Functional genomics approaches to understanding heterogeneous tissues in health and disease

Adam E. Handel

Green Templeton College

Thesis submitted for the degree of DPhil

Abstract

Functional genomics approaches offer an unparalleled window into cellular biology. It is particularly challenging to apply these methods to heterogeneous tissues, in which multiple cell types or transcriptomic identities co-exist. I applied single cell and bulk functional genomics approaches in an attempt to understand aspects of the biology underlying cerebral cortex and thymic epithelial cells. These are both extremely heterogeneous tissues but the nature of the heterogeneity differs: in cortex there are a large number of different cell types, whereas in thymic epithelial cells, despite a limited number of cell types, individual cells express virtually every gene in the transcriptome. In **Chapter 1**, I provide a general overview of the currently available functional genomics tools and briefly review relevant aspects of cortical and thymic biology. The first three results chapters examine functional genomics approaches to cerebral cortex. **Chapter 2** details the application of single cell functional genomics techniques to examine the extent to which stem cell-derived cortical neurons resemble primary human cortical neurons. In **Chapter 3**, I use RNA sequencing in cortical neurons derived from wild-type or *PSEN1* mutant stem cells in an attempt to identify an Alzheimer's disease-relevant gene signature and to assess how much variability there is between batches of stem cell differentiation. **Chapter 4** applies DNase-seq footprinting to primary brain tissue in order to understand how common variants associated with altered gene expression or susceptibility to neurological disease might exert their effects. The last two results chapters focus on thymic epithelial cells. In **Chapter 5**, I integrate multiple functional genomics datasets to identify a set of genes targeted by *Foxn1*, a criticial transcription factor in thymus development and function. **Chapter 6** use single cell RNA-seq from mature medullary thymic epithelial cells to gain an insight into how promiscuous gene expression in the thymus is orchestrated at the level of individual cells. Finally, in **Chapter 7**, I summarise the implications of my findings for cortical and thymic biology and speculate on the potential directions that functional genomics may take in the future to better understand heterogeneous tissues.

Acknowledgements

The production of this thesis would not have been possible without the kind help of many outstandingly talented and dedicated individuals. I have acknowledged their respective contributions at the end of each chapter. I am hugely grateful for the input I have received from all members of the Cader, Holländer and Ponting groups. Without their generosity of spirit and critical appraisal of my work, the body of research presented here would have been considerably impoverished.

I am also deeply appreciative of the financial and personal support afforded to me by the Wellcome Trust, in the guise of a Wellcome Trust Research Training Fellowship (100643/Z/12/Z). In particular, I would like to thank the Wellcome Trust for being supportive of unexpected changes in the direction of my programme of research. Without the freedom from clinical responsibilities that such an award permits, I would not have had the breadth of time to pursue my academic interests. I am very grateful too for the resources made available for my research by StemBANCC and the Wellcome Trust Strategic Award.

My supervisors, Prof. Chris Ponting, Prof. Zameel Cader and Prof. Georg Holländer have been extremely supportive through my DPhil. Their advice has been invaluable and the trust they placed in me by allowing me the freedom to choose my own research directions has been unparalleled. They provided me with encouragement when this was required and were always ready with an honest appraisal of my work. I look forward to continued collaboration with them in the future.

Finally, I must thank my family and friends, without whom this would have been frankly impossible. They have remained patient with me through all of the vicissitudes of a doctorate and have buoyed me up through any difficulties. In particularly, I would like to thank my parents, who have been incredibly supportive throughout, and my wife, Katharine, in whose company even the blackest day can seem brighter.

No DPhil can ever be conducted in isolation; without the contributions of all the individuals listed above this thesis would be almost entirely without merit.

Contents

Chapter 1 Introduction

1.1 Abstract

This thesis aims to explore the biology and pathophysiology of two extremely heterogeneous tissues, cerebral cortex and thymic epithelium, using functional genomics approaches. This chapter provides a brief overview of the current genomics methods available, highlighting the particular utility of these in understanding underlying cellular processes. It then reviews the current state of understanding relating to cellular identity within cerebral cortex and thymic tissue in health and disease. Challenges in applying functional genomics methods to these heterogeneous tissues are discussed and key questions that this thesis attempts to answer are identified.

1.2 Overview of genomics methods

1.2.1 Available genomics methods

The proliferation of genomics approaches that can be applied to biological systems has been extraordinary, fuelled particularly by developments in next generation sequencing technologies.[1,2] It is now possible to interrogate the genome, transcriptome, transcription factor binding, epigenome and three-dimensional structure of DNA using next generation sequencing (NGS)-based techniques (**Figure 1**).[3–9] NGS generates short reads by sequential single base extension that can be mapped against a reference genome or transcriptome.[10] Although based on quite different technology, whole proteome analysis has also made it possible to assess gene expression at the post-translational level.[11] Integrating multiple different measures of cellular physiology can provide insights into biology that any specific method could not deliver.[2]

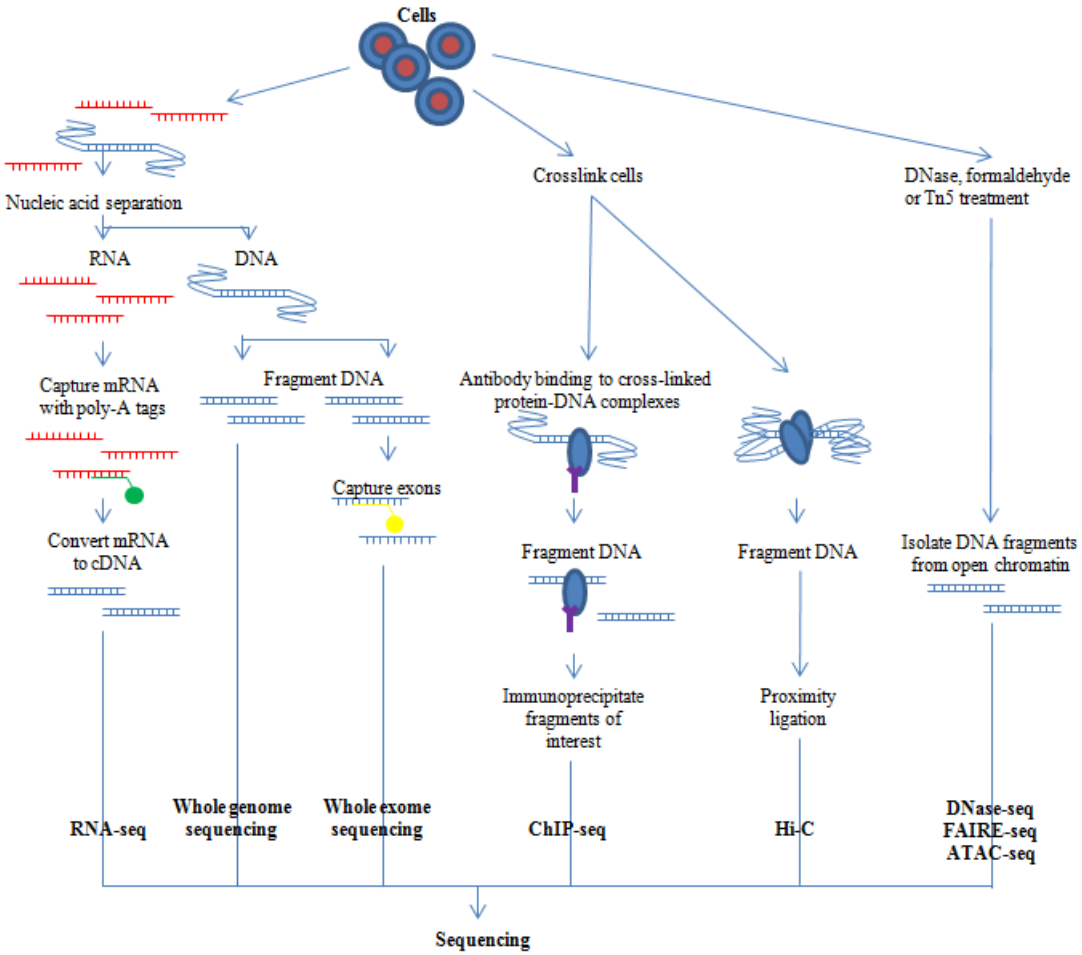

Figure 1 Overview of next generation sequencing methods. Modified from Handel *et al*.[1]

1.2.2 Genome sequencing

DNA can be sequenced using whole genome, whole exome or targeted approaches.[3] Variants can be inferred from short read assemblies derived from NGS. For the identification of previously identified common variants, targeted single nucleotide polymorphisms (SNP) arrays are likely to be more accurate than NGS-based variant calling.[12] Long read sequencing technologies will revolutionise genome sequencing although currently high error rates coupled with high costs preclude accurate long read eukaryotic genome assemblies without parallel NGS datasets.[13]

1.2.3 Transcriptome sequencing

RNA can be interrogated using NGS approaches by reverse transcribing this to cDNA prior to sequencing. Specific selection strategies can enable the enrichment of particular subsets of RNA, such as mature mRNA, long non-coding RNA, microRNA, nascent RNA, enhancer-associated RNA or ribosome-associated RNA.[4,14–16] The most common use of RNA-seq is transcript

quantification of mRNA, at which it frequently outperforms gene expression microarrays.[17] However, the correlation between levels of mRNA and protein is imperfect and non-uniform between different classes of genes.[18] In particular, parallel measurements of mRNA and protein turnover in fibroblasts demonstrated that transcription factors (TFs) were characterised by unstable mRNA and protein, RNA-processing genes were characterised by stable protein and unstable mRNA transcripts, and extracellular proteins were characterised by stable mRNA transcripts and unstable protein.[19] Taking into account sequence-specific features of genes that were predictive of the translated protein's rate of degradation improved the correlation between global mRNA levels and protein abundance in a human medulloblastoma cell line.[20] The relationship between mRNA and protein levels is further confounded by the recent observation in a human cell line that individual mRNA transcripts can flip dynamically between active translation and inactivity.[21] Taken together, these findings highlight the importance of considering the imperfect relationship between mRNA transcript abundance and protein levels when interpreting the results of any transcriptomic experiments.

1.2.4 Chromatin immunoprecipitation and sequencing

A combination of chromatin immunoprecipitation and sequencing (ChIP-seq) can infer the sites of transcription factor binding or histone modifications.[5] The addition of exonuclease treatment can improve the resolution of peaks.[22,23]

1.2.5 Epigenome sequencing

NGS-based approaches can interrogate a number of epigenomic features, including histone modifications (see ChIP-seq above), DNA methylation and hydroxymethylation, and DNA accessibility.

5-methylcytosine residues can be detected either by treating DNA with bisulphite treatment or the use of an antibody specific to 5-methylcytosine.[24,25] Combining bisulphite treatment with a specific oxidation step enables the identification of 5-hydroxymethylcytosine residues.[24] Typically methylated cytosines are enriched near repressed genes whereas hydroxymethylated

cytosines are thought to be an intermediate state between methylated and unmethylated cytosines related to genes playing a role in development and pluripotency.[24]

DNA accessibility can be assessed using a number of different techniques, including DNase (DNase-seq), micrococcal nuclease (MNase-seq), formaldehyde (FAIRE-seq) or transposase (ATAC-seq) treatment.[7,8,26,27] These approaches identify areas of open chromatin, which can be bound by transcription factors and RNA polymerases. Combining high depth sequencing with DNase-seq, MNase-seq or ATAC-seq can specifically identify regions of protected DNA within open chromatin likely to correspond to protein-DNA interactions.[28,29]

1.2.6 Three dimensional structure of DNA

Sequencing cross-linked fragments of DNA can identify DNA-DNA interactions and permit inference of the three dimensional organisation of the genome.[30,31] This can be important in the identification of long-range enhancer-promoter interactions.

1.2.7 Single cell approaches

Recently, protocols capable of capturing information on each of the above aspects of cellular physiology, admittedly often to a more limited extent, have also been developed for single cells.[32–38] These were able to identify transitory stages of cellular identity that previously had only been suspected and isolate the importance of paracrine effects on cellular specification.[39,40] Taken together, these findings suggested that specific cell types may occupy distinct transcriptomic spaces but with a large degree of variability between individual cells of a particular type within a tissue.[41] Single cell NGS methods provide a window into cellular heterogeneity that bulk NGS approaches cannot capture. Single cell RNA-seq approaches have already started dissecting out cell type populations from heterogeneous tissues including lymphocytes, lung epithelium, myoblasts, olfactory neurons and cerebral cortex.[42–45,39,40,46,47]

Protocols which combine multiple techniques can provide information on how genomic, epigenomic, proteomic and transcriptomic variation interact in individual cells.[48–50] This has the potential to redefine current understanding of cellular identity by integrating regulatory and transcriptomic signals within individual cells.

1.3 Cerebral cortex

1.3.1 Cortical development

The human cerebral cortex is a structure within the central nervous system involved with information processing and storage. It consists primarily of six layers of organised neurons and glia.[51] Pyramidal cells and glia are produced in a stereotypical fashion from radial glia located within the subventricular zone: initially deep layer pyramidal neurons, then upper layer pyramidal neurons and finally glia.[52] Conversely, cortical interneurons migrate tangentially through the developing cortex from transitory forebrain structures called ganglionic eminences.[53] There are specific transcriptomic and proteomic programmes that show temporal and spatial specificity within cortex (**Figure 2a**).[54,55,51,56,57] Specifically certain transcription factors are restricted to particular cortical layers.[58]

A powerful method of studying human corticogenesis is to use *in vitro* modelling techniques. Cortical neurons can be derived from embryonic stem cells (hES), induced pluripotent stem cells (iPSC) (**Figure 2b**) or direct transdifferentiation from other somatic cells.[59–63] Studying the temporal dynamics of hES-derived corticogenesis can identify sets of genes associated with different stages of cortical development and neuronal maturation.[64] However, the heterogeneity of cell types found in primary cerebral cortex limits the extent that bulk primary tissue samples can be meaningfully compared to stem cell-derived cortical neurons.[65]

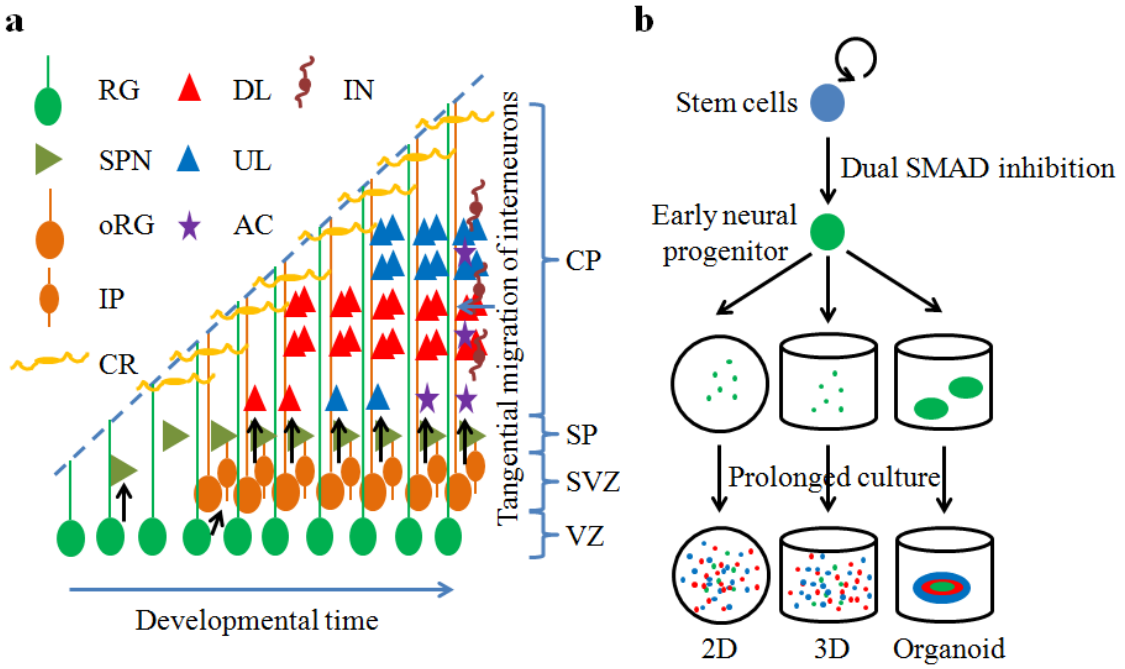

Figure 2 Overview of cortical development. (**a**) *In vivo* corticogenesis. RG = radial glia; SPN = subplate neurons; oRG = outer radial glia; IP = intermediate progenitors; CR = Cajal-Retzius cells; DL = deep layer pyramidal neurons; UL = upper layer pyramidal neurons; AC = astrocytes; IN = interneurons; VZ = ventricular zone; SVZ = subventricular zone; SP = subplate; CP = cortical plate. (**b**) *In vitro* models of corticogenesis.

1.3.2 Studying cellular heterogeneity in cerebral cortex

There is still regional and layer-specific transcriptomic diversity in the adult central nervous system.[51,54,55,57] This transcriptomic heterogeneity is reflected in a broad range of electrophysiological profiles.[66,67] This cellular heterogeneity has made transcriptomic characterisation of cerebral cortex particularly challenging.

One approach that can help dissect this mixture of multiple populations of distinct cell types is to sort cells by the expression of particular markers.[68] However, many of the specific markers currently known are intracellular, which can be problematic in species where transgenic models are unavailable. Furthermore, this will falsely combine cell types if there are no identified specific markers.

An approach that can obviate the problems generated by cortical cell type heterogeneity is single cell RNA-seq, since this can resolve individual cell types in an unbiased fashion. Single cell transcriptomics has been successfully applied to mouse and human brain tissue.[44,47,46] There were clear divisions observed between neurons, oligodendrocytes, astrocytes, endothelial cells and

microglia. Neurons from fetal and adult brains were clearly distinct, emphasising that the temporal changes in gene expression within cortex were not merely driven by different proportions of cell types but likely reflected genuine transcriptomic differences in the same cell type.[46,47] Identifying subpopulations of neurons is far more difficult, since the transcriptomic differences are more subtle and it is currently unclear how many subpopulations actually exist. Based on single cell RNA-seq data from mouse somatosensory neocortex and hippocampus, there are at least nine classes of excitatory neurons and 16 classes of interneurons.[47] In addition, this revealed largely unsuspected diversity amongst non-neuronal cells within the brain. This is highly likely to be a severe underestimation of the cellular complexity within cerebral cortex and larger single cell RNA-seq datasets will undoubtedly identify further subpopulations of cells.

Understanding the biological relevance of the subpopulations of cortical neurons identified is more difficult. However, an increasing understanding of specific marker genes provided by single cell RNA-seq studies can help correlate electrophysiological profiles with transcriptomic read-outs. Single cell approaches which can combine electrophysiology with single cell transcriptomics have already started to annotate neuronal classes identified from transcriptomic data with cellular phenotypes.[69] Similarly, methods of performing *in situ* transcriptomics at subcellular resolution enables expression profiles to be understood with relation to the native structure of cortical networks.[70–72]

1.3.3 The importance of cellular heterogeneity to neurological disease

Neurological diseases as a whole constitute a substantial burden of human morbidity and mortality.[73] Many of these show brain region- and cell type-specific effects, which can mean that transcriptomic, epigenomic or proteomic signatures of disease may be difficult to detect when analysing bulk samples.[74–77] Despite this, considerable progress has been made in untangling the genetic aetiology of neurological disorders through genome-wide association studies.[78,79] Integrating the location of common variants associated with disease susceptibility has provided clear evidence that many risk variants are located within open chromatin.[80,81] However, the precise mechanism by which these variants exert their effects is unclear.

In the case of Alzheimer's disease, a common cause of dementia, and responsible for high levels of morbidity and mortality worldwide, there is clear evidence for cell type-specific effects.[82] Alzheimer's disease is typified by the accrual of Aβ plaques and hyper-phosphorylated tau tangles with progressive neurodegeneration. All major brain cell types have been strongly implicated in Alzheimer's disease pathophysiology.[83–86] This may be one reason why the findings of transcriptomic studies examining Alzheimer's disease have rarely been replicable.[87] Even when considering only the effect of Alzheimer's disease on neurons, there is a predilection for neurodegeneration to occur in specific layers of the cortex, so even studying purified neurons would be unlikely to resolve matters completely.[88,89]

One potentially promising strategy in modelling Alzheimer's disease pathogenesis has been to apply stem cell models of corticogenesis to patient- and control-derived iPSC.[90] This has the advantage that the resultant cultures are almost entirely comprised of cortical neurons.[61,91] These neurons can show disease-relevant alterations in Aβ processing and tau phosphorylation but consistent transcriptomic differences still prove elusive.[92,93,85] This may be because the effective age of iPSC is reset and so the resultant neurons are fetal in nature. It is also true that the genetic background of iPSC is a strong driver of the resultant transcriptome.[94] Therefore, background variants may obscure the effect of any disease-associated variants. The use of genome engineering either to insert a disease-associated mutation into a control iPSC or to correct a disease-associated mutation in a patient iPSC should reduce variability caused by the genetic background.[95,96]

The combination of single cell transcriptomics and stem cell models may be able to identify cell type-specific disease effects and potentially suggest therapeutic targets.

1.3.4 Key questions for functional genomics in cerebral cortex

1.3.4.1 How closely do stem cell-derived cortical neurons resemble neurons from primary cerebral cortex?

1.3.4.2 Does the degree of variability between cortical neurons derived from different iPSC allow for meaningful comparisons to be made between patient- and control-derived iPSC models of corticogenesis?

1.3.4.3 Does the use of genome engineering reduce variability between iPSC-derived cortical neurons?

1.3.4.4 By what mechanism do common variants associated with neurological disease exert their effect on disease susceptibility?

1.4 Thymus

1.4.1 Thymic development and function

The thymus is a lymphoid organ that develops from the third pharyngeal pouch and T-cell progenitors called thymocytes.[97] Thymic size relative to body weight increases during the first year of life, at which point the thymus begins to involute.[98] The thymus is responsible for the positive and negative selection of developing T-lymphocytes, known as thymocytes (**Figure 3**).[99] Positive selection of thymocytes capable of interaction with T-cell receptors (TCRs) occurs within the cortex through interactions between cortical thymic epithelial cells (cTEC) and T-cell precursors. Negative selection then takes place through interactions between cells positively selected within the cortex and medullary thymic epithelial cells (mTEC). This process aims to ensure that the resultant mature T-cell population can function effectively but does not contain auto-reactive cells prone to cause autoimmunity.

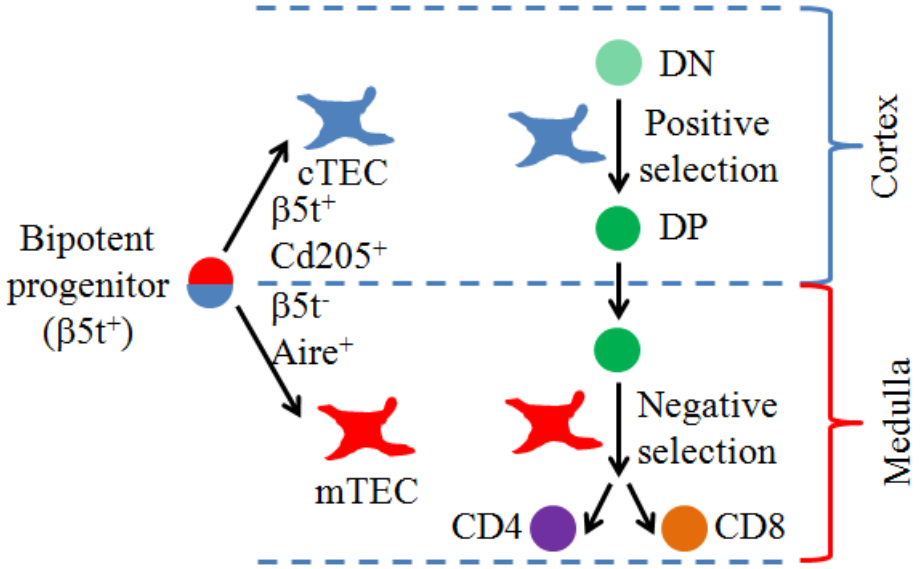

Figure 3 Schematic of thymic development and function. DN = double negative thymocyte ($Cd4^-Cd8^-$); DP = double positive thymocyte ($Cd4^+Cd8^+$); CD4 = $Cd4^+$ T-cell; CD8 = $Cd8^+$ T-cell. NB: negative selection may also occur in the cortex.[100,101]

cTEC and mTEC populations develop from a common $\beta5t^+$ progenitor cell.[97] This was shown in a mouse model in which green fluorescent protein (GFP) expression was dependent on previous expression of β5t through a Cre-LoxP system.[102] All thymic epithelial cell (TEC) populations were labelled by GFP, indicating a common bipotent TEC progenitor.

Successful TEC development depends critically upon the expression of a forkhead TF called *Foxn1*, the expression of which is restricted to TEC and keratinocytes.[103] The genetic ablation of *Foxn1* produces a "nude" mouse, which is athymic and has no visible coat.[104] Orthologous mutations in humans result in immunodeficiency, alopecia and nail dystrophy.[105] *Foxn1* plays an important role throughout TEC development and also affects TEC function and maintenance.[106,107] However, despite the importance of *Foxn1* to thymic biology, the molecular mechanisms by which it works and its gene targets have remained largely unknown.

The thymic microenvironment is extremely important for normal TEC development, and deficits within multiple receptor-ligand pairs can result in thymic dysfunction (*Rank*, *Cd40*, *Ltβr*, *Fgf2r*, *frizzled* and *Notch1*).[108] However, none of these show the profound thymic phenotype that characterises *Foxn1* deficiency.

1.4.2 Promiscuous gene expression

In order for thymocyte education by TEC to take place, virtually all proteins expressed within an organism that might be encountered by mature, peripheral T-cells need to be already expressed within the thymus. This process of promiscuous gene expression (PGE) in mTEC is orchestrated by another TF called *Aire*.[109] Its importance to thymic biology is illustrated by the fact that mutations in *Aire* or the human orthologue result in the development of autoimmunity, specifically autoimmune polyendocrine syndrome type 1 (APS1) in humans.[110–112] *Aire* binds to genes indicated by the repressive chromatin mark H3K27me3 to induce the expression of 3,980 genes, many of them showing tissue restricted expression.[113] The co-ordination of PGE within the mTEC compartment may involve stochastic expression within intrachromosomal regions in individual cells.[114,115] Correlative data suggest that chromatin accessibility, but not DNA methylation, is important in PGE. However, in the absence of models interfering with epigenomic processes, the precise role epigenomic remodelling plays in PGE remains uncertain. Recently, *Fezf2* was implicated in regulating PGE but the evidence supporting this is much less convincing than for *Aire* (Georg Holländer, personal communication).[116]

The heterogeneity manifest in the mTEC compartment due to PGE (and, to a lesser extent, in cTEC) means that, unlike in cerebral cortex in which transcriptomic heterogeneity is driven by different cell types, in mTEC individual cells of the same presumed cell type have wildly different transcriptomes.[113–115] The application of single cell RNA-seq in combination with conditional mutations in epigenome remodelling enzymes to this system has the potential to untangle the role epigenomics plays in cellular identity and PGE in the thymus.

1.4.3 Thymus pathophysiology and autoimmune disease

In addition to the syndromic effects associated with *Aire* deficiency, the thymus clearly plays a role in autoimmune diseases. One indirect method of assessing thymic function is to measure T-cell receptor excision circles (TREC), which are produced during T-cell maturation.[117] Studies have reported reduced levels of TREC in patients relative to controls for rheumatoid arthritis, systemic lupus erythematosus and multiple sclerosis.[118,119] Conversely, increased levels of TREC

were reported for type 1 diabetes mellitus and Grave's disease.[120,121] The interpretation of these studies are limited by the fact that TREC levels are altered by the rate of cell proliferation within the thymus.[122] However, with these caveats in mind, this suggests that the study of thymic biology and dysfunction may shed light on some processes that contribute to common autoimmune diseases.

1.4.4 Key questions for functional genomics in thymus

1.4.4.1 How does *Foxn1* function in the development and maintenance of TEC?

1.4.4.2 How do TEC co-ordinate promiscuous gene expression?

1.4.4.3 What role does dynamic remodelling of the epigenome play in promiscuous gene expression?

1.5 Conclusions

Studying the functional genomics of heterogeneous tissues, as exemplified by cerebral cortex and thymus, is challenging. Undoubtedly increasing the resolution of genomics methods to interrogate individual cells can identify important biological processes that would be missed by analysing bulk samples. This thesis will attempt to integrate multiple genomics datasets, including single cell RNA-seq, in order to understand the functional genomics of cerebral cortex and thymus in health and disease.

Chapter 2 Single cell transcriptomic phenotyping of iPSC-derived cortical neurons

2.1 Abstract

Induced pluripotent stem cell (iPSC)-derived cortical neurons potentially present a powerful new model to understand corticogenesis and neurological disease. Previous work has established that differentiation protocols can produce cortical neurons, but little has been done to characterise these at cellular resolution. In particular, it is unclear to what extent *in vitro* two-dimensional, relatively disordered culture conditions recapitulate the development of *in vivo* cortical layer identity. Single cell multiplex RT-qPCR was used to interrogate the expression of genes previously implicated in cortical layer or phenotypic identity in individual cells. 93.6% of single cells derived from iPSCs expressed genes indicative of neuronal identity. High proportions of single neurons derived from iPSCs expressed glutamatergic receptors and synaptic genes. 68.4% of iPSC-derived neurons expressing at least one layer marker could be assigned to a laminar identity using canonical cortical layer marker genes. I compared single cell RNA-seq of the iPSC-derived neurons to available single cell RNA-seq data from human fetal and adult brain and found that iPSC-derived cortical neurons closely resembled primary fetal brain cells. Unexpectedly a subpopulation of iPSC-derived neurons co-expressed canonical fetal deep and upper cortical layer markers. However this appeared to be concordant with data from primary cells. My results therefore provide reassurance that iPSC-derived cortical neurons are highly similar to primary cortical neurons at the level of single cells but suggest that current layer markers, although effective, may not be able to disambiguate cortical layer identity in all cells.

2.2 Hypotheses

2.2.1 iPSC-derived cortical neurons will express key neuronal genes relevant to cortical function and cortical layer identity.

2.2.2 iPSC-derived cortical neurons will resemble primary human cortical neurons at the level of the single cell transcriptome.

2.3 Introduction

Investigating the cellular basis of neurological diseases, especially those affecting the central nervous system (CNS), is rendered particularly challenging by the inaccessibility of the tissues involved. Induced pluripotent stem cell (iPSC)-based models have the potential to allow *in vitro* investigation of these tissues in human samples from patients affected by such diseases and, importantly, how disease progresses over time.[123] Protocols have been developed that are capable of generating cortical cells from human iPSCs, which appear to adopt specific cortical layer identities and develop functional synapses.[61,91,124–126] In this chapter, I used single cell functional genomics methods to assess the extent to which iPSC-derived cortical neurons resembled primary human cortical neurons.

Most transcriptomic studies of iPSC-derived cortical neurons have examined expression in samples pooled from a whole population of cells so would miss potential cell type-specific or layer-specific effects.[93,127] The development of single cell gene expression platforms, such as microfluidic chips, as well as evolving chip-free single cell RNA-seq technologies, make such studies a viable method to investigate iPSC-derived cortical neuron cultures at single cell resolution.[128,129] This has the advantage that the relative abundance of different cell types may be discerned and so comparisons between iPSC-derived and primary tissue can be made at the level of individual cells.

A core set of cortical layer markers has been used within the stem cell research community to establish the presence of neurons with different layer identities in iPSC-derived cortical neuronal cultures.[61,91,130] However, many of these markers were inferred from studies of mouse brain or

immunohistochemistry of human fetal brain, so the robustness of such markers in assigning layer identity to single neurons by single cell transcriptomics approaches is unknown.[54,58]

The degree of heterogeneity present in cortical neurons derived from iPSCs is a critically important aspect of *in vitro* models to understand. Layer-specific and phenotypic cellular identity is particularly relevant prior to applying such models to address disease-specific hypotheses.

Cortical neurons derived from iPSCs using such methods have been used to study a wide variety of neurodevelopmental and neurodegenerative conditions, and recapitulate disease-relevant phenotypes.[123] In the case of Alzheimer's disease, iPSC-derived cortical neurons displayed aberrant Aβ secretion and tau phosphorylation.[92,93] iPSC lines from autism spectrum disorder patients showed abnormalities in deep cortical layer formation and resulted in overproduction of GABAergic interneurons.[130,131] Studying the effect of disease pathology at a single cell level is an attractive approach as it may allow identification of cellular processes that cause cell type or layer-specific vulnerability.[89]

2.4 Methods

2.4.1 Derivation of iPSCs

iPSC lines were derived from skin biopsy fibroblasts in the James Martin Stem Cell Facility, University of Oxford, and all cultured under standardised protocols to minimise any potential variation attributable to laboratory differences and/or handling. Fibroblasts and derived iPSC lines tested negative for mycoplasma using MycoAlert (Lonza). iPSC-NHDF-1 (44 year-old female, reprogrammed with Yamanaka retroviruses *SOX2*, *KLF4*, *OCT3/4*, *c-MYC* and *NANOG*) has been described previously.[132] iPSC-AH017-3 and iPS-AH017-7 (67 year-old female) are described for the first time here. They were derived using the SeVdp(KOSM)302L Sendai virus system, containing genes for *KLF4, OCT3/4, SOX2, c-MYC* expressed from a single transcript, packaged into a single Sendai virus, ensuring consistency of gene dosage ratio.[133,134] The system also contains a target for mir302; mir302 is expressed in pluripotent cells, but not in the originating fibroblasts, ensuring complete removal of exogenous genetic material within a few passages.

Sendai-transduced fibroblasts were transferred onto mitomycin C-inactivated mouse embryonic feeder cells (MEFs; outbred Swiss mice established and maintained at the Department of Pathology, Oxford [135,136]) on 0.1% gelatin coated plates on day 2 and cultured in iPS medium consisting of knock-out DMEM (Invitrogen), 10% knock-out-Serum Replacement (Invitrogen), 2 mM Glutamax-I (Gibco), 100 U/mL penicillin (Invitrogen), 100 µg/mL streptomycin (Invitrogen), 1% non-essential amino acids (Invitrogen), 0.055 mM β-mercaptoethanol (R&D), 10 ng/mL bFGF (R&D). Medium was replaced (50%) daily, substituting with MEF-conditioned medium from day 10 onwards. Colonies displaying iPSC morphology were picked on ~day 21-28 and manually dissected onto fresh feeders every 5–7 days until stabilised, then adapted to feeder-free conditions onto Matrigel-coated plates (BD Matrigel hESC-qualified Matrix) in mTeSR™1 (StemCell Technologies, 100% change daily). Bulk passaging was by 0.5mM EDTA to make large-scale, quality-controlled, cryopreserved stocks (genomic DNA and RNA extraction for results shown is passage 22 for iPS-AH017-3 and passage 17 for iPS-AH017-7).[137] The number of feeder-free passages was kept to a minimum to ensure that differentiation experiments were set up within a consistent and very narrow window of passage numbers. The human iPSC lines used in this thesis were derived from human skin biopsy fibroblasts, following signed informed consent, with approval from a research ethics committee: National Health Service, Health Research Authority, NRES Committee South Central – Berkshire, UK - REC 10/H0505/71.

2.4.2 Assessment of Genome Integrity, and Tracking

Genomic DNA was extracted using Qiagen Blood and Tissue Kit. Genome integrity was assessed by an Illumina Human CytoSNP-12v2.1 beadchip array (~300,000 markers), analysed using KaryoStudio software (Illumina) and SNP deviations in the iPSC lines compared to the original pool of fibroblasts. This also enabled confirmation of the identity of the iPSC to the original fibroblasts. SNP datasets have been deposited in GEO under the accession number GSE69287 (Superseries GSE69302).

2.4.3 Sendai clearance assay

Clearance of Sendai virus from the reprogrammed cells was confirmed by quantitative RT-PCR. RNA was extracted using an RNeasy kit (Qiagen) from iPSC, from fibroblasts (negative control) and from fibroblasts infected with the Sendai reprogramming virus 5 days previously (positive control). Reverse transcription was carried out using a RetroScript kit (Ambion; 2 ug template RNA, 20 ul reaction volume), then 2 ul of 1:10 cDNA product in a 25 ul qRT-PCR reaction (Applied Biosystems StepOne Plus Real Time PCR machine, StepOne software), using Applied Biosystems 2xSYBR green PCR mix+ROX, 60^{o}C anneal and Sendai-specific primers (5'-AGACCCTAAGAGGACGAAGACAGA-3' and 5'-ACTCCCATGGCGTAACTCCATAG-3'). Target gene transcript levels were compared to actin B control (actin B primers, Eurogentec), and subsequently to the positive control.[138]

2.4.4 PluriTest

RNA was extracted from iPSC using an RNeasy kit (Qiagen) for Illumina HT12v4 transcriptome array analysis. Image data files were uploaded to www.pluritest.org and scored for pluripotency as previously described.[139] Transcriptome datasets have been deposited in GEO under the accession number GSE69288 (Superseries GSE69302).

2.4.5 Flow cytometry

iPSC lines were stained for expression of the pluripotency marker using TRA-1-60 Alexa Fluor 488 conjugate (B119983, Biolegend; with IgM-488 isotype-matched Alexa Fluor 488 control). Briefly, cells were fixed with 4% paraformaldehyde in phosphate buffer saline (PBS) and permeabilised with methanol at -20 degrees Celcius. 0.5–1×10^{6} cells were washed and stained in flow cytometry buffer (PBS, human IgG (10 µg/mL Sigma), FCS (1% Hyclone) sodium azide (0.01%)), for 30 min. The cells were assayed using a FACSCalibur (Becton Dickinson) – data from 10,000 cells were collected and analysed using FlowJo software, gated on FSC-SSC dot plots, and plotted as histograms (y axis normalised to mode).

2.4.6 Cortical neuronal differentiation

iPSCs were differentiated into cortical neurons using the Livesey protocol with some modifications.[61] In brief, iPSCs were cultured in an adherent monolayer on matrigel-coated plates. Neural induction was achieved using dual SMAD inhibition (1μM dorsomorphin and 10μM SB431542) in neural maintenance media (DMEM/F-12, neurobasal, N-2, B-27, 5μg/ml insulin, 1mM L-glutamine, 100μM non-essential amino acids, 100μM 2-mercaptoethanol, 50 units/ml penicillin and 50mg/ml streptomycin). After the formation of a neuroepithelial sheet, cells were passaged as small clusters onto wells coated with poly-L-ornithine and laminin. The subsequent differentiating cells were allowed to form neural rosettes in the presence of b-FGF and passaged until confluent and frozen as cortical neural progenitor stocks. RT-qPCR and immunofluorescence microscopy confirmed the adoption of cortical identity. Cortical cultures were maintained in neural maintenance media with additional laminin feeding (100μg/ml) every 10 days to avoid cell detachment. Neuronal age was reckoned from the day of neural induction. Cells were treated with 4μM cytosine arabinoside for 72 hours prior to single cell analysis.

2.4.7 Selection of genes for RT-qPCR

Candidate genes related to cortical layer or region identity and neuronal function were curated from multiple sources.[54,58,140,141,61,142] *GAPDH* and *ACTB* were included as housekeeping genes. Several glial genes were included to identify glia. Negative control genes normally expressed solely in kidney and liver were also included. Primer sequences are shown in **Table 1**.

Table 1 Primer sequences

Target	Forward primer	Reverse primer
ACTB	CCAACCGCGAGAAGATGAC	TAGCACAGCCTGGATAGCAA
ADAMTS3	CCTGGGGCTAGACATGTGTTAA	GGCCTGTAGCCTGGTTCTTAA
AKR1C3	AGCTGGGTTCCGCCATATA	GGTCGATGAAAAGTGGACCAA
ANK2	TGGATGTGGCATCAGTCCTA	ACATGCAGGGGAGTAAAACC
ANXA1	AAGTGCGCCACAAGCAAA	TGCCTTATGGCGAGTTCCA
B3GALT2	TGATGGCGTGCCAAAAGTA	GTCACATTGTCTCTCTTGTAGTCA
BCL11B	CAACCCGCAGCACTTGTC	CCTCGTCTTCTTCGAGGATGG
NDNF	CCAGATGGAGCTGAAATTAGCA	CTCCAAAGGCGCATCACA
CACNA1E	GCCTTTGAAGCTCGTGTCA	AAGAGGTACCATGGCCTTCA
CARTPT	TGAGAAGAAGTATGGCCAAGTCC	ATCCTTGCCCCTTTCCTCAC

CBLN1	GCGCAAAGGGATCTACAGTTTTAA	CCAGCGAAGGCTGAAATCAC
CBLN2	TCAGCTTCCACGTGGTCAAA	AGGCCGAGATCACTGGGTA
CCK	AGAACGGATGGCGAGTCC	ACGATGGACATTCGTCCAGAA
CDH24	TTTTCCCCTTGGGCCCTAC	GTGAGCAGTCACCTGGATCA
CHRNA3	AGACCAACCTGTGGCTCAA	ATGAACTCTGCCCCACCATA
CHRNA7	CCCAGATGGCCAGATTTGGAA	AGTGTGGAATGTGGCGTCAA
CNR1	ACGTCTGAGGATGGGAAGGTA	AGGACCAGGGTCTTGGCTAA
CPNE7	AGCCGAACGAGTACCTGAA	CAAACCCCAAAGCGGAAAAC
CUX1	ACGGACCTTGAAAGGGCAAA	CCGATGAGAGCTGTTCCCTTAA
CYP26A1	AGACCCTTCGACTGAATCCC	CAGCCCTTGGGAATCTGGTA
DBH	GAGTGGGAGATCGTGAACCA	ACCGACACGACCTTCTTCAA
DLG4	AGCTGGAGCAGGAGTTCAC	ACACGCTTCACCTTGTGGTA
EBF3	CTCCTGGGCGCTTTGTCTA	CTTGGGATCACTTTCTGCAACC
EOMES	CTGTGGCAAAGCCGACAATA	CTCATCCAGTGGGAACCAGTA
FEZF2	CGCTCTGCAGGCACAAAATTA	GCCGCACTGGTTGCATTTA
FOXG1	GCCAGCAGCACTTTGAGTTA	TGAGTCAACACGGAGCTGTA
GABRA2	GCCAATCAATCGGAAAGGAGAC	TCAGGTGGAAATGAGCTGTCA
GABRA5	GCTCTTGGATGGCTACGACAA	TGACGTAGATGTCGGTCCTCA
GABRB1	TGGGGCTTCTCTCTTTCCC	ATGACATGTTGCTGGGTTCA
GAD1	ATCCTGGTTGACTGCAGAGAC	CCAGTGGAGAGCTGGTTGAA
GAD2	CTGCTCCAAAGTGGATGTCAAC	AAAGTGGGCCTTTCTCCATCA
GAPDH	GAACGGGAAGCTTGTCATCAA	ATCGCCCCACTTGATTTTGG
GFAP	GCCAGTTGCAGTCCTTGAC	GCGCATCTGCCTCTCCA
GPHN	GTGGTTGCAGTCATGTCAACA	CGATTGCTGTCTCGAATCTTCC
GRIA1	CGCTCCACGTGATTGAAATGAA	TGGCTGCAGGGACAAACTTA
GRIA2	TGGAATGGGATGGTTGGAGAA	CACCTCTTCTCTCACAAGGGTA
GRIA3	TGGGGCTTTTGGGTCATTC	CTCCTGCACTGTGTTTCTCA
GRIA4	GCAAATCCTGCTGCTCCA	ATTCCCTGTCAGCCCTTGAA
GRIN1	GGCAACACCAACATCTGGAA	CCATCCGCATACTTGGAAGAC
GRIN2A	GACCGGCCTCAGTGACAA	AGGCACTGTCCCAAATCGAA
GRIN2B	TGCACCCGAAACTGGTGATA	TGCAGGGACTTGTCTTTCCA
HNF1A	TGGTACGTCCGCAAGCA	ACCTGTGGGCTCTTCAATCA
HOXB2	GCACGGCTTACACCAACAC	GGCCGGCACAGGTACTTATTA
HOXB3	CCCTTCGTCATGAATGGGATC	GGCAGGCGACAAATCTCC
HTR2C	AGACTGAAGCAATCATGGTGAAC	TGCTACTGGGCTCACAGAAA
KCNC1	GCTCTTCGAGGACCCGTAC	AGACCAGGATGAAGAAGAGGGAA
KCNC3	TCTTCGAGGACCCCTACTC	GAAGGTGGTGATGGAGATGA
KCNK2	AATCAGTCACTGGGATTTGGGAA	TGCGTGGTGAGATGTTTCCA
MAP2	CAACGGAGAGCTGACCTCA	CTACAGCCTCAGCAGTGACTA
MBP	GAGCTCCAGACCATCCAAGAA	GGGAGGGTCTCTTCTGTGAC
MET	TCCCCAATGACCTGCTGAAA	CTTTTCCAAGGACGGTTGAAGAA
MFGE8	ACCCAGCAGCAATGACGATA	TGCGTCACCACACCTGTTA

MIOX	CAAAGACAAGGCCAGCTTCC	GCGTGTGCATGAGCTTGTA
MKI67	AGAGTAACGCGGAGTGTCA	CTTGACACACACATTGTCCTCA
MPPED1	GCCCTGCTGGAGAAATGGAA	GGGGACCCAGTCCAGGAA
NANOG	TGCCTTGCTTTGAAGCATCC	TTTCTTCAGGCCCACAAATCAC
NCAM1	CTCCCAGTCCATGTACCTTGAA	GGTTCCCCTCCCAAGTGTAC
NEFL	ACGTGAAGATGGCTTTGGATA	TAGGCAGATCGGCCAAAGAC
NES	GCTGCGGGCTACTGAAAA	CTGAGCGATCTGGCTCTGTA
NEUROD6	GGAGACGATGCGACACTCA	ACAGACTCATCAAACGGTAGTGTTA
NR2F1	CAAAGCCATCGTGCTGTTCAC	GCTCCTCACGTACTCCTCCA
NR2F2	CTCAAGGCCATAGTCCTGTTCA	CAAGCTTTCCACATGGGCTAC
NTNG2	GTGCCACTGCAGGTTCC	GTCACCACATTCAGGAAGTCA
OLIG2	CGGAGCGAGCTCCTCAAA	ATGGCCCCAGGGGAAGATA
OPRK1	CCTTGGAGGCACCAAAGTCA	GGTCCCACCAGGAGTAGTCA
PAX6	CCCCACATATGCAGACACACA	GAACTGACACACCAGGGGAAA
PDE1A	GCATACAGGGACAACAAACAAC	TCTCAAGGACAGAGCGATCA
PDYN	CAGGATGGTCCCAAACCTATCA	AGAAAGCTCTGGCATCTCTCC
POU3F2	CGGATCAAACTGGGATTTACCC	CGAGAACACGTTGCCATACA
POU5F1	GGTATTCAGCCAAACGACCAT	CCGCAGCTTACACATGTTCT
PPP1R1B	TCTCAAGTCGAAGAGACCCAAC	TGCAGGTGAGACTCAGCAA
PRSS12	GTTAGCATTGCCTGCTACCC	TTTCTCCATCCATCAGTCTGACA
RASGRF2	CAAGGACTAAAGGCAAGGATAGCA	TGTTCAGCATGGGGAGCAAA
RORB	GGCAGACCCACACCTATGAA	ATTGTTGCCACAGTGCTTCC
RSPO3	GGAATACATCGGCAGCCAAA	TCCTTGGCAGCCTTGACTAA
RXFP1	AGTCGAATTTCCCCACCAACA	AAACGGGTGAGGACGTTATTCA
RXRG	ATGCCTGCGAGCCATTGTA	TCGCAGAGTCTCCACCTCA
SATB2	TTTGCCAAAGTGGCTGCAAA	TTTCTGGGCTTGGGTTCTCC
SCN4B	GAAGTGGACAACACAGTGACA	CGAGGAGCTCACGAGACA
SIX3	GGCCTCACTCCCACACAA	ATGCCGCTCGGTCCAA
SLC17A6	TGGGGCTACATCATCACTCA	GAAGTATGGCAGCTCCGAAA
SLC17A7	GCCATCTCCTTCCTGGTCCTA	GGCTATGTCCAGGTGGTTCAC
SLC32A1	AGGCTGGAACGTGACCAAC	GAGAAACAACCCCAGGTAGCC
SOX2	CATGAAGGAGCACCCGGATTA	CGGGCAGCGTGTACTTATCC
SV2C	GGACAGAATTGGGCGCTTAA	GTGCCGAACCAAAGGAAGAA
SYN1	GCAAGGACGGAAGGGATCA	TGTCTTCATCCTGGTGGTCAC
SYNDIG1L	CGGGACCAGGAGGATGAC	AGCGTGAGGAAGTTGTCTTCA
SYT2	TGCATCCTGGAGGCTAAGAA	CTGCATCAGGTGGATCTTCAC
SYT6	TCTGACCTGTCTCGGGAAAC	AGAACATGATCTCTCCCAAGTCC
TBR1	ACGAACAACAAAGGAGCTTCA	TGGTACTTGTGCAAGGACTGTA
TCF7L2	GCACCGTAGGACAAATCCC	TCCTGTCGTGATTGGGTACA
TENM2	TGGCAGGCAAGGATAGCA	TTGGCCTCGGATGAGAGAAA
TPH1	GCGGACTTGGCTATGAACTA	GTTCCCCAGGTCTTAATCTCC
TTR	CTGCCTTGCTGGACTGGTA	GCCACATTGATGGCAGGAC

TUBB3	GAGCGGATCAGCGTCTACTA	GGTTCCAGGTCCACCAGAA
VAT1L	TCCAACATGGTAACTGGAGAGAC	GATGGGGTTCACCTTCTCCA

2.4.8 Bulk RNA extraction and reverse transcription

RNA was extracted from entire wells by using the RNeasy Micro Kit (Qiagen) and reverse transcribed using SuperScript III (Invitrogen).

2.4.9 Single cell RT-qPCR

Single cell suspensions were generated using accutase dissociation followed by single cell filtration of iPSC-derived cortical neurons aged > 81 days. The success of the suspension was manually confirmed on a haemocytometer. The single cell suspension was sorted into a 96-well PCR plate containing a lysis mix (VILO reaction mix (Invitrogen), SUPERase-In (Ambion) and 10% NP40 (Fisher Scientific)) using a cell sorter (Becton Dickinson FACSAria II SORP or FACSAria Fusion). Sorting gates were set to include only live (DAPI negative) single cells. Stream alignment and sort efficiency was checked using Accudrop beads (Becton Dickinson). The PCR plate was vortexed and centrifuged before performing denaturation at 65ºC for 90 seconds. RT mix (SuperScript enzyme mix (Invitrogen) and T4 Gene 32 Protein (New England Biolabs)) was added to this and then this mix was reverse transcribed (25ºC for 5 minutes, 50ºC for 30 minutes, 55ºC for 25 minutes, 60ºC for 5 minutes and 70ºC for 10 minutes). Preamplification was then performed with TaqMan PreAmp Master Mix (Invitrogen), 500nM Deltagene assay primer mix (Fluidigm) and 0.5M EDTA, pH 8.0 (Invitrogen) on a thermocycler (95ºC for 10 minutes, and 20 cycles of 96ºC for 5 seconds and 60ºC for 4 minutes). The resultant product was treated with exonuclease I (New England Biolabs) to remove unincorporated primers (37ºC for 30 minutes and 80ºC for 15 minutes). Individual primers (100µM) were combined with assay loading reagent (Fluidigm) and TE DNA suspension buffer (Invitrogen) before being pipetted onto a Biomark chip (Fluidigm). The cDNA from the 96-well PCR plate was diluted with DNA suspension buffer 1 in 5, combined with sample loading reagent (Fluidigm) and loaded onto the same Biomark chip. This was run on a Biomark analyser system (Fluidigm). The results were visualised using Real Time PCR Analysis software version 4.1.2 (Fluidigm). Melting curves were

individually visualised for each reaction to ensure only a single peak was selected for expression analysis and that temperature ranges were common to all experiments. Detection thresholds (Tm) were automatically generated using a baseline linear correction model and a quality threshold of 0.65.

2.4.10 Single cell RNA-seq

Cells were dissociated as for single cell RT-qPCR at day 72 of neuronal differentiation. 300,000 DAPI negative cells were sorted into 200μL of neural maintenance media. These were loaded onto a small C_1 chip according to the Smarter-seq protocol detailed by Fluidigm.[44,143] Cells were co-stained with Hoechst and propidium iodide. Capture chambers were imaged on an Opera Imaging System. Lysis, reverse transcription, amplification and library prep were carried out in accordance with the Fluidigm Smarter-seq protocol. cDNA from 16 harvest chambers containing cells which were positive for Hoechst and negative for propidium iodide were sequenced using a rapid sequencing run (100 base paired-end reads) on an Illumina HiSeq system to a mean depth of ~5 million reads per cell (archived as accession number GSE69790). Additional single cell RNA-seq data were downloaded from Pollen *et al.* (accession number SRP041736) and Darmanis *et al.* (accession number GSE67835).[44,46]

2.4.11 Immunofluorescence microscopy

Cells were fixed with 4% paraformaldehyde and permeabilised with 0.3% Triton-X100. These were then incubated with blocking buffer before applying the primary antibody. After washing with PBS, fluorochrome-conjugated secondary antibodies were applied. This was further washed and mounted using Dako fluorescence mounting medium on a glass slide for confocal microscopy. Antibodies were used as detailed in Shi *et al.*[61] The specificity of these antibodies was tested by staining undifferentiated iPSC and demonstrating no detectable signal with antibodies against NANOG as a positive control. Analysis of layer marker co-expression was performed by Cytospin (Thermo Scientific) preparation of dissociated iPSC-derived cortical neurons as described by Shi and colleagues followed by the counting of five fields of view stained

by DAPI and for various combinations of layer markers.[61] Results are presented as mean proportions ± standard deviation.

2.4.12 Electrophysiology

A coverslip containing iPSC derived cortical neurons was placed in a recording chamber mounted onto the stage of an upright microscope and the somata of cortical neurons were targeted for recording using IR-DIC optics. Cells were continuously superfused with oxygenated aCSF (95% O_2/5% CO_2) containing 130 mM NaCl, 25 mM $NaHCO_3$, 2.5 mM KCl, 1.25 mM NaH_2PO_4, 2 mM $CaCl_2$, 2 mM $MgCl_2$, and 10 mM glucose. Patch-clamp electrodes (4–7 MΩ) were filled with an intracellular solution containing 120 mM K-gluconate, 10 mM KCl, 10 mM HEPES, 4 mM MgATP, 0.3 mM NaGTP, 10 mM Na-phosphocreatine. Recordings were obtained using a Multiclamp 700B amplifier and digitized at 10–20 kHz using Digidata 1550 acquisition board.

2.4.13 Data analysis

Single cell RT-qPCR data was analysed in R. In accordance with guidelines on single cell RT-qPCR analysis, Ct values were analysed without adjusting for housekeeping gene expression levels.[128] Genes with detectable expression were defined as those with a Ct value < 30. A well was defined as containing a cell which had been successfully reverse transcribed if there were detectable levels of *GAPDH* (Ct < 30). I tested primer efficacy with single cell cDNA by conducting a dilution series and found good linear amplification for detectable genes (r^2 = 0.98 ± 0.03). Genes were excluded from downstream analysis if these showed poor inter-chip reliability (the same single cell cDNA analysed on different chips). There was excellent correlation of gene expression between technical replicates (median r^2 = 0.96 ± 0.05). Ct values were adjusted between batches by the difference in mean *GAPDH* and *ACTB* expression relative to the first replicate of AH017-3. Cells were classified as neurons if there was detectable expression of *MAP2*, *NCAM1* or *TUBB3*. Cortical layer identity was established through the expression of candidate layer markers: deep layer markers: *BCLB11* and *TBR1*; upper layer markers: *CUX1*, *POU3F2* and *SATB2*. Internal controls were also used to estimate the magnitude of technical variation between Biomark analysis runs. The significance of principal component analysis

clustering was calculated by applying a Wilcoxon test to the distance between individual centroid locations for each group of points to be analysed compared to the distance to the centroid for the pooled points of both groups. Co-expression was analysed by calculating correlation co-efficients on the original dataset for pairs of cortical layer markers and then randomly permuting these Ct values 10,000 times. An empirical two-tailed p-value was estimated from this distribution and corrected for multiple hypothesis testing using the Benjamini-Hochberg method.

