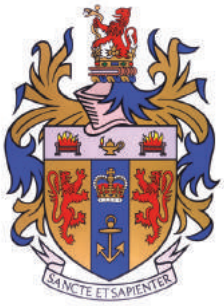
Single cell gene regulation in the epidermis



King's College London

THE FRANCIS CRICK INSTITUTE

Arsham Ghahramani MEng

Centre for Stem Cells and Regenerative Medicine, King's College London
The Francis Crick Institute

Co-supervisor: Prof Nicholas M Luscombe
Co-supervisor: Prof Fiona M Watt

Dissertation submitted towards the degree of Doctor of Philosophy

August 2018

Single cell gene regulation in the epidermis

Abstract

Mammalian epidermis comprises multiple cell subpopulations including cells residing in the interfollicular epidermis (IFE), hair follicles and sebaceous glands. Recent advances have led to gene expression profiling of single cells at lower cost, enabling unbiased identification of distinct subpopulations and cell states within these compartments.

To investigate the range of possible epidermal cell states, I studied the effect of a specific signalling perturbation, Wnt/beta-catenin signalling, on epidermal subpopulations. Although the bulk transcriptomic effect of Wnt signalling is well-studied, little is known about the effect of non-cell autonomous Wnt signalling at the single cell level. In particular, I wanted to address the mechanism by which keratinocytes in proximity to a Wnt-receiving cell are co-opted to undergo a change in cell fate.

To address this question, I performed single-cell RNA-sequencing on mouse keratinocytes co-cultured with and without beta-catenin-activated neighbouring cells. I identified five distinct cell states in cultures that had not been exposed to the beta-catenin stimulus and showed that the stimulus redistributes wild-type subpopulation proportions. Using temporal single-cell analysis, I reconstruct the cell fate

change induced by Wnt activation from neighbouring cells and identify transcription factors responsible for these changes.

As more single-cell data is produced there is an increasing need to integrate diverse datasets and better analyse underutilised data to gain biological insights. Analysis of single cell RNA-seq data is challenging due to biological and technical noise which not only varies between laboratories but also between batches.

To address these challenges, I applied a new generative deep learning approach called Generative Adversarial Networks (GAN) to biological data. I show that it is possible to integrate diverse skin (epidermal) datasets and in doing so, our generative model is able to simulate realistic scRNA-seq data that covers the full diversity of cell types. Using this generative model I am able to obtain a universal representation of epidermal differentiation and use this to predict the effect of cell state perturbations on gene expression at high time-resolution. This method is broadly applicable and can be used to analyse other single cell data types.

By integrating genomic, imaging and functional data I have uncovered new regulators of epidermal cell state and added to our understanding of non-cell autonomous Wnt signalling.

Preface

This dissertation describes research carried out jointly at The Francis Crick Institute, London, UK and King's College London, UK between between September 2014 and August 2018 under the supervision of Dr. Nicholas M. Luscombe and Dr. Fiona M. Watt.

Chapter 3 contains some previously published data and contents:

Arsham Ghahramani, Giacomo Donati, Nicholas M. Luscombe & and Fiona M. Watt (2018) Epidermal Wnt signalling regulates transcriptome heterogeneity and proliferative fate in neighbouring cells *Genome Biology* **201819**:3 https://doi.org/10.1186/s13059-017-1384-y

Chapter 4 contains some data available as a preprint and under review:

Arsham Ghahramani, Fiona M Watt & Nicholas M Luscombe (2018) Generative adversarial networks uncover epidermal regulators and predict single cell perturbations *bioRxiv* https://doi.org/10.1101/262501

Acknowledgments

Firstly I would like to thank Nick and Fiona for giving me the opportunity to jointly work between their labs. Your labs are filled with friendly and knowledgeable people. I will always be grateful for the skills I have developed and the lifelong friends I have made. Thank you to you both for being excellent mentors and inspirational scientists.

I would like to give special thanks to Giacomo Donati. You were a great mentor and guided me to shape my initial projects in the lab. When I joined the Watt lab I had little wet lab experience - you were not only patient with me but you also helped me become independent as soon as possible.

The work on Wnt signalling within this thesis was made possible by previous work by Cristina Lo Celso and Violeta Silva-Vargas in 2002-2006. I owe them my thanks for providing the building blocks to this research. I would also like to acknowledge the invaluable technical assistance provided by past and present members of the Crick Advanced Sequencing Facility, in particular Abdul Sesay and Mina

Bhaw.

I am grateful to everyone in the Watt and Luscombe labs who have made my time fun and to those who help to keep the labs running smoothly! Thank you to the remainder of the CSCRM, everyone has always been willing to listen, discuss and help at various moments throughout my PhD.

Oliver Culley, Ayelen Helling, Chloë Hurling, Blaise Louis and Victor Negri. You will never be just colleagues to me. We have laughed and we have cried from laughter. I could write a thesis on each of your brilliant qualities. In all seriousness, I didn't start my PhD expecting to find such incredible friends. You are each wonderful people and I am lucky to be surrounded by smart, positive and most importantly fun friends.

I'd like to thank Eve, Tristan, Elicka and all my friends in London. You each inspire and motivate me. Without your love, friendship and family I would not be where I am now.

Finally, thank you to my mother, father and Nick (Brett) for always being there for me. One of the greatest privileges in my life is that I have had support for each of my endeavours and I recognise that not everyone is so lucky.

Table of contents

Listing of figures

Abbreviations

APC	Adenomatous polyposis coli
BSA	Bovine serum albumin
CE	Core exon
CSV	Comma separated values
ChIP	Chromatin immunoprecipitation
DMEM	Dulbecco's modified Eagle medium
DMSO	Dimethyl sulfoxide
ECM	Extracellular matrix
EDTA	Ethylenediaminetetraacetic acid
ERCC	External RNA Controls Consortium
ESC	Embryonic stem cell
EdU	5-ethynyl-2'-deoxyuridine
FACS	Fluorescence-activated cell sorting
FBS	Fetal bovine serum
FCN	Fully connected network

FGF	Fibroblast growth factor
GAN	Generative adversarial network
GO	Gene Ontology
IFC	Integrated fluid circuit
IFE	Interfollicular epidermis
IVT	*in vitro* transcription
LOESS	Locally weighted scatterplot smoothing
LReLU	Leaky rectified linear unit
LSI	Latent space interpolation
NCA	Non-cell autonomous
NGS	Next generation sequencing
4OHT	4-Hydroxytamoxifen
PBS	Phosphate buffered saline
PCA	Principal component analysis
ReLU	Rectified linear unit
RT-qPCR	Real time quantitative polymerase chain reaction
scRNA-seq	Single cell RNA-sequencing
SG	Sebaceous gland
TCF	T-cell factor
TCOV	Transcriptome coefficient of variation
TF	Transcription factor
UMI	Unique molecular identifier

WGAN	Wasserstein generative adversarial network
WRE	Wnt responsive element
Wnt	Wingless-related integration site
WT	Wild type

Chapter 1

Introduction

1.1 Mammalian epidermis

The skin is the interface between the body and the external environment. Among its many functions is to act as an impermeable barrier throughout embryonic development and adult life. The outermost part of the skin is the interfollicular epidermis (IFE), a stratified multilayer epithelium with associated adnexal structures including hair follicles (HF) and sebaceous glands (SG). Keratinocytes in the IFE and HF are the most abundant cell type within the epidermis followed by less abundant cell types such as secretory sebocytes, melanocytes responsible for pigmentation, immune cells such as Langerhans cells and Merkel cells which facilitate transduction of mechanical force for sensation of light forces.

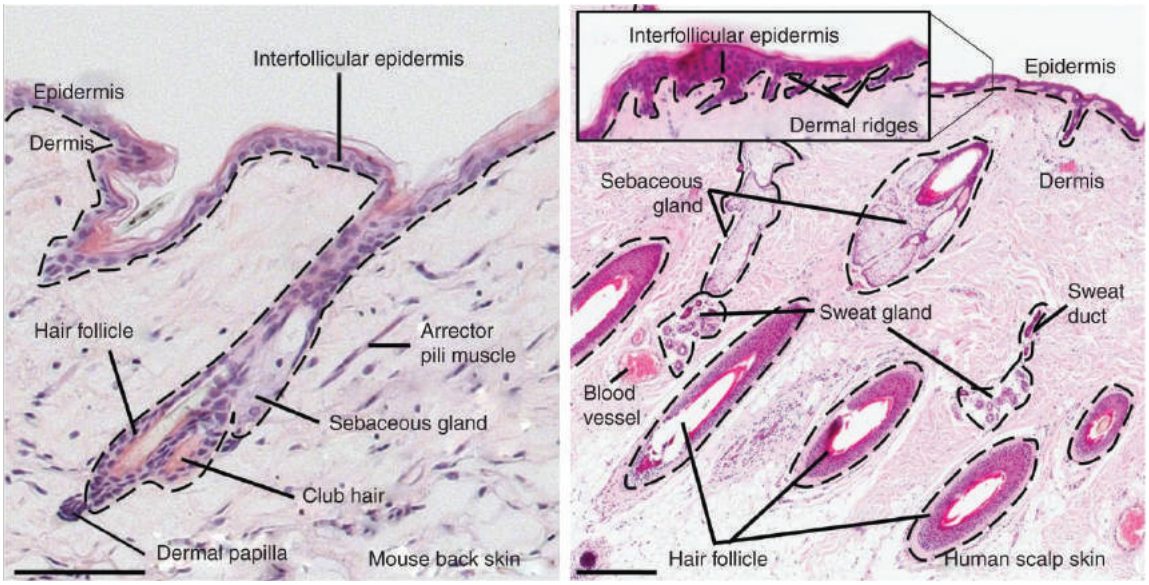


Figure 1.1: Comparative histology of mouse and human skin.
Histology of adult mouse (left) and human (right) skin stained with hematoxylin. Dashed line indicates the basement membrane boundary between dermis and epidermis. Note that rete (dermal) ridges and sweat glands are only present in the human skin. Scale bars, 100μm (mouse), 500μm (human). Figure adapted from Kretzschmar and Watt (2014).

1.1.1 Interfollicular epidermis

In humans the interfollicular epidermis is often characterised as comprising four histologically distinct layers of keratinocytes termed the basal layer, spinous layer, granular layer, and the cornified layer (McGrath and Uitto, 2010). The basal layer of the epidermis is composed of keratinocytes anchored to the basement membrane, a specialised layer of extracellular matrix (ECM) rich in collagens, laminins and proteoglycans. These layers are labelled on Figure 1.2 (right). Most IFE proliferation occurs in the basal layer and progeny of basal cells migrate upwards towards the surface as they differentiate. Cells in the intermediate spinous and granular layers are preparing for terminal differentiation and contain precursor proteins necessary for formation of the cornified envelope. The outermost layer of cells is a cornified layer of anuclear ker-

atinocytes. These terminally differentiated keratinocytes are filled with aggregrated keratin filaments, cross-linked proteins and lipids to aid the barrier function of the epidermis. As cells in the outermost layer become progressively more keratinised they are sloughed off and lost from the skin. Cells are continuously lost from the epidermis in this way, requiring that the epidermis possess a robust programme of homeostasis, balancing cell proliferation and differentiation to maintain the epidermis in a steady state.

Accompanying histological differences between IFE layers there are important molecular differences which distinguish these strata. Cells in the basal layer express high levels of integrins, primarily beta-1 (Itgb1) (Jones and Watt, 1993) and two keratins, keratin 5 and keratin 14 (Fuchs and Green, 1980). As keratinocytes commit to differentiation, expression of these basal marker genes is downregulated and markers of commitment to differentiation are upregulated. Two such genes are keratin 1 and keratin 10 which are exclusively expressed suprabasally (Schweizer et al., 1984; Joost et al., 2016). Furthermore, cells undergoing terminal differentiation express a distinct set of genes to enact the structural changes required for cornified envelope barrier function. Three molecules characteristic of terminally differentiating cells are envoplakin (Evl), periplakin (Ppl) and involucrin (Ivl) (Rice and Green, 1979; Ruhrberg et al., 1997). Cross-linking of these proteins with attached lipids to assemble the cornified envelope is triggered by a rise in intracellular Ca^{2+}, which activates the transglutaminase 1 (Tgm1) enzyme. In turn, transglutaminase catalyses formation of N□-(γ-glutamyl)lysine cross-links and the attachment of long chain ω-hydroxyceramides, a

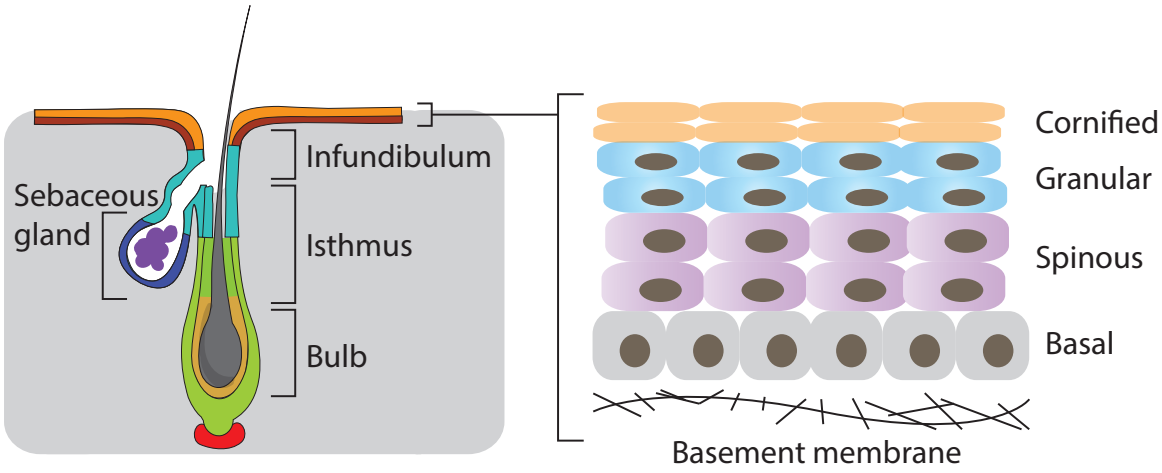


Figure 1.2: Schematic of a hair follicle and layers of the interfollicular epidermis.
Hair follicles (left) can be divided into three sections, the infundibulum, isthmus and bulb. Histologically, the interfollicular epidermis (right) is composed of four distinct layers, the basal, spinous, granular and cornified layers.

subset of lipids (Marekov and Steinert, 1998).

Mouse IFE is organised in a similar structure to human IFE, however, it is typically thinner and consists of 2-4 layers of supbrasal keratinocytes in contrast to 6-10 cell layers in humans. Another important difference is the topology of the basement membrane. In humans the dermal-epidermal junction undulates, termed rete ridges, whereas in mice this junction is flat.

1.1.2 Hair follicle

Hair follicles are the primary appendage connected to the interfollicular epidermis. They are multilayered cylindrical structures which extend into the dermis (see Figure 1.1). Hair follicles can be subdivided into three sections, labelled on Figure 1.2: (1) infundibulum, the upper portion of the hair follicle connecting the sebaceous gland

and interfollicular epidermis, (2) the isthmus is the middlle portion of the hair follicle, extending between the attachment of the arrector pilli muscle to the sebaceous gland duct. (3) The lowest portion of the hair follicle includes the hair bulb, hair matrix and dermal papilla.

Unlike the interfollicular epidermis, hair follicles have continuous cycles of growth (anagen), regression (catagen) and rest (telogen). In anagen, slow-cycling cells residing in the bulge (label retaining cells) produce rapidly proliferating matrix cells (Cotsarelis et al., 1990). The progeny of these cells differentiate into seven different lineages comprising the hair shaft and inner root sheath. On the start of catagen, the dermal papillae regresses through apoptosis and the hair follicle moves upwards. This phase is followed by hair follicle dormancy (telogen) where matrix cells cease to proliferate.

Hair follicles are a complex mini-organ and contain at least 11 different subpopulations (Yang et al., 2017). In mice, $CD34^{+}$, Keratin 15^{+}, label retaining cells residing in the bulge are considered as one of the main HF stem cell subpopulations. *In vitro* these cells are highly proliferative and are able to reconstitute both the interfollicular epidermis and hair follicle *in vivo* (Tumbar et al., 2004). Similarly, long term lineage tracing studies have shown that cells expressing Leu-rich repeat-containing G-protein-coupled receptor 5 (Lgr5) are able to differentiate into all lineages of the psilosebaceous unit (Jaks et al., 2008). Importantly, these cells have the capability to contribute to IFE maintenance, e.g. during wound injury response, however, they appear not to contribute to the IFE under homeostatic conditions (Ito et al., 2005).

1.1.3 Sebaceous gland

Sebaceous glands (SG) are exocrine glands usually attached to hair follicles. One of their primary roles is to secrete sebum, an oily substance, into hair follicles in order to lubricate hair and skin. SG have an acinar (multi-lobed) structure connected to the hair follicle via a secretory duct comprising stratified epithelial cells (Knutson, 1974). Cells on the outer periphery are undifferentiated whereas cells located towards the center contain large numbers of lipid droplets. Differentiated cells proximal to the duct break apart and release lipid-rich sebum with the help of lysosomal enzymes (Niemann, 2009).

The number of sebaceous glands remains approximately constant throughout life, however, their size and activity varies with age (Zouboulis, 2004). In humans, a strong increase in sebum production occurs immediately after birth and continues to increase for the first week before reducing thereafter. During this time sebaceous glands are enlarged; their size regresses after birth until adolescence when there is again an increase in gland size and activity.

1.1.4 Epidermal cell state and heterogeneity

Heterogeneity in keratinocytes has been studied for over thirty years; earlier studies focused on categorising differences in morphology and structural features (Lavker and Sun, 1982). Barandon and Green pushed forward understanding of keratinocyte

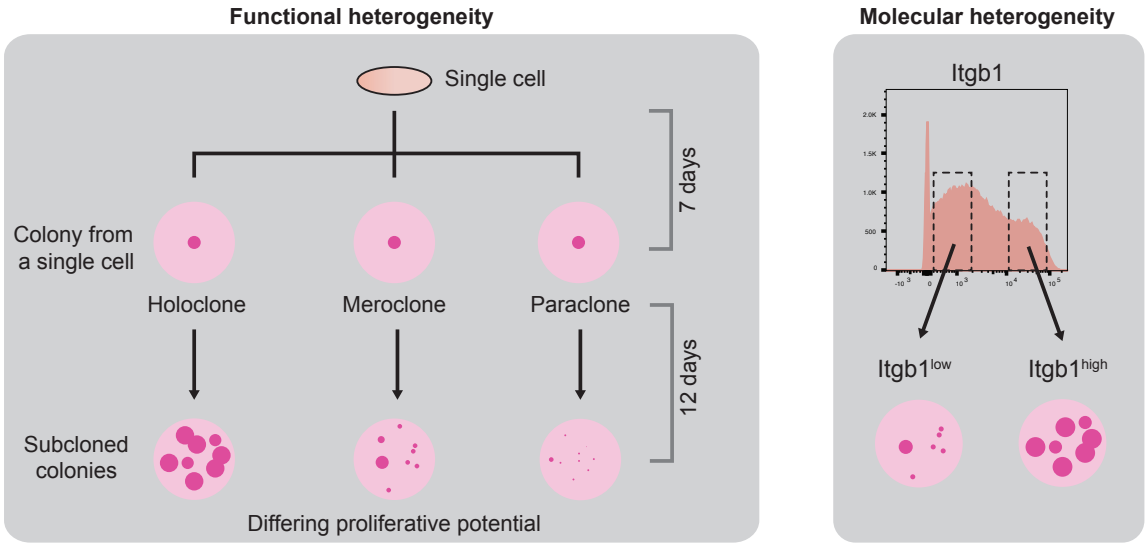


Figure 1.3: Epidermal cell heterogeneity.
Left: experiments by Barandon and Green demonstrated functional heterogeneity in cultured epidermal cells. Their pioneering experiments demonstrated that subclones from single cell colonies varied in their subsequent proliferative potential. They termed these cell subpopulations holoclones, meroclones and paraclones. Right: Jones and Watt built on this work and demonstrated a link between molecular heterogeneity, in this case heterogeneity in integrin beta-1 (Itgb1) cell surface abundance and proliferative potential. The Itgb1 FACS histogram was provided by Dr Christina Philippeos.

heterogeneity by demonstrating that not all keratinocytes were equal in their ability to proliferate *in vitro* (Barrandon and Green, 1987). Their study, depicted on the left side of Figure 1.3, interrogated the colony-forming ability of single human keratinocytes when cultured on a mouse fibroblast 3T3 feeder layer. Keratinocytes were cultured for seven days followed by a further subculture of 12 days. At the conclusion of the experiment they noted three categories of colonies termed holoclones, meroclones and paraclones with decreasing proliferative ability. Hence, their experiments established that keratinocytes are heterogeneous not only in appearance but also function.

One of the earliest studies to establish a link between molecular and functional heterogeneity examined integrin beta-1 (Jones and Watt, 1993; Jones et al., 1995). Cells

with higher levels of integrin beta-1 exhibit higher colony forming efficiency *in vitro* in comparison to cells with low expression of integrin beta-1. Furthermore, it was shown that integrin beta-1 expression is not uniformly distributed across the epidermis. The discovery of integrin beta-1 as a stem cell marker led to a search for further markers of epidermal subpopulations and lineages. Underlying this search is the hypothesis that the observed molecular heterogeneity corresponds to distinct cellular states of gene regulation. Hence, understanding the extent of keratinocyte heterogeneity will in turn lead to understanding the extent of gene regulatory states.

Technologies such as cell sorting, mass cytometry, mass spectrometry and single cell RNA-sequencing have allowed investigation of molecular heterogeneity in an unbiased manner. In particular, Joost and colleagues have performed a survey of messenger RNA expression in all murine epidermal cell types (Joost et al., 2016). This form of *a priori* investigation revealed a common differentiation and spatial gene expression signature for hair follicle and interfollicular epidermis cells.

1.2 Wnt/β-catenin signalling in the epidermis

Throughout the history of skin investigation a handful of signalling pathways have been recurrently identified as master regulators of multiple (often opposing) processes. Notch, Hippo, TGF-β/BMP, and Wnt/β-catenin signalling have all been implicated in regulation of self-renewal, lineage selection, differentiation and development of the epidermis (Watt and Jensen, 2009). Wnt signalling is particularly compelling to study

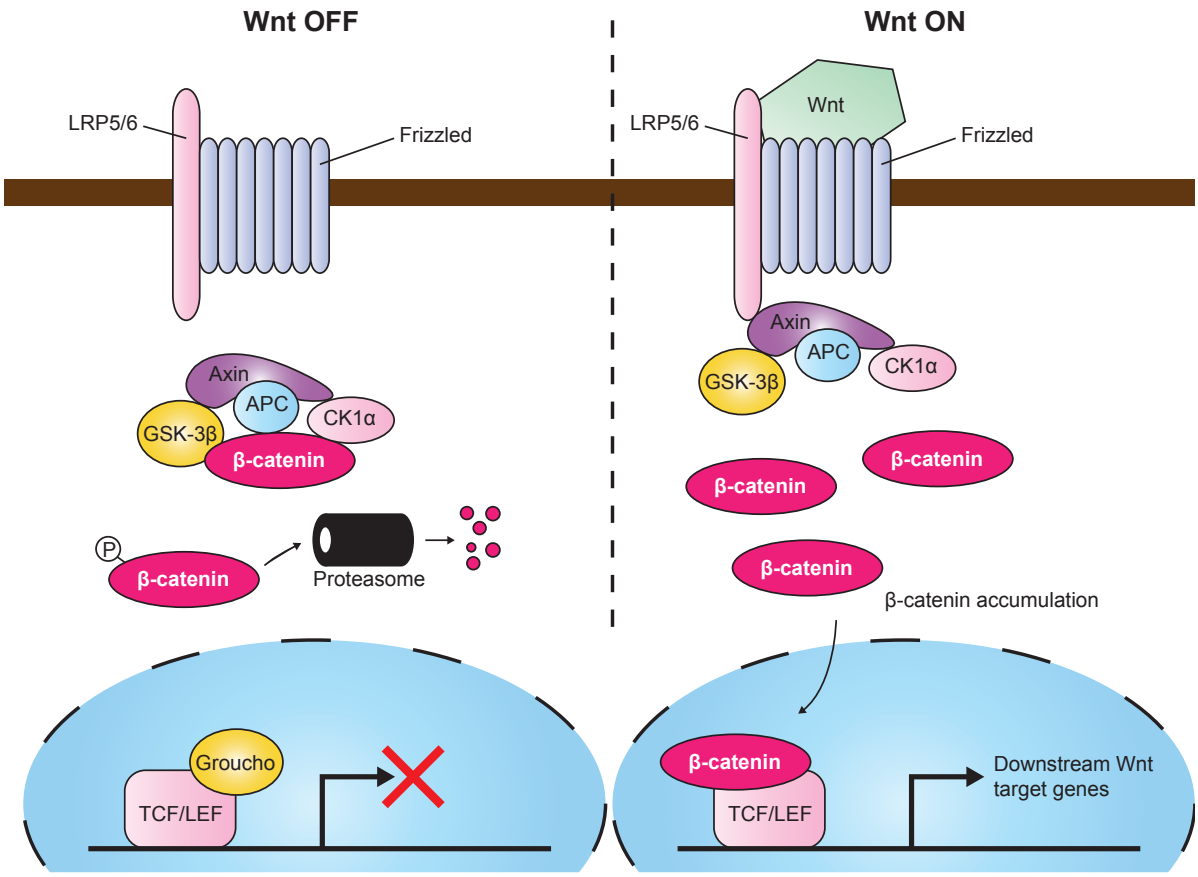


Figure 1.4: Wnt/β-catenin signalling pathway.
In the "off" state and in the absence of Wnt ligands β-catenin is bound by the destruction complex comprised of Axin, GSK3-β and CK1α. This leads to phosphorylation of β-catenin, ubiquitination by β-TrCP and proteasomal degradation of β-catenin. Hence, Wnt response DNA elements remain repressed by transcriptional corepressors such as Groucho. Wnt signalling is activated by the binding of Wnt ligands to the Frizzled receptor, causing recruitment of the destruction complex to LRP5/6. This inhibits the ubiquitination and degradation of β-catenin resulting in an accumulation in the cytoplasm and eventually increased β-catenin abundance in the nucleus. In the nucleus β-catenin acts as a transcriptional coactivator alongside TCF/LEF transcription factors, leading to the transcription of downstream Wnt target genes.

as many components of the Wnt signalling cascade are also present in other signalling pathways. Hence, observations and perturbations to Wnt signalling aids our understanding of how molecules in these interrelated signalling cascades coordinate cell and tissue state.

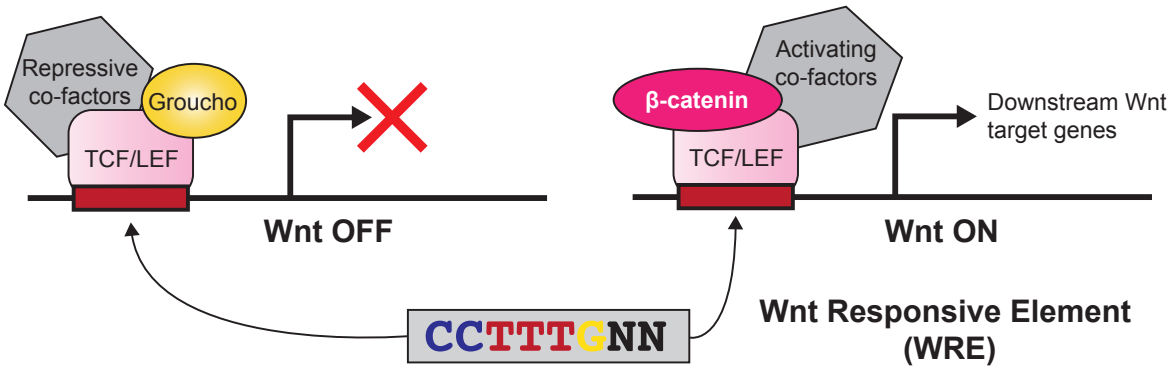


Figure 1.5: Wnt responsive elements activated by Wnt signalling.
Wnt responsive elements (WRE) with the motif CCTTTGNN are bound by TCF/LEF transcription factors. When Wnt signalling is off, these transcription factors are bound by Groucho and additional co-factors leading to transcriptional silencing. When Wnt signalling is on, there is an increase in nuclear beta-catenin which binds the WRE as a co-factor with TCF/LEF. Together this complex recruits other co-factors to activate transcription of Wnt target genes.

1.2.1 Wnt pathway

Wnt proteins are conserved in diverse organisms throughout the animal kingdom. There are 12 families of conserved Wnt proteins and the majority of mammalian genomes encode 19 Wnt genes. Few single cell organisms harbour Wnt proteins, however, Wnt homologs have been identified in sea sponges and sea anemones indicating that Wnt signalling may play a role in the evolution of multicellular gene regulation (Kusserow et al., 2005; Petersen and Reddien, 2009).

A key feature of these conserved signalling cascades is the Wnt effector protein β-catenin. In an unactivated state, β-catenin is predominantly cytoplasmic and continually degraded by the Axin complex consisting of *adenomatous polyposis coli* (APC), casein kinase 1 (CK1), and glycogen synthase kinase 3 (GSK3) (see Figure 1.4 for an overview of the Wnt pathway cascade). CK1 and GSK3 phosphorylation of β-catenin triggers β-catenin destabilisation (Peifer et al., 1994) by enabling recognition by βTrCP,

an E3 ubiquitin ligase subunit. Subsequently, β-catenin is ubiquitinated and degraded by the proteosome, preventing β-catenin export to the nucleus. Wnt/β-catenin signalling is activated when a Wnt ligand binds to the Frizzled receptor (Frz) leading to the recruitment of a complex of proteins comprising Dishevelled (Dvl), low density lipoprotein receptor-related protein 5/6 (LRP5/6). Dvl is next responsible for phosphorylation of the Wnt-Frx-LRP5/6 complex which triggers recruitment of the Axin complex. These events lead to inhibition of the β-catenin destruction complex resulting in elevated levels of nuclear β-catenin and accumulation of β-catenin in the nucleus (MacDonald et al., 2009). Within the nucleus β-catenin binds with an array of cofactors and transcription factors, most notably the TCF/LEF family (Arce et al., 2006), alongside other less well studied cofactors such as nuclear E-cadherin (CDH1) (Ferber et al., 2008).

1.2.2 β-catenin and TCF/LEF regulation of gene expression

Without β-catenin, TCF acts as a repressor of gene expression by interacting with Groucho (transducin-like enhancer of split 1, TLE1 in human). This complex binds to Wnt responsive elements (WRE) with a DNA binding consensus sequence of **CCTTTGNN** as shown in Figure 1.5. In mammals there are four TCF proteins TCF-1, TCF-3, TCF-4, and LEF-1 (in humans these genes are named TCF7, TCF7L1, TCF7L2 and LEF1 respectively) (Cadigan and Waterman, 2012). Upon Wnt signalling activation, TCF proteins complex preferentially with β-catenin, displacing Groucho and recruiting transcriptional coactivators. There is evidence that less common splice variants of

TCF-1 and TCF-4 have substantially different DNA binding characteristics to the predominant isoforms of the TCF family proteins (Atcha et al., 2007). Together, the range of binding sites and cofactors diversify the modes of gene regulation controlled by TCF/β-catenin proteins.

1.2.3 Non-canonical Wnt signalling

There are also pathways and downstream effectors of the Wnt pathway independent of β-catenin. All such pathways are termed the "non-canonical Wnt pathway" (reviewed in (Gó Mez-Orte et al., 2013)). Several non-canonical pathways exist and this remains an area of active research. The non-canonical pathways can be categorised into calcium-dependent (Ca^{2+}) and planar cell polarity (PCP) pathways. In Wnt/Ca^{2+} signalling, binding of the Wnt ligand to its receptor activates a release of intracellular calcium which further activates downstream calcium-dependent kinases. Alternatively, the Wnt/PCP pathway involves activation of of the Rho family of GTPases and downstream c-Jun N-terminal kinases (JNK) which amongst other processes regulate the arrangement of the cytoskeleton. Investigations in Chapters 2 and 3 of this study focus on the canonical Wnt pathway dependent on β-catenin, however, it is important to consider non-canonical pathways when interpreting the effects of Wnt activation through Wnt ligand rather than directly through β-catenin.

1.2.4 Role of Wnt/β-catenin signalling in the epidermis

One of the earliest studies to specifically dissect the role of β-catenin signalling in the epidermis was performed by Birchmeier and colleagues (Huelsken et al., 2001). In a pioneering study they showed that epidermal specific deletion of β-catenin during embryogenesis prevents the formation of hair follicle placodes. Additionally their study showed that in the absence of β-catenin, keratinocytes adopted an IFE fate and failed to differentiate into follicular lineages. Together these data implicated Wnt signalling in hair follicle placode patterning and epidermal lineage selection. Many later mouse studies of Wnt signalling would adopt a similar approach of transgenic abrogation or activation of the Wnt/β-catenin pathway.

To investigate the effect of epidermal Wnt activation, Watt and colleagues developed and applied a transgenic mouse model allowing transient and constitutive β-catenin activation targeted to the epidermis (K14ΔNβ-cateninER, see Chapter 2) (Lo Celso et al., 2004; Silva-Vargas et al., 2005). Constitutive activation was shown to lead to the formation of follicular tumours when β-catenin was specifically induced in the IFE. In contrast, short term activation at endogenous levels formed *de novo* hair follicles, suggesting that IFE lineage cells were redirected towards a hair follicle cell state.

On the basis of these results, later studies addressed the sensitivity of different epidermal compartments to Wnt activation. K15ΔNβ-cateninER and K5ΔNβ-cateninER transgenic mice were used to selectively activate HF bulge and SG compartments respectively (Baker et al., 2010). Intriguingly, only β-catenin activation of the SG but not

the HF bulge lead to ectopic hair follicle formation, however, the hair follicle bulge showed proliferation and expansion suggesting limited sensitivity to the β-catenin perturbation.

1.2.5 Wnt/β-catenin in epidermal diseases

Constitutive Wnt activation is a hallmark of a selection of cancers, most notably a subset of colorectal cancers. Loss of APC in these cancers leads to accumulation of β-catenin and transcription of downstream target genes (Clements et al., 2003). In a similar manner, constitutive activation of β-catenin signalling leads to developmental defects and diseases. Human pilomatricomas, which are benign tumours containing hair shaft and matrix cells, and trichofolliculomas which contain multilineage cells have both been shown to be caused by stabilising β-catenin mutations (Chan et al., 1999).

Perturbation of Wnt signalling resulting in lower β-catenin activity can also result in adverse effects. An examination of K14ΔNLef1 mice, a model for disruption of Lef1-β-catenin binding, demonstrated that reduction of Wnt signalling results in formation of sebaceous gland tumours (Niemann et al., 2002). The relevancy of this mouse model was later confirmed through sequencing of human sebaceous tumours, where approximately a third of tumours harbour mutations within the N-terminus of Lef1, preventing efficient binding with β-catenin (Takeda et al., 2006).

There is also evidence that some squamous cell carcinomas (SCC) are β-catenin

dependent. Using a mouse model of SCC formation, Huelsken and colleagues have shown that $CD34^+$ tumour-propagating cells display high levels of nuclear β-catenin (Malanchi et al., 2008). They further demonstrated functional importance to the β-$catenin^+$ squamous cell carcinoma cells by showing that loss of β-catenin led to tumour regression. Similarly, evidence from human basal cell carcinoma samples supports a role for β-catenin-mediated transformation of cell state (Salto-Tellez et al., 2006).

An extensive review of the role of Wnt signalling in the skin is provided by (Lim and Nusse, 2013).

1.3 Single cell RNA-sequencing

As discussed in section 1.1.4 epidermal heterogeneity has been a topic of investigation since the earliest studies which sought to categorise differences in cell morphological features (Lavker and Sun, 1982). In the wider field, a number of tools have proven indispensable for assaying single-cell heterogeneity at the RNA and protein level. Fluorescence activated cell sorting (FACS) allows the quantification of up to 18 proteins simultaneously (Chattopadhyay et al., 2006), although most commonly less than five proteins are quantified. Many FACS protocols have a minimal effect on cell viability and are therefore compatible with downstream quantification of RNA levels. Mass cytometry extends this to over 30 simultaneous protein channels (Bendall et al., 2011), although the process of protein quantification is destructive and therefore does not allow downstream measurements or experiments.

Single cell RNA quantification methods were first introduced by Coleman and colleagues who analysed expression of single live neurons using single cell qPCR (Eberwine et al., 1992). This is complemented by single molecule fluorescence *in situ* hybridisation (smFISH) which allows quantification of gene expression while retaining spatial information (Femino et al., 1998).

FACS, mass cytometry, single cell qPCR and smFISH all require a prior expectation of expressed and functionally relevant genes to choose as the subset of genes to assay. Multiplex methods of these techniques limit the maximum number of assayed genes to less than 50 genes or proteins whereas there are over 20,000 protein coding genes in the mouse or human genomes. The advent of RNA-sequencing allowed unbiased quantification of all expressed transcripts (Mortazavi et al., 2008). Recent developments in single cell capture methods and RNA-sequencing protocols have enabled the application of full transcriptome sequencing to single cells.

(Kolodziejczyk et al., 2015a) provides a thorough overview of current single-cell RNA sequencing technologies.

1.3.1 Single cell gene expression studies of skin

Gene expression studies of skin and specifically the epidermis have long recognised the need for single cell resolution. One of the earliest studies quantifying single cell RNA abundance was performed by Jensen and Watt in 2006. At the time, an increasing number of public microarray datasets from FACS-selected populations were avail-

able. Their study, enabled by advances in cDNA amplification methods, recognised the need to interrogate these subpopulations further. Subsequently, Tan et al. (2013) combined further advances in cDNA amplification coupled with increased sensitivity of microarrays to interrogate the full human epidermal transcriptome for 62 cells.

More recently, Joost et al. (2016) have taken advantage of advances in microfluidics based single cell methods (discussed later in this chapter) to survey all mouse epidermal cell states. Alongside this, the Kasper lab has released an online tool facilitating the study of cell states and gene expression markers of cell state. There are currently no publicly available human epidermal scRNA-seq datasets, although dermal data is available (Philippeos et al., 2018). However, there are plans to produce new epidermal scRNA-seq data as part of the Human Cell Atlas, an initiative to produce reference maps of all human cell types and subpopulations.

1.3.2 Single cell capture methods

The earliest cell capture methods involved manual isolation of single cells using a glass capillary and processing of the cell as an individual RNA-sequencing library (Tang et al., 2009). This low-throughput manual method had the advantage of being able to select cells from precise positions within a tissue and was initially used to study early embryonic development. In applications where targeted cell isolation from regions of tissues was desired, laser capture microdissection (LCM) proved to be a parallel low-throughput method (Keays et al., 2005).

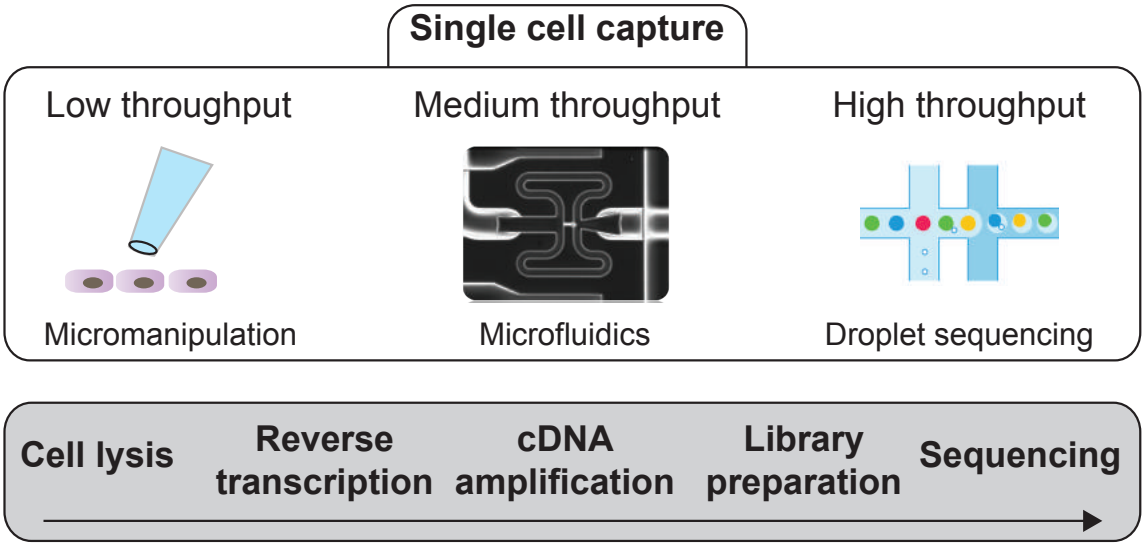


Figure 1.6: Single cell RNA-sequencing protocol overview.
Single cell capture methods vary in their cell throughput capacity. Methods such as capture by glass capillary allow control over location but it is difficult to process more than a few dozen cells. Medium throughput methods such as the Fluidigm C1 microfluidics platform allows capturing and processing of hundreds to a few thousand cells. More recent high-throughput methods such as droplet-based protocols facilitate easy processing of 100,000s of cells. All scRNA-seq protocols comprise five steps, cell lysis to release RNA, reverse transcription of RNA to cDNA, amplification of cDNA, library preparation and sequencing.

The first high-throughput methods enabled the capture of hundreds of cells; either utilising FACS to sort cells into individual wells or microfluidic platforms to capture cells in chambers (Shalek et al., 2014a). Both of these methods require the dissociation of cells into a suspension of single cells maintained in a buffer for cell viability, most commonly enzymatic dissociation through the action of trypsin. The Fluidigm C1 is the most commonly used microfluidic platform, currently allowing the capture and mRNA-sequencing of 96 cells per microfluidic integrated fluidic circuit (IFC) or "chip". Fluidigm provides three IFCs which differ in cell capture site size (5-10, 10-17 and 17-25μ), requiring a homogenously sized cell population with a known expected range of sizes.

