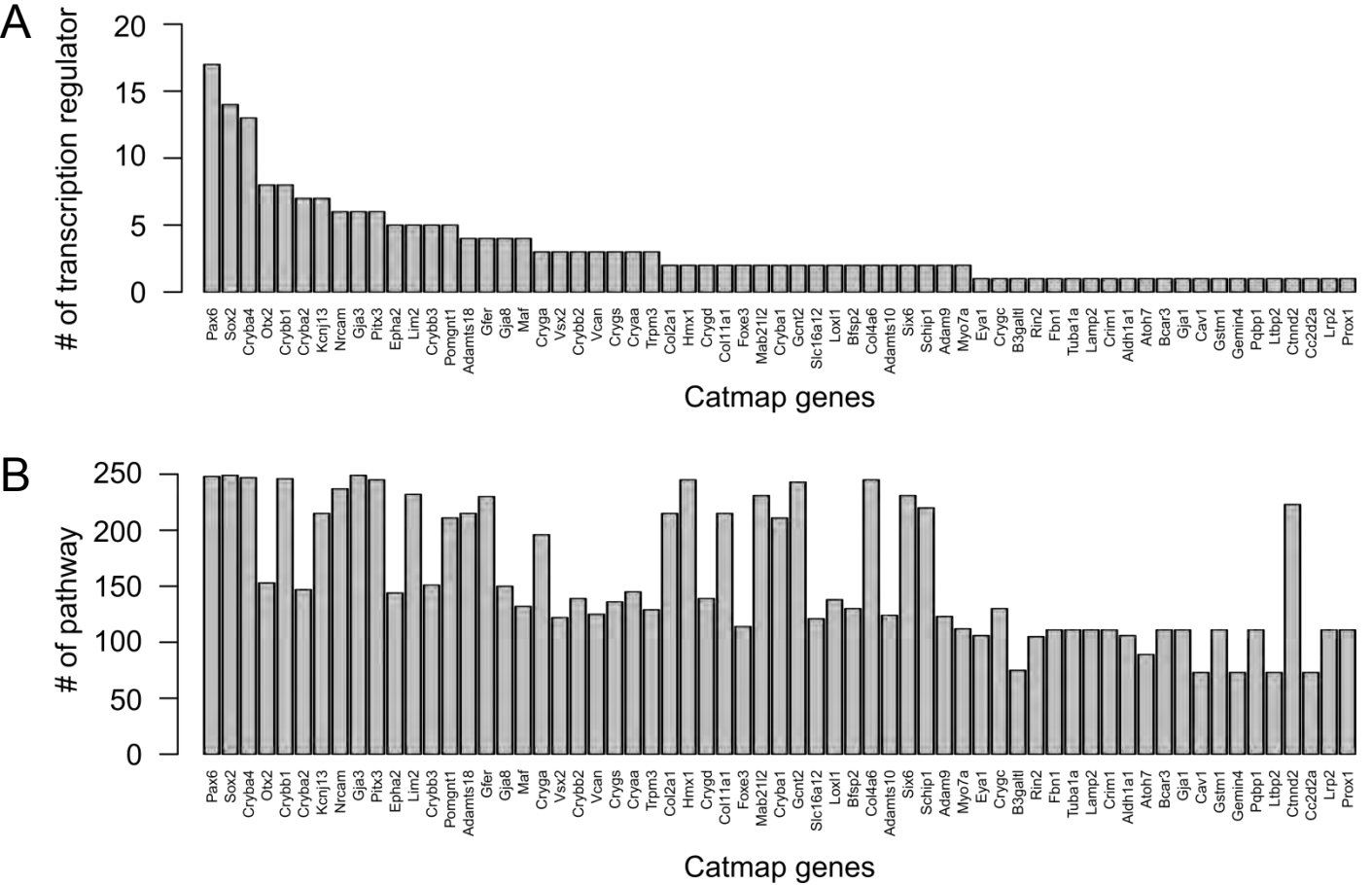


Figure 5

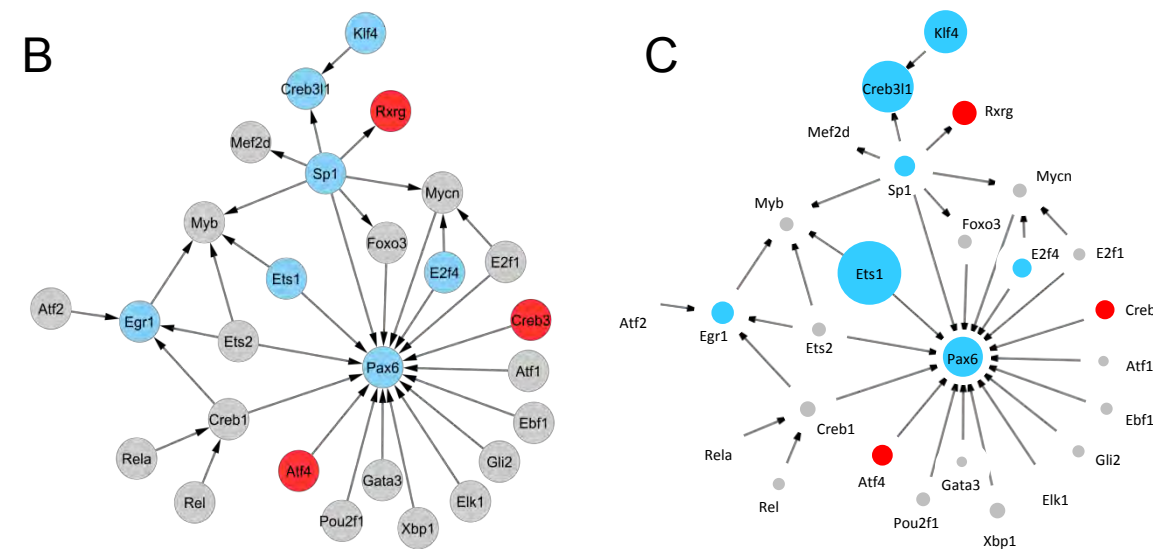
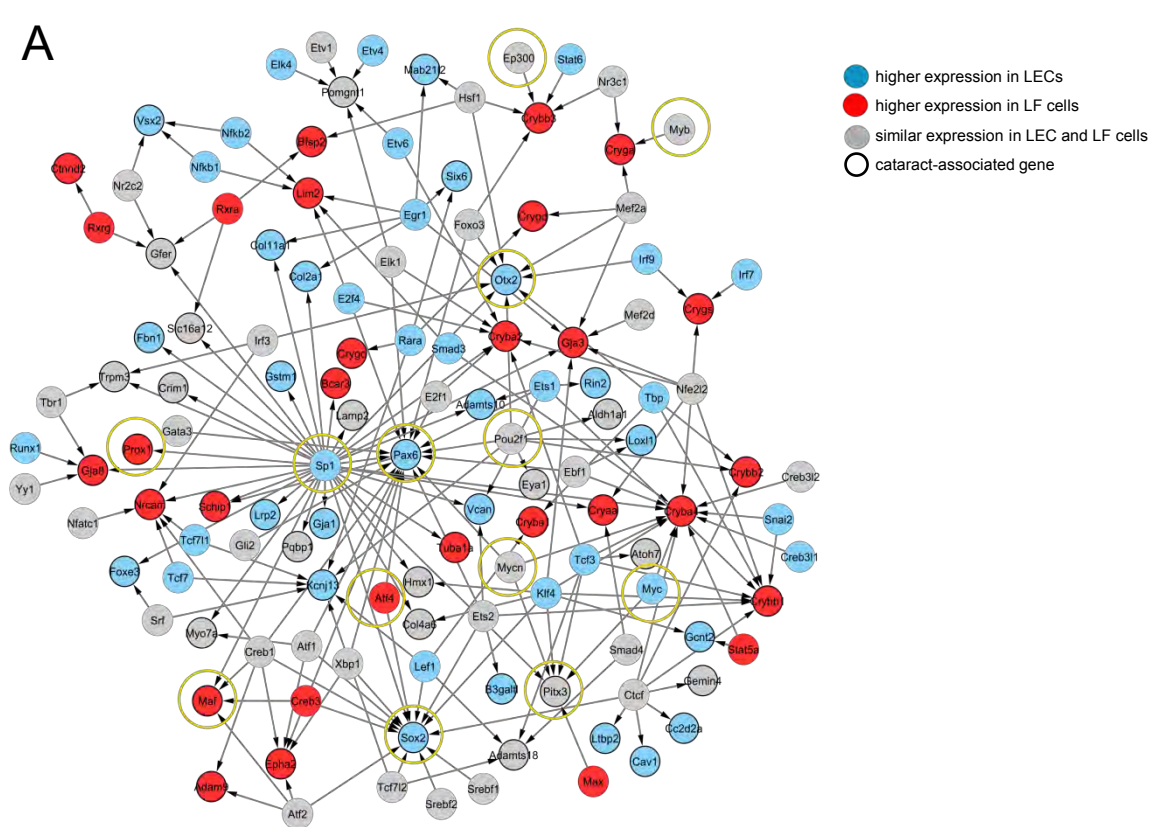


Figure 6

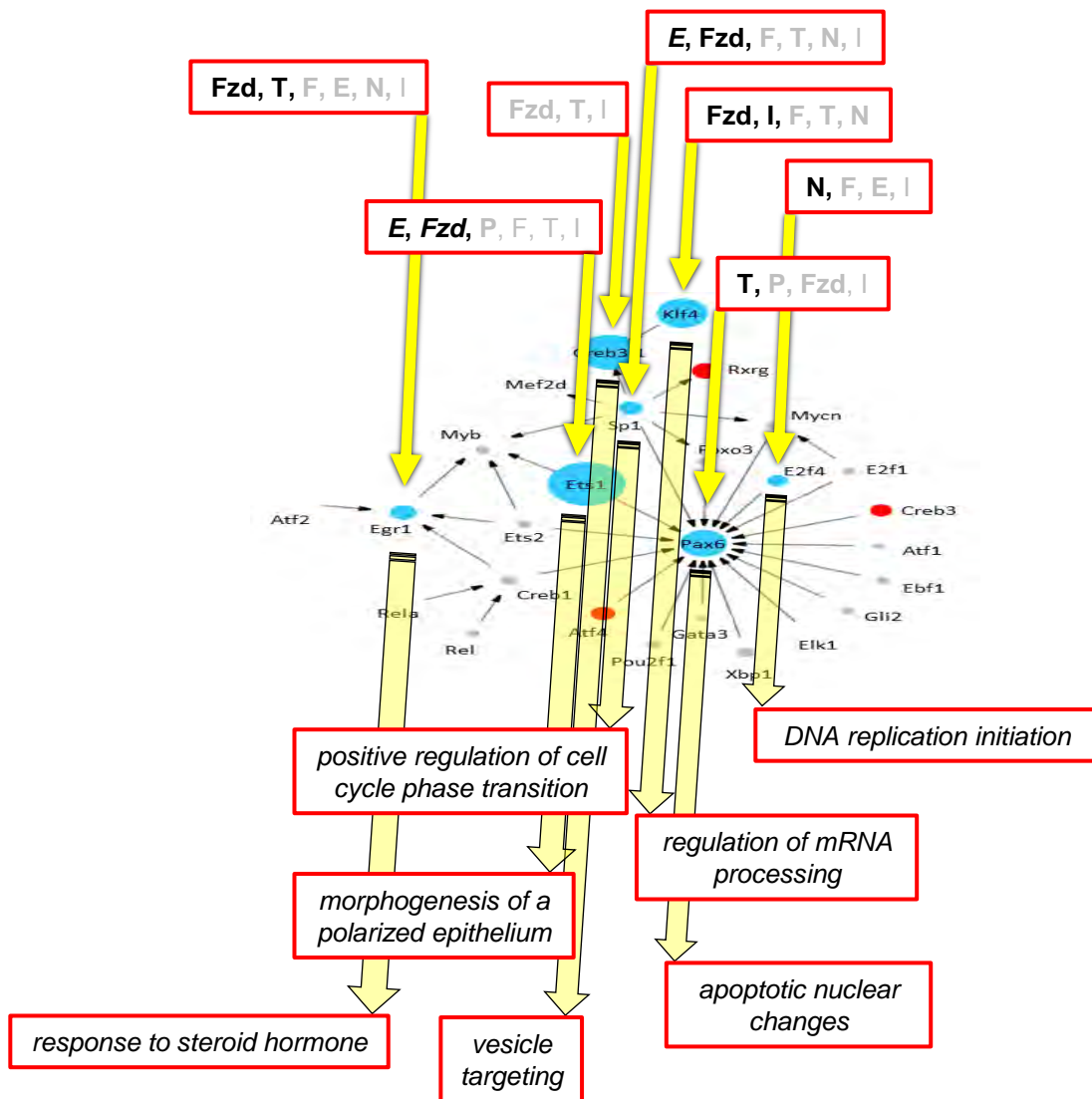
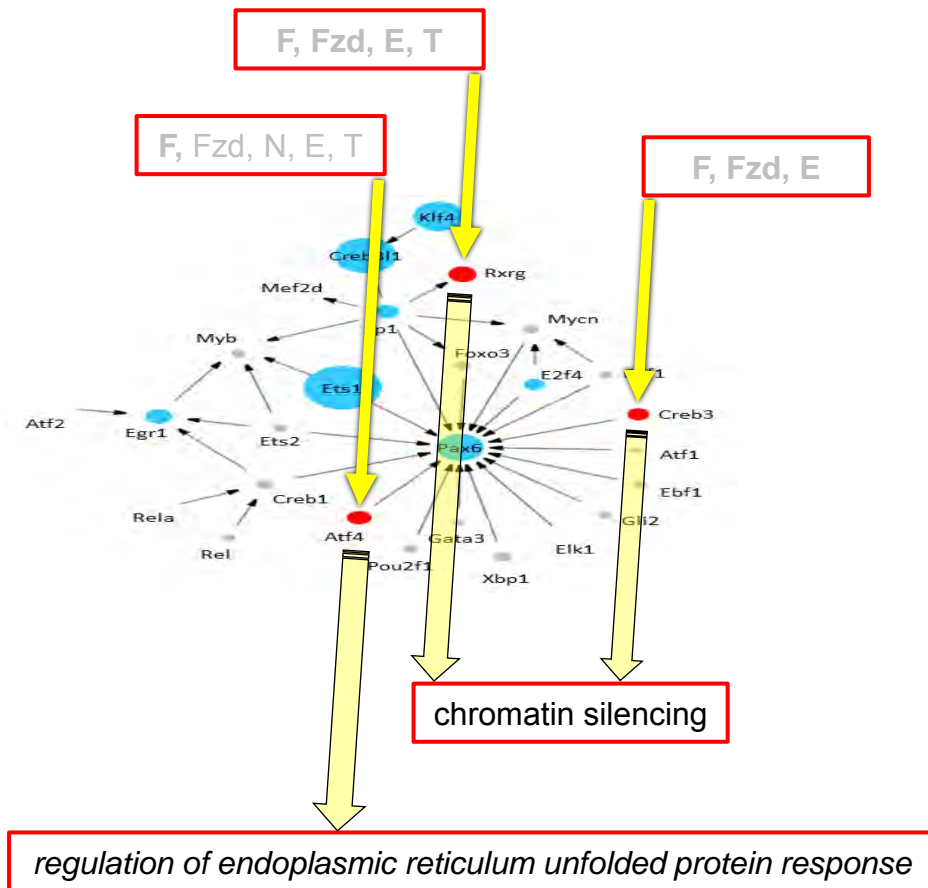
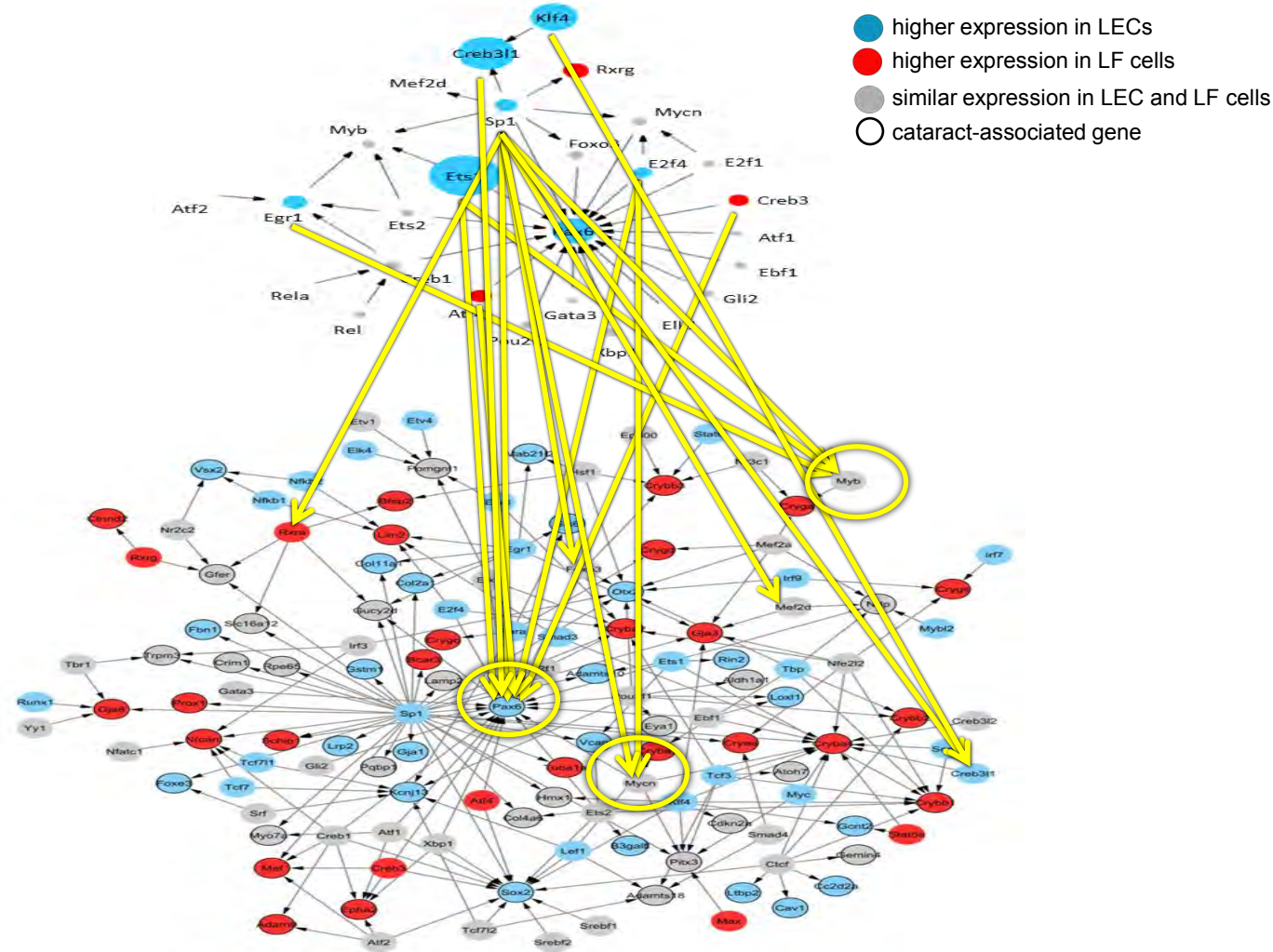