RNA-seq reads were processed using Trimmomatic with the arguments "LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50".[144] Trimmed reads were aligned against the hg19 UCSC genome assembly using Tophat2 (version 2.0.13, arguments "-p 8 -G <genome bt2 index> -1 <forward reads> -2 <backward reads>").[145] Overall alignment rates for iPSC-derived cortical neurons were ~67% with ~4000 detectable genes per cell. Fragments were assigned to UCSC known genes with Rsubread (version 1.16.1, arguments "isGTFAnnotationFile = TRUE, GTF.featureType="exon", GTF.attrType="gene_name", useMetaFeatures=TRUE, allowMultiOverlap=FALSE, isPairedEnd=TRUE, requireBothEndsMapped=TRUE, checkFragLength=FALSE, nthreads=8, strandSpecific=0, minMQS=0, countMultiMappingReads=FALSE, countChimericFragments=FALSE).[146] Counts were adjusted for library sizes using DESeq (version 1.20.0, estimateSizeFactorsForMatrix).[147] Cells were removed during QC if the number of expressed genes was lower that the boxplot lower limit (*i.e.* an outlier), if the proportion of fragments aligning to mitochondrial genes was high (> 0.15 as suggested in a previous study [148]) or if the cell was an outlier on initial principal component analysis (single cells of the same type were analysed by PCA and were excluded if > 5 times the median absolute deviation from the centroid). A gene was considered detectable if at least one fragment with both paired reads uniquely mapped was assigned to it. Principal component analysis was illustrated for the top 300 genes explaining the first two principal components as well as for the whole transcriptome. Cells types of single cells from primary tissue datasets were provided by the respective study authors.[44,46] Differential gene analysis of single cell data was conducted using Monocle with an FDR threshold of 0.05.[43] I only considered genes differentially expressed if the gene was detectable in > 10% of cells at a mean expression level FPKM > 1 in at

least one type of cell where the direction of change in the proportion of cells expressing the gene and the magnitude of mean expression were the same. Gene ontology analysis was conducted in DAVID using all detectable genes as the background.[149]

2.5 Results

2.5.1 Single cell RT-qPCR neuronal identity

I generated cortical neurons using a well-established protocol with small molecule dual SMAD inhibition for neural induction followed by plating of neuroepithelial cells for final differentiation.[61]

Over the course of neuronal differentiation, cultures showed the expected decrease in expression of pluripotency genes and increased expression of neuronal genes (**Figure 4a**). Staining of iPSC-derived cortical neurons showed the presence of synaptic markers, the deep layer marker TBR1 and the upper layer marker CUX1 (**Figure 4b-d**). Neurons demonstrated repetitive firing in response to depolarisation and spontaneous synaptic activity (**Figure 4e-f**), indicating functional maturation.

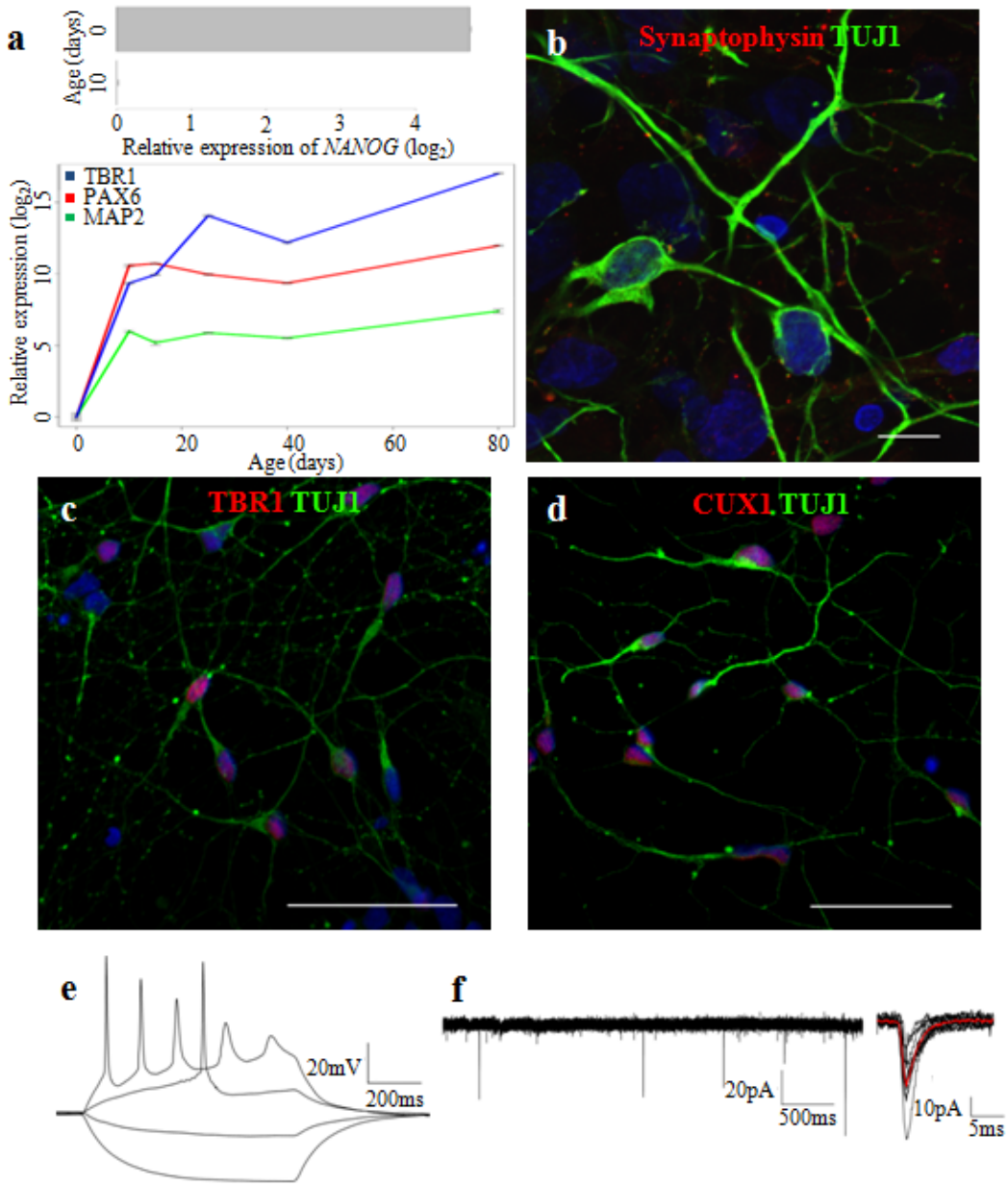

Figure 4 Validation of cortical neuronal phenotype. (**a**) RT-qPCR showing a reduction in *NANOG* expression (top) and increased expression of cortical identity markers (bottom) following neural induction in AH017-7. Error bars show the standard deviation from technical triplicates. Immunofluorescence microscopy for: (**b**) TUJ1 (green) and Synaptophysin (red), scale bar = 10µm; (**c**) TUJ1 (green) and TBR1 (red), scale bar = 50µm; (**d**) TUJ1 (green) and CUX1 (red), scale bar = 50µm; All images are also co-stained with DAPI (blue). (**e**) Repetitive firing evoked by current clamp protocol. (**f**) Spontaneous electrical activity.

The use of iPSC neurons for disease modelling and drug screening requires the production of cells emulating the cell types involved in the disease processes. I characterised the identity of cells produced by this protocol by performing single cell analysis on 3 iPSC lines (iPS-AH017-3,

iPS-AH017-7 and iPS-NHDF-1) following cortical differentiation for at least 81 days. Cells were dissociated into a single cell suspension and single cells (178 cells from AH017-3, 153 from AH017-7 and 75 from NHDF1) were sorted into PCR plates using FACS. Multiplex single cell RT-qPCR was carried out on 96 genes previously implicated as important for neuronal function or cortical layer identity, housekeeping genes and negative control genes.[61,54,58,140–142]

406 of 478 (85.3%) wells into which a single cell had been sorted were positive for expression of the housekeeper gene *GAPDH*. Of these, 380 (93.6%) expressed markers of neuronal identity (*MAP2*, *NCAM1* or *TUBB3*). The majority of these also expressed genes related to glutamatergic synapse function (e.g. *GRIA1* was detected in 63.9% of neurons, *DLG4* in 70.3% and *SYN1* in 67.6%). A small proportion (23.9%) of neurons expressed *GAD1*, suggesting the presence of GABAergic inhibitory neurons. This marker was inversely correlated with *SLC17A7*, a marker of excitatory identity ($r = -0.10$, $p = 0.03$). Almost no cells had any detectable expression of glia-specific genes (*GFAP* 1.0%, *OLIG2* 0.2%). A heatmap of all single cell RT-qPCR data is shown in **Figure 5a**. Co-expression analysis showed clusters associated with synaptic function and neuronal identity (**Figure 5b**). These included clusters containing multiple neuronal markers (e.g. *MAP2* and *NCAM1*), synaptic genes (e.g. *DLG4* and *ANK2*) and NMDA receptor genes (*GRIN2B*), with another cluster containing occipital lobe neocortical identity genes (*FEZF2*, *NEUROD6* and *ADAMTS3*) and AMPA receptor genes (*GRIA1*, *GRIA2* and *GRIA3*). I assessed whether there were any differences in the cellular composition of cultures aged between 81 days ($n = 109$) and 180 days ($n = 69$) in AH017-3 iPSC-derived cortical neurons and found no significant difference ($p = 0.44$).

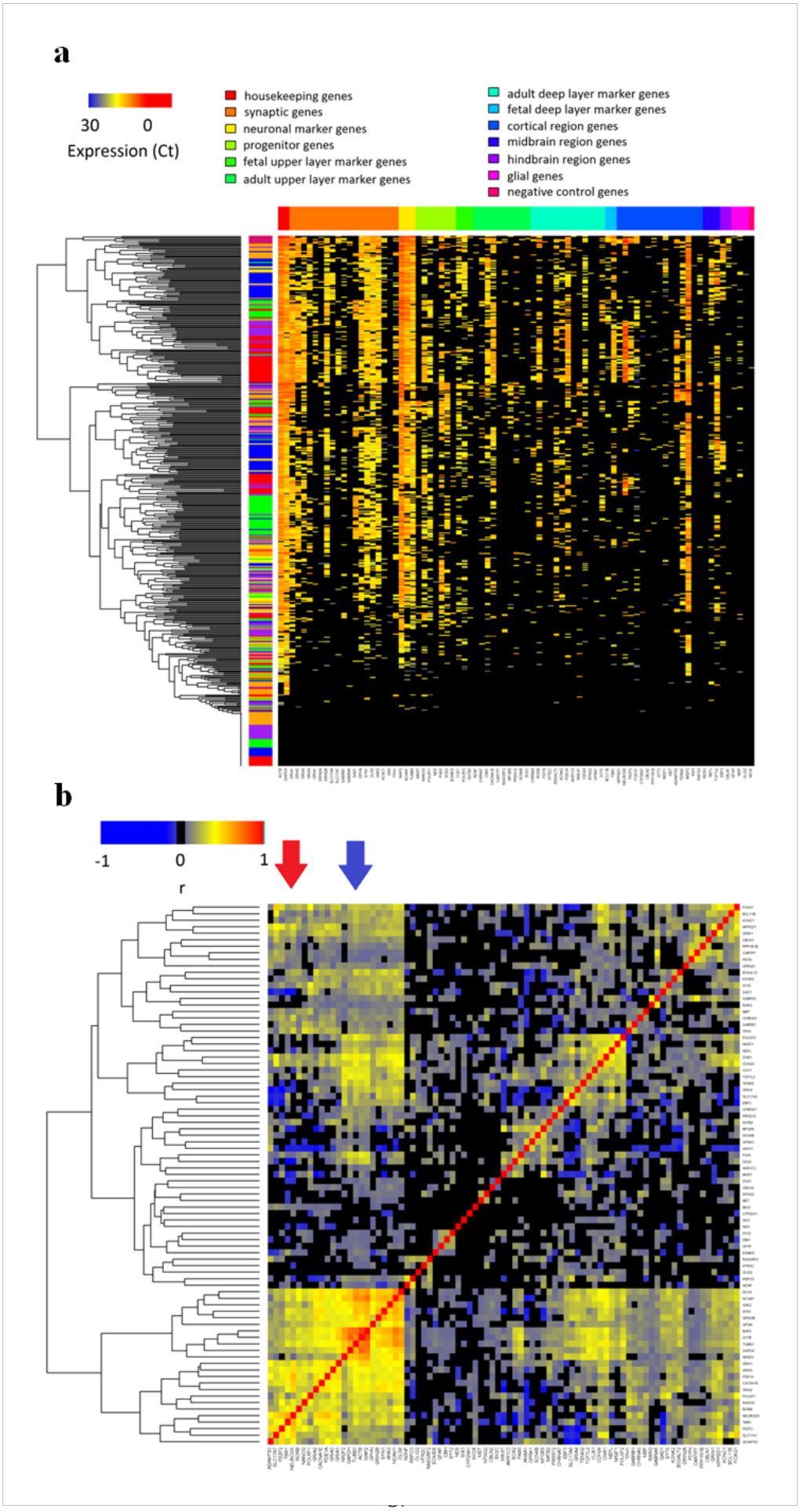
a
30
0
Expression (Ct)
housekeeping genes
synaptic genes
neuronal marker genes
progenitor genes
fetal upper layer marker genes
adult upper layer marker genes
adult deep layer marker genes
fetal deep layer marker genes
cortical region genes
midbrain region genes
hindbrain region genes
glial genes
negative control genes
b
-1
0
1
r

Figure 5 Single cell RT-qPCR data. (**a**) Heatmap of single cell gene expression data. Cells are shown in rows and genes in columns. The vertical colour key beside the dendrogram indicates the experiment (red = AH017-3 (repeat 1), blue = AH017-7, green = NHDF1, yellow = AH017-3 (repeat 2), purple = AH017-3 (repeat 3) and orange = AH017-7 (repeat 2)). The horizontal colour key indicates the functional category of genes. Cells are clustered by Euclidean distance. (**b**) Co-expression heatmap of single cell gene expression data. The shading represents the magnitude of the Pearson correlation co-efficient. Clusters of genes relevant to neuronal function (blue) and neocortical identity (red) are indicated by arrows. Genes not detected in bulk RT-qPCR samples were removed prior to analysis.

There was good correlation between bulk and pooled single cell gene expression (AH017-3: $r = 0.74$, $p < 0.0001$; AH017-7: $r = 0.78$, $p < 0.0001$; NHDF1: $r = 0.72$, $p < 0.0001$), suggesting that the process of dissociation did not greatly bias the resultant cell population analysed.

2.5.2 Comparison with single cell RNA-seq

In order to assess how well the transcriptomes of the iPSC-derived cortical neurons reflected those of single neurons from human fetal or adult neurons, single cell RNA-seq data was generated on sixteen 72-day old iPSC-derived cortical neurons. Single cell gene expression estimates from single cell whole transcriptomics correlated well with my previous single cell multiplex RT-qPCR data ($r = 0.84$, $p < 10^{-16}$). Single iPSC-derived neurons clustered closely with fetal but not adult cortical neurons and expressed high levels of genes specific for post-mitotic fetal neurons (**Figure 6a-c**).[44,46]

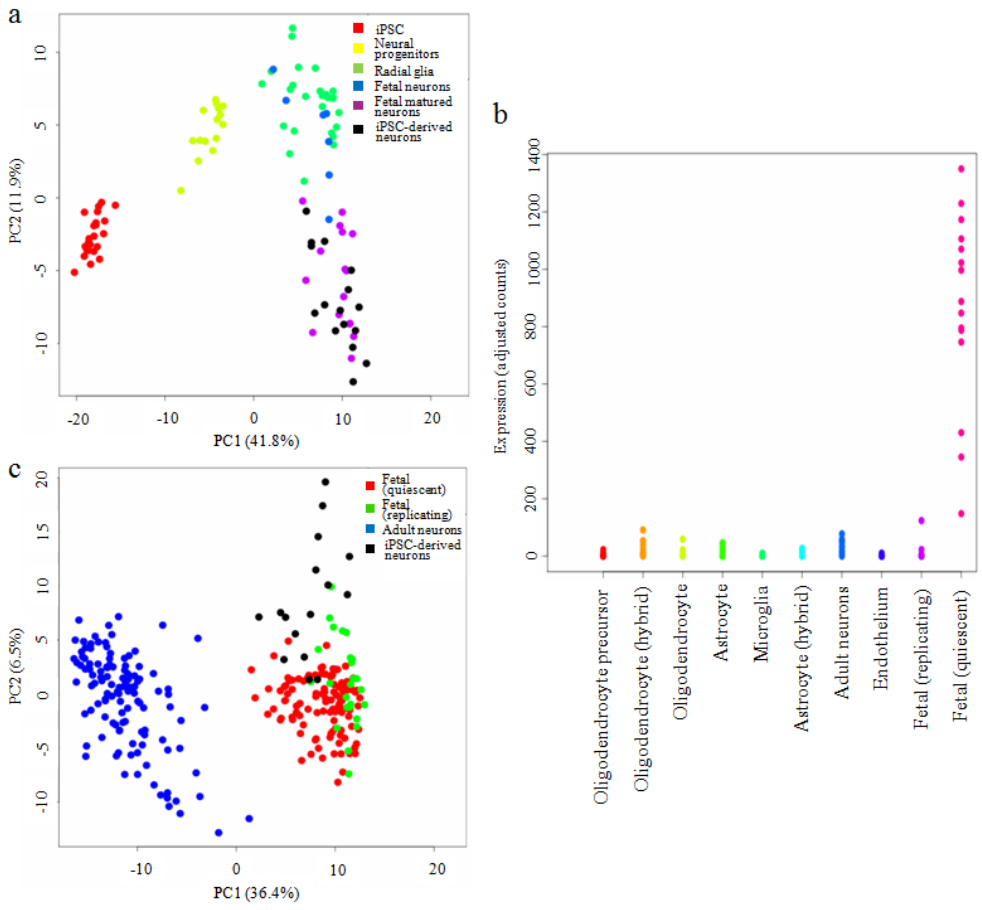

Figure 6 Single cell RNA-seq in iPSC-derived cortical neurons. (**a**) Principal component analysis of single cell RNA-seq on iPSCs (red), neural progenitors (yellow), fetal radial glia (green), fetal newborn cortical neurons (21-weeks post-conception; blue), fetal matured cortical neurons (21-weeks post-conception plus 3 weeks culture; purple) and iPSC-derived cortical neurons (black).[44] (**b**) Mean expression levels of genes demonstrated by Darmanis *et al.* to distinguish between different types of adult and fetal cells within the central nervous system.[46] (**c**) Principal component analysis of single cell RNA-seq on neurons from fetal quiescent cells (16-18-weeks post-conception; red), fetal replicating cells (16-18-weeks post-conception; green), adult temporal cortex (21 - 63 years old; blue) and iPSC-derived cortical neurons (black).

As clearly shown by clustering analysis, iPSC-derived cortical neurons show a striking resemblance to primary fetal cortical neurons at the single cell level. iPSC-derived cortical neurons clustered more closely with primary fetal neurons, which were cultured for three weeks prior to single cell transcriptomic analysis, than with freshly harvested fetal neurons (**Figure 6a**). This suggests that there is a considerable effect of *in vitro* maintenance on cellular transcriptomes. However, future protocols may refine this similarity and so I analysed the single cell RNA-seq data in order to identify functional pathways that showed significant differences between primary

neurons and iPSC-derived cortical neurons. Genes identified as significantly more highly or lowly expressed in primary brain cells than in iPSC-derived cortical neurons by Monocle were submitted to DAVID for gene ontology analysis.[149] Pathways more active in iPSC-derived neurons than fetal neurons included glycolysis and amino acid catabolism, whereas those more active in primary fetal cortical neurons than in iPSC-derived neurons included ribosomal and neuronal morphogenic pathways. As expected, there was considerably less expression of synaptic or ion channel-related pathways in iPSC-derived neurons than in adult cortical neurons, indicating that these cells remain electrophysiologically immature compared to mature brain tissue.

2.5.3 Cortical layer identity

Principal component analysis (PCA) showed that most experiments clustered closely together (**Figure 7a**), suggesting that this protocol is robust and reproducible. Prior work has established that there is continuous neurogenesis in these in vitro cultures with deep layer neurons emerging early and upper layer neurons appearing late.[61,91] I therefore examined whether cells clustered by cortical layer identity. Cortical layer markers implicated in layer specification as a result of bulk transcriptomic approaches in human brain were expressed only patchily in iPSC-derived cortical neurons and so were not considered as likely markers of layer identity in these cultures.[140] There was no obvious clustering by cortical layer identity as assessed by canonical layer markers (**Figure 7b**; deep layer markers: *BCL11B* (also known as CTIP2) and *TBR1*; upper layer markers: *CUX1, POU3F2* (also known as BRN2) and *SATB2*). There is a suggestion of separation between cells of deep and upper layer identity on PCA but this fell below statistical significance after correction for multiple hypothesis testing ($p = 0.07$).

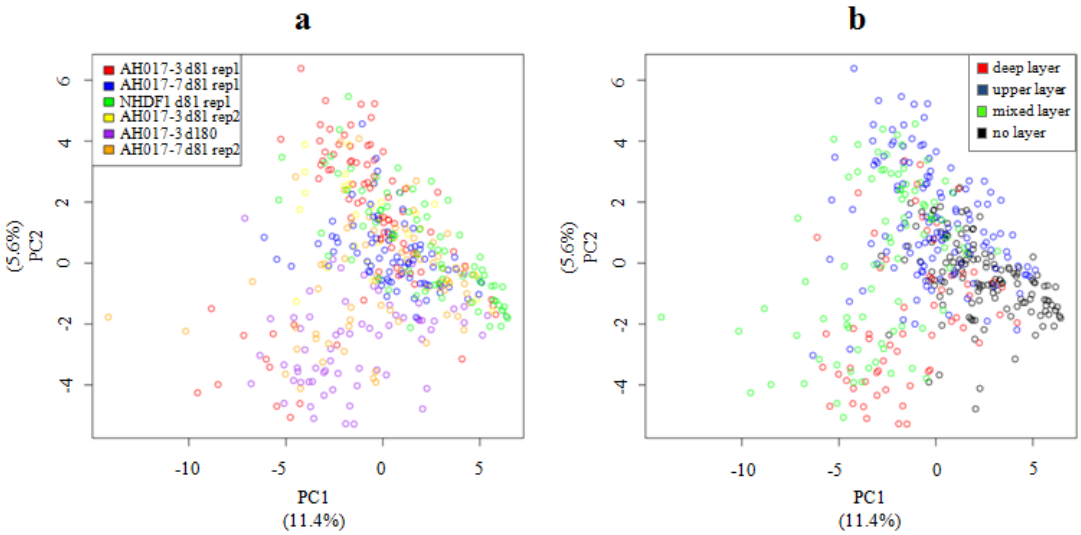

Figure 7 Principal component analysis of single cell RT-qPCR data. (**a**) Clustering of cells by experiment on single cell RT-qPCR data. The percentage of total variance accounted for by each principal component is shown beneath each axis. (**b**) Clustering of cells by cortical layer identity. The percentage of total variance accounted for by each principal component is shown beneath each axis.

I sought to investigate further this apparent lack of clustering. 184 of 380 (48.4%) iPSC-derived cortical neurons could be unambiguously assigned to either deep or upper layer identity on the basis of canonical layer markers with my single cell RT-qPCR dataset. 111 (29.2%) neurons had no detectable expression of any of the layer marker genes. 85 (22.4%) neurons expressed mixed cortical layer markers (**Figure 8a**). Potentially any clustering could have been obscured by cells with no or mixed layer identity. I considered whether cells of mixed layer identity could be those with low expression of one set of canonical layer markers in combination with otherwise well-established layer identity. However, the level at which layer markers were expressed did not appear to differ between neurons of single or mixed layer identity (**Figure 8b**; $p > 0.05$ for all; correlation between maximum deep and upper layer gene expression in cells of mixed layer identity: $r^2 = 0.001$, $p = 0.76$). Almost all combinations of canonical layer markers were expressed in cells (**Figure 8c)**.

Immunofluorescence microscopy demonstrated that this layer marker co-expression translated into protein co-expression (**Figure 8d-f**). I quantified the degree of layer co-expression for TBR1 and upper layer markers by immunofluorescence co-localisation using well-established antibodies

and found 26.5% ± 3.3 TBR1/BRN2 co-expression, 15.4% ± 4.2 TBR1/CUX1 co-expression, and 7.8% ± 2.8 TBR1/SATB2 co-expression.[61]

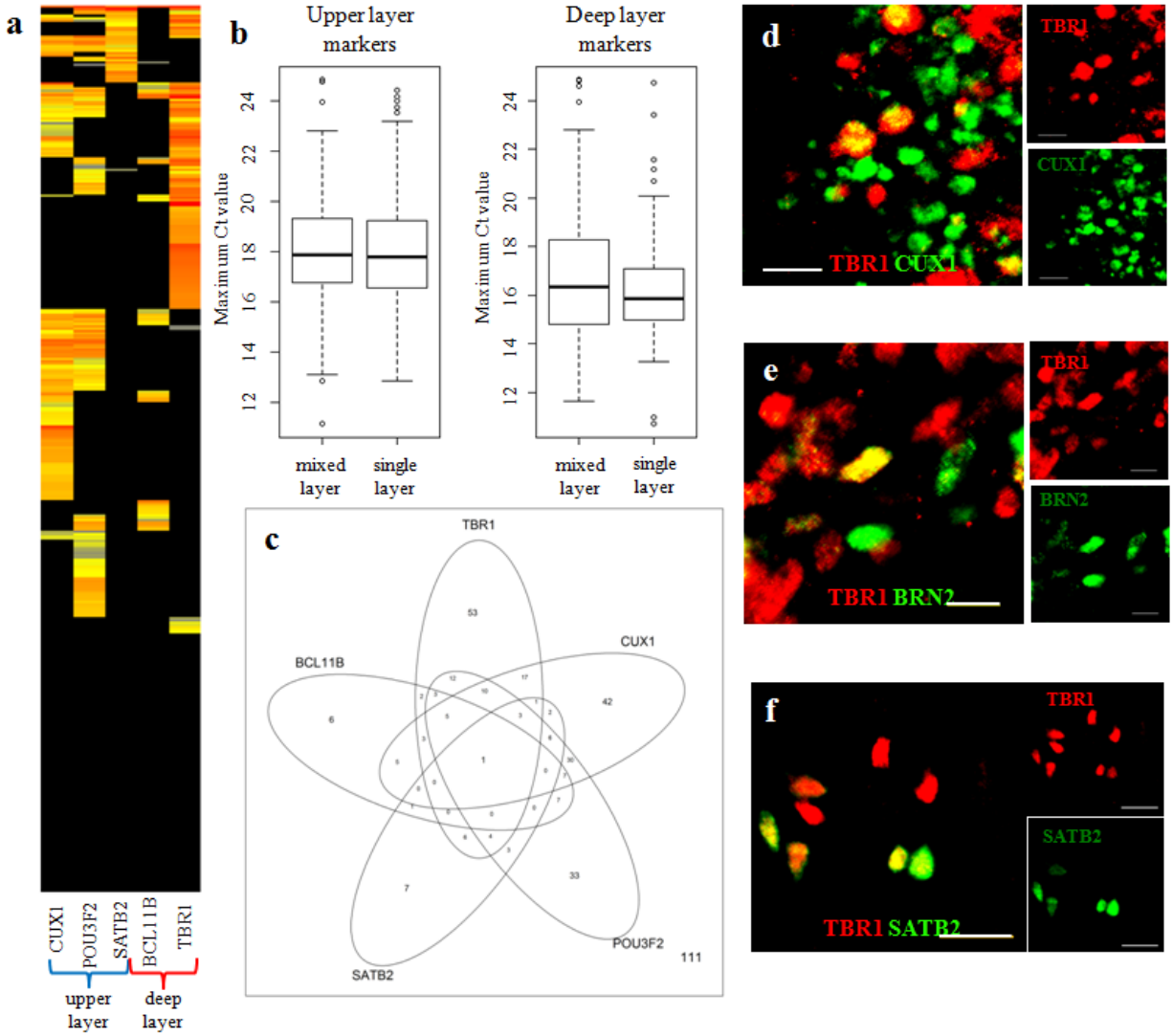

Figure 8 Cells with mixed cortical layer identity. (**a**) Heatmap of single cell gene expression data in neurons (n = 380) for deep and upper layer markers. Neurons (rows) are clustered by Euclidean distance. (**b**) Boxplots showing the magnitude of cortical layer marker expression by RT-qPCR in cells expressing both deep and upper layer markers or those with exclusively deep or upper layer identity. (**c**) The number of cells expressing each cortical layer marker by RT-qPCR. (**d-f**) Immunofluorescent images of cells co-expressing the deep layer marker TBR1 (red) and the upper layer markers CUX1 (**d**), BRN2 (**e**) or SATB2 (**f**) (green). Immunofluorescent images were from cortical neurons aged > 81 days old derived from AH017-3 (**c-d**) or NHDF1 (**e**) iPSCs. Scale bars represent 20μm.

Interestingly, although some combinations of markers showed the patterns of co-expression in my multiplex RT-qPCR dataset that would be expected from laminar identity (*CUX1 vs. POU3F2*: r

= 0.27, q < 0.0002; *POU3F2 vs. SATB2*: 0.14, q = 0.03), the canonical deep layer marker *BCL11B* was correlated with the upper layer markers *CUX1* (r = 0.14, q = 0.02) and *POU3F2* (r = 0.23, q = 0.001).

I then re-examined RNAseq data from primary human cells and made the unexpected observation that mixed layer identity is apparent in a proportion of single cell transcriptomes from fetal and adult brain (quiescent fetal cells: 66.4%; replicating fetal cells: 32.0%; adult neurons: 33.3%). There was no difference between the level of expression of canonical layer markers between cells of single layer and mixed layer identity ($p > 0.05$).

2.6 Discussion

I have confirmed that iPSC-derived neurons derived with a widely used protocol will generate neocortical cells principally expressing markers of neuronal excitatory glutamatergic phenotype at a single cell level. These cells accurately recapitulated the single cell transcriptomic signature of human fetal cortical neurons, and expressed many genes associated with neuronal and synaptic function. Widely used cortical layer markers were able to assign the majority of cells unambiguously to deep or upper layer identity but a considerable minority of cells co-expressed deep and upper layer markers.

Previous studies have compared bulk transcriptomic profiles of iPSC-derived cortical neurons to primary brain tissue.[65,150] When iPSC-derived cortical neurons were compared to donor-matched post-mortem brain, the similarity between iPSC-derived and primary cells increased with extended time in culture.[150] Another group showed that iPSC-derived neurons closely resembled fetal brain and that the equivalent "age" of the cultured cells was increased by changing from a two dimensional to three dimensional *in vitro* system.[65] However, any comparison between cultured cells and primary tissue will be hampered by alterations in the ratios between different cell types. For example, the size of the glial population in fetal brain will change markedly overtime and so the eventual production of glia in iPSC-derived cultures could well increase the apparent similarity of cultured cells to more mature primary tissue without any change in the underlying neuronal transcriptome.[91] Single cell transcriptomics offers a method of assessing the

similarity of iPSC-derived cells to primary tissue with the ability to remove the confounding introduced by cellular heterogeneity. This chapter shows that even at the level of individual cells, there is marked similarity to human fetal neurons.[44,46]

There are clearly still transcriptomic differences between iPSC-derived cortical neurons and primary fetal cells, as highlighted by my differential gene analysis. iPSC-derived cortical neurons show higher expression of genes related to glycolysis and catabolism of amino acids than is the case for fetal neurons *in vivo*, which may be a result of the abnormality of *in vitro* conditions. Similarly, the lack of structured cell migratory pathways *in vitro* was clear from the transcriptomic signature of iPSC-derived cells, with markedly lower expression of *DCX*, *ROBO2* and other genes involved in neuronal motility. Relatively simple 2D adherent monolayer neuronal cultures are more amenable for high throughput screens but potentially more organised; 3D models may improve matters. Single cell RNA-seq analysis of stem cell-derived cortical 3D organoid cultures suggested that differences between primary and cultured cortical neurons likely reflected elements within the culture media.[151] However, without a morphologically defined ventriculo-pial gradient *in vitro* there may still be deficits in differentiation and migration.[65] The fact that iPSC-derived cortical neurons cluster distinctly from adult cortical neurons may result in issues when attempting to model neurodegenerative disorders using these systems. This means that care should be taken when making molecular inferences regarding adult-onset neurodegenerative diseases from studies in iPSC-derived systems.[152] This is supported by the lack of detectable expression of many adult markers of cortical neurons and suggests that iPSC-derived cortical neurons are relatively immature compared to primary brain neurons. Regional specification of the iPSC-derived cortical neurons is also difficult to assess, as some of the key markers of neocortical regions are transcription factors that are expressed at relatively low level and may therefore be missed by a single cell transcriptomic approach. My gene ontology enrichment analysis suggested that the expression of some key genes involved in neuronal development were more lowly expressed in iPSC-derived cortical neurons than in primary adult or fetal neurons (*e.g. CDK5R1, LPPR4, PTPRZ1, BCL11B, MAP1B, ROBO2, NRXN1, DCX, GAS7, RPS27A*). Attention to these pathways may help guide improvements to existing

differentiation protocols. Despite these limitations, stem cell-derived models of corticogenesis are still likely to represent the best *in vitro* human model amenable to molecular manipulation of cells for the foreseeable future.[90]

Many of the canonical layer markers used to infer cortical layer identity in iPSC-derived cortical neurons have been informed primarily by rodent models of corticogenesis.[54] It is perhaps therefore surprising that they were able to assign the majority of single cell transcriptomes generated from iPSC-derived neurons to either deep or upper layer identity. However, there was clearly a subpopulation of neurons that had mixed layer identity.

Unexpectedly this appeared also to be the case when I examined layer identity in available single cell RNA-seq data from 21-weeks post-conception fetal and adult cortical neurons.[44,46] This may represent neurons with truly mixed deep and upper layer identity, or alternatively, the cortical layer markers widely used to assess layer identity do not function well in a minority of cells. For the primary neuron data, it is also possible that there was spill-over between C1 capture sites in the primary brain single cell RNA-seq experiments but I feel this is unlikely to be the case because two empty C1 capture sites had no detectable expression of any of the presumptive layer markers. This would also not explain the co-occurance of deep and upper layer markers in my multiplex RT-qPCR dataset, which was obtained by FACS.

Longitudinal bulk transcriptomic data obtained from cortical neurons-derived from human embryonic stem cells has shown that genes regarded as upper layer markers actually appear to peak early in differentiation.[64] Whether or not this gene expression signature is driven by neurons is impossible to ascertain from bulk transcriptomic approaches. My single cell data and analysis of published data suggests that some single cortical neurons can co-express multiple layer markers. When I extended the duration that the cells remained in culture for 180 days (~25 weeks) I still detected a high proportion of mixed layer identity, suggesting that this is not simply a transitory state in neuronal differentiation, at least *in vitro*. In fact, I found that the proportion of different subpopulations of cells detected did not vary significantly between 81 days and 180 days. I feel that the few non-neuronal cells detected in my cultures are likely to be neural progenitors.

Since RT-qPCR data from single cells was able to recapitulate gene expression in bulk samples lysed *in situ*, this suggests that minimal biases are introduced by the process of dissociating cells prior to downstream analysis. However, it must be noted that mRNA species show varying degrees of subcellular compartmentalisation in neurons and so undoubtedly the shearing off of axons and dendrites inevitably means that the transcriptome analysed is not identical to that present *in situ*.[153] Ultra-high resolution *in situ* RNA-seq methods will obviate this in the future.[70,71]

Many fetal layer markers show reciprocal regulation and in rodent models of corticogenesis mixed layer identity may indicate aberrant neural development.[154,155] Later stages of fetal corticogenesis demonstrate defined distributions of layer specific markers as detected by immunohistochemistry.[58] Whether this co-expression is detectable at the protein level in human cortex is unclear and should be examined in future studies. However, I observed the translation of layer marker co-expression into protein in iPSC-derived cortical neurons. This may represent neurons with truly mixed deep and upper layer identity, or alternatively, the cortical layer markers widely used in iPSC-derived cortical neuron cultures to assess layer identity do not function well in a minority of cells.

Mixed cortical layer identity is a potential limitation to consider when modelling neurological diseases, particularly those in which layer identity is important, such as Alzheimer's disease and many neurodevelopmental conditions, since, unlike *in vivo*, spatial distribution cannot be used as a proxy for cortical layer identity.[88,89,154,156] My findings suggest that layer identity may be better established by combining single cell transcriptomics with functional read-outs such as electrophysiology than by relying on the expression of single transcription factors.

2.7 Conclusion

iPSC-derived cortical neurons appear to express a multitude of neuronal marker genes. These neurons recapitulate accurately the transcriptome of human fetal corticogenesis but are distinct from adult neurons. Canonical layer markers traditionally used to assign cells to cortical layers are capable of assigning the majority of cells unambiguously to either deep or upper layer

identity. However, a subpopulation of cells shows mixed layer identity both at the level of RNA and protein expression. A degree of mixed layer identity also appears to be present in primary cortical neurons. Overall, these findings suggest that protocols to derive cortical neurons from iPSCs are a good model of fetal corticogenesis and produce cells that accurately mimic *in vivo* fetal cortical neurons.

2.8 Acknowledgements

Satyan Chintawar provided assistance with cell culture and immunofluorescence microscopy. Tatjana Lalic performed electrophysiology. Emma Whitely and Sarah Newey supplied NHDF1 iPSC-derived cortical neurons. Jane Vowles and Sally Cowley derived AH017-3, AH017-7 and NHDF1 iPSC from their respective fibroblasts and performed initial quality control measures on these lines. Paul Sopp provided FACS support. Karene Argoud performed C1-based cell capture for single cell RNA-seq and library preparation. Much of the data presented in this chapter has been published as Handel, A. E. *et al.* Assessing similarity to primary tissue and cortical layer identity in induced pluripotent stem cell-derived cortical neurons through single-cell transcriptomics. *Hum. Mol. Genet*. **25,** 989-1000 (2016).

Chapter 3 Stem cell-derived models of neurological disease

3.1 Abstract

Alzheimer's disease is a leading cause of excess morbidity and death. Around 5% of patients have the familial form of Alzheimer's disease and approximately half of these are caused by mutations in presenilin-1 (*PSEN1*). Investigating the cause of Alzheimer's disease is challenging due both to difficulties in accessing primary brain tissue and separating out the effects of genetic and environmental risk factors. Induced pluripotent stem cell (iPSC) models offer potential solutions to both issues; there are protocols capable of deriving cortical neurons from iPSCs and *in vitro* culture conditions enable tight control over potential environmental factors. Additionally, genome engineering techniques promise to help dissect out the contribution of single mutations to disease pathogenesis in the presence of an otherwise unchanged genetic background. In this chapter I aimed to investigate whether iPSC-derived cortical neurons differentiated from wild-type or *PSEN1* mutant lines could recapitulate transcriptomic aspects of Alzheimer's disease, whether any differences found were robust between isogenic and case-control analyses, and the extent to which the use of genome engineering techniques could reduce batch-to-batch variability between iPSC-derived cortical neurons. I differentiated iPSCs from a *PSEN1* isogenic series (wild-type, heterozygote and homozygote), a healthy control and an Alzheimer's disease patient with a *PSEN1* mutation into cortical neurons. These were then analysed using bulk RNA-seq. All cultures showed an upregulation of key genes associated with the production of cortical neurons, suggesting that the corticogenesis protocol was working well. Differential gene expression analysis using an FDR threshold of < 0.05 identified 129 differentially expressed genes in the *PSEN1* isogenic system and 2,597 in the case-control comparison. Several of these were known Alzheimer's disease-related genes, such as *APOE* and *CLU*. However, there was low correspondence between the two models, with only 28 differentially expressed genes showing a consistent direction of effect across all *PSEN1* genotypes relative to wild-type. Unexpectedly, there was no significant reduction in transcriptomic variability between isogenic *PSEN1* samples

compared with case-control samples. These results suggest that there are key differences between isogenic and case-control approaches to investing disease pathogenesis. Further research will be required in order to identify potential reasons for this and to develop methods that minimise batch-to-batch variability.

3.2 Hypotheses

3.2.1 iPSC-derived cortical neurons recapitulate transcriptomic aspects of Alzheimer's disease.

3.2.2 Disease-relevant findings are robust between case-control and isogenic iPSC-derived cortical neuron comparisons.

3.2.3 The use of isogenic genome-editing techniques reduces batch-to-batch variation, which enables the attribution of biological effects to a specific mutation.

3.3 Introduction

I demonstrated in **Chapter 2** that iPSC-derived cortical neurons are highly similar to primary fetal cortical neurons. This makes iPSC corticogenesis an attractive human system in which to model neurological disease.[90] However, in order for this model to provide useful biological insights, it is crucial that results should be robust to potential confounders, such as batch-to-batch variability. In addition, to make specific claims regarding the functional effect of particular disease-associated mutations, any observations must be attributable to the genetic mutation of interest rather than simply due to background genetic differences between stem cell lines.

Alzheimer's disease is an extremely common neurodegenerative dementia that is responsible globally for a high rate of morbidity.[157] Mutations in the presenilin-1 gene (*PSEN1*) account for the largest proportion of familial Alzheimer's disease.[158] *PSEN1* encodes a γ-secretase involved in the processing of multiple transmembrane proteins, including amyloid precursor protein (APP). Aβ fragments generated by APP processing are thought to contribute to Alzheimer's disease pathophysiology either directly or by forming insoluble extracellular plaques.[159]

The precise importance of APP processing in Alzheimer's disease is debated, since imaging measures of amyloidosis are not highly predictive of the rate of cognitive decline in carriers of familial Alzheimer's disease mutations.[160] Clinical trials of drugs designed to interfere with Aβ generation have proved disappointingly ineffective in patients with Alzheimer's disease.[161,162] This is perhaps unsurprising given that the presence of post-mortem Aβ plaques is not highly predictive of ante-mortem cognitive impairment.[163] This is in stark contrast to the other pathological hallmark of Alzheimer's disease, neurofibrillary tangles (NFTs) of phosphorylated MAPT (tau), which are much more highly correlated with the severity of disease both *post mortem* and in cerebrospinal fluid *in vivo*.[164] However, Aβ plaques are almost always a prerequisite for the development of NFTs, suggesting that the deposition of Aβ is a necessary but not sufficient process in the pathophysiology of most cases of Alzheimer's disease.

There have been multiple attempts to model Alzheimer's disease using iPSC-derived neurons, mainly through the use of highly penetrant, familial Alzheimer's disease mutations or trisomy 21 (chromosome 21 contains the *APP* gene).[92,93,165–168,90,169,85,152,170] In many cases, the authors identified defects in Aβ ratios ($A\beta_{1-42}$:$A\beta_{1-40}$ was increased), increased levels of phosphorylated tau or a redistribution of tau within cells. Links have been identified between increasing *APP* genetic load and downstream phosphorylated tau levels, which do not appear to be seen with *PSEN1* mutations.[85] This dichotomy may in part be due to subsequent enzymatic breakdown of APP *via* an η-secretase dependent pathway.[171]

Two studies did examine cortical neurons derived from the iPSCs of patients with sporadic Alzheimer's.[93,166] In these studies, only 50% of sporadic patient lines demonstrated a disease-relevant phenotype. Few studies have demonstrated convincing, unprovoked neurodegenerative changes in iPSC-derived neurons, although potentially methods of artificially aging iPSC-derived cells or deriving neurons directly from donor fibroblasts may be able to address this.[172,173] In addition, there are no reliable transcriptomic signatures of disease detectable in iPSC-derived cortical neurons.[92,93,166]

One reason why a transcriptomic signature of Alzheimer's disease might have proven elusive is that most studies used a limited number of replicates and different differentiation protocols. This

means that any disease-relevant signal could potentially be obscured by batch-to-batch or inter-laboratory variation. No studies have directly quantified the extent of variability between differentiations and genotypes in iPSC-derived cortical neurons.

A key limitation of the studies discussed above is that the genetic background between individual iPSC lines differs by more than just the causative mutation. Transcriptomic diversity amongst iPSC lines is driven, at least in part, by inter-individual genetic differences so background variants could well play a part in confounding a disease signature.[94] In order to obviate this, I used an isogenic series of TALEN-engineered iPSC lines with either one or two copies of the *PSEN1* locus modified in order to generate a splice variant (Δe9) known to be deficient in gamma-secretase activity.[95,174,175] These lines have already been biochemically phenotyped and demonstrate robust Aβ processing deficits in the absence of alterations in tau phosphorylation.

By using functional genomics techniques to compare cortical neurons derived from the iPSCs of an Alzheimer's disease patient with an intron 4 *PSEN1* mutation and a healthy control, I aimed to quantify the contributions of batch-to-batch variability to measures of gene expression in iPSC-derived cortical neurons.[176,177] In parallel, the use of the isogenic *PSEN1* mutation series enables the contribution of background genetic variants in driving any transcriptomic differences to be identified.

3.4 Methods

3.4.1 Derivation of iPSCs and quality control

The StemBANCC control (SBAd03-01) and Alzheimer's disease patient with a *PSEN1* intron 4 mutation (SFC808-04) iPSC lines were derived from skin fibroblasts using Sendai virus vectors and underwent normal StemBANCC quality control measures as described in **2.4.1**, **2.4.2**, **2.4.3**, **2.4.4** and **2.4.5**.

Isogenic *PSEN1* mutant iPSC lines were kindly supplied by Grace Woodruff from the University of California, San Diego (UCSD).[95] In brief, skin fibroblasts from a healthy control (J. Craig Venter) were reprogrammed using the Moloney Murine Leukemia Virus (MMLV) vectors *OCT4*, *SOX2*, *KLF4*, *c-MYC* and ± *EGFP* to iPSC and then cultured on mouse embryonic fibroblasts

(MEF).[178] Custom-designed TALEN constructs were used to insert *PSEN1* Δe9 mutations with single stranded oligodeoxynucleotide templates either hetero- or homozygotically. After TALEN-editing, clones were screened for successful homology directed repair using allele-specific PCR. iPSCs were checked for off-target effects or mutations in coding sequences using exome-sequencing. After expansion in Oxford, DNA was extracted using the QIAamp Mini Kit (Qiagen) and run on the HumanOmniExpress BeadChip Kit (Illumina). Karyograms were produced using GenomeStudio (Illumina) and showed no large-scale karyotypic abnormalities (data not shown). Sanger sequencing was used to confirm the presence of the expected mutation using the PCR primer pair, TTGCTGCTGGCATTTGGTTC and ACAGTGACCCTGAAAAATCAAGA, and the forward sequencing primer, GTGGAATCTTGCTGGCCTGA. Sequencing results were visualised in Chromas Lite (v. 2.1.1).

3.4.2 Cortical neuronal differentiation

iPSC lines were differentiated into cortical neurons as described in **2.4.6** with some modifications. UCSD *PSEN1* isogenic iPSC lines were passaged every 2-4 days from 10 days until 35 days post-neural induction and then plated for final differentiation (FP = final plating). Cells were cultured until 90 days post-neural induction (FP + 55).

SBAd03-01 and SFC808-04 iPSC lines were passaged every 2-4 days from 10 days post-neural induction until the appearance of neurons and then plated for final differentiation. Cells were cultured until 55 days after final plating (FP + 55). Neural maintenance medium was supplemented with 500μM pyruvate (Sigma). The differences resulted from the fact that these lines were part of a larger multi-centre experiment into inter-laboratory variability within StemBANCC following an externally specified standard operating procedure.

Each line was differentiated independently three times apart from SFC808-04, in which the third differentiation failed to produce neurons despite repeated passaging. The first batch of UCSD isogenic *PSEN1* lines were either immediately plated for final differentiation or kept frozen in liquid nitrogen for three months before being plated in tandem with the remaining differentiation batches.

3.4.3 Immunofluorescence microscopy

This was conducted as described in **2.4.11**.

3.4.4 Aβ secretion

Media was removed from cells, spun down at 400g for 5 minutes at 4°C and the supernatant flash frozen on dry ice. This was analysed using a MesoScale Discovery Aβ Peptide Panel1 (6E10) V-PLEX Plus Kit (K15200G-1).

3.4.5 RNA extraction

RNA was extracted from cells using the Mini RNeasy Kit (Qiagen) with on column DNase treatment. The quantity and RNA integrity number (RIN) of selected samples were estimated using the RNA Analysis ScreenTape (Agilent). In all tested samples, RIN was > 8.

3.4.6 RNA-seq

Libraries were prepared using polyA purification of mRNA, followed by reverse transcription and library preparation using the TruSeq RNA Library Preparation Kit (Illumina). In the case of UCSD isogenic *PSEN1* samples, 1.5μL of Exogenous RNA Control Consortium (ERCC) mix 1 were spiked-in at a 1:100 dilution prior to processing. Libraries were multiplexed and sequenced on an Illumina 2500 sequencing machine.

3.4.7 RNA-seq analysis pipeline

Quality metrics were derived from FASTQ files using FastQC.[179] 50 base-pair paired-end reads were aligned to the UCSC hg19 genome assembly and ERCC spike-ins using HISAT two-pass alignment (version 0.1.6 with the arguments; first pass: "-p 8 --known-splicesite-infile <UCSC spicesites> --novel-splicesite-outfile <novel splicesites> -x <genome index> -1 <forward reads> -2 <reverse reads>" and second pass: "-p 8 --known-splicesite-infile <UCSC spicesites> --novel-splicesite-infile <novel splicesites> -x <genome index> -1 <forward reads> -2 <reverse reads>".[180] Read pairs in which both reads were uniquely mapped were assigned to UCSC hg19 genes using HTSeq (htseq-count with the arguments "-s reverse -m intersection-nonempty").[181]

Exon-level expression data was generated using BEDtools coverage (with the argument "-split").[182] The ratio between spliced and non-spliced read counts at each exon was computed as the coverage using BEDtools coverage without "-split" divided by the coverage with "-split". Differential expression analysis of genes with non-zero expression was conducted using general linear modelling in edgeR, correcting for common, trended and tagwise dispersion and using effective library sizes adjusted by upper quartile or trimmed mean of M-values.[183] Highly variable genes were identified using technical modelling of ERCC spike-ins as described by Brennecke and colleagues for single cell RNA-seq experiments.[129] The one-way empirical significance of highly variable genes in the StemBANCC lines (lacking ERCC spike-ins) was estimated by comparing the median squared coefficient of variation (CV^2) of highly variable genes identified in the UCSD *PSEN1* isogenic lines with the median CV^2 of 1,000 randomly selected sets of genes matched by decile of mean expression level. Counts were adjusted by the removal of unwanted variation (RUVSeq), using either ERCC spike-ins where available or regressing out variation due to all but the top 5,000 differentially expressed genes based on *PSEN1* genotype.[184] Clustering analysis was performed using principal component analysis (the prcomp R package) on adjusted counts, limiting the analysis to genes expressed in at least 20% of samples. The biological coefficient of variation (BCV) was calculated for groups of samples using edgeR. One-way empirical significance was estimated by randomising experimental labels without replacement 1,000 times and recalculating BCVs. The two-way significance of differences between groups of samples was similarly calculated by comparison with this permuted dataset. Gene ontology enrichment was assessed using DAVID.[149]

3.4.8 Shotgun proteomics

Cells were dissociated with ice cold PBS and collected as a pellet. This was flash frozen on dry ice and stored at -80^{o}C. Pellets were treated with trypsin digestion overnight with solid phase extraction. Desalted peptides were resuspended in 3% Acetonitrile, 0.1% TFA and spiked in 1:10 ratio with yeast enolase (an internal standard used for quantitation) and injected into a nanoUPLC Acquity system mass spectrometry machine coupled with Waters Synapt G2-Si with ion mobility

separation. The resultant chromatogram was analysed using PLGS (version 3.0.2) as described in ref. [185].

3.4.9 Proteomics analysis

The abundance (in femtomoles) of all proteins detected at FDR < 0.05 was summated by gene and adjusted for the total amount of protein detected. Correlation with RNA expression was estimated using Spearman's correlation on genes detectable as RNA transcripts and proteins. Differential protein analysis was undertaken using an ANOVA test on all proteins detectable in at least 20% of samples. P-values were corrected for multiple hypothesis testing using the Benjamini-Hochberg method.

3.4.10 Genome-wide association study data

Single nucleotide polymorphism (SNP) data from genome-wide association studies (GWAS) were downloaded from the GWAS Catalogue (accessed on 5th March 2016).[186] These were manually curated for terms associated with Alzheimer's disease susceptibility, pathology or biomarkers, cognitive performance or MRI measures of brain volume. Intervals of 100kb on either side of all reported SNPs were then intersected with UCSC hg19 genes. These were intersected with the differential gene expression results of edgeR to obtain a vector of likelihood ratios (LR). The median LR for each GWAS trait was compared with the median LR from 10,000 permuted gene-sets matched by expression decile and a one-way empirical p-value estimated. P-values were adjusted for multiple hypothesis testing using Benjamini-Hochberg correction.[187]

3.4.11 Alzheimer's disease-related sets of genes

The enrichment of differential expression within specific sets of genes derived from published results of doubly transgenic mice (*APP*$^{KM670/671NL}$::*PSEN1*$^{\Delta e9}$) or integrative network analysis of multiple Alzheimer's disease datasets was undertaken in a similar manner to that described for GWAS SNPs.[188,189] Genes were intersected with edgeR differential gene analysis output and permutation analysis undertaken on the median LR with genes matched by expression decile.

One-way empirical p-values were estimated from 10,000 permutations and adjusted for multiple hypothesis testing using Benjamini-Hochberg correction.

3.5 Results

3.5.1 Cortical differentiation in iPSC lines

I included five iPSC lines in the research presented in this chapter (**Table 2**). Each iPSC line was differentiated into cortical neurons for > 90 days (until FP + 55) using a very similar method to that described in **Chapter 2**.

Table 2 Details of iPSC lines used. The age is the number of days post-neural induction

iPSC line	Mutation	Number of differentiations	Age at harvest (days)
PSEN1$^{UCSD::wt/wt}$	UCSD isogenic, wild-type	3	90, 90, 90
PSEN1 $^{UCSD::wt/\Delta e9}$	UCSD isogenic, heterozygote, G>T, exon 9 skipping	3	90, 90, 90
PSEN1 $^{UCSD::\Delta e9/\Delta e9}$	UCSD isogenic, homozygote, G>T, exon 9 skipping	3	90, 90, 90
PSEN1$^{SB::wt/wt}$	StemBANCC SBAd03-01, wild-type	3	116, 116, 115
PSEN1$^{SB::wt/int4}$	StemBANCC SFC808-04, G>del, multiple abnormal transcripts	2	119, 118

The isogenic *PSEN1* series was kindly supplied by collaborators at UCSD. I used Sanger sequencing to check that the sequence around the engineered *locus* was as expected prior to expansion of these iPSCs (**Figure 9a**). I then confirmed that each line differentiated successfully into cortical neurons by performing immunofluorescence microscopy for TUJ1 and TBR1 (**Figure 9b-f**). A previous publication had shown decreased $A\beta_{1\text{-}40}/_{1\text{-}42}$ ratios in *PSEN1*$^{UCSD::wt/\Delta e9}$ and *PSEN1* $^{UCSD::\Delta e9/\Delta e9}$ neural progenitors and neurons compared with *PSEN1*$^{UCSD::wt/wt}$, demonstrating the impairment of PSEN1 γ-secretase function in these mutant lines.[95] This was also the case in *PSEN1*$^{SB::wt/int4}$ compared with *PSEN1*$^{SB::wt/wt}$ iPSC-derived cortical neurons (**Figure 9g**).