More recently, droplet based microfluidic approaches have increased cell capture

throughput by more than an order of magnitude, enabling analysis of thousands to tens of thousands of cells. Drop-Seq was one of the first protocols to demonstrate the encapsulation of cells within nanoliter-scale aqueous droplets alongside an oligonucleutide-barcoded bead. Each cell and bead droplet acts as a small PCR chamber and distinct barcodes for each droplet allow downstream multiplexing (Macosko et al., 2015). This technology was later refined and commercialised by 10x Genomics with their Chromium single cell expression profiling system (Zheng et al., 2016).

1.3.3 Single cell sequencing protocols

Single cell RNA-sequencing is susceptible to noise due to the low starting amounts of input RNA. Studies have estimated a limit of detection between five and ten mRNA molecules which corresponds to an approximate capture efficiency of 10% (Ramsköld et al., 2012; Islam et al., 2011). Typical protocols involve four main steps: (1) cell lysis, (2) reverse transcription of RNA, (3) cDNA amplification, (4) library preparation for sequencing.

(1) Cell lysis: Cells are lysed in buffer to disrupt cell membranes and preserve RNA integrity.

(2) Reverse transcription: Ribosomal RNAs form the majority of cellular RNA and would therefore dominate if total RNA was sequenced. In order to select for the less abundant mRNAs the majority of protocols utilise polyT priming of reverse transcription. SmartSeq2 and STRT-Seq, the two most commonly utilised protocols achieve

this through first strand synthesis by a reverse transcriptase using a barcoded polyT primer followed by template switching of the reverse transcriptase to a template switching oligonucleotide (TSO) at the 5′ end of the RNA (Islam et al., 2011; Picelli et al., 2014). This approach allows full-length reverse transcription of the RNA in a sequence-independent manner. In contrast methods such as the Tang or QuartzSeq protocols result in a strong 3′ read bias.

(3) cDNA amplification: Most methods amplify the low amount of resultant cDNA at this point using PCR (SmartSeq2, STRT-seq, Tang). As with many applications of PCR, exponential amplification of transcripts can introduce artefacts and results must be interpreted cautiously. Such artefacts can be mitigated by the use of unique molecular identifier (UMI) barcodes allowing downstream computational removal of PCR duplicates. An alternative method is *in vitro* transcription (IVT) which results in a strong 3′ end bias, albeit avoiding biases assoicated with exponential PCR amplification.

(4) Library preparation and sequencing: Amplified cDNA libraries can be using either single or paired-end conventional manufacturer protocols.

1.3.4 Analysis of single cell RNA-seq

Owing to biological variability, the small amount of starting RNA, and low sensitivity of single cell protocols, scRNA-seq data is highly challenging to analyse. Analysis of hundreds to thousands of cells provides new opportunities for the study of cell and tissue biology. Single cell data is uniquely high-dimensional and sparse when

compared with most other biological abundance measurements. Whilst some bulk expression analysis methods equally apply to single cells, in many cases there has been a need for new algorithms and approaches. This section provides an overview of the necessary steps to analyse scRNA-seq data as shown in Figure 1.7.

Processing, quality control and normalisation

Initial processing of single cell RNA-sequencing data follows closely to bulk RNA-seq methods. Read data (typically in FASTQ format) must be demultiplexed, UMIs must be deduplicated and sequencing adapters trimmed. Some approaches such as dropEst estimate the probability of UMI barcodes erroneously colliding, particularly useful for very large numbers of cells (Petukhov et al., 2017). In experiments where external spike-in RNAs have been used at a known concentration (e.g. ERCC controls), these can be used to adjust for differences in cDNA amplification.

It is necessary to remove low quality cells before further downstream analysis to avoid erroneous cell clustering and pseudotemporal ordering. FASTQC and scPipe are two useful tools for examining the quality of single cell transcriptomes (Tian et al., 2018). Using a combination of features such as percentage of reads mapping to mitochondrial DNA and total number of reads it is possible to exclude poor quality outlier cells.

Single cell libraries are next aligned to the genome of choice. STAR and Kallisto are two modern alignment tools for full alignment and pseudoalignment respectively

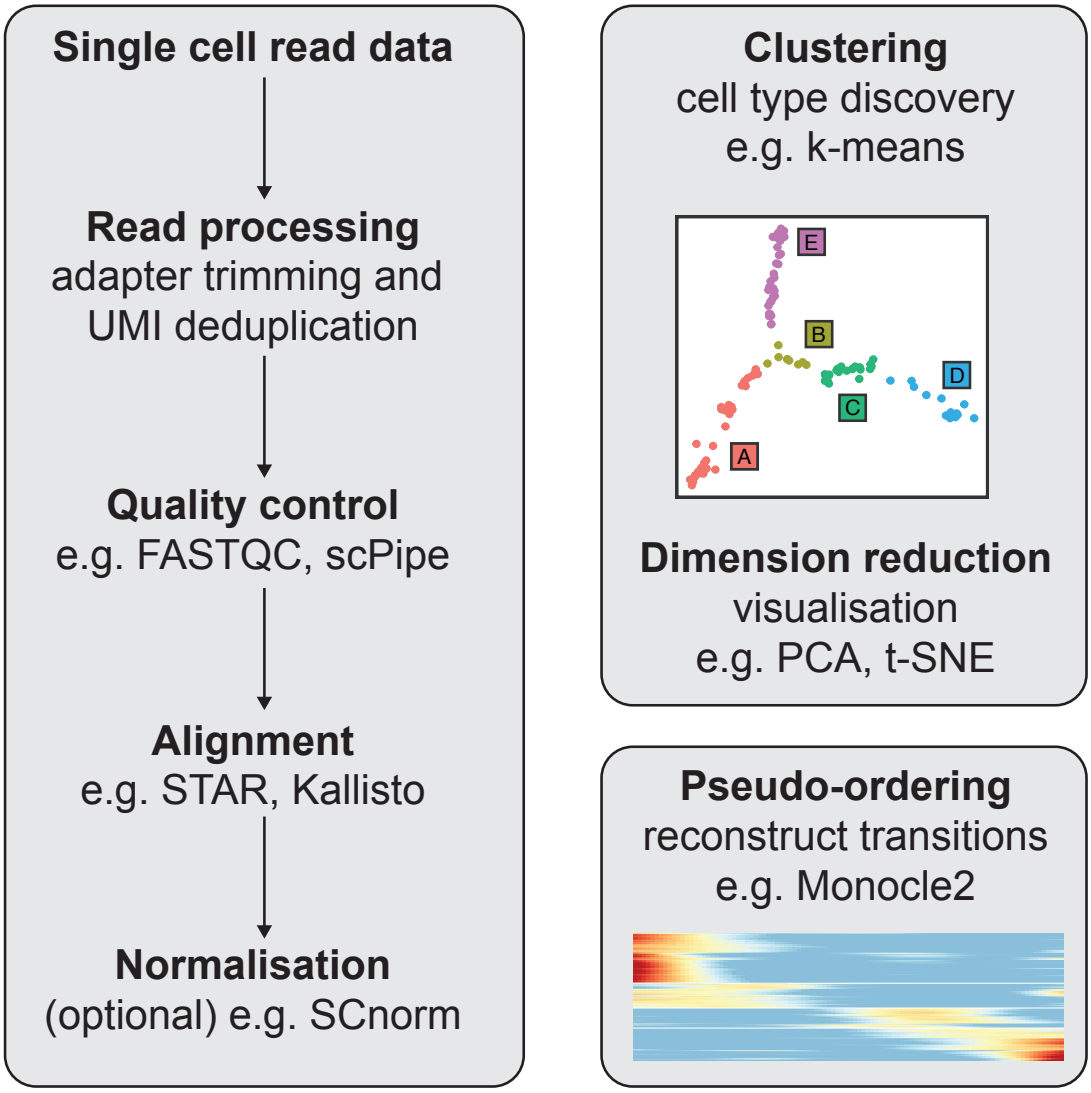


Figure 1.7: Analysis workflow for single cell RNA-seq data.
Single cell RNA-sequencing data must be carefully processed to ensure interpretable results. Raw read data is first processed by trimming adapter sequences and collapsing duplicate cDNA fragments if unique molecular identifier (UMI) barcodes were used. Next, read data is checked for a variety of quality metrics. Sequenced libraries which satisfy these metrics are then aligned to the genome of choice. Finally, aligned reads can be counted against genes or transcripts and optionally normalised to partially remove technical variation between sequencing libraries.

(Dobin et al., 2013; Bray et al., 2016). Once aligned, reads are counted against genes to obtain a raw count matrix consisting of unnormalised counts per gene and cell. There is no current consensus on a standard normalisation method for raw count data. Typically, cells are normalised for library size (e.g. transcripts per million, TPM).

When analysing data originating from multiple batches or even generated within different laboratories, it is desirable to remove unwanted batch-effects. Most existing bulk methods for batch-effect removal are ineffective when applied to single cell data, hence, development of new single cell specific methods for removal of batch effects is an active area of research. One example method is SCnorm, which applies regression to remove differences in sequencing depth, however, there is still debate as to the amount of biological variability mistakenly removed by these methods (Bacher et al., 2017; Camara, 2018). More recently, Haghverdi et al. (2018) have applied a mutual nearest neighbors matching approach for batch-effect-correction. This method requires that a subset of cells is shared between datasets and has the benefit of computation time scaling linearly with number of cells.

Clustering and pseudotemporal ordering

In many applications of scRNA-seq, cells are sampled from a heterogeneous mixture of subpopulations and it is desirable to cluster cells into distinct cell types or cell states. Once clustered, these assigned cell identities can be used to perform differential expression analysis between cell clusters to obtain marker genes and signature gene

expression profiles for each cluster.

It is now a *de facto* standard to apply t-distributed stochastic neighbour embedding (t-SNE) for the visualisation and dimensionality reduction of single cell transcriptomes (van der Maaten and Hinton, 2008a). This supersedes principal component analysis (PCA) which is more commonly used to visualise bulk gene expression data. PCA is a statistical method that performs a transformation to convert features (gene expression values) into a set of linearly uncorrelated principle components. More specifically, these principle components are obtained by calculating the eigenvectors of the covariance matrix for the input data. In contrast, t-SNE can only be used for visualisation and there is no interpretable distance metric or breakdown into meaningful components.

PCA and t-SNE have been widely applied across many fields. More recently, several methods have been introduced aimed specifically at single cell gene expression often combining dimensionality reduction, visualisation and clustering of cells to differing extents. In many applications of scRNA-seq, it is desirable to infer the number of distinct cell types and transcriptomic states present in the dataset without prior knowledge of the subpopulation composition - an unsupervised clustering problem. The majority of scRNA-seq unsupervised clustering methods are based on either hierarchical clustering (e.g. pcaReduce (žurauskienė and Yau, 2016) and SINCERA (Guo et al., 2015)) or k-means (e.g. SC3 (Kiselev et al., 2017) and RaceID (Grün et al., 2015)). Less commonly, graph based clustering algorithms can be applied; these have the disadvantage of requiring cells be mapped to a graph representation in order for graph

community detection algorithms to be applied (e.g. PhenoGraph (Levine et al., 2015) and early versions of Seurat (Butler et al., 2018)).

Many biological processes such as development and tissue differentiation exhibit smooth cell transitions. For these cases, it is useful to computationally reconstruct the transcriptional path taken by cells through possible cell states. Pseudo-ordering methods seek to reconstruct cell gene expression trajectories by estimating the temporal (pseudotemporal ordering) or spatial relationships between cells. Previously mentioned dimensionality reduction approaches or hierarchical clustering methods can be used to infer these relationships. Alternatively, Monocle2 (Qiu et al., 2017) utilises reverse graph embedding, i.e. transformation of gene expression vector space to a graph, followed by calculation of the minimum spanning tree through all cells. Pseudo-ordering methods have provided insights into the dynamics of single cell transcription, although they are particularly susceptible to technical noise as many methods assume that all cell-cell relationships are representative of a valid cell state transition (Dulken et al., 2017; Raj et al., 2018).

Analysis of cell-to-cell variability

Another area of investigation uniquely opened up by single cell RNA-seq is the characterisation of cell-to-cell transcriptional variability. Bulk differential gene expression methods often focus on determining whether there is a statistically significant difference in the mean gene expression between conditions or samples of interest. Single

cell gene expression information allows us to probe differences in the distribution of gene expression regardless of differences in mean expression level. SCDE developed by Kharchenko et al. (2014) and BASiCS developed by Vallejos et al. (2015) are two examples of a Bayesian approach to single cell differential distribution analysis. SCDE incorporates evidence provided by dropout events in individual cells with information on the average expression level for a gene within a subpopulation. This is a useful approach as although a dropout event does not exclude the possibility of expression for a gene, dropout events constrain the probability of expression at a high magnitude.

One example of the application of cell-to-cell variability analysis is for the analysis of transcriptional variability over ageing. Martinez-Jimenez et al. (2017) analysed differences in expression level and expression variability for immune cells in young and aged mice. Their analysis found that transcriptional variability increases with age and this contributes to a less efficient immune response. Relevant to our investigation, this study demonstrates that transcriptional variability is an important aspect of cell state and contributes to the function of tissues and systems.

1.4 Generative deep learning for analysis of genomic data

Deep learning refers to a collection of supervised or unsupervised machine learning methods which apply multiple stacked layers of nonlinear processing units to learn representations of data (Bengio et al., 2013). Deep learning has its origins in neural network research first carried out in 1943 (McCulloch and Pitts, 1943). Recent develop-

ments in neural network training algorithms and network architectures have resulted in state of the art performance in fields as diverse as computer vision (Esteva et al., 2017), speech recognition (Hinton et al., 2012) and genomics (Alipanahi et al., 2015).

1.4.1 Artificial neural networks

Artificial neural networks are inspired by the principles of neural information processing in the brain, where neurons integrate inputs through dendrites, process this input and activate under certain conditions by relaying an action potential along its (output) axon. The simplest artificial neural networks organise many such units in layers feeding forward from inputs to outputs with each neuron or unit in a layer taking inputs from all units in the previous layer. Figure 1.8a shows a fully connected neural network with two hidden layers. Each neuron calculates a weighted sum of its inputs, adds a "bias" value and then a non-linear activation function is applied (Figure 1.8a). Rectified linear unit or ReLU is the most popular applied non-linear activation function as it has been shown this function allows more reliable neural network training when compared to sigmoid or hyperbolic (tanh) activation functions (Krizhevsky et al.).

Neural network weights are parameters which are learned through a process of training with data. Most models are trained using an algorithm known as mini-batch stochastic gradient descent (MB-SGD) and back-propagation (Rumelhart et al., 1986). One typical use case of neural networks is to predict an output based on input data.

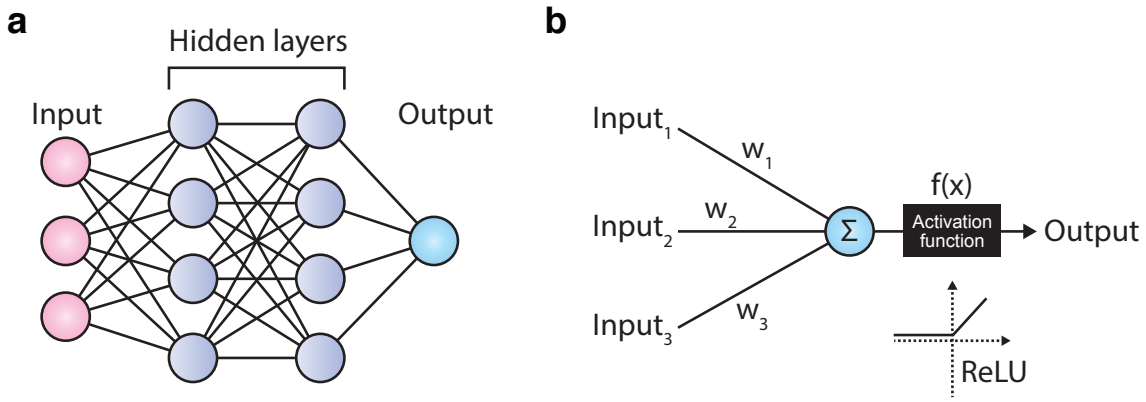


Figure 1.8: Artificial neural networks.
(a) A small fully connected neural network with two hidden layers. In a fully connected network, each neuron in a layer is connected to all neurons in the previous layer. Inputs and outputs feed forward to the final output layer. (b) Each neuron receives multiple inputs, a weighted sum of these is calculated followed by an activation function, typically ReLU, to calculate the corresponding output.

Training by gradient descent seeks to minimise the error between predicted and real output by calculating the update gradient for all neurons (gradients calculated by back-propagation) and updating neuron weights accordingly. Repeated rounds of gradient updating result in minimisation of neural network output error, also known as the loss function. Loss function minimisation can be visualised as finding the minimum value in the ”loss landscape”, corresponding to the optimal neural network weight parameters. One difficulty in training neural networks, as with all numerical optimisation techniques, is the intractable problem of determining whether a point is a local or global minimum.

1.4.2 Deep learning in genomics

Prodominant genomic analysis approaches follow the pattern of (a) processing raw genomic data (b) extraction of features associated with genomic sequences, cells or

tissues (c) dimensionality reduction and clustering of these features, (d) inference of biological function from dimensionally reduced representations. One potential advantage of machine learning approaches is the opportunity to combine (b)-(d) by training methods to learn dimensionally reduced representation and infer biological traits concurrently.

Two recent examples of this end-to-end data analysis paradigm are DeepCpG, a method for imputation of single cell DNA methylation, and DeepBind, a method for predicting DNA/RNA-protein binding specificity (Alipanahi et al., 2015; Angermueller et al., 2017a). Both methods allow input of nucleic acid sequence (alongside sparse methylation for DeepCpG) without the need for feature extraction. Dispensing the requirement for some data processing and feature building steps has two key advantages. Firstly, feature extraction relies on our prior understanding of the data and can result in unintentional disposal of information. There is a balance between capturing the minimum number of biological features sufficient for the prediction task and maintaining enough information content to achieve this. Incorporating feature extraction as part of the machine learning training process ensures a better balance. Secondly, neural networks are able to learn non-linear relationships between features and can perform this at multiple levels of feature abstraction. In the example of DeepCpG, single cell imputation is a difficult task for traditional computational methods as it is necessary to account for methylation at multiple genomic scales.

1.4.3 Generative adversarial networks

Generative adversarial networks (GANs) are an emerging method for unsupervised and semi-supervised machine learning first proposed by Ian Goodfellow (Goodfellow et al., 2014a). The simplest form of GANs are characterised by the concurrent training of two neural networks in competition (adversarial) with each other, resulting in the implicit modelling of a data distribution. These two networks are commonly referred to as the discriminator and the generator, due to their distinct roles (see schematic in Figure 1.9a). The generator (G) is tasked with transforming some input z (also known as latent space) into a realistic output $\hat{x}$ ($\hat{x} = G(z)$). GANs are often explained in the context of an art forgery analogy where the generator is tasked with producing a realistic forgery of an artwork. In contrast, the discriminator (D) is tasked with scoring the "authenticity" of an input to determine whether the input is real (x) or forged ($\hat{x}$), i.e. $D(x)$. Discriminator and generator networks are trained simultaneously, hence, as the generator network improves at producing a realistic output the discriminator network also improves at discriminating between authentic and generated samples. Each of the two GAN components can consist of any arbitrary network architecture, typically a combination of fully connected, convolution and recurrent layers.

A key powerful feature of GANs is the ability to learn a mapping from an underlying data distribution (z) to the real data distribution (x). This process is analogous to the reverse of dimensionality reduction techniques. For example, PCA constructs an invertible linear mapping from the real data distribution to the principle component

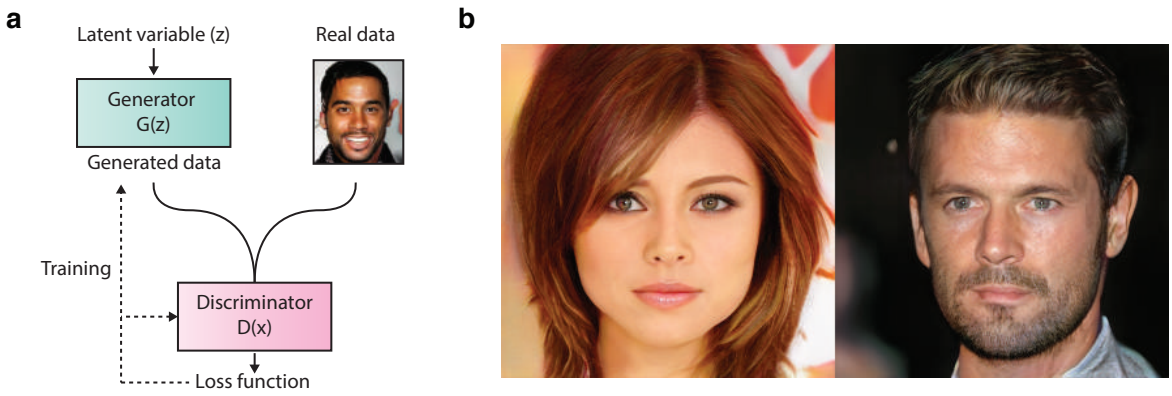


Figure 1.9: Generative adversarial networks.
(a) GANs consist of two competing neural networks. One network, termed the generator is tasked with transforming an arbitrary input z into a valid output, e.g. an image of a face. The second network, the discriminator, receives real and generated images as its input and must score these images on their authenticity. Both networks are trained sequentially using stochastic gradient descent resulting in incrementally improved performance for the generator and discriminator after each training batch. At the conclusion of training the generator network is capable of producing photorealistic images. (b) Examples of faces generated using ProGAN by (Karras et al., 2017a).

space where each component is orthogonal. Another example is the Fourier transform, widely used in signal processing, where data can be decomposed into a weighted combination of sinusoidal functions. Each of these examples require *a priori* knowledge regarding the statistical properties of the underlying data distribution. In contrast, GANs do not require any such assumptions in constructing the mapping learned by the generator network.

Training of GANs is equivalent to a minimax game in game theory, first formulated by the mathematician John von Neumann (v. Neumann, 1928). That is to say, an improved set of parameters for one of the networks leads to a worse loss for the other network. More formally GAN training attempts to solve: $\max_D \min_G V(D, G)$

where:

$$\max_{D} \min_{G} V(D, G) = \mathbb{E}_{x \sim p(x)_{\text{real data}}}[logD(x)] + \mathbb{E}_{z \sim p(z)_{\text{latent vector}}}[logD((G(z))] \tag{1.1}$$

i.e. the probability of the discriminator predicting the real data as authentic summed with the probability of the discriminator predicting that the generated data is not authentic.

There are many practical challenges to GAN training primarily derived from the instability of competitive optimisation for two neural networks. Optimisation of GAN training is an active area of research and it is expected that future advances will address current limitations (Salimans et al.). Common difficulties include **mode collapse**, where the generator network "collapses" to generate a small subset or even one output that scores highly with the discriminator network.

Applications of GANs

The GAN field is relatively nascent and new applications are developing. As is common with most machine learning methods, initial applications focus heavily on computer vision (Radford et al., 2015; Berthelot et al.; Antipov et al.). Current applications can be broadly categories into three categories **(1) classification and regression**, **(2) synthesis**, **(3) translation**.

(1) Classification and regression: GANs appear to process high-dimensional features

better than any existing unsupervised algorithm for many tasks (Reed et al., 2016b). Hence, in a parallel study Reed and colleagues reused layers from a trained GAN and applied these to a different (non-GAN) convolutional neural network (Reed et al., 2016a). Through this approach they demonstrated broad applicability of the unique characteristics of GANs.

(2) Synthesis: Once successfully trained, a GAN is by definition a trained model capable of synthesis of data from the output target distribution. A GAN variant named ProGAN (progressive growing of GANs) trained on images of celebrities (CelebA dataset) has demonstrated the ability to synthesise photorealistic images of human faces which are provably distinct from faces in the training dataset (Karras et al., 2017a). Two example generated images are shown in Figure 1.9a.

(3) Translation: Another feature of the generator network which has enabled many new applications is the assumption-free construction of a mapping from a latent space z to a desired data space x. GANs can be applied to construct any desired mapping. Promising use cases include super-resolution of undersampled images (Ledig et al., 2016) and image-to-image translation (Zhu and Park, CycleGAN, e.g. translate a photo taken in the summer to the same location in the winter).

1.5 Aims of this thesis

The aim of my research was to utilise a multidisciplinary approach from cell biology to deep learning in order to understand how epidermal cells make long and short term cell fate decisions at the single cell level.

I started by investigating the effect of Wnt/β-catenin signalling. My aim was to decouple the effects of autonomous and non-cell autonomous Wnt signalling. In order to do this I first investigated autonomous Wnt signalling and validated our method for inducing autonomous Wnt signalling (Chapter 2). Using a combination of single cell immunofluorescence and bulk gene expression analysis I established the transcriptional effects of autonomous Wnt activation.

Following these results I used single cell RNA-sequencing to determine the subpopulations present in mouse keratinocyte *in vitro* cultures (Chapter 3). This subpopulation information was used to map the single cell effects of neighbouring cell Wnt activation (non-cell autonomous Wnt activation) followed by validation of our findings.

Finally, during my research several public epidermal scRNA-seq datasets became available, including the dataset described in this thesis. I aimed to integrate these disparate datasets to uncover gene regulatory relationships governing epidermal cell state. To do this I developed a method applying generative adversarial neural networks which are able to integrate datasets with differing technical variation (Chapter

4). Using this deep learning approach I was able to simulate the effect of cell state perturbations, in particular differentiation, for single cells.

In total my experiments and analyses have investigated keratinocyte cell state and perturbations to cell state. My research on Wnt/β-catenin focused on one specific perturbation whereas the GAN research in Chapter 4 aimed to predict the effect of cell state perturbations and to define all possible epidermal cell states.

Chapter 2

Transient autonomous Wnt signaling in the epidermis

2.1 Introduction

Mammalian skin comprises many structures including nerves, blood vessels, hair follicles and the interfollicular epidermis (IFE). The IFE forms the outer covering of skin and functions as a barrier to the external environment through the continual shedding of terminally differentiated keratinocytes. IFE thickness and the position of associated adnexal structures differ according to location on the body. Epidermis has a stratified structure and is continually regenerated by stem cells residing in the basal layer. Epidermal stem cells undergo differentiation in the layers above this basal layer (suprabasal) and have a high capacity for self-renewal. It has been proposed that stem

cells divide infrequently and that their division can give rise to either more stem cells or transit amplifying (TA) cells (Potten, 1981). The latter, TA cells are capable of undergoing division several times before committing to terminal differentiation. Hence, they are responsible for amplifying the number of differentiated cells produced from one epidermal stem cell. While this concept has a number of attractive features, it has been challenged more recently (Jones et al., 2007).

Wnt signalling is involved in skin development at a very early stage (reviewed in Fuchs (2007)). Furthermore, it has been shown that Wnt/beta-catenin signalling is an important regulator of epidermal stem cells through its role in epidermal homeostasis (Choi et al., 2013). In all, Wnt signalling has been implicated in a diverse range of functions such as maintenance of stemness (Merrill, 2012), control of telomere length (Hoffmeyer et al., 2012), lineage selection (Lo Celso et al., 2008) and differentiation (Donati et al., 2014).

Previous studies have attempted to define the exact role of beta-catenin in the skin. Wnt activity in the epidermis has been shown to be spatially complex and temporally dynamic (Reddy et al., 2001), hinting that Wnt signalling outcome may depend on both extracellular environment and intracellular state. Transient nuclear accumulation of beta-catenin accompanied by upregulation of Wnt target genes has been observed during mouse embryogenesis concentrated in hair shaft precursor cells (Das-Gupta and Fuchs, 1999). Previous studies suggest that sustained beta-catenin activity can disrupt the normal reaction-diffusion mechanism responsible for hair follicle patterning. Additionally, it has been shown through in vivo xenopus (Moon et al., 1992)

and mouse (Lo Celso et al., 2004) inducible Wnt/beta-catenin models that varying the temporal nuclear accumulation of beta-catenin can lead to differing outcomes for epidermal stem cells. Taken together, these results suggest that there is a different optimal in vivo transient activation of beta-catenin signalling for each process regulated by Wnt such as epidermal stem cell maintenance, commitment to differentiation and hair follicle formation.

This chapter focuses on two main aims: **(1)** validation of an *in vitro* system for transient and constitutive Wnt activation (K14ΔNβ-cateninER, extensively used in Chapter 3), **(2)** characterisation of the transcriptional effects and differences between transient and constitutive Wnt activation. We define transient as less than three hours whereas constitutive represents continuous Wnt signalling. Using a combination of single cell immunofluorescence, mRNA-sequencing and analysis of alternative splicing events we show that the K14ΔNβ-cateninER transgene efficiently activates canonical Wnt signalling and identify a new role for β-catenin in regulation of splicing.

2.2 Results

2.2.1 Nuclear beta-catenin dynamics

K14ΔNβ-cateninER mice were previously generated in the lab (Lo Celso et al., 2004). The ΔNβ-cateninER construct was initially generated by in frame fusion of N-terminally truncated β-catenin (nucleotides 715-2604) to the ligand binding domain of a mutant

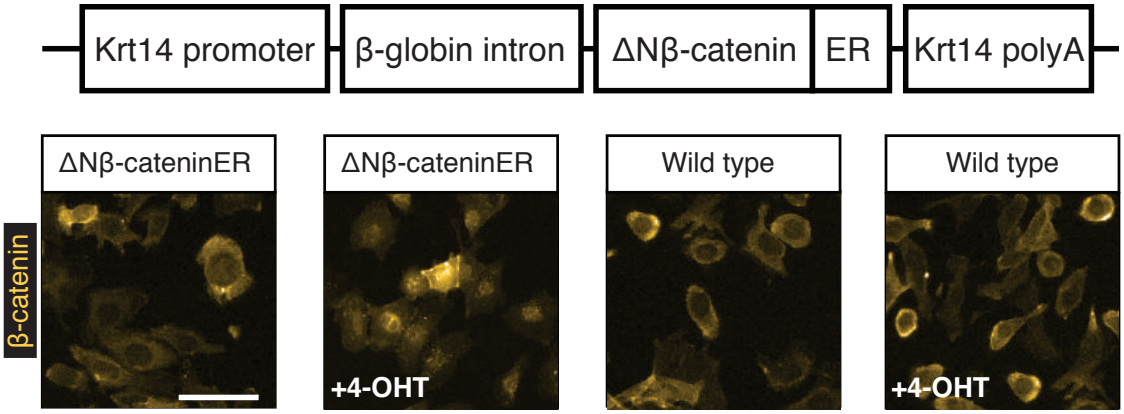


Figure 2.1: K14ΔNβ-cateninER Wnt activation construct.
Upper: K14ΔNβ-cateninER construct consists of the Keratin-14 promoter driving the expression of ΔNβ-cateninER. A β-globin intron was inserted 5' with respect to ΔNβ-cateninER and Keratin-14 polyA sequence inserted 3' in order to enhance expression of the fusion transcript. Lower: Immunofluorescence for β-catenin demonstrating increased nuclear abundance upon addition of 4OHT. Scale bar: 25μm.

murine estrogen receptor (ER), unable to bind endogenous oestrogen. Transgenic mice were generated by pronuclear injection of the construct into fertilised mouse embryos at day-1. Keratinocytes were isolated and cultured from the back skin of K14ΔNβ-cateninER mice as well as matched wild type mice. Both populations of cells underwent spontaneous immortalisation (see Romero et al. (1999)). It was confirmed that both sets of cells are able to grow in feeder-free conditions.

The upper portion of Figure 2.1 shows a schematic of the K14ΔNβ-cateninER construct. The beta-cateninER fusion protein is constitutively expressed under control of a Keratin-14 promoter, providing selective expression in basal epidermal cells. *In vitro*, we expect to constitutively express the K14ΔNβ-cateninER construct in cells maintaining an epidermal stem cell identity. Without activation of the fusion protein, the beta-cateninER protein accumulates in the cytoplasm. In the presence of 4-hydroxytamoxifen (4-OHT) the ER domain of the fusion protein changes confor-

mation to an activated state thus allowing translocation of beta-catenin from the cytoplasm to the nucleus.

We performed an initial experiment to validate the inducible beta-catenin construct *in vitro*. Wild type and ΔNβ-cateninER cells were treated for 48 hours in medium containing 4-OHT or an ethanol control. Function of our construct was confirmed using immunofluorescence for beta-catenin. Staining for beta-catenin showed a marked difference in nuclear intensity between treated ΔNβ-cateninER cells and control conditions (Figure 2.1, lower). These results confirmed that we are able to induce translocation of beta-catenin from the cytoplasm to the nucleus with this system.

We next investigated two important characteristics of the system: nuclear beta-catenin dynamics over time and how activation time affects markers of cell state. For these experiments immunofluorescence was used in combination with high content screening using the PerkinElmer Operetta. The Operetta allows high-throughout imaging of immunofluorescence stained 96-well plates. Coupling this approach with image processing software (Harmony and R) it is possible to quantify cytoplasmic and nuclear fluorescence at single-cell resolution.

In order to investigate the dynamics of beta-catenin translocation and its effects on cell state, we performed a time-course experiment. Four different activation scenarios were considered. We induced canconical Wnt signaling in the cells for 1, 3 and 24 hours alongside a control condition without activation. For each of these activation conditions, nuclear beta-catenin intensity was evaluated at +0, 3, 9 and 24 hours after

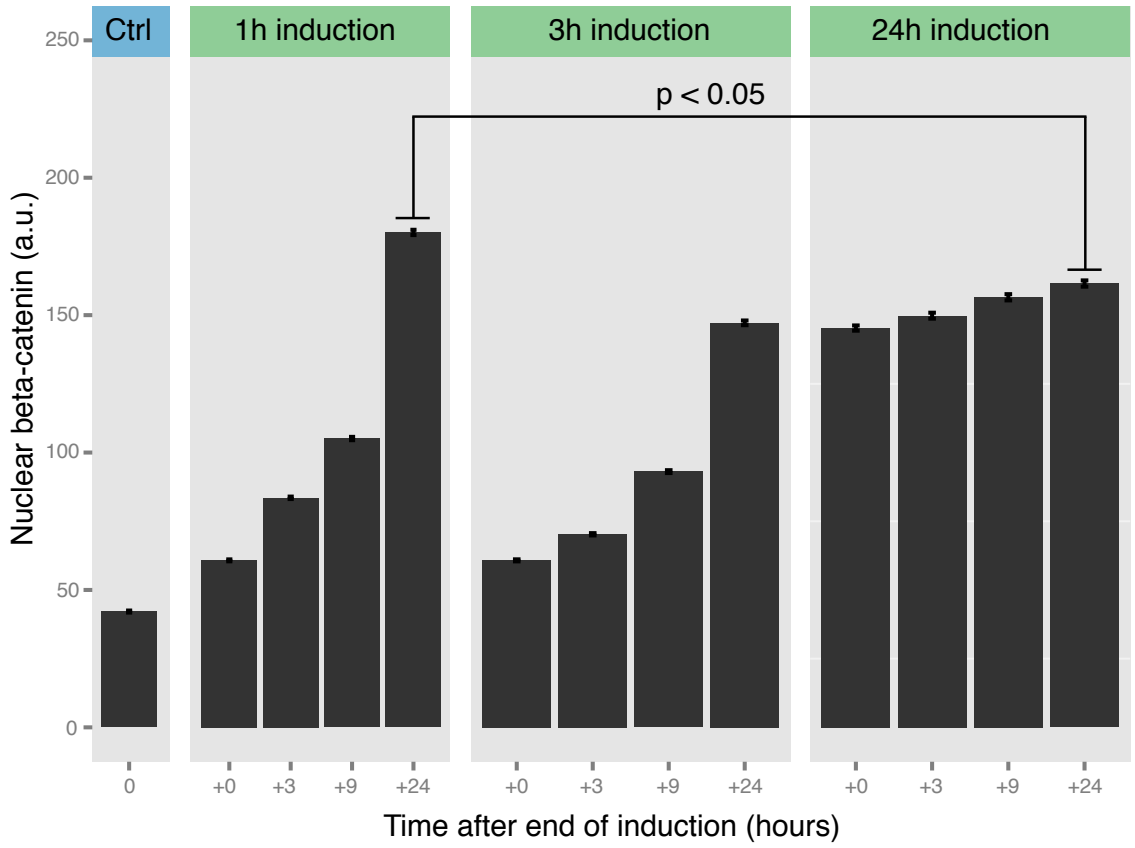


Figure 2.2: Nuclear B-catenin after Wnt activation.
Nuclear B-catenin dynamics after 1 hour, 3 hour and 24 hours of Wnt activation for K14ΔNB-cateninER keratinocytes compared to control. Errors bars are standard error (SE) ± mean. Nuclear beta-catenin measured in arbitrary units (a.u.) of fluorescence.

withdrawal of 4-OHT media as shown in Figure 2.2.

In the 1h and 3h activation conditions we observed an increase in nuclear beta-catenin until our final +24 hours timepoint. In comparison, there was little increase in the proceeding +24 hours of timepoints after the 24 hour Wnt activation condition. We interpreted this as meaning that 24 hours of constitutive activation saturates the maximum concentration of nuclear beta-catenin we can achieve using our system. Intriguingly, we observed that a 1h activation + 24 hours showed a higher abundance of

nuclear beta-catenin when compared to later timepoints in the 3 hour and 24 hour constitutive activations. We hypothesised that this was the result of a positive feedback gene expression circuit that allows a cell to obtain a higher level of nuclear beta-catenin when Wnt signalling is transiently activated for a short period of time. In contrast, constitutive activation results in a stronger negative feedback effect resulting in a lower concentration of nuclear beta-catenin.

2.2.2 Nuclear Lef1 and BLIMP1 dynamics following Wnt activation

Lef1 is a transcription factor which functionally interacts with beta-catenin as well as being directly downstream of Wnt and beta-catenin signaling activation. We sought to understand whether Lef1 nuclear abundance was differentially affected by transient Wnt activation. Figure 2.3a shows immunofluorescence of Lef1 and beta-catenin for control and Wnt activated ΔNβ-cateninER keratinoctyes 24 hours after beta-catenin activation. Nuclear fluorescence intensity was quantified for both Lef1 and beta-catenin for 1 hour, 3 hour and 24 hour induction times. All induction timepoints showed significantly increased nuclear Lef1 compared to control. The shortest activation, 1hr, showed the greatest increase in nuclear Lef1, supporting our hypothesis that a transient increase in nuclear beta-catenin leads to higher levels of nuclear beta-catenin and expression of downstream target genes.

We performed a similar analysis for BLIMP1, another transcription factor which has been shown to be downstream of beta-catenin activation. Unlike Lef1, BLIMP1

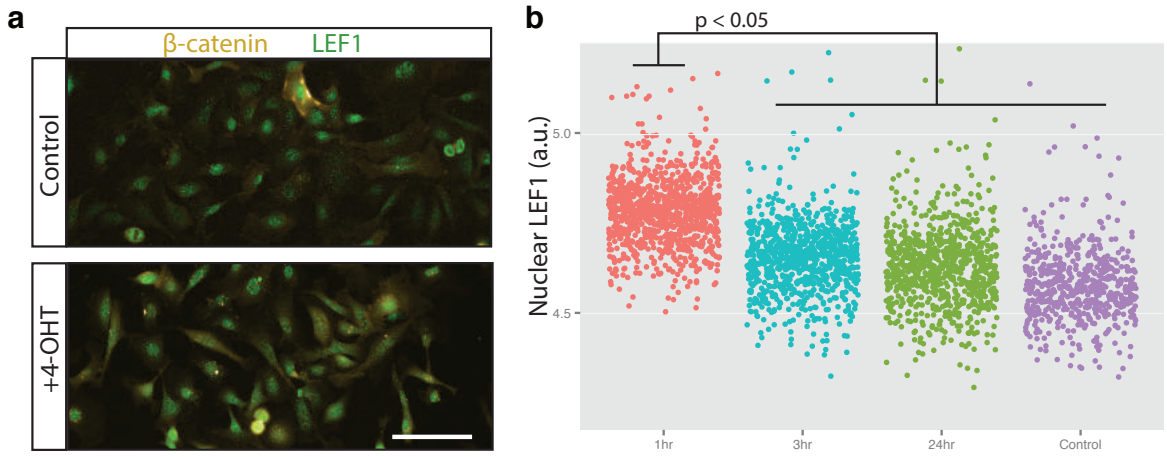


Figure 2.3: Nuclear LEF1 after Wnt activation.
(a) Immunofluorescence of β-catenin and LEF1 in control and Wnt activated (24hr) conditions. (b) Nuclear LEF1 abundance (arbitrary units) in single cells after 24 hours after 1hr, 3hr or 24hr Wnt activation compared to control. Scale, 100μm.

is not known to functionally interact with beta-catenin. Hence we sought to understand whether the transient positive feedback behaviour exhibited by beta-catenin and Lef1 extended to non-interacting proteins. All beta-catenin activation conditions showed significantly increased nuclear BLIMP1 in comparison to control (Figure 2.4a, $p < 0.05$, Kolmogorov-Smirnov test). Similarly to Lef1, the 1hr transient activation resulted in the greatest accumulation of nuclear BLIMP1. We further analysed the relationship between nuclear BLIMP1 and nuclear beta-catenin at the single cell level. Figure 2.4b shows the nuclear abundance of these two proteins in over 3000 cells under control and 1hr transient activation conditions. We observed that nuclear abundance of these two transcription factors is highly correlated (r^2 coefficient of determination of 0.85) indicating that these TFs are either co-regulated or one is directly regulating the other. After 1hr transient induction (+23 hr) this correlation is maintained (r^2 of 0.85) however the median nuclear abundance for both proteins increases, indicating

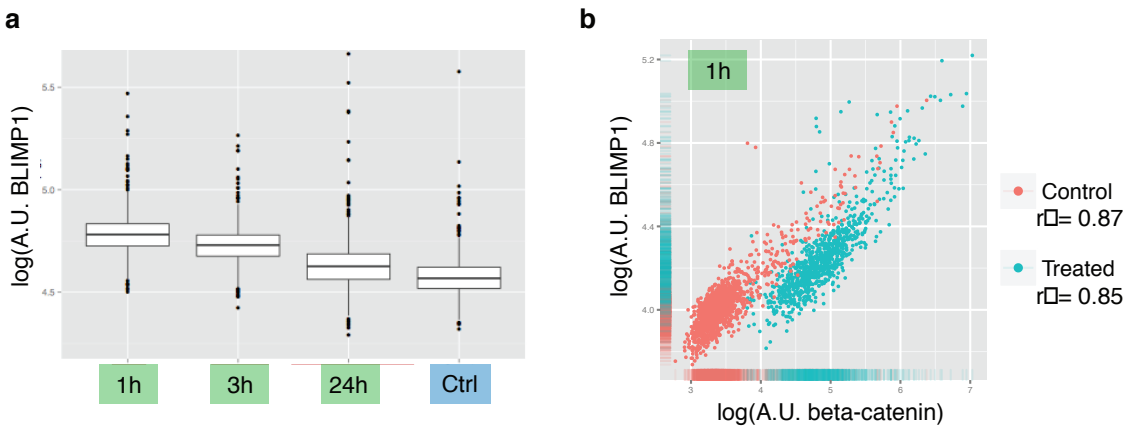


Figure 2.4: Relationship between BLIMP1 and Wnt activation.
(a) Nuclear BLIMP1 intensity (arbitrary units) 24 hours after 1hr, 3hr and 24hr activation compared to control (no activation). (b) Relationship between nuclear BLIMP1 and β-catenin in unactivated (control) and Wnt activated (treated, 24 hours after 1hr activation). r^2: coefficient of determination.

that BLIMP1 is likely to be a direct transcriptional target of beta-catenin.