Figure 7



Supplementary Figure S2



Supplementary Figure S3

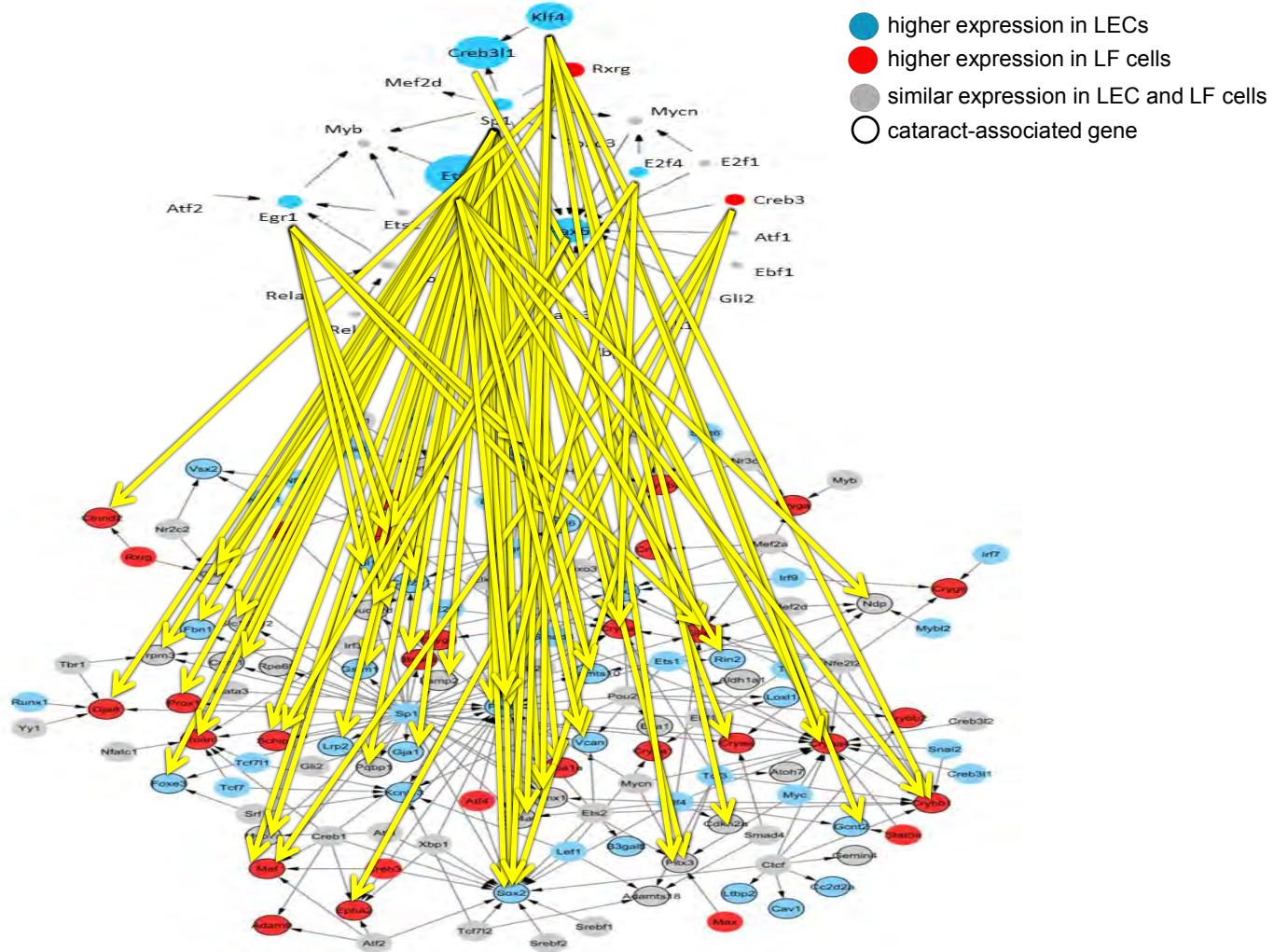
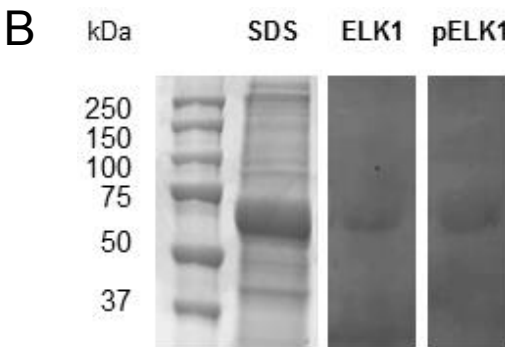
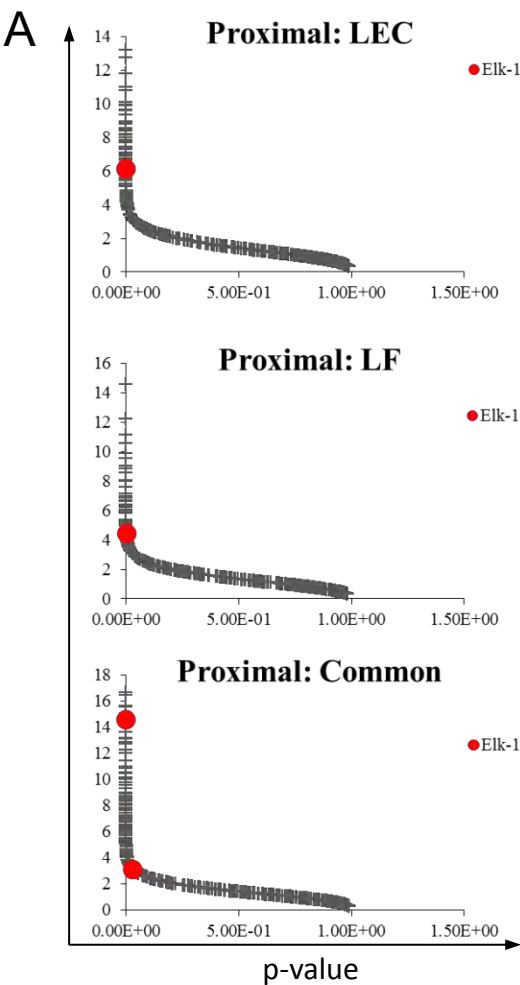


Figure 8



Supplementary Table S1. Selected GO terms arising from the TGs for the 10 TRs shown in Figure 6C.

TR	# cat. genes	# unique cat. genes	TG GO results excluding signalling pathways	raw p-value	Benjamini p-value
			Black text indicates Benjamini $p < 0.05$. <i>Italicised terms are specific to TGs of that TR.</i>		
Ets1	6	1	morphogenesis of a polarized epithelium establishment of planar polarity regulation of MAPK cascade regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle (+1 GO terms with Benjamini $p < 0.02$) establishment of spindle localization (+4 GO terms with Benjamini $p < 0.01$) cellular respiration	1.8E-03 1.1E-03 2.2E-04 7.9E-04 3.2E-05 1.4E-04	4.8E-02 3.2E-02 8.4E-03 2.5E-02 1.7E-03 5.9E-03
Pax6	2	0	apoptotic nuclear changes positive regulation of transcription, DNA-templated	5.9E-04 9.4E-04	3.3E-02 4.6E-02
E2f4	3	0	positive regulation of chromosome segregation (+1 GO term with Benjamini $p < 0.03$) DNA replication initiation DNA strand elongation regulation of glycolytic process (+1 GO term with Benjamini $p < 0.03$) response to hydrostatic pressure	3.1E-05 2.0E-06 1.6E-04 2.2E-04 1.3E-03	1.7E-03 1.3E-04 7.8E-03 9.8E-03 4.5E-02
Egr1	7	2	response to steroid hormone sterol biosynthetic process extrinsic apoptotic signaling pathway in absence of ligand regulation of establishment of protein localization to plasma membrane microtubule cytoskeleton organization	4.5E-05 7.3E-04 5.5E-04 1.7E-03 1.4E-03	2.2E-03 2.2E-02 1.8E-02 4.3E-02 3.7E-02
Sp1	35	16	positive regulation of cell cycle phase transition (+3 GO terms with Benjamini $p < 0.05$) regulation of cyclin-dependent protein kinase activity positive regulation of DNA biosynthetic process (+2 GO terms with Benjamini $p < 0.05$) regulation of gene expression (+4 GO terms with Benjamini $p < 1E-26$) RNA biosynthetic process (+3 GO terms with Benjamini $p < 1E-22$) positive regulation of cell morphogenesis in differentiation (+9 GO terms with Benjamini $p < 0.01$) adherens junction assembly TOR signaling (+1 GO term with Benjamini $p < 0.05$) ER-associated ubiquitin-dependent protein catabolic process mitochondrial outer membrane permeabilization involved in programmed cell death	1.8E-03 2.8E-03 1.6E-04 7.6E-34 9.7E-30 6.1E-07 2.3E-03 2.1E-04 2.9E-04 2.7E-05	3.4E-02 4.9E-02 4.3E-03 1.0E-30 7.7E-27 3.0E-05 4.2E-02 5.7E-03 7.5E-03 8.9E-04
Klf4	9	2	mitotic metaphase plate congression regulation of mRNA processing regulation of mRNA metabolic process intracellular protein transmembrane transport protein transmembrane transport	3.2E-05 1.1E-03 2.8E-04 6.2E-04 1.0E-03	2.0E-03 3.6E-02 1.2E-02 2.2E-02 3.3E-02
Creb3l1	1	0	vesicle targeting ER to Golgi vesicle-mediated transport vesicle targeting, rough ER to cis-Golgi Golgi vesicle transport	6.24E-06 8.62E-06 1.91E-05 2.14E-05	2.6E-02 1.2E-02 1.3E-02 1.3E-02
Atf4	2	0	negative regulation of response to endoplasmic reticulum stress (+1 GO term with Benjamini $p < 0.02$) positive regulation of ER-associated ubiquitin-dependent protein catabolic process negative regulation of endoplasmic reticulum stress-induced intrinsic apoptotic signaling pathway regulation of endoplasmic reticulum unfolded protein response regulation of endoplasmic reticulum stress-induced intrinsic apoptotic signaling pathway	1.5E-04 5.5E-04 5.8E-04 9.2E-04 1.2E-03	9.7E-03 2.6E-02 2.7E-02 4.0E-02 4.9E-02
Creb3	4	0	RNA metabolic process (+10 GO terms with Benjamini $p < 0.05$) chromatin silencing (+3 GO terms with Benjamini $p < 0.01$) chromatin assembly (+5 GO terms with Benjamini $p < 1E13$) regulation of transcription, DNA-templated (+9 GO terms with Benjamini $p < 1E-3$) telomere organization (+1 GO term with Benjamini $p < 1E-3$) beta-catenin-TCF complex assembly ER to Golgi vesicle-mediated transport Golgi vesicle transport vesicle targeting, rough ER to cis-Golgi	4.6E-22 2.0E-18 9.5E-17 9.4E-14 5.6E-11 1.5E-07 4.9E-06 7.1E-05 4.0E-04	1.1E-18 3.1E-15 5.9E-14 2.2E-11 9.2E-09 1.8E-05 4.9E-04 6.1E-03 2.9E-02
Rxrg	2	1	cellular macromolecule biosynthetic process (+8 GO terms with Benjamini $p < 0.05$) regulation of gene expression (+2 GO terms with Benjamini $p < 1E-3$) regulation of signal transduction (+9 GO terms with Benjamini $p < 0.05$) RNA biosynthetic process (+6 GO terms with Benjamini $p < 0.05$) negative regulation of cell death (+6 GO terms with Benjamini $p < 0.05$) positive regulation of cell communication (+1 GO terms with Benjamini $p < 0.02$) regulation of cell morphogenesis	1.2E-10 2.4E-08 1.2E-07 1.3E-07 3.1E-05 6.5E-05 1.8E-04	3.1E-07 3.1E-05 9.0E-05 8.5E-05 8.2E-03 1.3E-02 2.6E-02

Supplementary Table S2. Correlation of GO & LEC blueprint signal pathway predictions for 10TRs.

TR	TR is regulated by these LEC pathways	TG GO analysis results relating to signalling pathways Black text indicates Benjamini $p < 0.05$. <i>Italicised terms are specific to TGs of that TR.</i>	raw p-value	Benjamini p-value
Ets1	P, Fzd, C, N, A	<i>negative regulation of EGF R signaling pathway</i> <i>positive regulation of canonical Wnt signaling pathway</i> <i>positive regulation of Wnt signaling pathway</i> Wnt signaling pathway (+5 GO terms with Benjamini $p < 0.001$) regulation of EGF R signaling pathway response to TGF β (+1 GO term with Benjamini $p < 0.1$) response to FGF (+2 GO terms with Benjamini $p > 0.1$) response to PDGF response to EGF response to insulin (+2 GO terms with Benjamini $p > 0.1$)	2.3E-04 4.4E-04 1.1E-03 1.3E-07 2.3E-04 2.0E-03 2.2E-03 1.5E-02 2.2E-02 4.0E-02	8.8E-03 1.5E-02 3.2E-02 1.1E-05 3.9E-02 5.2E-02 5.7E-02 2.2E-01 2.8E-01 4.1E-01
Pax6	F, P, Fzd, N, C, B, A, T	response to TGF β (+1 GO term with Benjamini $p = 0.01$) negative regulation of TGF β R signaling pathway (+2 GO terms with Benjamini $p > 0.1$) response to PDGF regulation of Wnt signaling pathway (+7 GO terms with Benjamini $p > 0.1$) regulation of insulin R signaling pathway (+2 GO terms with Benjamini $p > 0.1$)	8.3E-05 9.4E-03 1.1E-02 1.7E-02 1.5E-02	1.0E-02 2.2E-01 2.4E-01 3.0E-01 3.1E-01
E2f4	F, P, Fzd, N	Notch signaling involved in heart development positive regulation of EGF R signaling pathway (+1 GO term with Benjamini $p > 0.1$) Notch signaling pathway positive regulation of FGF R signaling pathway positive regulation of insulin R signaling pathway	4.1E-03 6.3E-02 7.3E-02 8.1E-02 8.5E-02	1.1E-01 5.4E-01 5.8E-01 6.0E-01 6.2E-01
Egr1	F, P, Fzd, N, C, B, A, T	Wnt signaling pathway (+5 GO terms with Benjamini $p < 0.05$) response to TGF β (+1 GO term with Benjamini $p \leq 0.01$) negative regulation of EGF R signaling (+2 GO terms, 1 Benjamini $p < 0.1$, 1 $p > 0.1$) regulation of TGF β R signaling pathway (+3 GO terms with Benjamini $p > 0.1$) response to FGF response to insulin (+1 GO term with Benjamini $p > 0.1$) Notch signaling pathway (+2 GO terms with Benjamini $p > 0.1$)	8.9E-08 9.9E-05 2.3E-03 4.1E-03 8.8E-03 1.2E-02 1.6E-02	8.1E-06 4.1E-03 5.4E-02 8.5E-02 1.5E-01 1.8E-01 2.3E-01
Sp1	F, P, Fzd, N, C, A	<i>response to EGF (+1 GO term with Benjamini $p = 0.01$)</i> Wnt signaling pathway (+5 GO terms with Benjamini $p \leq 0.02$) response to TGF β (+1 GO term with Benjamini $p < 0.001$) response to insulin regulation of TGF β R signaling pathway (+5 GO terms with Benjamini $p > 0.1$) negative regulation of EGF R signaling pathway (+2 GO terms with Benjamini $p > 0.1$) positive regulation of Wnt signaling pathway (+1 GO term with Benjamini $p > 0.1$) regulation of Notch signaling pathway (+2 GO terms with Benjamini $p > 0.1$) regulation of insulin R signaling pathway (+1 GO term with Benjamini $p > 0.1$) response to FGF	5.6E-04 3.8E-15 7.8E-06 3.1E-04 3.1E-03 3.2E-03 3.3E-03 4.9E-03 3.9E-02 2.0E-02	1.3E-02 8.3E-13 2.9E-04 7.8E-03 5.3E-02 5.5E-02 5.5E-02 7.7E-02 3.6E-01 2.3E-01
Klf4	F, P, Fzd, N, C, B, T	Wnt signaling pathway (+1 GO term with Benjamini $p < 0.02$) response to insulin TGF β receptor signaling pathway (+1 GO term with Benjamini $p < 0.1$ and 3 with $p > 0.1$) regulation of Notch signaling pathway (+1 GO term with $p > 0.1$) regulation of Wnt signaling pathway (+2 GO terms with Benjamini $p > 0.1$) negative regulation of insulin R signaling pathway (+1 GO term with $p > 0.1$) response to EGF (+2 GO terms with Benjamini $p > 0.1$) response to FGF	4.3E-04 5.5E-04 2.1E-03 4.2E-03 4.8E-03 5.8E-03 9.7E-03 3.4E-02	1.6E-02 2.0E-02 6.1E-02 9.9E-02 1.1E-01 1.2E-01 1.8E-01 4.1E-01
Creb3 l1	F, P, Fzd, N, C, B, A, T	response to insulin Wnt signaling pathway (+5 GO categories with Benjamini $p > 0.1$) response to TGF β	9.3E-03 4.0E-02 5.8E-02	5.7E-01 7.4E-01 7.8E-01
Atf4	F	Wnt signaling pathway (+1 GO term with Benjamini $p < 0.1$ and 4 with $p > 0.1$) response to insulin response to FGF Notch signaling pathway (+1 GO term with Benjamini $p > 0.1$) response to EGF TGF β R signaling pathway	1.6E-03 3.0E-03 7.8E-03 3.3E-02 3.6E-02 7.2E-02	6.0E-02 9.9E-02 1.9E-01 4.0E-01 4.1E-01 5.6E-01
Creb3	F, P, Fzd, N	Wnt signaling pathway (+1 GO term with Benjamini $p < 0.1$ and 3 with $p > 0.1$) regulation of EGF R signaling pathway (+2 GO terms with Benjamini $p > 0.1$) response to FGF	8.3E-04 2.7E-02 7.4E-02	5.4E-02 5.5E-01 7.7E-01
Rxrg	F, P, Fzd, N, C, B, A, T	negative regulation of EGF R signaling pathway (+1 GO term with Benjamini $p > 0.1$) response to FGF (+1 GO term with Benjamini $p > 0.1$) TGF β R signaling pathway (+1 GO term with Benjamini $p > 0.1$) Wnt signaling pathway (+4 GO terms with Benjamini $p > 0.1$)	6.7E-04 2.6E-03 3.2E-03 3.7E-02	6.5E-02 1.7E-01 1.9E-01 5.7E-01

The SPAGI R package was applied here to get the signalling pathways for the lens epithelial cell data set. The results were then extended by incorporating other published data sets that were analysed by developing in-house R scripts. These results provide an interconnected, lens epithelial cell transcriptional blueprint of signalling pathways and associated target genes. Comparison of these target genes with the known cataract-associated genes identified three new gene regulatory networks and associated signal pathways predicted to control the networks.