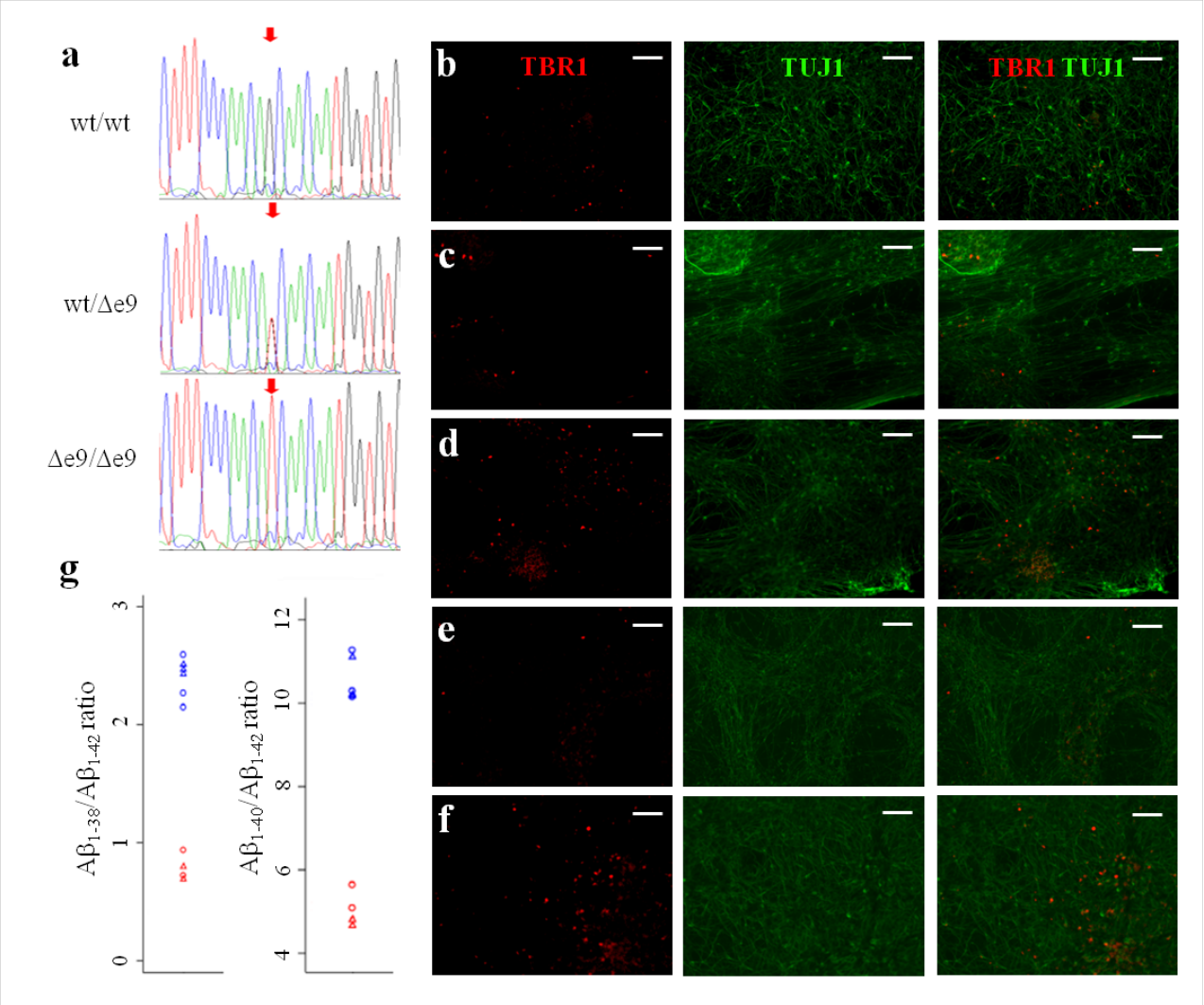

Figure 9 Validation of UCSD and StemBANCC iPSC-derived cortical neurons. (**a**) Sanger sequencing traces for *PSEN1*$^{UCSD::wt/wt}$ (top), *PSEN1*$^{UCSD::wt/\Delta e9}$ (middle) and *PSEN1*$^{UCSD::\Delta e9/\Delta e9}$ (bottom). The TALEN-engineered base (G>T) is indicated by the red arrow. Immunofluorescence microscopy images showing TBR1 (red) and TUJ1 (green) staining plus merged images on iPSC-derived cortical neurons at 49-50 days post-neural induction from: (**b**) *PSEN1*$^{UCSD::wt/wt}$, (**c**) *PSEN1*$^{UCSD::wt/\Delta e9}$, (**d**) *PSEN1*$^{UCSD::\Delta e9/\Delta e9}$ (**e**) *PSEN1*$^{SB::wt/wt}$ and (**f**) *PSEN1*$^{SB::wt/int4}$. Scale bars represent 40μm. (**g**) Aβ ratios for *PSEN1*$^{SB::wt/wt}$ (blue) and *PSEN1*$^{SB::wt/int}$ (red) samples at either FP+25 (circles) or FP+55 (triangles).

3.5.2 Clustering RNA-seq samples

Once I was convinced that the iPSC lines had successfully differentiated into cortical neurons and recapitulated aspects of Alzheimer's disease biology associated with Aβ secretion, I proceeded to perform transcriptomic profiling of iPSC-derived cortical neurons. There was a good rate of read alignment in all samples (> 94%). These samples had high level expression of genes associated with successful corticogenesis in a previous study of human embryonic stem cell-based (hES) cortical differentiation (**Figure 10a**).[64] It was also possible to confirm the expected abnormalities

of *PSEN1* exon-skipping in neurons derived from each of the UCSD isogenic iPSC lines, however exon skipping events were less obvious in *PSEN1*$^{SB::wt/int4}$ (**Figure 10b&c**).

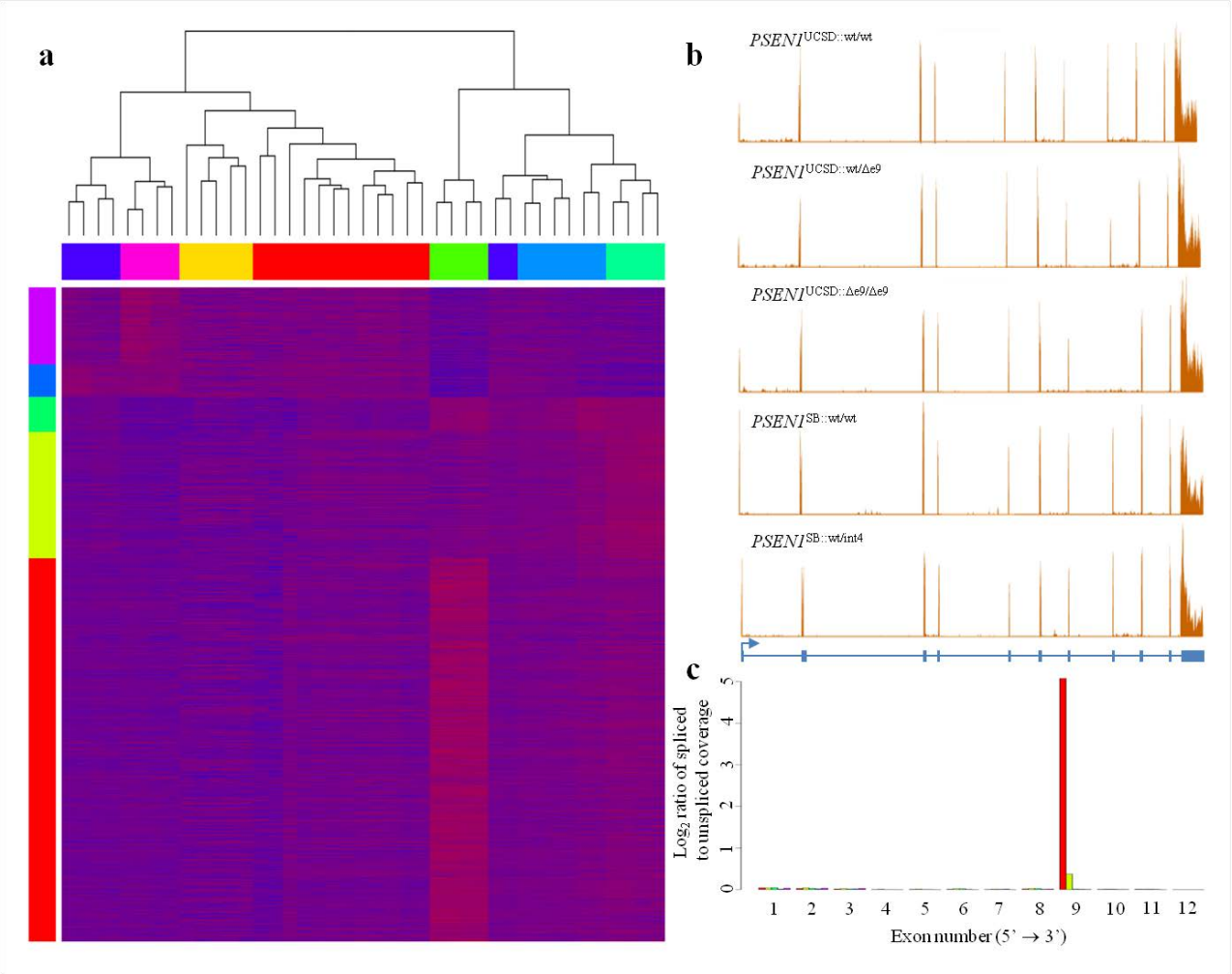

Figure 10 Transcriptomic characterisation of iPSC-derived cortical neurons. (**a**) A heatmap with rows showing between-sample scaled $\log_2$ expression (red = high and blue = low) for genes associated with different stages of stem cell corticogenesis and columns showing samples clustered by similarity. The vertical colour bar indicates genes associated with pluripotency (red), neural differentiation (yellow), cortical regionalisation (green), deep layer neuron generation (blue) and upper layer neuron generation (purple). The horizontal colour bar indicates each sample: green = hES (undifferentiated), turquoise = hES-derived neuroectoderm, light blue = hES-derived cortical neural progenitors, dark blue = hES-derived deep layer neurons, purple = hES-derived upper layer neurons, red = UCSD iPSC-derived cortical neurons and yellow = StemBANCC iPSC-derived cortical neurons. (**b**) Exon-level coverage of *PSEN1* for iPSC-derived cortical neurons at FP+55. (**c**) A barchart showing the $\log_2$ ratio of spliced to unspliced reads across *PSEN1*. More positive values indicate a higher relative level of exon skipping. Red = *PSEN1*$^{UCSD::\Delta e9/\Delta e9}$, yellow = *PSEN1*$^{UCSD::wt/\Delta e9}$, green = *PSEN1*$^{UCSD::wt/wt}$, blue = *PSEN1*$^{SB::wt/wt}$ and purple = *PSEN1*$^{SB::wt/int4}$.

3.5.3 Clustering analysis of iPSC-derived cortical neurons

I then performed clustering analysis on these samples in order to examine the extent to which *PSEN1* mutations might drive high-level differences between differentiations. This showed that

for the UCSD *PSEN1* isogenic series, different genotypes clustered closely together, with a clear batch effect resulting from whether neural progenitors from batch 1 were directly plated for final differentiation or frozen down and differentiated in parallel with later batches, even when adjusting for batch-to-batch variability (**Figure 11a**). In contrast, there was an obvious separation between *PSEN1*$^{SB::wt/wt}$ and *PSEN1*$^{SB::wt/int4}$ (**Figure 11b**). It was clear when attempting to cluster UCSD and StemBANCC lines together that the vast majority of the variation was inter-experimental even when limiting clustering to wild-type and *PSEN1* heterozygote samples and using RUVSeq to minimise variation unrelated to *PSEN1* genotype (**Figure 11c&d**).

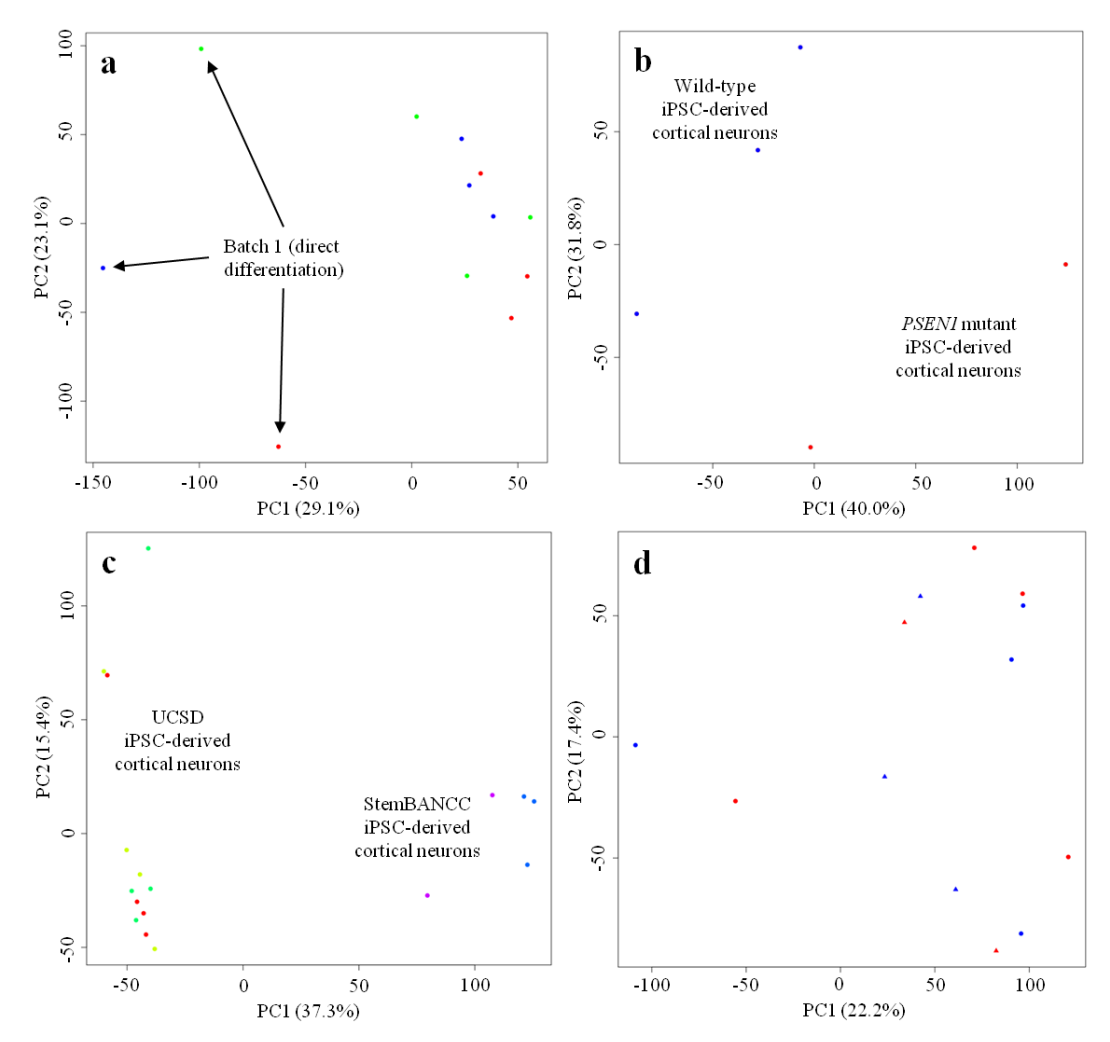

Figure 11 Principal component analysis of iPSC-derived cortical neurons. (**a**) *PSEN1*$^{UCSD::wt/wt}$ (blue), *PSEN1*$^{UCSD::wt/\Delta e9}$ (red) and *PSEN1*$^{UCSD::\Delta e9/\Delta e9}$ (green). (**b**) *PSEN1*$^{SB::wt/wt}$ (blue) and *PSEN1*$^{SB::wt/int4}$ (red). (**c**) *PSEN1*$^{UCSD::wt/wt}$ (green), *PSEN1*$^{UCSD::wt/\Delta e9}$ (yellow), *PSEN1*$^{UCSD::\Delta e9/\Delta e9}$ (red), *PSEN1*$^{SB::wt/wt}$ (blue), *PSEN1*$^{SB::wt/int4}$ (purple). (**d**) *PSEN1*$^{UCSD::wt/wt}$ (blue circles), *PSEN1*$^{UCSD::wt/\Delta e9}$ (red circles), *PSEN1*$^{SB::wt/wt}$ (blue triangles), *PSEN1*$^{SB::wt/int4}$ (red triangles). Counts were adjusted by trimmed mean of M-values (**a-c**) and by RUVSeq on genes

not differentially expressed by *PSEN1* genotype in (**d**).[184] For clarity, clusters have been labelled where possible.

3.5.4 Differentially expressed genes in isogenic *PSEN1* iPSC-derived cortical neurons

I used edgeR to assess which genes were differentially expressed between *PSEN1*$^{UCSD::wt/wt}$, *PSEN1*$^{UCSD::wt/\Delta e9}$ and *PSEN1* $^{UCSD::\Delta e9/\Delta e9}$.[183] In total, 357 genes were differentially expressed (FDR < 0.05) when using RUVSeq to remove batch-to-batch variability based on the expression level of ERCC spike-ins.[184] These genes were enriched for gene ontology terms associated with cellular adhesion and the extracellular matrix (**Figure 12a**). I wondered whether many differentially expressed genes could have been obscured by the fact that there was considerable variation between the first batch of directly differentiated neurons and subsequent batches. I therefore excluded the first batch of samples from further analysis. When I removed this set of samples, the number of differentially expressed genes *decreased* to 129. This set was highly enriched for terms associated with neuronal fate in the forebrain (driven mainly by *ARX*, *DLX1*, *DLX2* and *NKX2-2*; **Figure 12b**). 91 of the differentially expressed genes showed a consistent direction of difference in *PSEN1*$^{UCSD::wt/\Delta e9}$ and *PSEN1* $^{UCSD::\Delta e9/\Delta e9}$ relative to *PSEN1*$^{UCSD::wt/wt}$, implying that not all of these differences are likely to be driven simply by loss of γ-secretase activity. This was reinforced by the fact that the likelihood ratio (LR) estimates for the comparisons between *PSEN1*$^{UCSD::wt/wt}$ and either *PSEN1*$^{UCSD::wt/\Delta e9}$ or *PSEN1* $^{UCSD::\Delta e9/\Delta e9}$ were weakly correlated (Spearman's $r = 0.30$, $p < 10^{-20}$). The top gene ontology term enrichment in this set of genes was associated with proximal distal pattern formation (*HOXC10*, *HOXC11*, *HOXA10* and *HOXA9*).

One of the main benefits of iPSC-based models is the possibility of identifying genes involved in the aetiology and pathogenesis of human disease. In order to assess whether any of the identified differentially expressed genes may contribute to these processes, I used genome-wide association single nucleotide polymorphisms (SNPs) reported to show an association with Alzheimer's disease or an Alzheimer's disease-related trait to identify a list of candidate genes within 100kb of each SNP.[186] After intersecting this with the results of the differential expression analysis above, there was significantly higher magnitude differential expression than expected by chance near

regions associated with hippocampal volume (median LR = 2.64, permuted LR = 1.04, $p < 0.0001$, $q = 0.004$) (**Figure 12c**). The top differentially expressed genes contributing to this was *HOXC10* (adjusted counts $PSEN1^{UCSD::wt/wt} = 52.7$, $PSEN1^{UCSD::wt/\Delta e9} = 21$, $PSEN1^{UCSD::\Delta e9/\Delta e9} = 3$, LR = 22.6, $p < 0.0001$, $q = 0.004$) near rs2630772 This was a nominally significant ($p < 10^{-5}$) *locus* derived from a meta-analysis of two Australian neuroimaging GWAS cohorts.[190] *HOXC10* has never been linked to hippocampal development or function.

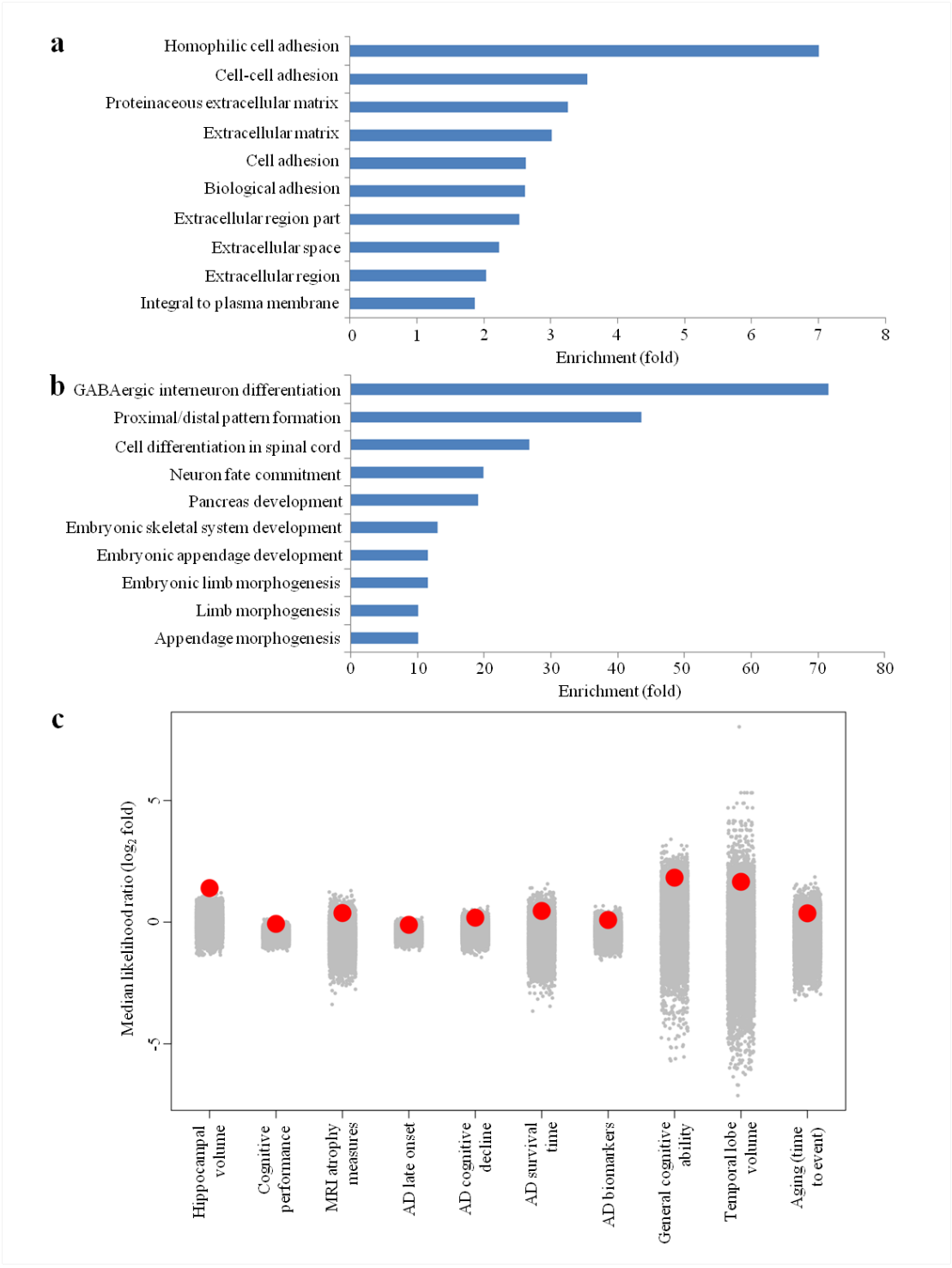

Figure 12 Differential expression analysis of UCSD *PSEN1* isogenic iPSC-derived cortical neurons. Top 10 enriched gene ontology terms (FDR < 0.05) in (**a**) all UCSD *PSEN1* isogenic differentiations and (**b**) after excluding the first batch plated directly for differentiation. (**c**) Enrichment of differential expression likelihood ratios (LR) within 100kb of GWAS SNPs. Red points indicate actual median LR and grey points indicate the median LR from 10,000 expression-matched permutations.

Intersecting my list of differentially expressed genes with GWAS intervals associated with Alzheimer's disease yielded eight genes (**Table 3**). *PNPO* and *FLT1* are particularly plausible candidates. *PNPO* encodes the enzyme pyridoxamine 5'-phosphate oxidase, which is expressed in the developing hippocampus and has a role in long term potentiation in the dentate gyrus.[191,192] *PNPO* is located near rs4794202, a SNP significantly associated with cognitive decline in Alzheimer's disease ($p < 10^{-7}$).[193] *FLT1* encodes a receptor for vascular endothelial growth factor (VEGF). *Flt1* is known to be expressed in both progenitors and post-mitotic neurons within the cerebral cortex and hippocampus.[194] VEGF signalling through Flt1 promotes survival and $A\beta_{1-42}$ clearance in cultured rodent cortical neurons, and is strongly upregulated following treatment with neurotoxic venom.[195,196] However, Flt1-signalling may promote microglial chemotaxis and inflammation, so the overall *in vivo* role of *FLT1* in Alzheimer's disease pathology is likely to be complex.[197] *FLT1* is located near rs17086609, which demonstrated a nominal association with cognitive performance ($p < 10^{-5}$).[198]

Table 3 Differentially expressed genes within 100kb of GWAS *loci* in UCSD *PSEN1* isogenic iPSC-derived cortical neurons. AD = Alzheimer's disease.

Gene	Adjusted counts (*PSEN1*^{UCSD})			LR	P-value	Q-value	GWAS trait
	wt/wt	wt/Δe9	Δe9/Δe9				
ELSPBP1	3.0	1.0	1.0	19.9	$< 10^{-4}$	0.01	Intelligence
FLT1	28.3	8.7	1.0	29.1	$< 10^{-6}$	0.0002	Cognitive performance
HOXB-AS3	6.7	1.7	1.0	37.0	$< 10^{-8}$	$< 10^{-5}$	Cognitive performance
HOXC10	113.0	13.7	2.7	31.6	$< 10^{-6}$	$< 10^{-4}$	Hippocampal volume
HOXC11	14.7	1.7	1.0	20.6	$< 10^{-4}$	0.008	Hippocampal volume
LIMS2	273.3	161.7	91.7	19.7	$< 10^{-4}$	0.01	AD cognitive decline
LOC100505738	3.0	308.7	3.3	297.8	$< 10^{-64}$	$< 10^{-60}$	Verbal declarative memory
PNPO	821.3	654.7	11.7	247.6	$< 10^{-53}$	$< 10^{-50}$	AD cognitive decline

It was possible that the differential gene signature may better reflect ongoing pathogenic processes in the brain rather than genetic signals associated with susceptibility or cognitive traits. In order to investigate whether this was the case, I intersected the dentate gyrus differential gene

expression profile of a doubly transgenic mouse model bearing mutations in *APP* and *PSEN1* (*APP*$^{KM670/671NL}$::*PSEN1*$^{\Delta e9}$) compared with wild-type with that obtained from this experiment.[188] There was a modest enrichment in the magnitude of differential expression at genes that were differentially expressed in the transgenic mouse model (median LR = 1.57, permuted LR = 0.77, $p < 0.0001$). However, only *TLL1* was in common between the mouse model and my iPSC-derived cortical neuron model. However the direction of expression of *TLL1* with *PSEN1* genotype was the opposite of that see in the mouse model.

One possible reason why so few differentially expressed genes were detected could be that *PSEN1* genotypes have a largely post-transcriptional effect. In an attempt to address this, I analysed shotgun proteomic data from the same differentiations. A median of 1,328 genes were represented amongst the detected proteins (range 1,201 - 1,412). Detectable proteins were mainly enriched for biological pathways associated with metabolism and cytoskeletal functions. However, there was also significant enrichment for terms related to axonal transport, clathrin-coated vesicle trafficking and neuronal synaptic plasticity, reinforcing the neuronal identity of iPSC-derived cortical neurons. There was a moderate correlation between RNA expression and protein abundance for paired samples from the same batch of differentiation (mean ± SD Spearman's r = 0.50 ± 0.02; **Figure 13a**). Detectable proteins were mostly confined to those genes with relatively high RNA expression (**Figure 13b**). Proteins that were detected in the absence of RNA expression (n = 72) were enriched for gene ontology pathways associated with glycolysis, keratin filaments and chromatin. The presence of keratin-related pathways suggested that at least some of these proteins were likely to represent contamination. None of the differentially expressed genes identified by RNA-seq were detected by shotgun proteomics. Two proteins were differentially expressed between different *PSEN1* genotypes: MACROD1 ($q < 10^{-5}$) and NAMPT ($q < 10^{-6}$) (**Figure 13c&d**). *MACROD1* has no known role in Alzheimer's disease or corticogenesis. *NAMPT* encodes the enzyme nicotinamide phosphoribosyltransferase, which is an intermediate step in the synthesis of NAD. *NAMPT* is an attractive candidate as a microdeletion in this gene has been reported in association with Alzheimer's disease.[199] NAMPT levels reduced with age in a transgenic mouse model of Alzheimer's disease and knocking out

Nampt in excitatory neurons resulted in widespread neurodegeneration.[200,201] Overexpression of NAMPT in astrocytes was also shown to mitigate motor neuron death in a mouse model of amyotrophic lateral sclerosis, suggesting that this may be a common pathway in neurodegenerative processes.[202] However, why this protein should have absent *only* in *PSEN1*$^{UCSD::\Delta e9/\Delta e9}$ is less clear, since intuitively it might have been expected that *PSEN1* mutants would have a compensatory *increase* in the abundance of neuroprotective proteins. In the case of both MACROD1 and NAMPT, there was no significant difference in RNA expression levels between *PSEN1* genotypes.

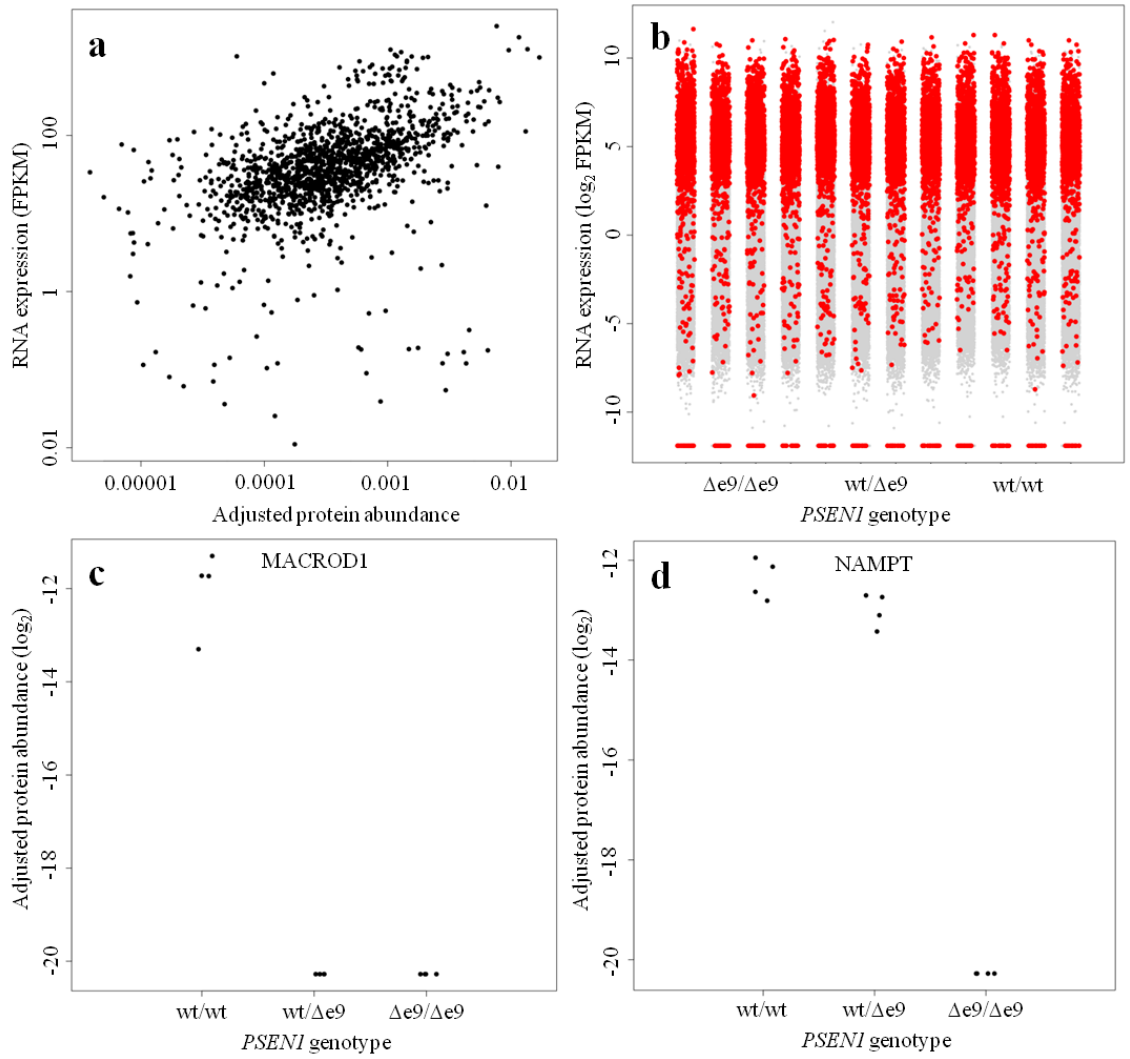

Figure 13 Shotgun proteomics of UCSD *PSEN1* isogenic iPSC-derived cortical neurons. (**a**) Scatterplot of protein abundance against RNA expression for genes detected by shotgun proteomics and RNA-seq in batch two of *PSEN1*$^{UCSD::wt/wt}$ iPSC-derived cortical neurons. (**b**) Plots of RNA expression levels (light gray) highlighting detected proteins in red. Plots of adjusted $\log_2$ protein abundance by *PSEN1* genotype for (**c**) MACROD1 and (**d**) NAMPT.

3.5.5 Differentially expressed genes between StemBANCC wild-type and mutant *PSEN1* iPSC-derived cortical neurons

Controlling for batch-to-batch variability in *PSEN1*$^{SB::wt/wt}$ and *PSEN1*$^{SB::wt/int4}$ iPSC-derived cortical neurons was more challenging, as these samples did not contain ERCC spike-ins. I used RUVSeq to find factors that maximally reduced the variation in genes not associated with *PSEN1* genotype and supplied edgeR with the resultant normalisation factors for differential expression analysis.[184,203] This identified 2,597 differentially expressed genes (FDR < 0.05) between *PSEN1*$^{SB::wt/wt}$ and *PSEN1*$^{SB::wt/int4}$. This set of genes was enriched for gene ontology terms involved with glial and neuronal fate specification (*ATOH1*, *LBX1*, *GSX2*, *NKX6-2*, *FOXA1*, *TLX3*, *NKX2-2*, *SHH* and *NKX6-1*), adhesion (*CDKN2A*, *ACVRL1*, *BCL6*, *COL1A1*, *MMP14*, *THBS1* and *PTENP1*) and tissue remodelling (*CALCA*, *CD38*, *P2RX7*, *TNFRSF11B*, *IAPP*, *CST3*, *CARTPT* and *INPP5D*) amongst others (**Figure 14a**). As with the *PSEN1* isogenic series, I intersected genes with Alzheimer's disease-associated GWAS SNPs to identify enriched disease-related processes. No GWAS terms were significantly enriched after correction for multiple hypothesis testing (**Figure 14b**). The top term was again hippocampal volume (median LR = 2.51, permuted LR = 1.18, p = 0.01, q = 0.46). Of the significantly differentially expressed genes, 159 were within 100kb of one of the GWAS SNPs. **Table 4** shows differentially expressed genes within 100kb of GWAS SNPs that were genome-wide significant ($p \leq 10^{-7}$).

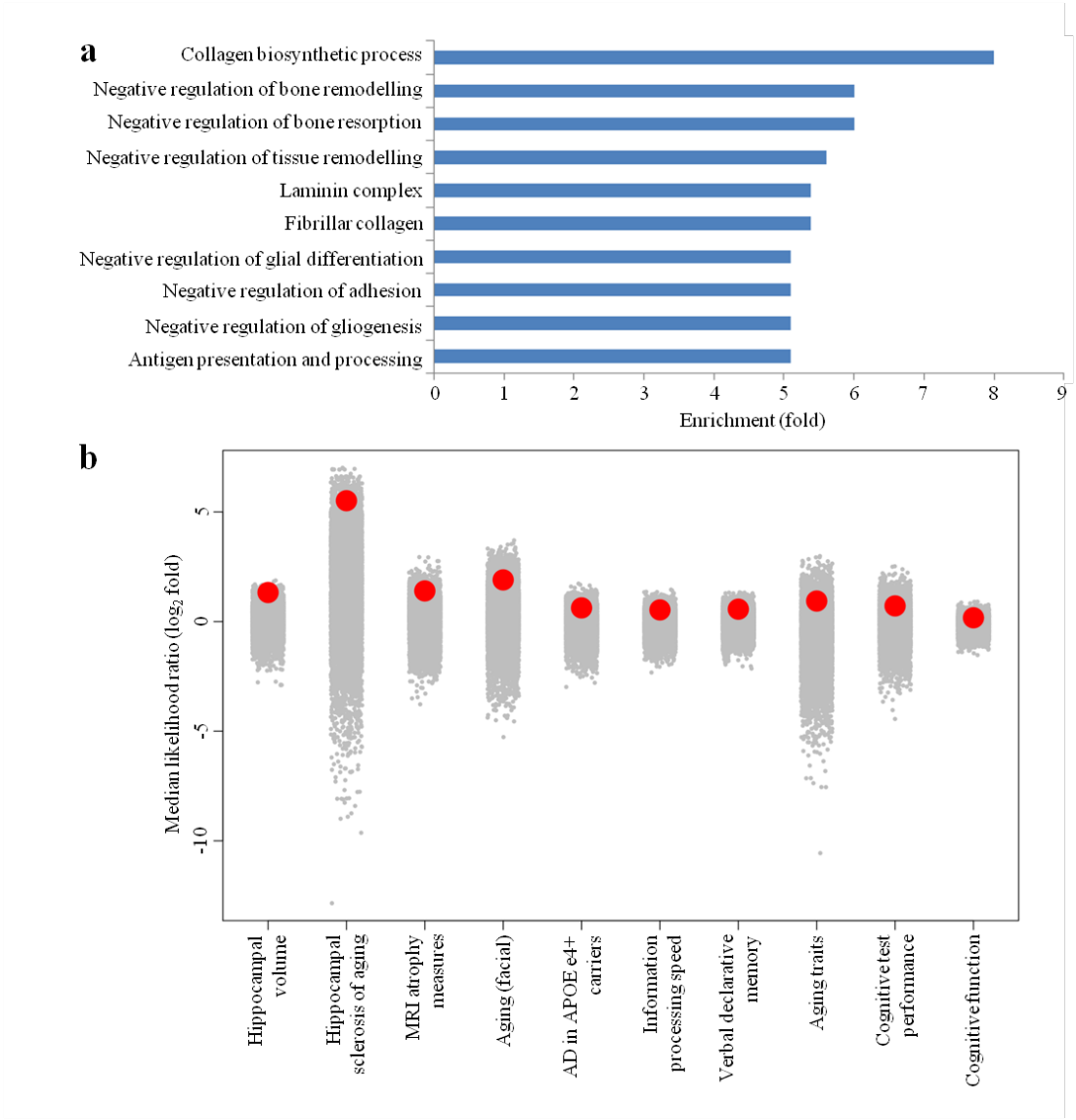

Figure 14 Differential expression analysis of StemBANCC iPSC-derived cortical neurons. (**a**) Top 10 enriched gene ontology terms (FDR < 0.05). (**b**) Enrichment of differential expression likelihood ratios (LR) within 100kb of GWAS SNPs. Red points indicate actual median LR and grey points indicate the median LR from 10,000 expression-matched permutations.

Table 4 Differentially expressed genes within 100kb of GWAS *loci* in StemBANCC iPSC-derived cortical neurons. A p-value threshold of $p \leq 10^{-7}$ was used to prioritise GWAS SNPs. AD = Alzheimer's disease.

Gene	Adjusted counts (*PSEN1*^{SB})		LR	P-value	Q-value	GWAS trait
	wt/wt	wt/int4				
ABCC9	13.7	70.5	51.6	$< 10^{-12}$	$< 10^{-10}$	Hippocampal sclerosis of aging
ACE	133.0	212.5	7.8	0.005	0.04	CSF levels of AD proteins
APOE	541.0	3725.0	103	$< 10^{-23}$	$< 10^{-21}$	AD, AD age of onset, AD biomarkers, AD late onset, CSF $A\beta_{1-42}$ in controls and AD, cognitive decline, cognitive age-related decline,

						cognitive function, dementia and AD neuropathology, verbal declarative memory
ASTN2	302.0	1116.0	67.6	$< 10^{-15}$	$< 10^{-13}$	Hippocampal volume
AURKA	255.3	107.5	15.1	0.0001	0.002	AD late onset
BAIAP2-AS1	3644.3	2331.0	8.0	0.005	0.04	Memory performance
CACNA1G	1176.0	1851.5	8.7	0.003	0.03	AD cognitive decline
CCZ1B	119.3	239.0	14.3	0.0002	0.002	AD APOE-ε4 interaction
CLU	13323.7	45985.5	47.1	$< 10^{-11}$	$< 10^{-9}$	AD, AD in APOE-ε4 -, AD in APOE-ε4 +, AD late onset
COL1A2	158.7	2082.0	26.0	$< 10^{-6}$	1.10E-05	Intelligence
CR2	1.0	5.0	15.0	0.0001	0.002	AD, AD in APOE-ε4 +, AD late onset
CSMD2	327.0	1497.0	20.8	$< 10^{-5}$	0.0001	Intelligence
CYB561	384.3	628.5	8.1	0.004	0.04	CSF levels of AD proteins
EPHX2	375.3	692.5	11.2	0.0008	0.01	AD, AD in APOE-ε4 -, AD in APOE-ε4 +, AD late onset
GLIS3	141.7	904.5	76.2	$< 10^{-17}$	$< 10^{-15}$	AD biomarkers
HLA-DQB1	1.0	9.0	26.9	$< 10^{-6}$	$< 10^{-5}$	Aging (facial)
HRK	41.7	116.0	21.1	$< 10^{-5}$	0.0001	Hippocampal volume
IL6R	15.0	60.5	36.1	$< 10^{-8}$	< 10-7	CSF levels of AD proteins
KANSL1-AS1	14.3	77.0	43.2	$< 10^{-10}$	$< 10^{-8}$	AD in APOE-ε4 -
KCNJ8	46.0	162.5	39.8	$< 10^{-9}$	$< 10^{-7}$	Hippocampal sclerosis of aging
LHFP	405.3	1071.5	14.9	0.0001	0.002	Hippocampal atrophy
MEF2C	707.3	296.5	22.1	$< 10^{-5}$	< 0.0001	AD late onset
MMP10	7.0	21.5	10.8	0.001	0.01	CSF levels of AD proteins
MYO16	266.7	634.5	26.3	$< 10^{-6}$	$< 10^{-5}$	AD cognitive decline
PNPO	163.0	397.5	10.9	0.0009	0.01	AD cognitive decline
PTK2B	516.7	885.5	11.8	0.0006	0.007	AD late onset
RIN3	51.7	96.5	9.9	0.002	0.02	AD in APOE-ε4 -, AD late onset
TREML3P	3.3	1.0	8.5	0.004	0.03	AD
WIF1	34.3	141.5	41.4	$< 10^{-9}$	$< 10^{-8}$	Hippocampal volume

As with the UCSD *PSEN1* isogenic lines, I assessed the enrichment of differential expression near the same doubly transgenic mouse model of Alzheimer's disease.[188] There was a significant enrichment of differential expression (median LR = 2.40, permuted LR = 1.13, $p < 0.0001$). This was driven mainly by 90 significantly differentially expressed genes, which were enriched for gene ontology terms associated with vesicle trafficking (*F13A1*, *SERPING1*, *THBS2*, *TGFB2* and *FN1*) and the extracellular matrix (*OGN*, *LAMA4*, *TNFRSF11B*, *ADAMTS16*, *COL3A1*, *COL1A2*,

CCDC80, *MGP*, *WNT6*, *PRSS12*, *TGFB2* and *FN1*). The majority of these (60 genes, 66.7%) showed the same direction of differential expression that was reported in the mouse model.

3.5.6 Intersecting differentially expressed genes from UCSD and StemBANCC iPSC-derived cortical neurons

The comparison of *PSEN1*$^{SB::wt/wt}$ and *PSEN1*$^{SB:wt/int4}$ alone cannot make any firm statements about disease biology as it is impossible to distinguish clone-to-clone variability from disease-related transcriptomic signatures. Any genetic signatures that derive purely from mutation within *PSEN1* should be common to both the isogenic and case-control analysis.

In order to enable a direct comparison, I performed RUVSeq-based adjustment in the isogenic *PSEN1* dataset using endogenous gene controls as described above, identifying 252 differentially expressed genes (FDR < 0.05).[184] The magnitude of LR estimates was quite strongly correlated between the differential expression results when using either ERCC spike-ins or endogenous controls for adjustment, suggesting that most results were robust against different adjustment methods (Spearman's r = 0.72, $p < 10^{-20}$; **Figure 15a**).

Overall, the correlation between measures of LR in both datasets was poor (Spearman's r = 0.23, $p < 10^{-20}$; **Figure 15b**). 98 genes were differentially expressed both between *PSEN1*$^{SB::wt/wt}$ and *PSEN1*$^{SB:wt/int4}$, and *PSEN1*$^{UCSD::wt/wt}$, *PSEN1*$^{UCSD::wt/\Delta e9}$ and *PSEN1*$^{UCSD::\Delta e9/\Delta e9}$ (**Figure 15c**). However, only 45 of these had the same direction of change in *PSEN1*$^{SB::wt/wt}$ *vs.* *PSEN1*$^{SB:wt/int4}$, and *PSEN1*$^{UCSD::wt/wt}$ *vs.* *PSEN1*$^{UCSD::wt/\Delta e9}$ and an even smaller number (28) had a consistent direction of change relative to wild-type in all *PSEN1* mutants. This poor correspondence was not driven solely by effects unique to a specific *PSEN1* Δe9 genotype since the correlation of *PSEN1*$^{SB::wt/wt}$ *vs.* *PSEN1*$^{SB:wt/int4}$ differential expression results with either genotype was equally low (*PSEN1*$^{UCSD::wt/wt}$ *vs.* *PSEN1*$^{UCSD::wt/\Delta e9}$: Spearman's r = 0.14, $p < 10^{-20}$; *PSEN1*$^{UCSD::wt/wt}$ *vs.* *PSEN1*$^{UCSD::\Delta e9/\Delta e9}$: Spearman's r = 0.19, $p < 10^{-20}$).

The genes which showed a consistent change in all *PSEN1* mutants were *BTNL9*, *CCDC152*, *CHRM1*, *CMTM7*, *CXCL5*, *EFCAB13*, *ESPNL*, *FAM66D*, *HMX3*, *HSPA2*, *LOC100132891*, *LOC400685*, *MEG3*, *NELL1*, *NKX2-1*, *NNAT*, *NPY*, *OTP*, *PCDHA10*, *PCDHA2*, *PCDHA6*,

PCDHGA10, *PCDHGB5*, *S100A6*, *TAF9B*, *TSPEAR*, *ZNF300* and *ZNF703*. *NPY* is known to have neuroprotective effects in mouse models of Alzheimer's neurodegeneration.[204,205] *S100A6*, an astrocytic marker, is also an interesting candidate, as it is located near a SNP associated with hippocampal volume and the distribution of S100A6 expression is altered in mouse models of Alzheimer's disease.[206]

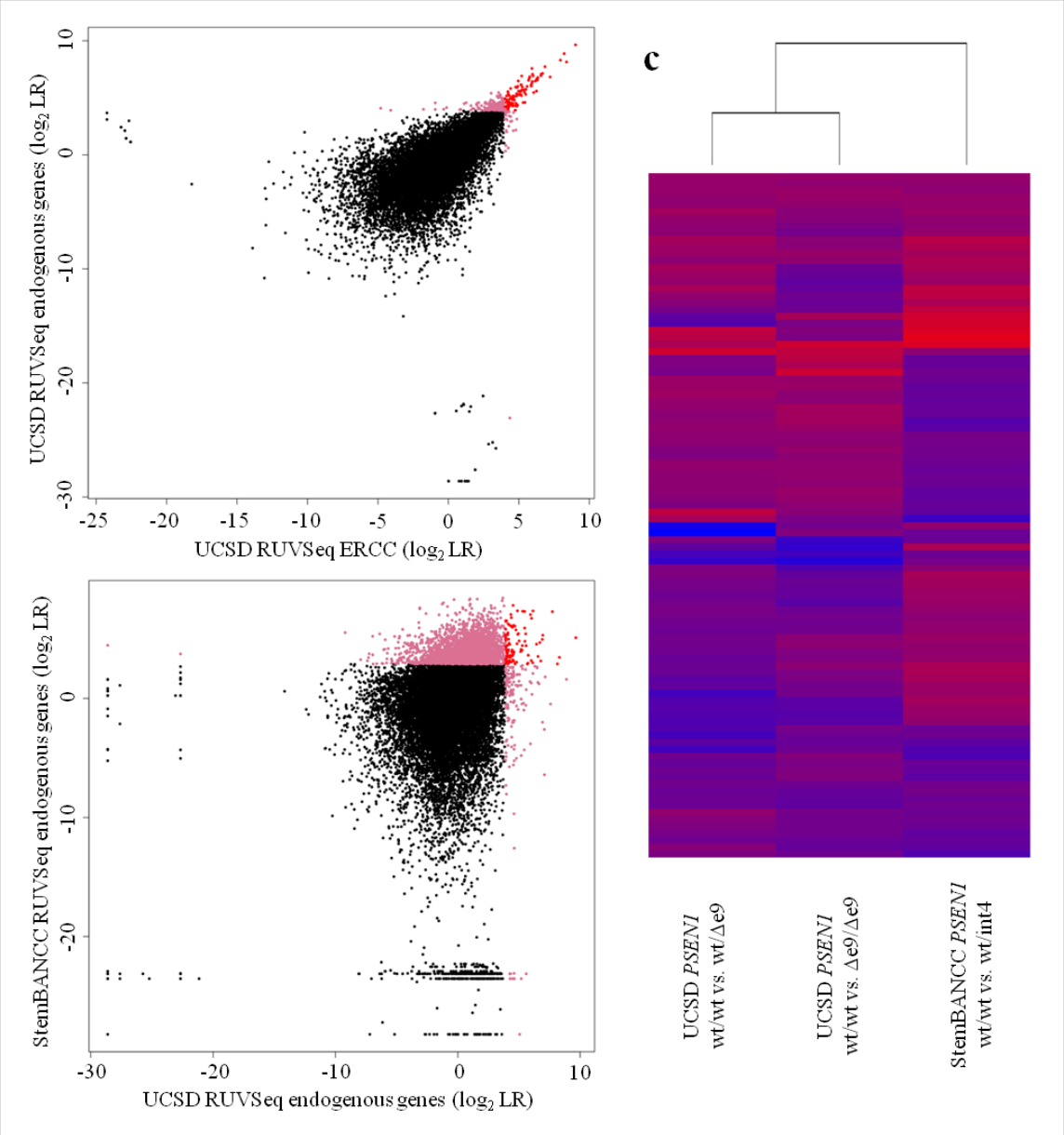

Figure 15 Comparing differential expression analysis results between UCSD and StemBANCC iPSC-derived cortical neurons. Scatter plots depicting the $\log_2$ likelihood ratio estimated by edgeR for: (**a**) UCSD *PSEN1* isogenic iPSC-derived cortical neurons adjusted for ERCC spike-ins (x-axis) or endogenous control genes (y-axis) (Spearman's r = 0.72, $p < 10^{-20}$), and (**b**) UCSD *PSEN1* isogenic (x-axis) and StemBANCC case-control (y-axis) iPSC-derived cortical neurons adjusted for endogenous control genes (Spearman's r = 0.25, $p < 10^{-20}$). Endogenous control genes were defined as those showing low variability between *PSEN1*

genotypes. Counts were adjusted using RUVSeq.[184] (**c**) A heat map showing $\log_2$ fold change for genes significant in both UCSD and StemBANCC differential expression analysis. Red = higher expression in wild-type, blue = higher expression in heterozygotes. Rows and columns are clustered by similarity.

It is possible that, despite few genes replicating between the UCSD and StemBANCC experiments, the underlying differential gene expression signature may be enriched for similar modules of an underlying gene connectivity network. To test this, I took advantage of a study that performed an integrative analysis of multiple sources of genetic and transcriptomic data in late onset Alzheimer's disease.[189] This indentified 40 modules of genes that showed greater connectivity (FDR < 0.05) in Alzheimer's disease datasets. I intersected the LR values calculated for *PSEN1* genotype in the current study with those modules to assess whether any were significantly enriched in both datasets. There was a moderate correlation between enrichment for specific modules in each dataset (Spearman's $r = 0.65$, $p < 10^{-5}$), and 13 modules were significant in both UCSD and StemBANCC analyses (**Table 5**). It is difficult to say whether these modules are enriched due to their role in Alzheimer's disease or because most are involved in differentiation pathways and may be driven by batch-to-batch or inter-line variation. I found 14 significantly differentiated genes within these modules that showed the same direction of effect in $PSEN1^{UCSD::wt/wt}$ *vs.* $PSEN1^{UCSD::wt/\Delta e9}$ and $PSEN1^{SB::wt/wt}$ and $PSEN1^{SB::wt/int4}$: *BMP6* (extracellular matrix and vasculature development), *C14orf23* (neurogenesis), *CAT* (myelination), *CMTM7* (immune function), *DLX1* (neurogenesis), *HMX1* (myelination), *HSPA2* (myelination), *MEG3* (synaptic transmission), *NKX2-2* (myelination), *NNAT* (dynein complex and extracellular matrix), *NPY* (neurogenesis), *PCDHGA10* (myelination and neuropeptide hormone), *TCERG1L* (synaptic transmission) and *VIPR2* (neuropeptide hormone).

Table 5 Enrichment of differential gene expression within gene modules of an Alzheimer's disease specific connectivity network. Module colours are those supplied by the original authors. Genes were matched for permutation analysis by expression decline. P-values were estimated from 10,000 permutations. P_c-value = Benjamini-Hochberg adjusted p-value.

Module	**Gene ontology**	**UCSD *PSEN1* isogenic**		**StemBANCC case-control**	
		Enrichment	**P_c-value**	**Enrichment**	**P_c-value**
Green 4	Mitotic cell cycle	2.11	0.006	7.35	0.0003
Magenta	Cell adhesion	2.37	0.0003	3.97	0.0003
Khaki	GABA biosynthesis	2.03	0.0003	3.61	0.0003
Tan	Extracellular matrix	2.44	0.0003	2.84	0.0003

Midnight blue	Neurogenesis	2.36	0.0003	2.51	0.0003
Chocolate	Dynein complex	2.18	0.0003	2.07	0.0003
Salmon	Synaptic transmission	2.22	0.0003	1.92	0.0003
Light green	Nerve ensheathment (myelination)	1.87	0.0003	1.98	0.0003
Cyan	Vasculature development	1.77	0.0003	1.98	0.0003
Green	Nerve ensheathment (myelination)	1.86	0.0003	1.82	0.0003
Red	Nerve ensheathment (myelination)	1.84	0.0003	1.80	0.0003
Yellow	Immune functions	1.59	0.0003	1.63	0.0003
Black	Neuropeptide hormone	1.22	0.007	1.27	0.02

3.5.7 Measuring batch-to-batch variability in iPSC-derived cortical neurons

It was possible that the low correspondence between the two *PSEN1* RNA-seq datasets could be due to the high level of batch-to-batch variability present between cortical differentiations. I attempted to explore this by borrowing a concept developed for single cell RNA-seq analysis that identifies genes with higher than expected variation between samples by using ERCC spike-ins to model technical noise (**Figure 16a**).[129] The resultant set of genes showing high variability (FDR < 0.05) between *PSEN1*$^{UCSD::wt/wt}$, *PSEN1*$^{UCSD::wt/\Delta e9}$ and *PSEN1*$^{UCSD::\Delta e9/\Delta e9}$ was strongly enriched for terms associated with GABAergic interneuron differentiation (*ARX*, *DLX2*, *DRD1*, *DLX1*, *NKX2-1* and *LHX6*), serotonergic receptor signalling (*HTR1A*, *HTR7*, *HTR2C* and *HTR3B*) and forebrain regionalisation (*WNT1*, *EMX1*, *EMX2*, *SIX3*, *NKX2-1* and *SHH*) (**Figure 16b**). This was not due to *PSEN1* genotypic difference because when I restricted the analysis only to wild-type lines, highly variable genes were enriched for similar gene ontology pathways (data not shown).

I hypothesised that the same set of highly variable genes may drive batch-to-batch variability in *PSEN1*$^{SB::wt/wt}$ and *PSEN1*$^{SB:wt/int4}$. This proved to be the case, with highly variable genes identified from the UCSD *PSEN1* isogenic system showing more variability amongst StemBANCC iPSC-derived cortical neurons than 1,000 random sets of genes matched for expression level (median squared coefficient of variation (CV^2) = 0.38, permuted CV^2 = 0.21, $p < 0.001$). The same was true when restricting this analysis to highly variable genes belonging to gene ontology terms associated with neuronal fate or regionalisation within the nervous system (median CV^2 = 0.78,

permuted $CV^2 = 0.27$, $p < 0.001$). This suggested that genes involved in neuronal differentiation and patterning of the nervous system were important in driving batch-to-batch variability in iPSC-derived cortical neurons.