2.2.3 Epidermal targets of Wnt/beta-catenin activation

We next sought to investigate the transcriptome-wide effects of transient and constitutive Wnt signaling. Downstream targets of Wnt/beta-catenin signaling are difficult to define in the skin using existing public data. One approach is to assume that targets that are differentially expressed upon activation of Wnt signaling in other model systems such as the intestine are also targets in the epidermis. Another approach is to utilise non-epidermal public data indicating where beta-catenin binds across the genome, i.e. chromatin immunoprecipitation sequencing (ChIP-seq). Both of these approaches have the disadvantage of ignoring skin-specific transcriptional targets, particularly important as beta-catenin is a co-factor and cannot activate transcription alone. Hence, beta-catenin binding of regulatory elements such as enhancers and pro-

moters can be the result of combinatorial binding with a variety of tissue-specific and non-specific co-factors including Lef1, Bcl9, Pygo (Cantù et al., 2017).

We performed bulk mRNA-sequencing of control, 1hr transient activation and 24hr constitutive activation conditions to determine transcriptome-wide gene expression changes. This allowed us to determine Wnt/beta-catenin target genes specific to either transient or constitutive activation as well as genes differentially expressed under both conditions. Figure 2.5 (upper) shows a volcano plot contrasting gene expression in the 1hr and 24hr activation conditions against control samples. We found 1404 genes differentially expressed after transient Wnt activation and 1492 genes differentially expressed under constitutive activation. We noted that both conditions were significantly enriched for differentially expressed long non-coding RNAs (lncRNA, 1hr n = 430, 24hr n = 439, $p < 0.05$, hypergeometric test). The locations of these non-coding transcripts can be categorised into three main groups: (1) located within the gene bodies of protein-coding genes, (2) located in intergenic regions nearby differentially expressed protein-coding genes and (3) located more than 200kb from the nearest differentially expressed gene. We excluded genes located within the gene bodies of differentially expressed genes to avoid misappropriating differential expression of these transcripts (Figure 2.5, lower).

Focusing on differentially expressed protein-coding genes, we were able to confirm efficient activation of the canonical Wnt pathway by examining expression of Axin2, a marker of Wnt activation and negative regulator of the Wnt pathway in multiple tissues (Jho et al., 2002). In both the 1hr and 24hr activation conditions Axin2 was

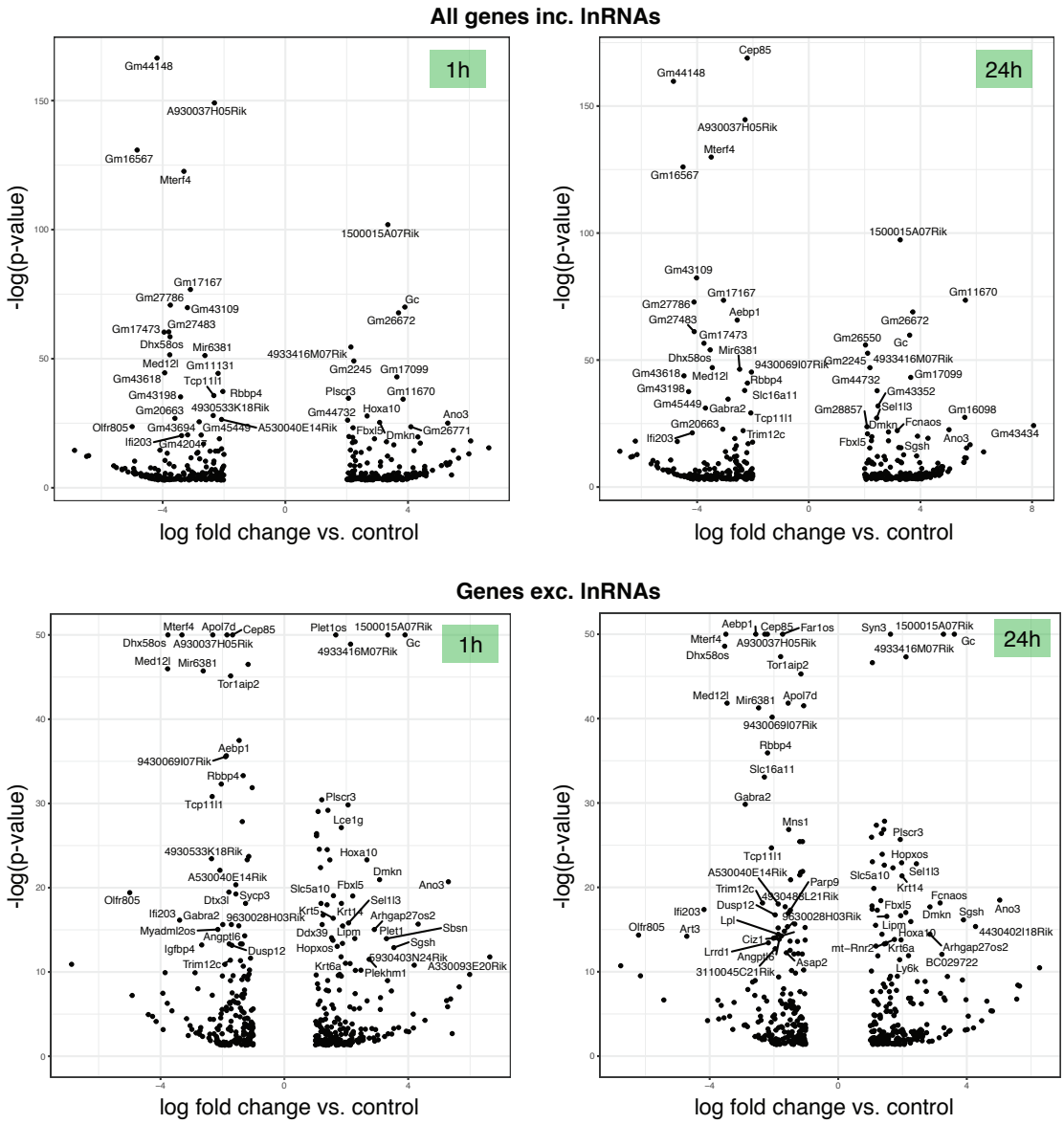


Figure 2.5: Genes differentially expressed in transient or constitutive Wnt activation.
Volcano plots showing differentially expressed genes for 1hr and 24hr Wnt activation relative to control. Upper panels show all genes, lower panels show genes excluding long non-coding RNAs.

upregulated over ten-fold relative to control (adjusted p-value = 3.2e-3). Furthermore, 72% of identified differentially expressed genes were also identified *in vivo* by a previous study in our lab, demonstrating the relevance of our *in vitro* activation model (Donati et al., 2014).

Both sets of differentially expressed genes were highly enriched for epidermal specific Gene Ontology terms such as epidermis development and epidermis morphogenesis (Figure 2.6a and 2.6b). Finally we examined Gene Ontology enrichment for genes differentially expressed in both activation condition (2.6c). This yielded similar results, however three RNA-binding related Gene Ontology terms were present amongst the most significant terms. This enrichment alongside the differentially expressed lncRNAs suggested that a common effect of Wnt activation in both conditions involved the perturbation of RNA-binding proteins and non-coding RNAs.

2.2.4 Transcriptional differences between transient and constitutive Wnt activation

In addition to determining targets of Wnt activation, we contrasted transient (1hr) and constitutive (24hr) activation of the Wnt pathway to determine transcriptional targets that are sensitive to duration of Wnt signalling. Figure 2.8 shows a volcano plot contrasting expression of genes in these two conditions. In contrast to our previous immunofluorescence assays of beta-catenin, LEF1 and BLIMP1 where we observed a strong difference, we found fewer than 200 differentially expressed genes with fewer

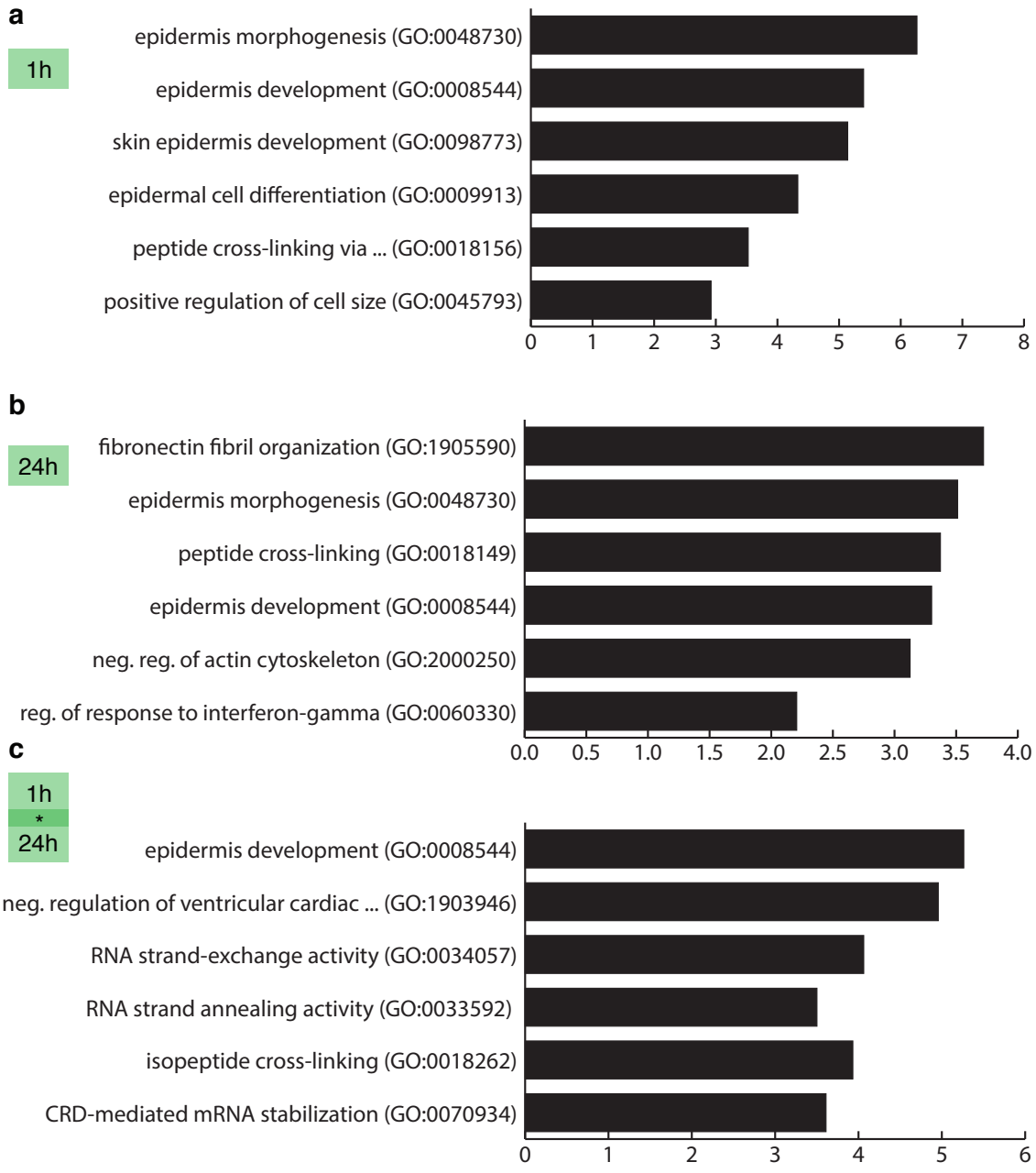


Figure 2.6: Gene ontology enrichment analysis for *beta*-catenin regulated genes.
Gene ontology enrichment analysis for genes differentially expressed between (a) 1hr and control, (b) 24hr and control, and (c) genes commonly differentially expressed in both 1hr and 24hr activation conditions vs. control.

RBM25, HSP90AB1, DDX49, MRPS14, CCDC124, GDI2, HNRNPU, FAM208A, TCF20, YBX1, ADARB2, IFIT1, SPOUT1, RRP9, CWC22, RPL36, RPS10, LRRC47, DSP, CCT3, CAST, EIF5B, ANXA2, ZCRB1, YWHAZ, FAM133B, EEF1A1, PKM, SLTM, HNRNPUL2, RPL26, SNRPC, FKBP3, SKIV2L2, STAU1, AHNAK, ANP32A, USP10, CSTF3, MAK16, NOL7, NOL8, RRBP1, ATXN2, S100A16, CTNNA1, RBBP6, EIF4B, RPL41, CPSF6, API5, KRR1, DEK, RPS26, ALDH6A1, H1F0, CAPRIN1, SERBP1, TSR1, HDGF, CEBPZ, SLIRP, EIF3A, EIF3B, METAP2

Figure 2.7: Differentially expressed mRNA-binding proteins.

than 10 genes up- or down-regulated at biologically meaningful log fold change and stringent statistical significance (log fold change > 1 and -log(p-value) > 5 represented by green region in Figure 2.8). From the few differentially expressed genes, one gene of interest is Notum, which was upregulated four-fold in the 24hr activation condition relative to 1hr activation. Notum is an extracellular protein with carboxyl oxoesterase activity and is known to deacylase Wnt ligands. Through this mechanism, Notum has been shown to strongly suppress Wnt signalling activity (Kakugawa et al., 2015).

Our results show that transcription of Notum is activated under constitutive Wnt signalling but not after transient Wnt signalling. We hypothesise that under constitutive Wnt activation, deacylation of endogenous Wnt ligands by Notum *in vitro* results in lower nuclear beta-catenin abundance. In contrast, transient Wnt signalling does not lead to upregulation of Notum, with the consequence of greater nuclear beta-catenin abundance. Furthermore, few genes are differentially expressed between constitutive and transient activation conditions. Hence, we hypothesise that post-transcriptional mechanisms are likely to be contributing to differences between transient and constitutive Wnt activation.

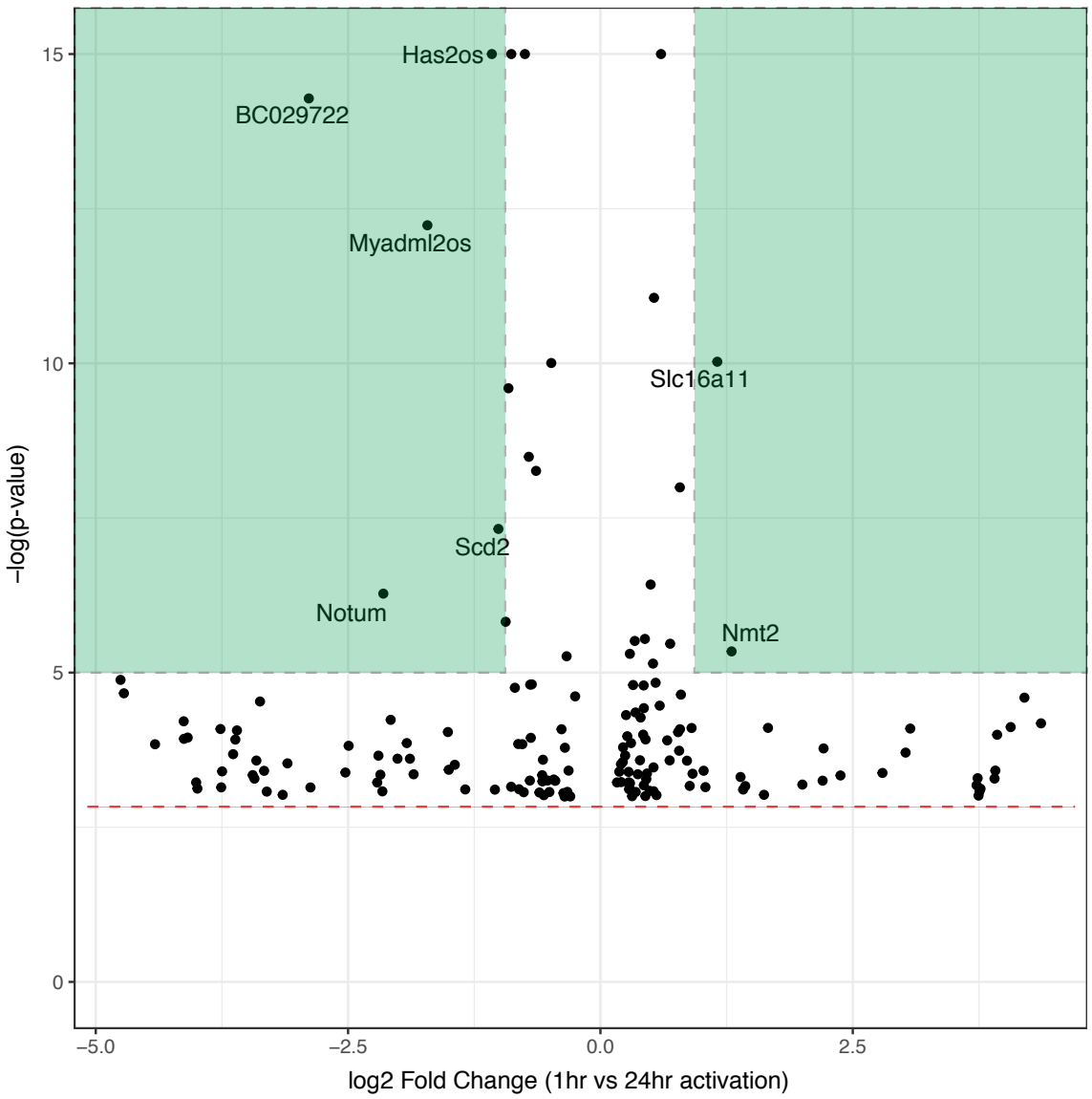


Figure 2.8: Genes differentially expressed between transient and constitutive Wnt activation.
Volcano plot showing differentially expressed genes between 1hr and 24hr Wnt activation (log fold change is 1hr vs. 24hr). Labelled genes are differentially expressed at log fold change > 1 and significant at -log(p-value) > 5. Red line denotes significant differential expression threshold.

2.2.5 Epidermal Wnt/beta-catenin activation regulates intron retention

There is a growing body of evidence that lncRNAs regulate mRNA splicing alongside established roles such as translational repression (Romero-Barrios et al., 2018). In addition, in colon cancer cells beta-catenin has been shown to regulate alternative splicing of oncogenic transcripts (Lee et al., 2006). The presence of multiple differentially expressed RNA-binding proteins for both Wnt activation conditions (see Box 2.7) indicates that post-transcriptional regulatory mechanisms are downstream of Wnt signalling. To examine whether Wnt activation regulates alternative splicing in keratinocytes we used Whippet, a recently developed method for detection and quantification of alternative splicing events (Sterne-Weiler et al., 2017). We identified similar numbers of alternative splicing events in control, 1hr and 24hr activation conditions (n=9950, 9624 and 9946 splice nodes respectively).

Alternative splicing (AS) events assessed by Whippet fall into six main categories: retained introns (RI), core exons (CE), alternative first exons (AF), alternative last exons (AL), alternative donor (AD) and alternative acceptor (AA) splice sites. In all samples, over half of all AS events comprise alternative acceptor and donor events which are 5′/3′ ungapped shortening or extension of exons. From these categories, retained introns showed the greatest change between control and Wnt activation conditions, with approximately 10% fewer retained intron events in both Wnt activation conditions relative to control. The 1hr and 24hr activation conditions showed similar

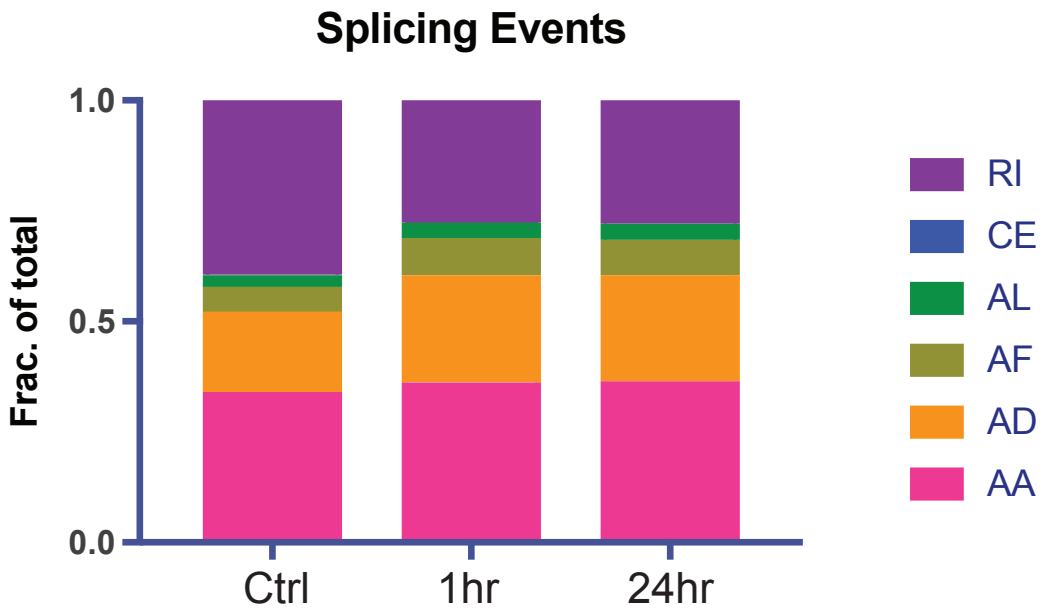


Figure 2.9: Proportion of alternative splicing events.
Proportion of alternative splicing events detected in control, 1hr and 24hr activation conditions. RI - retained intron, CE - core exon, AL - alternative last exon, AF - alternative first exon, AD - alternative donor splice site, AA - alternative acceptor splice site.

proportions of AS events for all categories.

Focusing on retained introns, we contrasted control samples against both 1hr and 24hr activated samples to determine retained intron events which were insensitive to Wnt activation time. We applied a high stringency threshold (p-value < 0.01, $\Delta\Psi > 0.3$, Ψ/PSI, percent spliced in) to obtain high confidence transcripts with differential retained introns relative to control. Figure 2.10a & 2.10b show two examples of identified retained intron events. Keratinocyte associated protein 3 (Krtcap3) shown in Figure 2.10a demonstrates increased retention of the intron between exon 4 and exon 5 with longer beta-catenin activation. Fused in sarcoma (FUS) shown in Figure 2.10b is another interesting example. Here, the intronic region between exons 14 and 15

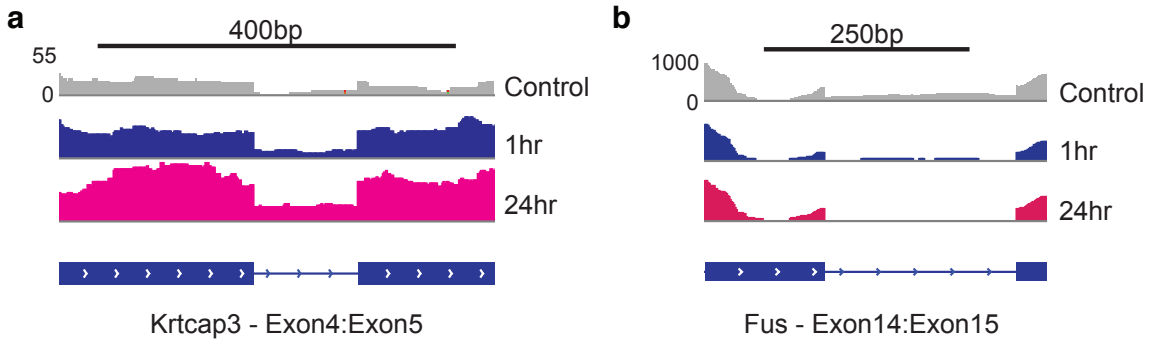


Figure 2.10: Example retained intron events.
Two examples of retained intron events, (a) shows the intronic region between exon 4 and exon 5 for Krtcap3, (b) shows the intronic region between exon 14 and exon 15 for Fus. Both examples show differentially retained introns between Wnt activation conditions and control (adjusted p-value < 0.05).

for Fus show highest intron retention in the control samples and this is reduced upon beta-catenin activation. Fus mutations have been causally linked to RNA processing defects in familial amyotrophic lateral sclerosis (ALS) (Kapeli et al., 2017). Furthermore, a recent study of intron retention in ALS demonstrated that RNA-binding proteins and regulators of RNA processing such as Fus are the targets of intron retention (Luisier et al., 2018).

We identified 482 transcripts for the 1hr activation and 508 transcripts for the 24hr condition with 202 introns differentially retained in common (Figure 2.11a). These genes include known functional epidermal proteins such as Sprr2h, Krt7, Itga7 and Krtcap3. Gene ontology enrichment showed the 202 genes to be highly enriched for mRNA-binding, regulation of splicing and for proteins localised to nuclear speckles as shown in Figure 2.11b. This suggests an unexpected feedback loop where activation of Wnt signaling results in an increase of retained intron events within transcripts that themselves are regulators of splicing. We next compared intron retention for the

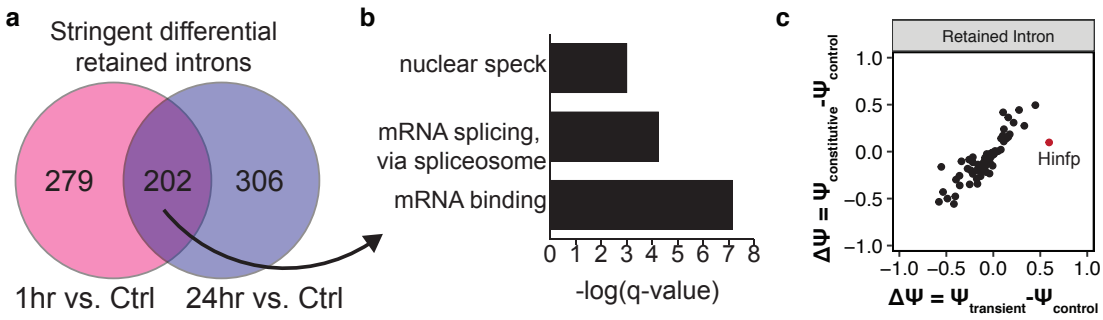


Figure 2.11: Retained introns insensitive to Wnt activation time.
(a) Venn diagram showing the overlap between stringent differential retained intron analysis for 1hr vs control and 24hr Wnt activation vs. control. Overlap represents introns which are differentially retained regardless of Wnt activation time. (b) Gene ontology enrichment analysis of common retained intron genes. (c) Relationship between ΔPSI for transient and constitutive activation.

202 common introns. Figure 2.11c shows differential intron usage between activation conditions and control. We found only one transcript, for the gene Histone H4 Transcription Factor (Hinfp), which showed significant differential intron usage when comparing 1hr and 24hr activation conditions. These results suggest that beta-catenin regulation of intron retention requires only a short duration of pathway activation (one hour) and is sustained under constitutive activation conditions.

2.3 Conclusions

In this chapter we validated an *in vitro* system for Wnt activation (K14ΔNβ-cateninER) and used this to investigate transcriptional differences between transient and constitutive activation. In line with transcriptional effects of Wnt activation in other tissues such as the intestine, we found upregulation of canonical Wnt target genes under both transient and constitutive activation. Previous *in vivo* data by Lo Celso et al. from our

lab showed that transient activation of Wnt signalling in murine epidermis leads to the formation of a new hair follicle growth, potentially via specification of susceptible cells to a hair follicle lineage.

Using K14ΔNβ-cateninER keratinocytes, we have shown that transient and constitutive activation of epidermal cells leads to distinct cell states by altering levels of key transcription factors (BLIMP1, LEF1). Furthermore, in the case of constitutive activation, negative feedback in the Wnt pathway is activated through upregulation of Notum, a suppressor of Wnt signalling. Notum's role in epidermal Wnt signalling requires further validation, in particular whether inhibition of Notum results in higher levels of nuclear beta-catenin under constitutive activation conditions. Unpublished data within the lab and observations by Gernot Walko and Angela Oliveira Pisco indicate that keratinocytes *in vitro* are insensitive to Wnt pathway activation through Wnt ligands. Notum acts by deacylating and inactivating Wnt ligands, and hence it would be of interest to investigate whether these observations are the result of Notum-mediated Wnt inactivation.

Finally, by analysing RNA-sequencing data for alternative splicing events, we have shown for the first time a link between Wnt activation and intron retention. Using Whippet, an alternative splicing bioinformatic tool, we established normal alternative splicing activity under control conditions for epidermal keratinocytes and contrasted these to constitutive and transient Wnt activation. Intriguingly, we observed that intron retention is most prominent under Wnt activated conditions for transcripts which are themselves regulators of splicing.

Taken together this chapter validates Wnt activation using the K14ΔNβ-cateninER system and elucidates how Wnt signalling is able to encode for multiple cell state outcomes by differing in outcome depending on time of activation. Our findings on intron retention further highlight how Wnt signalling is able to regulate multiple cell states through a combination of transcriptional activation and regulation of splicing.

2.4 Methods

2.4.1 Cell biology

Wnt activation in K14ΔNβ-cateninER keratinocytes

K14ΔNβ-cateninER transgenic mice were generated as previously described (Lo Celso et al., 2004). The ΔNβ-cateninER construct was initially generated by in frame fusion of N-terminally truncated β-catenin (nucleotides 715-2604) (Zhu and Watt, 1999) to the ligand binding domain of a mutant murine estrogen receptor (ER), unable to bind endogenous oestrogen (Littlewood et al., 1995). Transgenic mice were generated by pronuclear injection of the construct into fertilised mouse embryos at day 1.

Cell isolation and culture

The following cell isolation and culture protocols have been described previously by Jensen et al. (2010).

Culture medium for 3T3-J2 fibroblasts (feeder cells)

Feeder cells were cultured in Dulbecco's modified Eagle's medium (DMEM) containing 10% bovine serum (Gibco) supplemented with 100IU/mL penicillin (Life Technologies), 100μg/mL streptomycin (Life Technologies). Medium was stored at 4°C.

Mitomycin C for feeder cell mitotic inactivation

To prepare a 100x (o.4mg/mL) stock solution of Mitomycin C, 2mg of Mitomycin C powder (Sigma-Aldrich) was dissolved in 5mL of Milli-Q water. The solution was sterilised using a 0.22μm filter (Millipore). Stock solution was aliquoted and stored at -20°C.

Tamoxifen for activation of K14ΔNβ-cateninER cells

For activation of Wnt signalling, cells were treated with 4-OHT (200nM) or ethanol as a control. For 4-OHT stock, powder was dissolved in ethanol to form a 1000x (2mM) stock solution and aliquoted to be stored at -20°C. 4-OHT was added to room temperature medium immediately before experiments.

Culture medium for mouse keratinocytes

We cultured mouse keratinocytes in Calcium-free FAD medium (one part Ham's F12, three parts Dulbecco's modified Eagle's medium, 1.8×10^{-4} M adenine), supplemented

with 10% foetal calf serum (FCS) and a cocktail of 0.5 μg/ml hydrocortisone, 5 μg/ml insulin, 1×10^{-10} M cholera enterotoxin and 10 ng/ml epidermal growth factor (HICE cocktail) (Watt et al., 2006).

Cell isolation from mouse back skin

Shaved back skin was sterilised using 10% betadine solution for two minutes followed by two sterilisation baths in 70% ethanol solution for 1 minute each. Tissue was subsequently washed for 1 minute in sterile PBS. After this point we ensured that all further equipment coming into contact with the tissue is sterilised. Muscle and fat was scraped from the underside of the tissue using a scalpel. To facilitate separation of dermis and epidermis, skin was then incubated overnight (epidermal side facing up) in 0.25% trypsin (Life Technologies) without EDTA at 4°C. On the following day epidermis was scraped away from the dermis using sterile scalpels and subsequently minced using two scalpels. Minced epidermal tissue was resuspended in complete low calcium FAD medium and filtered through a 70μm cell strainer (BD Biosciences). The resulting cell suspension was centrifuged at 500g for 8 minutes at room temperature to form a cell pellet. Next the cell pellet was resuspended in low calcium complete FAD medium and plated for cell culture.

Culture of mouse keratinocytes

Mouse keratinocytes were either seeded onto mitotically inactivated feeder cells (if primary) or cultured in feeder-free conditions if immortalised. For routine cell culture keratinocytes were seeded at a density of 2.5×10^5 cells per T75 flask and cultured in complete low calcium FAD medium in an incubator with 8% CO_2 at 32°C. Medium was replaced every 48-72 hours and cells were passaged when confluent approximately 7 days after plating. In cultures with feeders, prior to passaging feeders were first removed by incubating cells in Versene for five minutes at room temperature followed by tapping on the side of the flask to free feeders form the flask surface leaving attached keratinocytes. Keratinoctyes were next incubated in 0.05% trypsin solution diluted in Versene for 3-8 minutes until keratinocytes detached from the flask surface. Detached cells were suspended in 5mL of complete FAD medium (to inactivate trypsin) and centrifuged for 5 minutes at 500g to form a cell pellet. Supernatant was aspirated above the cell pellet and cells were resuspended in fresh medium for counting with a haemocytometer to determine cell density. Cells were then replated as appropriate.

Storage and freezing of mouse keratinocytes

For long term storage, cells were detached and cell pellets formed as above. Supernatant was removed and replaced with FBS containing 10% DMSO (Sigma-Aldrich). Cells were resuspended at a density of 1×10^6 cells per mL. Cell suspension was tran-

ferred to 1mL cryovials. To minimise the stress of freezing on cells, cryovials were stored overnight in a container filled with isopropanol at -80°C to ensure a constant cooling rate. After 24 hours cells were transferred to a liquid nitrogen cell bank for long term storage.

2.4.2 Immunofluorescence and high-content imaging

Immunofluorescence staining

The following antibodies were used: β-catenin (1:250, Sigma C2206), Blimp1 (1:250, Santa Cruz, sc-47732), Lef1 (1:250, Abcam, ab137872).

Cultured cells were fixed with 4% PFA for 10 minutes followed by permeabilisation with 0.1% Triton X-100 for 10 minutes at room temperature. Cells were blocked for 1 hour at room temperature with 1% BSA in PBS. Primary antibody incubation was carried out for 90 minutes at room temperature. Samples were labelled with Alexa Fluor (488, 555, 647)- conjugated secondary antibodies for 1 hour at room temperature. Cells were imaged within 24 hours using an Operetta or Operetta CLS High-content Imaging System (PerkinElmer). Single cell cytoplasmic and nuclear fluorescence intensities were quantified with Harmony software (PerkinElmer) and analysed in R.

2.4.3 Bulk mRNA-sequencing and analysis

Bulk RNA extraction, library preparation and sequencing

Total RNA was purified with the RNeasy mini kit (Qiagen) with on-column DNaseI digestion, according to the manufacturer's instructions. RNA quality was assessed using the RNA ScreenTape system (Agilent), all libraries scored greater than 9.6 on the RNA integrity number (RIN) scale. RNA-sequencing libraries were made with a TruSeq RNA Sample Preparation Kit V2 according to the manufacturer's instructions and were sequenced using the HiSeq 2500 System with 75bp paired-end reads. Libraries were sequenced to a depth of 15-26 million reads per sample at the Advanced Sequencing Facility (Francis Crick Institute).

Processing of reads and quality control

Sequenced libraries were checked for quality control and common sequencing errors (e.g. high adapter contamination) using FastQC and Cutadapt to trim adapter sequences (Martin, 2011). Sequences were aligned to the Mus Musculus genome (GRCm38) using STAR (Dobin et al., 2013) discarding multiply-mapped reads. Gene level counts were extracted using featureCounts (Liao et al., 2014). Transcript levels were quantified as transcripts per million (TPM). Differentially expressed genes were determined using DESeq2 (Love et al., 2014).

Alternative splicing analysis

We used Whippet (v0.6, Sterne-Weiler et al. (2017)) to analyse RNA-Seq data to identify alternative splicing events. Whippet creates comprehensive splice graphs required for quantification of exon and intron usage. Genome annotation files were obtained from Ensembl and Whippet was run using default settings to obtain percent spliced in (PSI) usage for splice nodes. Nodes are defined as non-overlapping exons which together form the alternative splicing event graph of all possible alternative splicing combinations. Differential splice node usage was quantified using Whippet-delta.

Chapter 3

Non-cell autonomous Wnt signalling

3.1 Introduction

The mammalian epidermis comprises interfollicular epidermis (IFE), hair follicles, sebaceous glands and sweat glands. Under steady-state conditions, each of these compartments is maintained by distinct populations of stem cells. However, following wounding each stem cell subpopulation exhibits the capacity to contribute to all differentiated lineages (Page et al., 2013). Recent single-cell gene-expression profiling of adult mouse epidermis identified multiple epidermal subpopulations (Joost et al., 2016). Furthermore, in cultures of human and mouse keratinocytes there are three or more subpopulations with varying proliferative potential (Roshan et al., 2016; Jones and Watt, 1993).

One pathway that plays a key role in regulating stem cell renewal and lineage

selection in mammalian epidermis is Wnt/beta-catenin signalling, which is an important regulator of epidermal maintenance, wound repair and tumorigenesis (Lim and Nusse, 2013; Watt and Collins, 2008). Gene-expression profiling has identified a number of signalling pathways that are regulated by cell-intrinsic activation of beta-catenin. Wnt signalling is indispensable for adult epidermal homeostasis; loss of beta-catenin in the IFE causes a defect in stem-cell activation, resulting in reduced basal layer proliferation and IFE thinning (Choi et al., 2013; Lim et al., 2013) and loss of hair follicles. Conversely, transient activation of epidermal beta-catenin in the adult epidermis leads to expansion of the stem-cell compartment and results in the formation of ectopic hair follicles at the expense of the sebaceous glands and an increase in IFE thickness (Silva-Vargas et al., 2005; Lo Celso et al., 2004).

There is good evidence that intrinsic beta-catenin activation in epidermal keratinocytes leads to effects on neighbouring epidermal cells. For example, in the mouse hair follicle, activated mutant beta-catenin cells can co-opt wild type cells to form a new hair growth through secretion of Wnt ligands (Deschene et al., 2014). This form of non-cell autonomous (NCA) activation suggests that autonomous Wnt signalling has the capability of changing neighbour cell fate. Although the mechanisms of autonomous Wnt activation are well described, it is unclear how NCA effects differ to cell intrinsic effects and how beta-catenin can simultaneously regulate self-renewal while changing the fate of neighbouring cells.

In this study we set out to analyse NCA signalling in wild type mouse keratinocytes that were co-cultured with keratinocytes in which beta-catenin was activated. This has

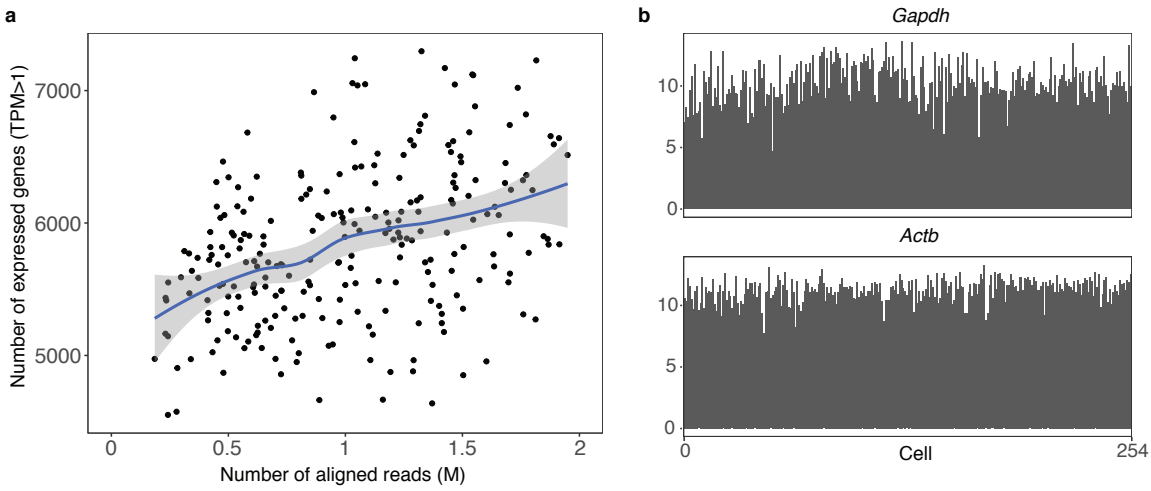


Figure 3.1: Quality control metrics for single cell libraries.
(a) Relationship between number of aligned reads and number of gene expressed above a threshold (TPM>1).
(b) Gene expression barplots for two example uniquitously expressed genes, Gapdh and Actb.

revealed previously unknown heterogeneity of wild type mouse keratinocytes and elucidated the effect of Wnt signalling on neighbouring cell state and heterogeneity.