Chapter 6

General discussion

General discussion

Summary of outcomes from this thesis

Four main outcomes have arisen through my research:

1. Development of a new, cross-species, compendium-based cell-type identification software, called C3. I demonstrated that C3 can accurately identify a cell type based on its gene expression profile through comparison to either a mouse or human compendium of gene expression profiles.
2. Development of a novel bioinformatics software for signalling pathway analysis, called SPAGI. I demonstrated that SPAGI can identify biologically-relevant signalling pathways using gene expression and protein-protein interaction (PPI) data.
3. Using the C3 method I showed that ROR1⁺ cells, derived from human pluripotent stem cells, molecularly and functionally resemble human lens epithelial cells (LECs).
4. Using the SPAGI method, together with data from the Fantom5 consortium, I generated a transcriptional blueprint for LECs and used it to identify both known and new interactions between key lens signalling pathways and their target genes.

Future work that can arise from these four outcomes is discussed below.

Advances in cell type identification

Compendium-based bioinformatics methods can provide a robust and objective approach to identify a cell type based on its gene expression profile. In these methods, the unknown sample's gene expression profile is used as a query profile against a large gene expression compendium consisting of many cell types. Most of the current methods implicitly assume there is a one-to-one correspondence between genes in the query and the compendium samples. This intrinsic assumption is violated when comparing data from different species, especially evolutionarily divergent organisms. Additionally, none of the current methods handle a cross-species query in a statistically rigorous fashion. Current methods are restricted to using a compendium generated from the same species as the query sample. For many model organisms, a compendium-based approach has been practically impossible as most publicly available data sets are only available for a small number of species (notably mouse and human).

Here, I showed that C3 accurately identified a range of cell types against compendia generated from different species. For example, C3 was applied to identify five different tissue types from 13 different species. It is expected that C3 will enable cell type identification for species other than mouse and human, thereby facilitating knowledge transfer from other species to understanding of human biology.

The ability of C3 to predict the identity of cells generated from pluripotent stem cells (i.e., ROR1⁺ LECs) was extensively validated by principal component analysis and a wide range of biological assays including *in vitro* regeneration of light-focusing lenses. Thus C3 offers a rigorous approach to validating cell types generated from stem cells, and therefore will be a useful tool for both research and industry/clinical applications in the stem cell field.

C3 is implemented as an open source R package to enable adoption and application of C3 by other groups. It should be noted though that the performance of C3 depends on the quality, variety and size of the compendium and whether or not a similar sample exists in the compendium. With 94 samples currently in the mouse compendium and 144 in the human compendium, a large variety of cell type identifications are currently possible. As more gene expression profiles are uploaded to public databases (e.g., GEO, ENCODE and GTEx), the breadth of identifications available to C3 will increase.

An interesting future research area is application of C3 to single cell RNA-seq (scRNA-seq) data. These datasets are increasingly gaining interest, and with this is a desire to identify/characterise individual cells within the single cell datasets. For this purpose, a compendium can be generated from either scRNA-seq data sets alone or by combining both bulk and scRNA-seq data sets. As scRNA-seq data sets consist of a large number of samples (typically tens of hundreds) some improvement of the method, such as improving the runtime and parallelisation, might be required to cope with both the query and compendium data sets. Application of C3 to scRNA-seq data could enable more accurate understanding of how individual cells contribute to the overall gene expression profiles within bulk-population gene expression data – thereby enabling better understanding of normal and or cancer cell behaviour (e.g., identification of cancer stem cells that may be present at a low frequency).

Advances in identification of active signalling pathways

Computational methods have been developed to identify the topological structures of signalling pathways using PPI data, but these methods are not designed for identifying active (i.e., biologically-relevant) signalling pathways from a gene expression profile. On the other

hand, there are statistical methods that use gene expression data to prioritize likely active signalling pathways. However, they typically only relate receptors and transcriptional factors and do not make full use of signalling pathway structures that link receptor, kinases and transcription factors. Additionally, most of the current methods were evaluated and applied to yeast PPI data, with only a few methods designed specifically to deal with the significantly greater complexity of mammalian data.

A generic signalling pathway database is available through the Kyoto Encyclopaedia of Genes and Genomes (KEGG), though this is of limited use for comprehensively obtaining a catalogue of cell type-specific signalling pathways. Commercial software are available (e.g., Ingenuity Pathway Analysis), however, the underlying assumptions used by these software are not fully disclosed due to the proprietary nature of the software. To our knowledge, the SPAGI method presented here is the only open-source method available that uses a gene expression profile to simultaneously and comprehensively identify an integrated set of candidate active (i.e., biologically-relevant) signalling pathways, including the likely pathway structures for every path (from receptor, through kinases, to transcriptional regulators).

The SPAGI method was used to generate a universe of possible signalling pathways based on known PPIs assuming one-to-one homology mapping of genes between human and mouse. This approach means that the SPAGI pathway universe can be continually updated as new PPI data is added to the STRING database. Cell type-specific pathway catalogues are then generated by using the list of cell-expressed receptors, kinases and transcriptional regulators. By ranking the output SPAGI pathways by expression level vs transcriptional regulators, SPAGI simultaneously predicts a set of candidate active/biologically-relevant signalling pathways together with their pathway structure from an input of a gene expression profile. Many of the highly ranked pathways from multiple cell types were pathways known to be biologically important. Additionally, the false positive identification rate for the ranked signalling pathways was low. However, given that mRNA expression levels do not guarantee protein expression or protein activity (e.g., phosphorylation status), it is not clear at this stage whether all the highly-ranked pathways are truly active. Further comparison of highly-ranked pathways against protein expression/protein activity data, and/or functional genomics analysis, would be useful to examine novel and highly-ranked pathways. As more human and mouse PPI data is uploaded to public databases, the SPAGI universe establishment process can be re-run, thereby improving the breadth of SPAGI predictions. Given the rate at which new data is added to the STRING PPI database, re-establishing the signalling pathway

universe could occur every 1 to 2 years. Similarly, as PPI data is obtained for proteins specific to other species, it will be possible to apply SPAGI to identify signalling pathways specific to other species.

Another worthy goal would be to convert the SPAGI method into a web tool or mobile phone app so that other researchers can perform SPAGI analysis on their datasets of interest. This is an attractive approach given that the SPAGI analysis performed for this thesis has already generated a universe of cell signalling paths suitable for use with other cell types. Additionally, as the Fantom5-based target gene analysis is available for 394 cell types (Marbach et al. 2016), there is a strong foundation for generating cell type-specific transcriptional blueprints as done here for LECs.

Future application of the SPAGI method to single cell RNA-seq data would be of interest, in order to assess the extent to which individual cells vary from the signalling pathway predictions obtained from the bulk expression profiling data. For example, optimizing the SPAGI approach for single cell gene expression profiling data could identify novel approaches for targeting low frequency cancer stem cells within a tumour population.

Finally, the SPAGI method could also be modified to identify cellular control mechanisms other than growth factor signalling pathways. For example, by utilizing PPIs involved in phagocytosis, cell division, cell death, migration, etc. This would require would be to change the input proteins (which are user-defined), and then deciding on starting and ending proteins (so that the method can build path for the defined starting and ending points).

A more detailed molecular understanding of lens epithelial cell biology

For my thesis, I developed and applied C3 and SPAGI to analyse gene expression data obtained from different populations of LECs including human pluripotent stem cell-derived ROR1⁺ LECs (Murphy et al. 2018), and published newborn mouse LECs (Hoang et al. 2014). These analyses were correlated with LEC target gene datasets derived through a published analysis of Fantom5 LEC data (Marbach et al. 2016), as well as a publicly available list of cataract-associated genes (Shiels et al. 2010). The resulting LEC transcriptional blueprint represents a comprehensive framework for investigating signalling and transcriptional control of LEC biology. Encouragingly, this blueprint was shown to include known key LEC transcriptional regulators (e.g., PAX6, SOX2, MYC, etc) and known transcriptional regulatory events (e.g., Ep300/Crybb3, Ets/Pax6 and Oct1/Pax6). Furthermore, promoter analyses and Western blotting support a role for ELK1 in human LEC biology as predicted

by the LEC transcriptional blueprint. The blueprint therefore provides a wealth of high-confidence, specific molecular hypotheses for functional genomics studies of lens and cataract development.

Interesting future bioinformatics-based extensions of this LEC blueprint include: making the data publically-accessible through a web interface (so that other lens researchers can use it to guide functional genomics studies); and expanding the blueprint to include lens fibre cells, (for example, through the generation of cap analysis of gene expression data for lens fibre cells).

Concluding remarks

Due to the advancement and availability of gene expression technologies, the generation of transcriptomic profile data for any given cell type is within reach for essentially any research group. Potential insights to be gained from this data include more detailed knowledge of normal and disease states, which in turn can lead to identification of better candidate treatments. This PhD thesis successfully developed and applied new gene expression analysis methods for cell type identification and the prediction of active/biologically-relevant signalling pathways. The utility of these new methods are supported by their application to identification of example cell types (including stem cell-derived LECs) and identification of new cell signalling/target gene networks in LECs. Thus this thesis has made strong contributions to the fields of bioinformatics, stem cell biology, and lens/cataract biology (as shown by the publications that have developed from each chapter). Equally importantly, the thesis results have led to clearly identifiable, interesting, relevant and achievable avenues for progressing the work contained within this thesis.

References

- Hoang TV, Kumar PK, Sutharzan S, Tsonis PA, Liang C, Robinson ML (2014) Comparative transcriptome analysis of epithelial and fiber cells in newborn mouse lenses with RNA sequencing *Mol Vis* 20:1491-1517
- Marbach D, Lamparter D, Quon G, Kellis M, Kutalik Z, Bergmann S (2016) Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases *Nat Methods* 13:366-370

- Murphy P et al. (2018) Light-focusing human micro-lenses generated from pluripotent stem cells model lens development and drug-induced cataract in vitro *Development* 145
- Shiels A, Bennett TM, Hejtmancik JF (2010) Cat-Map: putting cataract on the map *Mol Vis* 16:2007-2015

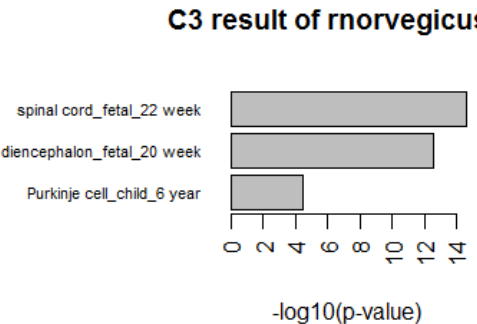
Appendix A

Supplementary material for Chapter 2

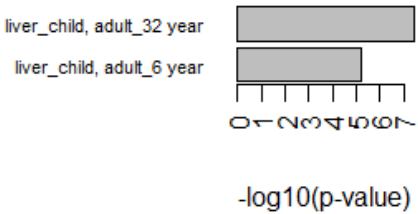
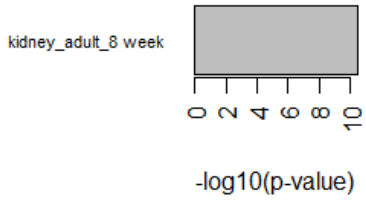
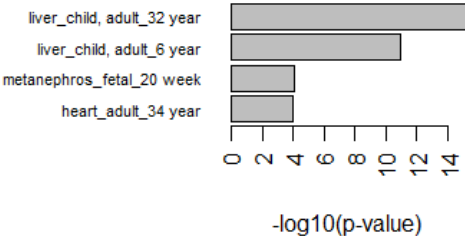
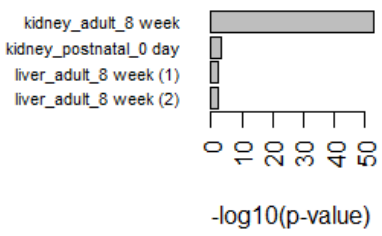
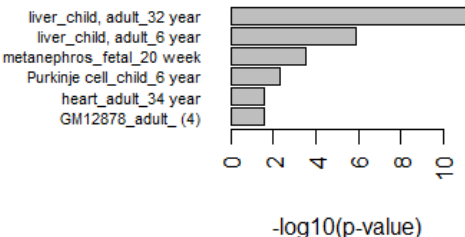
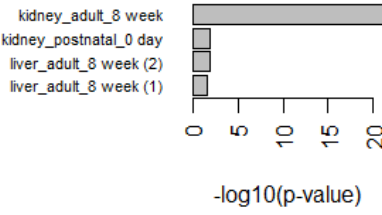
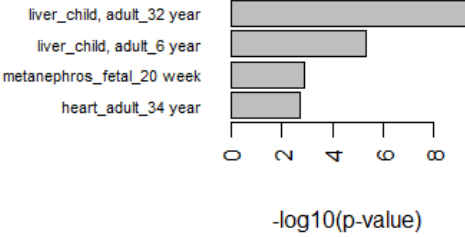
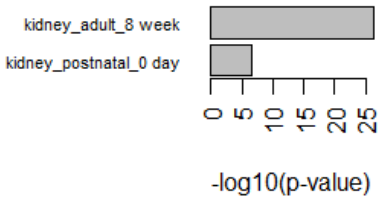
Supplementary Table 1 : Test result with different species' different cells/tissues with n=500, t=0.05

Data	Test result with human compendium	Test result with mouse compendium
Cattle (Bos taurus) brain GSE43013 PMID: 25677554	<p>C3 result of btaurus.br</p> <p>-log10(p-value)</p>	<p>C3 result of btaurus.br</p> <p>-log10(p-value)</p>
Dog (Canis lupus familiaris) GSE43013 PMID: 25677554	<p>C3 result of cfamiliaris.br</p> <p>-log10(p-value)</p>	<p>C3 result of cfamiliaris.br</p> <p>-log10(p-value)</p>
Domestic Guinea pig (Cavia porcellus) brain GSE43013 PMID: 25677554	<p>C3 result of cporcellus.br</p> <p>-log10(p-value)</p>	<p>C3 result of cporcellus.br</p> <p>-log10(p-value)</p>

<p>Horse (Equus caballus) brain</p> <p>GSE43013 PMID: 25677554</p>	<p>C3 result of ecaballus.br</p> <p>-log10(p-value)</p>	<p>C3 result of ecaballus.br</p> <p>-log10(p-value)</p>
<p>Hedgehog (Erinaceus europaeus) brain</p> <p>GSE43013 PMID: 25677554</p>	<p>C3 result of eeuropeus.br</p> <p>-log10(p-value)</p>	<p>C3 result of eeuropeus.br</p> <p>-log10(p-value)</p>
<p>Cat (Felis catus) brain</p> <p>GSE43013 PMID: 25677554</p>	<p>C3 result of fcatus.br</p> <p>-log10(p-value)</p>	<p>C3 result of fcatus.br</p> <p>-log10(p-value)</p>
<p>Mouse (Mus musculus) brain</p> <p>GSE43013 PMID: 25677554</p>	<p>C3 result of mmusculus.br</p> <p>-log10(p-value)</p>	<p>C3 result of mmusculus.br</p> <p>-log10(p-value)</p>

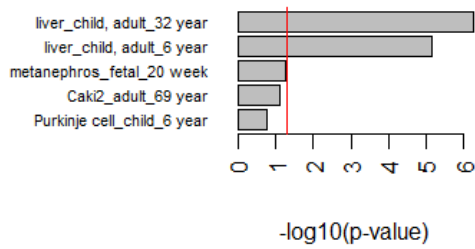
<p>Rabbit (<i>Oryctolagus cuniculus</i>) brain GSE43013 PMID: 25677554</p>	<p>C3 result of ocuniculus.br</p>  <p>-log10(p-value)</p>	<p>C3 result of ocuniculus.br</p>  <p>-log10(p-value)</p>
<p>Rat (<i>Rattus norvegicus</i>) brain GSE43013 PMID: 25677554</p>	<p>C3 result of rnorvegicus.br</p>  <p>-log10(p-value)</p>	<p>C3 result of rnorvegicus.br</p>  <p>-log10(p-value)</p>
<p>Pig (<i>Sus scrofa</i>) brain GSE43013 PMID: 25677554</p>	<p>C3 result of sscrofa.br</p>  <p>-log10(p-value)</p>	<p>C3 result of sscrofa.br</p>  <p>-log10(p-value)</p>
<p>Cattle (<i>Bos taurus</i>) kidney GSE43013 PMID: 25677554</p>	<p>C3 result of btaurus.kd</p>  <p>-log10(p-value)</p>	<p>C3 result of btaurus.kd</p>  <p>-log10(p-value)</p>

<p>Dog (<i>Canis lupus familiaris</i>) kidney GSE43013 PMID: 25677554</p>	<p>C3 result of cfamiliaris.kd</p> <p>liver_child, adult_32 year liver_child, adult_6 year hepatocyte_embryonic_5 day metanephros_fetal_20 week liver_fetal_22 week</p> <p>-log10(p-value)</p>	<p>C3 result of cfamiliaris.kd</p> <p>kidney_adult_8 week kidney_postnatal_0 day</p> <p>-log10(p-value)</p>
<p>Domestic Guinea pig (<i>Cavia porcellus</i>) kidney GSE43013 PMID: 25677554</p>	<p>C3 result of cporcellus.kd</p> <p>liver_child, adult_32 year liver_child, adult_6 year heart_adult_34 year liver_fetal_22 week metanephros_fetal_20 week</p> <p>-log10(p-value)</p>	<p>C3 result of cporcellus.kd</p> <p>kidney_adult_8 week kidney_postnatal_0 day</p> <p>-log10(p-value)</p>
<p>Horse (<i>Equus caballus</i>) kidney GSE43013 PMID: 25677554</p>	<p>C3 result of ecaballus.kd</p> <p>liver_child, adult_32 year liver_child, adult_6 year metanephros_fetal_20 week</p> <p>-log10(p-value)</p>	<p>C3 result of ecaballus.kd</p> <p>kidney_adult_8 week kidney_postnatal_0 day</p> <p>-log10(p-value)</p>
<p>Hedgehog (<i>Erinaceus europaeus</i>) kidney GSE43013 PMID: 25677554</p>	<p>C3 result of eeuropaeus.kd</p> <p>liver_child, adult_32 year liver_child, adult_6 year Purkinje_cell_child_6 year metanephros_fetal_20 week GM12878_adult_(2) GM12878_adult_(4)</p> <p>-log10(p-value)</p>	<p>C3 result of eeuropaeus.kd</p> <p>kidney_adult_8 week</p> <p>-log10(p-value)</p>