One theoretic advantage in using isogenic stem cell systems is that variability driven by the genetic background of the donor outside the gene of interest should be minimal. I attempted to test this by randomising experimental labels (*i.e.* whether a sample was derived from UCSD or StemBANCC iPSCs) in order to test whether the variability in gene expression (estimated as the biological coefficient of variation (BCV) calculated by edgeR on TMM adjusted read counts) was significantly lower for within experiment comparisons. I found that UCSD but not StemBANCC samples had significantly lower within experiment BCV than expected by chance (1,000 permutations - UCSD: BCV = 0.33, permuted BCV = 0.48, $p < 0.001$; StemBANCC: BCV = 0.39, permuted BCV = 0.47, $p = 0.08$). However, the difference in BCV between UCSD and StemBANCC samples was not significant (ΔBCV = -0.06, permuted ΔBCV = 0.01, $p = 0.30$; **Figure 16c**). In fact, when including all samples from the first batch of UCSD *PSEN1* isogenic differentiations, the trend for lower BCV in UCSD samples than StemBANCC disappeared (ΔBCV = 0.02, permuted ΔBCV = 0.01, $p = 0.80$), suggesting that the increase in transcriptomic noise from conducting differentiations at different times may overwhelm any reduction in variability due to an isogenic background.

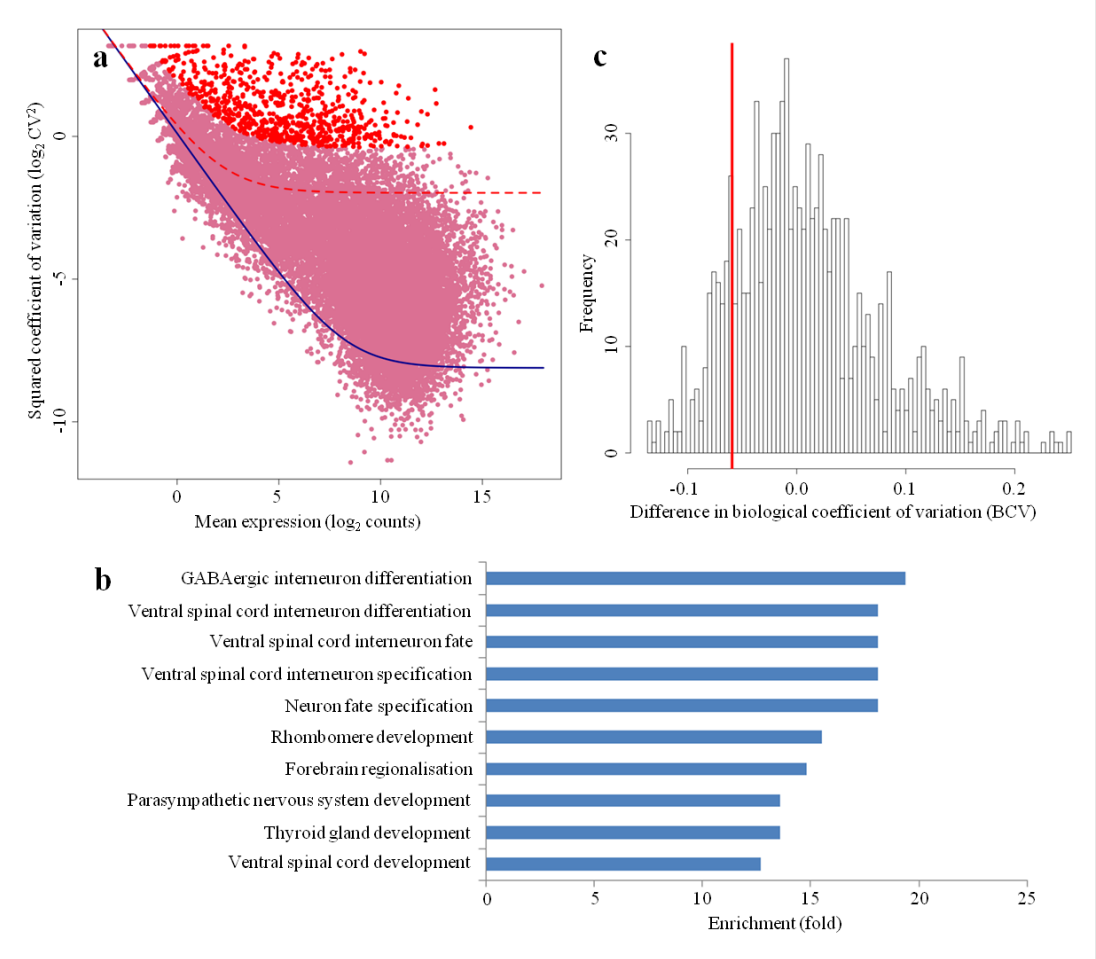

Figure 16 An assessment of variability in iPSC-derived cortical neurons. (**a**) Modelling technical variation to identify highly variable genes (FDR< 0.05, red). The blue line indicates the technical fit from ERCC spike-in expression and the red dashed line indicates the 95% confidence interval around this fit. (**b**) Gene ontology term enrichment within the set of highly variable genes. (**c**) A histogram showing the difference in biological coefficient of variation (BCV) between UCSD *PSEN1* isogenic and StemBANCC iPSC-derived cortical neurons calculated on 1,000 permuted datasets. The vertical red line indicates the true value of ΔBCV. Negative values indicate a lower BCV in UCSD *PSEN1* isogenic samples.

3.6 Discussion

In this chapter I have used bulk RNA-seq data generated on control and *PSEN1* mutant iPSC-derived cortical neurons to investigate batch-to-batch variability and the prospects for using this *in vitro* system to interrogate Alzheimer's disease-related biology. I was able to demonstrate that each iPSC line generated cells expressing multiple markers of cortical identity. A division by *PSEN1* genotype in clustering analysis was clearly seen for StemBANCC case-control iPSC-derived cortical neurons but was not observed in a *PSEN1* gene-edited isogenic series

differentiated into cortical neurons. Multiple genes that were differentially expressed by *PSEN1* genotype were detected in both case-control and isogenic neurons but there was very little overlap between the results of these experiments, suggesting that γ-secretase deficiency alone was unlikely to explain the observed effects. I found no significant difference in gene expression variability between cortical neurons derived from either the isogenic *PSEN1* lines or StemBANCC case-control lines.

In line with the results of **Chapter 2**, I found that cortical neurons derived from wild-type or *PSEN1* mutant iPSCs strongly expressed markers associated with cortical identity. Benchmarking iPSC-derived neurons against previously validated hES corticogenesis can be important, as cortical differentiation from iPSC has been found to be more variable than from hES.[207] Unsurprisingly, there were defects in Aβ processing associated with *PSEN1* mutations. This has been robustly observed in previous studies using various mutations associated with Alzheimer's disease and can be observed early in neural differentiation.[85,95]

I was able to identify only 129 genes that were differentially expressed by genotype in UCSD *PSEN1* isogenic iPSC-derived cortical neurons, including some with near SNPs associated with cognitive decline, such as *FLT1* and *PNPO*.[191,192,194,196] In contrast, there were 2,597 genes differentially expressed between *PSEN1*$^{SB::wt/wt}$ and *PSEN1*$^{SB:wt/int4}$ iPSC-derived cortical neurons, including well-known Alzheimer's disease-associated genes such as *APOE* and *CLU*.[208] This suggests that some transcriptomic aspects of Alzheimer's disease were recapitulated in the StemBANCC samples but not the UCSD *PSEN1* isogenic samples. This is despite the fact that many potential key drivers of Alzheimer's disease pathology, such as microglia and endothelium, were absent from the cultures.[86,209]

Only 45 differentially expressed genes demonstrated a consistent direction of effect in both datasets and only 28 in all *PSEN1* genotypes relative to wild-type. Lack of reproducibility has been a recurrent issue in applying transcriptomic methods to Alzheimer's disease *in vivo* or *in vitro*.[87,92,93,166] Analysis of the *PSEN1* isogenic iPSC-derived cortical neurons using shotgun proteomics identified two potential candidates, MACROD1 and NAMPT, which were

differentially expressed by *PSEN1* genotype, but showed no overlap with the RNA-seq results. There are several potential reasons why so few of the differentially expressed genes were robustly detected between experiments.

Firstly, it is possible that there was too much batch-to-batch variability in the UCSD *PSEN1* isogenic series to permit the identification of truly differentially expressed genes. It is unlikely that batch-to-batch variability masked genuinely differentially expressed genes specifically in the UCSD samples because there was no significant difference in the degree of transcriptomic variability between UCSD and StemBANCC iPSC-derived cortical neurons.

Secondly, the two types of *PSEN1* mutation could cause different forms of neurodegeneration. This is potentially supported by the observation that *PSEN1* intron 4 mutations cause Alzheimer's disease in isolation, whereas *PSEN1* Δe9 mutations are additionally associated with spastic paresis.[175,177] These phenotypic differences are mirrored by differences in the underlying neuropathological changes.[210,211] *PSEN1* intron 4 mutations result in classical Alzheimer's pathology with abundant neuritic plaques and neurofibrillary tangles, particularly in the hippocampus. In contrast, *PSEN1* Δe9 mutations have cotton wool spots, cerebral amyloid angiopathy and gliosis as well as the other classical neuropathological findings seen in Alzheimer's disease. However, one would have expected that more severe pathology would have led to the identification of *more* differentially expressed genes in the UCSD samples than the StemBANCC samples whereas the reverse was in fact true.

A third possibility is that the UCSD *PSEN1* isogenic iPSC maintained a "memory" of their wild-type status. This could arise through epigenetic modifications, such as DNA methylation, which were maintained through sequential cell divisions. There is some evidence that a small proportion of differentially methylated regions (DMRs) in iPSC do maintain the pattern observed in the donor fibroblasts, at least out to 17 passages in culture.[212] However most DMRs were near genes were associated with lineage priming rather than genes likely to be directly involved in Alzheimer's pathogenesis. Given that in both the UCSD and StemBANCC experiments differentially expressed genes were enriched for gene ontology pathways associated with

neuronal fate and differentiation, it is still possible that this could have contributed to differences observed between iPSC lines. This is particularly true for *PSEN1*$^{SB::wt/wt}$ and *PSEN1*$^{SB::wt/int4}$, in which the iPSCs were derived from different donors. Additionally, prior to and during reprogramming, *PSEN1*$^{UCSD::wt/wt}$, *PSEN1*$^{UCSD::wt/\Delta e9}$, *PSEN1*$^{UCSD::\Delta e9/\Delta e9}$ and *PSEN1*$^{SB::wt/wt}$ cells were exposed to normal levels of Aβ species, whereas *PSEN1*$^{SB::wt/int4}$ cells were exposed to abnormal $A\beta_{1-42}/A\beta_{1-40}$ ratios that could have altered the behaviour of fibroblasts and iPSCs. Certainly $A\beta_{1-40}$ is known to promote increased cell division in cultured fibroblasts and this could potentially lead to differences in reprogramming outcomes.[213,214]

If this final possibility is correct, it would have wide-reaching implications for iPSC-based research, since an implicit assumption of using iPSC-derived cells in causal modelling of diseases is that genetic differences are the main drivers of variation between lines. Should this variation instead turn out to be a function both of genetic *and* epigenetic differences driven by donors, interpreting the results of iPSC-based research would become much more challenging. It may also imply that care should be taken when generating isogenic iPSC series to ensure that exposure of cells to different molecular processes during reprogramming does not bias the experimental outcome.

Batch-to-batch variability is a key issue in iPSC-derived systems. The use of ERCC spike-ins in this study allowed me to identify a set of highly variable genes in the UCSD *PSEN1* isogenic samples.[129] This was enriched for pathways involved with neuronal fate and differentiation, suggesting that cortical differentiation protocols produce highly heterogeneous cellular populations in the final culture. The same set of genes was also highly variable in the StemBANCC differentiations, suggesting that this may be a common feature of cortical differentiation protocols. It is possible that alterations in differentiation tendancies may amplify variation between cultures; for example, the presence of GABAergic interneurons is known to alter hippocampal neurogenesis and therefore any bias towards the production of interneurons may in turn alter the behaviour of neural progenitors in the same culture.[215] There was a tendency for batch-to-batch variability to be lower in the isogenic series, although this did not reach

significance and this effect appeared to be outweighed by an increased variability if differentiations were not conducted in parallel. This variability is likely to limit the use of bulk transcriptomic and proteomic methods in these cultures and may require the application of single cell functional genomics methods in order to identify robust disease-related signatures.[151]

3.7 Conclusion

It was possible to identify transcriptomics signatures associated with *PSEN1* genotype in iPSC-derived cortical neurons. Some of the genes identified are thought to be key players in Alzheimer's disease pathogenesis. However, most differentially expressed genes did not replicate between a case-control comparison and an isogenic gene-edited series. This may have been due to the amount of noise inherent to cortical differentiation, due to differences between particular *PSEN1* mutations or because iPSCs maintain a memory of their original genotype. I also found that genes involved in neuronal differentiation were highly variable between different batches of iPSC-derived cortical neurons and that the degree of transcriptomic noise was high even in an isogenic iPSC system. My findings raise important issues in the design and interpretation of iPSC-based experiments aiming to implicate genetic variants in the aetiology or pathogenesis of disease.

3.8 Acknowledgements

Satyan Chintawar assisted in the expansion of the UCSD and StemBANCC iPSC lines used in this study. Bryan Adriaanse assisted with some of the immunofluroescence staining. James Smith performed Aβ quantification. Anna Baud at UCL generated the mass spectrometry data. The UCSD isogenic lines were a kind gift from Grace Woodruff.

Chapter 4 Functional genomics analysis of neurological disease

4.1 Abstract

Dense genotyping approaches have revealed much about the genetic architecture both of gene expression and disease susceptibility. However, assigning causality to genetic variants associated with a transcriptomic or phenotypic trait presents a far greater challenge. The development of epigenomic resources by ENCODE, the Epigenomic Roadmap and others has led to strategies that seek to infer the likely functional variants underlying these genome-wide association signals. It is known, for example, that such variants tend to be located within areas of open chromatin, as detected by techniques such as DNase-seq and FAIRE-seq. I aimed to assess what proportion of variants associated with phenotypic or transcriptomic traits in human brain are located within transcription factor binding sites. Wellington Footprints was used to infer transcription factor footprints from existing DNase-seq data derived from central nervous system tissues with high spatial resolution. This dataset was then employed to assess the likely contribution of altered transcription factor binding to both expression quantitative trait loci (eQTL) and genome-wide association study (GWAS) signals. Surprisingly, I show that most haplotypes associated with GWAS or eQTL phenotypes are located outside of DNase-seq footprints. This could imply that DNase-seq footprinting is too insensitive an approach to identify a large proportion of true transcription factor binding sites. Importantly, this suggests that prioritising variants for genome engineering studies to establish causality will continue to be frustrated by an inability of footprinting to identify the causative variant within a haplotype.

4.2 Hypotheses

4.2.1 Variants associated with eQTL or GWAS traits are enriched within DNase-seq footprints in brain tissue.

4.2.2 Many eQTL or GWAS haplotypes can be explained by inferred interference with TF binding through DNase-seq footprints in brain tissue.

4.3 Introduction

Chapter 3 interrogated a stem cell-based model for disease-related transcriptomic signatures arising from high penetrance mutations in *PSEN1*. However, most cases of Alzheimer's disease and neurological diseases in general are not believed to be caused by high penetrance, rare genetic mutations but are rather likely to be caused by the summative effect of multiple common genetic variants.[216–218] In this chapter I use functional genomics data from primary brain tissue in an attempt to understand how some of these common risk variants may be exerting their effects.

Genomic variation is a major exploratory variable for many phenotypes. These include traits and diseases, for which large volumes of genotyping data have become available.[219] They also include expression quantitative trait loci (eQTLs), which are genomic variants correlated with gene expression levels. A growing catalogue of eQTLs is being identified with the availability of genotyping datasets associated with whole transcriptome expression data. Recently, a wealth of eQTL data became available for tissues of the human central nervous system.[55,220–223]

Genome-wide association studies (GWAS) in large cohorts, performed on many of the prevalent brain disorders, have revealed new pathoetiological mechanisms. Nevertheless, it remains a substantial challenge to explain mechanistically how GWAS single nucleotide variants (SNVs) exert their effects. This is because over 95% of the sentinel SNVs of an associated locus fall in non-coding regions and any one is unlikely to be causal.[80,81] Whilst it is well-established that GWAS and eQTL variants occur preferentially within regions of open chromatin,[80,224] the precise functional consequences of these variants are less clear.

One or more variants in strong linkage disequilibrium (LD) with sentinel GWAS or eQTL variants are expected to be causal, in part by altering transcription factor (TF) binding and/or chromatin architecture.[225] However, variation within classical TF motifs was found to poorly predict changes in binding [226] and flanking sequences located far from the actual TF binding site also appear to be important.[227] Even when TF binding is altered, the level of expression from adjacent genes is often not substantially affected.[228] These observations imply the inadequacy of a simple model, that of altered gene expression resulting from distal DNA mutations lying directly within transcription factor binding sites.

The most direct methods for assessing TF binding are based on chromatin immunoprecipitation (e.g. ChIP-seq). These approaches are limited in that each TF must be interrogated individually with specific antibodies and can often be of relatively poor spatial resolution, although more recently developed approaches, such as ChIP-exo and ChIP-nexus, may improve upon this.[22,23] Methods of interrogating chromatin structure for TF binding sites collectively without specifying a particular TF include DNase-seq, FAIRE-seq and ATAC-seq.[8,229] Irreproducibility discovery rate (IDR) analysis is a powerful approach to reproducibly identifying regions with high DNase-seq accessibility that represent truly open chromatin.[230,231] Although DNase-seq and similar methods provide relatively coarse spatial resolution, nucleotides contained within TF-bound sites are relatively well protected from DNase digestion. This protection from cleavage produces 'footprints' which can aid in narrowing down a true TF binding site within a wider DNase hypersensitivity site.[28] Of all methods that identify DNase-seq footprints, Wellington Footprints has been proposed to provide the best estimates of true binding sites and also, unlike some other DNase-seq footprinting methods, is not reliant on the presence of TF motifs.[29,232] This approach relies on the imbalance in strand-specific alignment of DNase-seq reads. Limitations of DNase-seq footprinting include its lower power to identify TF binding that is dependent on short segments of non-colinear sequence.[233] Footprinting also captures a cross-sectional sample of TF binding and so may miss TF binding sites that are highly dynamic or induced by a particular stimulus.

I was interested in whether GWAS and eQTL associations in central nervous system tissues can best be explained by mutations directly within well-defined TF binding sites. In order to investigate this, I processed available DNase-seq datasets from the ENCODE and Epigenomic Roadmap projects to generate DNase-seq footprints at high spatial resolution. Combining this footprint data with reported brain eQTLs and brain-related GWAS signals allowed me to estimate the proportion of haplotypes that disrupt TF binding.

4.4 Methods

4.4.1 DNase-seq analysis

DNase-seq hg19 aligned reads were downloaded from the ENCODE (cerebellum, frontal cerebrum and frontal cortex) and Epigenomic Roadmap (fetal brain) projects for footprints from primary brain tissue.[231,234,235] For other tissue footprints, I downloaded multiple files of aligned reads from the ENCODE (A549 cells, aortic smooth muscle, Caco2 cells, Ecc-1 cells, Gc B-cells, H1-derived mesenchymal stem cells, H1-derived neuronal progenitor cultured cells, H1 cells, heart, Helas3 cells, hepatocarcinoma, hepatocytes, Ishikawa cells, K562 cells, keratinocytes, lung fibroblasts, medulloblastoma, monocytes, naive B-cells, neuroblastoma, olfactory neurospheres, renal glomerular endothelium, retinal pigment endothelium, skeletal muscle fibroblasts and urothelium) and Epigenomic Roadmap (fetal heart, fetal arm muscle and fetal abdominal skin) projects. DHS were called using F-seq with the arguments "-t 0 -of npf -f 0".[236] Irreproducibility discovery rate (IDR) analysis was used to assess whether it was appropriate to pool replicates for further analysis as described in [230]. In order to present a permissive set of DHS to the footprinting analysis pipeline, I used an IDR threshold of 0.05. Footprints within DHS were identified using Wellington Footprints (version 0.2.0) with the arguments "-p 8 -fp 11,22,2 -fdr 0.01" or "-p 8 -fp 6,40,2 -fdr 0.01" for the broader footprints.[232] Footprints were removed if these intersected ENCODE blacklisted regions.[231,237] I used a filtering method to remove likely artefacts that skewed the mean DNase profile within 100 base-pairs of the centre of each footprint by > 50%. Footprints used for analysis were restricted to autosomes.

4.4.2 GWAS and eQTL variants

GWAS variants were downloaded from the GWAS Catalog and classified into brain-related or adult neurological disorder-related as per [223]. GWAS variants were only included in downstream analysis if the associated p-value was $< 5 \times 10^{-8}$. eQTLs were obtained from a number of different studies [55,220–223]. *Trans*-eQTLs were not considered for further analysis. *Cis*-eQTLs reported as significant were pooled together for the combined analysis and also analysed individually. Haplotypes were imputed from 379 European 1000 Genomes phased haplotypes using vcftools with the arguments "--gzvcf in.file --hap-r2-positions snp.file --ld-window-bp 10000000 --min-r2 0.5".[238,239]

4.4.3 Functional annotation

BEDtools was used to intersect GWAS and eQTL variants at different r^2 thresholds and then custom Rscripts were used to analyse the proportion of haplotype blocks intersecting different features. Statistical analysis was conducted using the Genomic Association Tester (GAT), using 10,000 randomisations and a workspace based upon where annotations could fall (i.e. footprints were analysed relative to F-seq-identified DHS as by definition these could not be located elsewhere).[240] When the background chromosomes were used as a workspace, these were limited to autosomal chromosome arms minus known assembly gaps and blacklisted regions. When permuting one list of SNVs against another, the 1000 genome SNVs were used as a workspace. GC content was corrected for using 1Mb quintile isochores. Empirically determined standard deviations were plotted for all shown fold enrichment values. Nucleotide conservation scores (Phylo46way) were downloaded from ENCODE.[231] Motifs were identified using FIMO to search for JASPAR vertebrate motifs with a p-value threshold of 10^{-4}, repeats masked and a 1st order markov background on both the reference genome and the genome edited to contain all alternate single nucleotide variants.[241,242] MEDLINE was manually searched for each TF enriched within footprints to identify studies supporting a role in brain development or function. FANTOM5 permissive enhancers were downloaded directly from the FANTOM5 website.[243] Coverage histograms were obtained using Homer.[244]

4.5 Results

4.5.1 Functional validation of DNase-seq footprinting

DNase hypersensitivity sites (DHS) are typically several hundred base pairs long and encompass several predicted TF binding sites. I used FSeq to identify DHS within DNase-seq datasets using the irreproducibility discovery rate used by the ENCODE project.[230] Wellington footprints allows high precision and high confidence identification of true TF sites within a DHS by scanning within it for a region of DNase protected sequence.[232] 17,670 - 40,773 footprints between 11 - 22 bases in length were called by Wellington footprints for each of 4 brain DNase-seq datasets at FDR < 0.01. 20,468 (23.7%) of pooled brain DHS were found to contain at least one detectable footprint, a similar proportion to that found for the K562 cancer cell line from the original methods publication.[232] Footprints covered 0.6% of the total nucleotides underlying brain DHS. I used the Genomic Association Tester (GAT) to evaluate the statistical significance of footprints for genomic features that might be indicative of functional importance.[240] DNase-seq footprints were at least 2-fold enriched over DHS within regions upstream of genes (**Figure 17a**). Footprints also showed a high degree of central nucleotide conservation across mammalian evolution (**Figure 17b**).

FIMO-identified TF sequence motifs were enriched centrally in the footprints (1.38-fold, $p < 0.0001$; **Figure 17c**), as expected. There was also significant enrichment within footprints for FANTOM5-annotated enhancers and this, further, was significantly higher than the corresponding enrichment within the DNase-seq hypersensitivity peaks as a whole (1.78-fold, $p < 0.0001$).[243]

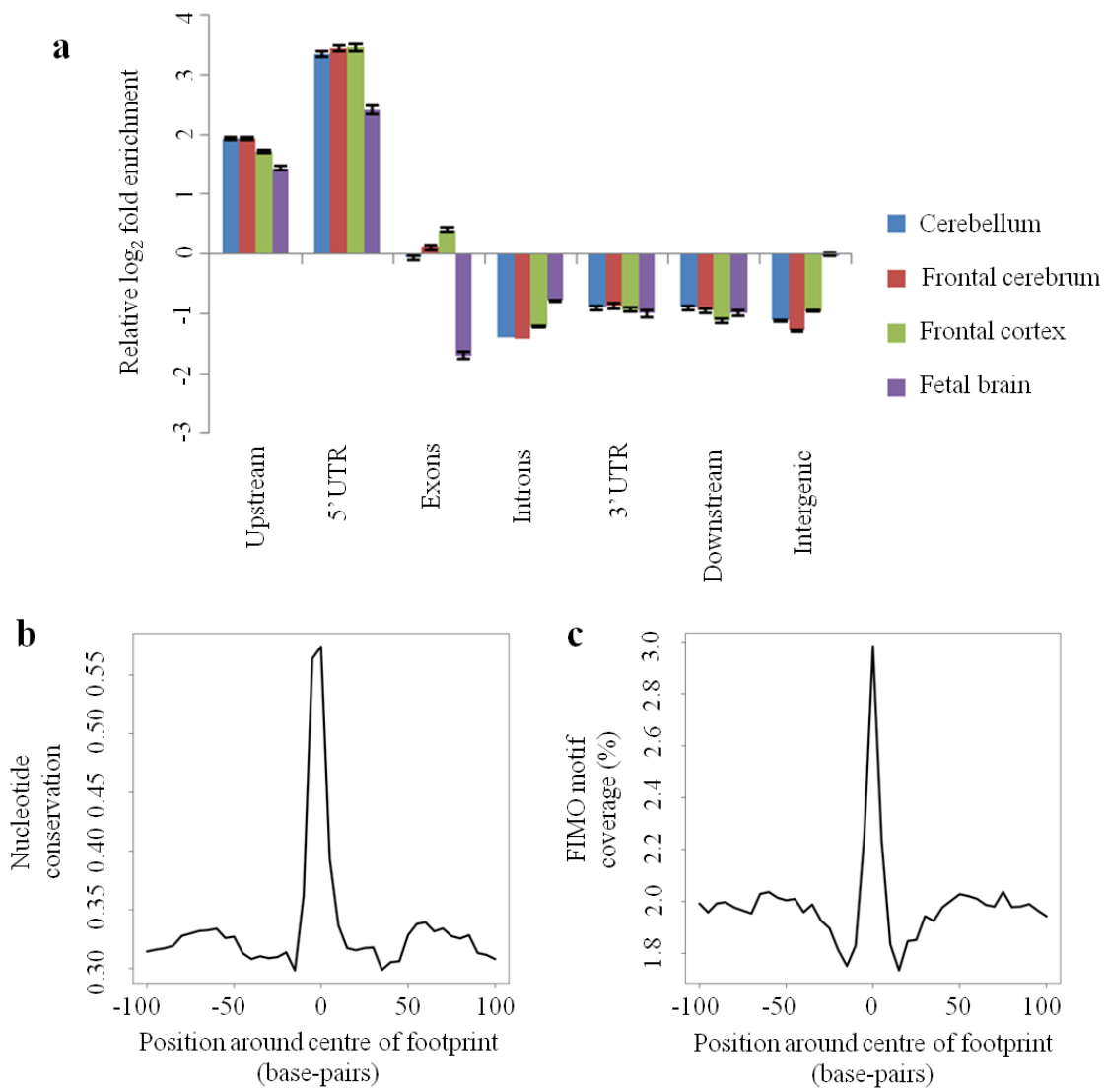

Figure 17 Functional annotation of DNase-seq footprints. (**a**) shows relative enrichment of footprint-containing DHS to DHS without footprints for different metagene regions. Error bars show the standard deviation of $\log_2$ fold enrichment. Histograms of (**b**) FIMO motif coverage and (**c**) nucleotide conservation score within mammalian species (PhyloP46way) around all brain footprints combined.

4.5.2 Integrating brain eQTLs with DNase-seq footprints

Brain *cis*-eQTL haplotypes are significantly enriched within DHS identified in brain tissue (67.7% of brain eQTLs at $r^2 > 0.5$ and 50.7% at $r^2 > 0.8$; **Figure 18a**). I defined eQTL haplotypes as those *cis*-SNVs in LD at $r^2 > 0.5$. The degree of LD between two alleles, A and B, is given by:

$$r^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A(1-p_A)p_B(1-p_B)}$$

It was then possible to annotate an eQTL-containing haplotype block with genomic features, such as DHS and DNase-seq footprints by intersecting those with the SNVs constituting the haplotype.

eQTLs typically contain multiple associated SNVs that all lie in strong LD. Most often it has not been possible to identify from among them the causal eQTL SNV, or indeed multiple causal SNVs.[245] Any SNV lying directly within the footprint may provide an accurate prediction of the causal SNV. This approach has the clear *caveat* that it will disregard sequence variation which, despite lying outside of the direct TF binding site, impacts on TF binding affinity.[226]

Brain eQTLs were significantly enriched (1.65-fold, $p < 0.0001$) in brain tissue DNase-seq footprints ($0.5 \leq r^2 \leq 1$) in excess of their enrichment within DHS overall (**Figure 18b**). Initially I restricted the analysis to only eQTLs and DNase-seq footprints that were matched by tissue. An eQTL haplotype was defined as underlying a DNase-seq footprint if at least one SNV within its haplotype at a particular LD threshold intersected a DNase-seq footprint. I then calculated the overall proportion ($\prod$) of tissue-matched eQTL SNVs that could be accounted for by DNase-seq footprints. This was defined as the proportion of haplotypes where at least one SNV intersected a DNase-seq footprint and was present within a haplotype at a given r^2 cut-off. This quantity reached a maximum of only $\prod = 16.4\%$ at a lax LD threshold of $r^2 > 0.5$ and fell to 9.5% at an LD threshold of $r^2 > 0.8$. When including data from all brain regions, the figure dropped markedly ($\prod = 9.6\%$ at $r^2 > 0.5$ and $\prod = 5.2\%$ at $r^2 > 0.8$; **Figure 18b**). 63 TF motifs showed significant enrichment for brain eQTLs at $r^2 > 0.8$, 48 (76%) of which are known to have effects on brain function (**Figure 19**).

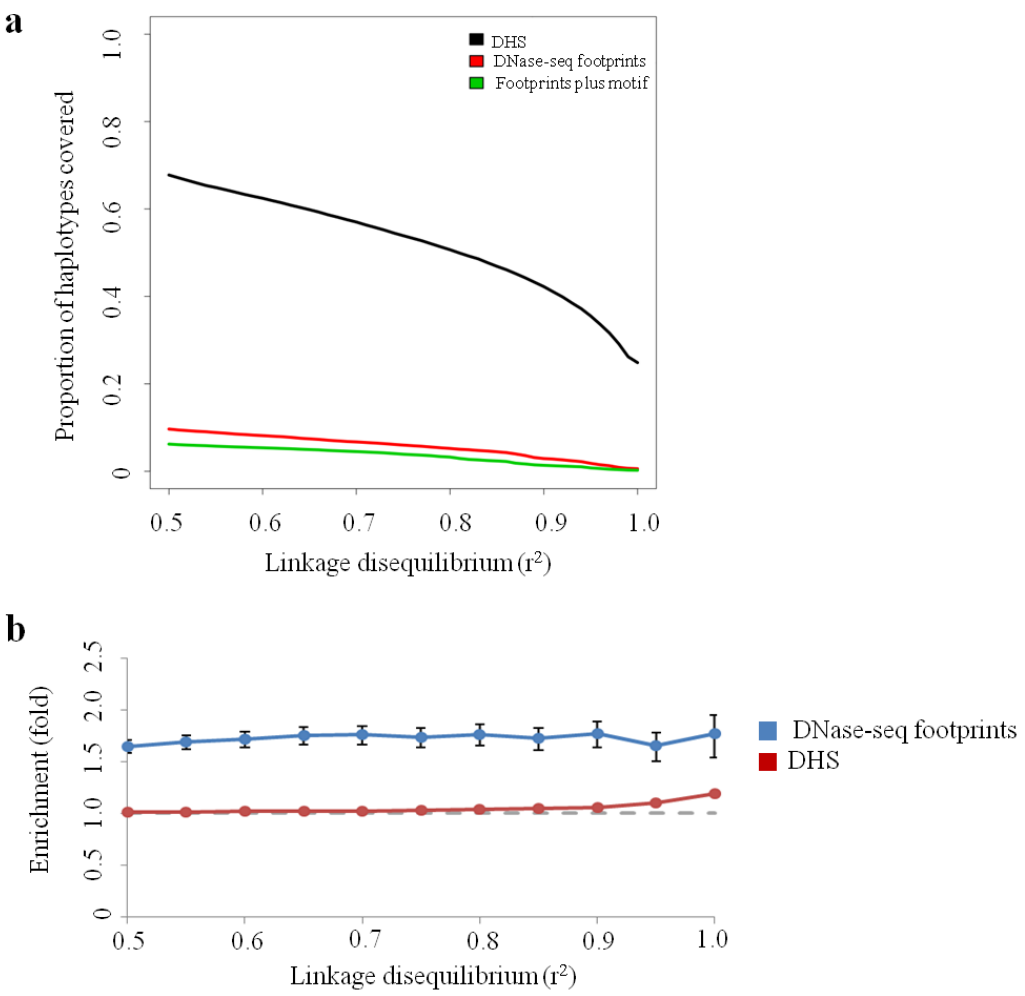

Figure 18 Brain eQTLs and DNase-seq footprints. (**a**) The proportion of brain eQTL haplotypes accounted for by either DHS (black line), footprints (red line) or footprints containing a FIMO-identified motif (green line). (**b**) The enrichment of eQTL haplotypes within brain footprints relative to brain DHS (blue) and within brain DHS relative to all autosomal chromosome arms (red). The dashed grey line indicates the value corresponding to no enrichment.

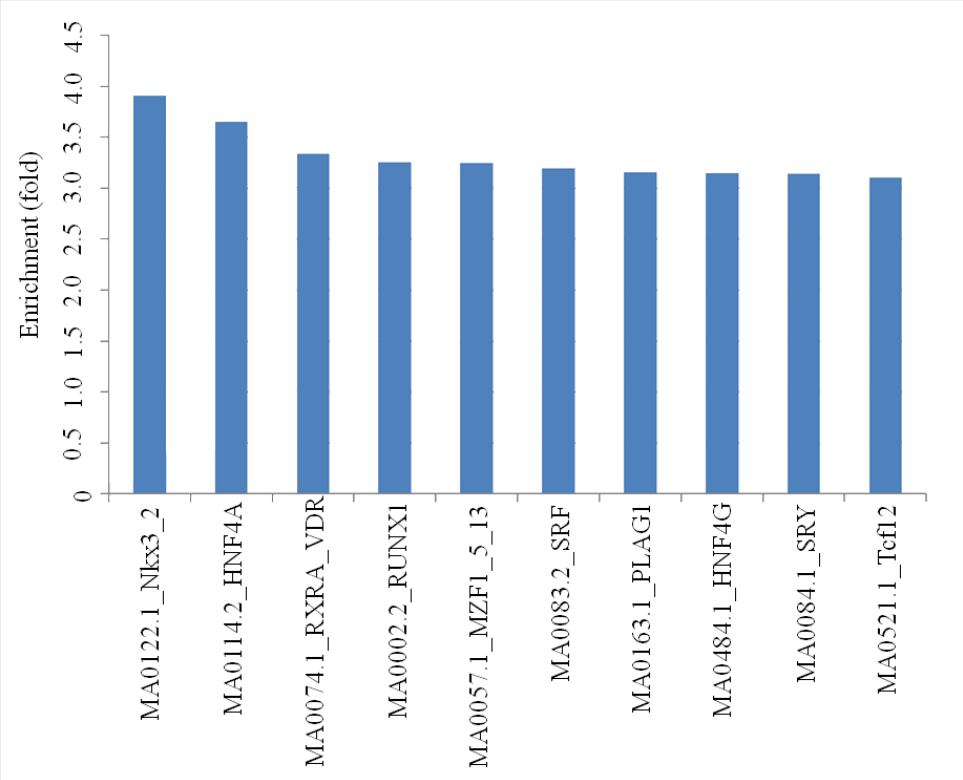

Figure 19 Enrichment of brain eQTLs within specific transcription factor motifs. All are significant at $q < 0.05$.

The precise footprint size chosen in this analysis might have had an undue influence on these results. Nevertheless, I found that it had only a modest effect on the proportion of eQTL SNVs accounted for by DNase-seq footprints ($\prod = 13.5\%$ at $r^2 > 0.5$, $\prod = 7.0\%$ at $r^2 > 0.8$). I also tested how sensitive my findings were to the significance threshold used to detect DNase-seq footprints. The previous results were generated using the p-value threshold reported by the original authors ($p < 10^{-20}$) to filter the input prior to FDR randomization.[232] However, because it is possible that this was too strict a threshold I reduced the threshold to $p < 0.01$. Even at this lax significance threshold $\prod$ remained low ($\prod = 25.1\%$ at $r^2 > 0.5$; $\prod = 13.2\%$ at $r^2 > 0.8$).

A small minority of SNVs may well be conferring their effects on target genes by interfering with TF binding. I generated a list of candidate SNVs in strong LD ($r^2 > 0.8$) with brain eQTLs also intersecting a DNase-seq footprint and TF motif. An illustrative example is an eQTL associated with the expression of *ROBO2* (**Figure 20**). *ROBO2* encodes a transmembrane receptor which is involved in axonal guidance within the central nervous system.[246] The lead eQTL SNP

(rs1447850) underlies a DNase-seq footprint found in fetal brain and also intersects motifs for *TAL1* and other TFs. *TAL1* is known to have a role in neuronal development.[247]

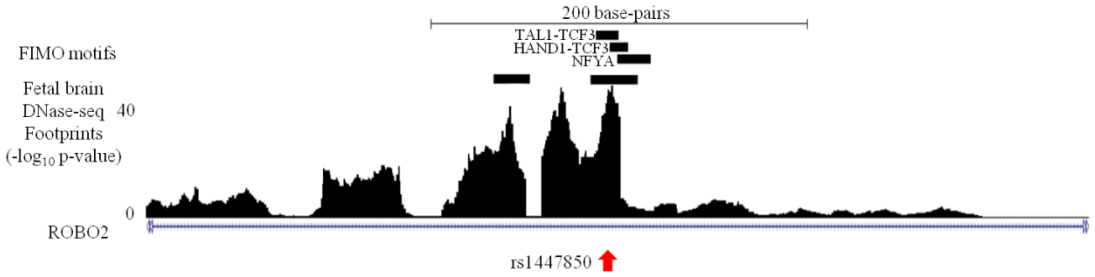

Figure 20 A brain eQTL SNV falling within a DNase-seq footprint and transcription factor recognition motif. Rs1447850 (red arrow) is significantly associated with expression of *ROBO2*. The location of motifs intersecting both the DNase-seq footprint and eQTL-associated SNV are indicated. The figure depicts chromosome 3 between positions 77,580,019 and 77,580,519.

4.5.3 Integrating brain-related GWAS SNVs with DNase-seq footprints

GWAS identify sentinel SNVs which are significantly associated with a phenotypic trait. Since GWAS SNVs rarely disrupt protein-coding regions, they are expected to alter gene expression regulation.[80] I therefore sought to identify whether brain-related GWAS (*e.g.* Alzheimer's disease, schizophrenia, *etc.*, n = 6,552) variants may be exerting their effect *via* a similar process to eQTLs. I found significant enrichment both of GWAS haplotypes within DHS and of GWAS SNVs within eQTL blocks (**Figure 21**). However, in contrast to brain eQTLs, there was no overall significant enrichment of GWAS haplotype blocks overlapping DNase footprints when compared to a background of DHS, even when reducing the LD threshold to relatively permissive levels (**Figure 21**). At the most relaxed LD threshold, there was nominal enrichment of GWAS-associated haplotypes within DNase-seq footprints but this was not significant after correction for multiple hypothesis testing (1.45-fold, p = 0.02, q = 0.21 at $r^2 > 0.5$).

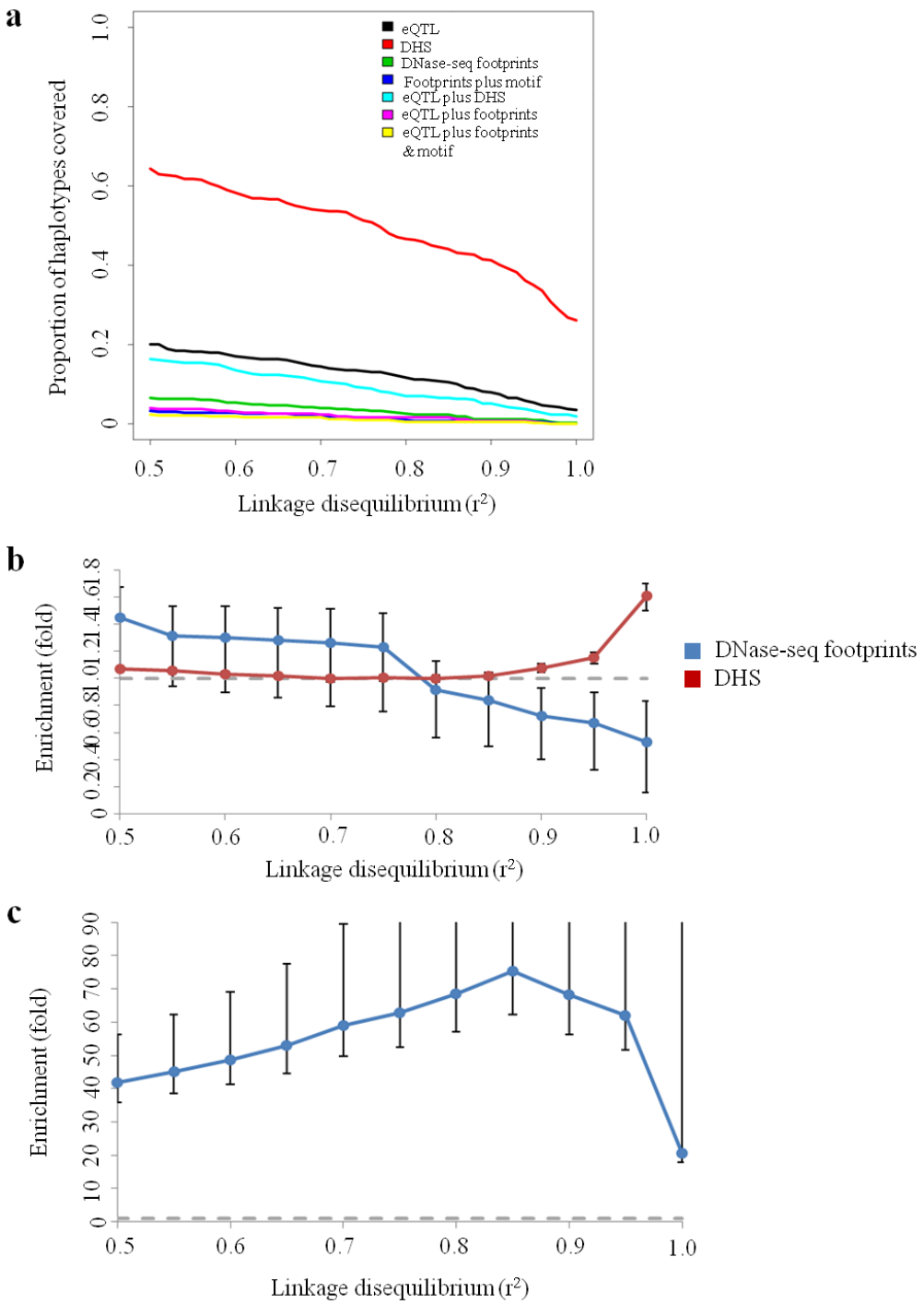

Figure 21 Brain-related GWAS haplotypes, eQTLs and DNase-seq footprints. (**a**) The proportion of brain GWAS haplotypes accounted for by various features. (**b**) The enrichment of GWAS haplotypes within brain footprints relative to brain DHS (blue) and within brain DHS relative to all autosomal chromosome arms (red). (**c**) GWAS haplotype enrichment for eQTLs. The dashed grey line indicates the value corresponding to no enrichment.

Despite the lack of globally significant enrichment for DNase-seq footprints, I identified a small minority of candidate variants (3.3% at $r^2 > 0.5$; 0.9% at $r^2 > 0.8$), in which a strongly linked SNV was contained within both a brain DNase-seq footprint and a FIMO-identified TF binding motif.

One such region is shown in **Figure 22**, in which a GWAS variant associated with susceptibility to schizophrenia lies in strong LD with an eQTL variant, both of which are in strong LD with a SNV located within a footprint containing a motif for ZFX. One possible link between *DCAF6*, associated with this LD block by eQTL, and susceptibility to psychiatric disorders is that its protein binds *NR3C1*, of which changes in methylation are linked with childhood abuse.[248] However, further work will be needed to explore this hypothesis.

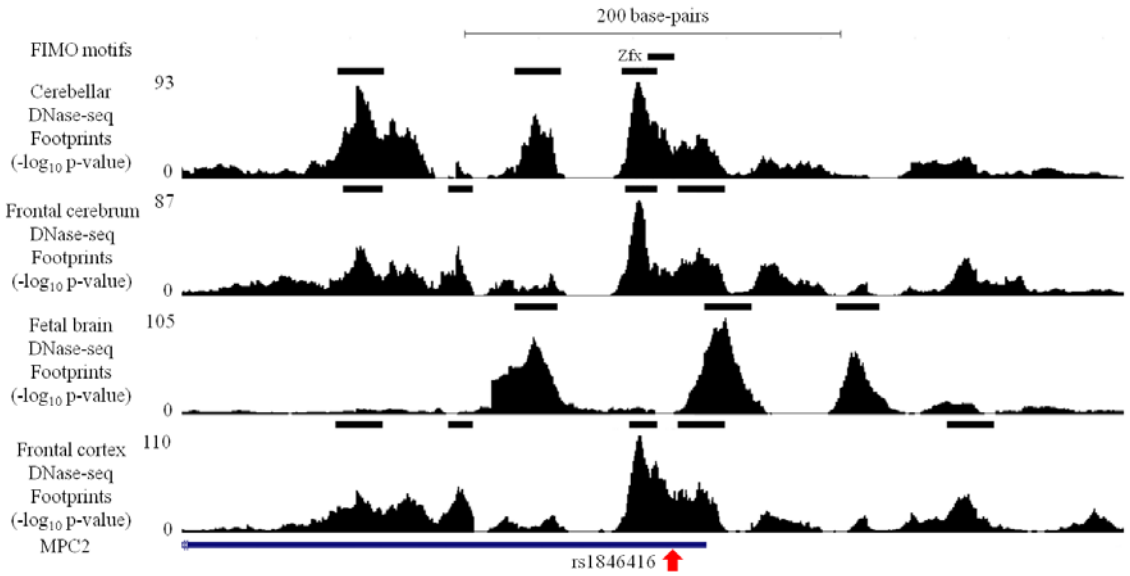

Figure 22 An example of a SNV within a schizophrenia GWAS haplotype falling within a DNase-seq footprint and transcription factor recognition motif. Rs1846416 (red arrow) is in strong linkage disequilibrium with a schizophrenia GWAS index SNV (rs10489202). The location of the Zfx motif intersecting both the DNase-seq footprint and GWAS-associated SNV are indicated. This figure depicts chromosome 1 between positions 167,905,050 and 167,905,750.

Consequently, whilst the majority of GWAS SNV haplotypes lie within DHS ($\prod$ = 64.3% at $r^2 > 0.5$; $\prod$ = 46.6% at $r^2 > 0.8$) very few of these are contained within a DNase-seq footprint ($\prod$ = 6.5% at $r^2 > 0.5$; $\prod$ = 2.6% at $r^2 > 0.8$). Furthermore, even when the GWAS SNV haplotype included an eQTL SNV, in few cases was a variant underlying a DNase-seq footprint identified (4.0% at $r^2 > 0.5$; 1.6% at $r^2 > 0.8$). This suggests that sequence variation underlying the biological effect of most GWAS haplotypes is mostly located outside of DNase footprints, and thus inferred TF binding sites, potentially through DNA-TF interactions that are not captured by DNase-seq footprint analysis.

Finally, in order to assess whether the lack of enrichment of GWAS haplotypes within DNase-seq footprints was specific for brain-related GWAS signals, I generated footprint data on all DNase-seq tracks available from the ENCODE project.[28] In contrast to the brain-related GWAS signals, all available GWAS haplotypes were significantly yet modestly enriched within DNase-seq footprints pooled from all ENCODE tissue types, when compared with all pooled DHS (**Figure 23**). Nevertheless, this still accounted for only a small minority of all GWAS haplotypes ($\prod$ = 17.7% at $r^2 > 0.5$; $\prod$ = 8.6% at $r^2 > 0.8$). This could indicate that some binding sites are highly dynamic or tissue-specific. However, this supports my earlier observations that DNase-seq footprints cannot explain the majority of GWAS signals even at extremely permissive LD thresholds.

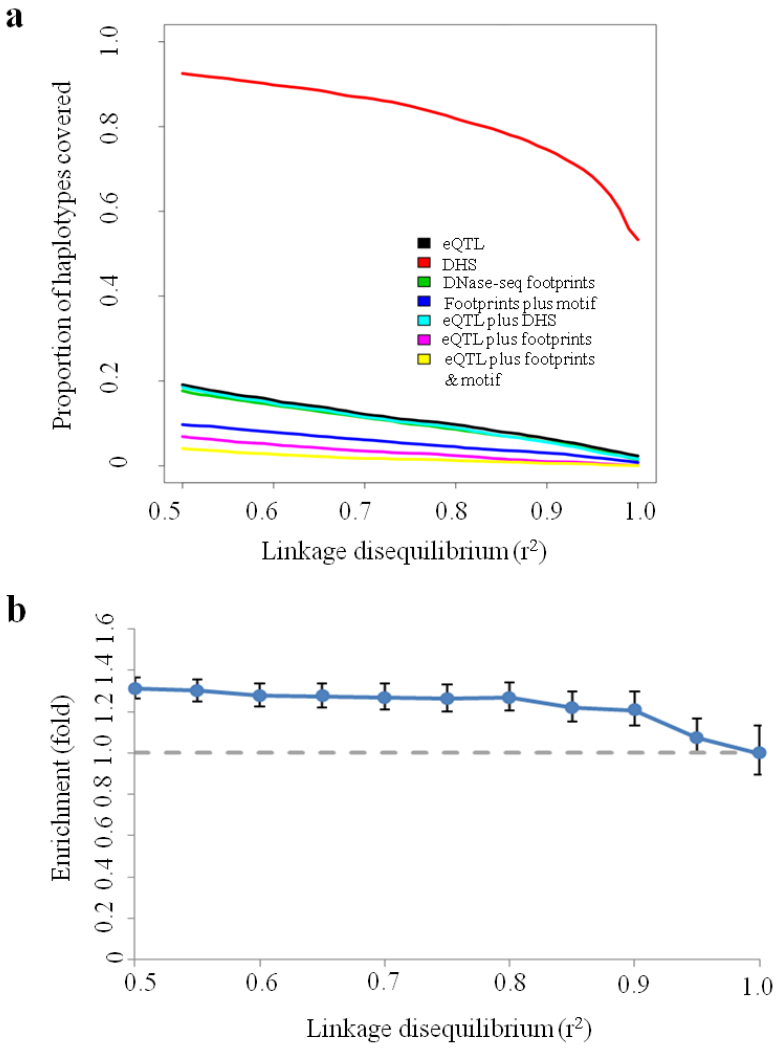

Figure 23 All GWAS haplotypes and DNase-seq footprints. (**a**) The proportion of GWAS haplotypes accounted for by various features. (**b**) The enrichment of GWAS haplotypes within all

footprints relative to DHS. The dashed grey line indicates the value corresponding to no enrichment.

4.6 Discussion

Given that many common SNVs typically lie in LD with any sentinel eQTL SNV (~27 at $r^2 > 0.5$; ~7 at $r^2 > 0.8$), it is challenging to reveal the variant responsible for the effect on gene expression. A prominent approach to achieve this is to identify TF binding sites, through DNase-seq and footprints, and to intersect these with known sequence variants that might alter the relevant TF's affinity.

My first observation was that although eQTL signals are enriched within DNase-seq footprints in excess of the previously reported association with open chromatin, this could only explain a very modest proportion of eQTL haplotypes. My second observation was that even this low level of enrichment was not observed for GWAS signals for brain-related traits. Based on the currently available brain DNase-seq and eQTL datasets I could estimate the proportions of eQTL and GWAS signals explicable by SNVs lying within TF binding sites as predicted by footprinting. Despite the proportion of eQTL variants falling within DHS being high (as observed previously), the proportion of either GWAS or eQTL haplotypes accounted for by direct disruption of TF binding sites predicted by footprints was minimal.[80] My findings on TF binding sites inferred from footprints are consistent with studies that found little enrichment of eQTLs within classical TF motifs.[224]

Molecular processes underlying genotype-phenotype relationships tend to be less proximal for GWAS than for eQTL studies. This is because eQTL variants explain variation in gene expression, typically in adult and specific tissues. By contrast, GWAS variants, although associated with disease, may exert their effects on gene expression or the epigenome only at specific developmental stages or ages or when subjects are exposed to specific exogenous or endogenous factors. Consequently, a potential disease-causative TF binding site may not be occupied in samples used in the GWAS study and therefore would not be considered by this study. The lower proportion of GWAS haplotypes within DHS that I observed compared with a previous study [80] could be due to my stricter definition of DHS. In particular I took advantage of

data from replicates through a robust irreproducibility discovery rate analysis in order that the DHS identified would be more likely to represent true open chromatin rather than background noise.[230,231]

The observed lack of enrichment in footprints for GWAS variants could reflect low predictive power caused by limited numbers of variants. This might be expected because there was a very modest but nonetheless significant enrichment of GWAS haplotypes for all traits within DNase-seq footprints generated from the much larger set of all available ENCODE data. Arguing against this, however, is that when I downsampled the number of eQTL haplotypes to match the number of brain-related GWAS haplotypes, a robust enrichment for DNase-seq footprints at $r^2 > 0.5$ remained (1.56-fold, $p < 0.002$). A further consideration is that the significance threshold I used to generate DNase footprints could be too conservative. However, when I tested different thresholds, my results did not change substantively.

There are four key limitations to inferring TF binding through DNase-seq footprint analysis. The first is that footprints may not be detected for TF binding which is dynamic, either through TFs migrating along DNA or through TFs with relatively rapid binding kinetics.[249] Transiently binding pioneer TFs capable of remodelling chromatin would not, for example, be detected by the method I used to identify footprints. Secondly, DNase-seq footprint analysis is also unlikely to detect TF binding events that require the co-ordination of multiple spatially, but not necessarily linearly, proximal DNA regions. Thirdly, TF binding in specific relatively rare cell types may be masked in bulk tissue samples.[250] Finally, binding sites that are disease-specific may not be detected in control datasets such as these. Evidence from an ATAC-seq study of $CD4^+$ T-cells suggests that many autoimmune disease causal variants are located preferentially in regions that show variable accessibility among individuals and over time, which implies that many true causal variants could be missed by DNase-seq footprinting in small numbers of control samples.[251]

With these caveats in mind, this chapter could also indicate that most eQTL and GWAS SNVs do not mediate their effects by directly disrupting classical TF binding events. In many GWAS, identification of an eQTL in LD with the GWAS SNV can help to prioritise the likely gene involved in conferring susceptibility at the associated genomic interval. However, even in this

situation where a GWAS haplotype contains an eQTL SNV, I did not in most cases identify a footprint-disruptive SNV (18.6% at $r^2 > 0.5$; 6.0% at $r^2 > 0.8$). It should also be noted that whilst I observed significant enrichment of eQTL haplotypes in footprints, even here, the majority (90.5% in the case of frontal cortex eQTLs at $r^2 > 0.8$) failed to disrupt a footprint. By combining multiple sources of DNase-seq footprinting and by assessing allelic imbalance in reads making up footprints, variants likely to alter TF binding site occupancy can be identified.[252–254] Even this large catalogue of tissue types suggests that only a small proportion of brain eQTL or brain-related GWAS haplotypes can be attributed to variants associated with altered TF binding in adult tissues: only 2.5% of eQTL and 3.0% of brain-related GWAS haplotypes at $r^2 > 0.8$ intersect one of the variants implicated in alteration of TF binding by Maurano and colleagues at FDR < 0.1%.[253] Even when using the most relaxed threshold in that study (FDR < 10%), this proportion only increased to 17.4% of eQTL and 18.1% of brain-related GWAS haplotypes.