3.2 Results

3.2.1 Single-cell mRNA-seq analysis of basal epidermal stem cells

To explore the effects of non-cell autonomous Wnt signalling on epidermal cell state we sequenced the transcriptomes of single wild type murine keratinocytes co-cultured with cells expressing an inducible form of stabilised beta-catenin (K14ΔNβ-cateninER) in a ratio of 9:1. We compared cells cultured in the absence of 4-hydroxy-tamoxifen (4OHT) with cells treated for 24h with Tamoxifen to induce beta-catenin. Cells were then disaggregated, loaded onto the C1 96-well microfluidic device (Fluidigm) and

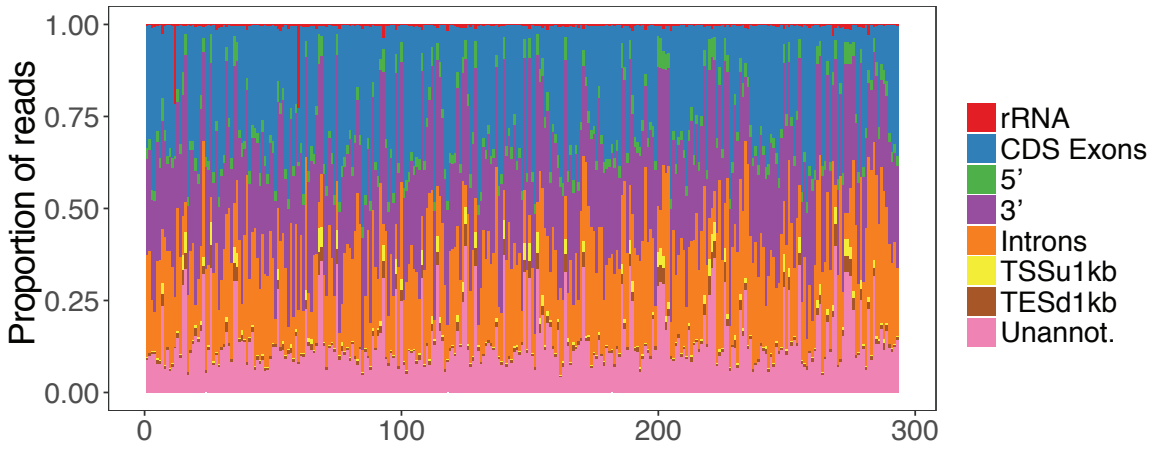


Figure 3.2: Read alignment distribution for single cell libraries.
Stacked proportional barplot showing read alignment coverage distribution for all cells before QC filtering. TSSu1kb - region up to 1kb upstream of the transcription start site. TESd1kb - region up to 1kb downstream of the transcription end site. Unannot. - unannotated intergenic regions.

captured for sequencing. Owing to the single-cell capture method used, highly keratinized and terminally differentiated cells over 20 microns in diameter were excluded. We identified K14ΔNβ-cateninER cells by aligning reads to the transgene sequence and subsequently removed these cells from analysis (10 untreated cells and 14 activated cells). After quality control we retained 125 wild type control cells and 129 wild type cells exposed to Wnt signalling neighbours. We recorded a median of 641,000 reads per cell equating to 4,000-8,000 genes expressed per cell as shown in Figure 3.1a (transcripts per million, TPM > 1). The majority of cells showed high expression of Gapdh and Actb, two genes we expect to be ubiquitously expressed regardless of cell type or cell state (Figure 3.1b). Read alignment distribution was in line with other single-cell RNA-seq datasets with minimal ribosomal and intergenic reads (Figure 3.2).

To explore cell-state heterogeneity in wild type keratinocytes that had not been

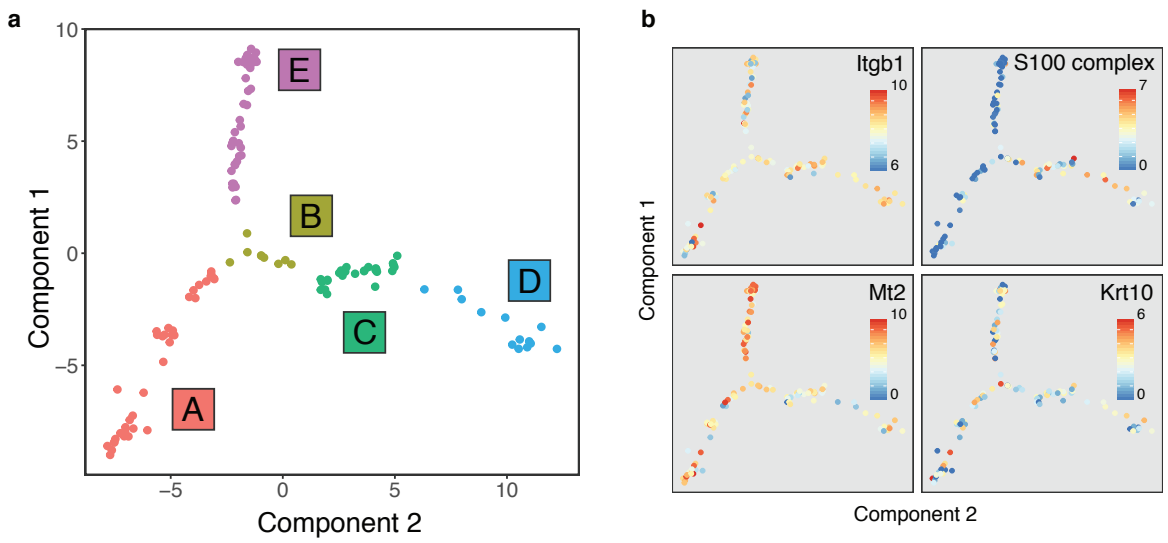


Figure 3.3: Molecular heterogeneity of epidermal cells in culture.
(a)Epidermal cell transcriptomes and cell state relation visualised using DDRTree and coloured according to unsupervised clustering. Each data point is one cell and axes are dimensionally reduced components of the cell transcriptomes. Colours and numbers represent the five identified cell states. (b) Gene expression for four marker genes shown for each cell on the state map. Top left: Integrin beta-1 (Itgb1), a basal IFE marker. Top right: S100 differentiation associated genes. Bottom left: Mt2, a basal IFE marker. Bottom right: Keratin 10, a suprabasal IFE marker of commitment to differentiation.

exposed to a neighbour in which beta-catenin was activated (untreated samples) we used reverse graph-embedding, a machine-learning technique. This enabled us to reconstruct a parsimonious tree connecting all observed epidermal cell states (DDRTree, Monocle 2, Trapnell et al. (2014)). We applied the DDRTree algorithm to wild type cells using expressed genes (TPM > 1) after removing cell-cycle associated genes. We identified five distinct *in vitro* cell states (Figure 3.3a) forming three major branches representing varying states of proliferation and differentiation.

States A and E showed highest expression of Mt2 a basal IFE marker alongside markedly low expression of S100 epidermal differentiation complex genes in comparison with the remaining subpopulations. These expression patterns indicate states A and E represent transcriptomic signatures prior to commitment to differentiation

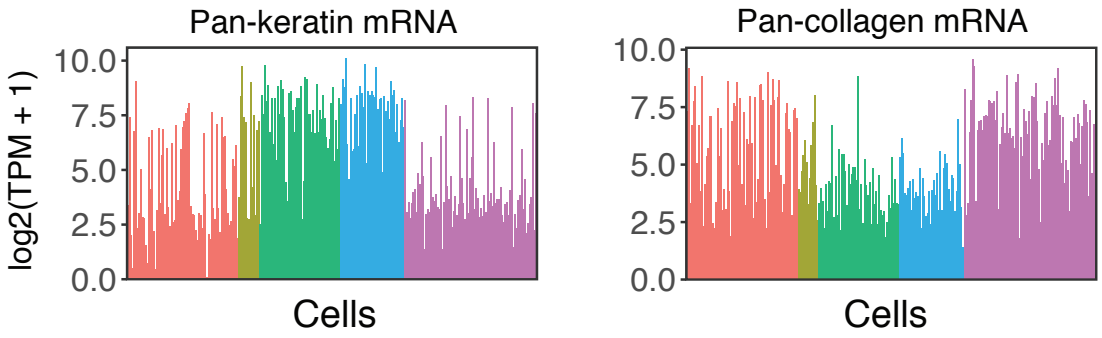


Figure 3.4: Pan-keratin and pan-collagen mRNAs.
Barplot showing mean expression per cell averaged over all keratin associated messenger RNAs (mRNA) (left) and all collagen mRNAs (right). Bars are coloured by cell state identity.

(Figure 3.3b; upper left) (Kypriotou et al., 2012). Itgb1, another marker of basal IFE cells showed variable expression *in vitro* in comparison to Mt2, but is expressed in all cells (Figure 3.3b; upper right). Similarly the differentiation marker Krt10 is variably expressed across all subgroups (Figure 3.3b; lower right) (Jones and Watt, 1993). Separation between pre- and post-commitment cell states is further apparent when looking at pan-keratin and pan-collagen gene expression. *In vivo,* keratinocytes commit to differentiation upon detaching from the basement membrane reducing the need for collagen expression and increase in overall keratin content. States B, C and D express significantly more keratin mRNAs and conversely states A and E are characterised by higher collagen mRNA levels (Figure 3.4; Kolmogorov-Smirnov test, $p < 0.05$).

For each cell state we determined genes differentially expressed versus the remainder of the population and identified between 6 (State B) and 101 (State C) markers (Figure 3.5a). We observed that the median expression of the top three markers for each state was sufficient to distinguish each state (Figure 3.5b)

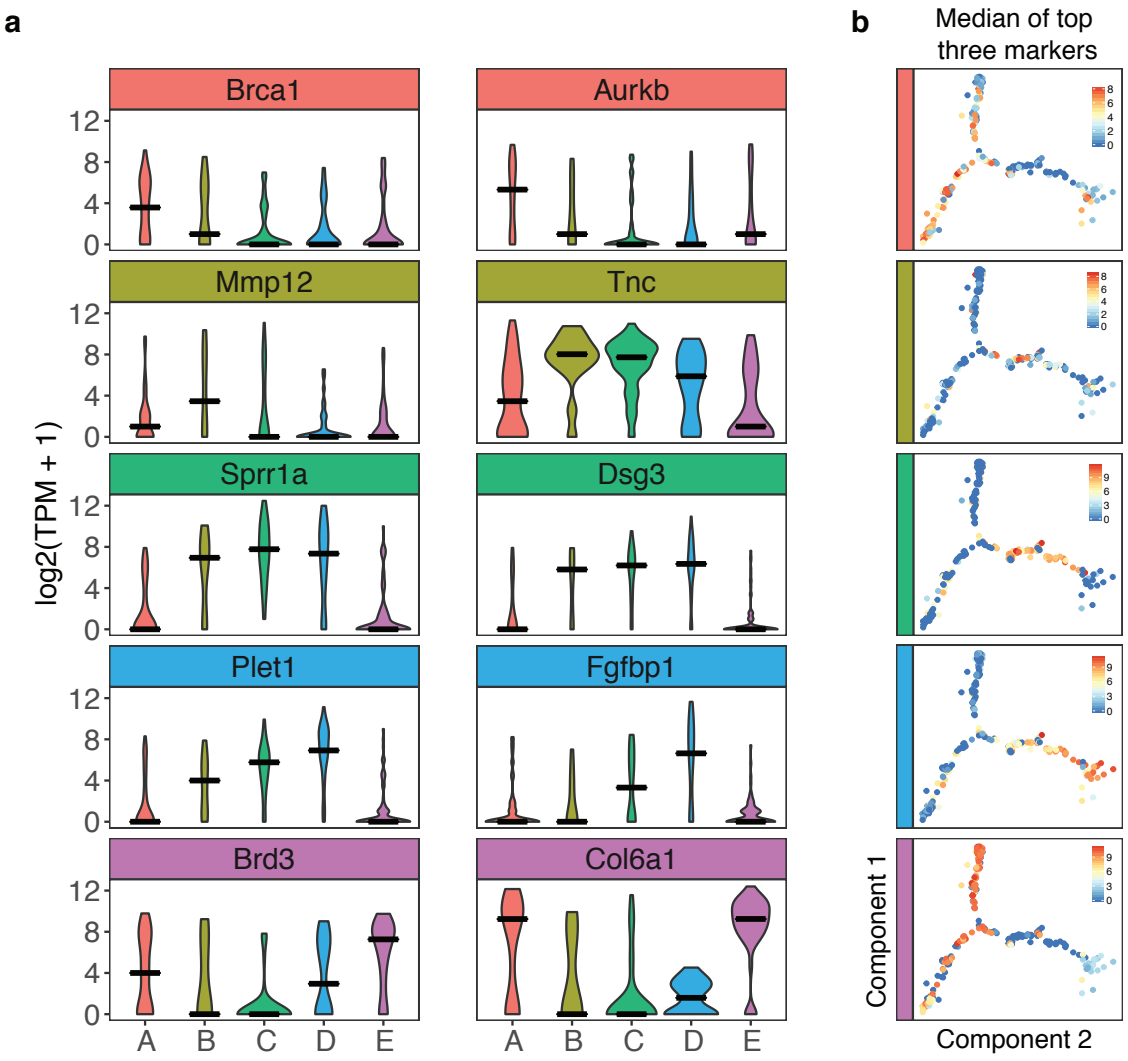


Figure 3.5: *In vitro* subpopulation markers.
(a) Distribution of the expression of cluster-specific markers coloured according to cell state. Two markers are shown per cell state. Black line represents the median expression of the cluster. (b) Median expression of the top three markers for each state per cell on the state map.

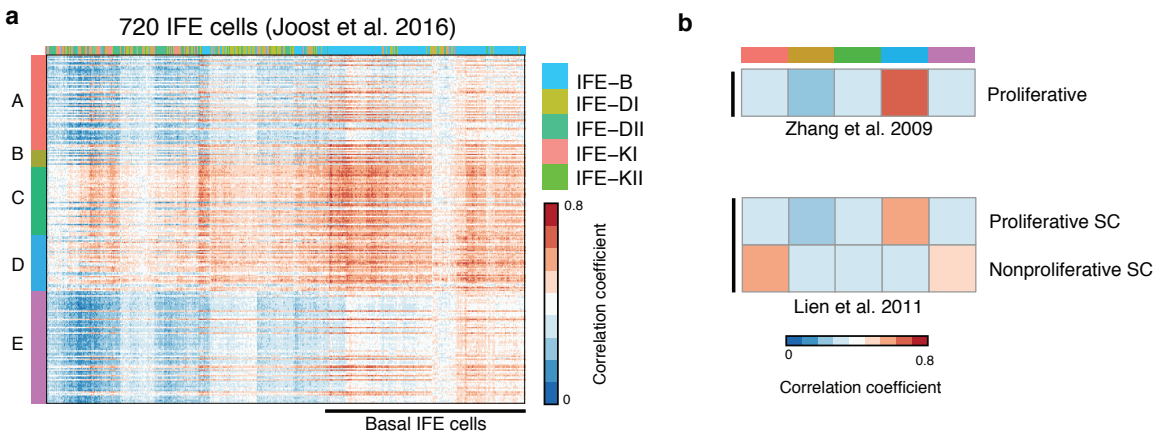


Figure 3.6: PCorrelating *In vitro* subpopulations with public bulk gene expression data.
(a) Heatmap showing the similarity between epidermal cells from this study and IFE cells from Joost et al. Similarity is measured by Pearson's correlation coefficient. Cells from this study are coloured by cluster along the vertical axis. Cells from Joost et al. are coloured by differentiation status along the horizontal axis. Joost et al. IFE cell legend is shown in order of differentiation status. B basal IFE cells, DI/DII differentiated suprabasal IFE cells, KI/KII keratinized IFE cells. (b) Heatmap showing similarity between cluster average transcriptomes from this study and proliferative IFE cells from Zhang et al. and activated vs quiescent IFE cells from Lien et al.

To determine the putative biological function of each cell state we correlated the single-cell signatures with a comprehensive set of IFE subpopulation expression profiles identified by Joost and colleagues. We further integrated two other bulk gene-expression studies which identified signatures for proliferative and nonproliferative epidermal stem cells (Zhang et al., 2009; Lien et al., 2011). From comparison with the Joost IFE subpopulations, all of our single cells correlate strongly with basal IFE stem cells, as expected since large (> 20μm) terminally differentiated cells were excluded from the analysis (Figure 3.6a). States A and to a lesser extent State E showed high correlation with an isolated subpopulation of non-proliferative self-renewing epidermal cells characterised by Lien and colleagues (Figure 3.6b).

We concluded that states A and E both exhibit a self-renewing cell gene expression signature but differ in proliferative state. States B, C and D formed a branch of the cell

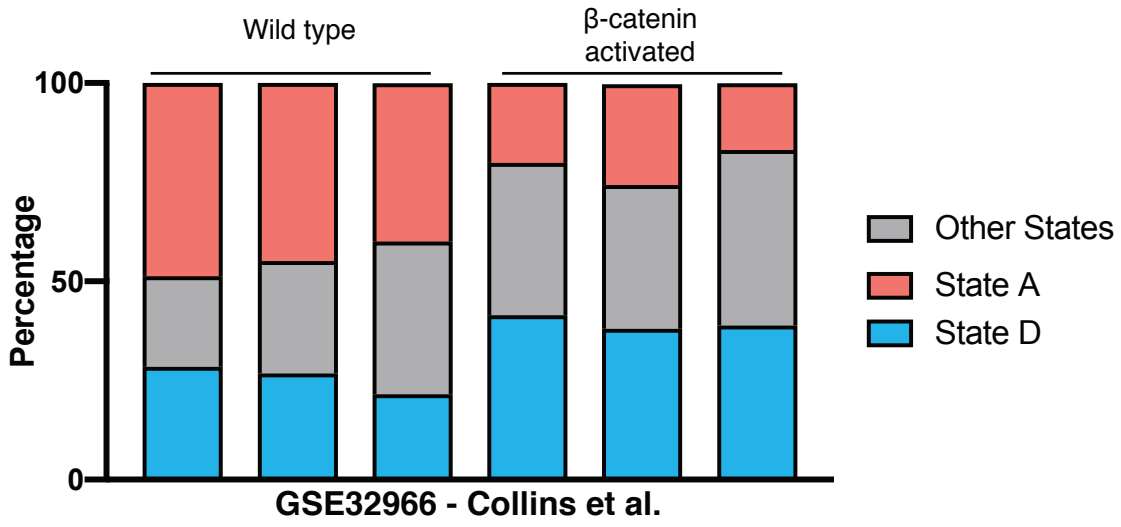


Figure 3.7: Deconvolving subpopulation mixture in Collins et al.
Stacked barplot showing estimated proportions of state A and state D cells in WT and beta-catenin activated epidermis samples from Collins et al. (GSE32966)

trajectory representing early commitment to differentiation, characterized by expression of proliferation associated genes from Roshan et al. (2016) (e.g. MRPL33, YY1) and correlated strongly with the expression profile of proliferative keratinocytes from Zhang et al. (2009). This branch of the state tree shows expression of early differentiation markers such as MXD1, Dsc2, Dsg3 (Salehi-Tabar et al., 2012) and highest expression of S100 early differentiation-associated genes. Figure 3.3 summarises the state classification of cells as determined by our cluster and DDRTree analysis highlighting the relationship between our three branches composed of five identified keratinocyte cell states.

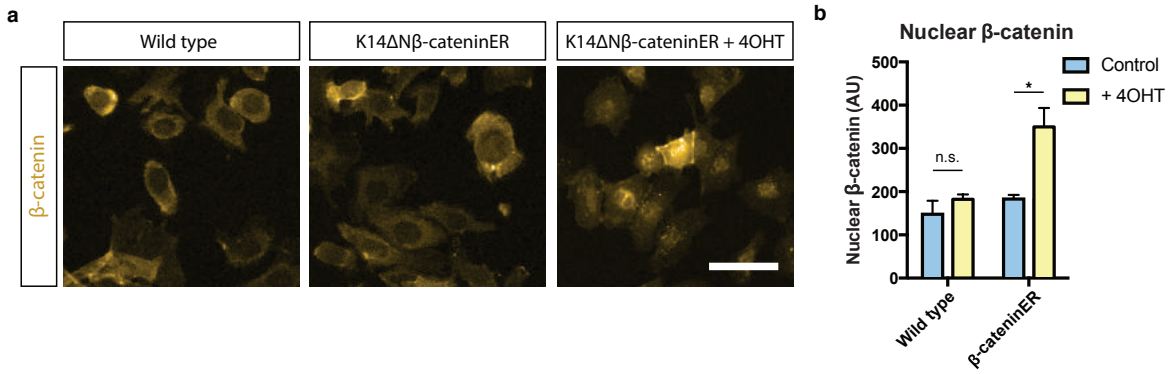


Figure 3.8: Induction of canonical Wnt signalling in a subpopulation of cells.
(a) Immunofluorescence showing cytoplasmic and nuclear beta-catenin in WT and K14ΔNβ-cateninER keratinocytes after induction of canonical Wnt signalling using 4OHT. Scale bar, 20 μm. Images are duplicated from Chapter 1, Figure 2.1. (b) Quantification of mean nuclear beta-catenin fluorescence intensity (n = 3 independent cultures).

3.2.2 Inducible Wnt signalling

In a previous study we generated gene-expression profiles from wild type and beta-catenin activated adult mouse epidermis (Collins et al., 2011). We reanalysed these data to estimate the relative proportion of cells in each of the cell states identified *in vitro* (Figure 3.3). We utilised CIBERSORT, a method for characterising the composition of tissue expression profiles resulting from mixtures of cells (Newman et al., 2015). Our reanalysis indicated that epidermal beta-catenin signalling results in a depletion of cells in State A and increases the abundance of cells in State D (Figure 3.7). This is consistent with the *in vivo* observation that intrinsic activation of epidermal beta-catenin results in proliferation and expansion of the stem cell compartment (Lo Celso et al., 2004).

Next to investigate whether epidermal cell states were altered by NCA Wnt signalling we examined the treated sample, wild type keratinocytes co-cultured with

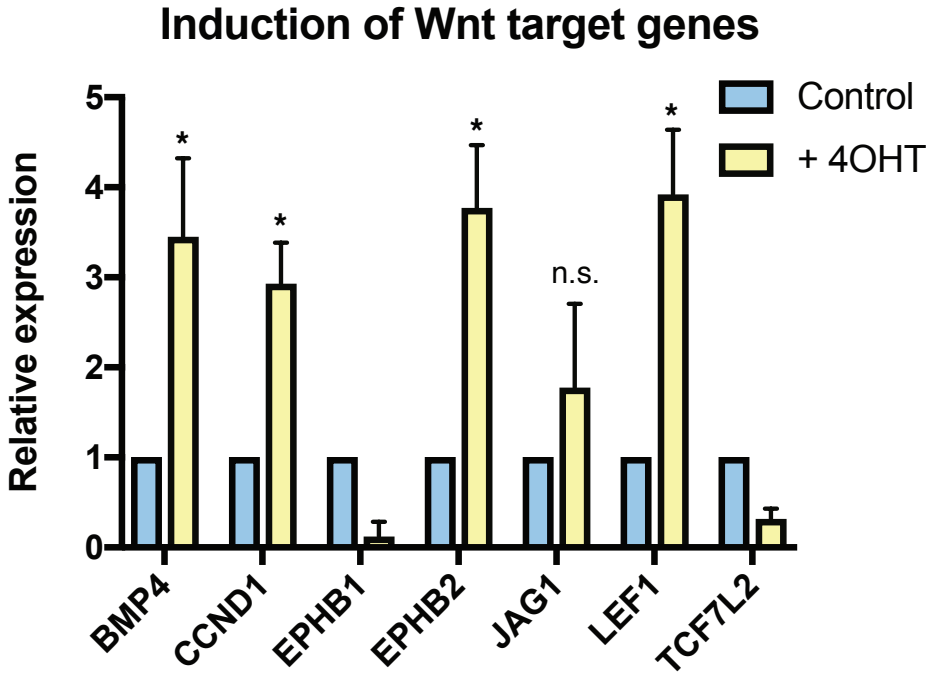


Figure 3.9: Activation of canonical Wnt target genes.
Seven canonical Wnt target genes in K14ΔNβ-cateninER cells upon induction with 4OHT quantified by quantitative reverse transcription polymerase chain reaction (n = 4). *P < 0.05, n.s. not significant. All data shown as mean ± SD.

K14ΔNβ-cateninER keratinocytes in the presence 4OHT. We confirmed that K14ΔNβ-cateninER cells intrinsically activated canonical Wnt signalling in response to 4OHT by detecting beta-cateninER translocation into the nucleus (Figure 3.8a and b). We also validated upregulation of canonical downstream target genes such as Bmp4, Cyclin-D1 and Lef1 in K14ΔNβ-cateninER keratinocytes using qRT-PCR (Figure 3.9).

3.2.3 Reconstruction of NCA Wnt induced state transition

Having identified several different states of wild type keratinocytes and validated the intrinsic effects of beta-catenin activation, we used single-cell RNA-seq to decon-

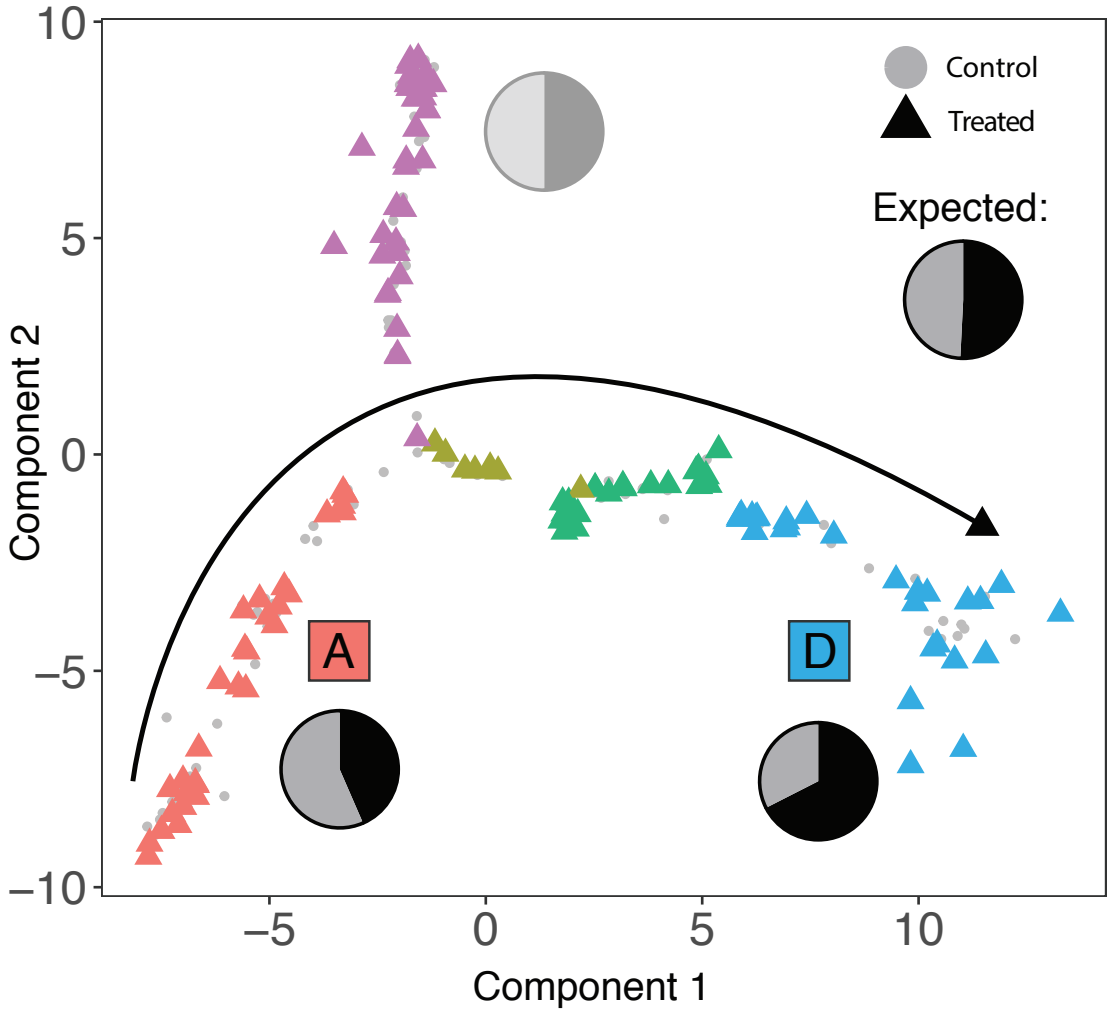


Figure 3.10: Neighbouring Wnt activation alters subpopulation proportions.
Keratinocytes exposed to Wnt signalling neighbours projected onto the WT epidermal cell state map. WT cells shown with grey circles, exposed cells shown coloured by cell cluster as triangles. Pie charts show relative proportions of WT vs signalling-exposed cells in each cell state cluster. States A and D are labelled as these states contain significantly different proportions of exposed vs unexposed cells.

volve the effects of NCA Wnt signalling. Single-cell transcriptomes from wild type keratinocytes co-cultured with 4OHT-activated K14ΔNβ-cateninER cells were compared with those of wild type cells co-cultured with uninduced K14ΔNβ-cateninER cells and mapped onto the same dimensionally reduced space (Figure 3.10). To exclude the possibility of transcriptional changes resulting from 4OHT treatment alone we screened bulk differential gene expression between the two cohorts of wild type cells. We found no evidence of estrogen receptor target genes among differentially expressed genes (Bourdeau et al., 2004).

We observed the same five distinct transcriptional states in wild type cells with or without NCA Wnt signalling (Figure 3.10), confirmed by independently analysing the treated cell population to reveal five equivalent subpopulations. However, exposure to non-cell autonomous Wnt signalling markedly changed the state distribution of keratinocytes (Fisher's exact test, $p<0.05$). Pie charts in Figure 3.10 show the observed ratio of control and signalling-exposed cells. States A and D significantly deviated from the expected ratio (binomial test, $p<0.05$). After exposure to NCA signalling there was a depletion of cells in the self-renewing, non-proliferative State A and a higher than expected proportion of cells in State D, representing a transition towards a proliferative and more differentiated transcriptional state (Figure 3.10).

Taking the states with altered cell proportions and the transition states in between (States A, B, C and D), we reconstructed the state transition induced by neighbouring Wnt+ keratinocytes using the Monocle pseudotime method (Trapnell et al., 2014). Wild type and exposed cells were ordered from State A to State D to reconstruct the

temporal order of gene expression changes for cells undergoing this transition, referred to as the pseudotransition. Figure 3.11a shows the proportion of control and Wnt signalling exposed cells along the reconstructed temporal transition from State A to State D; from this distribution it is clear that NCA Wnt exposed cells bias towards State D.

We next sought to understand why State A cells were uniquely depleted after neighbour Wnt signalling. Previous studies have shown that cell responses to extrinsic signalling are affected by intracellular and intercellular transcriptional noise (Kolodziejczyk et al., 2015b; Guo et al., 2016; Shalek et al., 2014b). We thus hypothesised that the response to NCA Wnt signalling involves changes in both the dynamic range of transcriptional variation (intracellular variation) and state-specific gene expression (intercellular variation).

3.2.4 NCA Wnt signalling reduces heterogeneity in protein synthesis-associated transcripts

We first examined whether there was a difference in intracellular transcriptomic heterogeneity between the three altered states and whether changes occurred along the pseudotransition. The resulting ordering of cells from State A to State D was used to examine the transcriptome coefficient of variation (TCOV) per cell (Figure 3.11b). Here TCOV is an intracellular measure of the spread of transcript abundance accounting for mean abundance. Notably, TCOV decreased over the state transition and was

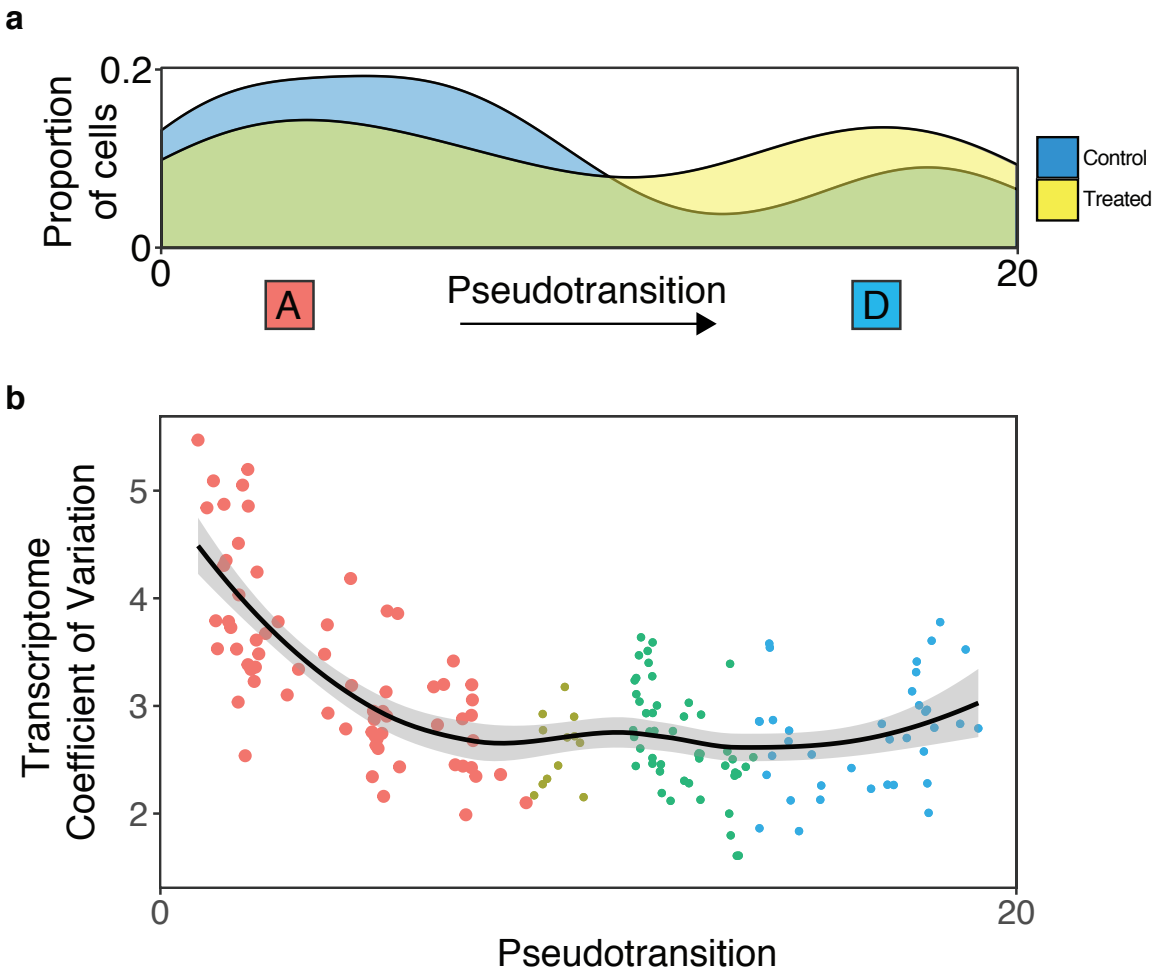


Figure 3.11: State A to D cell state transition and transcriptome coefficient of variation.
(a) Density of control vs NCA Wnt exposed cells along the reconstructed pseudotransition from states A to D. (b) Transcriptome coefficient of variation per-cell (TCOV), coloured by cell state and point size represents number of genes detected as expressed (TPM > 1) for that cell. Line depicts a loess-curve fit for the pseudotransition–TCOV relationship.

significantly higher in State A than States B, C and D (Kolmogorov-Smirnov test; p-value < 0.05). This reduction in dynamic range of gene expression is consistent with previous studies that have shown that progenitor cells have a higher rate of stochastic multilineage gene expression which reduces upon cell-fate commitment (Hu et al., 1997; Velten et al., 2017).

Next we contrasted the heterogeneity of genes that do not change in expression level between the transition states (Figure 3.12a-c). Figure 3.12c displays the relationship for the log-ratio of intercellular gene expression variation and expression level between the two extremes of the pseudotransition, States A and D, with the top 10 differentially dispersed genes labelled. Of interest are genes which change in expression heterogeneity from State A to State D while remaining at constant expression levels. Notably, Baz2a, Sox2, Col7a1 and Calcrl were amongst the genes with reduced COV in State A without significant differential expression (Supplementary Table S2). Baz2a has been previously established as part of the nucleolar remodelling complex that is important for establishing epigenetic silencing and transcriptional repression of rRNA genes (Gu et al., 2015; Santoro et al., 2002). Sox2 is an adult stem cell factor shown to be expressed in multiple epithelia (Arnold et al., 2011). Sox2 has been previously reported to be expressed in hair follicles but absent from the interfollicular epidermis (Driskell et al., 2009). Similarly Col7a1 and Calcrl are significantly upregulated in hair follicle bulge stem cells (Blanpain et al., 2004).

Figure 3.13 shows the expression variability for a selection of statistically significant (q < 0.05) differentially heterogeneous genes. We found multiple patterns of in-

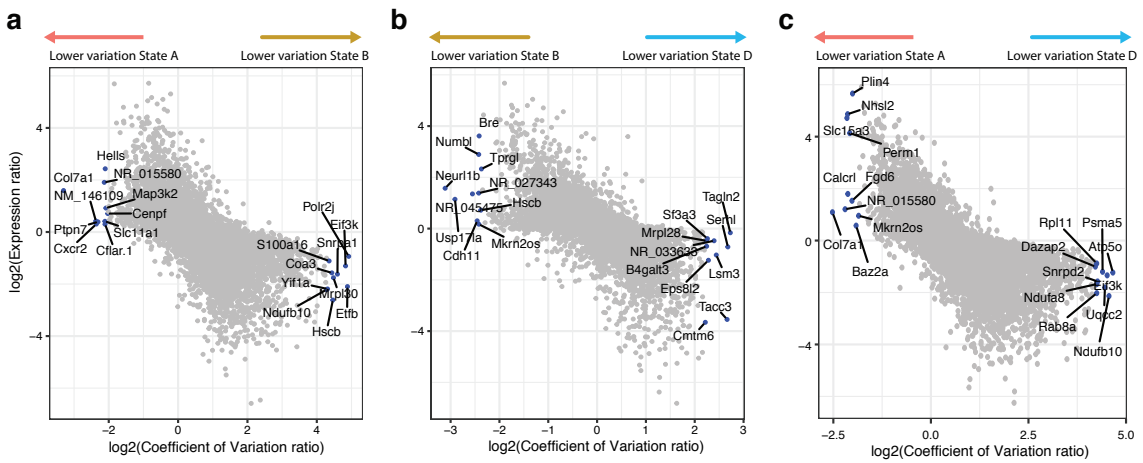


Figure 3.12: Relationship between expression and expression variability.
(a-c) Scatterplots showing the log-ratio of coefficient of variation versus the log-ratio of gene expression between pairs of pseudotransition states.

tercellular heterogeneity including genes which differed in variability across branches states and genes which were selectively more variable in a single state. Peroxiredoxin-1 (Prdx1) is one such gene which showed strongly variable expression in State A in comparison to the remaining states despite no statistically significant difference in median expression level. Overall, we observed many more heterogeneously expressed genes in State A (281 genes) than either States B, C or D. The contrast in number of differentially dispersed genes is demonstrated using the symmetric expression scale in Figure 3.14a. In comparison, we observed only 19 significantly differentially heterogeneous genes with lower heterogeneity in State D. A striking number of these are known regulators of stem cell identity such as Cdh11, Cav2 and Apc (Kim et al., 2014; Tan et al., 2013; Chan et al., 1999). They are typically upregulated in basal stem cells; however little is known about how their heterogeneity affects cell fate. Our analysis of intercellular variation suggests that in slow-cycling stem cells (State A) with low RNA and protein metabolism (Blanco et al., 2016), transcriptional heterogeneity is lowest

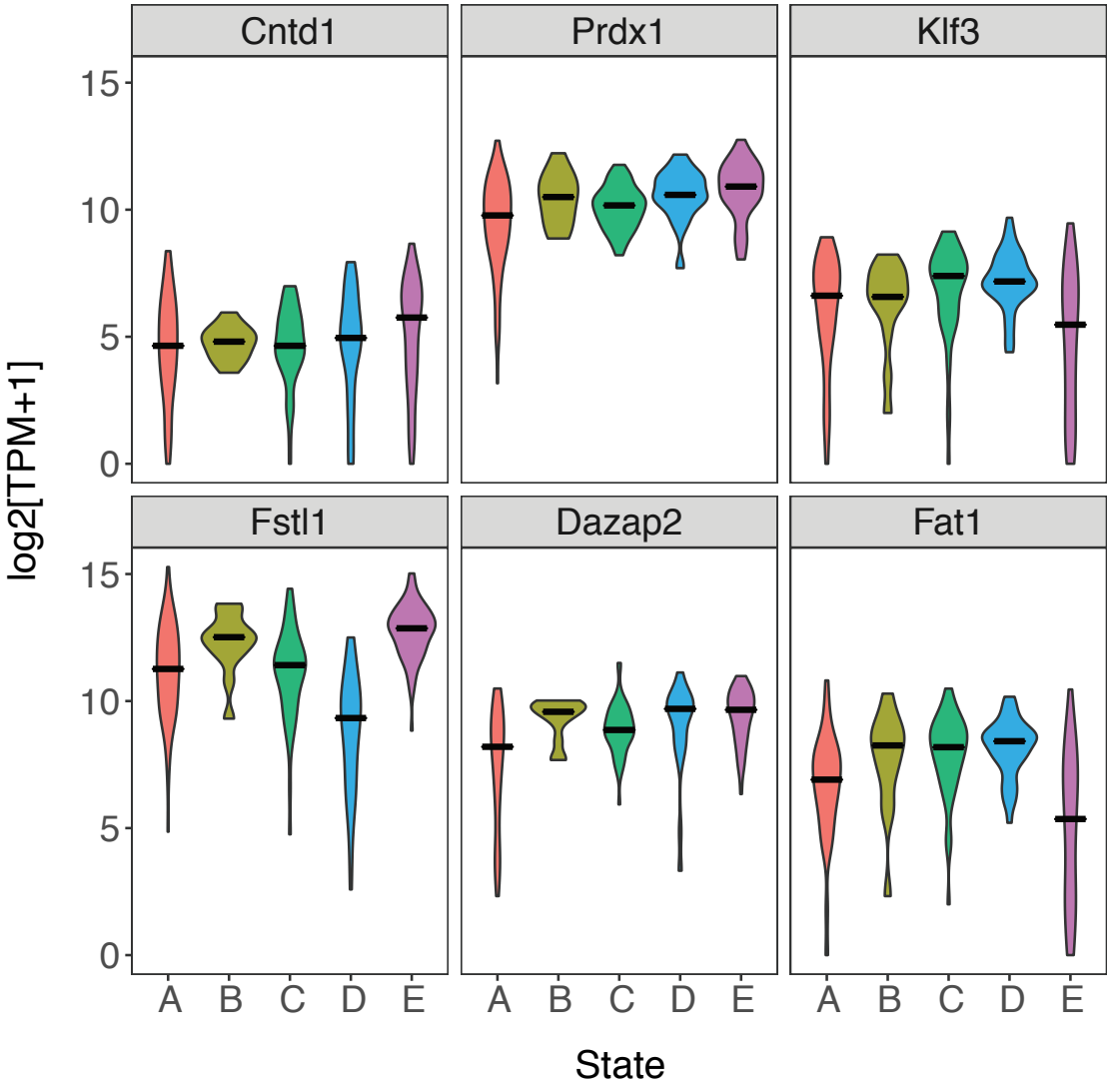


Figure 3.13: Gene markers of differential heterogeneity.
Violin plots depicting expression of cluster-specific genes showing differential dispersion between states.

for stem-cell marker genes, emphasising the importance of transcriptional noise in addition to transcriptional amplitude. These observations are consistent with our hypothesis that State A cells are responsive to NCA Wnt signalling due to greater transcriptional variability. Exposure to this coordinated extracellular stimulus reduces transcriptional heterogeneity for these cells and biases their fate towards State D.