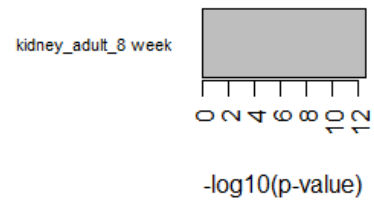
<p>Cat (<i>Felis catus</i>) kidney GSE43013 PMID: 25677554</p>	<p>C3 result of fcatus.kd</p>  <p>liver_child, adult_32 year liver_child, adult_6 year</p> <p>-log10(p-value)</p>	<p>C3 result of fcatus.kd</p>  <p>kidney_adult_8 week</p> <p>-log10(p-value)</p>
<p>Mouse (<i>Mus musculus</i>) kidney GSE43013 PMID: 25677554</p>	<p>C3 result of mmusculus.kd</p>  <p>liver_child, adult_32 year liver_child, adult_6 year metanephros_fetal_20 week heart_adult_34 year</p> <p>-log10(p-value)</p>	<p>C3 result of mmusculus.kd</p>  <p>kidney_adult_8 week kidney_postnatal_0 day liver_adult_8 week (1) liver_adult_8 week (2)</p> <p>-log10(p-value)</p>
<p>Rabbit (<i>Oryctolagus cuniculus</i>) kidney GSE43013 PMID: 25677554</p>	<p>C3 result of ocuniculus.kd</p>  <p>liver_child, adult_32 year liver_child, adult_6 year metanephros_fetal_20 week Purkinje cell_child_6 year heart_adult_34 year GM12878_adult_(4)</p> <p>-log10(p-value)</p>	<p>C3 result of ocuniculus.kd</p>  <p>kidney_adult_8 week kidney_postnatal_0 day liver_adult_8 week (2) liver_adult_8 week (1)</p> <p>-log10(p-value)</p>
<p>Rat (<i>Rattus norvegicus</i>) kidney GSE43013 PMID: 25677554</p>	<p>C3 result of rnorvegicus.kd</p>  <p>liver_child, adult_32 year liver_child, adult_6 year metanephros_fetal_20 week heart_adult_34 year</p> <p>-log10(p-value)</p>	<p>C3 result of rnorvegicus.kd</p>  <p>kidney_adult_8 week kidney_postnatal_0 day</p> <p>-log10(p-value)</p>

Pig (*Sus scrofa*)
kidney
[GSE43013](#)
PMID: [25677554](#)

C3 result of sscrofa.kd

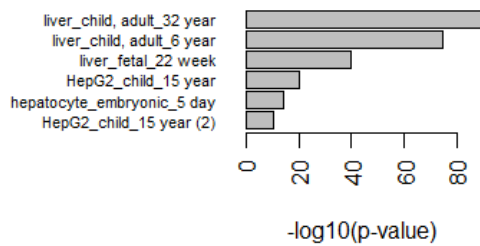


C3 result of sscrofa.kd

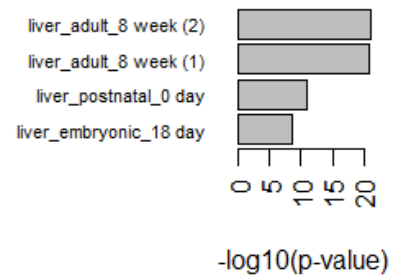


Cattle (*Bos taurus*) liver
[GSE43013](#)
PMID: [25677554](#)

C3 result of btaurus.lv

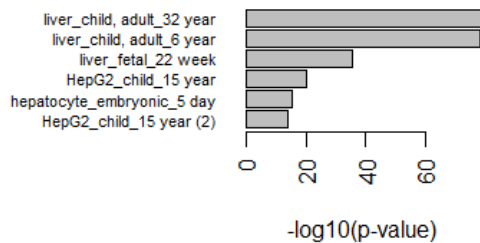


C3 result of btaurus.lv

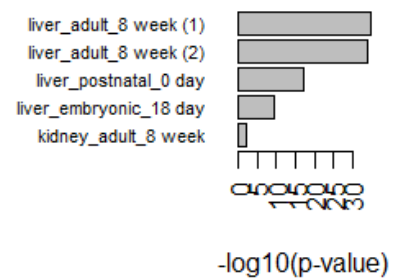


Dog (*Canis lupus familiaris*) liver
[GSE43013](#)
PMID: [25677554](#)

C3 result of cfamiliaris.lv



C3 result of cfamiliaris.lv

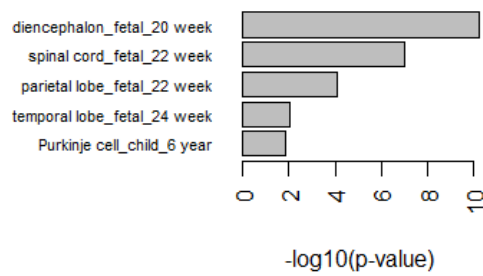


<p>Domestic Guinea pig (Cavia porcellus) liver</p> <p>GSE43013 PMID: 25677554</p>	<p>C3 result of cporcellus.lv</p> <p>liver_child_adult_32 year liver_child_adult_6 year liver_fetal_22 week hepatocyte_embryonic_5 day HepG2_child_15 year HepG2_child_15 year (2)</p> <p>-log10(p-value)</p>	<p>C3 result of cporcellus.lv</p> <p>liver_adult_8 week (1) liver_postnatal_0 day liver_adult_8 week (2) liver_embryonic_18 day</p> <p>-log10(p-value)</p>
<p>Horse (Equus caballus) liver</p> <p>GSE43013 PMID: 25677554</p>	<p>C3 result of ecaballus.lv</p> <p>liver_child_adult_6 year liver_child_adult_32 year liver_fetal_22 week HepG2_child_15 year hepatocyte_embryonic_5 day HepG2_child_15 year (2)</p> <p>-log10(p-value)</p>	<p>C3 result of ecaballus.lv</p> <p>liver_adult_8 week (1) liver_adult_8 week (2) liver_postnatal_0 day liver_embryonic_18 day kidney_adult_8 week</p> <p>-log10(p-value)</p>
<p>Hedgehog (Erinaceus europaeus) liver</p> <p>GSE43013 PMID: 25677554</p>	<p>C3 result of eeuropaeus.lv</p> <p>liver_child_adult_32 year liver_child_adult_6 year liver_fetal_22 week HepG2_child_15 year (2) HepG2_child_15 year hepatocyte_embryonic_5 day</p> <p>-log10(p-value)</p>	<p>C3 result of eeuropaeus.lv</p> <p>liver_adult_8 week (1) liver_adult_8 week (2) liver_postnatal_0 day liver_embryonic_18 day</p> <p>-log10(p-value)</p>
<p>Cat (Felis catus) liver</p> <p>GSE43013 PMID: 25677554</p>	<p>C3 result of fcatus.lv</p> <p>liver_child_adult_32 year liver_child_adult_6 year liver_fetal_22 week HepG2_child_15 year (2) HepG2_child_15 year hepatocyte_embryonic_5 day</p> <p>-log10(p-value)</p>	<p>C3 result of fcatus.lv</p> <p>liver_adult_8 week (2) liver_adult_8 week (1) liver_postnatal_0 day liver_embryonic_18 day</p> <p>-log10(p-value)</p>

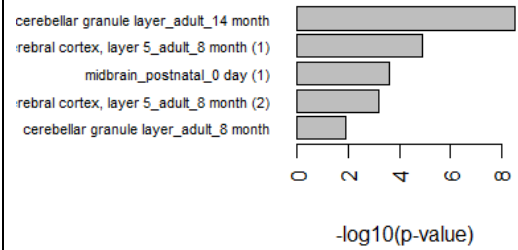
<p>Mouse (<i>Mus musculus</i>) liver GSE43013 PMID: 25677554</p>	<p>C3 result of mmusculus.lv</p> <p>liver_child_adult_32_year liver_child_adult_6_year liver_fetal_22_week HepG2_child_15_year hepatocyte_embryonic_5_day HepG2_child_15_year (2)</p> <p>-log10(p-value)</p>	<p>C3 result of mmusculus.lv</p> <p>liver_adult_8_week (1) liver_adult_8_week (2) liver_postnatal_0_day liver_embryonic_18_day</p> <p>-log10(p-value)</p>
<p>Rabbit (<i>Oryctolagus cuniculus</i>) liver GSE43013 PMID: 25677554</p>	<p>C3 result of ocuniculus.lv</p> <p>liver_child_adult_6_year liver_child_adult_32_year liver_fetal_22_week HepG2_child_15_year hepatocyte_embryonic_5_day HepG2_child_15_year (2)</p> <p>-log10(p-value)</p>	<p>C3 result of ocuniculus.lv</p> <p>liver_adult_8_week (2) liver_adult_8_week (1) liver_postnatal_0_day liver_embryonic_18_day</p> <p>-log10(p-value)</p>
<p>Rat (<i>Rattus norvegicus</i>) liver GSE43013 PMID: 25677554</p>	<p>C3 result of rnorvegicus.lv</p> <p>liver_child_adult_32_year liver_child_adult_6_year liver_fetal_22_week HepG2_child_15_year (2) hepatocyte_embryonic_5_day HepG2_child_15_year</p> <p>-log10(p-value)</p>	<p>C3 result of rnorvegicus.lv</p> <p>liver_adult_8_week (1) liver_adult_8_week (2) liver_postnatal_0_day liver_embryonic_18_day</p> <p>-log10(p-value)</p>
<p>Pig (<i>Sus scrofa</i>) liver GSE43013 PMID: 25677554</p>	<p>C3 result of sscrofa.lv</p> <p>liver_child_adult_6_year liver_child_adult_32_year liver_fetal_22_week HepG2_child_15_year hepatocyte_embryonic_5_day HepG2_child_15_year (2)</p> <p>-log10(p-value)</p>	<p>C3 result of sscrofa.lv</p> <p>liver_adult_8_week (2) liver_adult_8_week (1) liver_embryonic_18_day liver_postnatal_0_day</p> <p>-log10(p-value)</p>

Zebrafish
(Danio rerio)
control brain
[GSE74754](#)
PMID: [27935819](#)

C3 result of drerio.brain.ctl

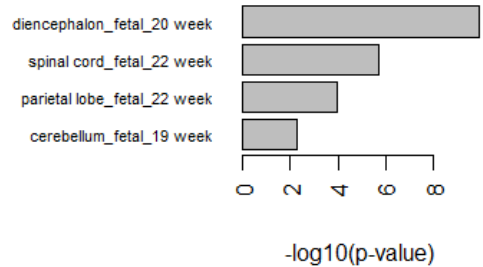


C3 result of drerio.brain.ctl

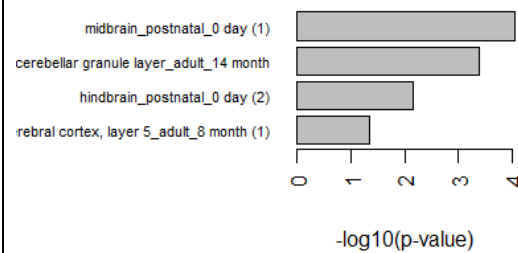


Zebrafish
(Danio rerio)
tumour brain
[GSE74754](#)
PMID: [27935819](#)

C3 result of drerio.brain.tumor



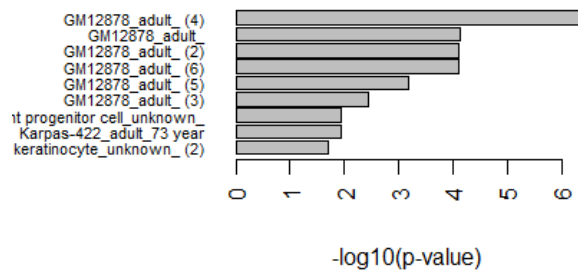
C3 result of drerio.brain.tumor



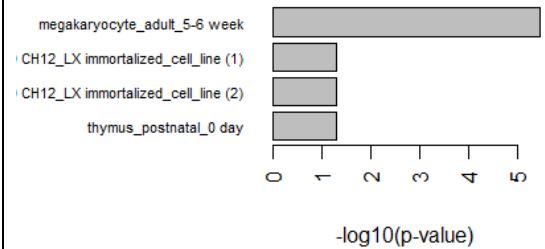
Dolphin
([Tursiops truncatus](#))
blood (hua)
[GSE78770](#)
PMID: [27608714](#)

Note:
GM12878
is a
lymphobla
stoid cell
line
produced
from the
blood of a
female
donor

C3 result of Hua



C3 result of Hua



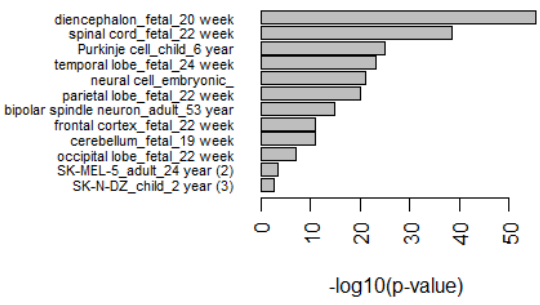
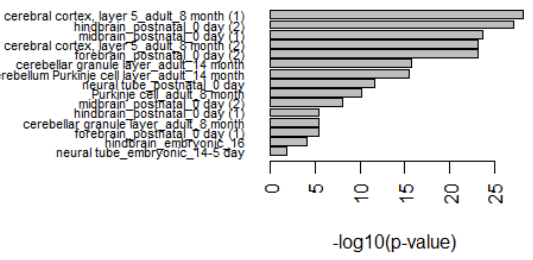
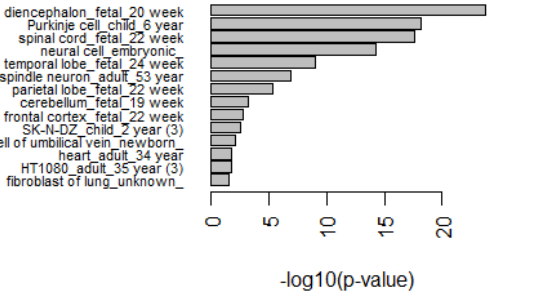
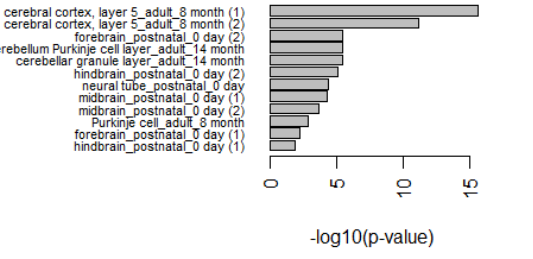
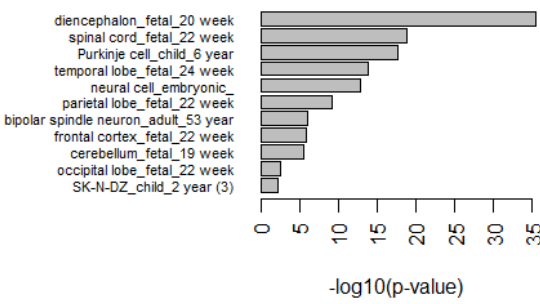
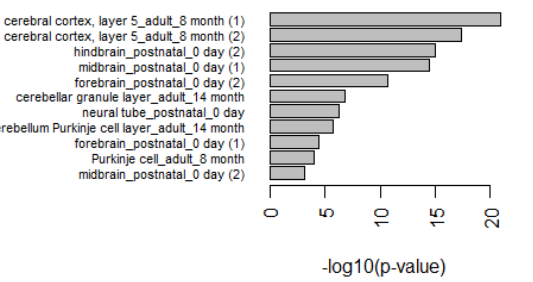
<p>Dolphin (Tursiops truncatus) blood (kai) GSE78770 PMID: 27608714</p>	<p>C3 result of Kai</p> <p>GM12878_adult_ (6) GM12878_adult_ (4) GM12878_adult_ (6) GM12878_adult_ (3) keratinocyte_unknown_ (2)</p> <p>-log10(p-value)</p>	<p>C3 result of Kai</p> <p>megakaryocyte_adult_5-6 week</p> <p>-log10(p-value)</p>
<p>Dolphin (Tursiops truncatus) blood (keo) GSE78770 PMID: 27608714</p>	<p>C3 result of Keo</p> <p>GM12878_adult_ (6) GM12878_adult_ (4) it progenitor cell_unknown_ (3) GM12878_adult_ (3) Karpas-422_adult_73 year GM12878_adult_ (2) GM12878_adult_ (5)</p> <p>-log10(p-value)</p>	<p>C3 result of Keo</p> <p>megakaryocyte_adult_5-6 week</p> <p>-log10(p-value)</p>
<p>Dolphin (Tursiops truncatus) blood (pele) GSE78770 PMID: 27608714</p>	<p>C3 result of Pele</p> <p>GM12878_adult_ (6) GM12878_adult_ (4) GM12878_adult_ (5) GM12878_adult_ (2) GM12878_adult_ (3) it progenitor cell_unknown_ (3) Karpas-422_adult_73 year</p> <p>-log10(p-value)</p>	<p>C3 result of Pele</p> <p>megakaryocyte_adult_5-6 week</p> <p>-log10(p-value)</p>
<p>Monkey (Macaca mulatta) skeletal muscle (early BPA) GSE53393 PMID: 24586524</p>	<p>C3 result of early.BPA</p> <p>skeletal muscle tissue_fetal_19 week tongue_fetal_20 week LHCN-MZ_adult_41 year heart_adult_34 year thyroid gland_fetal_40 week heart_fetal_19 week myotube_unknown_ (3) skeletal muscle myoblast_unknown_ (3) skin of body_fetal_22 week</p> <p>-log10(p-value)</p>	<p>C3 result of early.BPA</p> <p>skeletal muscle tissue_postnatal_0 day facial prominence_embryonic_14-5 day</p> <p>-log10(p-value)</p>

<p>Monkey (Macaca mulatta) skeletal muscle (early control) GSE53393 PMID: 24586524</p>	<p>C3 result of early.Ctl</p> <p>-log10(p-value)</p>	<p>C3 result of early.Ctl</p> <p>-log10(p-value)</p>
<p>Monkey (Macaca mulatta) skeletal muscle (late BPA) GSE53393 PMID: 24586524</p>	<p>C3 result of late.BPA</p> <p>-log10(p-value)</p>	<p>C3 result of late.BPA</p> <p>-log10(p-value)</p>
<p>Monkey (Macaca mulatta) skeletal muscle (late control) GSE53393 PMID: 24586524</p>	<p>C3 result of late.ctl</p> <p>-log10(p-value)</p>	<p>C3 result of late.ctl</p> <p>-log10(p-value)</p>

Note: Here, we have selected the top 500 highly expressed genes for each cell/tissue of the compendium and set the cut-off threshold value 5 percent of total number of cell/tissue; and then performed the test for each of the query data set with both the human and mouse compendium separately.