Previous studies have suggested that SNVs lying within classical TF binding motifs are unlikely to account for a large proportion of TF binding variation.[226,228] My findings extend these results by showing that this is further reflected by the low proportion of eQTL and GWAS haplotypes that can be directly accounted for by TF binding site disruption. There are two potential, not mutually exclusive, explanations for this: firstly, that eQTL and GWAS causal variants genuinely do not commonly interfere with TF binding *via* direct disruption of TF recognition sequences, and secondly, that DNase-seq footprint analysis methodology fails to identify a considerable proportion of true positive binding sites.

The first explanation raises questions as to how GWAS and eQTL SNVs modulate gene regulation. eQTL and GWAS variants could also affect gene expression or phenotypes *via* TF-independent mechanisms, such as by altering the rate of transcription of rare non-coding RNA transcripts which have yet to be identified.[255]

Definitive proof that particular variants account causally for the eQTL effect within particular *loci* would likely require scarless genome editing of candidate functional SNVs in relevant tissue types. However, the identification of candidate causal SNVs will require the integration of many lines of evidence simultaneously and, particularly in haplotypes with multiple SNVs intersecting

enhancers, may be challenging.[81,245] Zinc finger nuclease editing of enhancers surrounding a candidate GWAS interval associated with glucose metabolism demonstrates the potential power of this approach with the caveats that single base edits are extremely difficult to implement and that the phenotypic read-outs (particularly for GWAS signals occurring in the absence of eQTLs) are likely to be at best subtle.[256] However, new bioresources and technologies, such as large repositories of induced pluripotent stem cell lines from many subjects and CRISPR-Cas9 nucleases for rapid genome engineering, should help to galvanise mechanistic eQTL and post-GWAS studies.

Just as likely an explanation is that DNase-seq footprints derived from existing datasets may not greatly assist in the prioritisation of candidate causative variants within eQTL or GWAS haplotypes. Variants lying outside of classical TF binding motifs and of footprints could alter binding through mechanisms that are not captured by DNase-seq footprint analysis, such as 3D chromatin interactions and dynamic binding patterns.[257] Many of these 3D interactions may themselves be dynamic and show cell type-specific signatures.[258] Similarly, DNase-seq footprinting is likely to miss cooperative TF binding.[259]

If so, then this explanation has potentially important implications for genomic engineering approaches such as those discussed above. This is because if not all true TF binding sites are identifiable by current motif- or TF binding assay-agnostic methods such as DNase-seq footprinting, then this would result in the number of variants requiring investigation via genome editing remaining high for most eQTL or GWAS haplotypes.

4.7 Conclusion

Further work will be needed to extend my observations into other cell and tissue types to establish whether similar findings can be detected outside of brain tissue. If this indeed proves to be the case, then efforts will need to be redoubled to delineate the molecular mechanisms underlying haplotypes that fail to directly disrupt TF binding.

4.8 Acknowledgements

Giuseppe Gallone generated reference genome files for alternate alleles.

Chapter 5 The role of *Foxn1* in maintenance of thymic epithelial cell function

5.1 Abstract

The thymus is involved in the education and selection of developing thymocytes to ensure that the peripheral T-cell repertoire can respond vigorously to non-self antigens but does not react to auto-antigens. Thymic epithelial cells (TEC) constitute a key part of this process and deficits in the development or maintenance of TEC function can result in immune system dysfunction. One key regulator of TEC development is the transcription factor, *Foxn1*. The loss of *Foxn1* function in mice or humans results in severe immunodeficiency. However, the mechanism by which *Foxn1* exerts its effects post-development has been largely opaque. In this chapter, I analysed Foxn1 ChIP-seq and RNA-seq data in the setting of either *Foxn1*-Flag hypomorphic or inducible *Foxn1* deficient alleles. This permitted the genome-wide mapping of Foxn1 binding and identification of a set of 450 high confidence *Foxn1* target genes. The loss of *Foxn1* function in mature TEC resulted in severe defects in thymic cellularity and thymopoiesis. This stemmed from decreased expression of multiple key genes involved in antigenic processing and presentation. For two novel *Foxn1* target genes, *Psmb11* and *Cd83*, it was shown that mutation of these genes could partially recapitulate defects in thymopoiesis observed in *Foxn1* deficient mice. Using the *Foxn1* binding motif identified from Foxn1 ChIP-seq peaks, it was also shown that specific motifs in the promoter of *Psmb11* and *Cd83* were critical for Foxn1 binding and subsequent transcriptional control of these target genes. Thus the integration of multiple functional genomics methods was able to demonstrate that *Foxn1* is indispensible for maintaining normal TEC function in the mature thymus by regulating a core set of 450 genes.

5.2 Hypotheses

5.2.1 The intersection of Foxn1 ChIP-seq and transcriptomic targets enables the identification of direct *Foxn1* target genes.

5.2.2 *Foxn1* target genes are enriched within pathways critical in TEC function.

5.3 Introduction

Chapter 2, **Chapter 3** and **Chapter 4** dealt with the central nervous system, a heterogeneous tissue comprising a large number of constitutent cell types. This chapter and the following one concern thymic epithelial cells (TEC), a population of cells that shows extreme transcriptional heterogeneity despite consisting broadly of two cell types: cortical TEC (cTEC) and medullary TEC (mTEC). This is because TEC are involved in the education of developing thymocytes *via* positive and negative selection in order to ensure that the resultant peripheral population of T-cells is able to react strongly to a wide array of non-self antigens but is depleted of potentially auto-reactive cells.[260]

Thymic development is choreographed by a multitude of molecular pathways, of which one key transcription factor is *Foxn1*.[97,260] The absence of *Foxn1*, as occurs in the nude mouse model, results in an alymphoid, dysgenic thymic rudiment.[104] Similarly, mutations in the human form of *FOXN1* cause a T-cell immunodeficiency syndrome associated with alopecia and nail dystrophy.[105]

Foxn1 is not required for the initial fate specification of TEC, since thympoiesis can be restored by post-natal reversion of mutant *Foxn1* to wild-type function.[107,261–263] Continued *Foxn1* expression is critical for the maintenance of both cTEC and mTEC compartment, since post-natal ablation of *Foxn1* causes apoptosis of TEC and over-expression of *Foxn1* delays thymic senescence.[106,264,265] Thus, *Foxn1* plays important roles in TEC development and maintenance at multiple points during thymogenesis and ongoing thymic function.

Precisely how *Foxn1* performs these various functions is unclear; research has been difficult partly due to the fact that mutations of *Foxn1* in mouse models result in athymia and partly

because of a lack of ChIP-quality antibodies available for Foxn1. Several genes (including *Ccl25*, *Cd40*, *Ctsl*, *Cxcl12*, *Dll4*, *Fgf2*, *Fgfr2*, *H2-Aa*, *H2-Ab1*, *Ly75*, *Notch1*, *Pax1* and *Psmb11*) have been proposed as *Foxn1* targets but none have been directly shown to be bound by Foxn1 in TEC.[107,262,266–270] It is also unknown whether Foxn1 binds to a particular sequence *in vivo*; *in vitro* data suggests that it is likely to coordinate DNA binding through a short consensus motif.[271,272]

I aimed to address this lacuna by integrating Foxn1 ChIP-seq data with RNA-seq data from *Foxn1*-deficient TEC. This would permit the identification of genes that are both transcriptionally regulated and directly bound by *Foxn1*. These molecular pathways are likely to be important in TEC function.

5.4 Methods

5.4.1 Mice

C57BL/6 and *nu/nu* mice were obtained from Janvier. The generation of β5t-rtTA and tetO-Cre animals has been reported previously.[273,274] Mice transgenic for a bacterial artificial chromosome (BAC) encoding a Foxn1-Flag fusion protein under the regulatory control of the *Foxn1* locus were generated in a similar manner to the previously reported Foxn1-Cre mice.[275] (Ref) Foxn1-Flag transgenic mice were backcrossed onto a *nu/nu* background to obtain mice homozygous or heterozygous for the BAC transgene, designated wt*/wt*$^{nu/nu}$ and wt*/-$^{nu/nu}$ respectively. Homologous recombination was used to generate mice with a conditional *Foxn1* locus, in which exons 7 and 8 are flanked by loxP sites, and these were subsequently crossed to β5t-rtTA and tetO-Cre transgenic animals to obtain triple transgenic mice. All animals were kept under specific pathogen-free conditions and experiments were carried out in accordance with local and national regulations and permissions. For Cre induction, one week old iFoxn1^{fl} mice were treated with a single intraperitoneal (i.p.) injection of Dox (0.3mg).

5.4.2 Flow cytometry

Thymocytes were incubated with antibodies raised against TCR beta (H57-597; Biolegend), CD4 (GK1.5; BioLegend), CD8 (53-67; eBioscience), CD25 (PC61.5; eBioscience), CD44 (IM7;

eBioscience), c-kit (eBioscience), CD24 (eBioscience), CD69 (H1.2F3; BD Biosciences), PD1 (29F.1A12, Biolegend) and CCR7 (4B12, Sony). For intracellular staining, cells were fixed, permeabilized (Fixation/Permeabilization kit, eBioscience) and stained for Foxp3 (FJK-16S, eBioscience) and Helios (22F6, eBioscience). TECs were stained using antibodies against CD45 (30F11; eBioscience), EpCAM (G8.8; DSHB, University of Iowa), MHCII (M5/114.15.2; BioLegend), Ly51 (6C3; BioLegend), UEA-1 (Reactolab), CD80 (16-10A1, Biolegend), CD86 (GL-1, Biolegend) and CD83 (Michel 19, Biolegend). For intracellular staining of CD83, cells were fixed, permeabilized (Cytofix/Cytoperm Kit, BD Biosciences) and labelled for the expression of CD83. Stained samples were acquired on a FACSAria II or BD LSRFortessa flow cytometers, and the data were analyzed using the FlowJo (Treestar) software.

5.4.3 Quantitative PCR

Total RNA was isolated from sorted cells using the RNeasy Kit (Qiagen), reverse transcribed to cDNA using SuperScriptIII (Live Technologies) and assessed by qPCR (SensiMix; Bioline).

5.4.4 Foxn1 ChIP-seq

Thymic lobes were digested with 0.2mg/ml Liberase TM (Roche Diagnostics) and 30µg/ml DNaseI (Roche Diagnostics) in PBS at 37˚C for 60 min. TECs were enriched to 15-20% using magnetic beads (autoMACS Pro Separator, Miltenyi Biotech) and then DNA was crosslinked. Nuclear isolation and chromatin fragmentation were performed with 20-30 x 10^6 cells enriched for TEC using the *tru*ChIP High Cell Chromatin Shearing Kit and nonionic Shearing Buffer according to the manufacturer's recommendations (Covaris). Chromatin was immunoprecipitated using the M2 anti-FLAG antibody (F1804; Sigma).[276] DNA from multiple ChIPs was pooled to generate two samples (5ng each) for library generation and sequencing (MRC, UK).

5.4.5 ChIP-seq analysis

Contaminating adapter sequences were removed from fastq sequences using Trimmomatic (version 0.32).[144] Reads were aligned against the mouse genome (UCSC build mm10) using pre-alignment with BWA (version 0.7.5 with the options -q10 -t4) and final alignment was obtained

with Stampy (version 1.0.23 with the options -t4 --bamkeepgoodreads - M).[277,278] Peaks were called on deduplicated aligned sequences using MACS2 (version 2.0.10) with a relaxed p-value setting of 0.1.[279] No MAPQ threshold was used in order to preserve sequencing data from repeat-rich regions, which are known to harbour functional TF binding.[280,281] Peaks from replicates were then pooled and analysed using irreproducible discovery rate analysis (IDR < 0.05).[230,282] Consistency between ChIP and input replicate pairs was calculated and then the resultant threshold applied to merged ChIP and merged input data. This modified approach was intended to mitigate any influence of batch-to-batch variability in the input samples on peak-calling. Peaks were filtered against the ENCODE blacklist regions.[231,237] *De novo* motif discovery within Foxn1 ChIP-seq peaks was carried out using MEMEChIP on the 200 base-pairs surrounding the peak summit.[283] Motif location was identified using FIMO with a threshold of $p < 0.05$ on repeat-masked fasta files relative to a first-order Markov background.[241] Foxn1 ChIP-seq peaks were scanned for co-factor motifs using PScanChIP and a mixed background.[284] Genomic enrichment was calculated using GAT with 10,000 randomisations, controlling for GC content and using an appropriate workspace as background.[240] Genomic features were annotated using HOMER.[244] ChIP-seq peaks were allocated to genes if these were located within 5kb upstream or 100 bases downstream of the TSS using custom R scripts or BETA (using default settings).[285]

5.4.6 RNA-seq

TECs were enriched as detailed above and then sorted using a FACSAria II (BD Bioscience). RNA was isolated using RNeasy kit (Qiagen) and subjected to sequencing (TrueSeq, BGI).

5.4.7 RNA-seq analysis

Reads were aligned against the Ensembl transcriptome and genome (GRCm38) using Tophat2 (version 2.0.10).[286] Reads were allocated to protein-coding gene meta-features using featureCounts (requiring both pairs of reads to be aligned and excluding multi-mapping reads).[146] Exon-level expression data was estimated using BEDtools coverage.[182] Differential expression analysis on genes with at least 1 aligned fragment was conducted using general linear modelling

in edgeR, correcting for common, trended and tagwise dispersion, with additional batch correction for the conditional Foxn1 knock-out model.[203]

5.4.8 ATAC-seq

A cell pellet of ~10,000 cTEC was lysed, treated with transposase and DNA tagmented in accordance with a previously published protocol.[287] Tagmented fragments were amplified by PCR and sequenced using an Illumina MiSeq on 75bp paired end reads.

5.4.9 ATAC-seq analysis

Fragments were aligned to the mm10 genome using Bowtie2 (version 2.2.3 with the options --no-discordant --no-mixed -X 2000).[288] Read positions were corrected for transposon insertion offset.[8] ATAC-seq coverage 5kb upstream or 100 bases downstream of the TSS of *Foxn1* target genes was compared with that of all other protein-coding genes using an unpaired, two-tailed Wilcoxon test.

5.4.10 Co-expression analysis

Microarray expression data was downloaded from GSE56928.[289] This was quantile normalised and a thymic epithelial cell-specific signed co-expression network constructed using bidirectional weighted correlation in WGCNA.[290] The optimal soft threshold power was chosen to achieve $r^2 > 0.9$ for the network. Individual sets of genes were analysed by permutation analysis, controlling for expression deciles. This dataset was also used to generate a list of genes likely to be primarily driven by non-TEC contaminating thymic stromal cells. To this end, the dataset was screened for genes displaying significant differences on ANOVA testing between TECs and non-TECs, 4-fold higher expression in non-TECs than TECs and an expression level in TECs within the bottom quintile of genes.

5.4.11 Gene ontology analysis

Gene ontology analysis was conducted using DAVID.[149] Only enriched terms with Benjamini-Hochberg corrected p-values < 0.05 were reported.

5.4.12 Histology

Frozen thymus tissue sections (8µm) were fixed in acetone and stained using antibodies against Psmb11 (MBL), CD4 (GK1.5; BioLegend), CK5 (Covance), CK8 (Progen), ERTR7 (provided by W. van Ewijk, Netherlands), Foxn1 (provided by T. Amagai, Japan) and Aire (5H12, eBioscience). Image acquisition was undertaken on a Zeiss LSM510 (Carl Zeiss).

5.4.13 Luciferase assay

Promoter fragments of *Psmb11* and *Cd83* were subcloned into the pGL4.10 (luc2, Promega) reporter plasmid. These were then co-transfected along with an expression vector encoding Foxn1 into HEK293 cells using FuGENE HD (Promega). Luciferease activity was measured 24hr later using the Dual-Lucifearase Assay kit (Promega). Candidate Foxn1 motifs were mutated using the Q5 Site-Directed Mutagenesis Kit (NEB).

5.4.14 Statistical analysis of cytometric, qPCR and luciferase data

P-values were calculated using unpaired, two-tailed t-tests.

5.5 Results

5.5.1 Transgenic rescue of the nude mouse phenotype with the insertion of a BAC containing *Foxn1*-Flag

The insertion of a single copy of *Foxn1*-Flag into a mouse on the background of a nude phenotype (wt*/-$^{nu/nu}$) was insufficient to rescue the phenotype. Heterozygous mice demonstrated reduced thymic cellularity and a severe reduction in the mTEC compartment (**Figure 24a-d**). In contrast, homozygous mice (wt*/wt*$^{nu/nu}$) displayed almost normal thymic cellularity (**Figure 24a-d**).

The aberrant cellularity in wt*/-$^{nu/nu}$ mice translated into a deficit in thymopoiesis, with alterations in the proportion of maturing thymocytes at double negative (DN) and double positive (DP) stages (**Figure 24e&f**). In contrast, wt*/wt*$^{nu/nu}$ thymopoiesis was much closer to wild-type (**Figure 24e&f**).

This suggests that the *Foxn1*-Flag BAC is a hypomorphic allele, which can rescue the nude phenotype when present homozygously. This presented the option of interrogating *Foxn1* transcriptomic function in the hypomorphic (but, importantly, still extant) thymus of wt*/-$^{nu/nu}$ mice and comparing this to the near-normal wt*/wt*$^{nu/nu}$ thymus, as well as taking advantage of high affinity, specific antibodies raised to the Flag portion of the *Foxn1*-Flag transgene to perform ChIP-seq.

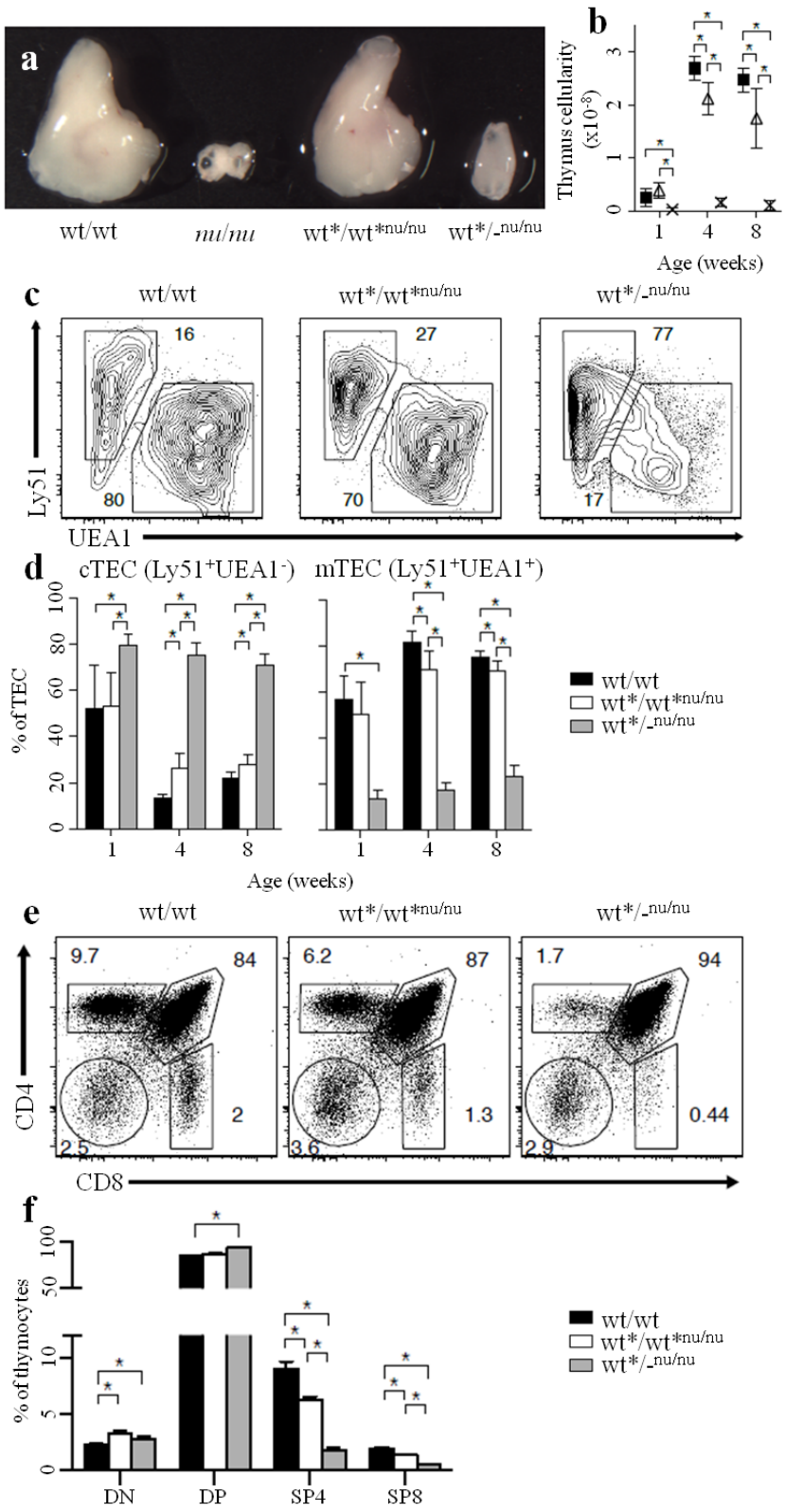

Figure 24 Transgenic rescue of nude phenotype in homozygous *Foxn1*-Flag mice. (**a**) Macroscopic analysis of thymic lobes from 4-5 week old mice of the indicated genotype. (**b**) Absolute cell numbers (mean ± SD) of a thymus from mice of the indicated genotype and age. The data shown is representative of two independent experiments, displaying the mean ±SD and analyzing ≥3 mice per group. Squares = wt/wt; triangles = wt*/wt*$^{nu.nu}$; and crosses = wt*/-$^{nu/nu}$. (**c**) Flow cytometric analysis of TEC (CD45^{-}EpCAM^{+}MHCII^{+}) subpopulations isolated from mice

with the indicated genotype at 4 weeks of age using the expression of Ly51 and reactivity with UEA1 to identify cortical and medullary epithelia, respectively. The numbers shown in the individual gates in each flow cytometry plot represent the frequencies observed for cTEC and mTEC in a representative experiment. (**d**) The frequencies of cTEC ($CD45^-EpCAM^+MHCII^+UEA1^-Ly51^+$) and mTEC ($CD45^-EpCAM^+MHCII^+UEA1^+Ly51^-$) in mice of the indicated genotype and age are displayed in the bar graphs. The data shown is representative of two independent experiments, displaying the mean ±SD and analyzing 3-4 mice per group. (**e**) Representative flow cytometric analysis of CD4 and CD8 expression on thymocytes. (**f**) The bar graph displays the frequencies (mean±SD) of the indicated subpopulations from at least two independent experiments. * = $p < 0.05$

5.5.2 *Foxn1* binding sites and *de novo* motif discovery

I used ChIP-seq to identify genome-wide binding sites for *Foxn1* in wt*/wt*$^{nu/nu}$ TEC. Irreproducibility discovery analysis was able to identify 9,012 peaks at an irreproducibility discovery rate (IDR) < 0.05. Approximately one third of these peaks were located within 5kb upstream and 100 bases downstream of transcriptional start sites (TSS) (**Figure 25a&b**). These peaks showed high levels of enrichment for previously identified promoter-associated histone modifications and open chromatin (H3K4me3: 23-fold, $p < 0.0001$; ATAC-seq open chromatin: 75-fold, $p < 0.0001$).[8,113] However, *Foxn1* did not appear to target tissue-restricted antigens (TRAs), showing significant impoverishment of binding near the TSS of these relative to all other genes (0.53-fold, $p < 0.0001$).

I used MEMEChIP to identify over-represented, centrally enriched motifs within 100 bases of Foxn1 ChIP-seq peak summits.[283] This found one short motif that show strong central enrichment, was present at 23% of binding sites (E-value < 10^{-129}; **Figure 25c**) and was highly enriched near the summit of Foxn1 peaks (2.9-fold, $p < 0.0001$; **Figure 25d**). It also was highly similar to a previously reported *in vitro* Foxn1 motif (ACGC) derived from an electrophoretic mobility shift assay.[271]

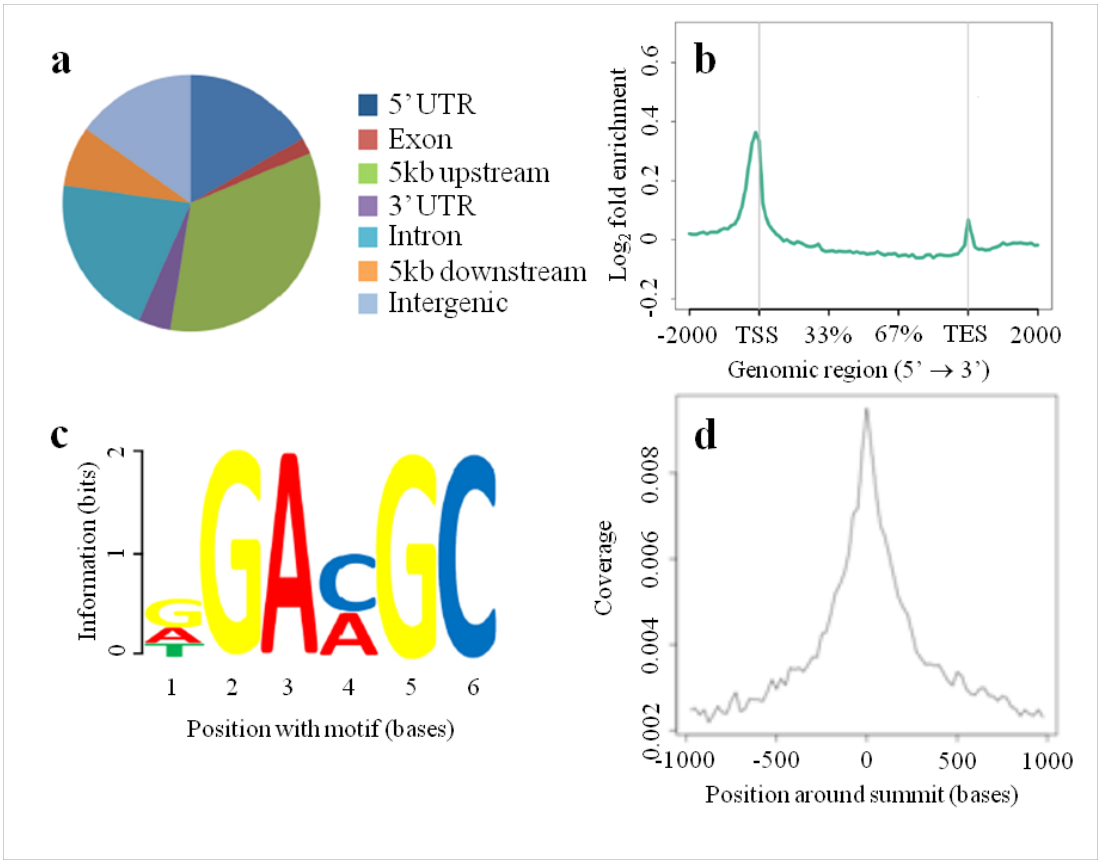

Figure 25 Foxn1 ChIP-seq analysis. (**a**) The pie-chart shows the proportion of Foxn1 ChIP-seq peaks falling within the different RefSeq meta-gene regions as indicated. (**b**) Enrichment of Foxn1 ChIP-seq signal across a meta-gene profile comprising all RefSeq mouse genes relative to control. (**c**) MEMEChIP-derived Foxn1 binding site motif (E-value < 10^{-129}). (**d**) Motif coverage relative to the summit of Foxn1 ChIP-seq peaks.

5.5.3 Identification of *Foxn1* direct target genes

Despite identifying a set of genes (5,415) bound by Foxn1, many of these may not be transcriptomically regulated by it.[228] In order to identify genes with altered expression in response to altered levels of functional *Foxn1*, I compared the transcriptomes of cTEC and mTEC from wt*/-$^{nu/nu}$ and wt*/wt*$^{nu/nu}$. This found that large numbers of genes were differentially expressed between the two genotypes (cTEC: 8,378; mTEC 11,690; FDR < 0.05). A subset of these were also directly bound by Foxn1 (cTEC: 1,148; mTEC: 968). However, just as genes bound by Foxn1 are unlikely all to be transcriptomically regulated by *Foxn1*, many of the genes identified as differentially expressed and direct Foxn1 targets in this model may represent indirect effects of *Foxn1* deficiency through secondary regulatory mechanisms.

This ambiguity was addressed by utilising an additional mouse model. In the iFoxn1$^{\Delta 7,8}$ model, exons 7 and 8 of the *Foxn1* locus are deleted in a Doxycycline (Dox)-induced, Cre-mediated and cTEC-restricted fashion. In the absence of Dox, iFoxn1$^{\Delta 7,8}$ mice displayed normal thymic cellularity and thymopoiesis (**Figure 26a**). 72 hours after treatment, thymic cellularity was reduced but thymopoiesis remained unaffected. RT-qPCR analysis of genes which were previously suspected to be *Foxn1* targets and which had a Foxn1 binding site near their TSS confirmed significant downregulation of these genes after treatment with Dox (**Figure 26b**). A comparison of RNA-seq data obtained on treated and untreated cTEC identified 2,506 differentially expressed genes (FDR < 0.05).

I was therefore able to intersect the results of the *Foxn1* ChIP-seq analysis with those of the RNA-seq comparison of wt*/wt*$^{nu/nu}$ *vs.* wt*/-$^{nu./nu}$ and treated *vs.* untreated iFoxn1$^{\Delta 7,8}$ TEC to identify a set of high confidence genes likely to be direct transcriptional targets of *Foxn1* (**Figure 26c**). This analysis suggested that *Foxn1* acts largely as a transcriptional activator rather than a repressor ($p < 10^{-7}$) and, indeed, genes showing decreased expression after Dox treatment were 1.92-fold more likely to have a nearby Foxn1 binding site than was the case for other genes ($p < 0.01$; **Figure 26d**). For this reason, I restricted my set of high confidence *Foxn1* targets to those which had a ChIP-seq peak within 5kb upstream or 100 bases downstream of the TSS, were decreased in expression in wt*/-$^{nu/nu}$ relative to wt*/wt*$^{nu/nu}$ cTEC and were decreased in expression in Dox-treated relative to untreated iFoxn1$^{\Delta 7,8}$ cTEC. This produced a final list of 450 genes (**Figure 26c**;

Table 6). The regions near the TSS of these genes were enriched for open chromatin in cTEC, confirming that these would be accessible to Foxn1 (**Figure 26e**; 1.54-fold, $p < 10^{-20}$).

However, many of the cellular changes observed in the thymi of wt*/-$^{nu/nu}$ mice suggested particular impairments in mTEC function. Since the Cre in iFoxn1$^{\Delta 7,8}$ was driven by the promoter of *Psmb11*, a specific marker of cTEC, it was not possible to formulate a list of direct *Foxn1* targets in mTEC. Reassuringly, though, many of the high confidence direct *Foxn1* target genes

were also highly expressed in mTEC, suggesting that these are likely to form a core of functionally important genes in both cTEC and mTEC (**Figure 26f**).

Gene ontology analysis suggested that these functions are likely to revolve around the processing of antigens for presentation to thymocytes (**Figure 26g**). Pathways related to thronine peptidases and proteasomal function were highly enriched in this high confidence set of genes (16.7-fold and 7.29-fold respectively; $p < 0.0001$). Given the significant association of *Foxn1* expression with pathways involved in antigen presentation, I next tested whether there was differential expression of *Aire*-regulated genes when comparing wt*/wt*$^{nu/nu}$ with wt*/-$^{nu/nu}$ mTEC. As a group, *Aire*-regulated genes tended to be more highly expressed in mTEC of wt*/wt*$^{nu/nu}$ mice than in wt*/-$^{nu/nu}$ animals compared to either *Aire*-independent tissue restricted antigens or all other protein-coding genes (median fold change: *Aire*-induced tissue restricted antigens 2.93-fold, other *Aire*-induced genes 2.61-fold, *Aire*-independent tissue restricted antigens 1.14-fold, and other protein-coding genes 1.00-fold; $p<0.05$ for all pair-wise comparisons). As the expression of *Aire* was comparable in mTEC with different *Foxn1* complementation (1.10-fold, $p=0.23$), these findings suggest that *Foxn1* expression facilitates promiscuous gene expression in a fashion that is independent but complementary to *Aire*.

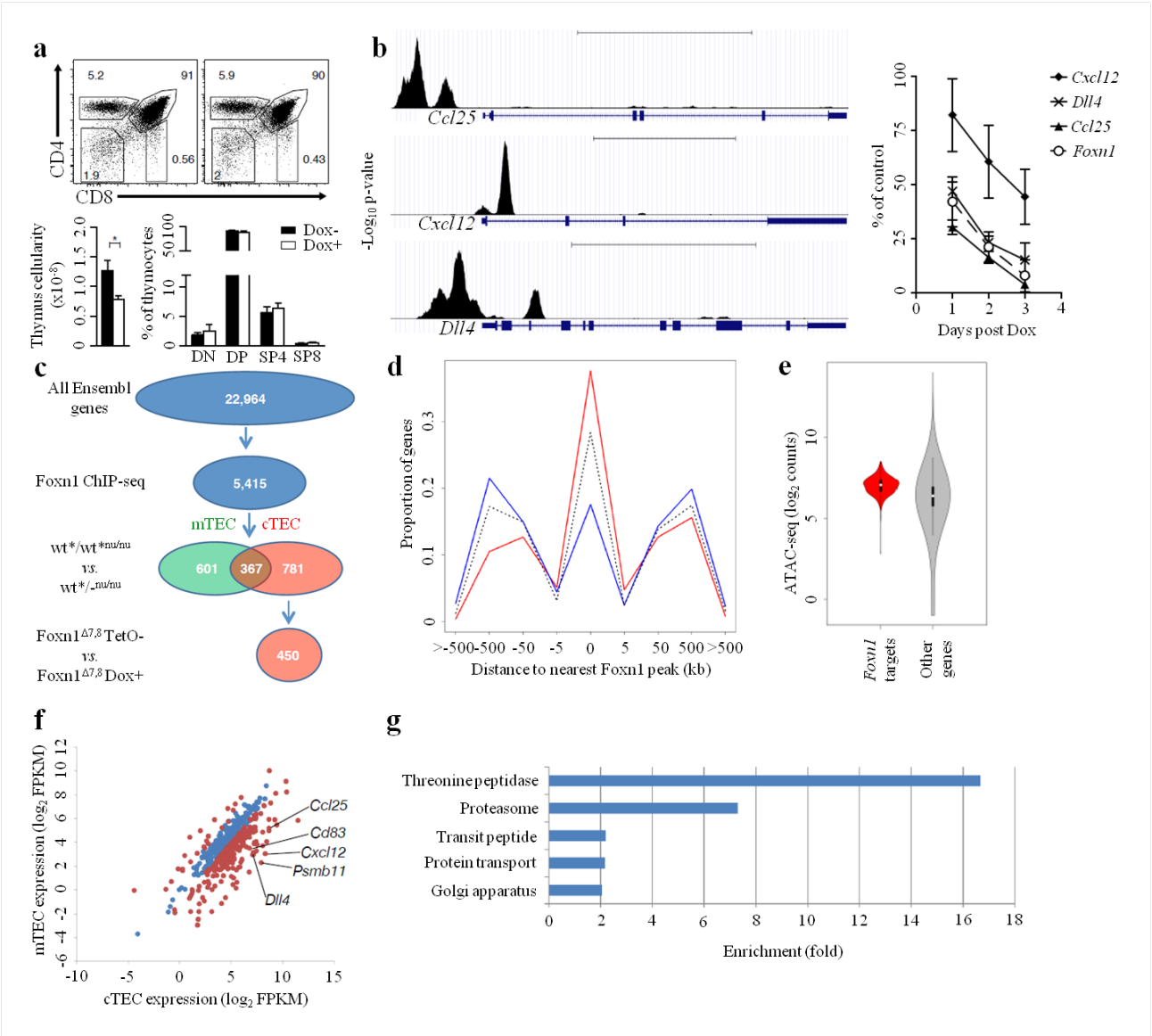

Figure 26 Intersection of Foxn1 ChIP-seq and RNA-seq analyses. (**a**) Thymic cellularity and intrathymic T cell development in one week old iFoxn1$^{\Delta7,8}$ mice 72hrs after exposure to Doxycycline (Dox+). Mice were treated i.p. with a single dose of doxycycline (Dox+) or PBS (Dox-) and analyzed 3 days later for the expression of CD4 and CD8 (upper panels and lower right bar graph) and total thymus cellularity. The numbers in the individual plots indicate the percentages of cells in the respective gates. The bar graphs show mean values ±SD and is representative of two independent experiments with ≥3 mice per group. (**b**) Left panels: Foxn1 ChIP-seq binding profiles (-log$_{10}$ p-value) for the suspected Foxn1 target genes *Ccl25* (top), *Cxcl12* (middle) and *Dll4* (bottom). Scale bars represent 5kb (top) and 2kb (middle and bottom). Right panel: RT-qPCR analysis for *Cxcl12*, *Dll4*, *Ccl25* and *Foxn1* transcripts in cTECs isolated at the indicated times after treatment of one week old iFoxn1$^{\Delta7,8}$ mice with a single dose of Dox or saline (control). (**c**) The flow chart shows the number of genes filtered from all protein coding Ensembl genes by sequential use of data obtained by Foxn1 ChIP-seq analysis, RNA-seq differential gene expression profiling in cTEC and mTEC comparing wt*/-$^{nu/nu}$ vs. wt*/wt*$^{nu/nu}$ mice, and transcriptome analyses of cTEC isolated from Dox-treated iFoxn1$^{\Delta7,8}$ and Dox-treated Foxn1$^{\Delta7,8}$ mice which lack the tetO-Cre transgene and therefore cannot delete exons 7 and 8 of the *Foxn1* locus in response to exposure to the drug. Genes were restricted to those downregulated with decreased Foxn1 activity. ChIP-seq targets were defined as genes with a Foxn1 peak (IDR < 0.05) 5kb upstream or 100 bases downstream of the TSS. Genes with FDR < 0.05 were called as

differentially expressed. (**d**) Distribution of Foxn1 ChIP-seq peaks around the transcriptional start site (TSS) of genes categorized by the direction of change following exposure of iFoxn1$^{\Delta 7,8}$ mice to Dox treatment. (**e**) ATAC-seq signal in cTEC near the TSS of high confidence Foxn1 target genes (red) and all other genes (gray). (**f**) Scatter plot analysis of expression levels of high confidence Foxn1 targets in cTECs and mTECs. Red indicates genes with at least 2-fold differential expression between cTECs and mTECs. Blue indicates genes not reaching this threshold. (**g**) Gene ontology analysis of high confidence Foxn1 gene targets. All terms with Benjaminin-Hochberg adjusted $p < 0.05$ and enrichment > 2 fold are shown. The number of high confidence targets in each term are: threonine peptidase (8, representing 38.1% of genes in that gene ontology category); proteasome (9, 16.7%); transit peptides (23, 5.0%); protein trafficking (23, 5.0%) and Golgi apparatus (26, 4.7%). * = $p < 0.05$.

Table 6 Top 30 high confidence *Foxn1* regulated genes ranked by $\log_2$ fold change. High confidence genes are those with a Foxn1 ChIP-seq peak 5kb upstream or 100 bases downstream of the transcription start site, and down-regulated in both wt*/-nu/nu relative to wt*/wt*nu/nu cTEC and Dox-treated TetO+ relative to TetO- iFoxn1$^{\Delta 7,8}$ cTEC at FDR < 0.05. Genes are ranked by $\log_2$ fold change (FC) in expression between wt*/-$^{nu/nu}$ and wt*/wt*$^{nu/nu}$ cTEC.

Gene symbol	wt*/-$^{nu/nu}$ *vs.* wt*/wt*$^{nu/nu}$ cTEC		Dox-treated TetO+ *vs.* TetO- iFoxn1$^{\Delta 7,8}$ cTEC	
	$\log_2$ FC	FDR	$\log_2$ FC	FDR
Ccl25	7.43	$< 10^{-20}$	3.99	$< 10^{-20}$
Prss16	4.77	$< 10^{-20}$	3.27	$< 10^{-20}$
Gpr25	4.33	$< 10^{-20}$	2.59	$< 10^{-15}$
Tspan10	4.30	$< 10^{-9}$	2.44	$< 10^{-8}$
AA467197	4.12	$< 10^{-20}$	2.55	$< 10^{-15}$
Cd83	3.99	$< 10^{-20}$	2.75	$< 10^{-20}$
Tbata	3.70	$< 10^{-20}$	3.88	$< 10^{-20}$
Ajap1	3.29	$< 10^{-20}$	2.35	$< 10^{-20}$
5031414D18Rik	3.27	$< 10^{-20}$	2.47	$< 10^{-15}$
Psmb11	3.25	$< 10^{-20}$	1.67	0.006
Apba2	3.20	$< 10^{-20}$	1.09	$< 10^{-11}$
Dtx4	3.09	$< 10^{-20}$	1.31	$< 10^{-11}$
Pds5b	3.08	$< 10^{-20}$	1.12	$< 10^{-9}$
Stc2	3.06	$< 10^{-20}$	1.55	$< 10^{-5}$
Zar1	2.98	$< 10^{-9}$	2.46	$< 10^{-7}$
Snap91	2.98	$< 10^{-20}$	2.36	$< 10^{-20}$
Srxn1	2.93	$< 10^{-20}$	1.12	$< 10^{-10}$
Ctnna2	2.89	$< 10^{-20}$	1.78	$< 10^{-4}$
Lhx3	2.78	0.02	2.34	0.03
Cecr6	2.73	$< 10^{-20}$	2.93	$< 10^{-19}$
Ankmy1	2.70	$< 10^{-18}$	2.43	$< 10^{-9}$
Tmie	2.68	$< 10^{-20}$	1.15	$< 10^{-10}$
Abca15	2.67	0.002	2.38	0.006
Gmfg	2.67	$< 10^{-11}$	0.74	0.02
Arhgef10l	2.66	$< 10^{-20}$	1.33	$< 10^{-16}$
Ptprn2	2.60	$< 10^{-20}$	2.01	$< 10^{-11}$
Lmx1a	2.59	$< 10^{-10}$	2.79	$< 10^{-5}$
Endou	2.55	$< 10^{-13}$	0.86	0.002
Mfsd12	2.45	$< 10^{-20}$	2.10	$< 10^{-20}$
Akap2	2.29	$< 10^{-20}$	0.82	0.01

The Cre used in the iFoxn1$^{\Delta 7,8}$ model was slightly leaky, with the result that some TetO- cTEC underwent Cre-mediated complementation of the *Foxn1* locus. This had the advantage that I could assess whether high confidence *Foxn1* gene targets demonstrated a linear relationship between the degree to which the LoxP-flanked exons were recombined in each sample and the expression of each gene. This indeed proved to be the case (r^2=0.92, $p < 0.0001$; **Figure 27**).

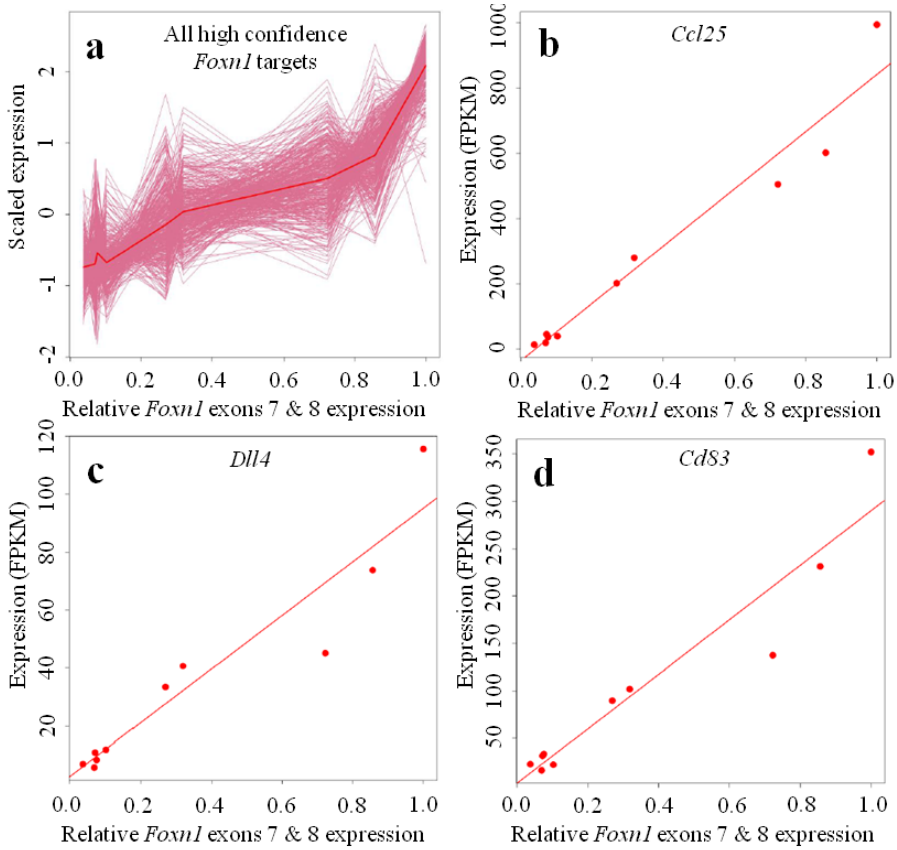

Figure 27 Linear relationship between relative *Foxn1* DNA binding exons expression and *Foxn1* target gene expression. (**a**) All high confidence *Foxn1* gene targets. Overall mean gene expression is shown in dark red ($r^2 = 0.92$, $p < 0.0001$). (**b**) *Ccl25* ($r^2 = 0.95$, $p < 0.0001$). (**c**) *Dll4* ($r^2 = 0.90$, $p < 0.0001$). (**d**) *Cd83* ($r^2 = 0.92$, $p < 0.0001$). Relative *Foxn1* DNA binding exon expression was calculated as the mean number of reads aligned to the *Foxn1* DNA binding exons 7 and 8 normalized by RNA-seq library size and overall *Foxn1* expression.

I next wanted to understand the context in which *Foxn1* target genes operate. In order to do this I used an existing microarray dataset from murine thymus cell subpopulations (**Figure 28a**).[289] Reassuringly, the expression of high confidence *Foxn1* targets was positively correlated with one another in a TEC-specific co-expression network (median $r = 0.13$, $p < 0.002$). Extending this to all TEC subsets allowed me to identify modules of co-expressed genes in which *Foxn1* targets

were over-represented. *Dll4*, *Ccl25* and several other putative high confidence *Foxn1* targets were significantly enriched within a cTEC-specific gene module associated with the regulation of T-cell activity (3.38-fold, $p < 0.002$). Other *Foxn1* targets, including *Psmb4*, *Psmb9* and *Tasp1*, were enriched within an mTEC^{hi} gene module of unknown biological function (2.24-fold, $p < 0.002$).

5.5.4 Potential Foxn1 co-factors

Foxn1 binding is unlikely to occur in isolation and so I used PScanChIP in an attempt to identify other TFs with recognition motifs enriched near Foxn1 binding sites.[284] This analysis identified several motifs that are significantly and centrally overrepresented when compared to a background of mixed promoter and enhancer regions likely to represent Foxn1 co-factors (**Figure 28b&c**). Among these were binding site motifs for transcription factors Creb1 and Tp63, which are known to regulate proliferation and TEC aging, respectively.[291,292] Supporting this computational prediction of co-factors, Foxn1 binding was significantly enriched near Tp63 binding sites that were previously identified in *Foxn1* expressing keratinocytes relative either to the whole genome or mTEC^{hi} H3K4me3 peaks (whole genome: 11.8-fold, $p < 0.0001$; H3K4me3: 2.03-fold, $p < 0.005$). This corroborated a Foxn1-Tp63 axis that could potentially regulate TEC homeostasis.[293–295]

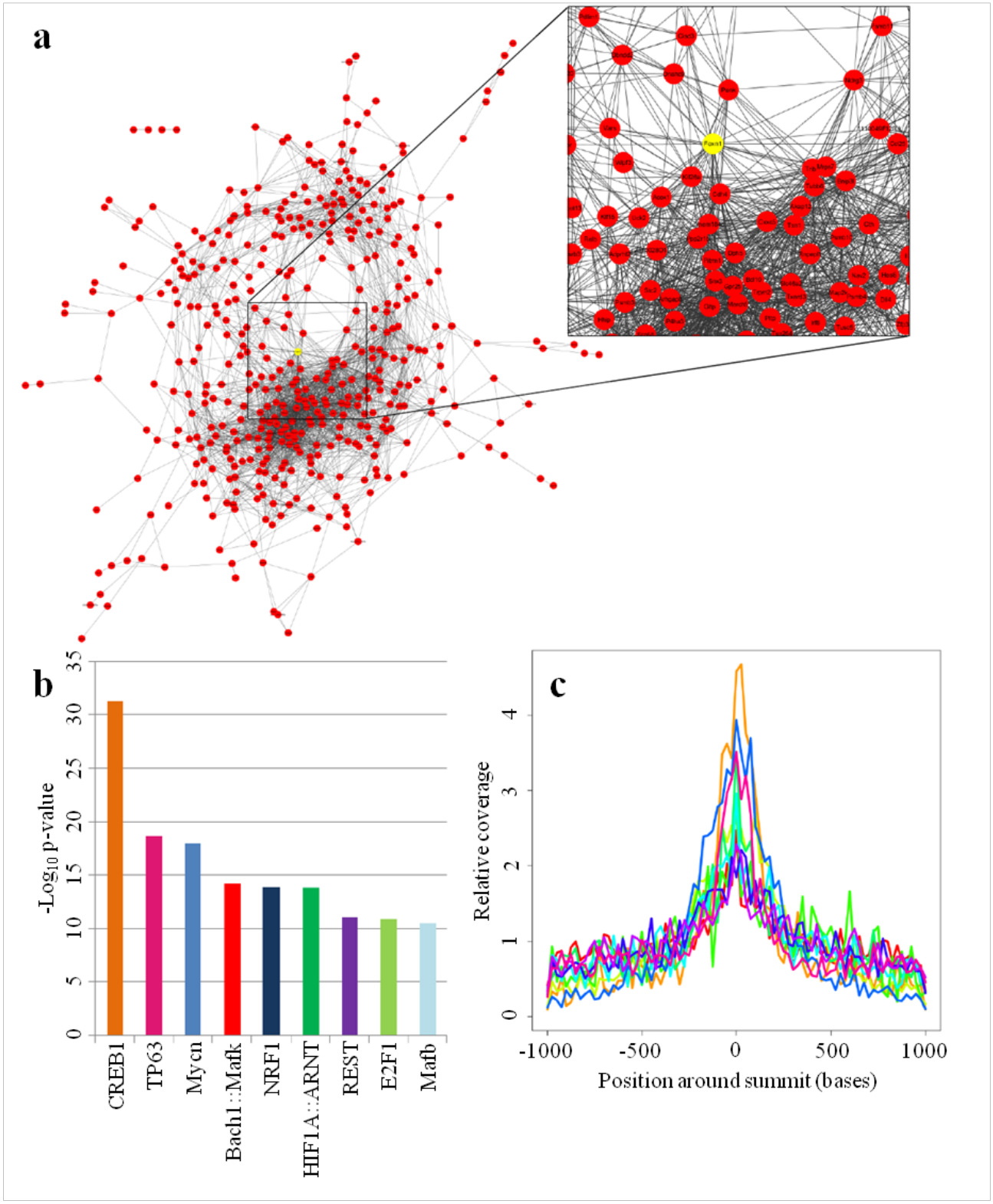

Figure 28 Network analysis and assessment of putative binding partners. (**a**) A direct seed co-expression network for cTEC generated from TEC-specific microarray data.[289] Interactions are shown for $r^2 > 0.7$. The local network surrounding Foxn1 (yellow) is shown in the magnified graphic view. (**b**) Likely binding partners of Foxn1 detected by PScan ChIP local motif enrichment. Specific motif IDs are CREB1_MA0018.1, TP63_MA0525.1, Mycn_MA0104.2, Bach1::Mafk_MA0591.1, NRF1_MA0506.1, HIF1A::ARNT_MA0259.1, REST_MA0138.2, E2F1_MA0024.2 and Mafb_MA0117.1. (**c**) Motif enrichment for binding motifs for several

PScan predicted binding partners near Foxn1 ChIP-seq peaks. Enrichment is shown relative to the mean coverage near Foxn1 ChIP-seq peaks. Colours are as in panel **b**.

5.5.5 Functional validation of the novel *Foxn1* motif

Two of the high confidence *Foxn1* target genes, *Psmb11* and *Cd83*, were chosen for further functional validation. These were selected because there is prior evidence that both *Psmb11* and *Cd83* are critical for the normal maturation of T-cells.[270,296–298] Foxn1 ChIP-seq peaks were located near the TSS of both *Psmb11* and *Cd83* (**Figure 29a**). qPCR analysis confirmed that the level of expression of *Psmb11* and *Cd83* was reduced after Dox treatment in iFoxn1$^{\Delta 7,8}$ cTEC (**Figure 29b**).

Analysis of thympoiesis in *Psmb11* knock-out mice demonstrated a clear defect in the production of SP8 thymocytes, which was very similar to that observed in wt*/-$^{nu/nu}$ mice (**Figure 29c**). This was associated with a deficit in negative selection, manifested as a reduction in the simultaneous expression of PD1 and Helios in CCR7^{-}Foxp3^{-} thymocytes, and a block in thymocyte maturation (**Figure 29c**).

In contrast, *Cd83* knock-out mice had a reduction in the proportion of SP4 thymocytes due to a failure in the SP4 maturation pathway (**Figure 29d**).