To determine genes essential for a cell to be receptive to neighbour Wnt activation we analysed the fold-change in heterogeneity between States A and D, comprising the majority of genes with differential heterogeneity. We found strong enrichment for translation and ribosome related genes, indicating a role for protein synthesis ($p < 1e\text{-}6$, Figure 3.14b). We hypothesise that cells in State A exhibit a multilineage primed transcriptional programme with stochastic expression of metabolism associated genes. Upon fate commitment, cells in the IFE steadily increase their translational rate in a proliferation independent manner (Blanco et al., 2016). Hence translation associated genes are subject to greater transcriptional regulation post-commitment independent of transcription level.

These data and our single-cell analysis identified an NCA Wnt-receptive subpopulation, State A, with greater dynamic range in gene expression (TCOV) and greater variation in the abundance of protein synthesis associated transcripts. Introduction of the NCA Wnt stimulus reduces variability in both aspects.

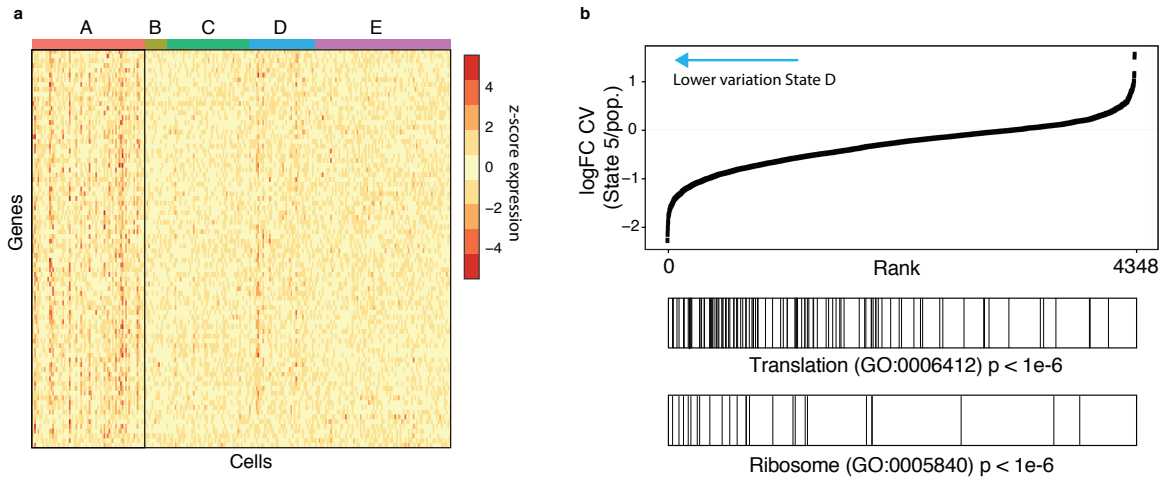


Figure 3.14: Protein-synthesis associated genes decrease in heterogeneity from state A to D.
(a) Heatmap with symmetric gene expression scale (z-score normalised log2(TPM + 1)) showing reduction in gene expression heterogeneity of 281 genes between state A and the remaining transition states. (b) Gene rank enrichment analysis of log-fold change in gene coefficient of variation (CV) between cells in state D and the remainder of the cell population. Genes with lower variation in state 5 are enriched for translation and ribosome-associated gene ontology (GO) annotations (p value < 1e-6)

3.2.5 Transcription factors driving cell fate change

To understand drivers of the observed differential heterogeneity we reconstructed transcriptional changes over time along the state trajectory. Expression of each gene was modeled as a nonlinear function of pseudotransition time (Trapnell et al., 2014). We found 632 genes that were dynamically regulated over the state transition (False discovery rate < 5%; Figure 3.15). Using hierarchical clustering we grouped these genes into four patterns of dynamic expression. Group I genes, most highly expressed in State A, were enriched for methylation associated genes and histone modifiers such as Setd3 and Kdm7a. These genes represent a pre-transition transcriptional profile of State A cells without exposure to signalling beta-catenin induced cells. Group III genes show highest expression in state C, an intermediate transition state, with enrichment for desmosome genes such as Dsc2, Dsc3, Dsg2 and Dsp, which are most highly

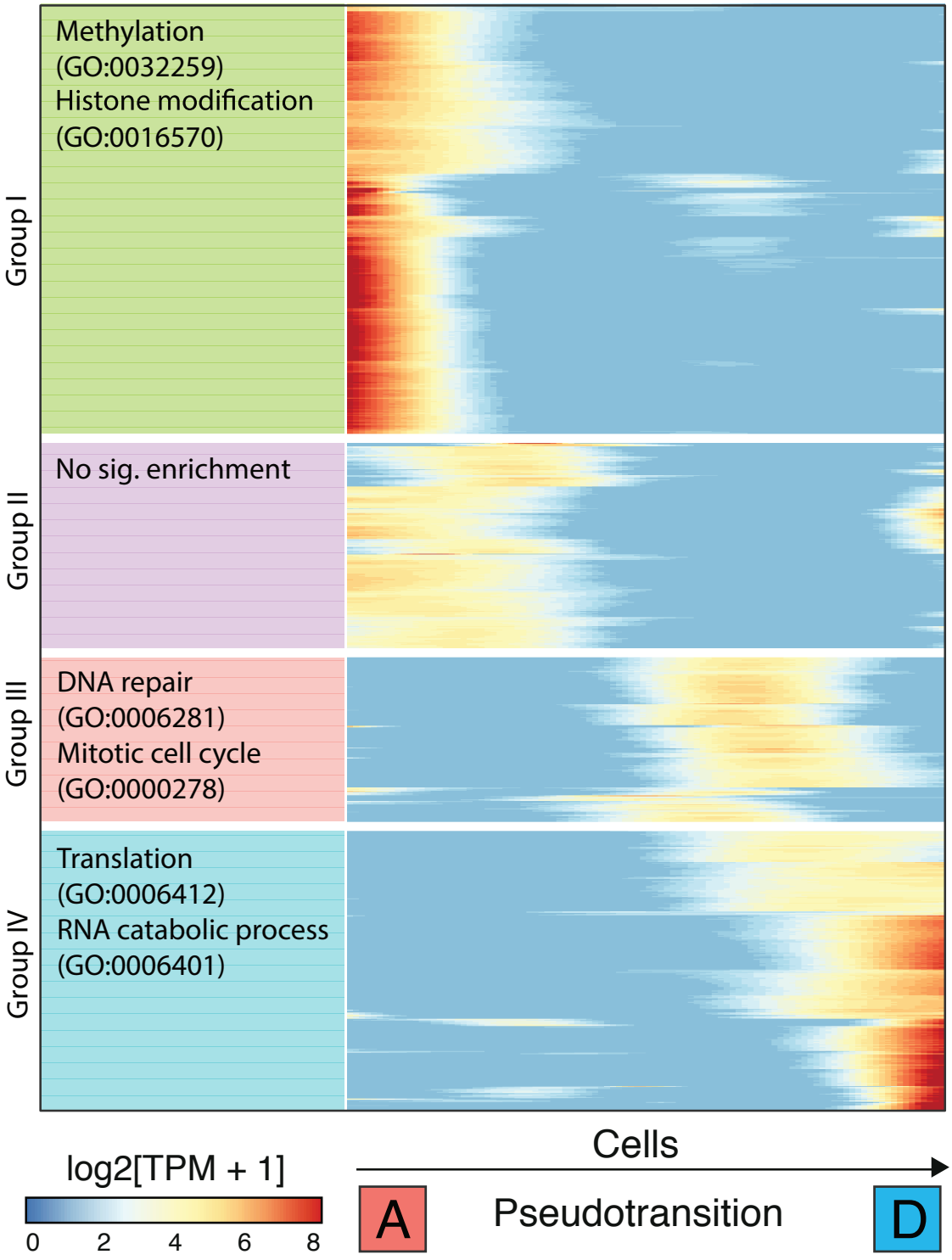


Figure 3.15: Reconstructing transcriptional changes in transition from state A to D.
Heatmap showing smoothed expression of pseudotransition-dependent genes (n = 632) ordered by hierarchical clustering and maximum expression. Top two enriched GO terms shown on left (all significant at q < 0.05). Genes (rows) are ordered by peak expression from state A to state D

expressed in the suprabasal layers of murine epidermis indicative of early commitment to differentiation (Joost et al., 2016). Group IV genes, predominantly expressed in State D, were enriched for protein synthesis associated genes and entry into mitosis, respectively. Notably group III includes the transcription factor Klf5, a regulator of proliferation in intestinal epithelial cells (Chanchevalap et al., 2004), and E2f1,which leads to epidermal hyperplasia when overexpressed in mice (Pierce et al., 1998).

To gain insight into the regulation of the dynamically expressed genes induced by Wnt+ cells we performed a transcription factor motif analysis (Figure 3.16a-c). We calculated enrichment of transcription factor binding sites from the ChEA database, removing transcription factors (TFs) which were not expressed in any of the seven cell states ($\log[TPM+1] > 1$). By analysing the promoters of the 632 dynamically expressed genes we identified 47 transcription factors putatively regulating the state transition. TFs were separated into three groups according to the directionality of gene expression from State A to D: positive, negative and neutral. We noted that the activities of some identified TFs such as Smad3 and Smad4 are only partially dependent on expression level. Hence we did not rule out TFs on the basis of expression.

From this analysis we predicted Smad3, Smad4, Kdm5b, E2f1 and E2f4 as previously unknown key regulators of the state transition, with Bcl3 as a likely regulator of the specific transition between State A and D. Known regulators of keratinocyte cell fate are shown in blue (Figure 3.16c). Of note are Gata6 and Foxm1, two TFs upregulated in State D and previously shown to mark cells with multi-lineage differentiation potential and increased proliferative capacity respectively (Donati et al., 2017;

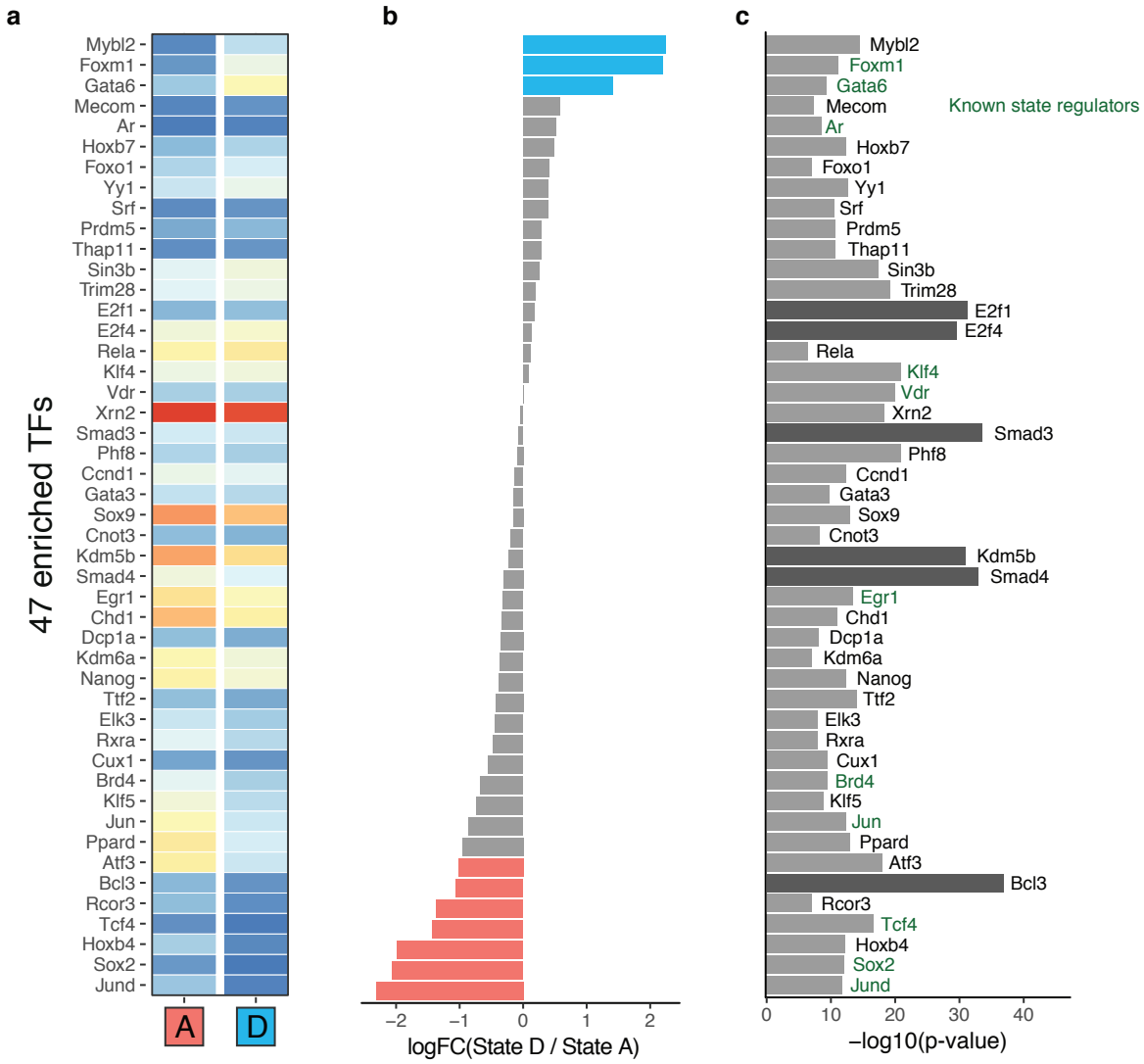


Figure 3.16: Transition from state A to D is regulated by 47 TFs.
(a) Heatmap showing expression of 47 TFs enriched as regulators of pseudotransition-dependent genes (b). Log-fold change in expression of TFs between state D and state A to show strongly directional TFs (c) and enrichment level of TFs.

Gemenetzidis et al., 2010; Molinuevo et al., 2017). We hypothesised that epidermal cells could be stratified based on the nuclear abundance of our identified novel state-regulating TFs, specifically Smad4 and Bcl3. Furthermore, our analysis on State D gene expression markers and transcriptome correlation indicated that this state shows a higher proliferation rate relative to states A and B. To investigate further, we calculated a cell proliferation index consisting of normalised expression of S-phase markers (Figure 3.17a). This index demonstrated that states C and D comprise proliferative cells with few cells in States A, B or E actively proliferating.

To confirm our findings, we used an EdU incorporation assay to distinguish proliferating cells and analysed whether keratinocytes positive for our predicted driver TFs (measured by nuclear intensity) were more likely to be proliferative. At the population level there was no significant difference in proliferation when epidermal cells were co-cultured with 4OHT induced K14ΔNβ-cateninER cells (Figure 3.17b). However, when cells were discriminated by nuclear intensity for Bcl3 or Smad4 we observed a significant difference between wild type cells exposed to NCA Wnt signalling and wild type or induced K14ΔNβ-cateninER cells alone (Figure 3.17c and 3.17d). On average 18% of $Bcl3^{+}$ cells were positive for EdU uptake in wild type or uninduced K14ΔNβ-cateninER cells; however, when wild type cells were co-cultured with induced K14ΔNβ-cateninER cells the EdU^{+} fraction rose to 34%. The proportion of proliferative $Smad4^{+}$ cells increased from 10-18% to 33% EdU^{+}.

Taken together these novel results indicate that Bcl3 and Smad4 are specific markers of the epidermal state transition and mark cells moving along the trajectory be-

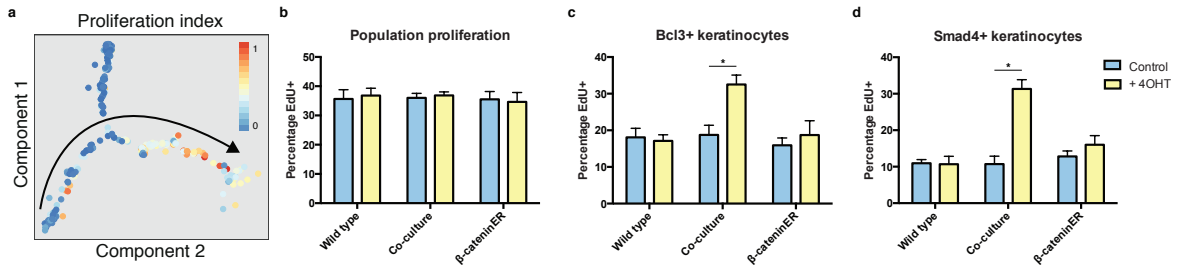


Figure 3.17: State D is more proliferative and Smad4+/Bcl3+.
(a) Normalised proliferation index projected onto the cell state map. Arrow denotes the direction of pseudo-transition. (b–d) Quantification of population proliferation by EdU assay in WT, K14ΔNβ-cateninER and co-cultured keratinocytes with and without 4OHT treatment (b). After stratification by Bcl3 nuclear abundance, cells in co-culture with activated K14ΔNβ-cateninER cells show a relative higher proliferation rate (c). Similarly, stratification by nuclear Smad4 shows higher proliferation in the treated co-culture condition (h) (n = 3 independent cultures). *P < 0.05. All data shown as mean ± SD

tween State A and State D (Figure 3.10) during the first 24 hours of exposure to a NCA Wnt signal.

3.2.6 NCA Wnt induced state transition is contact dependent

From our total population of 129 cells exposed to NCA Wnt signalling, cells in State E appeared to be unaffected. Our data suggest that State A comprises cells in a "responder" state permissive to NCA Wnt signalling due to the presence of key TFs and a more stochastic gene expression programme. We sought to address whether the reduction in ribosome-related gene expression heterogeneity and the induced expression of transition TFs are contact or distance dependent. To answer this question we labelled co-cultures of wild type and K14ΔNβ-cateninER cells with a cell reporter of protein synthesis (See methods and Figure 3.21).

We measured global protein translation by assaying incorporation of O-propargyl-

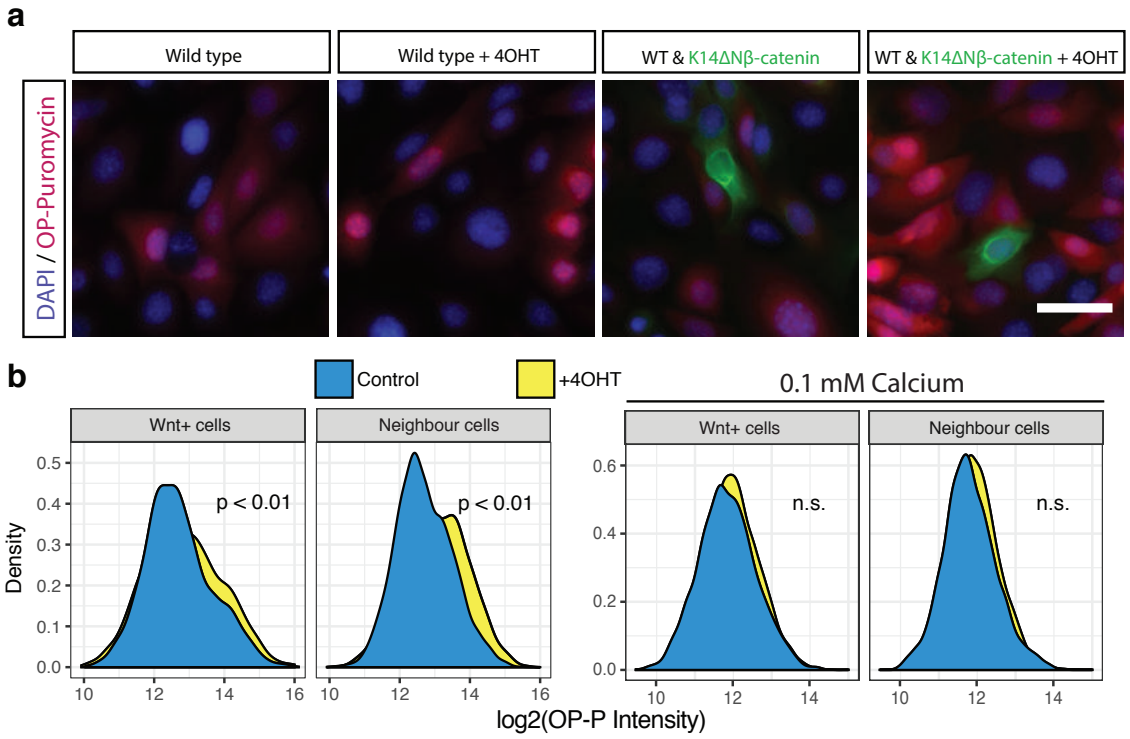


Figure 3.18: NCA Wnt activation increases translation in neighbouring cells.
(a) Appearance of typical WT cells and K14ΔNβ-cateninER co-cultures stained for OPP. (b) Left: Quantification of OPP in K14ΔNβ-cateninER cells and their neighbours in control and 4OHT treated conditions. Right: Similar quantification for cells co-cultured in low calcium medium (0.1 mM) to inhibit cell–cell contact (n = 3 independent cultures, n = 3 technical replicates, all pooled).

puromycin assay (OPP) and compared wild type cells in contact with induced K14ΔNβ-cateninER cells to untreated control cells (Figure 3.18a). We observed that wild type cells showed higher translational activity when in contact with a Wnt+ cell. To confirm this we analysed the neighbours of over 10,000 Wnt+ cells and compared the OPP fluorescence intensity distributions (Figure 3.18b). We found a small but statistically significant increase in translation rate for both Wnt+ cells and neighbour cells in the 4OHT treated condition. This suggested a contact-dependent mechanism for control of protein synthesis downstream of NCA Wnt signalling.

To confirm contact dependence and to rule out local diffusion of soluble factors we repeated the assay in low calcium conditions. Keratinocytes cultured in low calcium medium do not form adherens or desmosomal cell contacts (Hennings and Holbrook, 1983; O'Keefe et al., 1987). Strikingly we observed no NCA Wnt effect under these conditions (Figure 3.18b, right). Similarly we observed no increase in nuclear abundance of Smad4, our predicted TF downstream of NCA Wnt signalling, in Wnt+ cells, as predicted. However, in neighbouring cells there was a significant increase in nuclear Smad4 intensity, which is abrogated in low calcium conditions (Figure 3.19a and 3.19b).

Taken together these data suggest that the Smad4-mediated cell state transition is downstream of non-cell autonomous Wnt signalling. Furthermore, the induction of this transition is contact dependent and does not occur under conditions where desmosomal adherens junctions are inhibited.

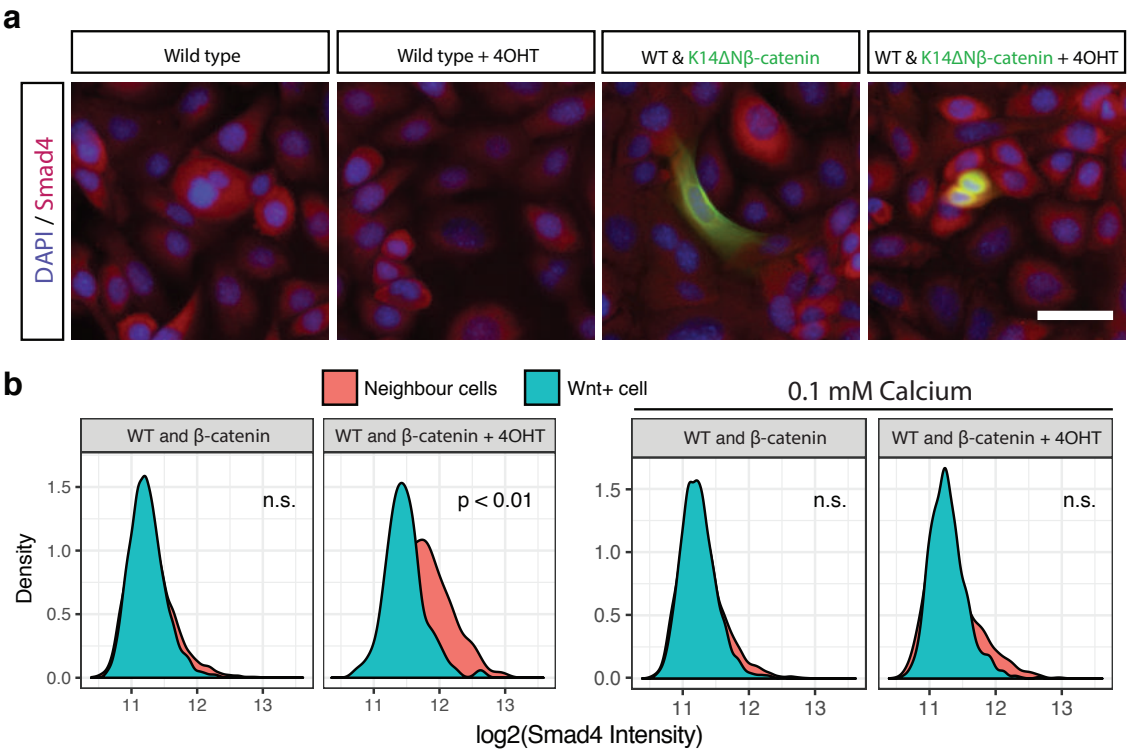


Figure 3.19: NCA Wnt activation increases nuclear Smad4 in neighbouring cells.
(a) Appearance of typical WT cells and K14ΔNβ-cateninER co-cultures stained for Smad4. (b) Left: Quantification of Smad4 in K14ΔNβ-cateninER cells and their neighbours in control and 4OHT treated conditions. Right: Similar quantification for cells co-cultured in low calcium medium (0.1 mM) to inhibit cell–cell contact (n = 3 independent cultures, n = 3 technical replicates, all pooled). All scale bars, 20 μm, n.s. not significant

3.3 Conclusions

Heterogeneity in the self-renewal and proliferative capabilities of keratinocytes has long been recognised. Previous analysis of clones and subclones of cultured human epidermal cells has demonstrated that there are at least three subpopulations, 'holoclones', 'meroclones' and 'paraclones' with descending self-renewal potential (Jones and Watt, 1993; Barrandon and Green, 1987). More recently, Roshan et al. have shown the existence of two in vitro states with differing proliferation rates and single cell transcriptomics have identified two distinct subpopulations of human keratinocytes in culture (Roshan et al., 2016). In this study, we have dissected molecular heterogeneity of epidermal cells at greater resolution and extended previous research by exploring the response of keratinocytes to neighbouring cells in which beta-catenin is activated. We have identified five distinct transcriptomic states and characterised their biological relevance in order to create a state map of keratinocytes in vitro.

Using the cell state map and inducible activation, we have shown that Wnt/beta-catenin signalling acts to perturb cell fate by co-opting neighbours to become biased towards a pre-existing proliferative fate (Figure 3.20). It is important to note that we found no evidence for a de novo cell state as a result of non-cell autonomous signalling. This highlights the relevance of transient Wnt/beta-catenin signalling to cell state and is consistent with a model of stochastic epidermal commitment where extrinsic cues alter the likelihood of a cell switching state (Rompolas et al., 2016). The observed difference in transcriptome variability between states A and D reflects a difference in

cell state stability. Only a modest increase in translational activity is observed in state D or neighbouring cells; however, there is a marked reduction in the variability of translation-associated genes, highlighting the importance of transcriptional noise as well as transcriptional volume for determination of cell state.

Combined transcriptomic and positional single cell analyses allowed us to resolve spatial and temporal effects. As a result of this, we identified a collection of TFs, many of which were not previously implicated in epidermal cell state. One example is Bcl3, which is expressed in murine and human basal IFE; however, its role in epidermal cell fate is poorly understood (Joost et al., 2016; Uhlén et al., 2015). In addition, we identified Smad4 and utilised this as a marker of cell state transition. Smad4-beta-catenin cross-talk has been previously identified as essential for hair follicle maintenance (Owens et al., 2008; Qiao et al., 2006; Yang et al., 2009). Here, we show that beta-catenin signalling activation leads to Smad4 activation in a non-cell autonomous manner.

Our study does not address the extracellular effectors of NCA Wnt signalling. We previously identified a diverse array of secreted signalling molecules downstream of canonical Wnt signalling, including Bmp6, Dkk3, Wnt ligands, cytokines and ECM components (Donati et al., 2014). Our analysis of TF effectors and previous evidence of epidermal self-renewal via autocrine Wnt signalling suggests that cells neighbouring a beta-catenin+ cell are simultaneously committed to a state of lesser self-renewal and greater proliferative abilities. This is achieved via a combination of Bmp signalling (effected through Smad3/4; Figure 3.20) and neighbouring Wnt inhibition (Dkk3, Lim et

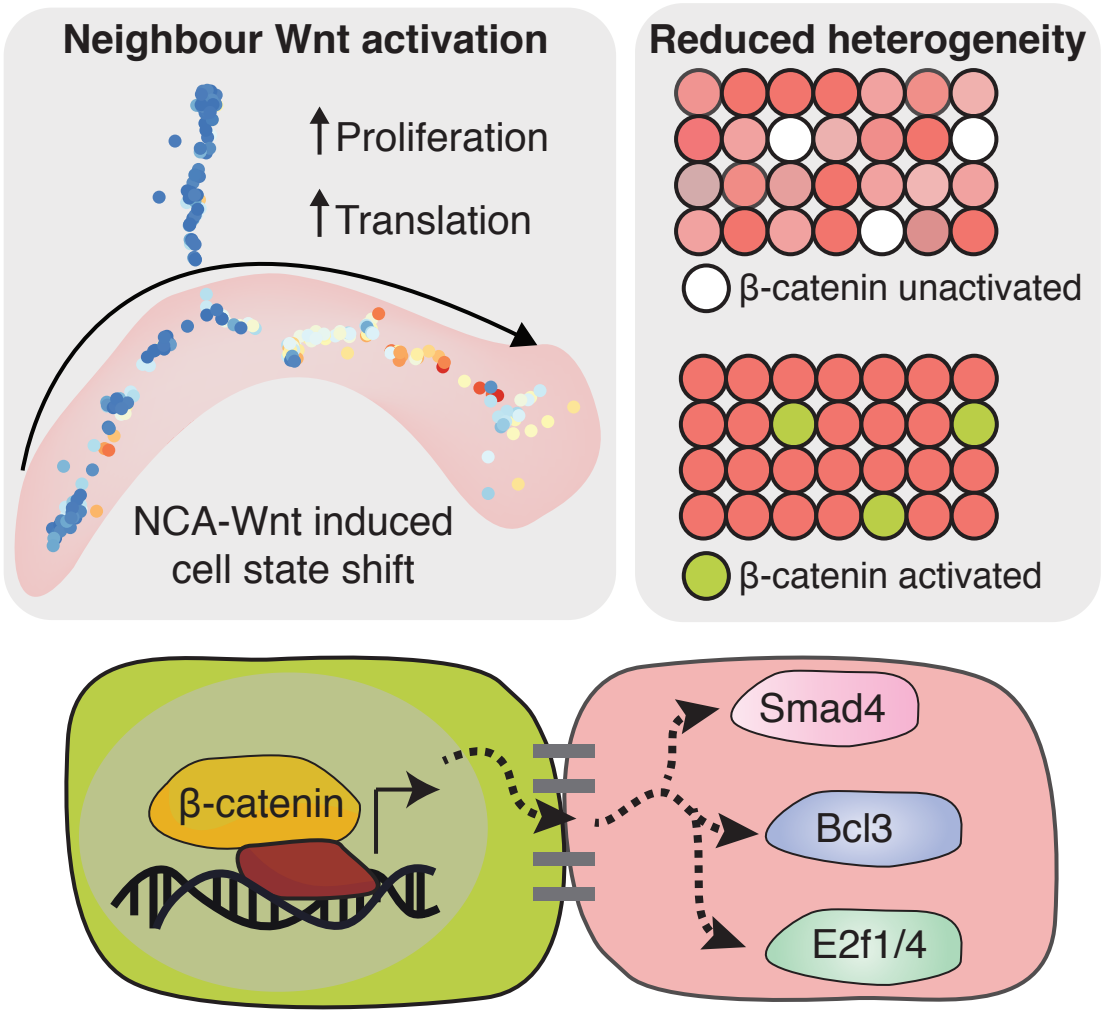


Figure 3.20: Effects of non-cell autonomous Wnt signalling.
Wnt/beta-catenin activation affects neighbouring cells. Epidermal cells in culture adopt one of five distinct transcriptomic states differing on the basis of proliferation and commitment to differentiation. Wnt/beta-catenin signalling acts as a non-cell autonomous signalling cue to activate a handful of TFs including Smad3/4, E2f1/4 and Bcl3 in neighbours. Concurrently, transcriptional heterogeneity is reduced as neighbour cells enter into a committed and proliferative transcriptional state

al.). Intriguingly these effects are contact-dependent, hinting at yet-unknown mechanisms of locally restricting these signalling molecules or major signalling contributions from other membrane-bound factors. The observed difference in translation rate and proliferation in neighbouring cells demonstrates asymmetric coupling of cell fates, an essential component of epidermal homeostasis to ensure a balance of cell fates and epidermal metabolism.

In conclusion, our data provide a framework for studying cell state in the interfollicular epidermis and extend our understanding of functional heterogeneity and NCA signalling. Using this knowledge, we demonstrate how Wnt/beta-catenin signalling, an orchestrator of regeneration, homeostasis and tumorigenesis in multiple tissues, influences neighbouring cell fate.

3.4 Methods

3.4.1 Cell biology

Cell isolation and culture

K14ΔNβ-cateninER transgenic mice were generated as previously described (Lo Celso et al., 2004). Keratinocytes were isolated and cultured from adult dorsal skin in FAD medium (one part Ham's F12, three parts Dulbecco's modified Eagle's medium, 1.8×10^{-4} M adenine), supplemented with 10% foetal calf serum (FCS) and a cocktail of 0.5

μg/ml hydrocortisone, 5 μg/ml insulin, 1×10^{-10} M cholera enterotoxin and 10 ng/ml epidermal growth factor (HICE cocktail) (Watt et al., 2006). For the co-culture scRNA-seq experiment, wild type and K14ΔNβ-cateninER keratinocytes were cultured on 12 well plates in a ratio of 9:1 for a total of 200,000 cells per well and allowed to attach for 24 hours. Subsequently, cells were treated with 4-OHT (200nM) or DMSO as a control. After 24 hours of treatment cells were trypsinised and resuspended as a single cell suspension.

3.4.2 Immunofluorescence, imaging and neighbour cell quantification

Immunofluorescence staining

The following antibodies were used: β-catenin (1:250, Sigma), Smad4 (1:250, Sigma), Bcl3 (1:250, Sigma). For EdU experiments (Molecular Probes; C10337), half of the cell culture medium was replaced with medium containing EdU for a final concentration of 10 μM EdU 30 minutes before fixation. Similarly, for OPP experiments (Molecular Probes; C10456) half of the cell culture medium was replaced 30 minutes before fixation with medium containing OPP for a final concentration of 20 μM OPP. Cultured cells were fixed with 4% PFA for 10 minutes followed by permeabilisation with 0.1% Triton X-100 for 10 minutes at room temperature. Cells were blocked for 1 hour at room temperature with 1% BSA in PBS. Primary antibody incubation was carried out for 90 minutes at room temperature. Samples were labelled with Alexa Fluor (488,

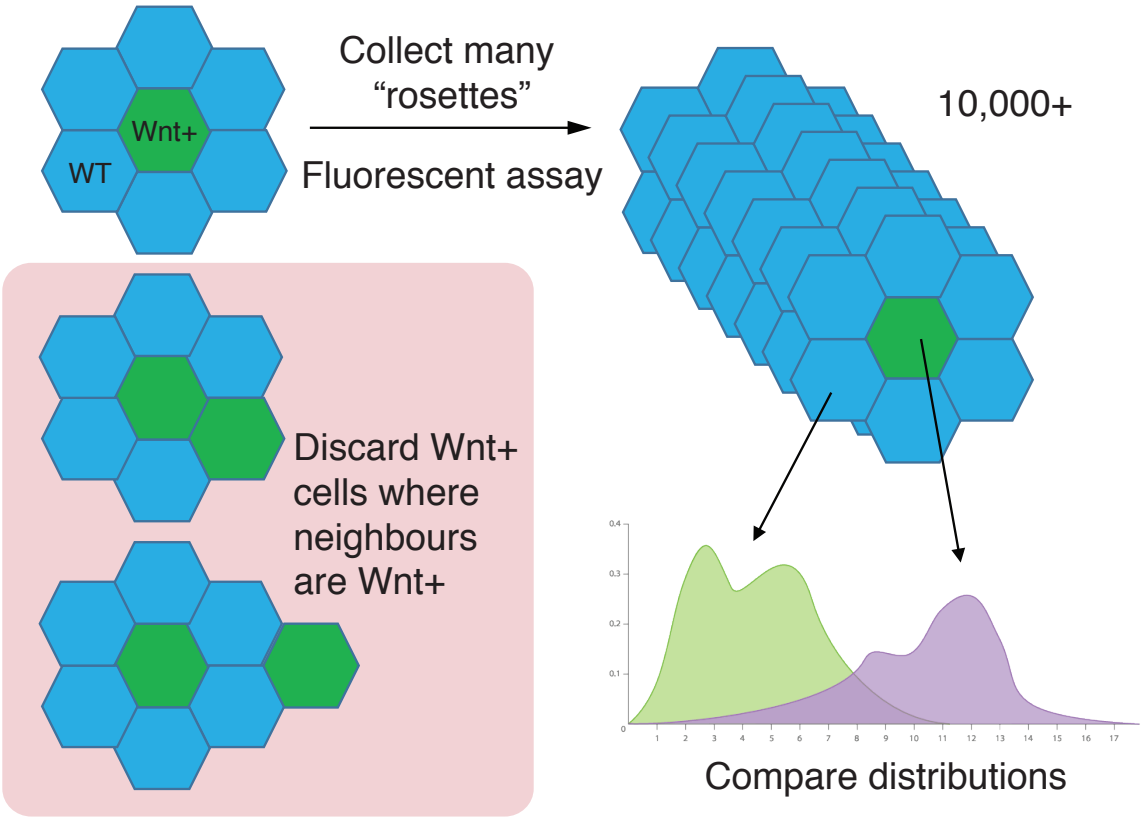


Figure 3.21: High-content single cell neighbour analysis overview.
Overview of method to analyse neighbouring cells using the Operetta high-content screening platform. Two cell types are co-cultured, here (90%) WT and (10%) Wnt^+ cells, each labelled with a fluorescent dye. A fluorescent assay is performed e.g. immunofluorescence for protein abundance or OPP assay for protein synthesis. Co-cultures are screened for ”rosettes” where a Wnt^+ is surrounded by WT cells. Tens of thousands of rosettes are collated and distribution of assay fluorescence is compared in Wnt^+ and neighbour cells.