**Supplementary Table 2 : Test result with different species' different cells/tissues with n=1000,
t=0.10**

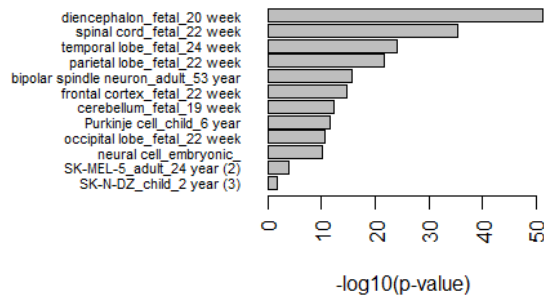
Data	Test result with human compendium	Test result with mouse compendium
Cattle (Bos taurus) brain GSE43013 PMID: 25677554	<p>C3 result of btaurus.br</p> <p>-log10(p-value)</p>	<p>C3 result of btaurus.br</p> <p>-log10(p-value)</p>
Dog (Canis lupus familiaris) GSE43013 PMID: 25677554	<p>C3 result of cfamiliaris.br</p> <p>-log10(p-value)</p>	<p>C3 result of cfamiliaris.br</p> <p>-log10(p-value)</p>
Domestic Guinea pig (Cavia porcellus) brain GSE43013 PMID: 25677554	<p>C3 result of cporcellus.br</p> <p>-log10(p-value)</p>	<p>C3 result of cporcellus.br</p> <p>-log10(p-value)</p>

<p>Horse (Equus caballus) brain</p> <p>GSE43013</p> <p>PMID: 25677554</p>	<p>C3 result of ecaballus.br</p>  <p>diencephalon_fetal_20 week spinal cord_fetal_22 week Purkinje cell_child_6 year temporal lobe_fetal_24 week neural cell_embryonic parietal lobe_fetal_22 week bipolar spindle neuron_adult_53 year frontal cortex_fetal_22 week cerebellum_fetal_19 week occipital lobe_fetal_22 week SK-MEL-5_adult_24 year (2) SK-N-DZ_child_2 year (3)</p> <p>-log10(p-value)</p>	<p>C3 result of ecaballus.br</p>  <p>cerebral cortex_layer 5_adult_8 month (1) hindbrain_postnatal_0 day (2) midbrain_postnatal_0 day (1) cerebral cortex_layer 5_adult_8 month (2) forebrain_postnatal_0 day (2) cerebellar granule layer_adult_14 month erebellum Purkinje cell layer_adult_14 month neural tube_postnatal_0 day Purkinje cell_adult_8 month midbrain_postnatal_0 day (2) hindbrain_postnatal_0 day (1) cerebellar granule layer_adult_8 month forebrain_postnatal_0 day (1) hindbrain_embryonic_16 neural tube_embryonic_14-5 day</p> <p>-log10(p-value)</p>
<p>Hedgehog (Erinaceus europaeus) brain</p> <p>GSE43013</p> <p>PMID: 25677554</p>	<p>C3 result of eeuropeus.br</p>  <p>diencephalon_fetal_20 week Purkinje cell_child_6 year spinal cord_fetal_22 week neural cell_embryonic temporal lobe_fetal_24 week bipolar spindle neuron_adult_53 year parietal lobe_fetal_22 week cerebellum_fetal_19 week frontal cortex_fetal_22 week SK-N-DZ_child_2 year (3) cell of umbilical vein_newborn heart_adult_34 year HT1080_adult_35 year (3) fibroblast of lung_unknown_</p> <p>-log10(p-value)</p>	<p>C3 result of eeuropeus.br</p>  <p>cerebral cortex_layer 5_adult_8 month (1) cerebral cortex_layer 5_adult_8 month (2) forebrain_postnatal_0 day (2) erebellum Purkinje cell layer_adult_14 month cerebellar granule layer_adult_14 month hindbrain_postnatal_0 day (2) neural tube_postnatal_0 day midbrain_postnatal_0 day (1) midbrain_postnatal_0 day (2) Purkinje cell_adult_8 month forebrain_postnatal_0 day (1) hindbrain_postnatal_0 day (1)</p> <p>-log10(p-value)</p>
<p>Cat (Felis catus) brain</p> <p>GSE43013</p> <p>PMID: 25677554</p>	<p>C3 result of fcatus.br</p>  <p>diencephalon_fetal_20 week spinal cord_fetal_22 week Purkinje cell_child_6 year temporal lobe_fetal_24 week neural cell_embryonic parietal lobe_fetal_22 week bipolar spindle neuron_adult_53 year frontal cortex_fetal_22 week cerebellum_fetal_19 week occipital lobe_fetal_22 week SK-N-DZ_child_2 year (3)</p> <p>-log10(p-value)</p>	<p>C3 result of fcatus.br</p>  <p>cerebral cortex_layer 5_adult_8 month (1) cerebral cortex_layer 5_adult_8 month (2) hindbrain_postnatal_0 day (2) midbrain_postnatal_0 day (1) forebrain_postnatal_0 day (2) cerebellar granule layer_adult_14 month neural tube_postnatal_0 day erebellum Purkinje cell layer_adult_14 month forebrain_postnatal_0 day (1) Purkinje cell_adult_8 month midbrain_postnatal_0 day (2)</p> <p>-log10(p-value)</p>

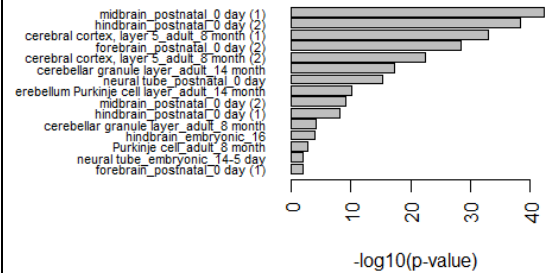
Mouse (*Mus musculus*)
brain

[GSE43013](#)
PMID: [25677](#)
[554](#)

C3 result of *mmusculus.br*



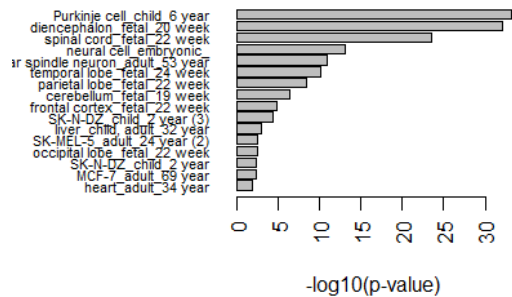
C3 result of *mmusculus.br*



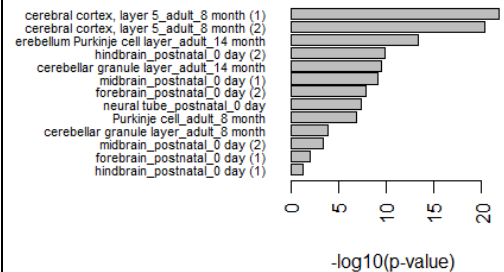
Rabbit (*Oryctolagus cuniculus*)
brain

[GSE43013](#)
PMID: [25677](#)
[554](#)

C3 result of *ocuniculus.br*



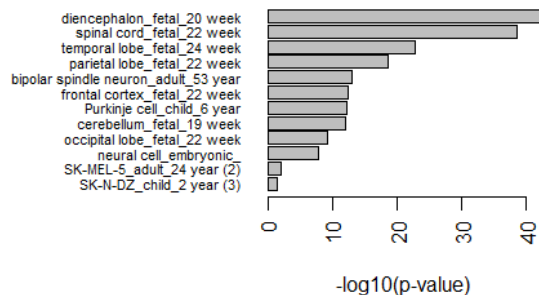
C3 result of *ocuniculus.br*



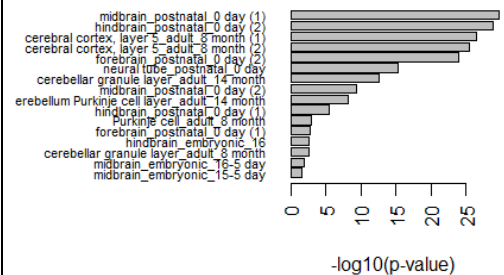
Rat (*Rattus norvegicus*)
brain

[GSE43013](#)
PMID: [25677](#)
[554](#)

C3 result of *rnorvegicus.br*

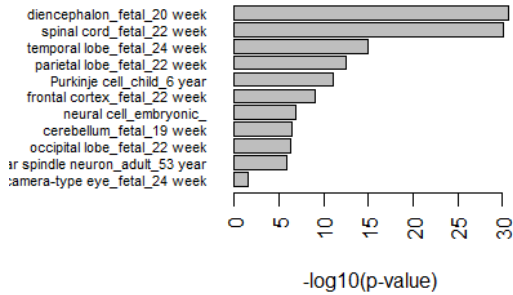


C3 result of *rnorvegicus.br*

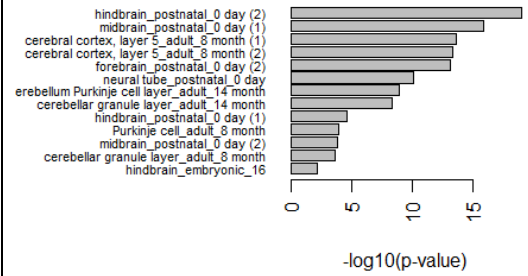


Pig (Sus
scrofa) brain
[GSE43013](#)
PMID: [25677](#)
[554](#)

C3 result of sscrofa.br

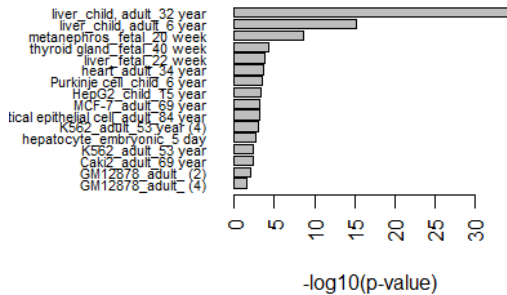


C3 result of sscrofa.br

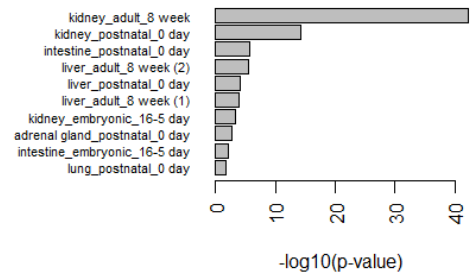


Cattle (Bos
taurus)
kidney
[GSE43013](#)
PMID: [25677](#)
[554](#)

C3 result of btaurus.kd

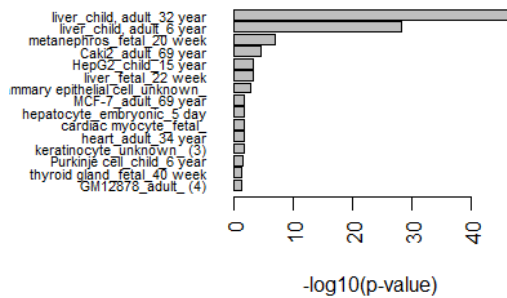


C3 result of btaurus.kd

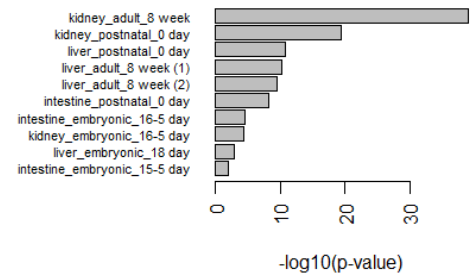


Dog (Canis
lupus
familiaris)
kidney
[GSE43013](#)
PMID: [25677](#)
[554](#)

C3 result of cfamiliaris.kd



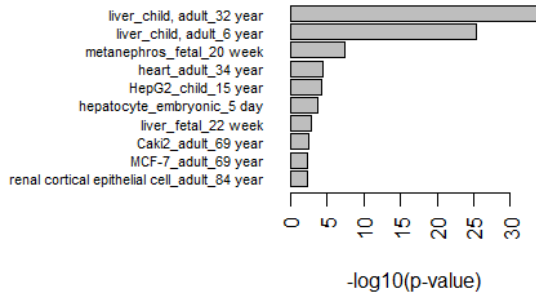
C3 result of cfamiliaris.kd



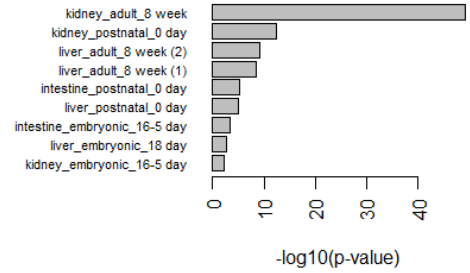
Domestic Guinea pig (Cavia porcellus) kidney

[GSE43013](#)
PMID: [25677554](#)

C3 result of cporcellus.kd



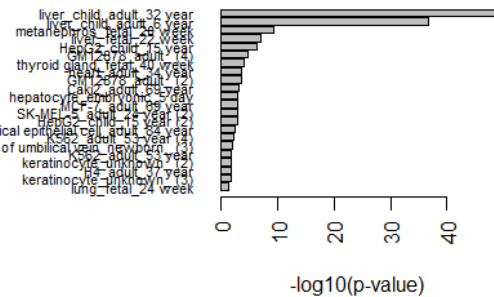
C3 result of cporcellus.kd



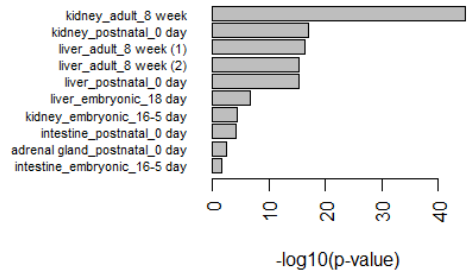
Horse (Equus caballus) kidney

[GSE43013](#)
PMID: [25677554](#)

C3 result of ecaballus.kd



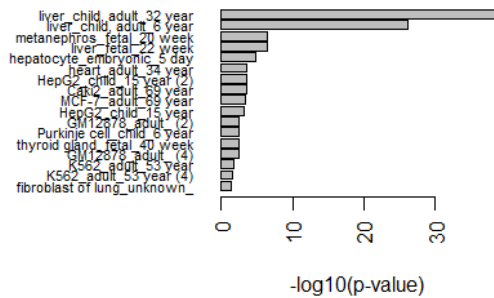
C3 result of ecaballus.kd



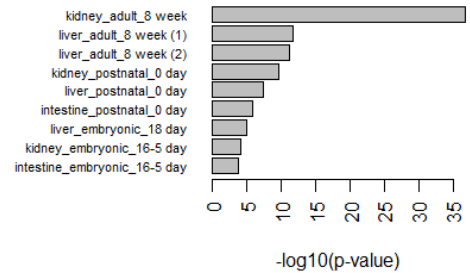
Hedgehog (Erinaceus europaeus) kidney

[GSE43013](#)
PMID: [25677554](#)

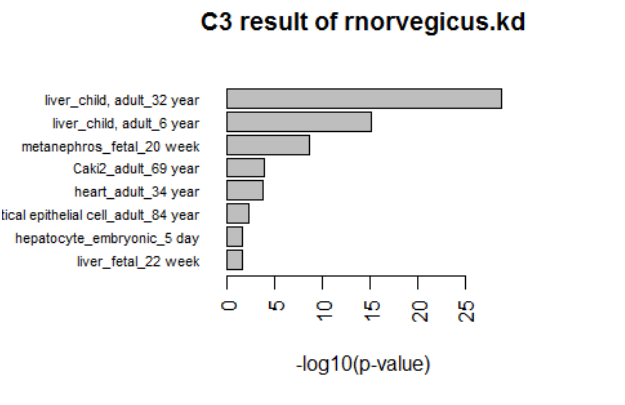
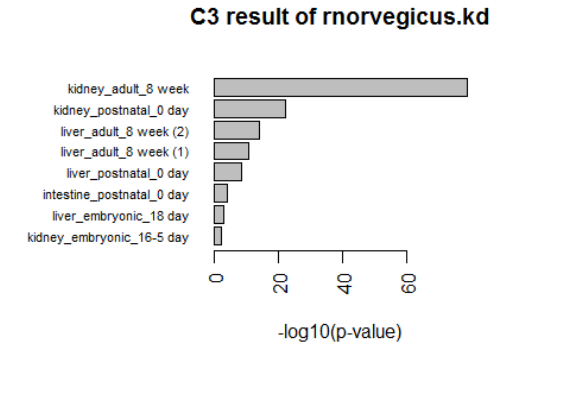
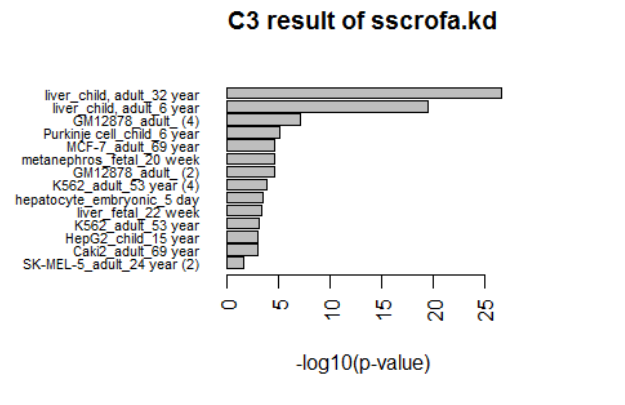
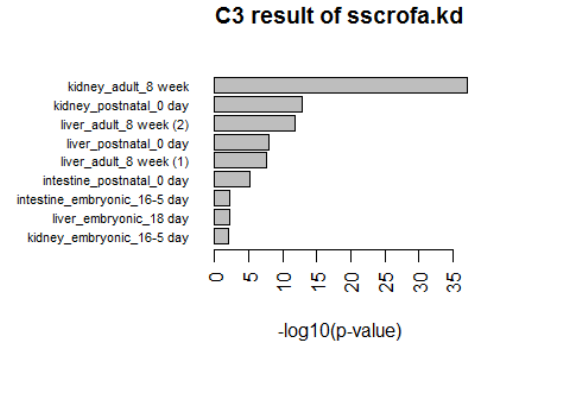
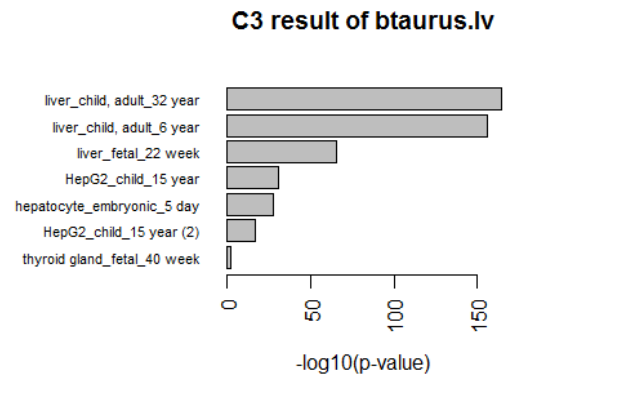
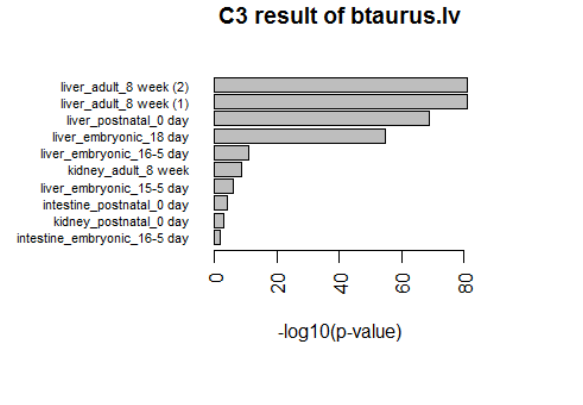
C3 result of eeuropaeus.kd