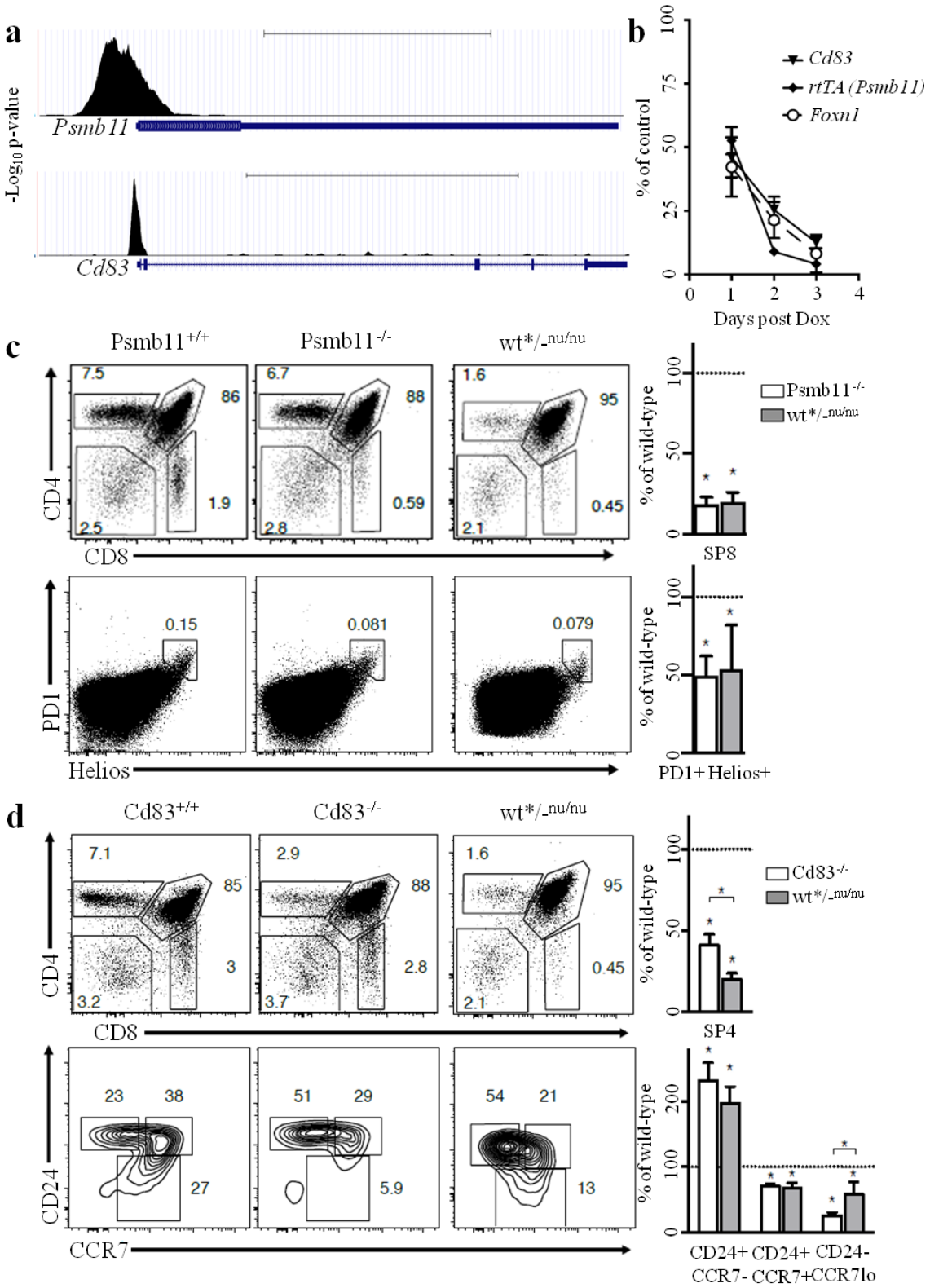

Figure 29 *Psmb11* and *Cd83* are direct targets of Foxn1. (**a**) Foxn1 ChIP-seq binding profiles (-$\log_{10}$ p-value) for the novel Foxn1 target genes *Psmb11* (top) and *Cd83* (bottom). Scale bars represent 2kb (top) and 5kb (bottom). (**b**) RT-qPCR analysis of gene expression in cTEC isolated from one week old iFoxn1$^{\Delta 7,8}$ mice at the indicated times after treatment with Dox or saline (control). Because the coding sequence of *Psmb11* is replaced homozygously in iFoxn1$^{\Delta 7,8}$ mice with the *rtTA* transgene, *rtTA* transcripts were measured in lieu of *Psmb11* encoding the proteasome component β5t. Data is representative of two independent experiments with 3 mice

per group and displays mean values±SD of three individual measurements per experimental group and time point. (**c**) Top left: Representative flow cytometric analysis of CD4 and CD8 expression on thymocytes isolated from 5 week old mice with the indicated genotype. Top right: Display of the relative frequency (mean ± SD) of single positive, mature thymocytes isolated from the indicated mouse strains as compared to wild-type controls whose values have been set as 100%. The data is representative of two independent experiments with ≥3 mice per group. Bottom left: Representative flow cytometric analysis of $Foxp3^{-}CCR7^{-}$ thymocytes for the expression of Helios and PD1. The cells were isolated from 5 week old mice with the indicated genotype. Bottom right: The bar graph displays the relative frequencies (mean ± SD) of $PD1^{+}Helios^{+}$ cells among Foxp3-CCR7- thymocytes as compared to wild-type controls whose values have been set as 100%. The data represents two independent experiments with at least 3 mice per group. (**d**) Top left: Representative flow cytometric analysis of CD4 and CD8 expression on thymocytes isolated from 5 ($wt^{*/-^{nu/nu}}$) and 7 week old ($CD83^{+/+}$ and $CD83^{-/-}$) mice. Top right: The bar graph displays the relative frequencies (mean ± SD) of single CD4- and CD8-positive thymocytes as compared to wild-type controls whose values have been set as 100%. The data represents two independent experiments with at least 3 mice per group. Bottom left: Representative flow cytometric analysis of CD24 and CCR7 expression on single CD4-positive (SP4; $CD4^{+}CD8^{-}TCR^{+}CD5^{+}$) thymocytes isolated from mice with the indicated genotype. Bottom right: The bar graph displays the relative frequencies (mean ± SD) of the indicated subpopulations as compared to wild-type controls whose values have been set as 100%. The data shown is representative of two independent experiments with at least 3 mice per group. * = $p < 0.05$.

Having confirmed that both *Psmb11* and *Cd83* are important in thympoiesis, the functional relevance of candidate Foxn1 recognition motifs intersecting the Foxn1 ChIP-seq peak in each was assessed in HEK293 cells co-transfected with a *Foxn1* expression plasmid and wild-type or mutagenised promoter-luciferase constructs.

A single Foxn1 binding peak was identified upstream of *Psmb11*'s TSS that contained two copies each of the Foxn1 binding motifs GACGC and GAAGC. Mutations of the proximal and, to a lesser degree, distal GACGC motifs significantly decreased luciferase activity whereas changes to both ablated *Psmb11* transcription altogether (**Figure 30a**). This appeared to exclude a contribution from Foxn1 binding to the alternative motif, GAAGC. However, although the replacement of the proximal GACGC motif with a GAAGC sequence significantly decreased promoter activity, this was not completely ablated in contrast to GAGGC, a sequence in which the A/C third base of the motif was replaced by a G, predicted from motif analysis to be non-functional. This suggests that either GACGC or GAAGC could promote Foxn1 binding to the *Psmb11* promoter but the precise position of the motif was an important factor in determining binding affinity.

ChIP-seq identified a single Foxn1 peak in the *Cd83* promoter containing a proximal GACGC and three distal GAAGC motifs. A deletion of the single GACGC motif decreased the transcriptional activity by half and removal of the promoter sequence containing the distal two GAAGC motifs reduced transcription by two thirds (**Figure 30b**). Deleting the remaining GAAGC and GACGC motifs in the truncated promoter further reduced transcription yet did not completely abolish luciferase activity. In the case of *Cd83*, there was some evidence that both versions of the motif contributed to the transcriptional activity of this minimal *Cd83* promoter but GACGC was more functionally important.

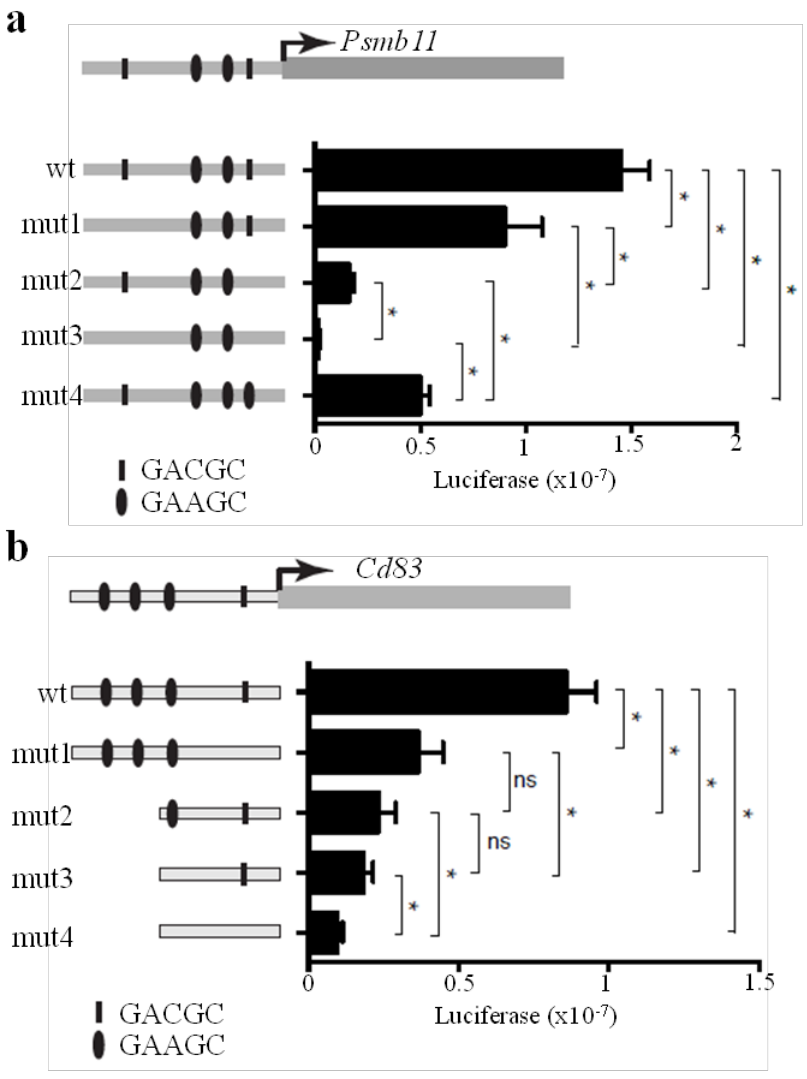

Figure 30 *Foxn1* motifs near *Psmb11* and *Cd83* are functional. (**a**) Luciferase assay for the Foxn1-dependent activation of the wild-type (wt) *Psmb11* promoter and of mutants of the *Psmb11* promoter containing alterations in the GACGC (black bars) and GAAGC motifs (black ovals) as indicated (mut1-4). The data shows mean values ±SD of luciferase activity and is representative of at least two independent experiments with a minimum of three repeats for each

reporter construct. (**b**) Luciferase assay for the Foxn1-dependent activation of the wild-type (wt) *Cd83* promoter containing a single GACGC and 3 distal GAAGC motifs and *Cd83* promoters with the indicated mutations in the GACGC and GAAGC motifs. The data shows mean values ±SD of luciferase activity and is representative of at least two independent experiments with a minimum of three repeats for each reporter construct. * = $p < 0.05$.

5.6 Discussion

Foxn1 is well known to be critical for thymic development.[104] There is accumulating evidence that it is also required for the maintenance of TEC function and thymopoiesis.[106,264] Previous knowledge of *Foxn1* targets was based on limited comparative studies of wild-type and *Foxn1* mutant mice. However, a key limitation of these results was the lack of any direct evidence of Foxn1 binding to these genes. In this chapter, I have attempted to identify high confidence *Foxn1* targets and demonstrate how these may be important for TEC function.

Foxn1 is a highly evolutionarily conserved gene consisting of a forkhead DNA binding domain in the middle of the protein and a transcriptional activation domain encoded by exons 8 and 9 at the C-terminus.[299] The use of a chimeric, full length *Foxn1* transcript extended at the C-terminus with a Flag octapeptide in the experiments presented here permitted assessment of Foxn1 binding in the presence of a fully intact DNA binding domain with the added advantage that the Flag domain could be used for chromatin immunoprecipitation.

Forkhead box (Fox) transcription factors represent a diverse set of genes evolutionarily descended from a single ancestral gene found in opisthokonts over one billion year ago.[300] This family of TFs modulate key biological processes, including development, metabolism, aging and cancer. *Foxn1* and the closely related *Foxn4* over the course of evolution have lost the ability to bind the classical Fox binding motifs.[272,301,302] *In vitro* methods indentified two highly similar candidate motifs (ACGC and GACGC[T/C/G]) mediating Foxn1 binding.[271,272] However, TF binding *in vivo* is frequently more complex, often involving interactions that are either sequence-independent or dependent on high-order, three-dimensional TF-DNA interactions.[226,233,259] By using Foxn1 ChIP-seq data from primary TEC, I was able to confirm that the *in vitro* motif also mediated Foxn1 binding *in vivo*.

Interestingly, despite differences in binding affinity to specific DNA motifs, the amino acids responsible for co-ordinating DNA interactions are highly conserved between Fox TFs.[272,300] This

suggests that it is likely that other co-factors, in combination with the epigenomic landscape of specific cell types, may determine tissue-specific Fox-mediated transcriptional control.[303–307] Currently there are no known *Foxn1* co-factors. However, motif analysis near the Foxn1 binding sites identified in this chapter has suggested several potential candidates. One of these, *Tp63*, was supported by available ChIP-chip data derived from keratinocytes, which showed significant enrichment within Foxn1 ChIP-seq peaks.[295] This represents a compelling potential co-factor, since it has a known role in the maintenance of TEC stemness.[294] It also physically interacts with Cbx4, a component of Polycomb Repressive Complex 1 (PRC1); deficiency of *Cbx4* results in gene dosage-dependent failure of TEC-mediated thymopoiesis.[308,309] This may explain some tissue-specific differences observed between wt*/-$^{nu/nu}$ and wt*/wt*$^{nu.nu}$ mice. Despite a severe deficit in thympoiesis and thymic cellularity in wt*/-$^{nu/nu}$ mice, these have a normal coat, which is lacking in *nu/nu* mice (data not shown). It is possible that either the presence of specific co-factors or differential chromatin accessibility may explain differential *Foxn1* dose requirements in different tissues.

Most previous approaches to identifying *Foxn1* targets have relied on comparing *nu/nu* to wild-type mice, with the implicit assumption that transcriptionally regulated genes are likely to be bound by Foxn1. However, I have demonstrated that some previously suggested *Foxn1* gene targets, *H2-Aa*, *H2-Ab*, *Pax1* and *Fgfr2* are neither directly bound nor regulated by Foxn1 in mature TEC.[107,268] Of course, this does not exclude the possibility that some of these genes could be regulated by *Foxn1* at an earlier developmental time-point. The identification of 450 high confidence *Foxn1* target genes, including the suspected *Foxn1* targets *Dll4*, *Cxcl12* and *Ccl25* and the novel targets *Psmb11* and *Cd83*, has provided potential molecular contexts that might explain the failure of *Foxn1*-deficient cells to initiate or support thymopoiesis.[310,311] For instance, TEC deficient for *Cd83* and *Psmb11*, respectively, reiterated defects in SP4 and SP8 thymocyte selection observed in wt*/-$^{nu/nu}$ mice.

The use of the conditional *Foxn1* knock-out model (Foxn1$^{\Delta 7,8}$) reinforced the importance of continued *Foxn1* expression in mature TEC to maintain their regular function. Activation of the Cre-LoxP system by Dox treatment in these mice with excision of both the forkhead and

transcriptional activation domains results in a 2-fold reduction in full-length *Foxn1* transcripts within 24 hours coupled with a sharp reduction in the expression of hundreds of cTEC genes ((≥ 3-fold: 406 genes; ≥ 2-fold: 1,192 genes) 72 hours later. Interestingly, this finding is seemingly at odds with recent observations suggesting a progressive decrease in the level of Foxn1 protein in mature TEC without any impairment of thymopoietic function, although this was seen in considerably older mice than were used in this chapter.[312]

5.7 Conclusion

Our findings demonstrated that *Foxn1* directly regulated the transcription of several hundred genes in cTEC, which collectively control multiple essential checkpoints throughout the process of intrathymic T-cell development. These genes are likely to form a vital hub in the maintenance of mature TEC function and thymopoiesis.

5.8 Acknowledgements

Saulius Zuklys and Saule Zhanybekova generated and bred the mouse models, performed ChIP-seq and RNA-seq experiments, luciferase assays, and FACS phenotyping. Stefano Maio conducted ATAC-seq experiments. Giuseppe Gallone produced the *Foxn1* meta-gene profile figure and provided invaluable advice on bioinformatics analysis approaches.

Chapter 6 Epigenomic mechanisms of promiscuous gene expression in thymus

6.1 Abstract

One of the most important roles of the thymus is to ensure that, through negative selection within the medulla and potentially cortex, autoreactive T-cells do not reach the periphery. This occurs by ensuring that developing T-cells are exposed to practically all genes expressed within the organism, a process called promiscuous gene expression (PGE). The cells responsible for this are mature mTEC, which express high levels of the major histocompatibility complex (mTEC^{hi}) frequently together with the transcription factor, *Aire*. This is known to target genes with high levels of the repressive chromatin mark, H3K27me3, and low levels of the promoter mark, H3K4me3. However, little is understood about how epigenomic remodelling controls PGE. In order to study the effects of the epigenome on PGE, I analysed single cell transcriptomic data from mTEC^{hi} derived either from mice with TEC-specific knock-out mutations in several key epigenomic remodelling enzymes, *Dnmt3a/b*, *Eed* and *Tdg*, or wild-type mice. Those mTEC^{hi} deficient for *Dnmt3a/b* expressed many more *Aire*-induced genes than seen in wild-type, whereas mTEC^{hi} deficient for *Eed* showed the reverse effect. There was weak evidence that *Aire*-induced gene expression was organised within chromosomal clusters, although this was unlikely to account for much of the total PGE observed in mature mTEC. Using ATAC-seq and ChIP-seq data on chromatin structure in mTEC^{hi}, I was able to expand the chromatin signature of Aire target genes to include several other chromatin modifications by constructing multiple logistic regression models. The final model showed a consistent tendency for *Aire*-induced genes to be characterised by high levels of H3K27me3 and H3K4me1, and low levels of H3K4me3 and H3K9me3. This provides strong evidence that epigenomic remodelling is critical for PGE in mature mTEC and suggests a complex histone code identifying *Aire*-induced genes. Further work will be required to uncover the immunological significance and function of epigenomic remodelling and the molecular mechanisms that recruit Aire and its co-factors to regions marked by multiple chromatin modifications.

6.2 Hypotheses

6.2.1 Remodelling of the epigenome is important for promiscuous gene expression.

6.2.2 Interfering with the process of epigenomic remodelling affects promiscuous gene expression.

6.2.3 Specific chromatin modifications are important for Aire recognition of promiscuously expressed genes.

6.3 Introduction

Promiscuous gene expression (PGE) is one of the key functions of TEC, as discussed in **Chapter 1**. This process is known to be co-ordinated by the *Autoimmune Regulator* gene (*Aire*).[109,313] *Aire* is a transcription factor, mutations of which are associated with the development of autoimmune polyendocrine syndrome type 1 (APS1) due to autoreactive T-cells surviving thymic selection.[110–112] *Aire* is mainly expressed by mTEC, although roles for *Aire* in other non-TEC thymic and extra-thymic cell populations have been identified.[314–317] At any one time, *Aire* is only expressed in a proportion of mTEC but the use of an *Aire* reporter construct demonstrated that almost all mTEC become *Aire*-positive at some point during differentiation.[313,318] How *Aire* expression is regulated is currently unclear, but is likely to involve interactions between ligands such as CD40 and RANK with their receptors, and transcription factors such as *Hipk2* with a regulatory element upstream of the *Aire* gene *locus*.[319,320]

The precise mechanism by which *Aire* orchestrates PGE is controversial. Combining ChIP-seq data on histone modifications in mTEC with RNA-seq comparisons between wild-type and *Aire* knock-out mTEC demonstrated that Aire target genes were characterised by low levels of the transcriptional activator mark H3K4me3 and high levels of the repressive mark H3K27me3.[113] This suggested that Aire regulated the transcription of genes that, at least on a population level, were characterised by repressive chromatin marks.[321–323] Purifying mTEC positive for the expression of particular tissue restricted antigens (TRAs) and analysing these cells using bulk ATAC-seq showed that the chromatin architecture surrounding the selected TRA and co-

expressed genes was more open than in mTEC negative for that TRA.[114] This supported a model of Aire activity, in which Aire bound to a repressed gene and induced local chromatin remodelling to an open configuration. The role DNA methylation plays in PGE is less clear; levels of methylation around genes was equivalent between wild-type and *Aire* knock-out mTEC but such experiments cannot directly demonstrate that interfering with DNA methylation would have no effect on PGE.[115,324] The heterogeneity of PGE between individual mTEC is such that bulk functional genomics approaches are limited, particularly in dissecting out any specific patterns of PGE regulation.

Single cell RNA-seq approaches offer a potential solution to this transcriptomic heterogeneity. A number of different studies analysed the mTEC single cell transcriptome.[113–115] Overall, these suggested that *Aire*-induced genes were expressed in clusters characterised by chromosomal bias. However, these clusters appeared to be specific to individual mice, raising the possibility that the number of mTEC analysed may simply be too low to capture the diversity of PGE in the mTEC compartment.[115]

A different approach to this problem is to iteratively sort mTEC to identify characteristic co-expression modules of TRAs. This identified six major co-expression modules, which were reported to be correlated with mTEC maturity.[325] However, this study was only able to identify cluster membership for a small minority (~16%) of TRAs, casting doubt on the contribution of their findings to overall PGE.

The epigenome is shaped by a multitude of different enzymes.[326] Repressive chromatin marks, including H3K27me3, are induced by Polycomb repressive complexes.[327] A key member of Polycomb repressive complex 2 is *Eed*.[328] This is known to be important in thymocyte maturation.[329] DNA methylation is regulated by the balance of multiple DNA methylases and demethylases. Two key DNA methylases involved in *de novo* DNA methylation are *Dnmt3a* and *Dnmt3b*.[330,331] Conversely, *Tdg* functions to return modified cytosine residues within CpG islands to a non-methylated state.[332] Given the profound importance of epigenomic regulation to development, studying the contribution of these genes to TEC development will require the generation of conditional knock-out models.[333]

I aimed to evaluate the importance of epigenomic remodelling in PGE by analysing mTEC sorted from mouse models with conditional knock-out of key enzymes involved in shaping the epigenome using single cell RNA-seq. The intersection of these findings with measures of chromatin accessibility from ATAC-seq and histone modificiations from ChIP-seq can provide further insight into the epigenomic landscape underlying PGE.

6.4 Methods

6.4.1 Mouse models

Mice carrying LoxP sites upstream of *Eed* exon 3 and downstream of exon 6 were crossed with β5t::Cre mice. I refer to these as Eed^{Δ} mice.[328]

Mice carrying LoxP sites flanking of *Dnmt3a* exon 19 and *Dnmt3b* exons 17 to 20 were crossed with Foxn1::Cre mice. I refer to these as $Dnmt3a/b^{\Delta}$ mice.[330,331]

Mice carrying LoxP sites flanking a stably integrated *Tdg* mini-gene on a knock-out background caused by cloning a neomycin cassette into the endogenous *Tdg* locus were crossed with Foxn1::Cre mice.[332] I refer to these as Tdg^{Δ} mice.

In each conditional knock-out model, the catalytically active exons were targeted for excision. C57BL/6 mice were used as wild-type controls.

6.4.2 FACS isolation of mTEC^{hi}

Thymi were isolated from 4 week old mice. Thymic lobes were digested in PBS containing 2.5mg/mL Liberase TM (Roche Diagnostics), 10mg/mL DNase I (Roche Diagnostics) for 15-20 minutes at 37ºC. TECs were enriched using AutoMACS (Miltenyi Biotec) and stained for the expression of epithelial cell adhesion molecule (EpCAM) (G8.8, eBioscience), CD45 (eBioscience), MHC class II (BioLegend), Ly51 (ebioscience), and UEA-1. Flow cytometric analysis and cell sorting were performed using FACSAria III and FACSDiva (BD Biosciences). Cells were sorted into 2.3µl of lysis buffer consisting of 0.2% triton plus RNase inhibitor (Takara).

6.4.3 Single cell RNA-seq

Reverse transcription, amplification and library preparation was performed according to the SMART-seq2 protocol incorporating 0.1µl of 1:250,000 ERCC spike-in mix 1 (Ambion) per well.[32] The resultant libraries from each plate were multiplexed and sequenced on Illumina HiSeq 2500 sequencing machines. Each plate was run down two separate lanes.

6.4.4 Single cell RNA-seq analysis pipeline

CRAM files were uploaded from the Sanger EBI servers and md5 checksums were calculated for all files to ensure no corruption had occurred on transfer. Single cell RNA-seq data was available on 1,440 single cells (384 *Dnmt3a/b*$^{\Delta}$, 384 *Eed*$^{\Delta}$, 384 *Tdg*$^{\Delta}$ and 288 wild-type mTEC^{hi}). FASTQ files were extracted from CRAM files using the CRAM toolkit (version 2.1 with the arguments "fastq -I <cram file> -F <reference fasta> --gzip --reverse").[334,335] Quality metrics were derived from FASTQ files using FastQC.[179] 125 base-pair paired-end reads were aligned to the GRChm38 top level DNA features (current on 15th June 2015) and ERCC spike-ins using HISAT two-pass alignment (version 0.1.6 with the arguments; first pass: "-p 8 --known-splicesite-infile <Ensembl spicesites> --novel-splicesite-outfile <novel splicesites> -x <genome index> -1 <forward reads> -2 <reverse reads>" and second pass: "-p 8 --known-splicesite-infile <Ensembl spicesites> --novel-splicesite-infile <novel splicesites> -x <genome index> -1 <forward reads> -2 <reverse reads>".[180] Read pairs in which both reads were uniquely mapped were assigned to genes using HTSeq (htseq-count with the arguments "-s no -m intersection-nonempty").[181]

Cells were eliminated from further analysis if the proportion of aligned reads was < 75%; if the number of detected genes was < 2,000 or > 10,000; if the proportion of fragments mapping to ERCC spike-ins was > 75% or < 1%; if the proportion of fragments mapping to mitochondrial genes was either an outlier or > 20%; if the proportion of fragments mapping to rRNA was an outlier; if the cell was classified as an outlier by PCA on raw counts; if the GC content of the library was < 40% or > 60%; or if the 5' to 3' bias was ≤ 0.2 (as assessed by Picard tools CollectRnaSeqMetrics). Outlier analysis was conducted in R using the boxplot function. PCA outliers were those where the ratio of median absolute deviation to median distance from the plate

centroid was > 2. Sequencing samples surviving quality control were combined by well identity. The proportion of fragments was estimated as $\Sigma y_{1\rightarrow j} / (\Sigma y_{1\rightarrow j} + \Sigma x_{1\rightarrow i})$, where y is the number of reads in the gene set of interest (e.g. ERCC spike-ins), x is the number of reads in the set of protein-coding genes, j is the number of genes in the gene set of interest and i is the number of protein-coding genes.

Size-factors were calculated for each cell surviving quality control on protein-coding genes and ERCC spike-ins using DESeq as described by Brennecke *et al.*[129,147] Counts were adjusted by two-factor correction: $([x_{1...i}] / sf_x) / sf_{ERCC}$, where sf_x is the size-factor of protein-coding genes and sf_{ERCC} is the size-factor of ERCC spike-ins. Counts were adjusted for cell cycle and *Aire* effects using single cell latent variable modelling (scLVM).[336] Exon-level coverage was assessed by using BEDtools with the arguments "-split -abam" with adjustment for protein-coding transcriptome size and exon size in each cell.[182]

Counts were converted into the absolute number of molecules by linear modelling of the known number of molecules of each ERCC spike-in present against the detected FPKM (with the fitted line forced through the origin).

Analysis of the number of detectable genes in different numbers of cells was conducted by randomly sampling either 100 combinations of cells with replacement at each number of cells (apart from 1 cell - at which point all cells were analysed separately). Empirical p-values were calculated by permuting the genotype label of cells 1,000 times. Multiple different FPKM (0, 1 and 10) and absolute number of molecules (1, 10 and 100) thresholds were used.

Clustering was conducted using PCA and t-SNE (implemented in the tsne R package).[337] In the case of clustering, genes were included in the model only if the standard deviation was > 0 and the gene was expressed in > 20% cells.

Differential expression was analysed using the SCDE package on genes with standard deviation > 0 and detectable at > 10 reads per cell.[203,338] The level of significance was set at FDR < 0.05.

Coexpression analysis was conducted on adjusted counts matrices using Spearman correlation. Genes were clustered by k-medoids with the R package "cluster".[339] The ideal number of clusters

was estimated by iteratively dividing genes into between 2 and 200 clusters using the function "pam". A Loess fit was used to model the trend in average dissimilarity between clusters with increasing numbers of clusters. The optimal number of clusters was set at the point at which the rate of average dissimilarity decrease fell to 50% of maximum. Clusters were then merged if the median Spearman correlation between clusters was ≥ 0.2. Intrachromosomal interactions were defined as those in which at least one of the maximally co-expressed genes within a cluster was located on the same chromosome as the gene-of-interest. Interchromosomal interactions were defined as those genes lacking a maximally co-expressed gene within the same cluster. The significance of intrachromomsal interactions was estimated by resampling the chromosomal location of all genes 1,000 times and calculating a one-way empirical p-value.

Gene ontology enrichment was assessed using DAVID with all mouse genes as a background.[149]

Single cell RNA-seq data was compared to previous datasets reported by Sansom and colleagues.[113] *Aire*-dependent, *Aire*-enhanced and tissue restricted antigen (TRA) gene identity were also obtained from that study. *Aire*-dependent genes were those undetectable in *Aire* deficient mTEC and *Aire*-enhanced genes were those with significantly increased expression in *Aire* positive relative to *Aire* deficient mTEC. Sets of genes were matched on proportional expression by binning expression values for wild-type mTEC^{hi} in 10% proportional expression windows and then resampling sets of genes to match the proportional expression profile of *Aire*-induced TRAs.

6.4.5 ATAC-seq

Two replicates of ~25,000 wild-type mTEC^{hi} sorted in the same manner as described above underwent lysis, tagmentation and PCR amplification as described in the ATAC-seq protocol.[287] ATAC-seq libraries were sequenced on an Illumina HiSeq 2500.

6.4.6 ATAC-seq analysis

I used FastQC to assess read quality and Trimmomatic to remove contaminating sequences (transpons or their reverse complements), crop the first and last 3 bases of each read based on

sequencing quality, and remove the 3' 10 bases of each read to remove partial transposon sequences.[144] I used Bowtie2 (version 2.2.3) to align 100 base-pair paired-end reads against the UCSC mm10 genome with the arguments "--no-mixed --no-discordant -X 2000" as in a previous study.[8] Reads were filtered to obtain concordantly mapping read pairs, in which at least one read mapped uniquely (without an XS:i BAM field). Picard Tools was used to remove duplicate fragments. The position of reads were passed into BEDtools and remapped taking into account transposon sequence insertion bias.[182] Fragment length distribution was plotted as a histogram and proportional coverage across the genome was estimated using 1kb windows. Peaks were called using MACS2 (version 2.0.10.20131028) with the arguments "--nomodel --nolambda --keep-dup all --call-summits" as in a previous study.[36,279] Peaks were filtered against the ENCODE mm10 blacklist and a set of mitochondrial pseudopeaks generated from 1,000,000 in silico 100 single-end reads produced from mitochondrial DNA aligned against non-mitochondrial DNA.[231,237] Enrichment of ATAC-seq peaks within genomic intervals was assessed using GAT with 10,000 randomisations, correcting for GC isochores.[240]

6.4.7 Histone ChIP-seq

Sorted wild-type mTEC^{hi} were pelleted and cross-linked in preparation for chromatin immunoprecipitation as previously described with the following modifications[340]: Protein G magnetic beads (Dynabeads, Life Technologies) were used to capture antibody-chromatin complexes. These complexes were washed four times with wash buffer (0.1 % SDS, 1 % Triton X-100, 0.1 % deoxycholate, 20 mM Tris-HCl (pH 8), 1 mM EDTA, 0.5 mM EGTA) containing 150 mM NaCl, once again with wash buffer containing 500 mM NaCl, and then once with LiCl wash buffer (250 mM LiCl, 1 % NP-40, 1% deoxycholate, 10 mM Tris- HCl (pH 8), 1 mM EDTA).[113] Antibodies directed against multiple histone marks were used: H3K4me3 diagenode #C15410003, H3K27me3 millipore #07-449, H3K4me1 abcam #ab8895, H3K9ac abcam #ab4441, H3K4ac millipore #07-539, H3K9me3 millipore #05-1242 and H3K27ac abcam #ab4729. Input samples were prepared in parallel. Libraries were prepared using the TruSeq ChIP

Preparation Kit (Illumina). Libraries were sequenced on Illumina HiSeq 2500 sequencing machines.

6.4.8 Histone ChIP-seq analysis

I used FastQC to assess read quality and Trimmomatic to remove adapter sequences (transpons or their reverse complements), crop the first and last 3 bases of each read based on sequencing quality, crop sequences based on a sliding window (4:15), and retain reads with a minimum length of 20 bases.[144] I used BWA (version 0.7.12) for pre-alignment of 100 base-pair paired-end reads against the UCSC mm10 genome with the arguments "bwa aln -t8 -q10 <forward/reverse reads>" and "bwa sampe <forward sai> <reverse sai>".[277] Pre-aligned bam files were further aligned with Stampy (version 1.0.23) with the arguments "-t 8 --process-part=n/10 --bamkeepgoodreads".[278] Reads were filtered to obtain concordantly mapping read pairs, in which at least one read mapped uniquely. Picard Tools was used to remove duplicate fragments. Peaks for punctuate marks (H3K4me1, H3K4ac and H3K9ac) were called using MACS2 (version 2.0.10.20131028) with the arguments "--keep-dup all" using pooled input samples as a control.[279] Peaks were called for broad marks (H3K9me3 and H3K27me3) using MACS2 with the arguments "--keep-dup all --broad". Peaks were filtered against the ENCODE mm10 blacklist.[231,237] Similarity between samples was assessed by Spearman's correlation of read coverage of regions within 1kb of protein-coding gene TSS. ChIP-seq signal for each gene was considered the maximum ratio between read coverage of regions within 1kb of protein-coding gene TSS in ChIP and input samples. Enrichment of ChIP-seq peaks within genomic intervals was assessed using GAT with 10,000 randomisations, correcting for GC isochores.[240]

6.4.9 Irreproducibility discovery rate (IDR) analysis

IDR analysis was undertaken as described by Kundaje.[230,282] Consistency between replicates and pseudoreplicates was estimated for peaks called with MACS2 as above using the additional arguments "-p 1e-1" and "--to-large" with ChIP-seq datasets. Peaks were called significant at IDR

< 0.01 and were filtered against the ENCODE mm10 blacklist (plus mitochondrial pseudopeaks in the case of ATAC-seq peaks).[231,237]

6.4.10 Multiple stepwise logistic regression modelling

ATAC-seq and ChIP-seq data was compiled either as a matrix where 0 indicated no binding site and 1 indicated a binding site intersecting the TSS of a gene +/- 1kb or as a matrix of log_2 transformed maximum signal intensity near the TSS of a gene +/- 1kb. The status of a gene as *Aire*-induced was encoded as 1 and *Aire*-independence encoded as 0. Multiple logical regression analysis was undertaken in R using the function glm (<*Aire* status> ~ <Σ ATAC/ChIP signal/peak>, family = binomial). Stepwise regression used step (direction = backward) on the fitted output of glm. Variance inflation factors (VIF) were calculated on independent variables and variables were then dropped in order to minimise Akaike Information Criterion (AIC) until all VIF scores were < 4.

6.5 Results

6.5.1 Single cell RNA-seq data on mTEC^{hi} accurately reflect mature mTEC

I removed mTEC^{hi} single cells that failed quality control checks on alignment rate, number of detected genes, ERCC spike-in proportion, mitochondrial reads, ribosomal reads, GC content, 5' to 3' bias and PCA outlier status for raw counts by plate. 899 out of 1,440 cells (62.4%; 186 *Dnmt3a/b*$^{\Delta}$, 265 *Eed*$^{\Delta}$, 263 *Tdg*$^{\Delta}$ and 185 wild-type cells) passed all quality control filters (**Figure 31a**). One plate in particular, *Dnmt3a*$^{\Delta}$ plate four, entirely failed quality control with extremely low alignment rates, low GC content and extreme 5' to 3' bias (excluding this plate, 66.9% of cells passed quality control). The detection of ERCC spike-ins showed a clear dependence on ERCC spike-in concentration (median Spearman's r > 0.84 for all genotypes).

I generated gene expression data from the remaining, high quality cells. In order to establish whether these reflected a relatively unbiased sample of mTEC^{hi} cells, I compared gene expression values obtained from this study with previous measures of bulk and single cell mTEC^{hi} transcriptomes.[113] The mean FPKM estimated from wild-type mTEC^{hi} single cell transcriptomes

correlated well with bulk FPKM from mature mTEC (Spearman's r = 0.87; **Figure 31b**). The proportions of mTEC^{hi} with detectable expression of genes (FPKM > 0) were also similar between the current dataset and a previous single cell RNA-seq study of mTEC^{hi} (Spearman's r = 0.95; **Figure 31c**).[113] As expected for highly variable, lowly expressed genes, the correlation between single cell datasets was lower for *Aire*-induced than *Aire*-independent genes (*Aire*-dependent: Spearman's r = 0.44; *Aire*-enhanced: Spearman's r = 0.73; *Aire*-independent TRA: Spearman's r = 0.88).

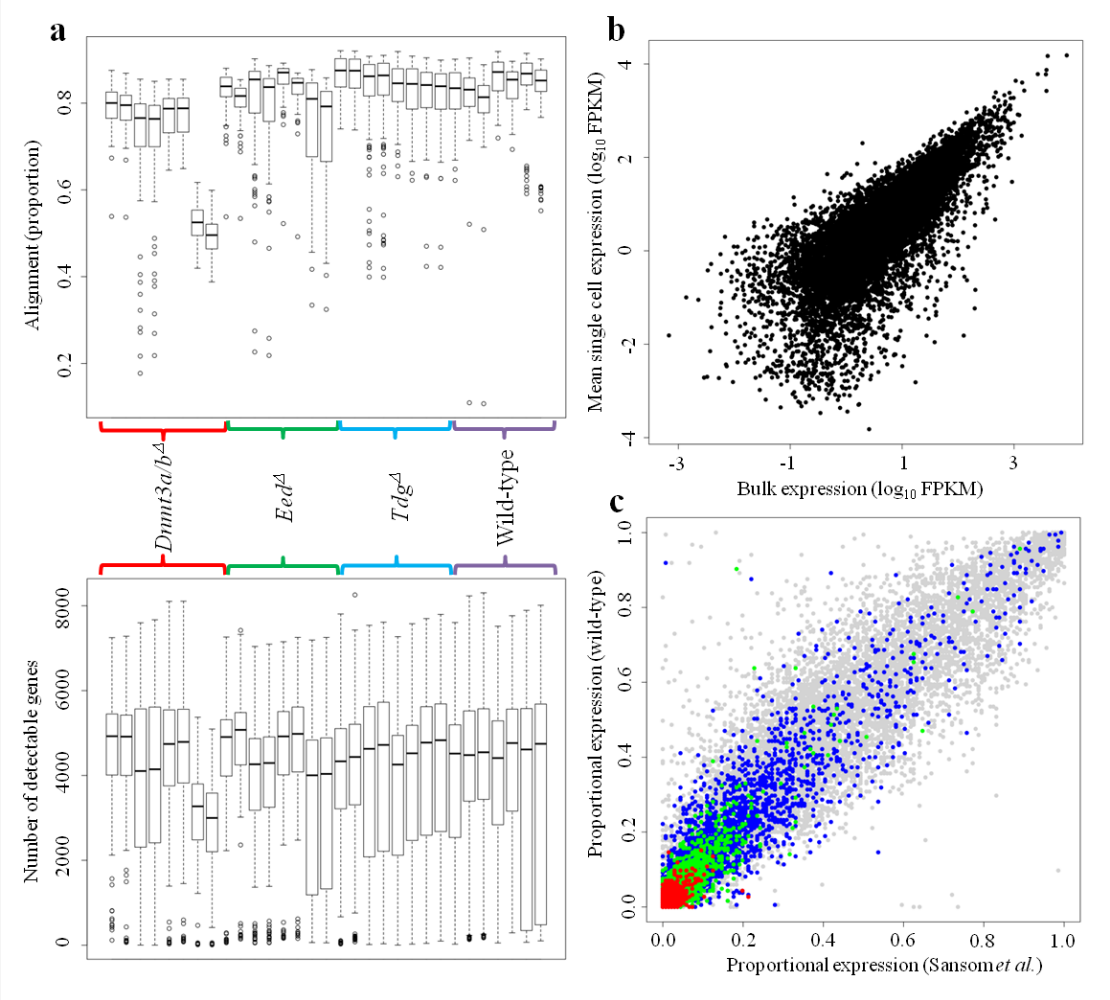

Figure 31 Quality assessment of single cell RNA-seq data. (**a**) The proportion of aligned reads (top) and number of detectable genes (bottom) by genotype. (**b**) Wild-type mean single cell expression values plotted against bulk expression values from mature mTEC. (**c**) Proportional single cell expression from an earlier publication (Sansom *et al.*) against the current dataset.[113] Colours indicate *Aire*-dependent genes (red), *Aire*-enhanced genes (green), *Aire*-independent TRAs (blue) and all other protein-coding genes (gray).

6.5.2 Assessment of Cre-Lox excision efficiency in single mTEC^{hi}

Having established that this dataset accurately reflected the mature mTEC transcriptome, I assessed the effectiveness of exon excision (**Figure 32**). Exon-level analysis is difficult in RNA-seq due to unpredictable biases in exon coverage, a problem which is particularly pronounced in single cell RNA-seq.[341] However, with these caveats, it was clear that *Dnmt3a* and *Tdg* excision had been successful in most cells even at the level of gene expression metrics (**Table 7**). *Dnmt3b* exon-level coverage demonstrated a clear reduction of coverage over exons 17 to 20 in *Dnmt3a/b*$^{\Delta}$ relative to wild-type mTEC^{hi}. The *Eed*$^{\Delta}$ model showed no convincing evidence either of reduced *Eed* expression or the expected pattern of exon excision.

Table 7 The differential expression of genes in conditional knock-out mTEC^{hi}. In each case, mean expression values and non-zero proportional expression are shown for the relevant conditional knock-out model in comparison with wild-type mTEC^{hi}.

Gene	Conditional knock-out		Wild-type		P-value
	Mean FPKM	Proportion	Mean FPKM	Proportion	
Dnmt3a	19.6	0.33	78.8	0.73	$< 10^{-8}$
Dnmt3b	12.1	0.15	17.6	0.16	0.27
Eed	35.5	0.54	58.5	0.53	0.22
Tdg	5.2	0.35	82.7	0.62	$< 10^{-12}$

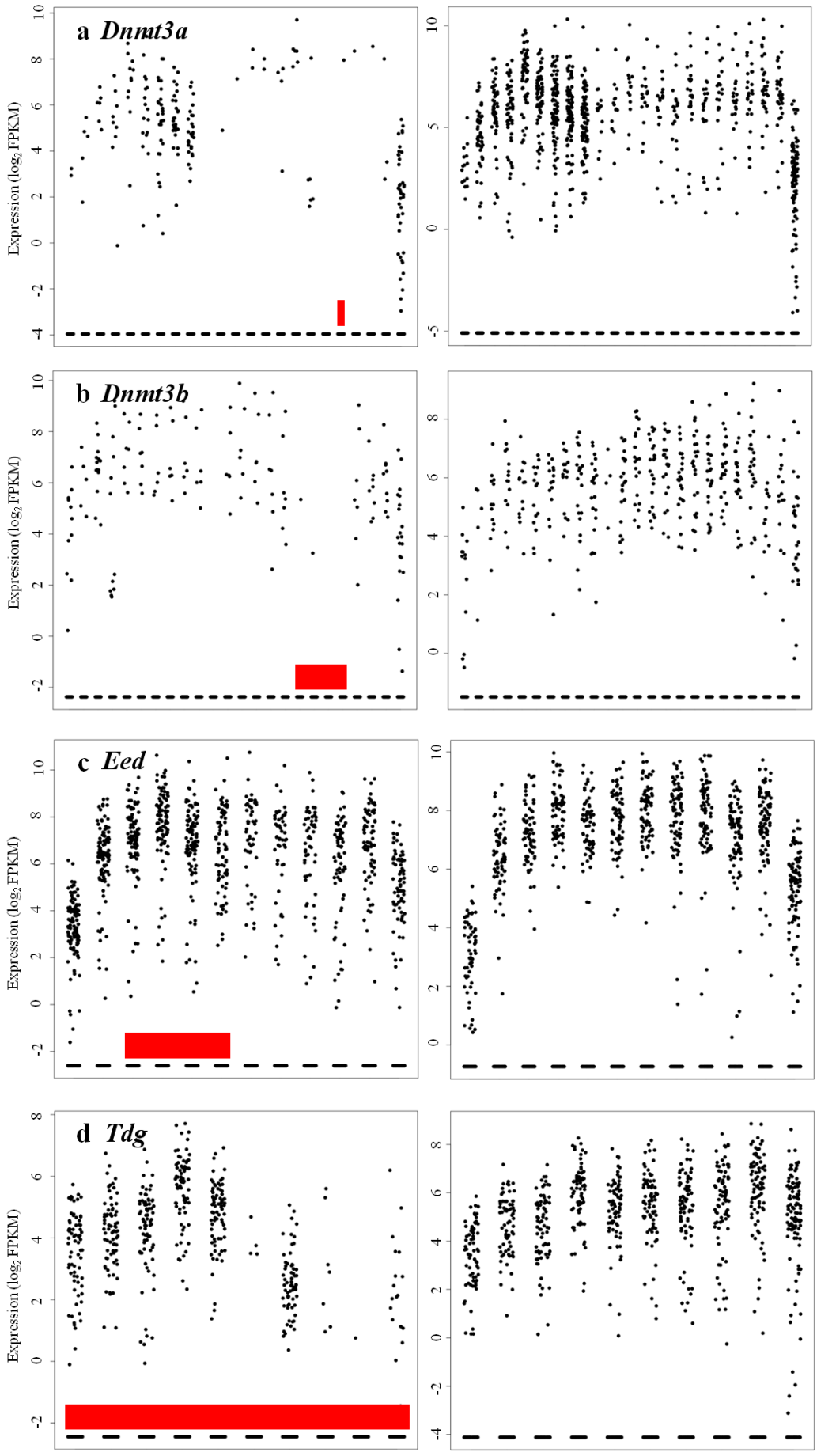
a Dnmt3a
b Dnmt3b
c Eed
d Tdg
Expression (log2 FPKM)

Figure 32 Evidence of exon excision in conditional knock-out models. (**a**) *Dnmt3a*, (**b**) *Dnmt3b*, (**c**) *Eed* and (**d**) *Tdg*. Each point represents expression of an exon in an individual cell for the appropriate conditional knock-out model (left) and wild-type (right). The red bar indicates exons flanked by LoxP sites.

6.5.3 Differences between conditional knock-out models and wild-type mTEC^{hi}

I assessed whether there were high level differences apparent between wild-type and conditional knock-out models by clustering analysis (**Figure 33a&b**). Most mTEC^{hi} clustered closely together regardless of genotype but there was a clear subcluster predominantly consisting of *Eed*$^{\Delta}$ cells. This subclustering was mainly driven by *Aire*-induced genes as, when expression values were corrected for cell cycle and *Aire*-induced genes, the subcluster disappeared (**Figure 33c**). This was despite that fact that cell cycle and *Aire*-induced genes contributed little overall to transcriptomic variance (**Figure 33d&e**).

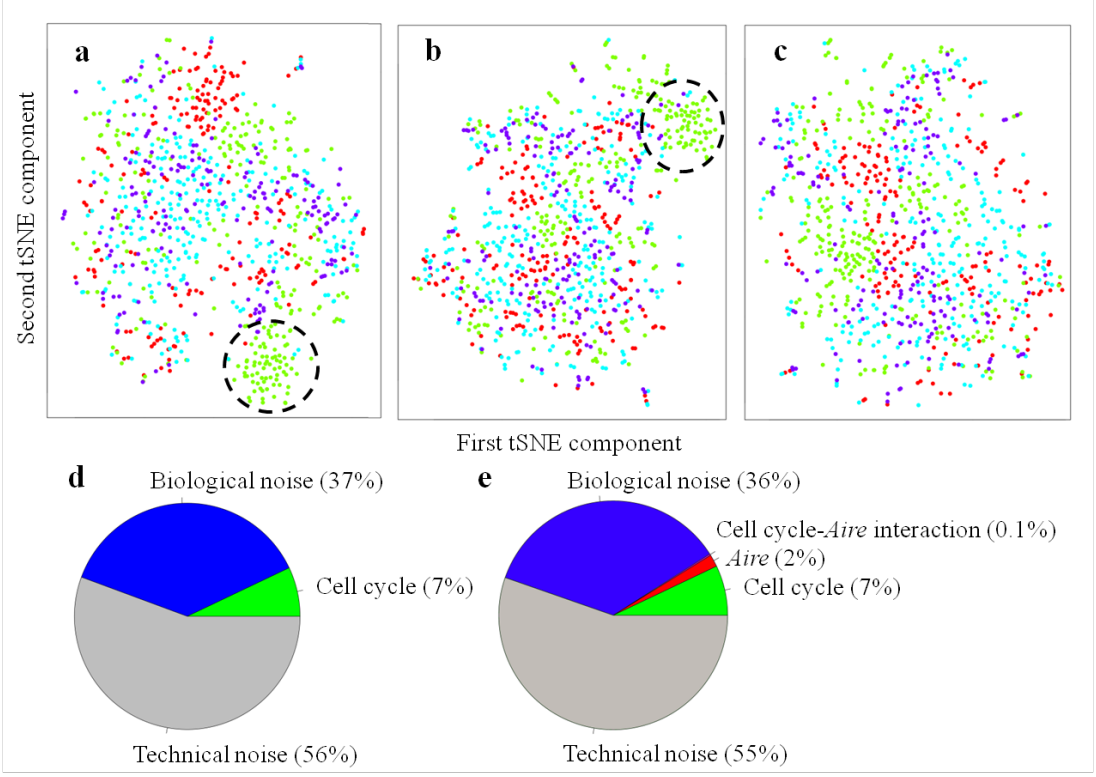

Figure 33 Clustering analysis of single mTEC^{hi}. (**a**) Clustering with all protein-coding genes. (**b**) Clustering after adjusting for cell cycle effects. (**c**) Clustering after adjusting for cell cycle effects and *Aire*-induced genes. Genotype is indicted by colour: *Dnmt3a/b*$^{\Delta}$ (red), *Eed*$^{\Delta}$ (green), *Tdg*$^{\Delta}$ (blue) and wild-type (purple). Where visible, the distinct subcluster of *Eed*$^{\Delta}$ is indicated by a dashed circle. The estimated contribution of latent variables to total transcriptomic variance is shown when adjusting for either (**d**) cell cycle or (**e**) both cell cycle and *Aire*-induced genes.

Analysis of differentially expressed genes at the level of individual transcriptomes identified several differences between wild-type mTEC^{hi} and the various epigenomic conditional knock-out models after controlling for cell cycle effects (**Table 8**). Interestingly, some of the differences between *Dnmt3a/b*$^{\Delta}$ and wild-type mTEC^{hi} were driven by higher expression of genes containing HIN-200 domains in wild-type cells (*Mnda*, *Ifi203*, *Ifi205*, *Aim2*). This class of genes is known to play a role in the regulation of autoimmunity.[342] In the case of *Eed*$^{\Delta}$ mTEC^{hi}, the major differences were driven by higher expression of genes involved in epithelial development (*e.g. Fgfr2* and *Bmpr1a*), suggesting that those cells may be more immature that those from wild-type mice. *Tdg*$^{\Delta}$ mTEC^{hi} were highly similar to wild-type mTEC^{hi}, differing only in a small number of genes enriched within pathways associated with extracellular space (*Grp*, *Gc*, *Cxcl3*, *Il1rn* and *Serpinb2*).

Table 8 The number of differentially expressed genes between single mTEC^{hi} from epigenomic conditional knock-out and wild-type mice.

	***Dnmt3a/b*$^{\Delta}$ *vs.* wild-type**	***Eed*$^{\Delta}$ *vs.* wild-type**	***Tdg*$^{\Delta}$ *vs.* wild-type**
Increased in wild-type	52	89	4
Decreased in wild-type	51	716	21

6.5.4 Interfering with epigenomic remodelling impacts on *Aire*-induced gene expression

Having established that mTEC^{hi} from epigenomic conditional knock-out mice differed significantly from those from wild-type mice, particularly in the case of *Dnmt3a/b*$^{\Delta}$ and *Eed*$^{\Delta}$, I assessed whether this was likely to have an impact on antigen presentation in the thymus. Intuitively interfering with epigenomic remodelling could affect the overall proportion of TRAs expressed by mTEC^{hi}. There was a clear effect of genotype on the expression of *Aire*-induced genes with a less marked effect on other protein-coding genes (**Table 9**).

Table 9 The difference in proportional expression between conditional knock-out and wild-type mTEC^{hi} for categories of *Aire*-induced and *Aire*-independent genes. A gene was considered expressed if it was detected at > 1 FPKM. Fold change is the ratio between median proportional expression in conditional knock-out and wild-type cells. P-values were adjusted for multiple hypothesis testing using the Benjamini-Hochberg method.

Gene category (number of genes expressed in wild-type)	***Dnmt3a/b*$^{\Delta}$ *vs.* wild-type**		***Eed*$^{\Delta}$ *vs.* wild-type**		***Tdg*$^{\Delta}$ *vs.* wild-type**	
	Fold change	**P-value**	**Fold change**	**P-value**	**Fold change**	**P-value**
***Aire*-induced TRA** (61)	1.62 (99)	$< 10^{-5}$	0.74 (45)	< 0.01	0.98 (60)	0.99
***Aire*-induced non-TRA** (47)	1.35 (63.5)	$< 10^{-5}$	1.10 (52)	0.52	0.93 (44)	0.52
***Aire*-independent TRA** (503)	1.01 (506)	0.92	0.87 (438)	$< 10^{-7}$	0.97 (490)	0.20
All other protein-coding genes (3,898)	0.98 (3,831)	0.12	0.95 (3,706)	< 0.01	0.96 (3,754)	< 0.05

To model the potential effect of this on TRA presentation within the thymic medulla, I resampled increasing numbers of mTEC^{hi} single cell transcriptomes from each of conditional knock-outs and wild-type mice. There was significantly higher numbers of *Aire*-induced TRAs and *Aire*-induced non-TRAs in *Dnmt3a/b*$^{\Delta}$ compared with wild-type mTEC^{hi} (**Figure 34a&b**), while *Aire*-independent genes were unaffected (**Figure 34c&d**). The reverse effect on *Aire*-induced genes was seen in *Eed*$^{\Delta}$. The representation of *Aire*-induced genes was very similar between *Tdg*$^{\Delta}$ and wild-type mTEC^{hi}. The effect on *Aire*-induced genes was robust with different expression thresholds, based either on FPKM or estimated absolute molecule counts, and was clear even when analysed at the level of individual PCR plates. A proportion of this effect could be due to the fact that protein-coding genes in general cover a larger range of proportional expression values than *Aire*-induced genes, meaning that lowly expressed, *Aire*-induced genes are more highly affected by stochastic detection differences. This is unlikely to be the case because when I matched categories of genes based on proportional expression bins of 10% size in wild-type mTEC^{hi}, the difference in *Aire*-induced gene detection remained significant.

Aire expression was unlikely to be responsible for this effect, as *Aire* mRNA abundance did not differ significantly between wild-type and conditional knock-out mTEC^{hi} ($p > 0.05$ for all) and the removal of *Aire* negative cells prior to analysis did not change the results of this analysis.