555, 647)- conjugated secondary antibodies for 1 hour at room temperature. Cells were imaged within 24 hours using an Operetta or Operetta CLS High-content Imaging System (PerkinElmer). Single cell cytoplasmic and nuclear fluorescence intensities were quantified with Harmony software (PerkinElmer) and analysed in R.

Neighbour cell quantification (see Figure 3.21)

For neighbouring cell quantification K14ΔNβ-cateninER cells were labeled with CellTracker Green CMFDA dye (Molecular Probes) according to the manufacturer's instructions. Single cell fluorescence intensity data and positional information were analysed in R. For each K14ΔNβ-cateninER CellTracker+ cell the mean fluorescence intensity of neighbouring cells was calculated. Neighbouring cells were defined as the nearest cell within 20μm (nucleus-to-nucleus distance). The mean number of neighbours was 5.4, as expected from a hexagonal packing model below confluence with mean cell diameter of 8μm. K14ΔNβ-cateninER cells were excluded if more than two neighbouring cells were also CellTracker+ (See Figure 3.21). We collected over 10,000 hexagonally packed "rosettes" of Wnt activated (Wnt+) and wild type cells per condition. Fluorescence intensity distributions from biological and technical replicates were pooled and contrasted between conditions using the non-parametric Kolmogorov-Smirnov test.

3.4.3 Bulk gene expression analysis

Bulk RNA extraction and real-time qPCR

Total RNA was purified with the RNeasy mini kit (Qiagen) with on-column DNaseI digestion, according to the manufacturer's instructions. RNA was reverse transcribed with SuperScript III (Invitrogen). PCR reactions were performed with TaqMan Fast

Universal PCR Master Mix and Taqman probes purchased from Invitrogen.

Processing of reads and quality control

Reads were preprocessed using FastQC and Cutadapt (Martin, 2011). Sequences were aligned to the Mus Musculus genome (GRCm38) using Tophat (Kim et al., 2013) discarding multiply-mapped reads. Gene level counts were extracted using featureCounts (Liao et al., 2014). Transcript levels were quantified as transcripts per million (TPM). Genes with a TPM greater than 1 were considered as expressed. We filtered cells for analyses on the basis of number of aligned reads (> 200,000), percentage of ribosomal reads (< 2%) and number of genes expressed (> 2000). 254 cells were taken forward for analysis.

3.4.4 Single cell transcriptomics

Single cell capture, library preparation and RNA-sequencing

Single keratinocytes were captured on a medium-sized (10-17μm) microfluidic chip (C1, Fluidigm). Cells were assessed for viability (LIVE/DEAD assay, Life Technologies) and C1 capture sites were imaged by phase contrast to determine empty and doublet capture sites. Cells were loaded onto the chip at a concentration of 300 cells μL^{-1}. Doublet or non-viable cells were excluded from later analysis. Cell lysis, reverse transcription, and cDNA amplification were performed on the C1 Single-Cell Auto Prep

IFC, as per the manufacturer's instructions. For cDNA synthesis the SMART-Seq v4 Ultra Low Input RNA Kit (Clontech) was used. Single cell Illumina NGS libraries were constructed with Nextera XT DNA Sample Prep kit (Illumina). Sequencing was performed on Illumina HiSeq4000 (Illumina) using 100bp paired-end reads.

Identification of K14ΔNβ-cateninER cells

K14ΔNβ-cateninER cells were identified by aligning RNA-seq reads to the transgene locus using bowtie2 (Langmead and Salzberg, 2012). Subsequently, cell identity was confirmed using qRT-PCR with Fast SYBR Green Master Mix (ThermoFisher Scientific) using the primers ATGCTGCTGGCTGGCTATGGTCAG (forward) and ATAGATCATGGGCGGTTCAGC (reverse) spanning the beta-catenin estrogen-receptor junction.

Dimensionality reduction, cell state map and pseudotransition

We performed dimensionality reduction and constructed the principal graph representing transitions between all possible cell states using DDRTree from Monocle2 (Trapnell et al., 2014). All DDRTree dimensionality reduction was performed using default parameters and a final dimensionality of two. We initially performed this analysis on wild type cells to determine the unperturbed cell state map. Subsequently we applied the DDRTree algorithm on the Wnt+ cell exposed group to confirm that we independently achieve a similar cell state map. We used all 254 cells for the final

transcriptomic state map and differential expression to obtain cell state marker genes. Cell clusters obtained from Monocle were confirmed by a combination of dimensionality reduction of the cells using t-distributed stochastic neighbour embedding (tSNE) (van der Maaten and Hinton, 2008b) and cluster identification with DBSCAN (Ester et al., 1996). Differential gene expression analysis was performed using Monocle 2 and VGAM using a likelihood ratio test controlling for batch effects and number of aligned reads per cell. Genes were filtered for log-2-fold change > 0.5 and an adjusted p-value < 0.05. Expression profiles from this study were correlated with expression profiles from Joost et al. (single cell RNA-seq, GSE67602), Zhang et al. (bulk microarray, GSE16516) and Lien et al. (bulk microarray, GSE31028) using pearson correlation coefficient of all genes expressed greater than TPM>1 in more than 5 cells.

Heterogeneity analysis

Differential gene dispersion was performed using the Kolmogorov-Smirnov test after subtracting group mean expression from each group. Differentially dispersed genes were defined as q-value < 0.05. We filtered for genes with a coefficient of variation (CV) fold change > 2 between the state in question and the remainder of the population. Enrichment of gene log-fold change in heterogeneity was performed using a mean-rank gene set enrichment test on GO Biological Process terms as described previously (Michaud et al., 2008).

Pseudotransition gene expression and transcription factor enrichment

Pseudotransition cell ordering was determined by applying the Monocle pseudotime algorithm to the states with significantly different proportions of control and Wnt+ exposed cells (states A, B, C and D). Gene ontology enrichment was performed on the resulting clusters of temporal gene expression using enrichR (Kuleshov et al., 2016). Transcription factor enrichment was performed by quantifying overrepresentation of target genes in the set of temporally regulated genes using the ChEA ChIP-X transcription factor binding database (Lachmann et al., 2010).

Chapter 4

Generative adversarial neural networks for analysis of scRNA-seq data

4.1 Introduction

The development of affordable single cell RNA-seq has enabled the measurement of transcript abundance in hundreds to thousands of individual cells (Hashimshony et al., 2012; Tang et al., 2009; Zilionis et al., 2017). These profiles provide an opportunity to define cell state and indirectly measure the signals and factors influencing cell fate. However, despite the richness of single cell measurements, these data are computationally challenging to analyse due to technical and biological noise meaning that traditional bulk expression analysis approaches are frequently not applicable. Current single cell RNA-seq dimensionality reduction methods can successfully reveal

clustering and structure within data when technical noise is low; however, they cannot easily integrate diverse datasets produced using distinct protocols (Pierson and Yau, 2015; Wang et al., 2017; Lin et al., 2017). Furthermore, current approaches focus on differential expression and marker gene identification but do not yield functional gene regulatory relationships.

Deep learning algorithms enabled by advances in computational power have demonstrated the capability to analyse diverse datasets from images to genomics (Esteva et al., 2017; Angermueller et al., 2017b). In particular, generative adversarial networks (GAN), first introduced in 2014, have shown promise in the field of computer vision (Goodfellow et al., 2014b). Since their introduction, GANs have become an active area of research with multiple variants developed that have resulted in improved performance and training (Radford et al., 2015; Chen et al., 2016; Miyato et al., 2017; Arjovsky et al., 2017; Gulrajani et al., 2017). Common to all of these variants (Creswell et al., 2017) is the concurrent training of two neural networks competing against one another, referred to as the generator and discriminator (Figure 4.1). The generator is tasked with generating simulated data, whereas the discriminator is tasked with evaluating whether data is authentic or not. Notably, only the discriminator directly observes real data while the generator improves its simulations through interaction with the discriminator. As training progresses both neural networks learn key features of the training data. Additionally, both discriminator and generator performance improves as they compete against each other.

In the field of computer vision, GANs have proved capable of generating visually

convincing and novel images such as faces or furniture (Karras et al., 2017b; Zamyatin and Filchenkov, 2017). Furthermore, it has been shown that the generator network learns a meaningful latent space where visually similar data such as similar faces are clustered. A meaningful latent space for single cell RNA-seq is particularly appealing as this can be used in conjunction with existing dimensionality reduction methods for improved and more meaningful subpopulation extraction. Additionally, an advantageous feature of generative neural networks is that non-linear relationships between features of the data are learned through training and can be later extracted to provide further insights.

Here, we integrate disparate epidermal datasets produced by three separate laboratories. The mammalian epidermis comprises interfollicular epidermis (IFE), hair follicles, sebaceous glands and further associated adnexal structures: together they form a protective interface between the body and external environment. Hence, epidermal cell fate is determined through the integration of extracellular cues with transcriptional activity (Watt and Hogan, 2000; Hsu et al., 2014). Under steady-state conditions, each epidermal compartment is maintained by distinct stem cell populations. However, under certain conditions such as wounding, each stem cell subpopulation is able to contribute to all differentiated lineages (Page et al., 2013). These distinct yet seemingly interchangeable subpopulations hint at a common, but as yet undefined epidermal gene regulatory network. For the first time, we have applied GANs to genomic data to uncover previously unknown gene regulatory relationships and regulators of epidermal cell state. We show that GANs produce biologically meaningful

dimensionality reduction and using the trained generator neural network we demonstrate that GANs can be used to predict the effect of cell state perturbations on unseen single cells.

4.2 Results

4.2.1 Generative adversarial networks integrate diverse datasets

We applied a generative adversarial network (GAN) to integrate multiple diverse mouse epidermal single cell RNA-seq datasets. These datasets originated from three labs, spanning mouse epidermal cells *in vitro* (Ghahramani et al., 2018), whole epidermis *in vivo* (Joost et al., 2016) and isolated subpopulations *in vivo* (Yang et al., 2017) (Figure 4.1, upper right. See methods and materials for GEO accession numbers). Each dataset has different characteristic technical and biological variation. No adjustments were made to account for batch-to-batch variations within datasets and across labs. After removing non-epidermal cell types and outlier cells we retained 1763 cells for training and 500 unseen cells for testing and validation. Similarly to other analysis methods applied to scRNA-seq, we filtered genes by expression level, resulting in 6605 potentially informative genes (see methods).

GANs consist of two neural networks competing against each other. The generator is tasked with producing realistic output data from a random input vector z named the input latent variable. In our case the latent variable z is of lower dimension (100

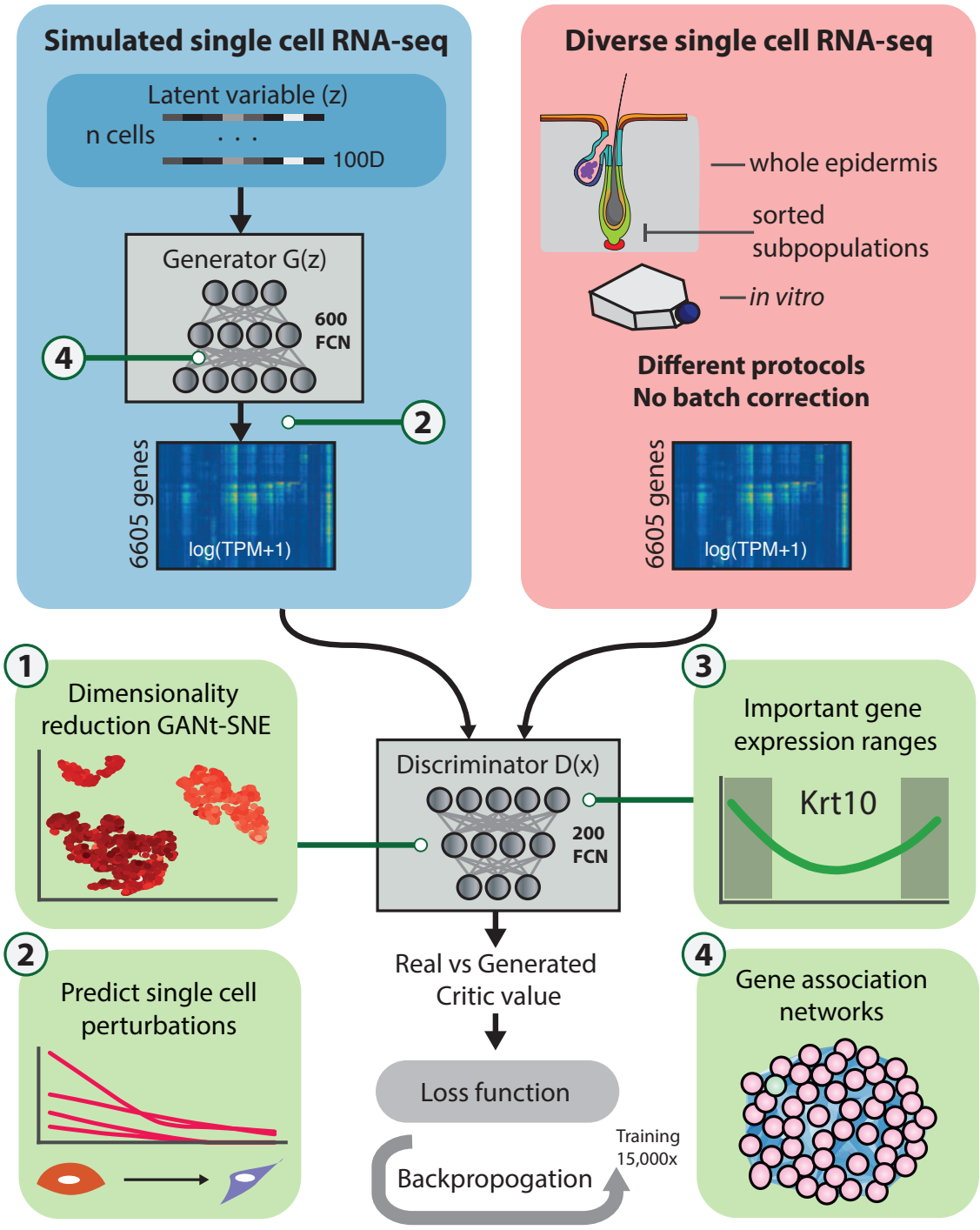


Figure 4.1: Overview of generative adversarial networks applied to scRNA-seq.
Generative adversarial networks consist of two neural networks concurrently training whilst competing against one another. These networks referred to as the generator and discriminator networks each have a distinct task. The generator is tasked with generating data by transforming a 100-dimensional latent variable into a single cell gene expression profile. In turn the discriminator is tasked with evaluating whether data is authentic or generated. Only the discriminator network sees the real single cell RNA-seq data, which have not been corrected for batch effects or technical variation. Once trained, we are able to extract biologically meaningful information from the generator and discriminator networks such as gene association networks, predict single cell time-courses, important gene expression ranges and biologically meaningful dimensionality reduction. Fully connected neurons, FCN.

dimensions) than the output scRNA-seq data (6605 genes), hence the generator represents a mapping from a lower dimensional space to gene expression space. During development of the GAN we varied the dimensionality of the input latent variable and finalised on a 100-dimensional vector, arbitrarily chosen as this is much lower than the desired output dimensionality (6605 genes) and we could not achieve stable training with latent space dimensionality an order of magnitude lower (10). Each 100-dimensional latent vector maps to 6605 transcripts per million (TPM, log(TPM+1)) normalised gene expression values for one cell. The second neural network, named the discriminator, is tasked with discriminating between the real and generated data. The discriminator takes log(TPM+1) gene expression values as input and outputs a score related to its assessment of the gene expression input. These two networks are trained using a combined loss function leading to improved performance of both neural networks (Figure 4.1).

We tested several variations of GANs ultimately utilising an Improved Wasserstein GAN with added gradient penalty term in the loss function as described previously (Gulrajani et al., 2017) due to training stability. Other variations tested include a Spectral Normalisation GAN (Miyato et al., 2017) and Loss-Sensitive GAN (Qi, 2017). We trained our generative neural network for approximately 15,000 epochs per full training run. We evaluated the generator network output performance at multiple checkpoints using t-distributed stochastic neighbour embedding (t-SNE, Figure 4.2) and correlation between real samples and generated samples as shown in Figure 4.3a and b. The t-SNE dimensionality reduction visually demonstrates the number of clus-

ters within real and generated cell populations. At each evaluation we generated 500 cells using 500 random latent variables and compared these to cells used for neural network training and cells witheld from training to be used as validation cells. Early in training, the GAN struggles to produce a varied output representative of the breadth of cell types and states; generated cells are closely correlated and form a single cluster in the t-SNE plot (Figure 4.2 1000 steps, Figure 4.3a). After 5,000 steps the generator begins to produce a varied output with generated cells mapping to multiple clusters in the t-SNE plot covering the different cell types, cell states, cell origin and experimental batches present in our combined dataset. After 10,000 steps, the generator network is capable of producing a subset of cells with similar transcriptomic profiles to those of Yang and colleagues that form a distinct cluster. Further GAN training broadens the distribution of correlations between generated samples, indicating an increasingly diverse generator output (Figure 4.3a).

We observed that after 13,000 steps continued training does not continue to increase the output diversity of the generator network, as defined by the median distance between cell gene expression profiles. Figure 4.4 shows that GAN output diversity reaches a maximum between 12,500 and 15,000 training steps. Furthermore, additional training between 12,500 and 15,000 steps does not change the generated cell correlation distribution (Figure 4.3a). At this point the discriminator loss function has also converged and therefore we cease further training (Figure 4.5).

We computationally validate our GAN by examining the correlation between generated samples and unseen samples (Figure 4.3). The distributions obtained at differ-

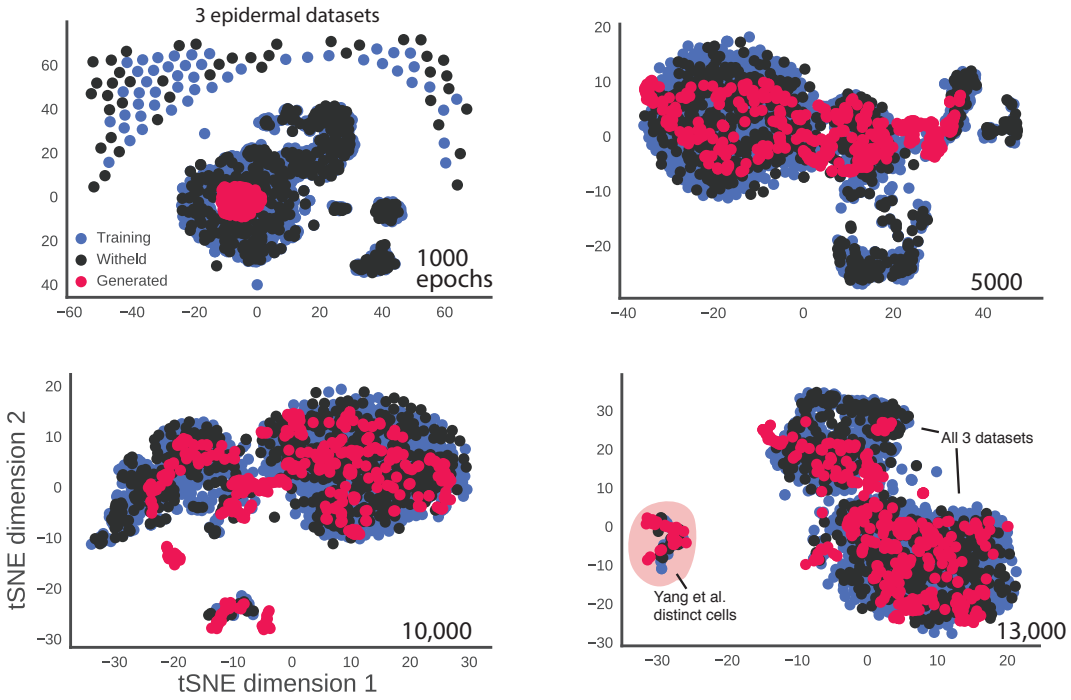


Figure 4.2: Generated cells at four training steps.
t-SNE projection of 1763 real training cells, 500 real withheld cells and 500 generated single cell expression data: each data point represents a cell and cells with similar expression profiles are positioned close together. Generated cells are clustered at the beginning, but with training gradually occupy the entire range of expression outputs for different cell types across three studies. Cells are visualised at 1000, 5000, 10,000 and 13,000 training steps.

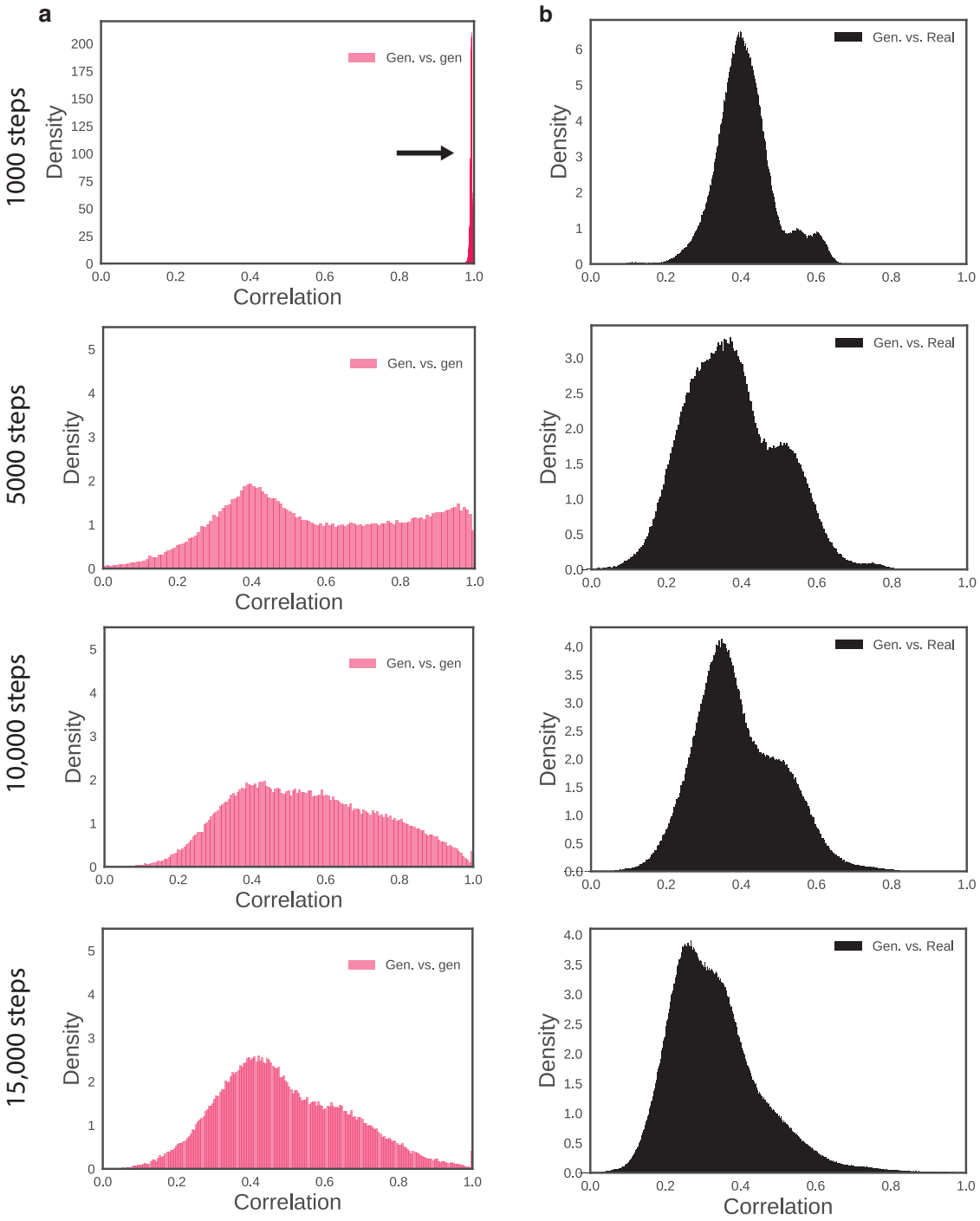


Figure 4.3: GAN training improves expression output diversity.
(a) Distribution of pairwise correlation between pairs of real or generated cells after 1000, 5000, 10,000 and 15,000 training steps. (b) Distribution of pairwise correlation between pairs of real and generated cells after 1000, 5000, 10,000 and 15,000 training steps.

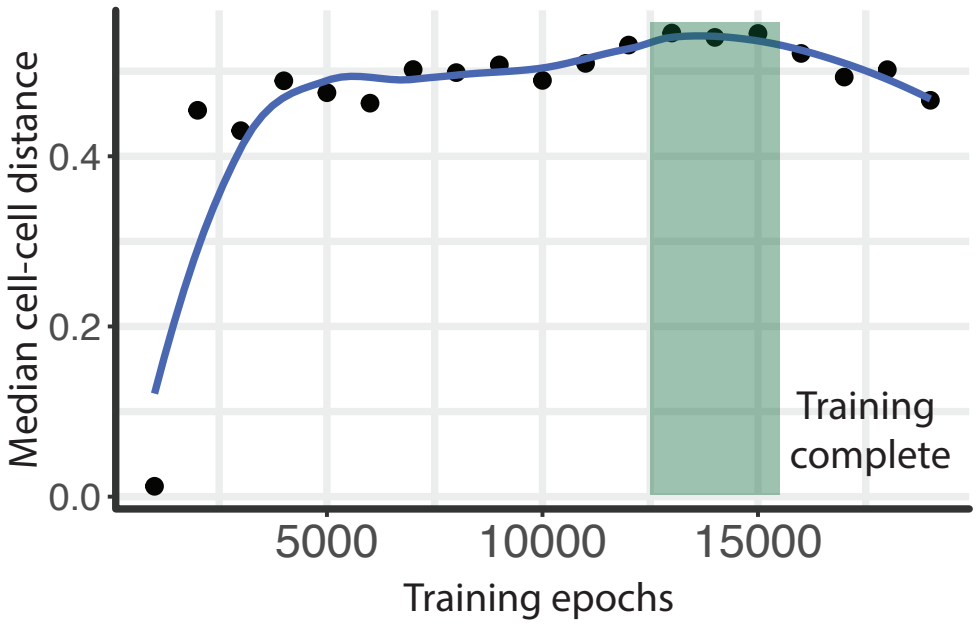


Figure 4.4: GAN output diversity over training time.
Relationship between GAN training epochs and output diversity as measured by median cellcell distance. Blue line is a LOESS regression. The greatest diversity of gene expression is achieved between 13,000 and 15,000 steps.

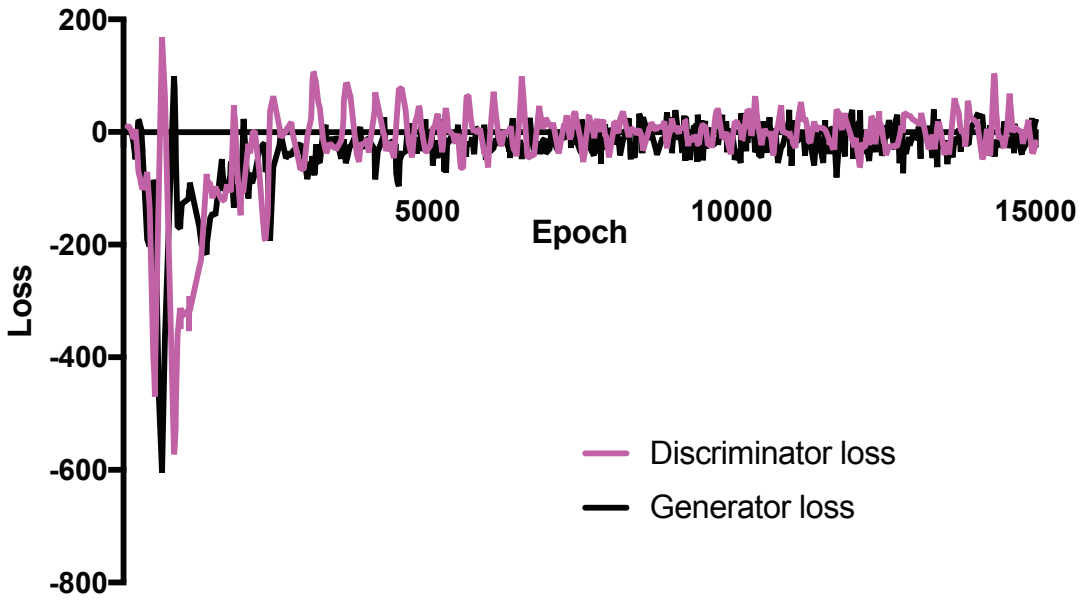


Figure 4.5: Generator and discriminator network loss curves.
Pink line shows discriminator network loss curve, black line shows generator network loss curve. Loss values for both neural networks converge to zero, indicating increasing performance for the networks over training. However, unlike most neural networks, convergence of loss for GANs is necessary but not sufficient to indicate that training is complete. Hence, we use loss in addition to other metrics set out in this section.

ent training steps show that generated samples are correlated with, but distinct from, real samples. At early training steps this distribution is centred around a Pearson correlation of 0.4. After 10,000 steps this distribution broadens, indicating the generator learns to simulate a diverse population of cells. This can also be seen from the expression of individual genes. Figure 4.6 shows single cell expression bar plots for three genes that distinguish subpopulations in our cell cohort, Ap1s3, Bcl2 and Nucks1. For each gene we show 500 unseen cells and 500 closely related generated cells with an average correlation of 0.71. Our generated cohort follows a similar pattern of expression to the real unseen cells, with Ap1s3 and Bcl2 highly expressed in differentiating cell types, and Nucks1 expression absent in vitro. In real cells where Ap1s3 or Bcl2 are not detected, the corresponding generated cells show low non-zero expression indicative of a form of imputation performed by the generator. Across all genes, expression variability in the simulated cells is similar to the real data. Together these results demonstrate that the generator network is not memorising and reproducing training samples but is instead inferring relationships between gene expression values in order to output convincing heterogeneous generated cells.

Focusing on the data from Joost and colleagues, the 1422 cells comprise eleven epidermal cell types originating from whole dorsal epidermis including previously uncharacterised cells. The generator neural network is able to simulate cells clustering with all *in vivo* epidermal cell types, ranging from sebaceous gland cells with distinct gene expression profiles to upper hair follicle and interfollicular epidermal cells (IFE) with similar transcriptomes but distinct spatial positions (Figure 4.7). This range of

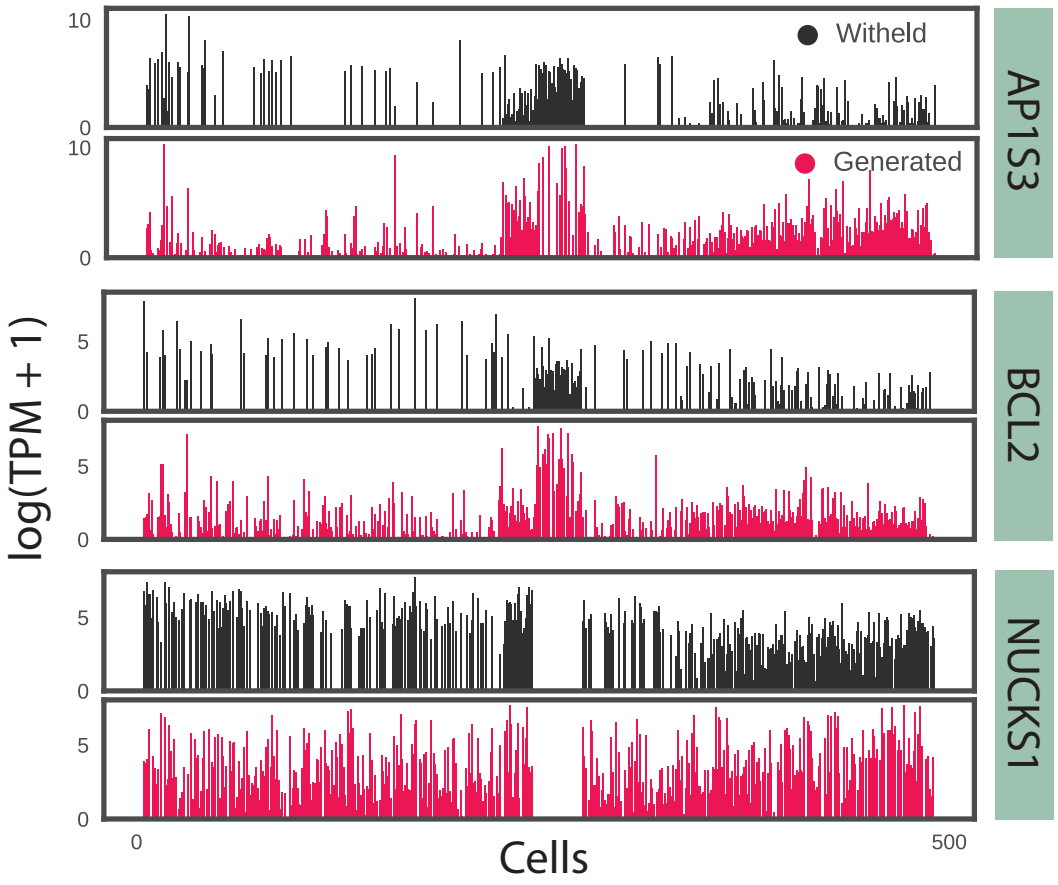


Figure 4.6: Generated and real expression values for three genes.
Bar plots showing expression of Ap1s3, Bcl2 and Nucks1 in 500 unseen real cells and 500 closely related generated cells. Cells are ordered by hierarchical clustering of the unseen real cells using all 6605 genes.

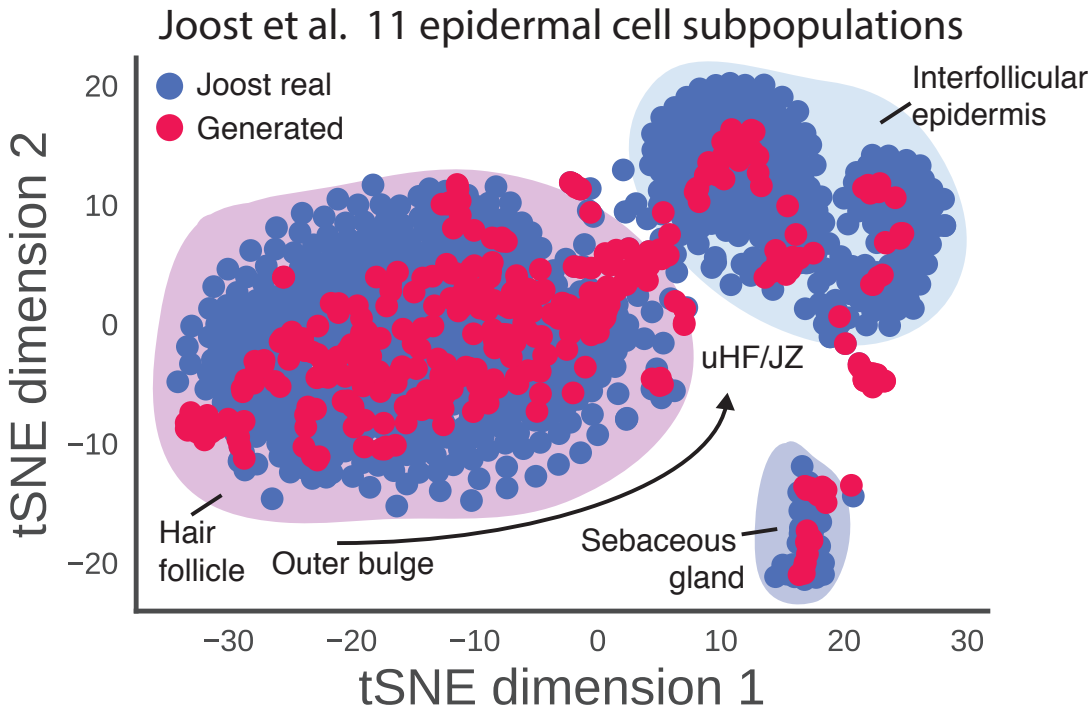


Figure 4.7: Generating cells from the Joost et al. dataset.
t-SNE projection of subset of real cells from Joost et al. and corresponding generated cells. Despite training with a larger and diverse dataset, the GAN captures cell diversity at a detailed level described by Joost. Labels show the corresponding cell type for each cluster. Upper hair follicle, uHF. Junctional zone, JZ.

cell type simulation is achieved despite training the GAN using three diverse datasets.

4.2.2 GAN training on non-epidermal datasets

We next sought to apply the GAN algorithm to non-epidermal datasets to ascertain whether the method can generalise to a more diverse range of cell types. We focused on the generative capabilities of the GAN and assessed whether it was equally capable of generating cells from non-epidermal datasets. We selected three datasets on the basis of their diversity.

Firstly, we applied the GAN to a timecourse of motor neuron differentiation (Sagner et al., 2018). Briefly, Sagner and colleagues started from human embryonic stem cells (ESCs) and over seven days directed their differentiation towards motor neurons using a combination of signalling factors including Wnt and FGF. For this dataset we trained the GAN using 2342 cells and 14,000 genes, substantially more than for the epidermal datasets owing to the more diverse range of cell types present from ESC to motor neuron. Figure 4.8a shows a t-SNE projection of real and generated cells resulting from training the GAN on this motor neuron dataset. The GAN is able to generate cells from all cell type clusters present in the real dataset and produces a similar number of outlier cells to the real scRNA-seq.

Subsequently, we applied the GAN to two droplet-sequencing based single cell datasets. We first analysed data from Greenleaf and colleagues alone, training the GAN on 7,400 cells which passed QC filtering (Buenrostro et al., 2018). This dataset covers the full spectrum of human hematopoietic differentiation comprising a minimum of 10 phenotypically distinct cell types sequenced using the 10x Genomics Chromium system. A key feature of the Chromium system is the high-throughput of sequenced cells, typically in the tens of thousands. Furthermore, sequenced libraries are of lower depth than FACS or Fluidigm C1 based methods with typically 50-100,000 aligned reads per cell in contrast to up to 1 million for the Fluidigm C1 epidermal data from Joost et al. (2016). As before, the fully trained generator neural network is able to generate cells clustering with the full spectrum of gene expression profiles as demonstrated in Figure 4.8b. We next retrained on the combined dataset of Buenrostro et al.

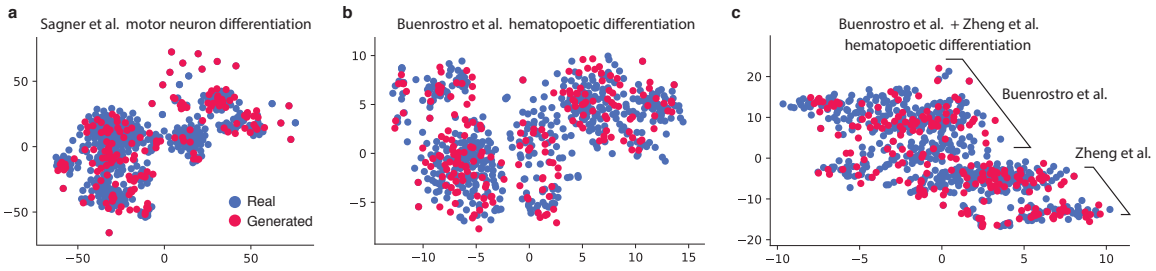


Figure 4.8: Applying GAN to non-epidermal datasets.
(a-c) t-SNE projection of a subset of real and generated cells from (a) Sagner et al. (2018) profiling motor neuron differentiation, (b and c) Buenrostro et al. (2018); Zheng et al. (2017) profiling hematopoietic differentiation using high-throughput droplet-sequencing based methods. For each dataset we visualised 500 real and 300 generated cells.

(2018) and Zheng et al. (2017) to understand whether the trained generator network could successfully simulate gene expression profiles from two large (> 100,000 cells) datasets. Figure 4.8c shows the t-SNE projection after GAN training on the combined dataset. Here again, one trained generator network demonstrates the ability to simulate cells clustering with two datasets and two distinct sets of technical and biological variation.

4.2.3 Dimensionality reduction by combining the discriminator network with t-SNE

Focusing again on the three integrated epidermal datasets, we hypothesised that after completion of GAN training the discriminator network learned biologically relevant features of scRNA-seq data, discarding uninformative genes in order to discriminate successfully between the generated and real samples. The discriminator neural network consists of a single hidden layer whose output is transformed to a discriminator

or "critic" output value. Therefore the discriminator transforms 6605 gene expression values into 200 learned internal features, effectively a reduced dimensional representation. We sought to understand whether the discriminator hidden layer output contained learned features that could be used for dimensionality reduction. We performed t-SNE on the discriminator hidden layer output to visualise these learned features (Figure 4.9) and compared this with two other approaches; PCA alone and t-SNE alongside PCA-based batch effect removal.

First we performed a principal component analysis (PCA) on the combined epidermal training data (Figure 4.9a, middle column); the main source of linearly separable variance in the first two components is the batch effect of dataset origin and laboratory and the approach generally fails to identify biologically meaningful clusters even at higher principal components. We removed this technical source of variation by subtracting the first two principal components and performing a t-SNE on the resulting dataset. The resulting t-SNE (Figure 4.9a, right column) no longer clusters cells by dataset origin, so removing variation caused by different experimental protocols, at the cost of concurrently removing informative biological variation in gene expression.

Two major sources of variation in epidermal cells are their differentiation status and spatial position within the epidermis: a useful dimensionality reduction approach should capture at least one of these features, ideally both. To investigate this we overlaid expression of Krt14, a marker of undifferentiated keratinocytes, Krt10 a marker of commitment to differentiation and Rps29, a ubiquitously expressed ribosomal protein upregulated in the IFE in comparison to hair follicles.