C3 result of eeuropaeus.kd



<p>Cat (<i>Felis catus</i>) kidney GSE43013 PMID: 25677554</p>	<p>C3 result of fcatus.kd</p> <p>liver_child_adult_32 year liver_child_adult_6 year metanephros_fetal_20 week liver_fetal_22 week HepG2_child_15 year K562_adult_53 year (4) GM12878_adult_ (4) SK-MEL-5_adult_24 year (2) thyroid_gland_fetal_40 week MCF-7_adult_69 year renal_cortical_epithelial_cell_adult_84 year fibroblast_of_lung_unknown heart_adult_34 year HepG2_child_15 year (2)</p> <p>-log10(p-value)</p>	<p>C3 result of fcatus.kd</p> <p>kidney_adult_8 week liver_adult_8 week (2) liver_adult_8 week (1) liver_postnatal_0 day kidney_postnatal_0 day intestine_postnatal_0 day liver_embryonic_18 day kidney_embryonic_16-5 day</p> <p>-log10(p-value)</p>
<p>Mouse (<i>Mus musculus</i>) kidney GSE43013 PMID: 25677554</p>	<p>C3 result of mmusculus.kd</p> <p>liver_child_adult_32 year liver_child_adult_6 year metanephros_fetal_20 week heart_adult_34 year liver_fetal_22 week HepG2_child_15 year hepatocyte_embryonic_5 day Caki2_adult_69 year</p> <p>-log10(p-value)</p>	<p>C3 result of mmusculus.kd</p> <p>kidney_adult_8 week liver_adult_8 week (2) liver_adult_8 week (1) kidney_postnatal_0 day liver_postnatal_0 day intestine_postnatal_0 day liver_embryonic_18 day kidney_embryonic_16-5 day</p> <p>-log10(p-value)</p>
<p>Rabbit (<i>Oryctolagus cuniculus</i>) kidney GSE43013 PMID: 25677554</p>	<p>C3 result of ocuniculus.kd</p> <p>liver_child_adult_32 year liver_child_adult_6 year metanephros_fetal_20 week HepG2_child_15 year liver_fetal_22 week heart_adult_34 year Purkinje_cell_child_6 year thyroid_gland_fetal_40 week Caki2_adult_69 year tical_epithelial_cell_adult_84 year keratinocyte_unknown_(3) G401_child_3 month</p> <p>-log10(p-value)</p>	<p>C3 result of ocuniculus.kd</p> <p>kidney_adult_8 week kidney_postnatal_0 day liver_adult_8 week (2) liver_adult_8 week (1) intestine_postnatal_0 day kidney_embryonic_16-5 day liver_postnatal_0 day intestine_embryonic_16-5 day</p> <p>-log10(p-value)</p>

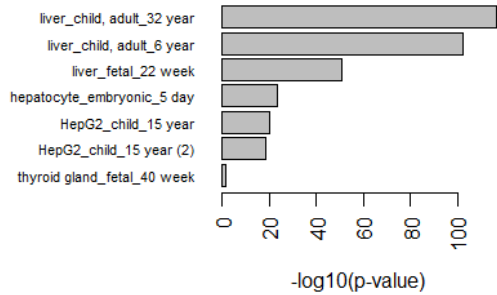
<p>Rat (<i>Rattus norvegicus</i>) kidney GSE43013 PMID: 25677554</p>	<p>C3 result of rnorvegicus.kd</p>  <p>liver_child_adult_32 year liver_child_adult_6 year metanephros_fetal_20 week Caki2_adult_69 year heart_adult_34 year liver_epithelial_cell_adult_84 year hepatocyte_embryonic_5 day liver_fetal_22 week</p> <p>-log10(p-value)</p>	<p>C3 result of rnorvegicus.kd</p>  <p>kidney_adult_8 week kidney_postnatal_0 day liver_adult_8 week (2) liver_adult_8 week (1) liver_postnatal_0 day intestine_postnatal_0 day liver_embryonic_18 day kidney_embryonic_16-5 day</p> <p>-log10(p-value)</p>
<p>Pig (<i>Sus scrofa</i>) kidney GSE43013 PMID: 25677554</p>	<p>C3 result of sscrofa.kd</p>  <p>liver_child_adult_32 year liver_child_adult_6 year GM12878_adult_(4) Purkinje_cell_child_6 year MCF-7_adult_69 year metanephros_fetal_20 week GM12878_adult_(2) K562_adult_53 year (4) hepatocyte_embryonic_5 day liver_fetal_22 week K562_adult_53 year HepG2_child_15 year Caki2_adult_69 year SK-MEL-5_adult_24 year (2)</p> <p>-log10(p-value)</p>	<p>C3 result of sscrofa.kd</p>  <p>kidney_adult_8 week kidney_postnatal_0 day liver_adult_8 week (2) liver_postnatal_0 day liver_adult_8 week (1) intestine_postnatal_0 day intestine_embryonic_16-5 day liver_embryonic_18 day kidney_embryonic_16-5 day</p> <p>-log10(p-value)</p>
<p>Cattle (<i>Bos taurus</i>) liver GSE43013 PMID: 25677554</p>	<p>C3 result of btaurus.lv</p>  <p>liver_child_adult_32 year liver_child_adult_6 year liver_fetal_22 week HepG2_child_15 year hepatocyte_embryonic_5 day HepG2_child_15 year (2) thyroid_gland_fetal_40 week</p> <p>-log10(p-value)</p>	<p>C3 result of btaurus.lv</p>  <p>liver_adult_8 week (2) liver_adult_8 week (1) liver_postnatal_0 day liver_embryonic_18 day liver_embryonic_16-5 day kidney_adult_8 week liver_embryonic_15-5 day intestine_postnatal_0 day kidney_postnatal_0 day intestine_embryonic_16-5 day</p> <p>-log10(p-value)</p>

<p>Dog (<i>Canis lupus familiaris</i>) liver</p> <p>GSE43013 PMID: 25677554</p>	<p>C3 result of cfamiliaris.lv</p> <p>liver_child_adult_32 year liver_child_adult_6 year liver_fetal_22 week HepG2_child_15 year hepatocyte_embryonic_5 day HepG2_child_15 year (2)</p> <p>-log10(p-value)</p>	<p>C3 result of cfamiliaris.lv</p> <p>liver_adult_8 week (1) liver_adult_8 week (2) liver_postnatal_0 day liver_embryonic_18 day liver_embryonic_16-5 day kidney_adult_8 week liver_embryonic_15-5 day intestine_embryonic_16-5 day kidney_postnatal_0 day intestine_postnatal_0 day</p> <p>-log10(p-value)</p>
<p>Domestic Guinea pig (<i>Cavia porcellus</i>) liver</p> <p>GSE43013 PMID: 25677554</p>	<p>C3 result of cporcellus.lv</p> <p>liver_child_adult_32 year liver_child_adult_6 year liver_fetal_22 week HepG2_child_15 year hepatocyte_embryonic_5 day HepG2_child_15 year (2)</p> <p>-log10(p-value)</p>	<p>C3 result of cporcellus.lv</p> <p>liver_adult_8 week (1) liver_adult_8 week (2) liver_postnatal_0 day liver_embryonic_18 day liver_embryonic_16-5 day kidney_adult_8 week liver_embryonic_15-5 day intestine_embryonic_16-5 day intestine_postnatal_0 day kidney_postnatal_0 day</p> <p>-log10(p-value)</p>
<p>Horse (<i>Equus caballus</i>) liver</p> <p>GSE43013 PMID: 25677554</p>	<p>C3 result of ecaballus.lv</p> <p>liver_child_adult_32 year liver_child_adult_6 year liver_fetal_22 week HepG2_child_15 year hepatocyte_embryonic_5 day HepG2_child_15 year (2) thyroid_gland_fetal_40 week</p> <p>-log10(p-value)</p>	<p>C3 result of ecaballus.lv</p> <p>liver_adult_8 week (1) liver_adult_8 week (2) liver_postnatal_0 day liver_embryonic_18 day liver_embryonic_16-5 day kidney_adult_8 week liver_embryonic_15-5 day kidney_postnatal_0 day</p> <p>-log10(p-value)</p>

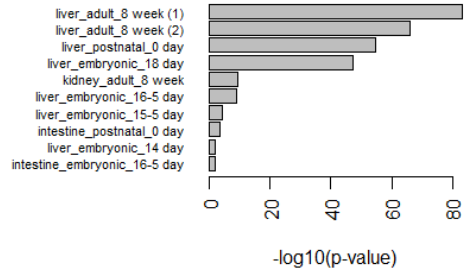
Hedgehog
(*Erinaceus europaeus*)
liver

[GSE43013](#)
PMID: [25677554](#)

C3 result of eeuropeaus.lv



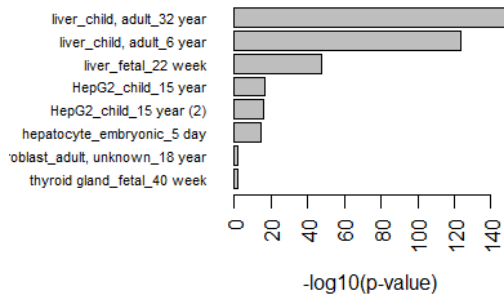
C3 result of eeuropeaus.lv



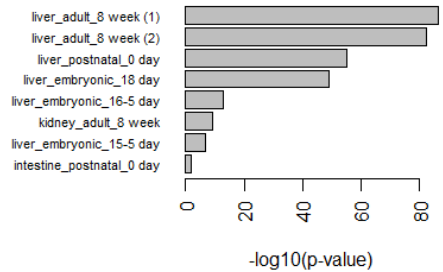
Cat (*Felis catus*) liver

[GSE43013](#)
PMID: [25677554](#)

C3 result of fcatus.lv



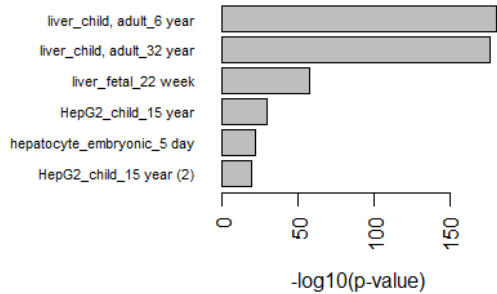
C3 result of fcatus.lv



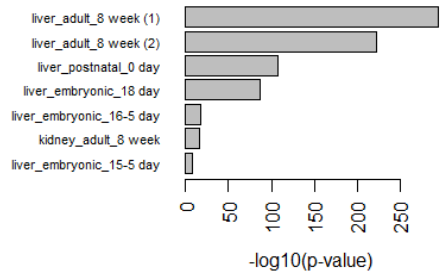
Mouse (*Mus musculus*)
liver

[GSE43013](#)
PMID: [25677554](#)

C3 result of mmusculus.lv



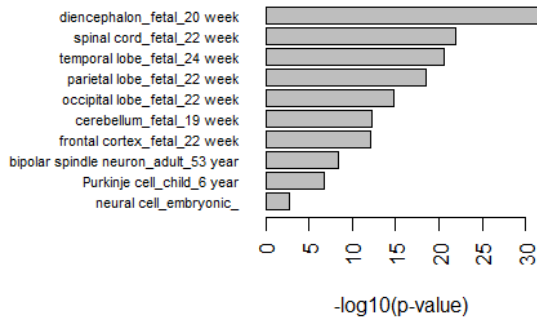
C3 result of mmusculus.lv



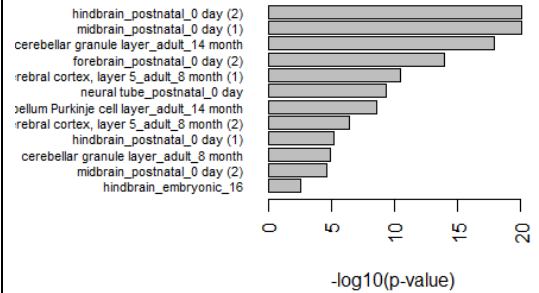
<p>Rabbit (<i>Oryctolagus cuniculus</i>) liver GSE43013 PMID: 25677554</p>	<p>C3 result of ocuniculus.lv</p> <p>liver_child_adult_6 year liver_child_adult_32 year liver_fetal_22 week HepG2_child_15 year hepatocyte_embryonic_5 day HepG2_child_15 year (2)</p> <p>-log10(p-value)</p>	<p>C3 result of ocuniculus.lv</p> <p>liver_adult_8 week (1) liver_adult_8 week (2) liver_postnatal_0 day liver_embryonic_18 day kidney_adult_8 week liver_embryonic_16-5 day liver_embryonic_15-5 day intestine_postnatal_0 day intestine_embryonic_16-5 day intestine_embryonic_15-5 day kidney_postnatal_0 day</p> <p>-log10(p-value)</p>
<p>Rat (<i>Rattus norvegicus</i>) liver GSE43013 PMID: 25677554</p>	<p>C3 result of rnorvegicus.lv</p> <p>liver_child_adult_32 year liver_child_adult_6 year liver_fetal_22 week hepatocyte_embryonic_5 day HepG2_child_15 year HepG2_child_15 year (2)</p> <p>-log10(p-value)</p>	<p>C3 result of rnorvegicus.lv</p> <p>liver_adult_8 week (1) liver_adult_8 week (2) liver_postnatal_0 day liver_embryonic_18 day liver_embryonic_16-5 day kidney_adult_8 week liver_embryonic_15-5 day kidney_postnatal_0 day</p> <p>-log10(p-value)</p>
<p>Pig (<i>Sus scrofa</i>) liver GSE43013 PMID: 25677554</p>	<p>C3 result of sscrofa.lv</p> <p>liver_child_adult_6 year liver_child_adult_32 year liver_fetal_22 week HepG2_child_15 year hepatocyte_embryonic_5 day HepG2_child_15 year (2) thyroid gland_fetal_40 week</p> <p>-log10(p-value)</p>	<p>C3 result of sscrofa.lv</p> <p>liver_adult_8 week (1) liver_adult_8 week (2) liver_postnatal_0 day liver_embryonic_18 day kidney_adult_8 week liver_embryonic_16-5 day liver_embryonic_15-5 day kidney_postnatal_0 day</p> <p>-log10(p-value)</p>

Zebrafish
(Danio rerio)
control brain
[GSE74754](#)
PMID: [27935819](#)

C3 result of drerio.brain.ctl

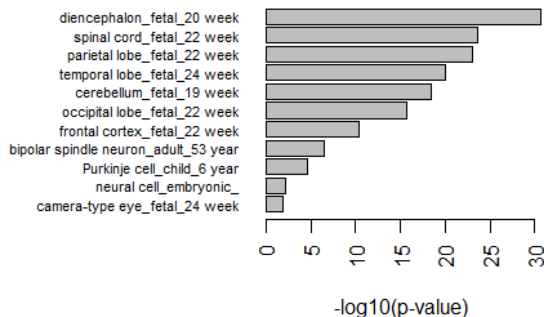


C3 result of drerio.brain.ctl

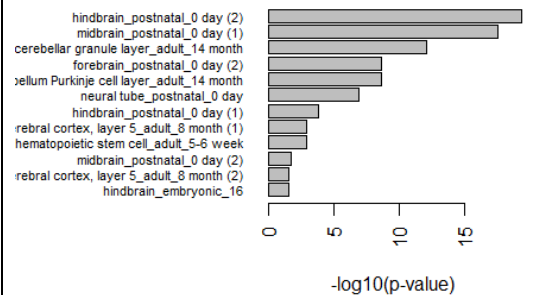


Zebrafish
(Danio rerio)
tumour brain
[GSE74754](#)
PMID: [27935819](#)

C3 result of drerio.brain.tumor



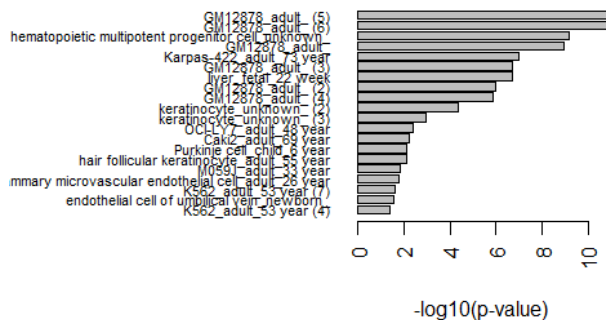
C3 result of drerio.brain.tumor



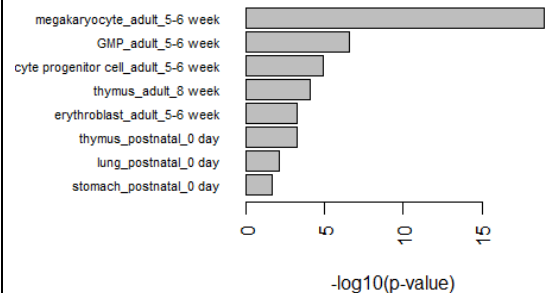
Dolphin
([Tursiops truncatus](#))
blood (hua)
[GSE78770](#)
PMID: [27608714](#)