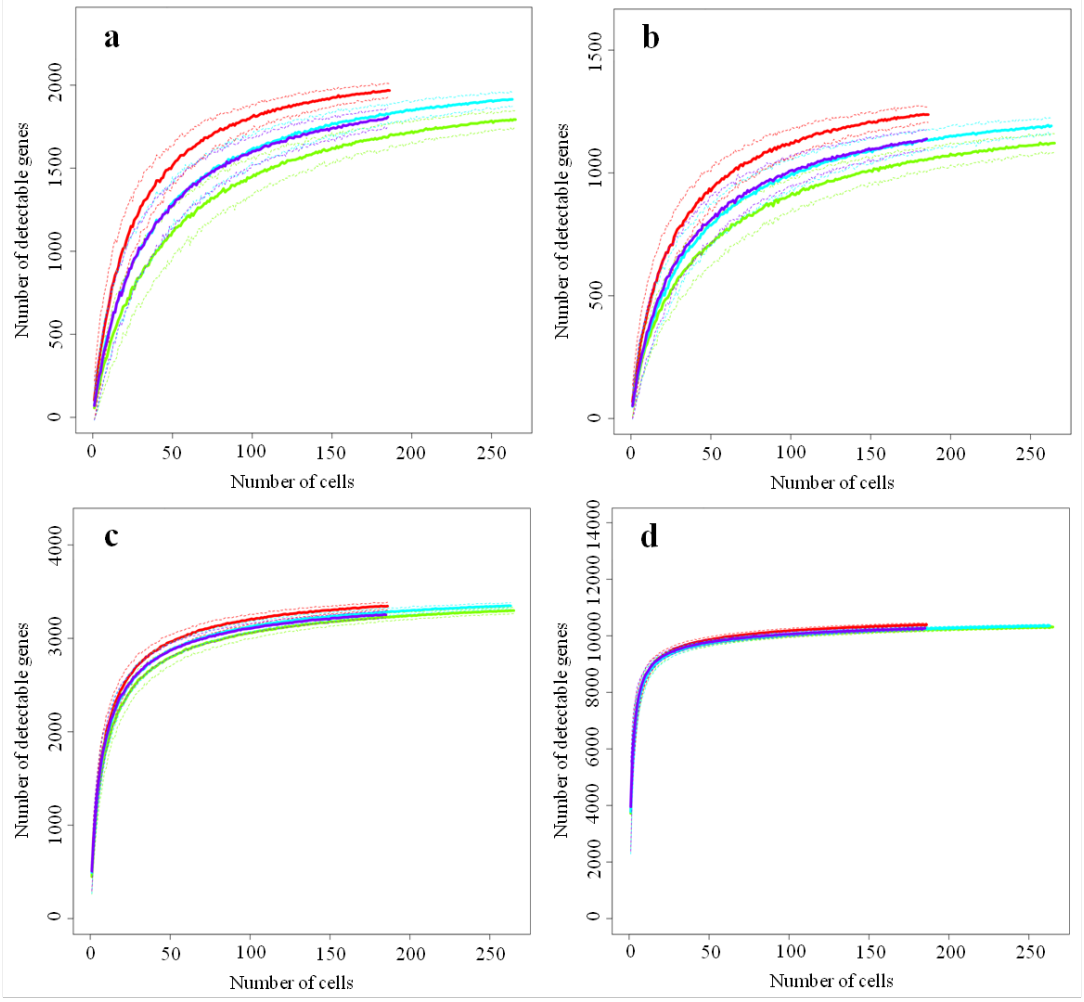

Figure 34 Total number of detectable genes for different numbers of mTEC^{hi} pooled. (**a**) *Aire*-induced TRAs, (**b**) *Aire*-induced non-TRAs, (**c**) *Aire*-independent TRAs and (**d**) all other protein-coding genes. A detection cut-off of FPKM > 1 was used. Genotype is indicted by colour: *Dnmt3a/b*$^{\Delta}$ (red), *Eed*$^{\Delta}$ (green), *Tdg*$^{\Delta}$ (blue) and wild-type (purple). Dashed lines indicate 95% confidence intervals estimated from 100 resampled sets of cells.

6.5.5 Evidence for intrachromosomal co-expression for *Aire*-induced genes

Previous studies have reported that co-expressed *Aire*-induced genes tended to cluster in intrachromosomal domains.[115] I performed co-expression analysis of single cell RNA-seq data from wild-type mTEC^{hi} to assess whether this was also seen in this dataset. There were clear co-expression clusters of *Aire*-induced TRAs (**Figure 35a**). When I analysed the proportion of genes in each cluster for which at least one of the top co-expressed genes in the same cluster was

located on the same chromosome, no clusters had significantly more intrachromosomal interactions over a permuted background after correction for multiple hypothesis testing (**Figure 35b**). However, when co-expression interactions for all clusters were considered together, there was a significant enrichment for intrachromosomal interactions (observed = 229, expected = 117.0, $p < 0.001$), but this only accounted for a minority of total *Aire*-induced TRAs (~9.6%).

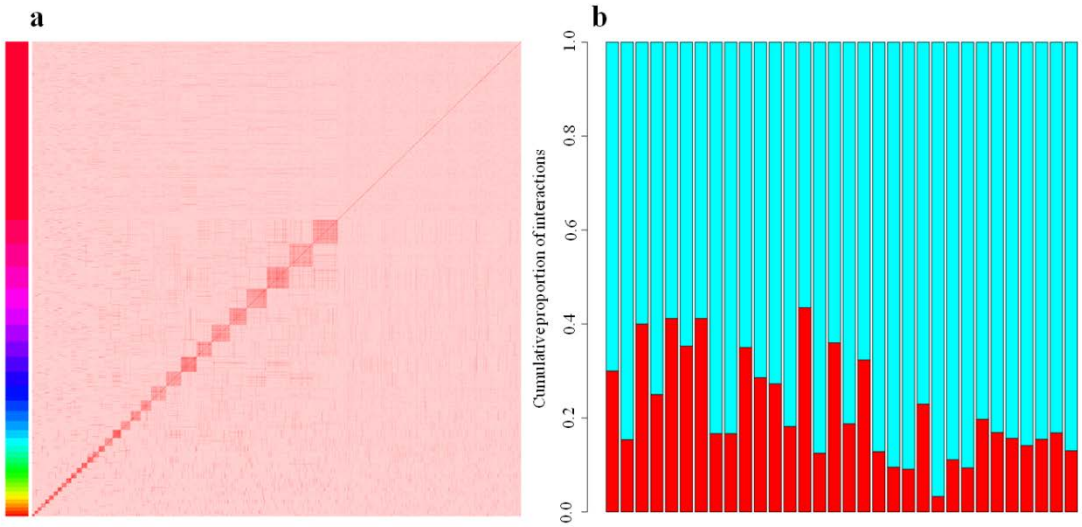

Figure 35 Intrachromosomal interactions amongst co-expression clusters of *Aire*-induced TRAs. (**a**) Co-expression heatmap of *Aire*-induced TRAs in wild-type mTEC^{hi}. The intensity of red indicates the magnitude of correlation. The vertical bar is coloured by cluster membership. (**b**) The cumulative proportion of genes for which at least one top co-expressed cluster member is located on the same chromosome (red) or for which none of the top co-expressed cluster members are on the same chromosome (blue). Individual clusters (y-axis) are ordered by size from smallest to largest.

6.5.6 Chromatin accessibility is weakly predictive of *Aire*-induced gene expression

ATAC-seq was used to delineate regions of the genome in mTEC^{hi} that were accessible. Transposase-accessible DNA was strongly enriched upstream of the TSS of genes (**Figure 36a**). I analysed chromatin accessibility and proportional expression in wild-type mTEC^{hi}.

For protein-coding genes as a whole, ATAC-seq signal near the TSS of genes was strongly correlated with the proportion of single cells with detectable expression (Spearman's $r = 0.70$, $p < 10^{-20}$; **Figure 36b**). However, the correlation was much weaker for *Aire*-dependent (Spearman's $r = 0.16$, $p < 10^{-9}$) or *Aire*-enhanced (Spearman's $r = 0.19$, $p < 10^{-19}$) genes but still reasonably strong for *Aire*-independent TRAs (Spearman's $r = 0.51$, $p < 10^{-20}$).

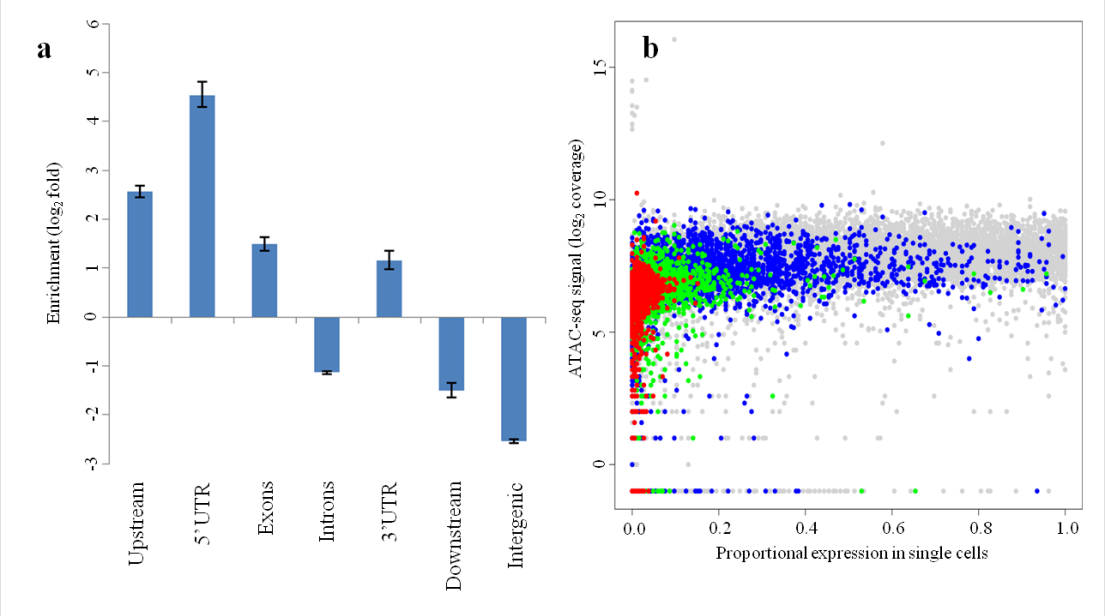

Figure 36 ATAC-seq and proportional expression in mTEC^{hi}. (**a**) Enrichment of mTEC^{hi} ATAC-seq peaks (IDR < 0.01) within different meta-gene regions. Error bars represent 95% confidence intervals. (**b**) Scatter plot of ATAC-seq signal against proportional expression in wild-type single mTEC^{hi}. Colours represent *Aire*-dependent genes (red), *Aire*-enhanced genes (green), *Aire*-independent TRAs (blue) and all other protein-coding genes (gray)

The same trend was seen when intersecting significant ATAC-seq peaks with genes. The presence of an ATAC-seq peak near the TSS of a gene was associated with higher proportional expression, particularly for *Aire*-independent genes (fold-change between median proportional expression for genes with *vs.* genes without an ATAC-seq peak: *Aire*-enhanced TRAs: 2.0-fold, $p < 10^{-7}$; *Aire*-enhanced non-TRAs: 2.8-fold, $p < 10^{-8}$; *Aire*-independent TRAs: 2.7-fold, $p < 10^{-20}$; all other protein-coding genes: 4.7-fold, $p < 10^{-20}$). When each category of genes was matched for wild-type proportional expression of *Aire*-induced TRAs in 10% expression bins as described in **6.5.4**, there was markedly less difference in proportional expression between genes with or without ATAC-seq peaks (*Aire*-enhanced TRAs: 2.0-fold, $p < 10^{-7}$; *Aire*-enhanced non-TRAs: 3.2-fold, $p < 10^{-6}$; *Aire*-independent TRAs: 2.0-fold, $p < 10^{-20}$; all other protein-coding genes: 3.3-fold, $p < 10^{-20}$).

When genes were matched for proportional expression as described above, it was clear that much of the observed difference in accessibility resulted from the degree of gene expression rather than differences between *Aire*-responsiveness (**Figure 37**). In fact, the accessibility of *Aire*-

independent non-TRAs matched for proportional expression was no different to that of *Aire*-induced TRAs ($p = 0.35$) or non-TRAs ($p = 0.53$), whereas *Aire*-independent TRAs did show higher accessibility than *Aire*-induced TRAs ($p < 10^{-10}$), *Aire*-induced non-TRAs ($p < 10^{-9}$) and *Aire*-independent non-TRAs ($p < 10^{-6}$).

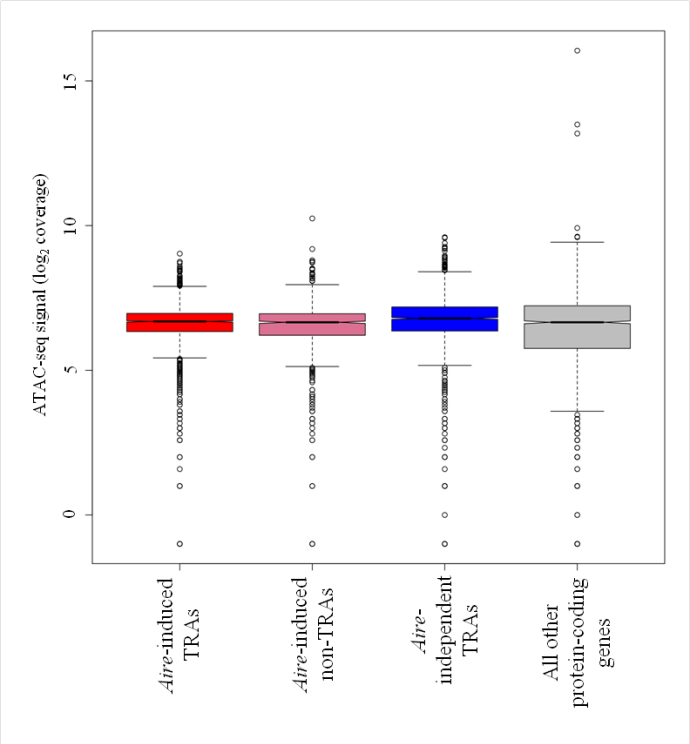

Figure 37 A boxplot of ATAC-seq signal within 1kb of the TSS of genes. Gene categories were matched on proportional expression with *Aire*-induced TRAs in 10% proportional expression bins.

6.5.7 Refining the histone signature of *Aire*-induced genes

It is possible that DNA accessibility in mTEC^{hi} is confounded by the extreme transcriptomic heterogeneity seen in these cells. Therefore, histone marks may provide better insight into the mechanisms of promiscuous gene expression. Previously H3K27me3 has been recognised as a marker of *Aire*-induced genes and this led to the hypothesis that Aire may bind, likely *via* other co-factors, to H3K27me3-rich regions of the genome to permit transcription of otherwise repressed genes.[113] However, the histone architecture in mTEC^{hi} has not been extensively studied and so it is possible that other histone marks also contribute to the recognition of Aire-target genes.

In order to address this question, I analysed ChIP-seq data from wild-type mTEC^{hi} for the activator histone marks H3K4ac, H3K4me1, H3K4me3 and H3K9ac and repressive histone marks H3K27me3 and H3K9me3. Many of these were strongly correlated with protein-coding gene expression but most were less predictive of the magnitude of *Aire*-induced genes (**Table 10**).

Table 10 The correlation between histone ChIP-seq signal and proportional expression of genes in wild-type mTEC^{hi}. All were significant at $p < 0.01$. H3K4me3 data was taken from Sansom *et al.*[113]

Histone mark	Correlation (Spearman's r)			
	***Aire*-dependent**	***Aire*-enhanced**	***Aire*-independent TRAs**	**All protein-coding genes**
H3K27me3	0.09	-0.09	-0.34	-0.54
H3K4ac	0.24	0.38	0.67	0.78
H3K4me1	0.15	0.21	0.33	0.23
H3K4me3	0.13	0.26	0.65	0.80
H3K9ac	0.18	0.31	0.68	0.82
H3K9me3	-0.11	-0.11	-0.38	-0.47

I extended this analysis to assess whether ChIP-seq signals of particular histone marks were increased in *Aire*-induced genes when matched for proportional expression (**Figure 38**). The clearest signal near *Aire*-induced genes was for high levels of H3K27me3 ($p < 10^{-20}$ relative to *Aire*-independent genes), and low levels of both H3K4me3 ($p < 10^{-14}$) and H3K9ac ($p < 10^{-13}$). There was additionally a small magnitude increase of H3K4ac signal near *Aire*-independent genes relative to *Aire*-induced genes ($p < 10^{-15}$). Interestingly, there was a small increase in H3K4me1 signal near TRAs (both *Aire*-induced and *Aire*-independent) relative to non-TRAs ($p < 0.01$), potentially suggesting a common mechanism of TRA expression involving H3K4me1-rich elements near the TSS of TRAs.

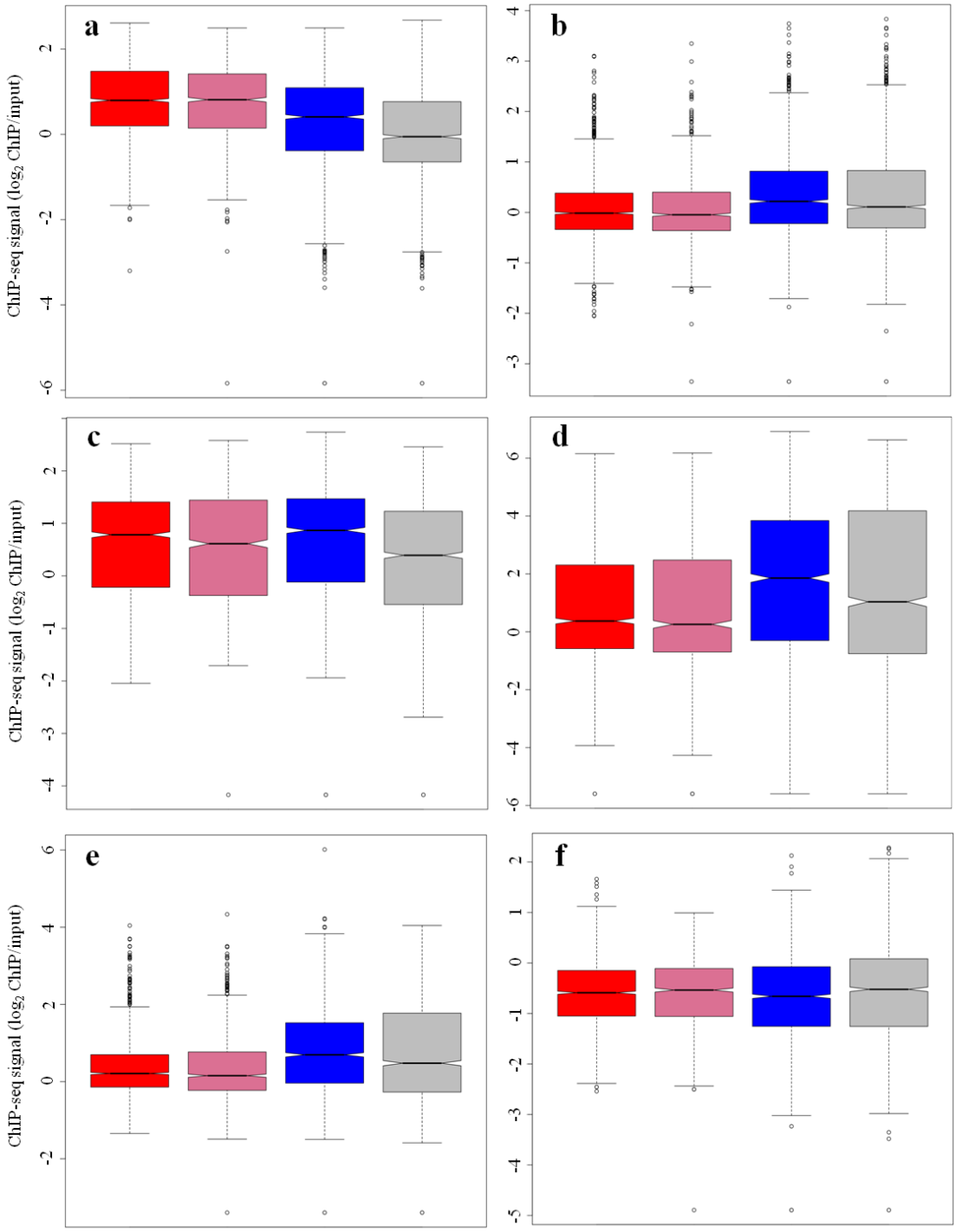

Figure 38 ChIP-seq signal for histone modifications in mTEC^{hi} relative to the *Aire* status of genes. (**a**) H3K27me3, (**b**) H3K4ac, (**c**) H3K4me1, (**d**) H3K4me3, (**e**) H3K9ac and (**f**) H3K9me3. Colours represent *Aire*-induced TRAs (red), *Aire*-induced non-TRAs (light red), *Aire*-independent TRAs (blue) and all other protein-coding genes (gray). Gene categories were matched on proportional expression with *Aire*-induced TRAs in 10% proportional expression bins. In each case ChIP-seq and input coverage were adjusted for library size.

6.5.8 Multiple chromatin states predict *Aire*-dependence of genes

In an attempt to identify an underlying histone signature of *Aire*-induced genes, I combined all mTEC^{hi} histone data into a multiple stepwise logical regression model to reduce collinearity for both peaks called using irreproducibility discovery analysis (IDR < 0.01) or the maximum ATAC-seq/ChIP-seq signal near the TSS of all genes (**Table 11**). As previously, genes were matched for proportional expression in the wild-type mTEC^{hi} dataset.

Table 11 Multiple stepwise logistic regression on *Aire* status with histone data. Odds ratios > 1 or β > 0 indicate the variable is increased in *Aire*-induced genes. Variables with high VIF collinearity scores were dropped and then stepwise backward regression was performed on the surviving variables. Gene categories were matched on proportional expression with *Aire*-induced TRAs in 10% proportional expression bins. OR = odds ratio, CI = confidence interval, β = regression co-efficient, SE = standard error and NA = not applicable (*i.e.* not included in the final regression model).

Dataset	**Peaks (IDR < 0.01)**		**Signal (log_2)**	
	OR (95% CI)	**P-value**	**β (SE)**	**P-value**
(Intercept)	1.23 (1.16-1.31)	$< 10^{-10}$	-1.33 (0.11)	$< 10^{-20}$
ATAC-seq	0.56 (0.44-0.70)	$< 10^{-6}$	0.19 (0.02)	$< 10^{-20}$
H3K27me3	1.96 (1.67-2.31)	$< 10^{-15}$	0.84 (0.03)	$< 10^{-20}$
H3K4ac	0.41 (0.32-0.52)	$< 10^{-12}$	NA	NA
H3K4me1	1.43 (1.25-1.64)	$< 10^{-6}$	0.09 (0.04)	0.03
H3K4me3	0.66 (0.56-0.79)	$< 10^{-5}$	-0.37 (0.02)	$< 10^{-20}$
H3K9ac	0.63 (0.53-0.75)	$< 10^{-6}$	NA	NA
H3K9me3	0.40 (0.20-0.76)	0.007	-0.19 (0.04)	< 0.00001

All types of ATAC-seq and ChIP-seq peaks contributed to a multiple logistic regression model with *Aire*-induced gene status as a dependent variable. There was minimal collinearity between independent variables. The presence of H3K27me3 and H3K4me1 peaks near the TSS of genes was predictive of *Aire*-induced status, whereas ATAC-seq, H3K4ac, H3K4me3, H3K9ac and H3K9me3 were predictive of *Aire*-independence.

When the model was built using log_2 transformed ATAC-seq or ChIP-seq signal near the TSS of genes instead, several histone signatures demonstrated high collinearity scores (H3K9ac: VIF = 8.46; H3K4me3: VIF = 5.75). H3K9ac was highly correlated with H3K4ac and H3K4me3 and this was dropped from the model. H3K4ac made no independent contribution to the model (β = -0.01, SE = 0.05, p = 0.85) and so was also dropped. In the final model, ATAC-seq, H3K27me3 and H3K4me1 were independent predictors of *Aire*-induced status, with H3K4me3 and H3K9me3 being independent predictors of *Aire*-independence.

6.6 Discussion

Single cell RNA-seq of mTEC^{hi} derived from wild-type and TEC-specific conditional knock-out of genes encoding the epigenome remodelling enzymes, Dnmt3a/b, Eed and Tdg, recapitulated previously published transcriptomic data from bulk and single cell mature mTEC samples. When key epigenomic remodelling enzymes were knocked out in mTEC^{hi}, this had complex effects on *Aire*-induced gene expression. Concentrating further analysis on wild-type mTEC^{hi} demonstrated low magnitude but significant intrachromosomal interactions between *Aire*-induced TRAs. Intersecting single cell transcriptomes with data on chromatin accessibility and histone modifications revealed a complex histone code associated with the expression of Aire target genes.

To my knowledge, no previous studies have profiled the transcriptomic changes associated with epigenomic remodelling enzymes mutations in TEC. Previous research has suggested that *Aire*-induced gene expression is associated with a pattern of histone modifications (high levels of H3K27me3 and low levels of H3K4me3) and that mature mTEC selected to express a particular TRA demonstrate higher accessibility around the TSS of that TRA and its co-expressed partners.[113,114] Additionally, the fact that *Aire* knock-out mice do not show global changes in DNA methylation in mTEC was taken as evidence that DNA methylation is unimportant in PGE.[115] However, without direct disruption of epigenomic remodelling mechanisms it is impossible to claim causation.

The data presented here suggest that *de novo* DNA methylation is likely to be an important process in controlling *Aire*-induced gene expression. *Dnmt3a/b* knock-out in mouse embryonic fibroblasts caused global demethylation and uncontrolled proliferation.[331] In this mTEC^{hi} dataset there was no evidence of deregulated growth, as cell cycle latent variable scores did not differ significantly between *Dnmt3a/b*$^{\Delta}$ and wild-type cells ($p = 0.13$). However, there were robust effects on the overall proportion of *Aire*-induced genes detected. A key difficulty when studying PGE biology is that many of the genes of interest are expressed extremely infrequently and so technical biases surrounding detection are a potential issue. I attempted to control for this by

demonstrating that the observed effect remained detectable at multiple expression cut-offs and when estimating the number of molecules present in each cell based on the expression of ERCC spike-ins. If this finding were solely due to the presence of many lowly expressed *Aire*-induced genes, then it should have disappeared when I selected a matched set of *Aire*-independent genes based on binned proportional expression. The molecular mechanisms underlying this increased *Aire*-induced gene expression are unclear but may involve altered Aire interaction with heterochromatin or Polycomb repressive complexes. Generally DNA methylation and repressive marks such as H3K27me3 are mutually exclusive.[343,344] *Dnmt1* knock-out models causing global DNA demethylation demonstrated that the methylation state across the genome was critical for the correct association of Polycomb repressive complexes with their target genes.[345] Global loss of DNA methylation resulted in the accumulation of ectopic H3K27me3 at many genomic *loci*. Conversely, insufficiency of *DNMT3B*, and thus *de novo* methyation, results in DNA demethylation and the disappearance of repressive chromatin modifications, including H3K27me3, from the promoters of many genes important in immune function but relatively minimal alteration outside of centromeric and X chromosomal regions.[346,347] It is likely that in this *Dnmt3a/b*$^{\Delta}$ mTEC^{hi} model, with highly efficient knock-out of functional *Dnmt3a* and *Dnmt3b*, the pattern of H3K27me3 histone modifications has been severely altered. Since H3K27me3 is highly enriched near *Aire*-induced genes, it is perhaps not unexpected that these show higher proportional expression. In fact, one might predict that previously *Aire*-dependent genes become *Aire*-independent following the knock-out of *Dnmt3a/b*, since these could be expressed without H3K27me3-mediated *Aire* de-repression.

Surprisingly there were few differences apparent between *Tdg*$^{\Delta}$ and wild-type mTEC^{hi} despite quite efficient *Foxn1*-driven Cre excision of the *Tdg* locus. This is remarkable given the critical role that *Tdg* is known to play in DNA demethylation of many promoters, particularly those sensitive to retinoic acid.[348] There are two main potential reasons why *Tdg* knock-out may have had little effect on the transcriptome of mTEC^{hi}. Firstly, it is possible that the loss of *Tdg* expression occurred after the time window when retinoic acid was contributing to thymic

development. There is conflicting evidence as to whether or not retinoic acid is important in thymogenesis.[349–352] However, it is clear that any effect of retinoic acid on thymic development occurs at an early stage during the pattern of the third pharyngeal pouch (mouse E11), with Foxn1^{+} cells appearing only later (mouse E11.5-12.5).[349–351] The role of retinoic acid in mature TEC is unclear. Therefore, it is possible that *Tdg* was excised by the *Foxn1*-driven Cre after the genes most sensitive to its absence had already had their effect. Secondly, there was still detectable expression of *Tdg* mRNA in *Tdg*$^{\Delta}$ mTEC^{hi}, albeit at a much lower proportion and level (**Table 9**). It is possible that only a low level of *Tdg* was required for mTEC development and function; a more complete knock-out may have been needed to see any transcriptomic effect.

Interpretation of the effects seen in *Eed*$^{\Delta}$ mTEC^{hi} is more difficult. It was clear that *β5t*-driven Cre-mediated excision of *Eed* was largely ineffective. Despite this, mTEC^{hi} from *Eed*$^{\Delta}$ mice had lower levels of *Aire*-induced and, to a lesser extent, *Aire*-independent TRA expression. It is surprising that *any* mTEC^{hi} expressed *Eed*, since in normal thymic development > 99% of mTEC are derived from β5t^{+} bipotent precursors.[102] This could either be due to inefficient Cre activity, which was not the case because cTEC from the same mice had no detectable *Eed* (data not shown), or to the expansion of an mTEC population derived from β5t^{-} precursor cells. If the latter explanation is correct, this may indicate that *Eed* expression is critical for normal mTEC development and that the medullary compartment of the thymus is replaced by mTEC that proceed through a different developmental trajectory. Such cells may not be able to recapitulate normal mTEC maturation and function in its entirety as suggested by the deficient PGE observed in *Eed*$^{\Delta}$ mTEC^{hi}.

There are some important limitations of the transcriptomic data presented here. Even though this is a relatively large-scale single cell RNA-seq study, the proportion of total mTEC^{hi} sampled is still low. Where PGE is the main area of interest and these genes are generally expressed at the lower limit of detection, this is a cause for concern. I attempted to mitigate this by using multiple expression cut-offs and analysing data separately using ERCC spike-in estimates of molecular abundance as well as more traditional measures of gene expression. When pooling all wild-type

mTEC^{hi}, there was detectable expression of ~77% of known *Aire*-induced TRAs. This means that, even with the inclusion of 185 single cells, some TRAs were not observed. The trajectory of the graph depicting the number of genes detected against the number of cells sampled suggests that another magnitude of cells will be needed to represent all *Aire*-induced TRAs (**Figure 34**). The depth of sequencing per cell required to interrogate the *Aire*-induced transcriptome means that this experiment would only be feasible if sequencing costs were drastically reduced or with massively multiplexed targeted sequencing methodologies. This would also require the development of new differential analysis tools, since Bayesian priors do not work well at the limits of detection.[338]

There was some evidence that *Aire*-induced gene expression may be co-ordinated within intrachromosomal domains. However, this was the case only for a minority (~9.6%) of *Aire*-induced TRAs. Interestingly, this is of the same order of magnitude (~16%) as the proportion of *Aire*-induced genes known to be co-expressed from sequential purification of mTEC populations for particular TRAs.[325] It is possible that a larger dataset of single mTEC^{hi} transcriptomes will be required to exclude a dominant role of intrachromosomal co-expression domains in the co-ordination of *Aire*-induced genes; the fact that co-expression modules are unique to individual mice argues against a strong and consistent role for this mechanism in controlling PGE.[115]

Integration of multiple datasets relating to chromatin structure in mTEC^{hi} reiterated the importance of H3K27me3 and H3K4me3 in determining whether or not a specific gene is a target of Aire. However, both when using peaks or signal magnitude independent of peak-calling, multiple other dimensions of chromatin architecture appear to be involved in gene regulation by Aire. The most consistent finding from my multiple logistic regression models was for H3K27me3 and H3K4me1 to characterise *Aire*-induced genes, and H3K4me3 and H3K9me3 to characterise *Aire*-independent genes. Interestingly, this suggests that the idea that repressive chromatin *per se* indentifies Aire target genes is overly simplistic, since the presence of H3K9me3, a repressive chromatin modification, is actually modestly predictive of *Aire*-independent gene status. This may mean that mTEC use repressive histone modifications other than H3K27me3 to control the expression of some *Aire*-independent genes.

The contrasting effect on Aire target status of ATAC-seq peaks compared with coverage near the TSS of genes is curious. As shown in **Figure 37** ATAC-seq signal differs minimally between *Aire*-induced and *Aire*-independent genes in a univariate analysis. A multiple logistic regression model built on peaks called by irreproducibility discovery analysis suggested that the presence of a peak was predictive of a gene being *Aire* independent; conversely, in the model constructed from ATAC-seq and ChIP-seq signal intensity, ATAC-seq signal was predictive of a gene being regulated by Aire. There is a strong correlation between ATAC-seq signal and the peak p-value called by MACS2 (Spearman's r = 0.96), so it is likely that the two measures are equivalent. The ATAC-seq protocol is still relatively novel and it may be that peak-callers have yet to be optimised for such datasets, particularly if the pattern of accessible chromatin is not classical of enhancer or promoter peaks.[36] Thus, although the contribution of histone modifications to *Aire*-induced gene status is clear, the relationship of this to chromatin accessibility is less clear.

6.7 Conclusion

Deficiencies of *Dnmt3a/b* or *Eed* resulted respectively in an increase or decrease of the number of *Aire*-induced TRAs detectable in single mTEC^{hi}, suggesting a critically important role for epigenomic remodelling in PGE. *Aire*-induced genes had a weakly significant tendancy to form intrachromosomal co-expression clusters, but this was likely to account only for a small proportion of Aire target gene expression. *Aire*-induced genes were characterised by a complex histone code rich in H3K27me3 and H3K4me1 and poor in H3K4me3 and H3K9me3. This chapter has illustrated how important epigenomic states are in determining the expression of Aire target genes.

6.8 Acknowledgements

Stefano Maio, Carlos Meyer and Hong Ying Teh performed mouse breeding and sorted mTEC^{hi} for single cell RNA-seq. Iain Macaulay and Annina Graedel performed single cell library preparation. Stefano Maio prepared samples and libraries for ATAC-seq. Noriko Shikama-Dorn and Saule Zhanybekova prepared samples and libraries for ChIP-seq.

Chapter 7 Conclusion

7.1 Cerebral cortex and neurological disease

In **Chapter 2**, **Chapter 3** and **Chapter 4,** I detailed the functional genomics approaches I adopted to address a series of specific key questions regarding cerebral cortex biology and pathophysiology posed in **Chapter 1**. **Chapter 2** employed an iPSC-derived model of corticogenesis and applied single cell genomics methods to this in order to investigate the extent to which iPSC-derived cortical neurons resembled primary human cortical neurons. **Chapter 3** extended this model into a study of cortical neurons derived from a series of iPSC bearing different *PSEN1* mutations known to cause Alzheimer's disease in an attempt to explore disease-relevant biology. In contrast, **Chapter 4** used functional genomics data from primary brain samples to investigate how common variants associated with neurological disease may exert their effects. Below, I evaluate the impact of my research within the framework of the specific questions I aimed to answer and suggest the potential directions that further work could take in exploring the remaining gaps in current scientific knowledge.

7.1.1 How closely do stem cell-derived cortical neurons resemble neurons from primary cerebral cortex?

Before making any attempt to infer disease biology from stem cell-derived models, it is imperative to understand the extent to which the resultant cells reflect primary tissues. Surprisingly, few studies directly addressed this, and most protocol papers reported the presence of markers by RT-qPCR or immunofluorescence microscopy as a proxy for tissue identity.[60,61] Where attempts were made to assign a transcriptomic identity to iPSC-derived cortical neurons, this was hampered by the heterogeneity of cultures and the likely presence of cell types in primary tissues that were not present *in vitro*.[65] An obvious solution to this was to apply single cell technology. This had been done to a limited extent by the use of single cell RT-qPCR in relatively small numbers of cells and a generally uncritical approach was taken to assigning cells to specific functional categories within the cerebral cortex.[124,130] I aimed to address this gap in documented knowledge by undertaking a relatively large-scale RT-qPCR study of multiple genes

considered important to regionalisation and functional subspecification within cortical neurons and glia along with whole transcriptomic data from a smaller number of cortical neurons derived from the same iPSC line. This demonstrated that, even at the level of single cells, iPSC-derived cortical neurons were highly similar to primary human fetal cortex. Another recently published study performing single cell RNA-seq in cerebral organoids also found a high degree of similarity between stem cell-derived cortical neurons and primary human fetal cortex.[151] However, I found that iPSC-derived cortical neurons were very different to adult cortical neurons. This is hardly unexpected, since the transcriptome of adult human cortical neurons will be shaped by decades of life, hormonal signals, long-term modulation of synaptic plasticity and interactions with a multitude of other cell types.[353,354]

Surprisingly, 22.4% of neurons co-expressed canonical cortical layer markers from *different* layers of the cortex. This was confirmed by assessing co-expression of the cortical layer markers by immunofluorescence microscopy and showing similar patterns of mixed layer identity. Despite that fact that deep and upper layer cortical markers are reportedly mutually exclusive, the fact that paracrine cues from surrounding cells will be very different within disordered iPSC-derived cultures from those experienced by cells *in vivo* means that non-canonical layer identity is not unexpected under such conditions.[39]

Further exploration of the heterogeneity present in these cultures will require much larger scale single cell RNA-seq studies. It is quite likely that there are subpopulations of cells that were missed by previous single cell transcriptomic approaches, either due to the relatively small numbers of analysed cells or biases intrinsic to the sample preparation protocols.

Narrowing the transcriptomic gap between iPSC-derived cortical neurons and primary adult cortical neurons presents a far greater challenge. Clearly the time-scale of scientific projects precludes “natural” aging of iPSC-derived neurons *in vitro*, even assuming cells could be kept viable for a sufficient number of years. Potentially artificial activation of aging-related pathways could help, although it is difficult to see how this could replicate the complex endocrine changes and synaptic pruning that take place during the lifetime of an individual.[172]

7.1.2 Does the degree of variability between cortical neurons derived from different iPSC allow for meaningful comparisons to be made between patient- and control-derived iPSC models of corticogenesis?

This was a more difficult question to answer than I had anticipated. There are actually at least two separate threads to this issue: firstly, is the batch-to-batch noise too high to detect any transcriptomic signal, and secondly, is it possible to classify any resultant transcriptomic signature as disease-related (which in most cases will be equivalent to genotype-related)?

By comparing bulk RNA-seq data from StemBANCC wild-type and *PSEN1* mutant iPSC-derived cortical neurons, I was able to identify over two thousand genes that were differentially expressed between the two lines. The degree of transcriptomic noise was high (biological coefficient of variation = 0.3-0.4) and was approximately equivalent to levels of noise reported in previous RNA-seq studies of primary tissue samples.[203] However, the fact that I was able to detect a large number of differentially expressed genes despite this suggests that batch-to-batch variability is unlikely to overwhelm any differences between iPSC lines completely.

However, assessing whether these differences were a result of *PSEN1* genotype or simply a result of using different iPSC lines was more challenging. There was very little overlap between differentially expressed genes in this StemBANCC case-control experiment and one in which I used *PSEN1* isogenic lines generated by collaborators in UCSD. This suggested that *PSEN1* genotype alone has a limited effect on the transcriptome of iPSC-derived cortical neurons. This does not mean that the differentially expressed genes were unrelated to disease status, since it is possible that background genome variation or a "memory" of disease status could have differentially driven the observed transcriptomic changes. Given that any Alzheimer's disease transcriptomic signature would be likely to involve many key neuronal pathways, it is extremely difficult to dissect out variability in differentiation capacity between lines from genuine, disease-related differences. Applying well-defined exogenous stimuli, such as Aβ oligomers, to cultures may help to uncover disease-related signatures.

Future studies should concentrate on identifying an expected set of high confidence genes that are robustly differentially expressed in primary cortical neurons from Alzheimer's disease relative to control brains at all stages of disease.[189] This would enable a rapid "sanity check" of *in vitro* Alzheimer's disease models. If these genes were not detected as differentially expressed in a particular model then this would clearly lead one to question its validity. The poor replicability observed in the two experiments on *PSEN1* genotype I presented in this thesis argue strongly that any novel findings obtained from stem cell-based systems should be replicated in primary tissue wherever possible. This clearly creates a problem for a model designed to examine extremely early disease biology, for which, unless a reasonable number of individuals with familial Alzheimer's disease are included in brain banks prior to the onset of overt disease, there is no obvious solution in the absence of good animal models of neurodegeneration.[90,160]

A better understanding of the impact of differentiation protocols on transcriptomic noise in iPSC-derived cortical neurons will be essential in designing well-powered experiments capable of identifying disease-relevant signatures. I am involved in a study assessing inter-laboratory variability of iPSC corticogenesis which may help address this. Similarly, studies examining the degree of differentiation propensity in different lines may also help distinguish differentiation-related gene signatures from those genuinely related to disease. The use of single cell RNA-seq approaches in these situations would permit cell type-specific analyses, which would reduce the effect of variable proportions of different cell types in iPSC-derived cultures.

7.1.3 Does the use of genome engineering reduce variability between iPSC-derived cortical neurons?

There is convincing evidence that the genetic background of iPSC lines is a key determinant of the subsequent transcriptomic diversity between lines.[94] For this reason, it was widely assumed that the use of isogenic, genome-engineered iPSC lines would reduce gene expression noise in iPSC-derived differentiated cells.[355]

Surprisingly, I could not demonstrate a significant decrease in transcriptomic noise in the genome-engineered iPSC-derived cortical neurons compared with the case-control iPSC-derived

cortical neurons. In fact, there was some evidence that the specific time at which cells were plated down for differentiation was a larger contributor to the final transcriptome of differentiated cells than the genetic background. Given that the duration of neuronal differentiation was so long (55 days from final plating to harvesting cells) it could be that differences in the proportion of neuronal progenitor subtypes or the effective age of cultures overwhelmed any initial contributions of donor genotype.

In order to answer this definitively, future work could look at multiple timepoints across the course of cortical differentiation to assess when the transcriptomes of cultures diverge from one another. It will be interesting to see the results of similar approaches taken with different gene edits and protocols in order to identify whether this is a *PSEN1*- and cortical-specific effect, or whether it is a more general limitation of iPSC-based models. If this turns out to be a general feature of iPSC-based biology, this means that the assumption that studies using isogenic iPSC lines will necessarily be better powered to detect differences than iPSC lines not matched for genetic background may be unwarranted.

7.1.4 By what mechanism do common variants associated with neurological disease exert their effect on disease susceptibility?

The mechanistic link between common sequence variants and phenotypic traits is one of the major issues stemming from the fact that most eQTL and GWAS signals fall in non-coding DNA. It is clear that many of the trait-associated haplotypes are within regions of accessible chromatin, but precisely how this translates to altered expression of genes or disease-related biological processes is far less clear.[80]

I used brain DNase-seq data to assess whether a large proportion of these variants could actually be exerting their effects by disrupting TF binding. In order to do this, I generated a map of high resolution DNase-seq footprints, which corresponded to regions of DNA protected from DNase digestion by bound proteins.[232] This showed that only a minority of haplotypes contained at least one variant intersecting a DNase-seq footprint, suggesting either that DNase-seq footprinting was

insufficiently sensitive to capture most TF binding events or that most eQTL and GWAS variants did not affect TF binding.

It is possible that improvements in the detection of footprints by accounting for sequence bias may alter these estimates (although likely only to a small extent).[356] It is also possible that functional genomics studies in larger numbers of individuals or in cells of specific types may help uncover the molecular mechanisms linking sequence variation to expression or phenotypic traits.[251]

7.2 Thymus

In **Chapter 5** and **Chapter 6** I addressed a number of issues in thymic biology. **Chapter 5** integrated ChIP-seq and RNA-seq datasets in order to identify biological pathways underlying the effects of *Foxn1* on TEC differentiation and function. **Chapter 6** used single cell RNA-seq data from mTEC^{hi} from either wild-type mice or those with conditional TEC mutations in *Dnmt3a/b*, *Eed* or *Tdg* to study the role of epigenomics in PGE. I also used histone ChIP-seq and ATAC-seq data to refine a histone code for *Aire*'s regulation of genes. In the next paragraphs, I consider whether these experiments were able to address some of the key questions I raised in **Chapter 1**.

7.2.1 How does *Foxn1* function in the development and maintenance of TEC?

Despite the fact that *Foxn1* is a critical factor in thymic development, relatively little is known about its functions *in vivo*.[104] At the time this research was first carried out, functional genomics technologies did not permit Foxn1 ChIP-seq in the developing thymus due to the number of cells required. For this reason, ChIP-seq and RNA-seq approaches were employed to identify *Foxn1* target genes in the mature thymus, in which the requirements for material could be met by pooling TEC isolated from multiple mice. There is evidence that *Foxn1* is critical for the maintenance of TEC competency and so biological pathways identified by this approach were likely to represent critical pathways for TEC function.[106,357]

For the first time, I was able to use ChIP-seq to identify an *in vivo* binding motif for Foxn1 and generate a genome-wide map of Foxn1 binding sites in TEC. The intersection of this dataset with RNA-seq data from constitutive or inducible *Foxn1* deficient cTEC and mTEC permitted the

identification of 450 high confidence *Foxn1* gene targets. These were enriched within pathways involved with antigen processing and presentation. Two of these *Foxn1* target genes, *Cd83* and *Psmb11*, recapitulated specific thymopoietic defects seen in the *nude* mouse model when knocked-out.

My findings should help understand the role *Foxn1* plays in TEC function in the mature thymus. Some of the binding sites identified are likely to contribute to TEC development, although definitive proof of that awaits the use of ultra-low input ChIP-seq protocols.[37,358]

7.2.2 How do TEC co-ordinate promiscuous gene expression?

One of the most important functions of TEC is positive and negative selection of thymocytes; this process is dependent on PGE.[109] The major TF responsible for PGE in mTEC is *Aire*, although clearly other processes are active both in mTEC and cTEC that promote PGE.[113] Some studies found that individual TEC co-ordinate PGE through the transcriptional activation of specific genomic regions, whereas other evidence suggested that PGE in individual TEC was actually stochastic.[113–115] Interpreting the results of these studies is also complicated by the high rate of doublets recently identified in the C1 single cell RNA-seq platform. An iterative approach to this problem using bulk transcriptomic data, in which TEC were sequentially purified for specific antigens in order to identify modules of co-expressed genes, found some evidence of co-ordinated PGE but only for a small minority of total *Aire*-induced genes (~16%).[325]

My analysis showed that there was significantly higher intra-chromosomal co-expression than expected by chance but that this could only account for a small proportion of *Aire*-induced PGE (~9.6%). It was also noteable that co-expression patterns did not replicate between individual mice, let alone different mouse models, or even between different sequencing plates from the same mice.

This suggests that PGE is partially co-ordinated within intra-chromosomal domains but that this is a minor component of overall PGE and far from the dominant factor in co-ordinating PGE. It may be that far larger numbers of single cells, possibly with retained spatial information, will be required in order to fully understand PGE in TEC.[71,359,360] However, the fact that *Aire*-induced

TRAs are often lowly expressed will mean that the depth of sequencing required for this is currently economically prohibitive.

7.2.3 What role does dynamic remodelling of the epigenome play in promiscuous gene expression?

The epigenome is known to be important for PGE as *Aire*-induced genes are more frequently highly enriched for the repressive chromatin mark H3K27me3 and depleted in the activator mark H3K4me3 than is the case for *Aire*-independent genes.[113] However, the role dynamic remodelling of the epigenome plays in PGE is far less well understood.

In order to address this question, I analysed single cell RNA-seq data generated on mTEC^{hi} derived from either wild-type mice or mice with TEC-specific mutations in *Dnmt3a/b*, enzymes involved in *de novo* DNA methylation, *Eed*, a component of Polycomb Repressive Complex 2, or *Tdg*, an enzyme involved in modifying methylated cytosine bases. I found that mutation of *Tdg* had little effect on the mTEC^{hi} transcriptome. In contrast, the inactivation of *Dnmt3a/b* caused increased proportional expression of *Aire*-induced genes, potentially due to H3K27me3 redistribution in response to reduced DNA methylation.[345] The picture was more complex for *Eed* mutation, as most mTEC^{hi} appeared to escape Cre-mediated excision of the LoxP-flanked exons. However, contrary to expectations, mTEC^{hi} demonstrated reductions in proportional expression of *Aire*-induced and *Aire*-independent TRAs. Since mTEC targeted gene loss was specific for β5t^{+} cells and since most mTEC are produced by β5t^{+} bipotent precursor cells in the thymus, this suggests that these mTEC^{hi} must have an as yet undescribed developmental origin.[102] Given the defects in TRA expression, this hypothetical developmental pathway would appear to produce hypofunctional mTEC^{hi}.

Additionally, through the use of an expanded dataset of histone ChIP-seq and ATAC-seq in mTEC^{hi}, I was able to demonstrate that the histone code underlying *Aire*-induced gene expression was more complex than simply high levels of H3K27me3 and low levels of H3K4me3. Additionally, high levels of H3K4me1 and low levels of H3K9me3 were characteristic of *Aire*-induced genes.

Further work will be required to understand the functional implications of these epigenome-remodelling mutations, particularly in relation to the potentially novel origin of mTEC^{hi} in *Eed* mutant mice. In order to test some of the hypotheses raised by analysis of the single cell RNA-seq data, ATAC-seq and histone ChIP-seq from mutant mTEC^{hi} would be invaluable. Techniques that can interrogate the epigenome and transcriptome in single cells will undoubtedly prove instrumental in understanding the role of the epigenome in PGE.[49]

7.3 Future functional genomics approaches to tissue heterogeneity

7.3.1 Methodological developments

Functional genomics methodologies are advancing at a phenomenal pace. The leap from single cell multiplexed RT-qPCR to wide-spread use of high throughput single cell RNA-seq took place almost within the duration of my DPhil studies.[130,359] In parallel, single cell methods offering information on chromatin accessibility, the three-dimensional structure of DNA, DNA methylation and targeted protein abundance, often in parallel with transcriptomics, have been developed.[33,36,49,50]

These new technologies are still in the optimisation phase. For example, even in the case of single cell RNA-seq, now considered a routine technique, the protocols typically capture only ~20% of the total transcriptome. For highly expressed genes this is probably adequate but for lowly expressed genes, particularly transcription factors, this is far from ideal.[361] Sparse sampling of the underlying biological substrate is a recurrent theme in all single cell approaches; it is difficult to imagine a situation in which this would not be the case unless the efficiency of all steps within protocols approaches 100%. One example where this may be the case, was illustrated by MERFISH, in which highly multiplexed single molecule FISH was used to probe the expression of hundreds or thousands of RNA species simultaneously.[72] The authors reported single molecule resolution when measuring the expression of 140 multiplexed transcripts but this fell significantly when extending this to > 1,000 transcripts. This is an important study as it highlighted that, even in the absence of reverse transcription capture and amplification biases, it was difficult to achieve high sensitive, multiplex assays of RNA expression. It also demonstrated the importance of

retaining spatial information in samples, which is now possible using *in situ* single cell RNA-seq protocols.[71] There are ways of attempting to reconstruct spatial information after dissociation of single cells but these are likely to be useful only in relatively simple biological systems.[362]

Single cell RNA-seq research appears to be splitting into three complementary approaches. The first is based on ultra-high throughput profiling of many thousands of cells, using technologies such as Drop-seq.[359] The second concentrates on integrating multiple forms of phenotypic data in smaller numbers of single cells, either dynamic cellular information using newly developed platforms such as the Fluidigm Polaris, or orthogonal measurements of either epigenomic state or protein abundance.[49,50] The third focuses on non-destructive, *in situ* analyses of single cells, where measurements can be acquired over time to an even greater extent than is permitted by the Polaris system.[363] It is likely that dissecting the complex interplay of factors affecting cell type specification in heterogenous tissues will require all three approaches. This is because the identification of novel, rare subpopulations requires information on large numbers of cells but understanding how cellular specification occurs mechanistically and how this fits within developmental trajectories needs more detailed, dynamic datasets.

The situation is even more challenging for single cell DNA protocols, in which there can by definition only be two copies of each allele present (assuming a diploid karyotype). The results of ATAC-seq and DNA methylation assays in single cells have demonstrated how sparse such datasets can be, to the extent where analysis is challenging without aggregating single cells; this results in the concomitant loss of much of the single cell information, which likely motivated the study in the first place.[34,36] Refinements to ChIP-seq protocols have both reduced the number of cells required and have improved the spatial resolution of binding sites by incorporating a DNA exonuclease step.[22,23,37,358]

In contrast to RNA-seq and DNA-seq-based methods, single cell proteomics is relatively underdeveloped. Multiplexed protein detection is possible using FACS analysis, single cell mass spectrometry, proximity extension assays or Western blotting.[50,364,365] However, there is currently no tool to analyse the whole proteome at the level of single cells (and, as shown in **Chapter 3**, even in bulk samples sensitivity is limited). Without drastic improvement in spectrometer

sensitivity or a paradigm shift allowing effective amplification of all peptide species, it is difficult to conceive of a method that could permit single cell proteomics to interrogate the entire proteome.

7.3.2 Bioinformatic developments

At least as important as the technological refinement of functional genomics methods are improvements to analysis tools. The expansion of single cell approaches such as Drop-seq is leading to the production of unprecedented amounts of sequencing data. The bottleneck of such studies will soon be (and probably already is) the interpretation of such rich and complex datasets. The challenge involved in this can be placed into stark context by the fact that methods for analysing even *bulk* RNA-seq datasets are far from ideal. The particular method used can lead to marked differences in the final estimates of gene expression.[366] It is likely that pseudo-alignment approaches such as Sailfish and Kallisto, which have been shown on *in silico* and some RNA-seq datasets to provide more accurate transcript quantification than more traditional approaches.[367,368] However, the optimum pipeline will likely remain project-specific. For example, in the case of TEC biology, assigning multi-mapping reads to genes could inappropriately inflate the estimated expression level of many lowly expressed TRAs.

The analysis of single cell RNA-seq data is rendered even more complex than bulk RNA-seq data by virtue of the fact that drop-out events and zero-inflation are extremely common.[361] The optimum methods for quality control, clustering or differential expression in this situation is still unclear and, indeed, may be cell-type or dataset specific.[338,369,370] The use of spike-ins as a method of modelling technical noise is critical, and many of the downstream analytical tools make use of this information to estimate transcriptomic noise and drop-out rates.[129,336,371] Batch effect and latent variables have a large role in shaping many aspects of the single cell transcriptome; methods of accounting for these factors allow the identification of features otherwise concealed within the data.[336]

Bioinformatic approaches to ChIP-seq, DNase-seq and (particularly) ATAC-seq are likewise far from standardised, despite the best intentions of the ENCODE consortium.[372] In the case of ChIP-

seq, the use of irreproducibility discovery rate-based analysis is able to leverage the power of replicates to improve the likelihood of identified peaks being true positives.[230,282] However, it is likely that many functional binding sites lie within repeat-rich regions, which conventional approaches to peak-calling simply discard.[280,281] For broad chromatin marks, peak-calling is mostly unsatisfactory, suggesting that new methods will be needed to improve on current analytical pipelines.[230] I discussed the limitations of analysing DNase-seq data in depth in **Chapter 4** so will not reiterate that here; however, the analysis of ATAC-seq datasets is even more problematic, with different approaches used to call regions of accessibility in two papers from the same group.[8,36] Undoubtedly, the increased use of ATAC-seq in multiple biological contexts will stimulate bioinformatic efforts in this area.

7.3.3 Final thoughts

Although I have concentrated on single cell functional genomics methods in the above discussion, bulk functional genomics approaches are still extremely useful. I suspect that, as the novelty of single cell RNA-seq reduces, technical and analytic pipelines will become standardised. This can only be a good thing, as it means that experiments will employ single cell approaches only when these are likely to provide valuable biological insights. The proliferation of single cell datasets will also enable cross-validation of analytical methods between datasets. I predict that this will lead to an overhaul of how biologists interpret cellular identity. Future work is likely to focus more on how cellular identity varies by spatial location within a heterogeneous tissue or how cells interact with one another than on simply cataloguing large numbers of single cells.[71,373]

References

1. Handel, A. E., Disanto, G. & Ramagopalan, S. V. Next-generation sequencing in understanding complex neurological disease. *Expert Rev. Neurother.* **13,** 215–227 (2013).
2. Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A. & Kim, D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.* **16,** 85–97 (2015).
3. Foo, J.-N., Liu, J.-J. & Tan, E.-K. Whole-genome and whole-exome sequencing in neurological diseases. *Nat. Rev. Neurol.* (2012). doi:10.1038/nrneurol.2012.148
4. Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* **12,** 87–98 (2011).
5. Barski, A. & Zhao, K. Genomic location analysis by ChIP-Seq. *J. Cell. Biochem.* **107,** 11–18 (2009).
6. Laird, P. W. Principles and challenges of genomewide DNA methylation analysis. *Nat. Rev. Genet.* **11,** 191–203 (2010).
7. Song, L. & Crawford, G. E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.* **2010,** pdb.prot5384 (2010).
8. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10,** 1213–1218 (2013).
9. Dekker, J., Marti-Renom, M. A. & Mirny, L. A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.* **14,** 390–403 (2013).
10. Metzker, M. L. Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11,** 31–46 (2010).
11. Altelaar, A. F. M., Munoz, J. & Heck, A. J. R. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat. Rev. Genet.* **14,** 35–48 (2013).
12. Yi, M. *et al.* Performance comparison of SNP detection tools with illumina exome sequencing data—an assessment using both family pedigree information and sample-matched SNP array data. *Nucleic Acids Res.* gku392 (2014). doi:10.1093/nar/gku392
13. Pendleton, M. *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12,** 780–786 (2015).
14. Takahashi, H., Lassmann, T., Murata, M. & Carninci, P. 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat. Protoc.* **7,** 542–561 (2012).
15. Ingolia, N. T., Brar, G. A., Rouskin, S., McGeachy, A. M. & Weissman, J. S. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.* **7,** 1534–1550 (2012).
16. Nojima, T. *et al.* Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell* **161,** 526–540 (2015).
17. Wang, C. *et al.* The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nat. Biotechnol.* **32,** 926–932 (2014).
18. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **13,** 227–232 (2012).
19. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473,** 337–342 (2011).
20. Vogel, C. *et al.* Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol. Syst. Biol.* **6,** 400 (2010).
21. Yan, X., Hoek, T. A., Vale, R. D. & Tanenbaum, M. E. Dynamics of Translation of Single mRNA Molecules In Vivo. *Cell* **165,** 976–989 (2016).