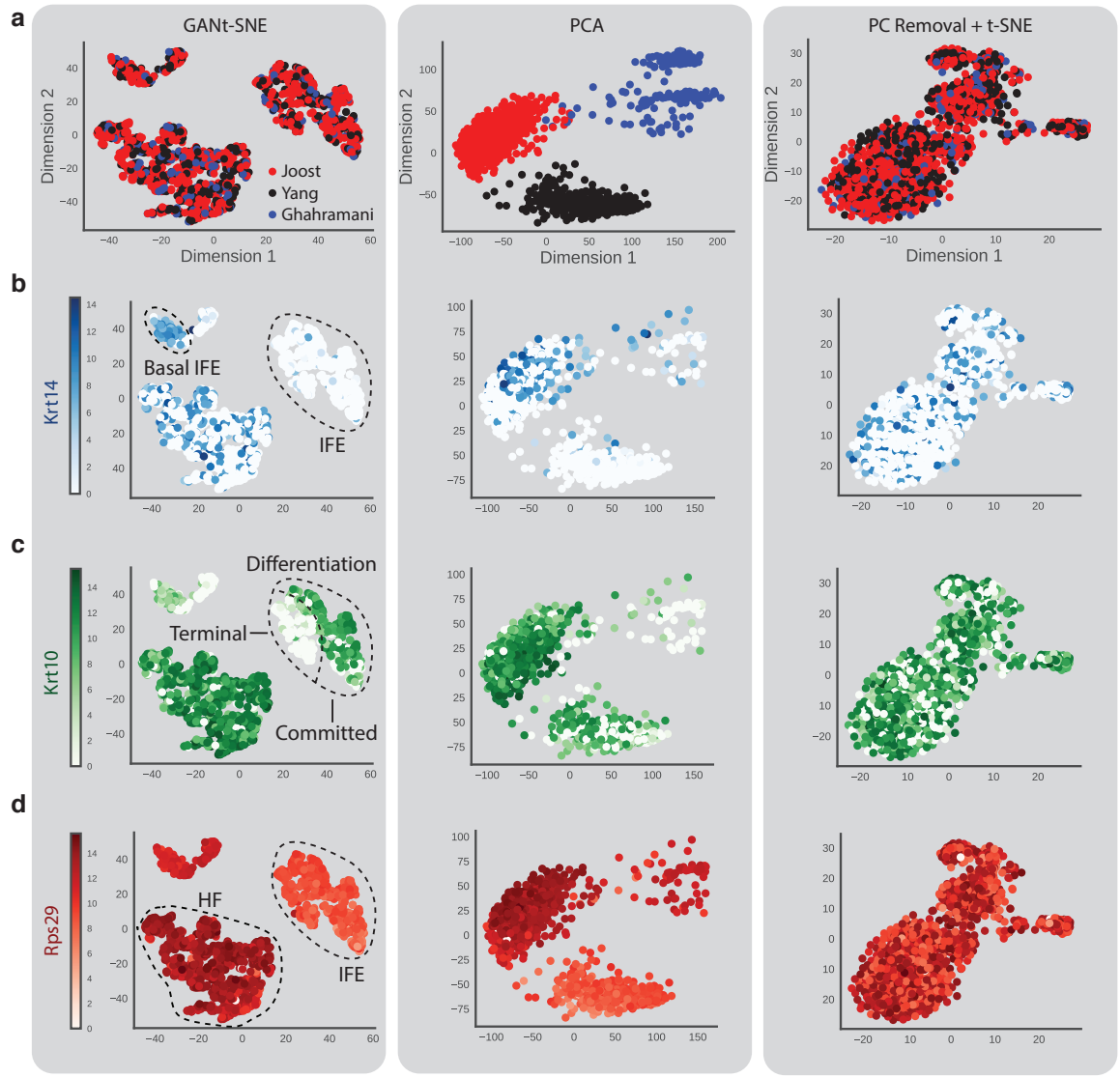


Figure 4.9: GANt-SNE clusters biologically similar cells.
(a) GANt-SNE visualises 2263 real cells by performing t-SNE on 200 features captured by the discriminator neural network's hidden layer. This is compared with principal component analysis (PCA) and t-SNE after removal of linear batch effects using PCA (PC removal + t-SNE). PCA is heavily affected by batch effect and clusters the three datasets separately. (b) Keratin-14 expression overlaid onto projections. Krt-14 is a marker for basal interfollicular epidermis (IFE). GANt-SNE distinguishes between basal IFE and non-basal IFE cells, whereas PCA and PC removal + t-SNE fail to distinguish between these cell types. (c) Keratin-10 is a marker for differentiation interfollicular epidermis (IFE). . GANt-SNE successfully clusters separates supra-basal and terminally differentiated IFEs. (d) Ribosomal Protein S29 is known to be expressed highly in interfollicular epidermis (IFE) and at medium levels in hair follicles (HF). GAN-tSNE successfully distinguishes between these two major cell types.

Biologically, we expect cells with similar expression levels of these marker genes to cluster. PCA and the combined approach of PC removal + t-SNE perform poorly in this regard and expression of all three markers fails to distinguish cells across the dimensionally reduced space (Figure 4.9 middle and right column). In contrast, our GANt-SNE approach clusters Krt14^{high} cells deriving from the basal layer of the IFE (Figure 4.9b, Basal IFE) which are separated from the remaining IFE cells. A third cluster of hair follicle cells shows sporadic expression of Krt14 as this is not a functional marker for these cells. Our dimensionality reduction separates differentiated IFE cells highly expressing Krt10 from terminally differentiated and keratinized Krt10^{low} cells (Figure 4.9c). Furthermore, GANt-SNE correctly separates IFE-derived cells from hair follicle cells as seen by Rps29 expression (Figure 4.9d). These biological distinctions are not identified by the two alternative methods of dimensionality reduction.

These results lend credence to our hypothesis that the discriminator learns biologically relevant features of the data. Using the GANt-SNE approach we successfully separate cells by differentiation status and spatial position (IFE vs. hair follicle and differentiating vs. undifferentiated). This is achieved without a priori knowledge of technical variation and batch effects by training on the commonalities between training data. In contrast the PCA and t-SNE approaches fail to separate cells into these biologically meaningful clusters, as they crudely extract the strongest sources of variation - often a non-linear mixture of biological and technical variation.

4.2.4 Simulating cellular perturbations using latent space interpolation

To take advantage of the gene expression rules learned by the generator network we devised an algorithm for retrieving the latent space vector z from an arbitrary gene expression profile x, such that $x = G(z)$. For each target cell we randomly sampled latent space vectors (z) and generated cells until a sufficiently similar cell is obtained (see methods). In other fields where GANs have been applied, such as image generation, the latent space representation of an image meaningfully represents visually similar images. Furthermore, vector arithmetic in the latent space leads to meaningful outputs. For example Radford and colleagues (Radford et al., 2015) have shown that subtracting the latent space vectors of a face wearing glasses from a face without glasses results in a differential vector representing glasses; adding this to a different face outputs a face with glasses (i.e.$G(z')$ produces a person wearing glasses where $z' = (z_{\text{glasses}} - z_{\text{no glasses}}) + z_{\text{face}}$).

We hypothesised that latent space arithmetic can be extended to cell types and cell states. To investigate this we simulated the process of differentiation at the single cell level using latent space arithmetic, as shown in Figure 4.10. We sampled 30 terminally differentiated and undifferentiated pairs of cells from the unseen pool of cells in the Joost dataset using the original study labels, obtained their latent space vectors $z_{\text{differentiated}}$ and z_{basal} and calculated the difference between these vectors $\delta = z_{\text{differentiated}} - z_{\text{basal}}$. We sampled cells from both the interfollicular epidermis and hair

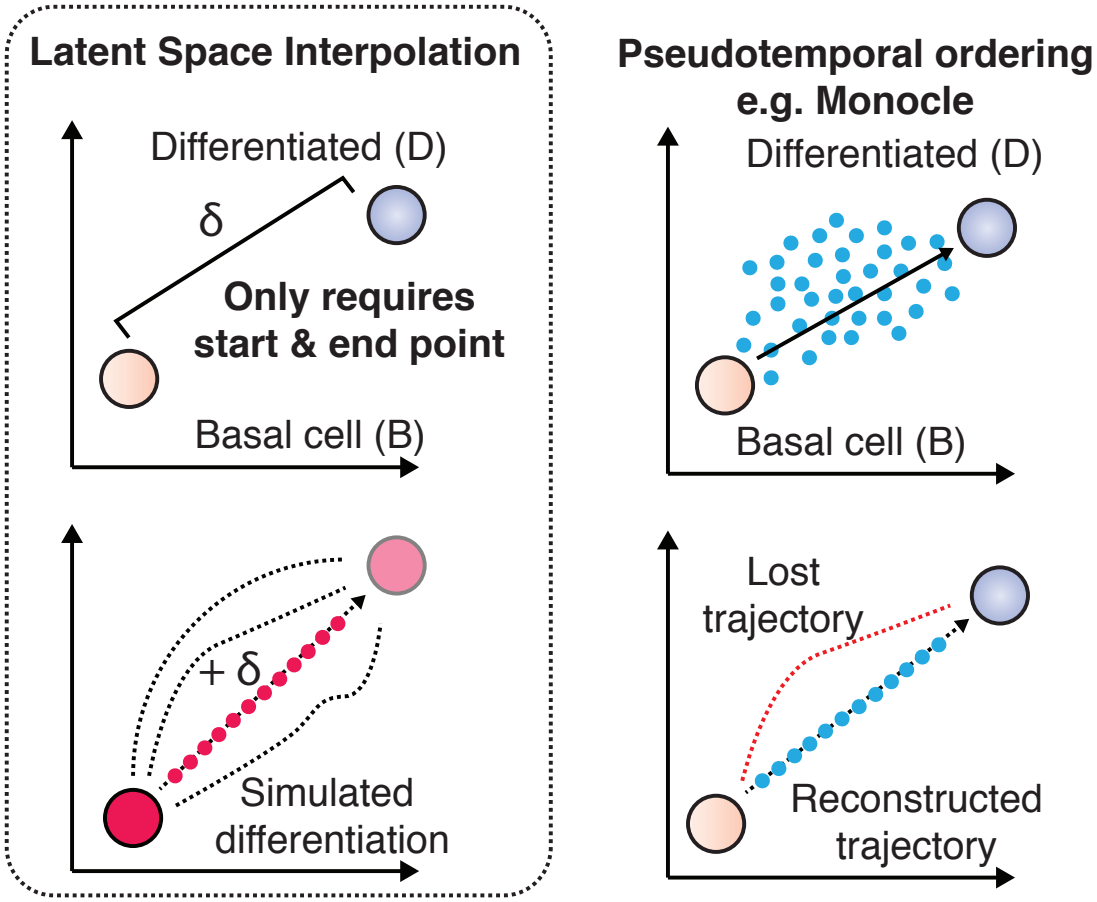


Figure 4.10: Simulating cell state transitions using latent space interpolation.
Schematic showing how to simulate differentiation for single cells using latent space interpolation. Expression data for pairs of undifferentiated and differentiated cells are transformed to their respective latent space vectors (z_{basal}, $z_{differentiated}$)) The difference between the vectors (⍰) represents differentiation in latent space (upper panel). This difference vector is then applied to the latent space vector of another unseen cell in order to predict the terminally differentiated state (lower panel). We also simulate 1000 differentiation timepoints.

follicles in order to obtain a universal differentiation latent vector. Unseen cells were used to demonstrate that the generator's learned latent space could generalise to previously unseen gene expression profiles. We then added δ, the latent space differentiation vector, to an unseen undifferentiated cell and interpolated 1000 timepoints between the undifferentiated latent space point and the simulated differentiation latent space endpoint. We used the generator network to produce gene expression for all 1000 timepoints and repeated this process for 30 cells with heterogeneous undifferentiated starting profiles. It is important to note that only the starting point (the undifferentiated cell gene expression profile) is real, whereas the differentiated endpoint and all timepoints in between are generated by the neural network.

Figure 4.11 shows simulated time-series gene expression profiles for a selection of genes. Two differentiation markers, Periplakin (Ppl) and Grainyhead-like protein 3 homolog (Grhl3), demonstrate the ability of our latent space arithmetic to simulate successfully simulate differentiation of single cells. Over differentiation time points the mean expression level of these two markers increases, spanning the expression levels observed in the undifferentiated and differentiated subpopulations (B - basal start point, D - differentiated end point median expression line). Similar to real scRNA-seq gene expression profiles, the generated time points for individual cells display substantial cell-to-cell variation in gene expression over time, with cells showing different gene expression curves dependent on initial expression level and cell state. Similarly, for basal IFE markers such as the examples in Figure 4.11, Integrin-beta 1 (Itgb1) and Metallothionein 2A (Mt2) the reduction in expression is successfully captured by the

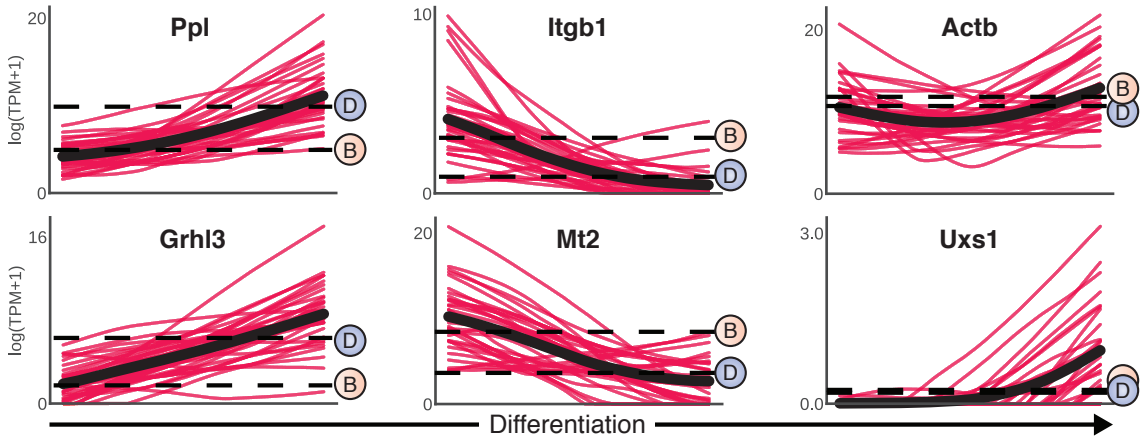


Figure 4.11: Simulated single cell expression profiles (30 cells) for six genes.
Simulated expression profiles for 6 representative genes in 30 cells over 1000 differentiation timepoints. Red lines indicate expression profiles in individual cells and lack lines provide the mean expression over all cells. Dotted lines correspond to the median expression of basal undifferentiated (B) and differentiated (D) cells from the Joost dataset. Periplakin (Ppl) and Grainyhead like transcription factor 3 (Grhl3) are known differentiation whose simulations show increasing expression. Integrin beta-1 (Itgb1) and Metallothionein 2 (Mt2) are basal IFE marker which display gradually decreasing expression Actin beta (Actb) and UDP-Glucuronate Decarboxylase 1 (Uxs1) are not traditionally considered to be markers but display non-linear expression profiles with substantial cell to cell variability

GAN latent space interpolation.

Moreover, the expression curve is not necessarily monotonic or linear. We also observed a range of highly non-linear expression profiles such as genes only expressed in early or late timepoints (Uxs1), parabola-like expression (Actb, Ube2f) and genes which change in expression variation but not mean expression. These non-linear expression profiles demonstrate that latent space interpolation produces meaningful gene expression predictions based on gene expression interdependencies and does not simply average between gene expression profiles.

Next, we applied Monocle to the Joost et al. (2016) dataset to contrast our predicted gene expression profiles with pseudo-ordering derived observations. We ap-

plied Monocle to IFE cells only in order to capture IFE differentiation dynamics. Figure 4.12 shows the predicted cell state trajectories when Monocle is applied using default settings. Using this cell ordering, we reconstructed gene expression dynamics for Ppl, Grhl3, Itgb1 and Uxs1 as shown in 4.13. For Ppl and Grhl3 we expect increasing expression over differentiation. Monocle's pseudo-ordering captures this increase, however, the reconstructed expression dynamics over differentiation are constructed from the mean of ordered cells. Hence, the Monocle method does not predict the likely difference in dynamics from cell to cell.

Focusing on Itgb1, we expect expression of this basal integrin to decrease over differentiation. Here, pseudo-ordering has mis-ordered cells early and mid-differentiation, hence failing to capture that cells earliest in differentiation highly express Itgb1. This can be partially rectified by guiding pseudo-ordering using Itgb1 as a marker gene. Finally, few cells express Uxs1, hence Monocle predicts that Uxs1 is invariant to differentiation, as the $Uxs1^{+}$ cells are averaged out by the majority of $Uxs1^{-}$. Observing the pseudo-ordering it is apparent that Uxs1 is only expressed by differentiating cells, a feature captured by the GAN latent space interpolation.

To investigate further the dynamics of gene expression over epidermal differentiation we clustered the simulated time-series profiles for all 6605 genes. Using k-means clustering we were able to distinguish eight types of dynamic expression profile (Figure 4.14). Genes expressed early in differentiation were enriched for epidermal stem cell related gene ontology and concordantly late-expressed genes were enriched for epidermal differentiation (q-value<0.01 for both, Figure 4.15). We focused

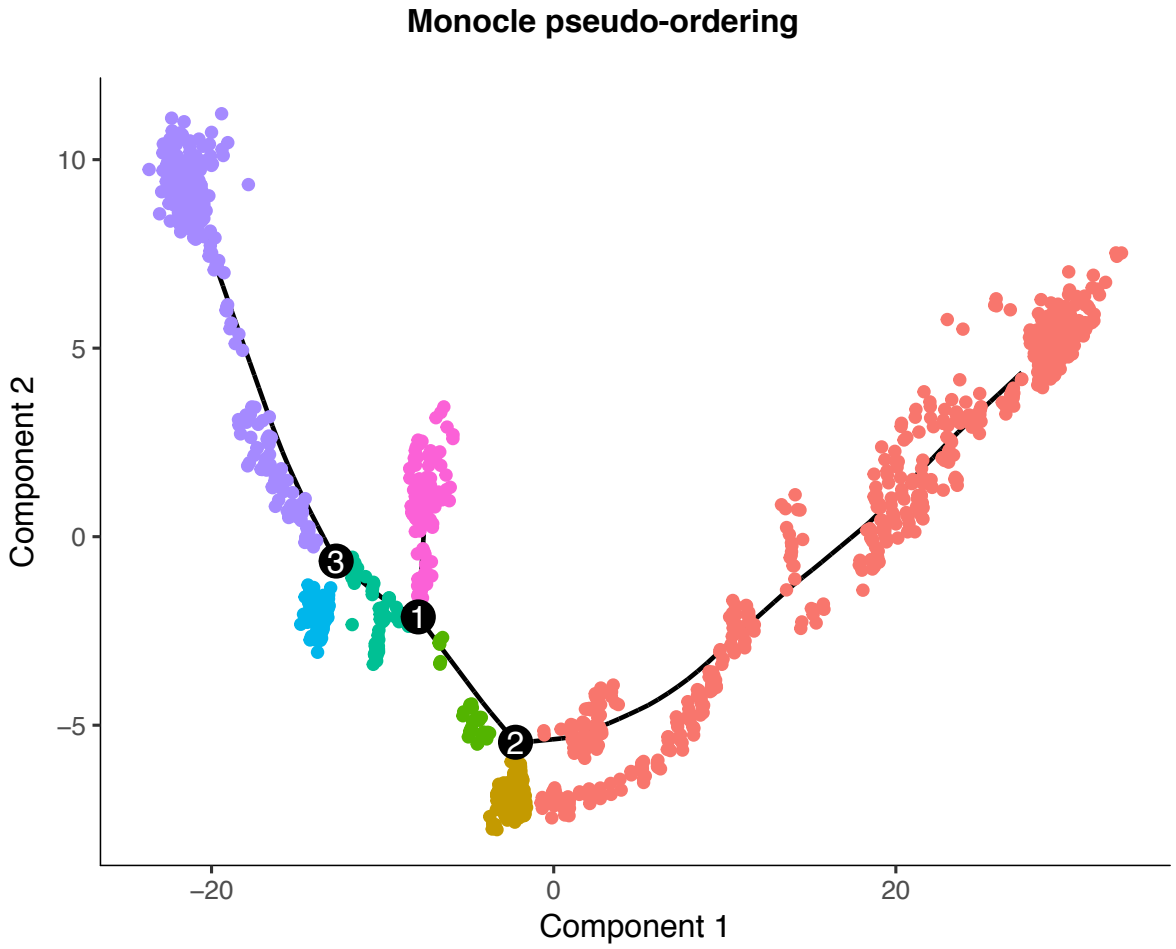


Figure 4.12: Monocle predicted pseudo-order of IFE cells.
Monocle dimensionality reduction with superimposed trajectory of IFE cells from Joost et al. (2016). Cells are coloured by Monocle-predicted clusters. Labels 1, 2 and 3 indicate Monocle-predicted cell state trajectory branch points.

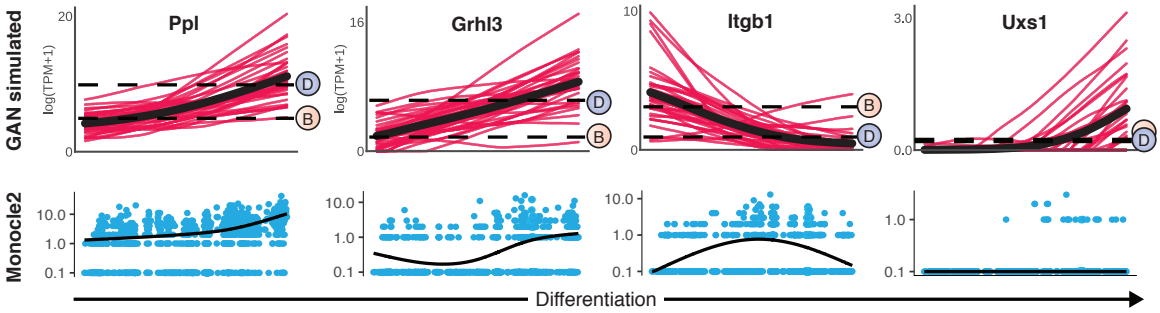


Figure 4.13: LSI simulated expression compared to Monocle.
Simulated expression profiles for 4 representative genes in 30 cells over 1000 differentiation timepoints (as before in Figure 4.11. (Upper) red lines indicate expression profiles in individual cells and lack lines provide the mean expression over all cells. Dotted lines correspond to the median expression of basal undifferentiated (B) and differentiated (D) cells from the Joost dataset. (Lower) blue represents IFE cells as pseudo-ordered by Monocle and the predicted expression dynamics derived from the pseudo-ordering.

on four clusters, two corresponding to increasing and decreasing expression (Figure 4.14, Groups I and II) and two subsets of genes only transiently up- or down-regulated during commitment to differentiation (Figure 4.14, Group III and IV genes). To validate these findings we used a recently published dataset investigating human epidermal differentiation (Mishra et al., 2017). Mishra and colleagues utilised methylcellulose suspension-induced differentiation to obtain temporal differentiation gene expression data from human keratinocytes at 0, 4, 8 and 12 hours. A majority of our predicted transiently expressed genes are dynamically expressed in this dataset with Groups I and II showing decreasing and increasing bulk expression over the differentiation time course (Figure 4.16, Group I and II). Focusing on Group IV, these 223 genes are of particular biological interest as they are predicted to be transiently expressed during commitment to differentiation and are therefore likely to play a functional role in this process. From our predicted epidermal commitment genes, this group clusters into three subgroups based on bulk peak expression around 0 hours, between 4 and 8 hours, and 12 hours or later. We hypothesise that three groups are observed as our

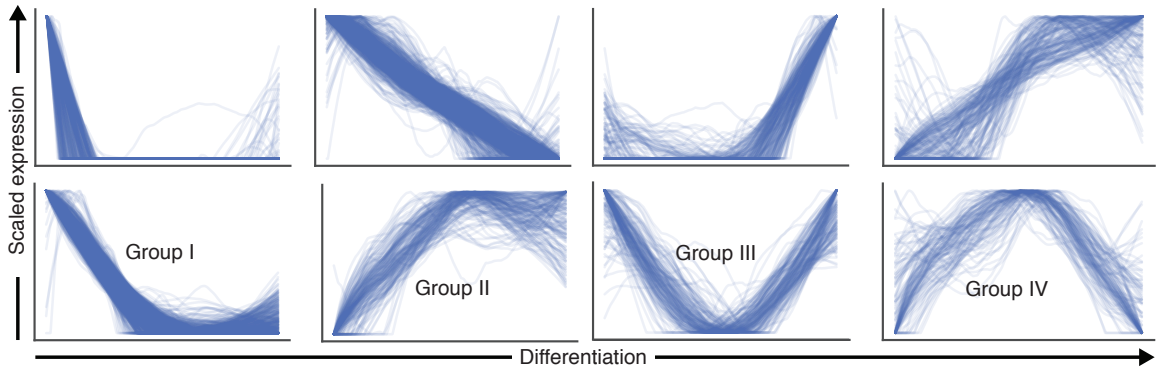


Figure 4.14: Clustering of simulated expression profiles.
K-means clustering of simulated differentiation expression profiles for 6605 genes. We identify eight broad types of temporal gene expression, of which four of the dynamic expression patterns are labelled Groups I-IV.

simulated differentiation assumes a synchronised population of differentiating cells. The asynchronous nature of *in vitro* differentiation and heterogeneous starting population reduces our ability to detect transiently expressed and subsequently downregulated genes. From the subgroup with peak expression between 4 and 8 hours we identified two protein phosphatases, PTPN1 and PTPN13 which are also identified and extensively validated by Mishra and colleagues. This is followed by transient MAF expression - a member of the AP1 subfamily of differentiation transcription factors - predicted by our latent space interpolation approach and observed in the Mishra dataset along with previous studies (Lopez-Pajares et al., 2015).

In summary, using our generative model we have simulated a perturbation to cell state in the form of epidermal differentiation and subsequently obtained high-resolution predicted gene expression profiles. Using the generator neural network we have obtained information on switch-like expression of genes, an aspect of gene expression which is difficult to infer from previous methods of single cell RNA-seq analysis such as pseudotime ordering. Pseudo-ordering of cells provides an under-

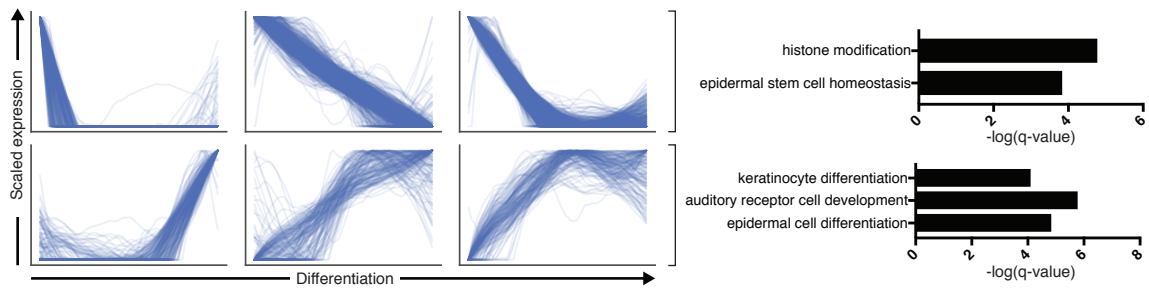


Figure 4.15: Gene ontology enrichment for predicted differentiation genes.
Gene ontology enrichment for genes that generally decrease (upper) or increase (lower) over differentiation as predicted by latent space interpolation.

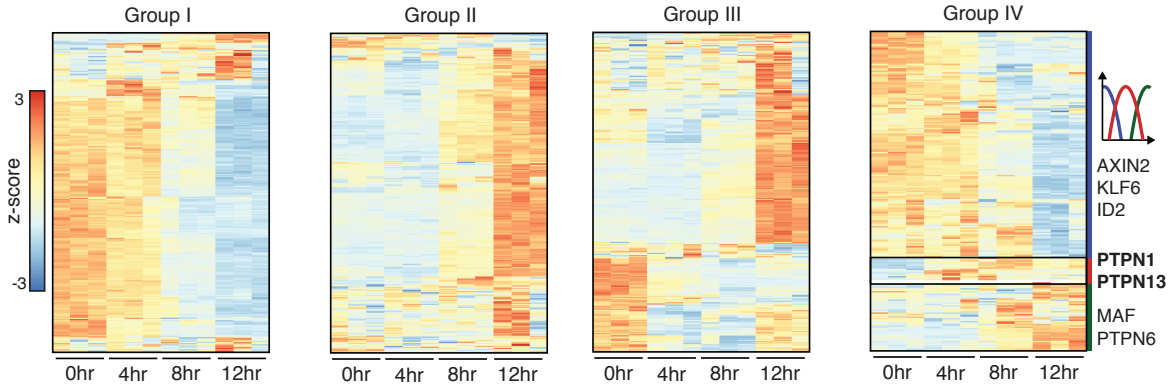


Figure 4.16: Validation of LSI prediction with bulk gene expression data.
Heatmaps showing bulk expression levels of genes in Groups I-IV from Mishra et al. at 0, 4, 8 and 12 hours of suspension-induced differentiation. Most genes display similar expression profiles to the simulations, despite the differences in experimental set ups.

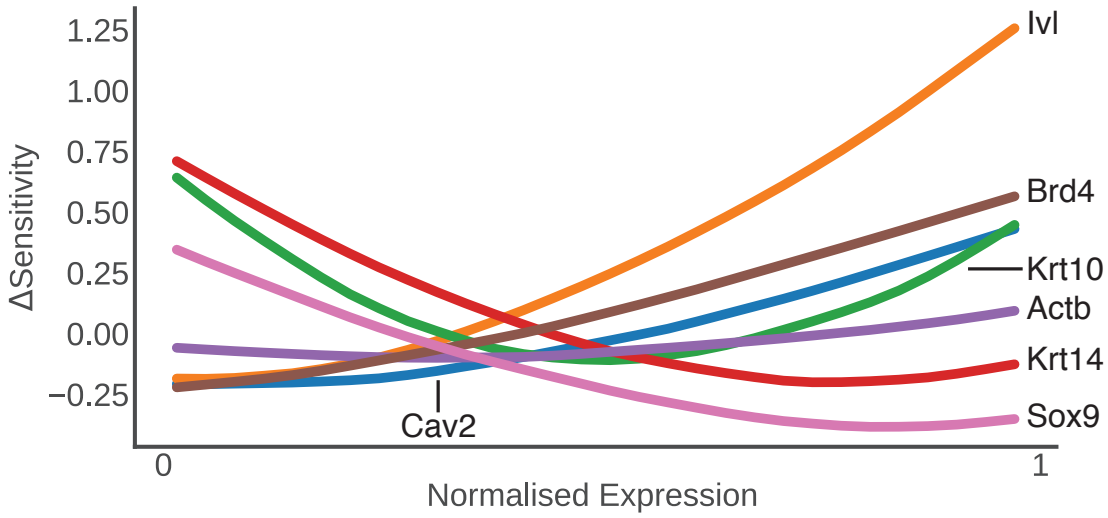


Figure 4.17: Sensitivity analysis of discriminator network.
Sensitivity curves for seven known markers for epidermal or determinants of epidermal state: Actin beta (Actb), Involucrin(Ivl), Keratin 10 (Krt10), Keratin 14 (Krt14), Caveolin 2 (Cav2), Bromodomain-containing protein 4 (Brd4), SRY-Box 9 (Sox9). Some genes such as Ivl an IFE differentiation marker display higher sensitivity at increased expression levels whereas others display the opposite trend.

standing of the sequence of gene expression events, however, this can be confounded by highly transient gene expression events which may lead to erroneously missing or identifying cell state transitions. In comparison, our generative method is guided by gene expression rules learned by the GAN, hence producing valid gene expression profiles at all timepoints. These results can be extended to any other perturbation captured by latent space.

4.2.5 Discriminator network identifies state-determining gene expression ranges

As GAN training progresses, the discriminator network learns to identify compatible ranges of gene expression levels and interrelationships for all genes. An advantage

of our approach is that these relationships can be highly non-linear, for example the discriminator can learn gene expression interdependencies which only apply when genes are expressed in a certain range or combination. To extract this learned information we performed a sensitivity analysis on the discriminator network by taking real gene expression profiles from the unseen cell group and varying expression of genes individually from the lowest observed expression level in the cohort to the maximum observed expression level. Figure 4.17 shows the relationship between adjusted expression level and change in discriminator network output critic value. This analysis produced a sensitivity curve for each gene where absolute sensitivity value indicates a strong change in discriminator critic value and the gradient of the sensitivity curve denotes ranges of gene expression that the discriminator is sensitive to. For genes with no known role in epidermal cell state such as Actb there is little change in sensitivity across all expression levels. In contrast, for known markers or determinators of epidermal state, such as Ivl, Krt10, Krt14, Cav2, Brd4 and Sox9, there is a strong relationship between expression level and sensitivity. For some genes such as Ivl, an IFE differentiation marker, an increase in expression showed increased sensitivity, i.e. a greater effect on discriminator critic value. However, for other genes such as Krt10 medium levels of transcription resulted in a lower sensitivity than low and high levels indicating that for this gene there are two important ranges of transcription. The discriminator network identification of Krt10 import expression ranges is supported by strong expression of Krt10 in the interfollicular epidermis above the basal layer and subsequent downregulation in terminally differentiating cells (Joost et al., 2016). Krt10 misexpression has been shown to cause epidermal barrier defects (Müller et al.,

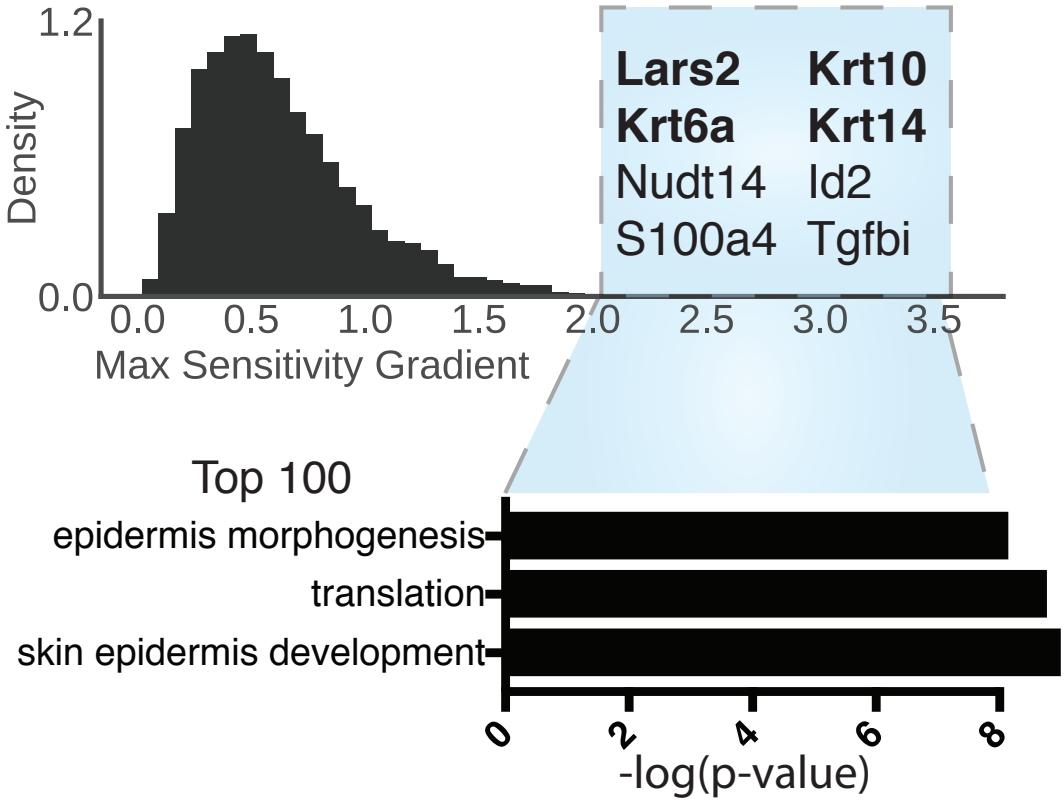


Figure 4.18: Sensitivity analysis identifies epidermal regulators.
(Upper) histogram of maximum sensitivity curve gradients of 6605 genes, with the top eight genes labelled; known markers are highlighted in bold. (Lower) Gene Ontology enrichment for the top 100 genes by maximum sensitivity gradient as scored by -log10(p-value).

2006; Cheng et al., 1992). Furthermore, a cellular need for binary expression of Krt10 is clear as the protein constitutes approximately 40% of all cellular proteins in the suprabasal layer of the epidermis (Fuchs and Green, 1980; Fuchs et al., 1992).

We sought to compare the relative importance of genes in determining cell transcriptional state across all epidermal cell types. Sensitivity varies non-linearly with expression level, hence, to compare all genes we calculated the maximum gradient of the sensitivity curve for each gene (Figure 4.18 upper). This metric is an indication there is at least one range of expression levels which has a large effect on discriminator net-

work output. Strikingly, the top 100 genes identified are highly enriched for known epidermal regulators and comprise both low and highly expressed genes (Figure 4.18 lower). Amongst the top 10 are three keratin genes with spatially-restricted expression; Krt6a (inner bulge), Krt10 (suprabasal IFE) and Krt14 (basal IFE). Our analysis predicted Lars2 expression as one of the most important indicators of pan-epidermal cell state. Lars2 is a mitochondrial leucyl-tRNA synthetase and is upregulated during differentiation as seen by our differentiation predictions and from the Joost dataset (Joost et al., 2016). Considering these analyses we hypothesise that Lars2 is one of the key regulators of cell metabolic considerations during differentiation. This sensitivity analysis approach enables identification of state-regulating genes without bias for transcript abundance.

4.2.6 GAN-derived gene association networks predict Gata6 targets

Finally we examined the internal features of the generator network to extract gene expression interdependencies to complement our sensitivity analysis ranking of state-determining genes by providing context on how these genes are coregulated. The final layer of the generator network non-linearly transforms an arbitrary number of internal features to the final 6605 gene expression values using a leaky rectified linear unit activation function. We found empirically that 600 internal features produced stable and diverse generator output, representing a maximum of 600 features that, when non-linearly combined, produce each gene expression output value. Since these internal features are an order of magnitude fewer than the number of gene expression values,

the generator network is forced to learn the most salient relationships between genes in order to produce a convincing output. Hence, genes with correlated generator final layer values are predicted to correspond to co-regulated genes. This is distinct from correlation analysis of gene expression, which infers linear and directly correlated regulation.

We visualised the correlation of the final layer weights for the 600 features between all genes using a force directed network (Figure 4.19) and also examined the structure and clustering of the genes using a hierarchically clustered heatmap. On a macroscopic scale genes are segregated by their overall positive or negative effect on gene expression. This can be seen from the polarity of the generator derived gene-gene network in Figure 4.19

On a local scale the majority of genes cluster into small groups of between 10 and 50 closely associated genes. We used these local gene association networks to examine Lars2, a gene predicted from our analysis to be highly important for epidermal cell state. Lars2 is a mitochondrial leucyl-tRNA synthetase and our analysis shows it to be a member of an extremely closely regulated group of ribosomal and translation related genes (Figure 4.20). We also examined the local gene association networks for three known epidermal regulators examined in our previous sensitivity analysis Krt14, Itga6 and Ivl, also shown in Figure 4.20. Itga6 and Krt14 are both known basal IFE markers; the local network for these genes highlights the power of our machine learning approach to identify coregulated genes. Several known basal IFE regulators are present in both local networks, such as Krt5, Krt14, Itga3, Cav2 and Col17a1. We

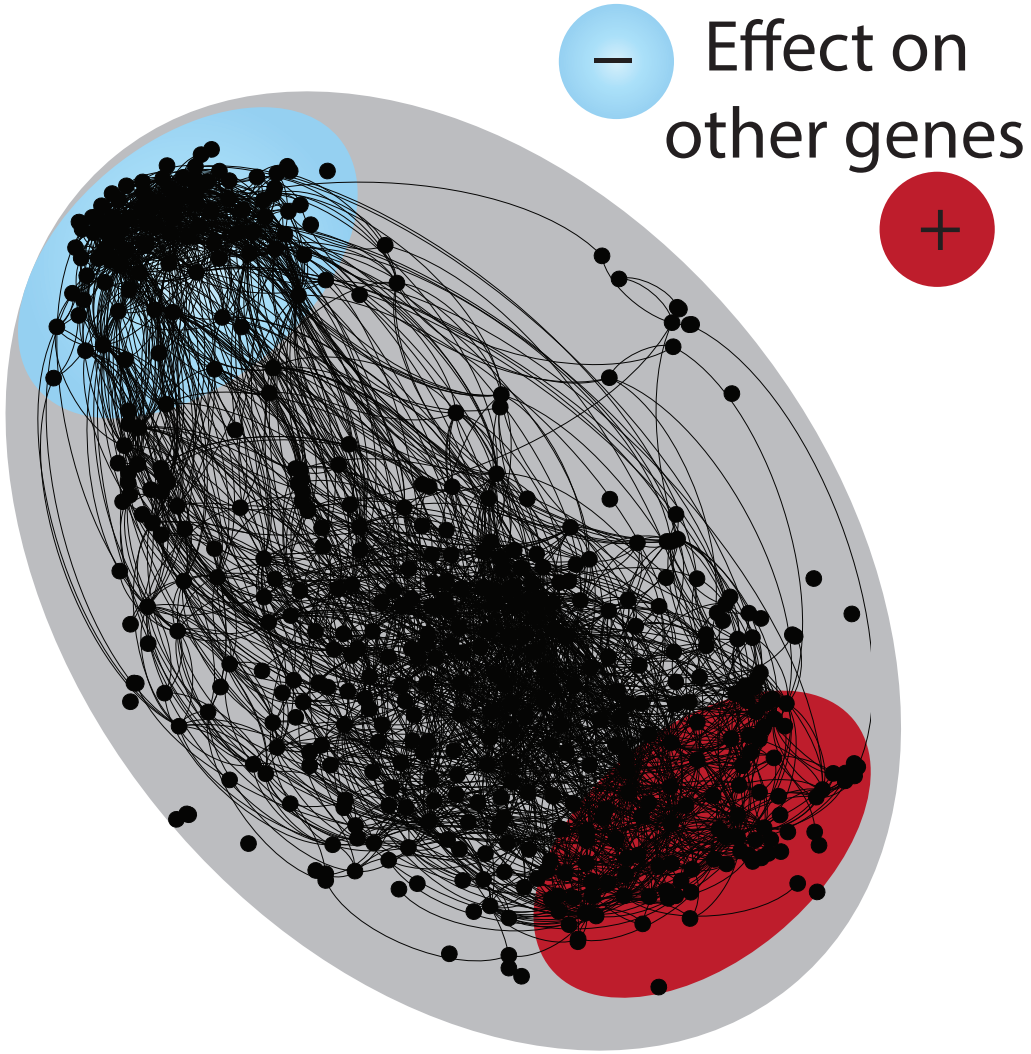


Figure 4.19: Generator-derived gene-gene association network (global view).
Network representation of the correlation between discriminator neural network final layer values for 6605 genes (centre panel); genes are depicted as nodes and correlations between them are shown as edges (threshold final layer correlation: 0.5). Network regions containing genes with overall positive and negative effect on other genes are highlighted in red and blue respectively.

observed a similar pattern of enriched associated genes for Ivl, an IFE differentiation marker which is co-associated with several other differentiation genes and Rps29 a marker of IFE-derived cells which is exclusively associated with ribosomal genes in our local networks. Furthermore, the local network for Notch ligand Jag2 contains several hair follicle bulge and basal IFE regulators, corresponding to its role in these epidermal locations (Powell et al., 1998). Strikingly, Wnt3 is present in this local network indicative of Wnt-Notch signalling interplay (Estrach et al., 2006; Shi et al., 2015), which cannot be seen at the RNA level using conventional gene expression analysis. These co-regulated epidermal genes are robustly identified despite little correlation of expression in the scRNA-seq data.