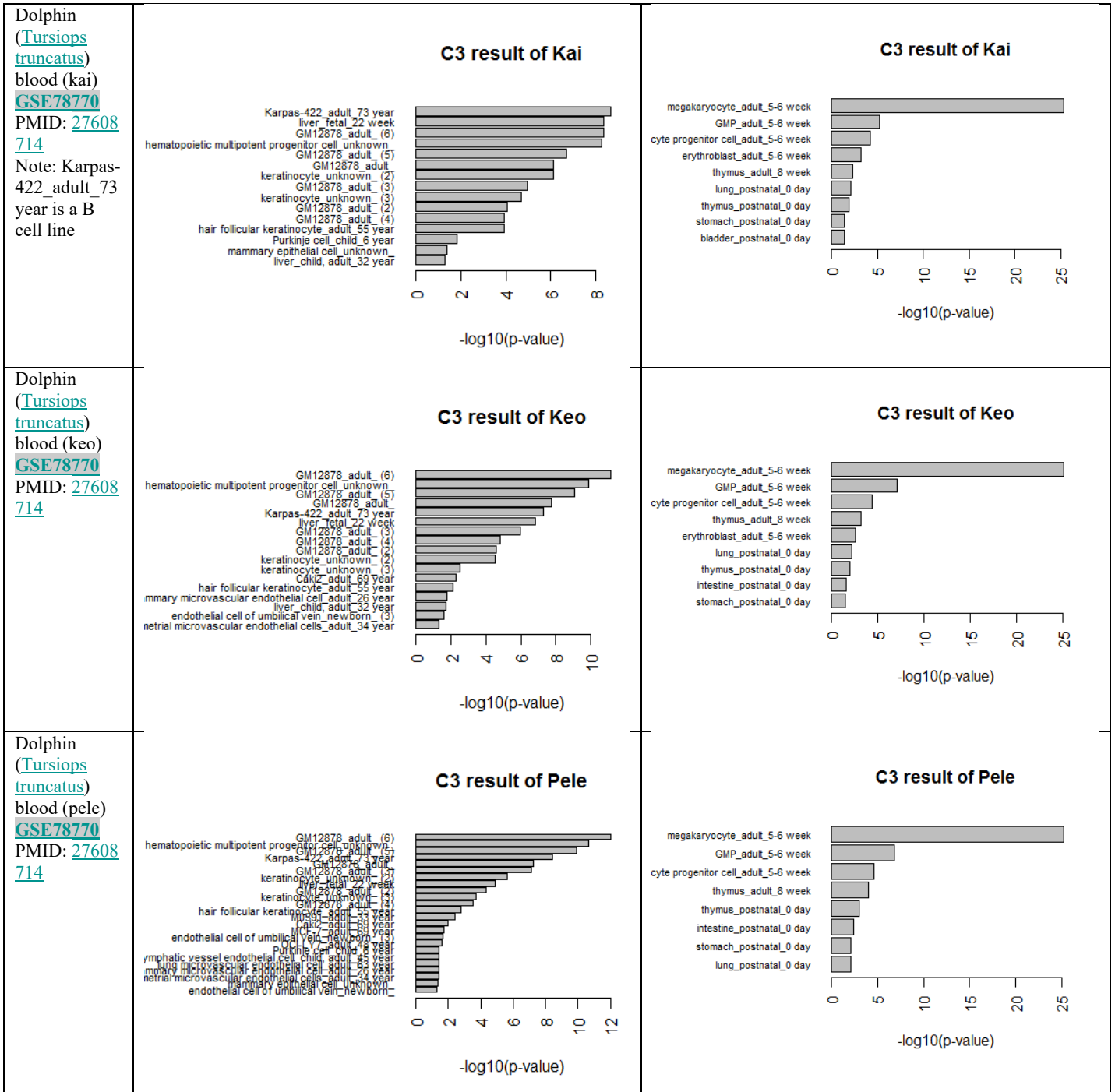
Note:
GM12878
is a
lymphoblastoid cell
line
produced
from the
blood of a
female
donor

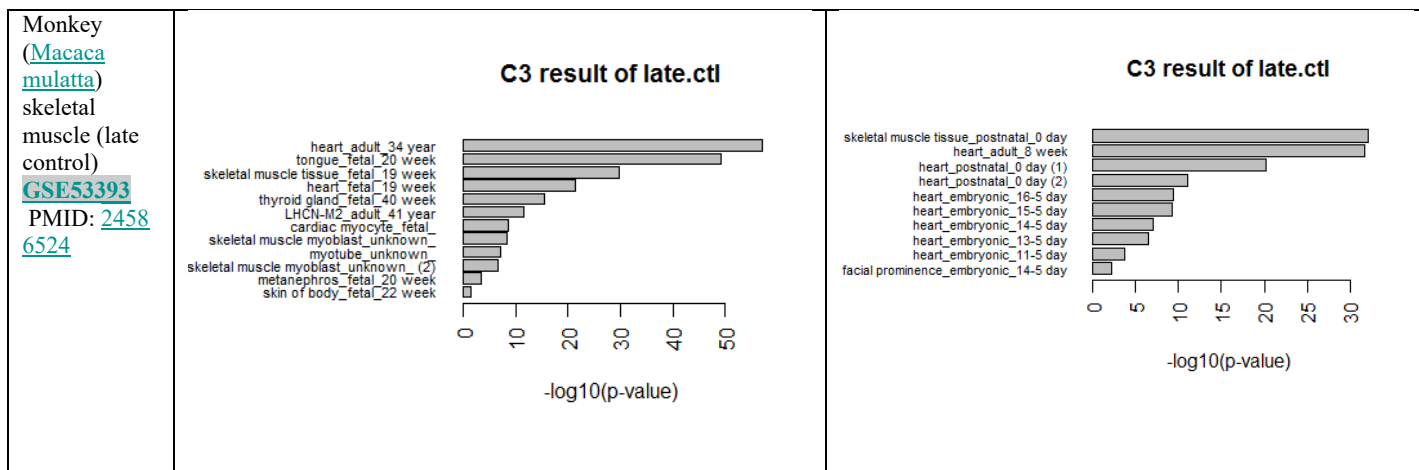
C3 result of Hua



C3 result of Hua







Note: Here, we have selected the top 1000 highly expressed genes for each cell/tissue of the compendium and set the cut-off threshold value 10 percent of total number of cell/tissue; and then performed the test for each of the query data set with both the human and mouse compendium separately.

Supplementary Table 3: Summary test results of selected five samples after quantile normalization of the data sets

	Sample name	<i>n</i> =500, <i>t</i> =0.05		<i>n</i> =1000, <i>t</i> =0.10	
		Human	Mouse	Human	Mouse
Data set 1 (GSE43013)	<i>R. norvegicus</i> kidney	2	1	2	1
	<i>B. taurus</i> liver	1	1	1	1
Data set 2 (GSE74754)	<i>D. rerio</i> brain (control)	1	1	1	1
Data set 3 (GSE78770)	<i>T. truncatus</i> blood (hua)	1	1	1	1
Data set 4 (GSE53393)	<i>M. mulatta</i> skeletal muscle (early BPA)	1	1	2	1

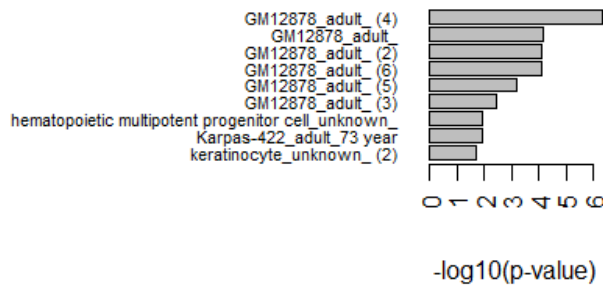
Supplementary Table 4: Detail test results of selected five samples after quantile normalization of the data sets with parameter settings $n=500$, $t=0.05$

Data	Test result with human compendium	Test result with mouse compendium
Rat (<i>Rattus norvegicus</i>) kidney GSE43013 PMID: 25677554	<p>C3 result of rnorvegicus.kd</p> <p>liver_child_adult_32 year liver_child_adult_6 year metanephros_fetal_20 week heart_adult_34 year</p> <p>-log10(p-value)</p>	<p>C3 result of rnorvegicus.kd</p> <p>kidney_adult_8 week kidney_postnatal_0 day</p> <p>-log10(p-value)</p>
Cattle (<i>Bos taurus</i>) liver GSE43013 PMID: 25677554	<p>C3 result of btaurus.lv</p> <p>liver_child_adult_32 year liver_child_adult_6 year liver_fetal_22 week HepG2_child_15 year hepatocyte_embryonic_5 day HepG2_child_15 year (2)</p> <p>-log10(p-value)</p>	<p>C3 result of btaurus.lv</p> <p>liver_adult_8 week (2) liver_adult_8 week (1) liver_postnatal_0 day liver_embryonic_18 day</p> <p>-log10(p-value)</p>
Zebrafish (<i>Danio rerio</i>) control brain GSE74754 PMID: 27935819	<p>C3 result of drerio.brain.ctl</p> <p>diencephalon_fetal_20 week spinal cord_fetal_22 week parietal lobe_fetal_22 week temporal lobe_fetal_24 week Purkinje cell_child_6 year</p> <p>-log10(p-value)</p>	<p>C3 result of drerio.brain.ctl</p> <p>cerebellar granule layer_adult_14 month rebral cortex_layer 5_adult_8 month (1) midbrain_postnatal_0 day (1) rebral cortex_layer 5_adult_8 month (2) cerebellar granule layer_adult_8 month</p> <p>-log10(p-value)</p>

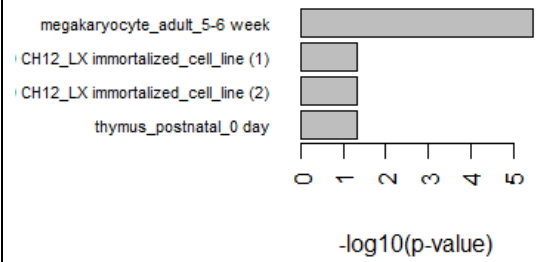
Dolphin
([Tursiops truncatus](#))
blood (hua)
[GSE78770](#)
PMID: [27608714](#)

Note:
GM12878
is a
lymphobla
stoid cell
line
produced
from the
blood of a
female
donor

C3 result of Hua

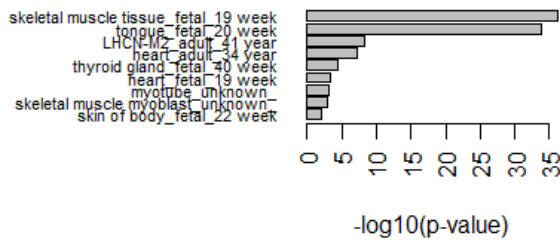


C3 result of Hua

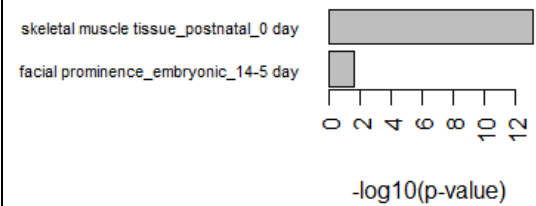


Monkey
([Macaca mulatta](#))
skeletal
muscle (early
BPA)
[GSE53393](#)
PMID: [24586524](#)

C3 result of early.BPA



C3 result of early.BPA

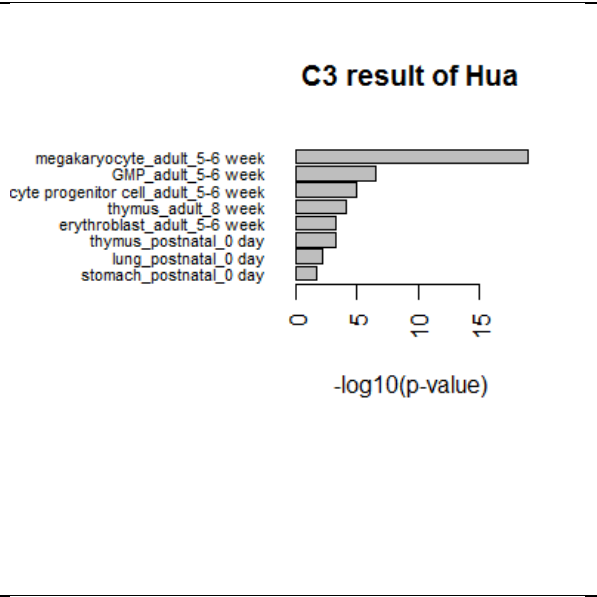
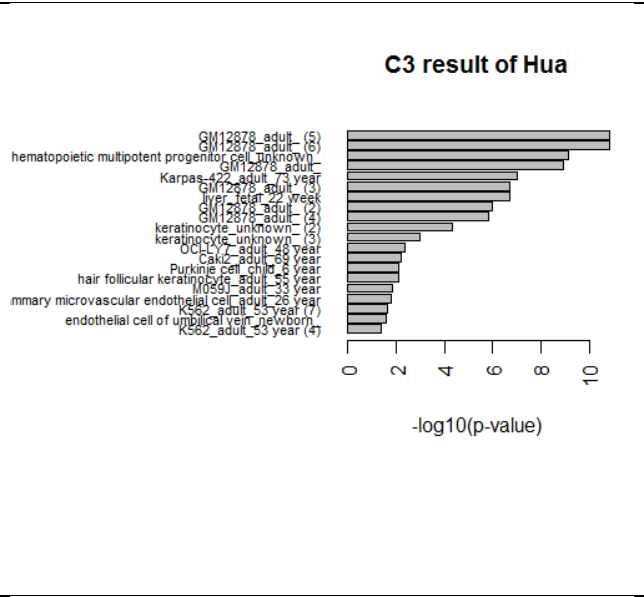


Supplementary Table 5: Detail test results of selected five samples after quantile normalization of the data sets with parameter settings $n=1000$, $t=0.10$

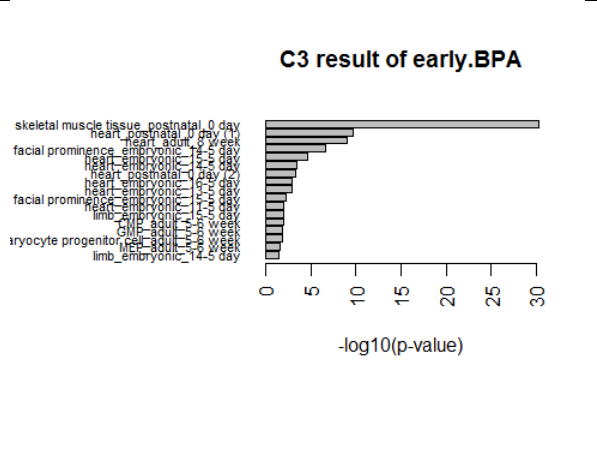
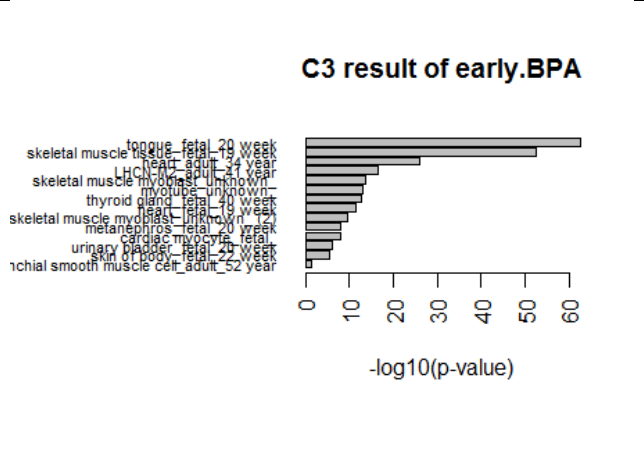
Data	Test result with human compendium	Test result with mouse compendium
Rat (<i>Rattus norvegicus</i>) kidney GSE43013 PMID: 25677554	<p>C3 result of rnorvegicus.kd</p> <p>-log10(p-value)</p>	<p>C3 result of rnorvegicus.kd</p> <p>-log10(p-value)</p>
Cattle (<i>Bos taurus</i>) liver GSE43013 PMID: 25677554	<p>C3 result of btaurus.lv</p> <p>-log10(p-value)</p>	<p>C3 result of btaurus.lv</p> <p>-log10(p-value)</p>
Zebrafish (<i>Danio rerio</i>) control brain GSE74754 PMID: 27935819	<p>C3 result of drerio.brain.ctl</p> <p>-log10(p-value)</p>	<p>C3 result of drerio.brain.ctl</p> <p>-log10(p-value)</p>

Dolphin
([Tursiops truncatus](#))
blood (hua)
PMID: [27608714](#)

Note:
GM12878
is a
lymphoblastoid cell
line
produced
from the
blood of a
female
donor



Monkey
([Macaca mulatta](#))
skeletal
muscle (early
BPA)
PMID: [24586524](#)



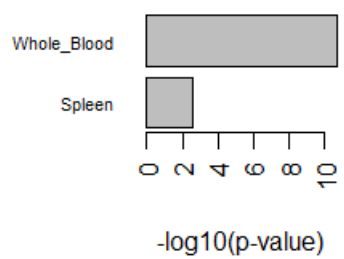
Supplementary Table 6: Detail test results of selected five samples of the data sets with GTEx human compendium