22. Rhee, H. S. & Pugh, B. F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147,** 1408–1419 (2011).

23. He, Q., Johnston, J. & Zeitlinger, J. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nat. Biotechnol.* **33,** 395–401 (2015).

24. Booth, M. J. *et al.* Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nat. Protoc.* **8,** 1841–1851 (2013).

25. Taiwo, O. *et al.* Methylome analysis using MeDIP-seq with low DNA concentrations. *Nat. Protoc.* **7,** 617–636 (2012).

26. Schones, D. E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132,** 887–898 (2008).

27. Simon, J. M., Giresi, P. G., Davis, I. J. & Lieb, J. D. Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. *Nat. Protoc.* **7,** 256–267 (2012).

28. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489,** 83–90 (2012).

29. Barozzi, I., Bora, P. & Morelli, M. J. Comparative evaluation of DNase-seq footprint identification strategies. *Front. Genet.* **5,** 278 (2014).

30. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (2009).

31. Fullwood, M. J., Han, Y., Wei, C.-L., Ruan, X. & Ruan, Y. Chromatin interaction analysis using paired-end tag sequencing. *Curr. Protoc. Mol. Biol. Ed. Frederick M Ausubel Al* **Chapter 21,** Unit 21.15.1-25 (2010).

32. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10,** 1096–1098 (2013).

33. Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502,** 59–64 (2013).

34. Smallwood, S. A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11,** 817–820 (2014).

35. Wang, Y. & Navin, N. E. Advances and Applications of Single-Cell Sequencing Technologies. *Mol. Cell* **58,** 598–609 (2015).

36. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523,** 486–490 (2015).

37. Rotem, A. *et al.* Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* **33,** 1165–1172 (2015).

38. Jin, W. *et al.* Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature* **528,** 142–146 (2015).

39. Shalek, A. K. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510,** 363–369 (2014).

40. Hanchate, N. K. *et al.* Single-cell transcriptomics reveals receptor transformations during olfactory neurogenesis. *Science* aad2456 (2015). doi:10.1126/science.aad2456

41. Eberwine, J. & Kim, J. Cellular Deconstruction: Finding Meaning in Individual Cell Variation. *Trends Cell Biol.* **25,** 569–578 (2015).

42. Arsenio, J. *et al.* Early specification of CD8+ T lymphocyte fates during adaptive immunity revealed by single-cell gene-expression analyses. *Nat. Immunol.* **15,** 365–372 (2014).

43. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32,** 381–386 (2014).

44. Pollen, A. A. *et al.* Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32,** 1053–8 (2014).

45. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509,** 371–375 (2014).

46. Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci.* 201507125 (2015). doi:10.1073/pnas.1507125112

47. Zeisel, A. *et al.* Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347,** 1138–1142 (2015).

48. Macaulay, I. C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12,** 519–522 (2015).

49. Angermueller, C. *et al.* Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* (2016). doi:10.1038/nmeth.3728

50. Darmanis, S. *et al.* Simultaneous Multiplexed Measurement of RNA and Proteins in Single Cells. *Cell Rep.* **14,** 380–389 (2016).

51. Bernard, A. *et al.* Transcriptional architecture of the primate neocortex. *Neuron* **73,** 1083–1099 (2012).

52. Toma, K. & Hanashima, C. Switching modes in corticogenesis: mechanisms of neuronal subtype transitions and integration in the cerebral cortex. *Front. Neurosci.* **9,** 274 (2015).

53. Wonders, C. P. & Anderson, S. A. The origin and specification of cortical interneurons. *Nat. Rev. Neurosci.* **7,** 687–696 (2006).

54. Belgard, T. G. *et al.* A transcriptomic atlas of mouse neocortical layers. *Neuron* **71,** 605–616 (2011).

55. Kang, H. J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478,** 483–489 (2011).

56. Pletikos, M. *et al.* Temporal specification and bilaterality of human neocortical topographic gene expression. *Neuron* **81,** 321–332 (2014).

57. Sharma, K. *et al.* Cell type- and brain region-resolved mouse brain proteome. *Nat. Neurosci.* (2015). doi:10.1038/nn.4160

58. Saito, T. *et al.* Neocortical layer formation of human developing brains and lissencephalies: consideration of layer-specific marker expression. *Cereb. Cortex N. Y. N 1991* **21,** 588–596 (2011).

59. Gaspard, N. *et al.* An intrinsic mechanism of corticogenesis from embryonic stem cells. *Nature* **455,** 351–357 (2008).

60. Chambers, S. M. *et al.* Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. *Nat. Biotechnol.* **27,** 275–280 (2009).

61. Shi, Y., Kirwan, P. & Livesey, F. J. Directed differentiation of human pluripotent stem cells to cerebral cortex neurons and neural networks. *Nat. Protoc.* **7,** 1836–1846 (2012).

62. Pang, Z. P. *et al.* Induction of human neuronal cells by defined transcription factors. *Nature* **476,** 220–223 (2011).

63. Zou, Q. *et al.* Direct conversion of human fibroblasts into neuronal restricted progenitors. *J. Biol. Chem.* **289,** 5250–5260 (2014).

64. van de Leemput, J. *et al.* CORTECON: A Temporal Transcriptome Analysis of In Vitro Human Cerebral Cortex Development from Human Embryonic Stem Cells. *Neuron* **83,** 51–68 (2014).

65. Paşca, A. M. *et al.* Functional cortical neurons and astrocytes from human pluripotent stem cells in 3D culture. *Nat. Methods* **12,** 671–8 (2015).

66. Zaitsev, A. V. *et al.* Interneuron Diversity in Layers 2–3 of Monkey Prefrontal Cortex. *Cereb. Cortex* **19,** 1597–1615 (2009).

67. Zaitsev, A. V., Povysheva, N. V., Gonzalez-Burgos, G. & Lewis, D. A. Electrophysiological classes of layer 2/3 pyramidal cells in monkey prefrontal cortex. *J. Neurophysiol.* **108,** 595–609 (2012).

68. Zhang, Y. *et al.* An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *J. Neurosci. Off. J. Soc. Neurosci.* **34,** 11929–11947 (2014).

69. Fuzik, J. *et al.* Integration of electrophysiological recordings with single-cell RNA-seq data identifies neuronal subtypes. *Nat. Biotechnol.* (2015). doi:10.1038/nbt.3443

70. Lovatt, D. *et al.* Transcriptome in vivo analysis (TIVA) of spatially defined single cells in live tissue. *Nat. Methods* **11,** 190–196 (2014).

71. Lee, J. H. *et al.* Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat. Protoc.* **10,** 442–458 (2015).

72. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348,** aaa6090 (2015).

73. Silberberg, D., Anand, N. P., Michels, K. & Kalaria, R. N. Brain and other nervous system disorders across the lifespan — global challenges and opportunities. *Nature* **527,** S151–S154 (2015).

74. Seifert, G., Schilling, K. & Steinhäuser, C. Astrocyte dysfunction in neurological disorders: a molecular perspective. *Nat. Rev. Neurosci.* **7,** 194–206 (2006).

75. Hanisch, U.-K. & Kettenmann, H. Microglia: active sensor and versatile effector cells in the normal and pathologic brain. *Nat. Neurosci.* **10,** 1387–1394 (2007).

76. Zlokovic, B. V. The Blood-Brain Barrier in Health and Chronic Neurodegenerative Disorders. *Neuron* **57,** 178–201 (2008).

77. Zhang, F., Lin, Y.-A., Kannan, S. & Kannan, R. M. Targeting specific cells in the brain with nanomedicines for CNS therapies. *J. Control. Release Off. J. Control. Release Soc.* (2015). doi:10.1016/j.jconrel.2015.12.013

78. Gubitz, A. K. & Gwinn, K. Mining the genome for susceptibility to complex neurological disorders. *Curr. Mol. Med.* **9,** 801–813 (2009).

79. Tan, M.-S., Jiang, T., Tan, L. & Yu, J.-T. Genome-wide association studies in neurology. *Ann. Transl. Med.* **2,** 124 (2014).

80. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337,** 1190–1195 (2012).

81. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518,** 337–343 (2015).

82. Ballard, C. *et al.* Alzheimer's disease. *Lancet* **377,** 1019–1031 (2011).

83. Winkler, E. A. *et al.* GLUT1 reductions exacerbate Alzheimer's disease vasculo-neuronal dysfunction and degeneration. *Nat. Neurosci.* **18,** 521–530 (2015).

84. Pekny, M. *et al.* Astrocytes: a central element in neurological diseases. *Acta Neuropathol. (Berl.)* (2015). doi:10.1007/s00401-015-1513-1

85. Moore, S. *et al.* APP Metabolism Regulates Tau Proteostasis in Human Cerebral Cortex Neurons. *Cell Rep.* **11,** 689–696 (2015).

86. Gjoneska, E. *et al.* Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature* **518,** 365–369 (2015).

87. Winkler, J. M. & Fox, H. S. Transcriptome meta-analysis reveals a central role for sex steroids in the degeneration of hippocampal neurons in Alzheimer's disease. *BMC Syst. Biol.* **7,** 51 (2013).

88. Gómez-Isla, T. *et al.* Profound loss of layer II entorhinal cortex neurons occurs in very mild Alzheimer's disease. *J. Neurosci. Off. J. Soc. Neurosci.* **16,** 4491–4500 (1996).

89. Romito-DiGiacomo, R. R., Menegay, H., Cicero, S. A. & Herrup, K. Effects of Alzheimer's disease on different cortical layers: the role of intrinsic differences in Abeta susceptibility. *J. Neurosci. Off. J. Soc. Neurosci.* **27,** 8496–8504 (2007).

90. Livesey, F. J. Human stem cell models of dementia. *Hum. Mol. Genet.* **23,** R35-39 (2014).

91. Shi, Y., Kirwan, P., Smith, J., Robinson, H. P. C. & Livesey, F. J. Human cerebral cortex development from pluripotent stem cells to functional excitatory synapses. *Nat. Neurosci.* **15,** 477–486, S1 (2012).

92. Yagi, T. *et al.* Modeling familial Alzheimer's disease with induced pluripotent stem cells. *Hum. Mol. Genet.* **20,** 4530–4539 (2011).

93. Israel, M. A. *et al.* Probing sporadic and familial Alzheimer's disease using induced pluripotent stem cells. *Nature* **482,** 216–220 (2012).

94. Rouhani, F. *et al.* Genetic background drives transcriptional variation in human induced pluripotent stem cells. *PLoS Genet.* **10,** e1004432 (2014).

95. Woodruff, G. *et al.* The presenilin-1 ΔE9 mutation results in reduced γ-secretase activity, but not total loss of PS1 function, in isogenic human stem cells. *Cell Rep.* **5,** 974–985 (2013).

96. Haimovich, A. D., Muir, P. & Isaacs, F. J. Genomes by design. *Nat. Rev. Genet.* **16,** 501–516 (2015).

97. Hamazaki, Y. Adult thymic epithelial cell (TEC) progenitors and TEC stem cells: Models and mechanisms for TEC development and maintenance. *Eur. J. Immunol.* (2015). doi:10.1002/eji.201545844

98. Savino, W., Mendes-da-Cruz, D. A., Lepletier, A. & Dardenne, M. Hormonal control of T-cell development in health and disease. *Nat. Rev. Endocrinol.* (2015). doi:10.1038/nrendo.2015.168

99. Shah, D. K. & Zúñiga-Pflücker, J. C. An overview of the intrathymic intricacies of T cell development. *J. Immunol. Baltim. Md 1950* **192,** 4017–4023 (2014).

100. Goldman, K. P., Park, C.-S., Kim, M., Matzinger, P. & Anderson, C. C. Thymic cortical epithelium induces self tolerance. *Eur. J. Immunol.* **35,** 709–717 (2005).

101. Stritesky, G. L. *et al.* Murine thymic selection quantified using a unique method to capture deleted T cells. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 4679–4684 (2013).

102. Ohigashi, I. *et al.* Aire-expressing thymic medullary epithelial cells originate from β5t-expressing progenitor cells. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 9885–9890 (2013).

103. Coffer, P. J. & Burgering, B. M. T. Forkhead-box transcription factors and their role in the immune system. *Nat. Rev. Immunol.* **4,** 889–899 (2004).

104. Nehls, M., Pfeifer, D., Schorpp, M., Hedrich, H. & Boehm, T. New member of the winged-helix protein family disrupted in mouse and rat nude mutations. *Nature* **372,** 103–107 (1994).

105. Pignata, C. *et al.* Congenital Alopecia and nail dystrophy associated with severe functional T-cell immunodeficiency in two sibs. *Am. J. Med. Genet.* **65,** 167–170 (1996).

106. Cheng, L. *et al.* Postnatal Tissue-specific Disruption of Transcription Factor FoxN1 Triggers Acute Thymic Atrophy. *J. Biol. Chem.* **285,** 5836–5847 (2010).

107. Nowell, C. S. *et al.* Foxn1 regulates lineage progression in cortical and medullary thymic epithelial cells but is dispensable for medullary sublineage divergence. *PLoS Genet.* **7,** e1002348 (2011).

108. Sun, L., Li, H., Luo, H. & Zhao, Y. Thymic epithelial cell development and its dysfunction in human diseases. *BioMed Res. Int.* **2014,** 206929 (2014).

109. Abramson, J. & Goldfarb, Y. AIRE: From promiscuous molecular partnerships to promiscuous gene expression. *Eur. J. Immunol.* (2015). doi:10.1002/eji.201545792

110. Finnish-German APECED Consortium. An autoimmune disease, APECED, caused by mutations in a novel gene featuring two PHD-type zinc-finger domains. *Nat. Genet.* **17,** 399–403 (1997).

111. Nagamine, K. *et al.* Positional cloning of the APECED gene. *Nat. Genet.* **17,** 393–398 (1997).

112. Liston, A., Lesage, S., Wilson, J., Peltonen, L. & Goodnow, C. C. Aire regulates negative selection of organ-specific T cells. *Nat. Immunol.* **4,** 350–354 (2003).

113. Sansom, S. N. *et al.* Population and single cell genomics reveal the Aire-dependency, relief from Polycomb silencing and distribution of self-antigen expression in thymic epithelia. *Genome Res.* (2014). doi:10.1101/gr.171645.113

114. Brennecke, P. *et al.* Single-cell transcriptome analysis reveals coordinated ectopic gene-expression patterns in medullary thymic epithelial cells. *Nat. Immunol.* **16,** 933–941 (2015).

115. Meredith, M., Zemmour, D., Mathis, D. & Benoist, C. Aire controls gene expression in the thymic epithelium with ordered stochasticity. *Nat. Immunol.* **16,** 942–949 (2015).

116. Takaba, H. *et al.* Fezf2 Orchestrates a Thymic Program of Self-Antigen Expression for Immune Tolerance. *Cell* **163,** 975–987 (2015).

117. Verschuren, M. C. *et al.* Preferential rearrangements of the T cell receptor-delta-deleting elements in human T cells. *J. Immunol. Baltim. Md 1950* **158,** 1208–1216 (1997).

118. Kayser, C., Alberto, F. L., da Silva, N. P. & Andrade, L. E. C. Decreased number of T cells bearing TCR rearrangement excision circles (TREC) in active recent onset systemic lupus erythematosus. *Lupus* **13,** 906–911 (2004).

119. Thewissen, M. *et al.* Analyses of immunosenescent markers in patients with autoimmune disease. *Clin. Immunol. Orlando Fla* **123,** 209–218 (2007).

120. Hofer, J. *et al.* Elevated proportions of recent thymic emigrants in children and adolescents with type 1 diabetes. *Rejuvenation Res.* **12,** 311–320 (2009).

121. Strawa, K. *et al.* Increased concentration of T-cell receptor rearrangement excision circles (TREC) in peripheral blood in Graves' disease. *Clin. Endocrinol. (Oxf.)* **81,** 769–774 (2014).

122. Hazenberg, M. D. *et al.* Increased cell division but not thymic dysfunction rapidly affects the T-cell receptor excision circle content of the naive T cell population in HIV-1 infection. *Nat. Med.* **6,** 1036–1042 (2000).

123. Qiang, L., Fujita, R. & Abeliovich, A. Remodeling neurodegeneration: somatic cell reprogramming-based models of adult neurological disorders. *Neuron* **78,** 957–969 (2013).

124. Zhang, Y. *et al.* Rapid single-step induction of functional neurons from human pluripotent stem cells. *Neuron* **78,** 785–798 (2013).

125. Espuny-Camacho, I. *et al.* Pyramidal neurons derived from human pluripotent stem cells integrate efficiently into mouse brain circuits in vivo. *Neuron* **77,** 440–456 (2013).

126. Mariani, J. *et al.* Modeling human cortical development in vitro using induced pluripotent stem cells. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 12770–12775 (2012).

127. Lin, M. *et al.* RNA-Seq of human neurons derived from iPS cells reveals candidate long non-coding RNAs involved in neurogenesis and neuropsychiatric disorders. *PloS One* **6,** e23356 (2011).

128. Ståhlberg, A., Rusnakova, V., Forootan, A., Anderova, M. & Kubista, M. RT-qPCR work-flow for single-cell data analysis. *Methods San Diego Calif* **59,** 80–88 (2013).

129. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* (2013). doi:10.1038/nmeth.2645

130. Pasca, S. P. *et al.* Using iPS cell-derived neurons to uncover cellular phenotypes associated with Timothy Syndrome. *Nat. Med.* **17,** 1657–1662 (2011).

131. Mariani, J. *et al.* FOXG1-Dependent Dysregulation of GABA/Glutamate Neuron Differentiation in Autism Spectrum Disorders. *Cell* **162,** 375–390 (2015).

132. Hartfield, E. M. *et al.* Physiological characterisation of human iPS-derived dopaminergic neurons. *PloS One* **9,** e87388 (2014).

133. Nishimura, K. *et al.* Development of defective and persistent Sendai virus vector: a unique gene delivery/expression system ideal for cell reprogramming. *J. Biol. Chem.* **286,** 4760–4771 (2011).

134. Nishimura, T. *et al.* Generation of rejuvenated antigen-specific T cells by reprogramming to pluripotency and redifferentiation. *Cell Stem Cell* **12,** 114–126 (2013).

135. Chia, R., Achilli, F., Festing, M. F. W. & Fisher, E. M. C. The origins and uses of mouse outbred stocks. *Nat. Genet.* **37,** 1181–1186 (2005).

136. Gardner, R. L. Investigation of cell lineage and differentiation in the extraembryonic endoderm of the mouse embryo. *J. Embryol. Exp. Morphol.* **68,** 175–198 (1982).

137. Beers, J. *et al.* Passaging and colony expansion of human pluripotent stem cells by enzyme-free dissociation in chemically defined culture conditions. *Nat. Protoc.* **7,** 2029–2040 (2012).

138. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods San Diego Calif* **25,** 402–408 (2001).

139. Müller, F.-J. *et al.* A bioinformatic assay for pluripotency in human cells. *Nat. Methods* **8,** 315–317 (2011).

140. Zeng, H. *et al.* Large-scale cellular-resolution gene profiling in human neocortex reveals species-specific molecular signatures. *Cell* **149,** 483–496 (2012).

141. Hawrylycz, M. J. *et al.* An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489,** 391–399 (2012).

142. Miller, J. A. *et al.* Transcriptional landscape of the prenatal human brain. *Nature* **508,** 199–206 (2014).

143. Fluidigm. Using C1 to Generate Single-Cell cDNA Libraries for mRNA Sequencing.

144. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinforma. Oxf. Engl.* **30,** 2114–2120 (2014).

145. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).

146. Liao, Y., Smyth, G. K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* **41,** e108 (2013).

147. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).

148. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11,** 163–166 (2014).

149. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4,** 44–57 (2009).

150. Hjelm, B. E. *et al.* In vitro-differentiated neural cell cultures progress towards donor-identical brain tissue. *Hum. Mol. Genet.* **22,** 3534–3546 (2013).

151. Camp, J. G. *et al.* Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 15672–15677 (2015).

152. Sposito, T. *et al.* Developmental regulation of tau splicing is disrupted in stem cell-derived neurons from frontotemporal dementia patients with the 10 + 16 splice-site mutation in MAPT. *Hum. Mol. Genet.* (2015). doi:10.1093/hmg/ddv246

153. Buxbaum, A. R., Yoon, Y. J., Singer, R. H. & Park, H. Y. Single-molecule insights into mRNA dynamics in neurons. *Trends Cell Biol.* **25,** 468–475 (2015).

154. Sakakibara, T. *et al.* Delayed maturation and differentiation of neurons in focal cortical dysplasia with the transmantle sign: analysis of layer-specific marker expression. *J. Neuropathol. Exp. Neurol.* **71,** 741–749 (2012).

155. Knock, E. *et al.* The methyl binding domain 3/nucleosome remodelling and deacetylase complex regulates neural cell fate determination and terminal differentiation in the cerebral cortex. *Neural Develop.* **10,** 13 (2015).

156. Stoner, R. *et al.* Patches of disorganization in the neocortex of children with autism. *N. Engl. J. Med.* **370,** 1209–1219 (2014).

157. Alzheimer's Association. 2015 Alzheimer's disease facts and figures. *Alzheimers Dement. J. Alzheimers Assoc.* **11,** 332–384 (2015).

158. Tanzi, R. E. & Bertram, L. Twenty years of the Alzheimer's disease amyloid hypothesis: a genetic perspective. *Cell* **120,** 545–555 (2005).

159. Hardy, J. A. & Higgins, G. A. Alzheimer's disease: the amyloid cascade hypothesis. *Science* **256,** 184–185 (1992).

160. Yau, W.-Y. W. *et al.* Longitudinal assessment of neuroimaging and clinical markers in autosomal dominant Alzheimer's disease: a prospective cohort study. *Lancet Neurol.* **14,** 804–813 (2015).

161. Hardy, J. The amyloid hypothesis for Alzheimer's disease: a critical reappraisal. *J. Neurochem.* **110,** 1129–1134 (2009).

162. Giacobini, E. & Gold, G. Alzheimer disease therapy--moving from amyloid-β to tau. *Nat. Rev. Neurol.* **9,** 677–686 (2013).

163. Nelson, P. T. *et al.* Correlation of Alzheimer disease neuropathologic changes with cognitive status: a review of the literature. *J. Neuropathol. Exp. Neurol.* **71,** 362–381 (2012).

164. Li, J.-Q. *et al.* Risk factors for predicting progression from mild cognitive impairment to Alzheimer's disease: a systematic review and meta-analysis of cohort studies. *J. Neurol. Neurosurg. Psychiatry* (2015). doi:10.1136/jnnp-2014-310095

165. Shi, Y. *et al.* A human stem cell model of early Alzheimer's disease pathology in Down syndrome. *Sci. Transl. Med.* **4,** 124ra29 (2012).

166. Kondo, T. *et al.* Modeling Alzheimer's disease with iPSCs reveals stress phenotypes associated with intracellular Aβ and differential drug responsiveness. *Cell Stem Cell* **12,** 487–496 (2013).

167. Briggs, J. A. *et al.* Integration-free induced pluripotent stem cells model genetic and neural developmental features of down syndrome etiology. *Stem Cells Dayt. Ohio* **31,** 467–478 (2013).

168. Weick, J. P. *et al.* Deficits in human trisomy 21 iPSCs and neurons. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 9962–9967 (2013).

169. Liu, Q. *et al.* Effect of potent γ-secretase modulator in human neurons derived from multiple presenilin 1-induced pluripotent stem cell mutant carriers. *JAMA Neurol.* **71,** 1481–1489 (2014).

170. Iovino, M. *et al.* Early maturation and distinct tau pathology in induced pluripotent stem cell-derived neurons from patients with MAPT mutations. *Brain J. Neurol.* (2015). doi:10.1093/brain/awv222

171. Willem, M. *et al.* η-Secretase processing of APP inhibits neuronal activity in the hippocampus. *Nature* **526,** 443–447 (2015).

172. Miller, J. D. *et al.* Human iPSC-based modeling of late-onset disease via progerin-induced aging. *Cell Stem Cell* **13,** 691–705 (2013).

173. Mertens, J. *et al.* Directly Reprogrammed Human Neurons Retain Aging-Associated Transcriptomic Signatures and Reveal Age-Related Nucleocytoplasmic Defects. *Cell Stem Cell* (2015). doi:10.1016/j.stem.2015.09.001

174. Perez-Tur, J. *et al.* A mutation in Alzheimer's disease destroying a splice acceptor site in the presenilin-1 gene. *Neuroreport* **7,** 297–301 (1995).

175. Hutton, M. *et al.* Complete analysis of the presenilin 1 gene in early onset Alzheimer's disease. *Neuroreport* **7,** 801–805 (1996).

176. Tysoe, C. *et al.* A presenilin-1 truncating mutation is present in two cases with autopsy-confirmed early-onset Alzheimer disease. *Am. J. Hum. Genet.* **62,** 70–76 (1998).

177. De Jonghe, C. *et al.* Aberrant splicing in the presenilin-1 intron 4 mutation causes presenile Alzheimer's disease by increased Abeta42 secretion. *Hum. Mol. Genet.* **8,** 1529–1540 (1999).

178. Gore, A. *et al.* Somatic coding mutations in human induced pluripotent stem cells. *Nature* **471,** 63–67 (2011).

179. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. (Accessed: 8th July 2015)

180. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12,** 357–360 (2015).

181. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinforma. Oxf. Engl.* **31,** 166–169 (2015).

182. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

183. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2009).

184. Peixoto, L. *et al.* How data analysis affects power, reproducibility and biological insight of RNA-seq studies in complex datasets. *Nucleic Acids Res.* **43,** 7664–7674 (2015).

185. Heywood, W. E. *et al.* Identification of novel CSF biomarkers for neurodegeneration and their validation by a high-throughput multiplexed targeted proteomic assay. *Mol. Neurodegener.* **10,** 64 (2015).

186. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** D1001–D1006 (2014).

187. Hochberg, Y. & Benjamini, Y. More powerful procedures for multiple significance testing. *Stat. Med.* **9,** 811–818 (1990).

188. Readhead, B. *et al.* Molecular systems evaluation of oligomerogenic APP(E693Q) and fibrillogenic APP(KM670/671NL)/PSEN1(Δexon9) mouse models identifies shared features with human Alzheimer's brain molecular pathology. *Mol. Psychiatry* (2015). doi:10.1038/mp.2015.167

189. Zhang, B. *et al.* Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* **153,** 707–720 (2013).

190. Mather, K. A. *et al.* Investigating the Genetics of Hippocampal Volume in Older Adults without Dementia. *PLOS ONE* **10,** e0116920 (2015).

191. Hwang, I. K. *et al.* Age-dependent changes of pyridoxal phosphate synthesizing enzymes immunoreactivities and activities in the gerbil hippocampal CA1 region. *Mech. Ageing Dev.* **126,** 1322–1330 (2005).

192. Kwak, S.-E. *et al.* Pyridoxine 5'-phosphate oxidase, not pyridoxal kinase, involves in long-term potentiation induction in the rat dentate gyrus. *Hippocampus* **19,** 45–56 (2009).

193. Sherva, R. *et al.* Genome-wide association study of the rate of cognitive decline in Alzheimer's disease. *Alzheimers Dement. J. Alzheimers Assoc.* **10,** 45–52 (2014).

194. Kim, H.-Y., Choi, J.-S., Cha, J.-H., Choi, J.-Y. & Lee, M.-Y. Expression of vascular endothelial growth factor receptors Flt-1 and Flk-1 in embryonic rat forebrain. *Neurosci. Lett.* **425,** 131–135 (2007).

195. Sanchez, A., Wadhwani, S. & Grammas, P. Multiple neurotrophic effects of VEGF on cultured neurons. *Neuropeptides* **44,** 323–331 (2010).

196. Mendonça, M. C. P., Siqueira Soares, E., Miguel Stávale, L., Pierre Irazusta, S. & da Cruz-Höfling, M. A. Upregulation of the vascular endothelial growth factor, Flt-1, in rat hippocampal neurons after envenoming by Phoneutria nigriventer; age-related modulation. *Toxicon Off. J. Int. Soc. Toxinology* **60,** 656–664 (2012).

197. Ryu, J. K., Cho, T., Choi, H. B., Wang, Y. T. & McLarnon, J. G. Microglial VEGF receptor response is an integral chemotactic component in Alzheimer's disease pathology. *J. Neurosci. Off. J. Soc. Neurosci.* **29,** 3–13 (2009).

198. Need, A. C. *et al.* A genome-wide study of common SNPs and CNVs in cognitive performance in the CANTAB. *Hum. Mol. Genet.* **18,** 4650–4661 (2009).

199. Villela, D., Schlesinger, D., Suemoto, C. K., Grinberg, L. T. & Rosenberg, C. A microdeletion in Alzheimer's disease disrupts NAMPT gene. *J. Genet.* **93,** 535–537 (2014).

200. Ghosh, D., Levault, K. R. & Brewer, G. J. Relative importance of redox buffers GSH and NAD(P)H in age-related neurodegeneration and Alzheimer disease-like mouse neurons. *Aging Cell* **13,** 631–640 (2014).

201. Stein, L. R. *et al.* Expression of Nampt in hippocampal and cortical excitatory neurons is critical for cognitive function. *J. Neurosci. Off. J. Soc. Neurosci.* **34,** 5800–5815 (2014).

202. Harlan, B. A. *et al.* Enhancing NAD+salvage pathway reverts the toxicity of primary astrocytes expressing amyotrophic lateral sclerosis-linked mutant SOD1. *J. Biol. Chem.* (2016). doi:10.1074/jbc.M115.698779

203. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinforma. Oxf. Engl.* **26,** 139–140 (2010).

204. Rose, J. B. *et al.* Neuropeptide Y fragments derived from neprilysin processing are neuroprotective in a transgenic model of Alzheimer's disease. *J. Neurosci. Off. J. Soc. Neurosci.* **29,** 1115–1125 (2009).

205. Spencer, B. *et al.* Systemic Central Nervous System (CNS)-targeted Delivery of Neuropeptide Y (NPY) Reduces Neurodegeneration and Increases Neural Precursor Cell Proliferation in a Mouse Model of Alzheimer Disease. *J. Biol. Chem.* **291,** 1905–1920 (2016).

206. Boom, A. *et al.* Astrocytic calcium/zinc binding protein S100A6 over expression in Alzheimer's disease and in PS1/APP transgenic mice models. *Biochim. Biophys. Acta* **1742,** 161–168 (2004).

207. Hu, B.-Y. *et al.* Neural differentiation of human induced pluripotent stem cells follows developmental principles but with variable potency. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 4335–4340 (2010).

208. Lambert, J.-C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45,** 1452–1458 (2013).

209. Lyros, E., Bakogiannis, C., Liu, Y. & Fassbender, K. Molecular links between endothelial dysfunction and neurodegeneration in Alzheimer's disease. *Curr. Alzheimer Res.* **11,** 18–26 (2014).

210. Mann, D. M. A., Pickering-Brown, S. M., Takeuchi, A. & Iwatsubo, T. Amyloid Angiopathy and Variability in Amyloid β Deposition Is Determined by Mutation Position in Presenilin-1-Linked Alzheimer's Disease. *Am. J. Pathol.* **158,** 2165–2175 (2001).

211. Ringman, J. M. *et al.* Neuropathology of Autosomal Dominant Alzheimer Disease in the National Alzheimer Coordinating Center Database. *J. Neuropathol. Exp. Neurol.* (2016). doi:10.1093/jnen/nlv028

212. Kyttälä, A. *et al.* Genetic Variability Overrides the Impact of Parental Cell Type and Determines iPSC Differentiation Potential. *Stem Cell Rep.* **6,** 200–212 (2016).

213. Theda, L., Drews, M. K., Zitnik, G., Oshima, J. & Martin, G. M. Aβ 1-40 enhances the proliferation of human diploid fibroblasts. *Neurobiol. Aging* **38,** 11–13 (2016).

214. Xu, Y. *et al.* Proliferation rate of somatic cells affects reprogramming efficiency. *J. Biol. Chem.* **288,** 9767–9778 (2013).

215. Wang, B. *et al.* The amyloid precursor protein controls adult hippocampal neurogenesis through GABAergic interneurons. *J. Neurosci. Off. J. Soc. Neurosci.* **34,** 13314–13325 (2014).

216. Gibson, G. Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13,** 135–145 (2011).

217. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88,** 294–305 (2011).

218. Marei, H. E. *et al.* Common and Rare Genetic Variants Associated With Alzheimer's Disease. *J. Cell. Physiol.* **231,** 1432–1437 (2016).

219. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 9362–9367 (2009).

220. Heinzen, E. L. *et al.* Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS Biol.* **6,** e1 (2008).

221. Liu, C. *et al.* Whole-genome association mapping of gene expression in the human prefrontal cortex. *Mol. Psychiatry* **15,** 779–784 (2010).

222. Gibbs, J. R. *et al.* Abundant Quantitative Trait Loci Exist for DNA Methylation and Gene Expression in Human Brain. *PLoS Genet* **6,** e1000952 (2010).

223. Ramasamy, A. *et al.* Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nat. Neurosci.* **17,** 1418–1428 (2014).

224. Gaffney, D. J. *et al.* Dissecting the regulatory architecture of gene expression QTLs. *Genome Biol.* **13,** R7 (2012).

225. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482,** 390–394 (2012).

226. Levo, M. *et al.* Unraveling determinants of transcription factor binding outside the core binding site. *Genome Res.* (2015). doi:10.1101/gr.185033.114

227. Andrabi, M. *et al.* Genome-wide transcription factor activities are explained by intrinsic conformational dynamics of binding-sites and distal flanking-regions. *bioRxiv* 20602 (2015). doi:10.1101/020602

228. Cusanovich, D. A., Pavlovic, B., Pritchard, J. K. & Gilad, Y. The functional consequences of variation in transcription factor binding. *PLoS Genet.* **10,** e1004226 (2014).

229. Bell, O., Tiwari, V. K., Thomä, N. H. & Schübeler, D. Determinants and dynamics of genome accessibility. *Nat. Rev. Genet.* **12,** 554–564 (2011).

230. Kundaje, A. ENCODE: TF ChIP-seq peak calling using the Irreproducibility Discovery Rate (IDR) framework. Available at: https://sites.google.com/site/anshulkundaje/projects/idr. (Accessed: 22nd March 2014)

231. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

232. Piper, J. *et al.* Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res.* **41,** e201 (2013).

233. Slattery, M. *et al.* Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.* **39,** 381–399 (2014).

234. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28,** 1045–1048 (2010).

235. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).

236. Boyle, A. P., Guinney, J., Crawford, G. E. & Furey, T. S. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinforma. Oxf. Engl.* **24,** 2537–2538 (2008).

237. Kundaje, A. A comprehensive collection of signal artifact blacklist regions in the human genome. *ENCODE* (2013).

238. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinforma. Oxf. Engl.* **27,** 2156–2158 (2011).

239. Consortium, T. 1000 G. P. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

240. Heger, A., Webber, C., Goodson, M., Ponting, C. P. & Lunter, G. GAT: a simulation framework for testing the association of genomic intervals. *Bioinforma. Oxf. Engl.* **29,** 2046–2048 (2013).

241. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinforma. Oxf. Engl.* **27,** 1017–1018 (2011).

242. Mathelier, A. *et al.* JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42,** D142-147 (2014).

243. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507,** 455–461 (2014).

244. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38,** 576–589 (2010).

245. Corradin, O. *et al.* Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res.* **24,** 1–13 (2014).

246. Sundaresan, V. *et al.* Dynamic expression patterns of Robo (Robo1 and Robo2) in the developing murine central nervous system. *J. Comp. Neurol.* **468,** 467–481 (2004).

247. Lahti, L. *et al.* Differentiation and molecular heterogeneity of inhibitory and excitatory neurons associated with midbrain dopaminergic nuclei. *Dev. Camb. Engl.* (2015). doi:10.1242/dev.129957

248. McGowan, P. O. *et al.* Epigenetic regulation of the glucocorticoid receptor in human brain associates with childhood abuse. *Nat. Neurosci.* **12,** 342–348 (2009).

249. Elf, J., Li, G.-W. & Xie, X. S. Probing Transcription Factor Dynamics at the Single-Molecule Level in a Living Cell. *Science* **316,** 1191–1194 (2007).

250. Mo, A. *et al.* Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. *Neuron* **86,** 1369–1384 (2015).

251. Qu, K. *et al.* Individuality and variation of personal regulomes in primary human T cells. *Cell Syst.* **1,** 51–61 (2015).

252. Wen, X., Luca, F. & Pique-Regi, R. Cross-Population Joint Analysis of eQTLs: Fine Mapping and Functional Annotation. *PLOS Genet* **11,** e1005176 (2015).

253. Maurano, M. T. *et al.* Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat. Genet.* **47,** 1393–1401 (2015).

254. Moyerbrailean, G. A. *et al.* Which Genetics Variants in DNase-Seq Footprints Are More Likely to Alter Binding? *PLOS Genet* **12,** e1005875 (2016).

255. Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. Beyond GWASs: illuminating the dark road from association to function. *Am. J. Hum. Genet.* **93,** 779–797 (2013).

256. Savic, D. *et al.* Alterations in TCF7L2 expression define its role as a key regulator of glucose metabolism. *Genome Res.* **21,** 1417–1425 (2011).

257. Webber, J. L., Zhang, J., Mitchell-Dick, A. & Rebay, I. 3D chromatin interactions organize Yan chromatin occupancy and repression at the even-skipped locus. *Genes Dev.* **27,** 2293–2298 (2013).

258. Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518,** 331–336 (2015).

259. Jolma, A. *et al.* DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527,** 384–388 (2015).

260. Holländer, G. *et al.* Cellular and molecular events during early thymus development. *Immunol. Rev.* **209,** 28–46 (2006).

261. Nehls, M. *et al.* Two genetically separable steps in the differentiation of thymic epithelium. *Science* **272,** 886–889 (1996).

262. Bleul, C. C. *et al.* Formation of a functional thymus initiated by a postnatal epithelial progenitor cell. *Nature* **441,** 992–996 (2006).

263. Bredenkamp, N., Nowell, C. S. & Blackburn, C. C. Regeneration of the aged thymus by a single transcription factor. *Dev. Camb. Engl.* **141,** 1627–1637 (2014).

264. Corbeaux, T. *et al.* Thymopoiesis in mice depends on a Foxn1-positive thymic epithelial cell lineage. *Proc. Natl. Acad. Sci.* **107,** 16613–16618 (2010).

265. Zook, E. C. *et al.* Overexpression of Foxn1 attenuates age-associated thymic involution and prevents the expansion of peripheral CD4 memory T cells. *Blood* **118,** 5723–5731 (2011).

266. Reeh, K. A. G. *et al.* Ectopic TBX1 suppresses thymic epithelial cell differentiation and proliferation during thymus organogenesis. *Dev. Camb. Engl.* **141,** 2950–2958 (2014).

267. Weiner, L. *et al.* Dedicated epithelial recipient cells determine pigmentation patterns. *Cell* **130,** 932–942 (2007).

268. Tsai, P. T., Lee, R. A. & Wu, H. BMP4 acts upstream of FGF in modulating thymic stroma and regulating thymopoiesis. *Blood* **102,** 3947–3953 (2003).

269. Cai, J., Lee, J., Kopan, R. & Ma, L. Genetic interplays between Msx2 and Foxn1 are required for Notch1 expression and hair shaft differentiation. *Dev. Biol.* **326,** 420–430 (2009).

270. Ripen, A. M., Nitta, T., Murata, S., Tanaka, K. & Takahama, Y. Ontogeny of thymic cortical epithelial cells expressing the thymoproteasome subunit β5t. *Eur. J. Immunol.* **41,** 1278–1287 (2011).

271. Schlake, T., Schorpp, M., Nehls, M. & Boehm, T. The nude gene encodes a sequence-specific DNA binding protein with homologs in organisms that lack an anticipatory immune system. *Proc. Natl. Acad. Sci. U. S. A.* **94,** 3842–3847 (1997).

272. Nakagawa, S., Gisselbrecht, S. S., Rogers, J. M., Hartl, D. L. & Bulyk, M. L. DNA-binding specificity changes in the evolution of forkhead transcription factors. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 12349–12354 (2013).

273. Perl, A.-K. T., Wert, S. E., Nagy, A., Lobe, C. G. & Whitsett, J. A. Early restriction of peripheral and proximal cell lineages during formation of the lung. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 10482–10487 (2002).

274. Mayer, C. E. *et al.* Dynamic spatio-temporal contribution of single β5t+ cortical epithelial precursors to the thymus medulla. *Eur. J. Immunol.* (2015). doi:10.1002/eji.201545995

275. Zuklys, S. *et al.* Stabilized beta-catenin in thymic epithelial cells blocks thymus development and function. *J. Immunol. Baltim. Md 1950* **182,** 2997–3007 (2009).

276. Vokes, S. A. *et al.* Genomic characterization of Gli-activator targets in sonic hedgehog-mediated neural patterning. *Dev. Camb. Engl.* **134,** 1977–1989 (2007).

277. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **25,** 1754–1760 (2009).

278. Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* **21,** 936–939 (2011).

279. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).

280. Chung, D. *et al.* Discovering Transcription Factor Binding Sites in Highly Repetitive Regions of Genomes with Multi-Read Analysis of ChIP-Seq Data. *PLOS Comput Biol* **7,** e1002111 (2011).

281. Wang, J. *et al.* Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* **22,** 1798–1812 (2012).

282. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5,** 1752–1779 (2011).

283. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37,** W202-208 (2009).

284. Zambelli, F., Pesole, G. & Pavesi, G. PscanChIP: Finding over-represented transcription factor-binding site motifs and their correlations in sequences from ChIP-Seq experiments. *Nucleic Acids Res.* **41,** W535-543 (2013).

285. Wang, S. *et al.* Target analysis by integration of transcriptome and ChIP-seq data with BETA. *Nat. Protoc.* **8,** 2502–2515 (2013).

286. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7,** 562–578 (2012).

287. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol. Ed. Frederick M Ausubel Al* **109,** 21.29.1-9 (2015).

288. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).

289. Ki, S. *et al.* Global transcriptional profiling reveals distinct functions of thymic stromal subsets and age-related changes during thymic involution. *Cell Rep.* **9,** 402–415 (2014).

290. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 559 (2008).

291. Impey, S. *et al.* Defining the CREB regulon: a genome-wide analysis of transcription factor regulatory regions. *Cell* **119,** 1041–1054 (2004).

292. Burnley, P. *et al.* Role of the p63-FoxN1 regulatory axis in thymic epithelial cell homeostasis during aging. *Cell Death Dis.* **4,** e932 (2013).

293. Candi, E. *et al.* DeltaNp63 regulates thymic development through enhanced expression of FgfR2 and Jag2. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 11999–12004 (2007).

294. Senoo, M., Pinto, F., Crum, C. P. & McKeon, F. p63 Is essential for the proliferative potential of stem cells in stratified epithelia. *Cell* **129,** 523–536 (2007).

295. Della Gatta, G. *et al.* Direct targets of the TRP63 transcription factor revealed by a combination of gene expression profiling and reverse engineering. *Genome Res.* **18,** 939–948 (2008).

296. Murata, S. *et al.* Regulation of CD8+ T cell development by thymus-specific proteasomes. *Science* **316,** 1349–1353 (2007).

297. Fujimoto, Y. *et al.* CD83 Expression Influences CD4+ T Cell Development in the Thymus. *Cell* **108,** 755–767 (2002).

298. García-Martínez, L. F. *et al.* A novel mutation in CD83 results in the development of a unique population of CD4+ T cells. *J. Immunol. Baltim. Md 1950* **173,** 2995–3001 (2004).

299. Schorpp, M., Hofmann, M., Dear, T. N. & Boehm, T. Characterization of mouse and human nude genes. *Immunogenetics* **46,** 509–515 (1997).

300. Hannenhalli, S. & Kaestner, K. H. The evolution of Fox genes and their role in development and disease. *Nat. Rev. Genet.* **10,** 233–240 (2009).

301. Overdier, D. G., Porcella, A. & Costa, R. H. The DNA-binding specificity of the hepatocyte nuclear factor 3/forkhead domain is influenced by amino-acid residues adjacent to the recognition helix. *Mol. Cell. Biol.* **14,** 2755–2766 (1994).

302. Pierrou, S., Hellqvist, M., Samuelsson, L., Enerbäck, S. & Carlsson, P. Cloning and characterization of seven human forkhead proteins: binding site specificity and DNA bending. *EMBO J.* **13,** 5002–5012 (1994).

303. Schuur, E. R. *et al.* Ligand-dependent interaction of estrogen receptor-alpha with members of the forkhead transcription factor family. *J. Biol. Chem.* **276,** 33554–33560 (2001).

304. Cirillo, L. A. *et al.* Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Mol. Cell* **9,** 279–289 (2002).

305. Zhang, L., Rubins, N. E., Ahima, R. S., Greenbaum, L. E. & Kaestner, K. H. Foxa2 integrates the transcriptional response of the hepatocyte to fasting. *Cell Metab.* **2,** 141–148 (2005).

306. Gao, N. *et al.* Forkhead box A1 regulates prostate ductal morphogenesis and promotes epithelial cell maturation. *Dev. Camb. Engl.* **132,** 3431–3443 (2005).

307. Carroll, J. S. *et al.* Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell* **122,** 33–43 (2005).

308. Liu, B. *et al.* Cbx4 regulates the proliferation of thymic epithelial cells and thymus function. *Dev. Camb. Engl.* **140,** 780–788 (2013).

309. Mardaryev, A. N. *et al.* Cbx4 maintains the epithelial lineage identity and cell proliferation in the developing stratified epithelium. *J. Cell Biol.* **212,** 77–89 (2016).

310. Bleul, C. C. & Boehm, T. Chemokines define distinct microenvironments in the developing thymus. *Eur. J. Immunol.* **30,** 3371–3379 (2000).

311. Tsukamoto, N., Itoi, M., Nishikawa, M. & Amagai, T. Lack of Delta like 1 and 4 expressions in nude thymus anlages. *Cell. Immunol.* **234,** 77–80 (2005).

312. Rode, I. *et al.* Foxn1 Protein Expression in the Developing, Aging, and Regenerating Thymus. *J. Immunol. Baltim. Md 1950* (2015). doi:10.4049/jimmunol.1502010

313. Yano, M. *et al.* Aire controls the differentiation program of thymic epithelial cells in the medulla for the establishment of self-tolerance. *J. Exp. Med.* **205,** 2827–2838 (2008).

314. Gardner, J. M. *et al.* Deletional Tolerance Mediated by Extrathymic Aire-Expressing Cells. *Science* **321,** 843–847 (2008).

315. Gardner, J. M., Fletcher, A. L., Anderson, M. S. & Turley, S. J. AIRE in the thymus and beyond. *Curr. Opin. Immunol.* **21,** 582–589 (2009).

316. Gardner, J. M. *et al.* Extrathymic Aire-Expressing Cells Are a Distinct Bone Marrow-Derived Population that Induce Functional Inactivation of CD4+ T Cells. *Immunity* **39,** 560–572 (2013).

317. Yamano, T. *et al.* Thymic B Cells Are Licensed to Present Self Antigens for Central T Cell Tolerance Induction. *Immunity* **42,** 1048–1061 (2015).

318. Kawano, H. *et al.* Aire Expression Is Inherent to Most Medullary Thymic Epithelial Cells during Their Differentiation Program. *J. Immunol. Baltim. Md 1950* **195,** 5149–5158 (2015).

319. Rattay, K. *et al.* Homeodomain-interacting protein kinase 2, a novel autoimmune regulator interaction partner, modulates promiscuous gene expression in medullary thymic epithelial cells. *J. Immunol. Baltim. Md 1950* **194,** 921–928 (2015).

320. LaFlam, T. N. *et al.* Identification of a novel cis-regulatory element essential for immune tolerance. *J. Exp. Med.* **212,** 1993–2002 (2015).

321. Oven, I. *et al.* AIRE recruits P-TEFb for transcriptional elongation of target genes in medullary thymic epithelial cells. *Mol. Cell. Biol.* **27,** 8815–8823 (2007).

322. Org, T. *et al.* AIRE activated tissue specific genes have histone modifications associated with inactive chromatin. *Hum. Mol. Genet.* **18,** 4699–4710 (2009).

323. Giraud, M. *et al.* Aire unleashes stalled RNA polymerase to induce ectopic gene expression in thymic epithelial cells. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 535–540 (2012).

324. Wu, G. *et al.* DNA methylation profile of Aire-deficient mouse medullary thymic epithelial cells. *BMC Immunol.* **13,** 58 (2012).

325. Rattay, K., Meyer, H. V., Herrmann, C., Brors, B. & Kyewski, B. Evolutionary conserved gene co-expression drives generation of self-antigen diversity in medullary thymic epithelial cells. *J. Autoimmun.* (2015). doi:10.1016/j.jaut.2015.10.001

326. Handel, A. E., Ebers, G. C. & Ramagopalan, S. V. Epigenetics: molecular mechanisms and implications for disease. *Trends Mol. Med.* **16,** 7–16 (2010).

327. Simon, J. A. & Kingston, R. E. Mechanisms of polycomb gene silencing: knowns and unknowns. *Nat. Rev. Mol. Cell Biol.* **10,** 697–708 (2009).

328. Yu, M. *et al.* Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis. *Mol. Cell* **36,** 682–695 (2009).

329. Richie, E. R. *et al.* The Polycomb-group gene eed regulates thymocyte differentiation and suppresses the development of carcinogen-induced T-cell lymphomas. *Oncogene* **21,** 299–306 (2002).

330. Kaneda, M. *et al.* Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. *Nature* **429,** 900–903 (2004).

331. Dodge, J. E. *et al.* Inactivation of Dnmt3b in Mouse Embryonic Fibroblasts Results in DNA Hypomethylation, Chromosomal Instability, and Spontaneous Immortalization. *J. Biol. Chem.* **280,** 17986–17991 (2005).

332. Cortázar, D. *et al.* Embryonic lethal phenotype reveals a function of TDG in maintaining epigenetic stability. *Nature* **470,** 419–423 (2011).

333. Ferguson-Smith, A. C. Genomic imprinting: the emergence of an epigenetic paradigm. *Nat. Rev. Genet.* **12,** 565–575 (2011).

334. Fritz, M. H.-Y., Leinonen, R., Cochrane, G. & Birney, E. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.* **21,** 734–740 (2011).

335. Cochrane, G., Cook, C. E. & Birney, E. The future of DNA sequence archiving. *GigaScience* **1,** 2 (2012).

336. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33,** 155–160 (2015).

337. Amir, E. D. *et al.* viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* **31,** 545–552 (2013).

338. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11,** 740–742 (2014).

339. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. & Hornik, K. cluster: Cluster Analysis Basics and Extensions. *R Package Version 203* (2015).

340. Adli, M. & Bernstein, B. E. Whole-genome chromatin profiling from limited numbers of cells using nano-ChIP-seq. *Nat. Protoc.* **6,** 1656–1668 (2011).

341. Lahens, N. F. *et al.* IVT-seq reveals extreme bias in RNA sequencing. *Genome Biol.* **15,** R86 (2014).

342. Mondini, M. *et al.* The interferon-inducible HIN-200 gene family in apoptosis and inflammation: implication for autoimmunity. *Autoimmunity* **43,** 226–231 (2010).

343. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454,** 766–770 (2008).

344. Lindroth, A. M. *et al.* Antagonism between DNA and H3K27 methylation at the imprinted Rasgrf1 locus. *PLoS Genet.* **4,** e1000145 (2008).

345. Reddington, J. P. *et al.* Redistribution of H3K27me3 upon DNA hypomethylation results in de-repression of Polycomb target genes. *Genome Biol.* **14,** R25 (2013).

346. Jin, B. *et al.* DNA methyltransferase 3B (DNMT3B) mutations in ICF syndrome lead to altered epigenetic modifications and aberrant expression of genes regulating development, neurogenesis and immune function. *Hum. Mol. Genet.* **17,** 690–709 (2008).

347. Martins-Taylor, K., Schroeder, D. I., LaSalle, J. M., Lalande, M. & Xu, R.-H. Role of DNMT3B in the regulation of early neural and neural crest specifiers. *Epigenetics* **7,** 71–82 (2012).

348. Cortellino, S. *et al.* Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. *Cell* **146,** 67–79 (2011).

349. Sun, X. *et al.* Directed differentiation of human embryonic stem cells into thymic epithelial progenitor-like cells reconstitutes the thymic microenvironment in vivo. *Cell Stem Cell* **13,** 230–236 (2013).

350. Wendling, O., Dennefeld, C., Chambon, P. & Mark, M. Retinoid signaling is essential for patterning the endoderm of the third and fourth pharyngeal arches. *Dev. Camb. Engl.* **127,** 1553–1562 (2000).

351. Blackburn, C. C. & Manley, N. R. Developing a new paradigm for thymus organogenesis. *Nat. Rev. Immunol.* **4,** 278–289 (2004).

352. Li, P., Pashmforoush, M. & Sucov, H. M. Mesodermal retinoic acid signaling regulates endothelial cell coalescence in caudal pharyngeal arch artery vasculogenesis. *Dev. Biol.* **361,** 116–124 (2012).

353. Casey, B. J., Jones, R. M. & Hare, T. A. The adolescent brain. *Ann. N. Y. Acad. Sci.* **1124,** 111–126 (2008).

354. Bale, T. L. Epigenetic and transgenerational reprogramming of brain development. *Nat. Rev. Neurosci.* **16,** 332–344 (2015).

355. Li, M., Suzuki, K., Kim, N. Y., Liu, G.-H. & Belmonte, J. C. I. A Cut above the Rest: Targeted Genome Editing Technologies in Human Pluripotent Stem Cells. *J. Biol. Chem.* **289,** 4594–4599 (2014).

356. Gusmao, E. G., Allhoff, M., Zenke, M. & Costa, I. G. Analysis of computational footprinting methods for DNase sequencing experiments. *Nat. Methods* **advance online publication,** (2016).

357. Chen, L., Xiao, S. & Manley, N. R. Foxn1 is required to maintain the postnatal thymic microenvironment in a dosage-sensitive manner. *Blood* **113,** 567–574 (2009).

358. Brind'Amour, J. *et al.* An ultra-low-input native ChIP-seq protocol for genome-wide profiling of rare cell populations. *Nat. Commun.* **6,** 6033 (2015).

359. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161,** 1202–1214 (2015).

360. Rotem, A. *et al.* High-Throughput Single-Cell Labeling (Hi-SCL) for RNA-Seq Using Drop-Based Microfluidics. *PLoS ONE* **10,** e0116328 (2015).

361. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **16,** 133–145 (2015).

362. Achim, K. *et al.* High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat. Biotechnol.* **33,** 503–509 (2015).

363. Bintu, L. *et al.* Dynamics of epigenetic regulation at the single-cell level. *Science* **351,** 720–724 (2016).

364. Bendall, S. C. *et al.* Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. *Science* **332,** 687–696 (2011).

365. Hughes, A. J. *et al.* Single-cell western blotting. *Nat. Methods* **11,** 749–755 (2014).

366. Robert, C. & Watson, M. Errors in RNA-Seq quantification affect genes of relevance to human disease. *Genome Biol.* **16,** 177 (2015).

367. Patro, R., Mount, S. M. & Kingsford, C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.* **32,** 462–464 (2014).

368. Bray, N., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal RNA-Seq quantification. *ArXiv150502710 Cs Q-Bio* (2015).

369. Ilicic, T. *et al.* Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* **17,** 29 (2016).

370. Delmans, M. & Hemberg, M. Discrete distributional differential expression (D3E) - a tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinformatics* **17,** 110 (2016).

371. Kim, J. K., Kolodziejczyk, A. A., Illicic, T., Teichmann, S. A. & Marioni, J. C. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nat. Commun.* **6,** 8687 (2015).

372. ENCODE Project. ENCODE and modENCODE Guidelines For Experiments Generating ChIP, DNase, FAIRE, and DNA Methylation Genome Wide Location Data: Version 2.0, July 20, 2011. (2011).

373. Dura, B. *et al.* Profiling lymphocyte interactions at the single-cell level by microfluidic cell pairing. *Nat. Commun.* **6,** 5940 (2015).