Data	Paramaters: $n=500, t=0.05$	Paramaters: $n=1000, t=0.10$																						
Rat (<i>Rattus norvegicus</i>) kidney GSE43013 PMID: 25677554	<p>C3 result of rnorvegicus.kd</p> <table><tr><th>Sample</th><th>$-\log_{10}(\text{p-value})$</th></tr><tr><td>Kidney_Cortex</td><td>5.5</td></tr><tr><td>Heart_Left_Ventricle</td><td>4.5</td></tr><tr><td>Heart_Atrial_Appendage</td><td>2.5</td></tr><tr><td>Liver</td><td>2.0</td></tr></table>	Sample	$-\log_{10}(\text{p-value})$	Kidney_Cortex	5.5	Heart_Left_Ventricle	4.5	Heart_Atrial_Appendage	2.5	Liver	2.0	<p>C3 result of rnorvegicus.kd</p> <table><tr><th>Sample</th><th>$-\log_{10}(\text{p-value})$</th></tr><tr><td>Kidney_Cortex</td><td>18</td></tr><tr><td>Liver</td><td>15</td></tr><tr><td>Heart_Left_Ventricle</td><td>8</td></tr><tr><td>Adrenal_Gland</td><td>4</td></tr><tr><td>Heart_Atrial_Appendage</td><td>2</td></tr></table>	Sample	$-\log_{10}(\text{p-value})$	Kidney_Cortex	18	Liver	15	Heart_Left_Ventricle	8	Adrenal_Gland	4	Heart_Atrial_Appendage	2
Sample	$-\log_{10}(\text{p-value})$																							
Kidney_Cortex	5.5																							
Heart_Left_Ventricle	4.5																							
Heart_Atrial_Appendage	2.5																							
Liver	2.0																							
Sample	$-\log_{10}(\text{p-value})$																							
Kidney_Cortex	18																							
Liver	15																							
Heart_Left_Ventricle	8																							
Adrenal_Gland	4																							
Heart_Atrial_Appendage	2																							
Cattle (<i>Bos taurus</i>) liver GSE43013 PMID: 25677554	<p>C3 result of btaurus.lv</p> <table><tr><th>Sample</th><th>$-\log_{10}(\text{p-value})$</th></tr><tr><td>Liver</td><td>48</td></tr></table>	Sample	$-\log_{10}(\text{p-value})$	Liver	48	<p>C3 result of btaurus.lv</p> <table><tr><th>Sample</th><th>$-\log_{10}(\text{p-value})$</th></tr><tr><td>Liver</td><td>115</td></tr><tr><td>Pancreas</td><td>5</td></tr><tr><td>Inney_Cortex</td><td>3</td></tr></table>	Sample	$-\log_{10}(\text{p-value})$	Liver	115	Pancreas	5	Inney_Cortex	3										
Sample	$-\log_{10}(\text{p-value})$																							
Liver	48																							
Sample	$-\log_{10}(\text{p-value})$																							
Liver	115																							
Pancreas	5																							
Inney_Cortex	3																							
Zebrafish (<i>Danio rerio</i>) control brain GSE74754 PMID: 27935819	<p>C3 result of drerio.brain.ctl</p> <table><tr><th>Sample</th><th>$-\log_{10}(\text{p-value})$</th></tr><tr><td>Brain_Cerebellum</td><td>1.5</td></tr></table>	Sample	$-\log_{10}(\text{p-value})$	Brain_Cerebellum	1.5	<p>C3 result of drerio.brain.ctl</p> <table><tr><th>Sample</th><th>$-\log_{10}(\text{p-value})$</th></tr><tr><td>Brain_Frontal</td><td>14</td></tr><tr><td>Brain_Cerebellar_Hemisphere</td><td>10</td></tr><tr><td>Brain_Cortex</td><td>8</td></tr><tr><td>Brain_Cerebellum</td><td>7</td></tr></table>	Sample	$-\log_{10}(\text{p-value})$	Brain_Frontal	14	Brain_Cerebellar_Hemisphere	10	Brain_Cortex	8	Brain_Cerebellum	7								
Sample	$-\log_{10}(\text{p-value})$																							
Brain_Cerebellum	1.5																							
Sample	$-\log_{10}(\text{p-value})$																							
Brain_Frontal	14																							
Brain_Cerebellar_Hemisphere	10																							
Brain_Cortex	8																							
Brain_Cerebellum	7																							

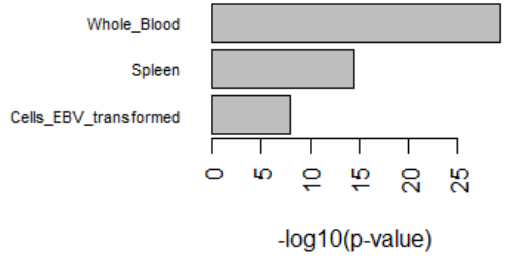
Dolphin
([Tursiops truncatus](#))
blood (hua)
[GSE78770](#)
PMID: [27608714](#)

Note:
GM12878
is a
lymphobla
stoid cell
line
produced
from the
blood of a
female
donor

C3 result of Hua

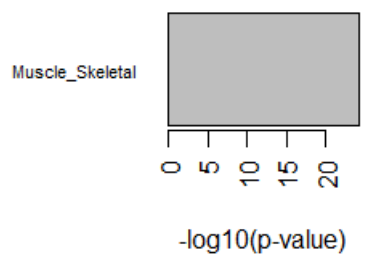


C3 result of Hua

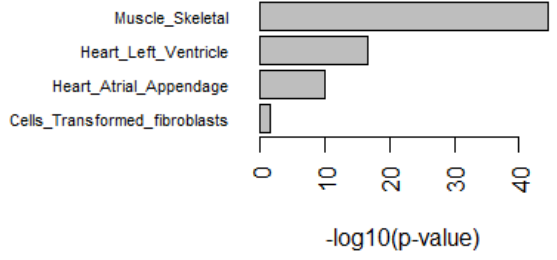


Monkey
([Macaca mulatta](#))
skeletal
muscle (early
BPA)
[GSE53393](#)
PMID: [24586524](#)

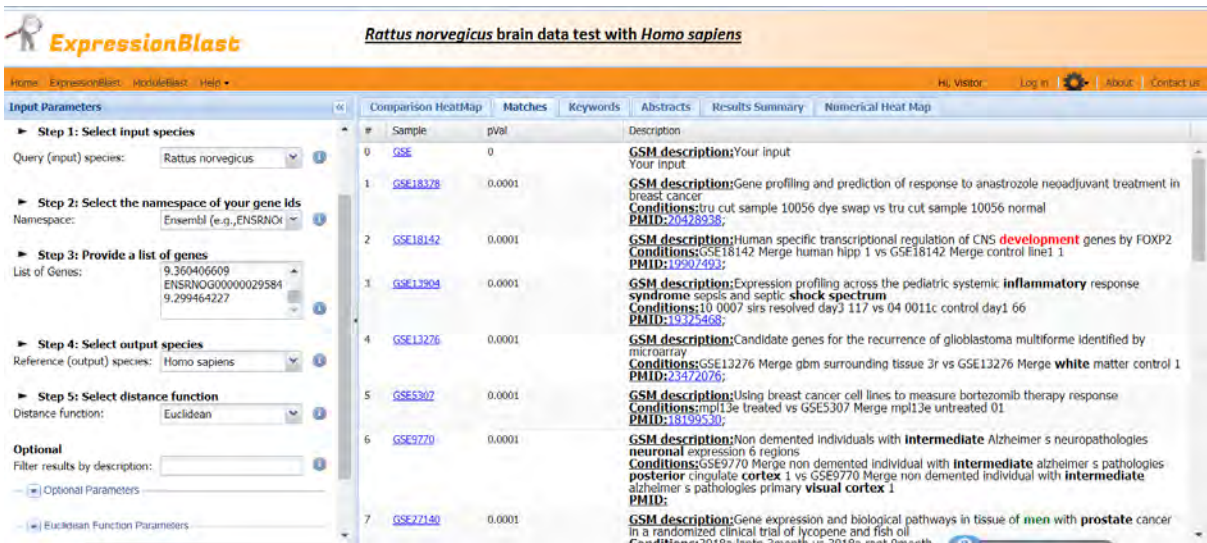
C3 result of early.BPA



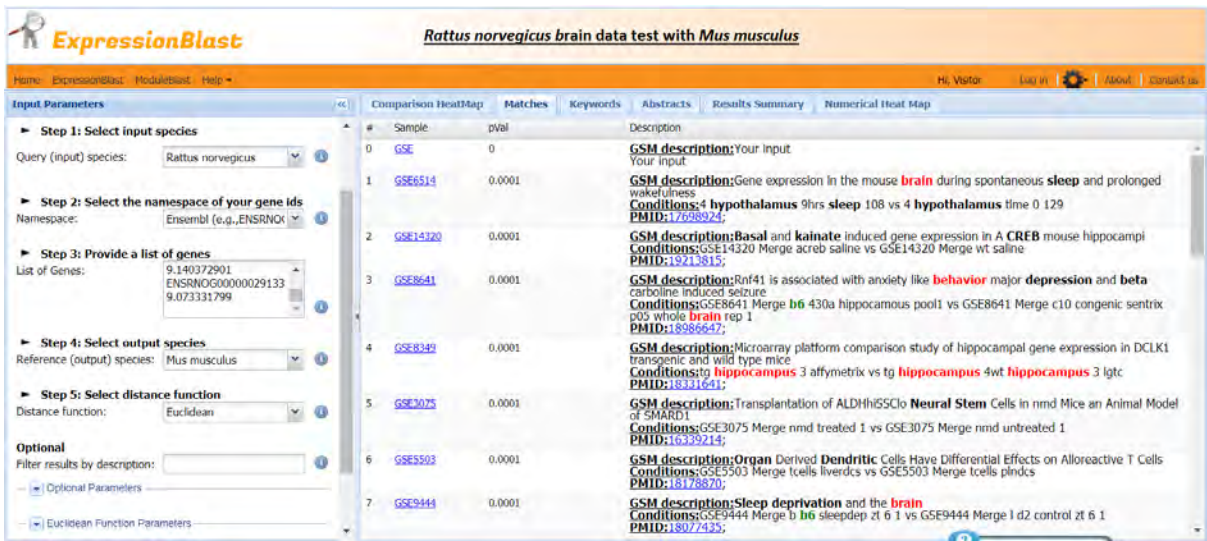
C3 result of early.BPA



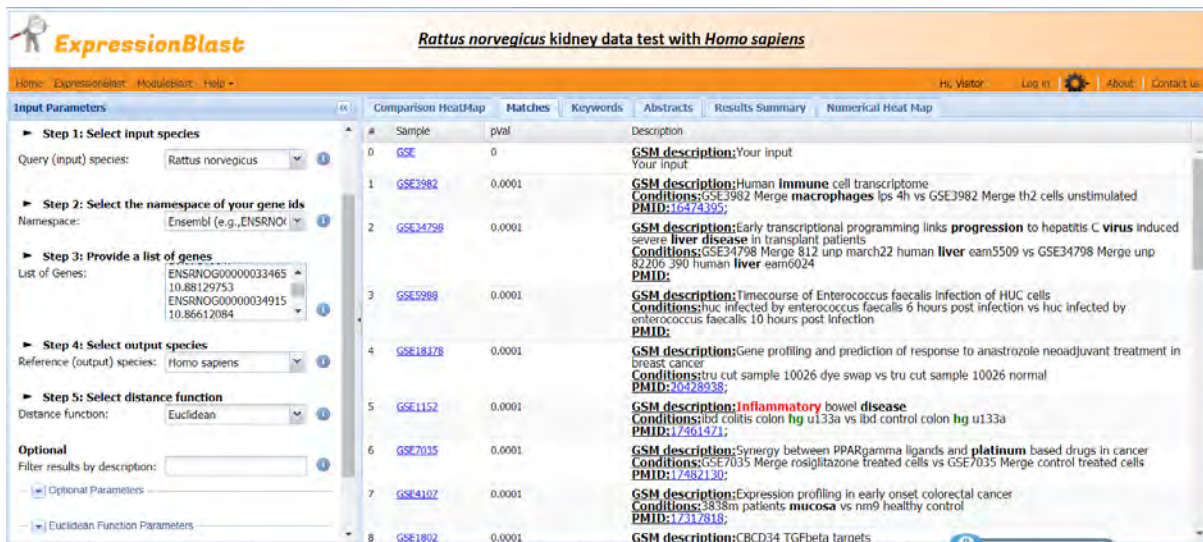
Supplementary Figure 1 : ExpressionBlast test result (screenshot) for 3 different tissues of *Rattus norvegicus*



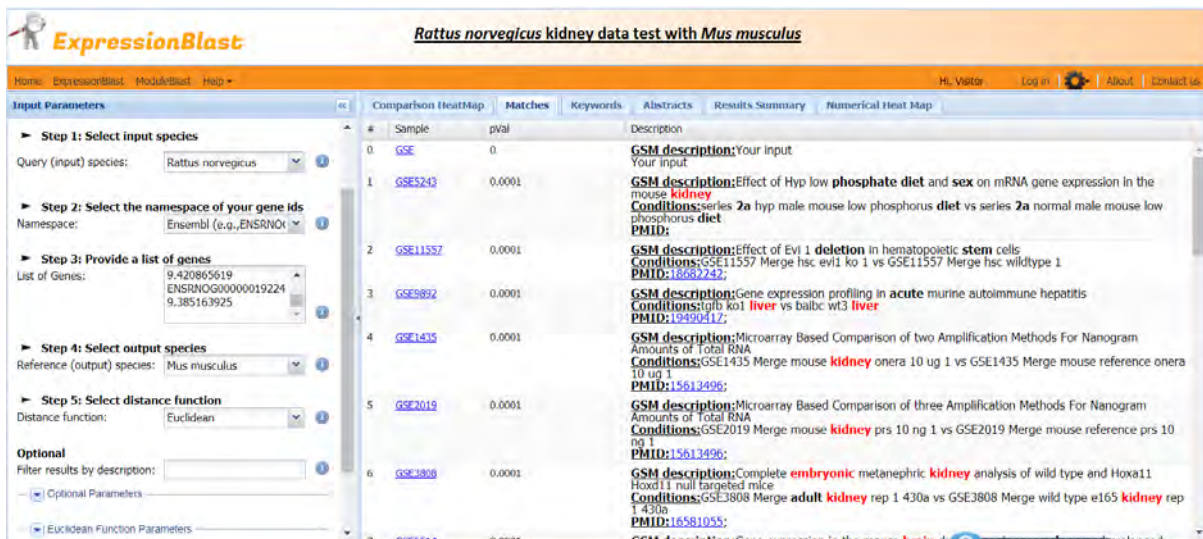
Supplementary Figure 1(a): Test result of *R. norvegicus* brain sample data with *H.sapiens*



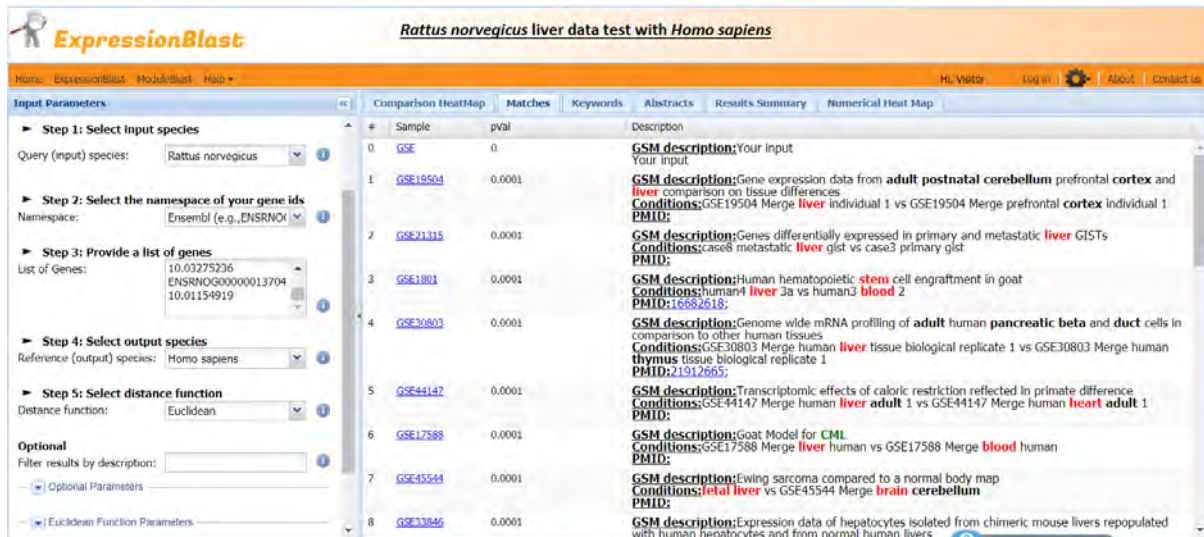
Supplementary Figure 1(b): Test result of *R. norvegicus* brain sample data with *M.musculus*



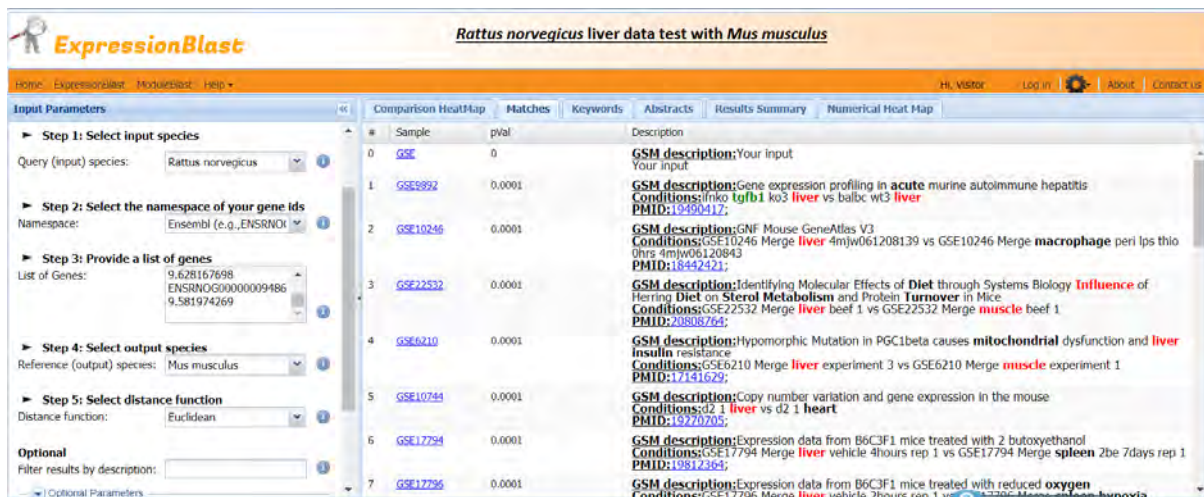
Supplementary Figure 1(c): Test result of *R. norvegicus* kidney sample data with *H.sapiens*



Supplementary Figure 1(d): Test result of *R. norvegicus* kidney sample data with *M.musculus*



Supplementary Figure 1(e): Test result of *R. norvegicus* liver sample data with *H.sapiens*



Supplementary Figure 1(f): Test result of *R. norvegicus* liver sample data with *M.musculus*