Finally, in order to explore the predictive power of our neural network approach we derived the local gene association network for the GATA-binding factor 6 transcription factor, GATA6 (Figure 4.21a). We predicted that as a transcription factor, the GATA6 network would contain many genes which are directly regulated by GATA6. Hence, we hypothesised that genes in the network with a positive final layer correlation should be upregulated in GATA6^{+} cells. Using a previous dataset from the lab, we contrasted bulk gene expression of GATA6^{+} and GATA6^{-} cells derived from the junctional zone and sebaceous duct of the epidermis (Figure 4.21a) (Donati et al., 2017). Figure 4.21c shows the log fold-change distribution of genes between the GATA6^{+}/GATA6^{-} subpopulations. Genes derived from our GATA6 gene regulatory network show significantly higher expression in the GATA6^{+} cells when compared to all other genes ($p < 0.05$, Kolmogorov-Smirnov test). Taken together these results suggest GANs are

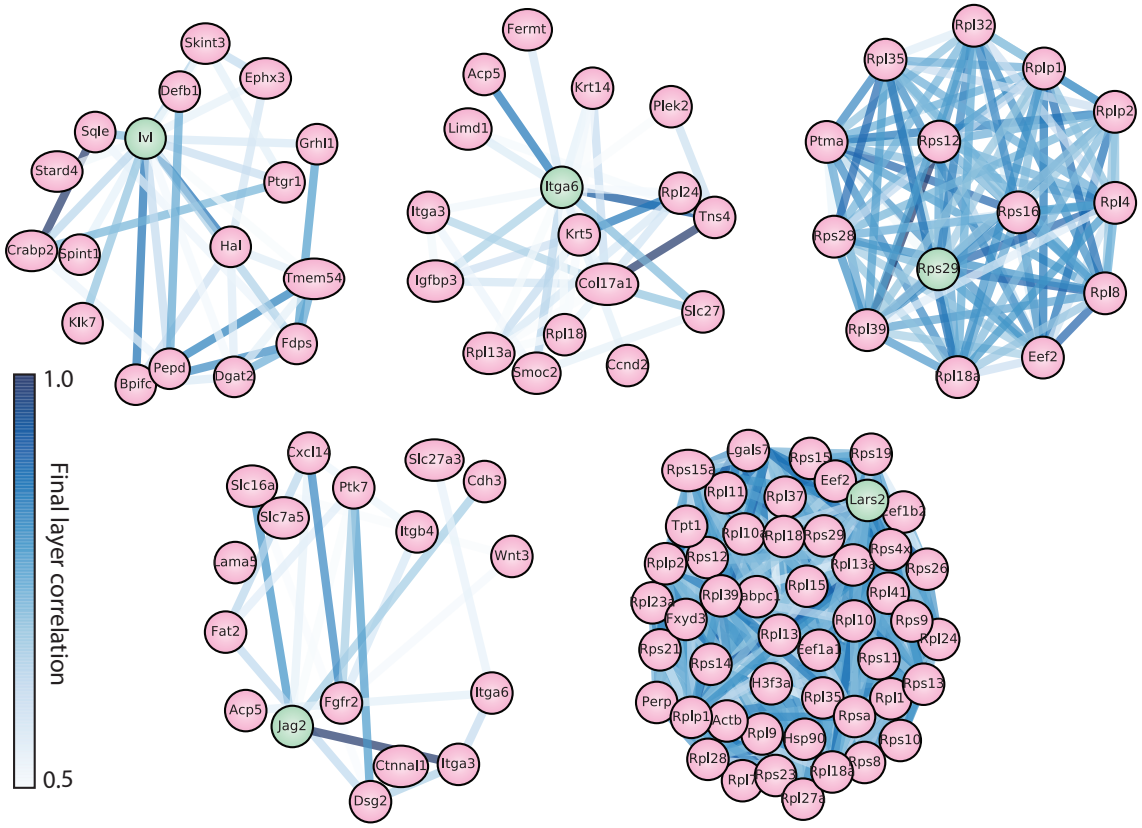


Figure 4.20: Generator-derived gene-gene association network (local view).
Local gene association networks for Lars2, Itga6, Ivl, Rps29 and Jag2; edges are colored by generator final layer correlation. Genes are depicted as nodes and correlations between them are shown as edges (threshold final layer correlation: 0.5).

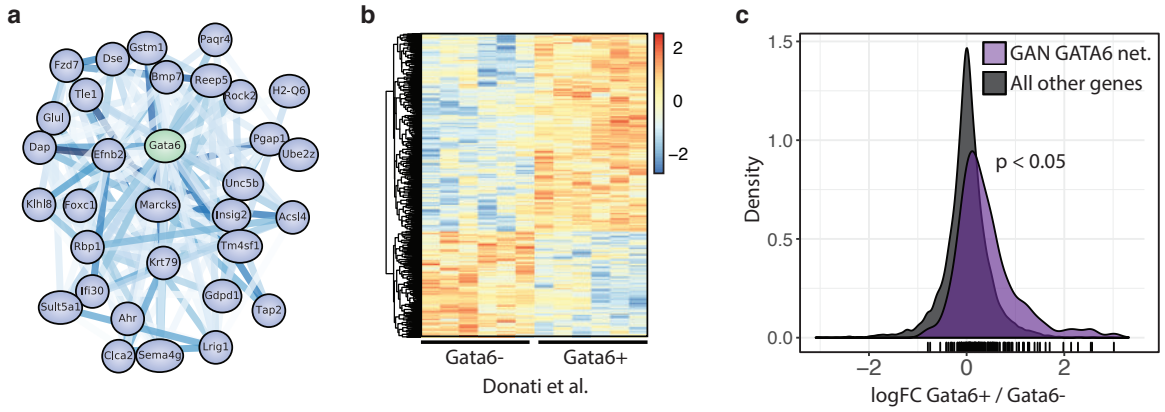


Figure 4.21: Generator-derived gene-gene association network for Gata6.
(a) Local gene association network for Gata6 displayed as described in Figure 4.20. (b) Heatmap showing differentially expressed gene expression z-scores for Gata6+ and Gata6- junctional zone and sebaceous duct cells from Donati et al. (c) Probability density plot of the log-fold change in gene expression between Gata6+ and Gata6- cells for genes in the predicted GATA6 local association network (violet) and all other genes (grey). Network genes display greater expression changes between Gata6+ and Gata6- cells (pvalue = 1.3e-4).

capable of elucidating complex gene interrelationships beyond the limits of linear correlation analyses.

4.3 Conclusions

We have addressed a major goal of single cell gene expression analysis; the desire to obtain functional gene relationships and a predictive model of transcriptional state in single cells. This serves as a framework for understanding cell fate transitions, disease-causing perturbations and state-determining genes across tissue types.

Here we demonstrate the effectiveness of a new deep learning approach in integrating multiple gene expression datasets, despite no prior adjustment for technical and batch-to-batch variation. To the best of our knowledge, this is the first time GANs have

been applied to genomic data. Our approach of combining data from cells in different environments with diverse technical and biological noise allows robust inference of gene associations intrinsic to epidermal cells rather than experimental conditions. We focused on skin due to the wealth of knowledge available regarding subpopulations and cell state. However, our generative model can be extended to integrate cells from multiple tissues and we anticipate this will advance our understanding of gene expression relationships across tissues.

As further scRNA-seq data becomes available, particularly through large scale projects such as the Human Cell Atlas (Rozenblatt-Rosen et al., 2017) we envisage that GANs will be a viable strategy to analyse all human cell types en masse. In our study we have shown that a relatively small dataset of under 2000 cells is sufficient to uncover state-determining genes and gene expression networks in the epidermis. GANs applied to a greater dataset covering multiple tissues could be used to resolve organism-wide gene expression relationships and to predict previously unseen cell state transitions.

4.4 Methods

4.4.1 Deep learning and neural networks

Datasets and data preparation

All datasets analysed in this chapter are publicly available under the GEO accessions GSE90848, GSE67602 and GSE99989. We combined these three datasets using transcripts per million (TPM) normalisation and removed labelled non-epithelial cells from the Joost dataset. Data was filtered for: cells with more than 1000 genes detected $\log_2(\text{TPM}+1) > 1$ and genes expressed in more than 500 cells at $\log_2(\text{TPM}+1) > 1$. All neural network input data was $\log_2(\text{TPM}+1)$ and the combined dataset is available as a CSV matrix at https://github.com/luslab/scRNAseq-WGAN-GP/tree/master/data.

Tools and code availability

Data preparation was performed using R. Gene ontology enrichment was evaluated using enrichR (Kuleshov et al., 2016). All other analysis and implementation of the generative adversarial networks was performed using Python (numpy, sklearn and networkx). We used Google's Tensorflow (Abadi et al., 2016) deep learning framework to implement, train and monitor our neural networks. We have provided a Jupyter Notebook containing our implementation of the generative adversarial network available at https://github.com/luslab/scRNAseq-WGAN-GP.

Generative adversarial network algorithm

We adapted our generative adversarial network algorithm from four previous works (Arjovsky et al., 2017; Gulrajani et al., 2017; Goodfellow et al., 2014b; Radford et al., 2015). Both generator and discriminators are fully connected neural networks with one hidden layer. We used a Leaky Rectified Linear Unit (LReLU) as the activation function for both networks using a coefficient of 0.2.

The generator input consists of a 100-dimensional latent variable (z) and a hidden layer size of 600 fully connected units. In order to reduce training time and data required for the generator network to simulate the Poisson distributed nature of scRNA-seq count data we trained the generator network using an additive Poisson and Gaussian distributed latent variable where $z = N(\mu = 0, \sigma = 0.1 * max) + P(\lambda = 1)$. Generator output is a 6605 dimensional vector representing the 6605 genes in the training cohort.

The discriminator network input is a 6605 dimensional vector (gene expression profile) and has a hidden layer size of 200 fully connected units. The final layer of the discriminator does not apply an activation function, in line with other Wasserstein GANs using a gradient penalty loss function.

We used backpropagation and RMSProp to train the two neural networks (learning rate of 5e-5) and a batch size of 32 cells. Additionally, data were augmented during training by randomly permuting the expression of 10 genes expressed at $\log_2$(TPM+1) < 3. We used several loss functions during this study before finalising on using the

Wasserstein-GAN (WGAN) with gradient penalty loss function. WGAN replaces the original GAN where data distributions of generated and real data are compared using the Jensen-Shannon divergence for the loss function (Goodfellow et al., 2014b). In WGANs the loss function is calculated using the Wasserstein distance which gives improved training stability (Arjovsky et al., 2017). Gulrajani and colleagues further improved on this variant of GANs by eliminating weight-clipping in favor of a gradient penalty in the form of the gradient between pairs of real and generated samples (Gulrajani et al., 2017).

The two loss functions to minimise are:

$$L_{\text{Discriminator}} = E[D(G(z))] - E[D(x)] + \lambda E[(||\nabla D(x')||_2 - 1)^2]$$

$$L_{\text{Generator}} = -E[D(G(z))] ,$$

where G is the generator network and $D$is the discriminator network such that $G : z \rightarrow \hat{x}$, $z$is the generator latent variable, $\hat{x}$ is a generated gene expression profile, x is a real gene expression profile and x' is uniformly sampled between pairs of generated and real gene expression profiles.

Convergence of the loss function is not sufficient to evaluate completed training of GANs. For each timepoint we evaluated the diversity of cells produced by the generator by simulating 500 cells and calculating the median distance (Pearson), i.e. median cell-cell distance.

For a list of GAN parameters used see Appendix B.

4.4.2 Dimensionality reduction and clustering for the GAN

We evaluated clustering and structure of real and generated single cell RNA-seq profiles in three ways 1) PCA (principal component analysis), 2) t-distributed stochastic neighbour embedding (t-SNE) on data with the top two principal components removed. 3) t-SNE on the hidden layer output values of the discriminator network. For all t-SNEs we used a default perplexity parameter of 30.

Simulated time-series gene expression profiles were clustered by k-means clustering with 20 expected clusters. Clustered dynamic gene expression profiles were subsequently visualised and two pairs of clusters were manually merged where they were deemed to be visually similar.

4.4.3 Latent space mapping and interpolation

Terminally differentiated cells and basal IFE cells were sampled from the combined dataset on using labels from the original studies. In order to generate corresponding latent space variable z_{basal} and $z_{\text{differentiated}}$ for cells x_{basal} and $x_{\text{differentiated}}$ we randomly generated cell expression profiles until the correlation between x and the generated cell $G(z)$reached a threshold of 0.7. Figure 4.3 shows that a correlation of 0.7 between two cells is rare, hence, we assume that this threshold is sufficient to generate a gene ex-

pression profile representative of the original cell state. We next calculated the difference between the basal and differentiated cell in latent space i.e. $\delta = z_{\text{basal}} - z_{\text{differentiated}}$. This latent space representation of differentiation was used to simulate differentiation on a further unseen cell where we obtained $z_{\text{unseen}}$and calculated a predicted terminally differentiated latent space point $z_{\text{unseen differentiated}} = z_{\text{unseen}} + \delta$. To obtain simulated gene expression time series data we interpolated 1000 points between $z_{\text{unseen}}$and $z_{\text{unseen differentiated}}$and generated 1000 intermediate gene expression profiles from these.

4.4.4 Discriminator network sensitivity analysis

Contribution of genes to output and sensitivity of the discriminator network was performed by taking unseen cells and varying the output of each gene between the minimum and maximum observed across all cells. For important genes this results in a change of the discriminator output $D(x_p)$where $x_p$is a perturbed gene expression profile. We performed 100 linearly interpolated perturbations. Sensitivity curves were calculated by taking the mean normalised $D(x_p)$at each expression level as unseen cells have can differ in their baseline or unperturbed discriminator score $D(x)$. Max sensitivity gradient (see Figure 4.18) was calculated by taking the absolute value of maximum gradient of these curves.

4.4.5 Local gene association networks

Gene co-regulatory relationships were inferred by analysing the weights of the generator neural network final layer. The Pearson correlation matrix of the final layer was used as a network adjacency matrix. To construct association networks for a gene we selected all local genes with a final layer correlation > 0.5 and retained any connections within the local network also with final layer correlation > 0.5.

Chapter 5

Conclusions and Future Perspective

5.1 Key conclusions

In this thesis I aimed to understand how keratinocytes make long and short term cell fate decisions. I focused on two approaches. Firstly, I investigated perturbations to cell state through a signalling pathway known to regulate epidermal cell state in an autonomous and non-autonomous manner, the Wnt/β-catenin pathway. Secondly, using an *a priori* approach, I applied a new generative method to integrate single cell gene expression data and uncover gene regulatory relationships governing cell state.

In Chapter 2 I characterised Wnt activation using the ΔNβ-cateninER inducible activation system in order to investigate non-cell autonomous Wnt signalling in Chapter 3. I found differences in nuclear β-catenin abundance when Wnt signalling is activated transiently (1hr) or constitutively (24hrs). Unexpectedly, I also found that one

of the effects of autonomous Wnt activation is upregulation of mRNA-binding proteins and a change in intron retention.

In Chapter 3 I used the ΔNβ-cateninER system to dissect non-cell autonomous Wnt activation effects from the better studied autonomous activation effects. I started by applying single cell RNA-seq to examine the molecular heterogeneity of wild type keratinocytes *in vitro*. Next I used this as a reference to understand the effect of exposing a keratinocyte to a Wnt-activated neighbour. This analysis showed that Wnt^{+} cells can induce their neighbours to transition towards a more proliferative transcriptional state. Using a pseudo-ordering approach I reconstructed the transcriptional changes occurring in this transition and determined the responsible transcription factors. In addition, single cell analysis identified a perturbation in protein-synthesis associated genes. To validate this finding I developed a high-throughput neighbour-cell image analysis method combined with a protein synthesis assay to show that neighbours of Wnt^{+} cells increase in translational activity.

Finally, in Chapter 4 I developed a new generative method for the analysis of single cell RNA-seq data. Using generative adversarial neural networks I have shown it is possible to synthesise scRNA-seq data indistinguishable from real data. As a consequence of using a generative approach I was able to simulate differentiation in single epidermal cells and predict gene expression changes at high time resolution. In addition this method predicts gene-gene regulatory relationships in an unsupervised manner, which I validated by examining genes perturbed when the transcription factor Gata6 is expressed.

5.2 Future directions and open questions

Beyond the aims of this thesis, several exciting questions remain open and below I highlight potential directions for future research.

5.2.1 Role of intron retention in Wnt signalling

Intron retention is an emerging field and its role in epidermal transcriptional regulation, studied in Chapter 2, leaves several avenues for future research (J-L Wong et al.). From this work, the mechanism of intron retention downstream of Wnt signalling is currently unclear. One hypothesis is indirect regulation of intron retention by activation of splicing regulators downstream of beta-catenin activation. In Chapter 2 , several RNA-binding proteins were observed to be upregulated after Wnt activation. Future work could focus on investigating the role of these RNA-binding proteins by selective knock-down (e.g. siRNA-mediated downregulation) of these splicing components and observing the resultant effect on intron retention under Wnt activated conditions.

An alternative hypothesis is direct regulation of intron retention by beta-catenin. Within the nucleus beta-catenin is primarily thought to bind with TCF/LEF transcription factors. However, it is possible that beta-catenin can also directly bind RNA-binding proteins and hence regulate intron retention directly. A future study could perform mass spectrometry of beta-catenin in complex with its nuclear partners to

characterise whether this form of regulation is present.

5.2.2 Mechanism for transduction of non-cell autonomous Wnt activation

Non-cell autonomous Wnt signalling clearly regulates the state of neighbouring cells. In Chapter 3 I showed that NCA Wnt effects are abrogated by culturing keratinocytes in low-calcium medium and hence likely to be dependent on desmosome (or other intercellular junctions) mediated cell-cell contact. Although downstream activated transcription factors were identified, I did not identify the molecular basis for transduction of the NCA Wnt signal from autonomously activated cell to neighbouring cell. It is of interest to identify the molecular mechanism of this signal transduction to provide further insight into Wnt^{+} stem cell niches. To investigate this further it may be necessary to selectively knock-down components of the desmosome and observe the effect on NCA Wnt signalling. Furthermore, profiling autonomously activated cells for a range of cell surface proteins using mass cytometry (e.g. CyTOF, Fluidigm) could provide clues regarding the molecular mechanism.

5.2.3 Further application of GANs

Additional resources should be placed into applying GANs to two immediate applications. Firstly, GANs should be applied to upcoming large-scale single cell projects

such as the Human Cell Atlas (Rozenblatt-Rosen et al., 2017). Futher details regarding the project can be found at https://www.humancellatlas.org/. My work demonstrates that GANs are able to integrate diverse datasets. Projects such as the Human Cell Atlas have the long term goal of acquiring single cell gene expression data for all human cell types. It would be interesting to attempt to create a generative model capable of synthesising scRNA-seq expression profiles from all human cell types. If successful, such a model could be used to explore all possible cell state transitions and examine the similarities among cell states in different tissues. For example, both the hair follicle and intestinal stem cell niches share molecular features such as $Lgr5^{+}$ stem cells (Haegebarth and Clevers, 2009). Using an all-cell integrative approach, it would be valuable to determine whether there is a common gene regulatory network in these and other tissues.

Secondly, GANs should be applied to additional single cell data types. In particular, single cell chromatin accessibility (scATAC-seq, Buenrostro et al. (2015)) and methylation (Farlik et al., 2015) are two areas where there is an increasing amount of data available for neural network training. Furthermore, both of these datatypes share many characteristics with scRNA-seq and are therefore more likely to be successfully analysed with a generative model.

5.2.4 Neural network structures incorporating gene properties

In Chapter 4 of this thesis I applied a fully connected neural network structure for both generator and discriminator neural networks. In these neural networks, all genes are connected to all other genes making them computationally expensive to use for large numbers of genes (> 20,000). Furthermore, in current applications of neural networks to genomic data there is no incorporation of prior knowledge regarding genes such as length, chromosome and proximity to other genes. In contrast, most neural networks applied to image data do incorporate prior information in the form of spatial relations between image pixels. Future research into neural network structures which incorporate known properties of genes and their interrelationships could improve their ability to model gene regulatory relationships and may require less data for training.

5.3 Concluding remarks

In this thesis I investigated several aspects of epidermal cell state and gene regulation in single cells. Chapter 3 focuses on one perturbation to epidermal cells (non-cell autonomous Wnt signaling) and how this affects heterogeneous subpopulations *in vitro*. This is an example of one specific induced cell state transition. Work presented in Chapter 4 extends this to incorporate multiple states and conditions of epidermal cells *in vitro* and *in vivo*. One current open question is the extent to which epidermal cells can transition between states. Epidermal cell differentiation is usually considered

to be a one-way process. However, recent evidence suggests that under comparatively rare circumstances, such as wounding, cells are able to dedifferentiate and contribute to wound repair (Donati et al., 2017).

An exciting future application of this work is to map all possible cell state transitions and to predict circumstances under which rare cell state transitions can occur. In the case of dedifferentiation of epidermal cells it may be possible to apply this information to modulate wound healing and repair *in vivo*.

Finally, although the focus of this thesis has been on gene expression, another exciting future prospect is multi-omic integration of single cell data. In particular, chromatic accessibility information (e.g. single cell ATAC-seq, Buenrostro et al. (2015)) and three-dimensional organisation of DNA in the nucleus (e.g. single cell Hi-C, Stevens et al. (2017)) are two additional aspects of transcriptional regulation which are distinct from gene expression. Integration of these data types with single cell RNA-seq using methods such as the GAN will further our understanding of cell state and transcriptional regulation of cell state transitions.

Appendix A

TFs regulating state A to state D NCA

Wnt transition

Below are transcription factors regulating the state A to state D transition relating to Chapter 3.

Rest, Smad1, Ctcf, Asxl1, Rcor2, Mef2a, Top2b, Nr3c1, Pbx1, Tbp, Rad21, Clock, Smarcd1, Rarb, Padi4, Bcl6, Elf1, Rela, Rcor3, Kdm6a, Foxo1, Mecom, Elk3, Rxra, Dcp1a, Cnot3, Ar, Klf5, Gata6, Brd4, Cux1, Gata3, Srf, Prdm5, Thap11, Chd1, Foxm1, Jund, Sox2, Hoxb4, Ccnd1, Hoxb7, Jun, Nanog, Yy1, Ppard, Sox9, Egr1, Ttf2, Mybl2, Tcf4, Sin3b, Atf3, Xrn2, Trim28, Vdr, Phf8, Klf4, E2f4, Kdm5b, E2f1, Smad4, Smad3, Bcl3

Figure A.1: Transcription factors regulating non-cell autonomous Wnt activation.

Appendix B

GAN training parameters

Below are parameters used for the generative adversarial network related to Chapter 4.

Table B.1: Parameters used for creating GAN neural networks and training

Parameter	**Value**
Batch size	30
Generator input size (z)	100
Generator layer 1 units	100
Generator layer 2 units	600
Generator layer 3 units	6605 (num. genes)
Generator initialised weights range	: -0.5 to 0.5
Generator initial learning rate	5.00E-05
Discriminator input size	6605 (num. genes)
Discriminator layer 1 units	6605 (num. genes)
Discriminator layer 2 units	200
Discriminator layer 3 units	1
Discriminator initialised weights range	: -0.05 to 0.05
Discriminator initial learning rate	5.00E-05
L2 regularisation scale	0.8

References

Mart\'\in Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. {TensorFlow}: {Large-Scale} Machine Learning on Heterogeneous Distributed Systems. mar 2016.

Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838, aug 2015. ISSN 1087-0156. doi: 10.1038/nbt.3300. URL http://www.nature.com/articles/nbt.3300.

Christof Angermueller, Heather J. Lee, Wolf Reik, and Oliver Stegle. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome*

Biology, 18(1):67, dec 2017a. ISSN 1474-760X. doi: 10.1186/s13059-017-1189-z. URL http://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1189-z.

Christof Angermueller, Heather J Lee, Wolf Reik, and Oliver Stegle. {DeepCpG}: accurate prediction of single-cell {DNA} methylation states using deep learning. *Genome Biol.*, 18(1):67, apr 2017b.

Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. FACE AGING WITH CONDITIONAL GENERATIVE ADVERSARIAL NETWORKS. URL https://arxiv.org/pdf/1702.01983.pdf.

L Arce, N N Yokoyama, and M L Waterman. Diversity of LEF/TCF action in development and disease. *Oncogene*, 25(57):7492–7504, dec 2006. ISSN 0950-9232. doi: 10.1038/sj.onc.1210056. URL http://www.ncbi.nlm.nih.gov/pubmed/17143293http://www.nature.com/articles/1210056.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein {GAN}. jan 2017.

Katrin Arnold, Abby Sarkar, Mary Anna Yram, Jose M Polo, Rod Bronson, Sumitra Sengupta, Marco Seandel, Niels Geijsen, and Konrad Hochedlinger. Sox2(+) adult stem and progenitor cells are important for tissue regeneration and survival of mice. *Cell Stem Cell*, 9(4):317–329, oct 2011.

F. A. Atcha, A. Syed, B. Wu, N. P. Hoverter, N. N. Yokoyama, J.-H. T. Ting, J. E. Munguia, H. J. Mangalam, J. L. Marsh, and M. L. Waterman. A Unique DNA Binding Domain Converts T-Cell Factors into Strong Wnt Effectors. *Molecular and Cellular Biology*, 27(23):8352–8363, dec 2007. ISSN 0270-7306. doi: 10.1128/MCB.02132-06.

URL http://www.ncbi.nlm.nih.gov/pubmed/17893322http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2169181http://mcb.asm.org/cgi/doi/10.1128/MCB.02132-06.

Rhonda Bacher, Li-Fang Chu, Ning Leng, Audrey P Gasch, James A Thomson, Ron M Stewart, Michael Newton, and Christina Kendziorski. SCnorm: robust normalization of single-cell RNA-seq data. *Nature Methods*, 14(6):584–586, apr 2017. ISSN 1548-7091. doi: 10.1038/nmeth.4263. URL http://www.nature.com/doifinder/10.1038/nmeth.4263.

Christopher M. Baker, Annemieke Verstuyf, Kim B. Jensen, and Fiona M. Watt. Differential sensitivity of epidermal cell subpopulations to β-catenin-induced ectopic hair follicle formation. *Developmental Biology*, 343(1-2):40–50, jul 2010. ISSN 00121606. doi: 10.1016/j.ydbio.2010.04.005. URL http://www.ncbi.nlm.nih.gov/pubmed/20398648http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3098388http://linkinghub.elsevier.com/retrieve/pii/S0012160610002356.

Y Barrandon and H Green. Three clonal types of keratinocyte with different capacities for multiplication. *Proc. Natl. Acad. Sci. U. S. A.*, 84(8):2302–2306, apr 1987. ISSN 0027-8424. URL http://www.ncbi.nlm.nih.gov/pubmed/2436229http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC304638.

S. C. Bendall, E. F. Simonds, P. Qiu, E.-a. D. Amir, P. O. Krutzik, R. Finck, R. V. Bruggner, R. Melamed, A. Trejo, O. I. Ornatsky, R. S. Balderas, S. K. Plevritis, K. Sachs,

D. Pe'er, S. D. Tanner, and G. P. Nolan. Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. *Science*, 332(6030):687–696, may 2011. ISSN 0036-8075. doi: 10.1126/science.1198704. URL http://www.ncbi.nlm.nih.gov/pubmed/21551058http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3273988http://www.sciencemag.org/cgi/doi/10.1126/science.1198704.

Y. Bengio, A. Courville, and P. Vincent. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, aug 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2013.50. URL http://www.ncbi.nlm.nih.gov/pubmed/23787338http://ieeexplore.ieee.org/document/6472238/.

David Berthelot, Thomas Schumm, and Luke Metz. BEGAN: Boundary Equilibrium Generative Adversarial Networks. URL https://arxiv.org/pdf/1703.10717.pdf.

Sandra Blanco, Roberto Bandiera, Martyna Popis, Shobbir Hussain, Patrick Lombard, Jelena Aleksic, Abdulrahim Sajini, Hinal Tanna, Rosana Cortés-Garrido, Nikoletta Gkatza, Sabine Dietmann, and Michaela Frye. Stem cell function and stress response are controlled by protein synthesis. *Nature*, 534(7607):335–340, jun 2016.

Cedric Blanpain, William E Lowry, Andrea Geoghegan, Lisa Polak, and Elaine Fuchs. Self-renewal, multipotency, and the existence of two cell populations within an epithelial stem cell niche. *Cell*, 118(5):635–648, sep 2004.

Véronique Bourdeau, Julie Deschênes, Raphaël Métivier, Yoshihiko Nagai, Denis Nguyen, Nancy Bretschneider, Frank Gannon, John H White, and Sylvie Mader. Genome-wide identification of high-affinity estrogen response elements in human and mouse. *Mol. Endocrinol.*, 18(6):1411–1427, jun 2004.

Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527, may 2016. ISSN 1087-0156. doi: 10.1038/nbt.3519. URL http://www.nature.com/articles/nbt.3519.

Jason D. Buenrostro, Beijing Wu, Ulrike M. Litzenburger, Dave Ruff, Michael L. Gonzales, Michael P. Snyder, Howard Y. Chang, and William J. Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523 (7561):486–490, jun 2015. ISSN 0028-0836. doi: 10.1038/nature14590. URL http://www.nature.com/doifinder/10.1038/nature14590.

Jason D Buenrostro, M Ryan Corces, Caleb A Lareau, Beijing Wu, Alicia N Schep, Martin J Aryee, Ravindra Majeti, Howard Y Chang, and William J Greenleaf. Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell*, 173(6):1535–1548.e16, may 2018. ISSN 1097-4172. doi: 10.1016/j.cell.2018.03.074. URL http://www.ncbi.nlm.nih.gov/pubmed/29706549.

Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies,

and species. *Nature Biotechnology*, 36(5):411–420, apr 2018. ISSN 1087-0156. doi: 10.1038/nbt.4096. URL http://www.nature.com/doifinder/10.1038/nbt.4096.

Ken M Cadigan and Marian L Waterman. TCF/LEFs and Wnt signaling in the nucleus. *Cold Spring Harbor perspectives in biology*, 4(11):a007906, nov 2012. ISSN 1943-0264. doi: 10.1101/cshperspect.a007906. URL http://www.ncbi.nlm.nih.gov/pubmed/23024173http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3536346.

Pablo G. Camara. Methods and challenges in the analysis of single-cell RNA-sequencing data. *Current Opinion in Systems Biology*, 7:47–53, feb 2018. ISSN 2452-3100. doi: 10.1016/J.COISB.2017.12.007. URL https://www.sciencedirect.com/science/article/pii/S2452310017301944{#}bib38.

Claudio Cantù, Pierfrancesco Pagella, Tania D Shajiei, Dario Zimmerli, Tomas Valenta, George Hausmann, Konrad Basler, and Thimios A Mitsiadis. A cytoplasmic role of Wnt/b-catenin transcriptional cofactors Bcl9, Bcl9l, and Pygopus in tooth enamel formation. *Science Signaling*, 10(465):eaah4598, feb 2017. ISSN 19379145. doi: 10.1126/scisignal.aah4598. URL http://www.ncbi.nlm.nih.gov/pubmed/28174279.

Edward F. Chan, Uri Gat, Jennifer M McNiff, and Elaine Fuchs. A common human skin tumour is caused by activating mutations in $"β$-catenin. *Nat. Genet.*, 21(4):410–413, apr 1999. ISSN 10614036. doi: 10.1038/7747. URL http://www.ncbi.nlm.nih.gov/pubmed/10192393http://www.nature.com/doifinder/10.1038/7747.

Sengthong Chanchevalap, Mandayam O Nandan, Didier Merlin, and Vincent W Yang. All-trans retinoic acid inhibits proliferation of intestinal epithelial cells by inhibiting expression of the gene encoding Krüppel-like factor 5. *FEBS Lett.*, 578(1-2): 99–105, dec 2004.

Pratip K Chattopadhyay, David A Price, Theresa F Harper, Michael R Betts, Joanne Yu, Emma Gostick, Stephen P Perfetto, Paul Goepfert, Richard A Koup, Stephen C De Rosa, Marcel P Bruchez, and Mario Roederer. Quantum dot semiconductor nanocrystals for immunophenotyping by polychromatic flow cytometry. *Nature Medicine*, 12(8):972–977, aug 2006. ISSN 1078-8956. doi: 10.1038/nm1371. URL http://www.nature.com/articles/nm1371.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. {InfoGAN}: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. jun 2016.

J Cheng, A J Syder, Q C Yu, A Letai, A S Paller, and E Fuchs. The genetic basis of epidermolytic hyperkeratosis: a disorder of differentiation-specific epidermal keratin genes. *Cell*, 70(5):811–819, sep 1992.

Yeon Sook Sook Choi, Yuhang Zhang, Mingang Xu, Yongguang Yang, Mayumi Ito, Tien Peng, Zheng Cui, Andras Nagy, Anna-Katerina Katerina Hadjantonakis, Richard A. a. Lang, George Cotsarelis, Thomas Andl, Edward E. E. Morrisey, and Sarah E. E. Millar. Distinct functions for {Wnt/$"β$-catenin} in hair follicle stem cell proliferation and survival and interfollicular epider-

mal homeostasis. *Cell Stem Cell*, 13(6):720–733, dec 2013. ISSN 19345909. doi: 10.1016/j.stem.2013.10.003. URL http://dx.doi.org/10.1016/j.stem.2013.10.003http://www.ncbi.nlm.nih.gov/pubmed/24315444http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3900235http://linkinghub.elsevier.com/retrieve/pii/S1934590913004499.

Wilson M Clements, Andrew M Lowy, and Joanna Groden. Adenomatous polyposis coli/beta-catenin interaction and downstream targets: altered gene expression in gastrointestinal tumors. *Clinical colorectal cancer*, 3(2):113–20, aug 2003. ISSN 1533-0028. URL http://www.ncbi.nlm.nih.gov/pubmed/12952568.

Charlotte A Collins, Kai Kretzschmar, and Fiona M Watt. Reprogramming adult dermis to a neonatal state through epidermal activation of $"β$-catenin. *Development*, 138(23):5189–5199, dec 2011.

George Cotsarelis, Tung-Tien Sun, and Robert M. Lavker. Label-retaining cells reside in the bulge area of pilosebaceous unit: Implications for follicular stem cells, hair cycle, and skin carcinogenesis. *Cell*, 61(7):1329–1337, jun 1990. ISSN 0092-8674. doi: 10.1016/0092-8674(90)90696-C. URL https://www.sciencedirect.com/science/article/pii/009286749090696C?via{%}3Dihub.

Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative Adversarial Networks: An Overview. oct 2017.

R DasGupta and E Fuchs. Multiple roles for activated LEF/TCF transcription complexes during hair follicle development and differentiation. *Development (Cambridge, England)*, 126(20):4557–4568, 1999. ISSN 0950-1991.

Elizabeth R Deschene, Peggy Myung, Panteleimon Rompolas, Giovanni Zito, Thomas Yang Sun, Makoto M Taketo, Ichiko Saotome, and Valentina Greco. {$"β$-Catenin} activation regulates tissue growth non-cell autonomously in the hair stem cell niche. *Science*, 343(6177):1353–1356, mar 2014.

Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, jan 2013. ISSN 1460-2059. doi: 10.1093/bioinformatics/bts635. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts635.

Giacomo Donati, Valentina Proserpio, Beate Maria Lichtenberger, Ken Natsuga, Rodney Sinclair, Hironobu Fujiwara, and Fiona M Watt. Epidermal Wnt/β-catenin signaling regulates adipocyte differentiation via secretion of adipogenic factors. *Proceedings of the National Academy of Sciences of the United States of America*, 111(15):E1501–9, apr 2014. ISSN 1091-6490. doi: 10.1073/pnas.1312880111. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3992657{&}tool=pmcentrez{&}rendertype=abstracthttp://www.ncbi.nlm.nih.gov/pubmed/24706781http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3992657.

Giacomo Donati, Emanuel Rognoni, Toru Hiratsuka, Kifayathullah Liakath-Ali, Esther Hoste, Gozde Kar, Melis Kayikci, Roslin Russell, Kai Kretzschmar, Klaas W Mulder, Sarah A Teichmann, and Fiona M Watt. Wounding induces dedifferentiation of epidermal Gata6(+) cells and acquisition of stem cell properties. *Nat. Cell Biol.*, 19(6): 603–613, jun 2017.

Ryan R Driskell, Adam Giangreco, Kim B Jensen, Klaas W Mulder, and Fiona M Watt. Sox2-positive dermal papilla cells specify hair follicle type in mammalian epidermis. *Development*, 136(16):2815–2823, aug 2009.

Ben W. Dulken, Dena S. Leeman, Stéphane C. Boutet, Katja Hebestreit, and Anne Brunet. Single-Cell Transcriptomic Analysis Defines Heterogeneity and Transcriptional Dynamics in the Adult Neural Stem Cell Lineage. *Cell Reports*, 18(3):777–790, jan 2017. ISSN 2211-1247. doi: 10.1016/J.CELREP.2016.12.060. URL https://www.sciencedirect.com/science/article/pii/S2211124716317673.

J Eberwine, H Yeh, K Miyashiro, Y Cao, S Nair, R Finnell, M Zettel, and P Coleman. Analysis of gene expression in single live neurons. *Proceedings of the National Academy of Sciences of the United States of America*, 89(7):3010–4, apr 1992. ISSN 0027-8424. URL http://www.ncbi.nlm.nih.gov/pubmed/1557406http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC48793.

M Ester, H P Kriegel, J Sander, and X Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD*, 1996.

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, feb 2017. ISSN 0028-0836. doi: 10.1038/nature21056. URL http://www.nature.com/articles/nature21056.

Soline Estrach, Carrie A Ambler, Cristina Lo Celso, Katsuto Hozumi, and Fiona M Watt. Jagged 1 is a beta-catenin target gene required for ectopic hair follicle formation in adult epidermis. *Development*, 133(22):4427–4438, nov 2006.

Matthias Farlik, Nathan C Sheffield, Angelo Nuzzo, Paul Datlinger, Andreas Schönegger, Johanna Klughammer, and Christoph Bock. Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell reports*, 10(8):1386–97, mar 2015. ISSN 2211-1247. doi: 10.1016/j.celrep.2015.02.001. URL http://www.ncbi.nlm.nih.gov/pubmed/25732828http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4542311.

A M Femino, F S Fay, K Fogarty, and R H Singer. Visualization of single RNA transcripts in situ. *Science (New York, N.Y.)*, 280(5363):585–90, apr 1998. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/9554849.

Emma C. Ferber, Mihoko Kajita, Anthony Wadlow, Lara Tobiansky, Carien Niessen, Hiroyoshi Ariga, Juliet Daniel, and Yasuyuki Fujita. A Role for the Cleaved Cytoplasmic Domain of E-cadherin in the Nucleus. *Journal of Biological Chemistry*, 283(19):12691–12700, may 2008. ISSN 0021-9258. doi: 10.1074/jbc.M708887200. URL http://www.ncbi.nlm.nih.gov/pubmed/18356166http://www.

pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2442316http://www.jbc.org/lookup/doi/10.1074/jbc.M708887200.

E Fuchs, R A Esteves, and P A Coulombe. Transgenic mice expressing a mutant keratin 10 gene reveal the likely genetic basis for epidermolytic hyperkeratosis. *Proc. Natl. Acad. Sci. U. S. A.*, 89(15):6906–6910, aug 1992.

Elaine Fuchs. Scratching the surface of skin development. *Nature*, 445(7130): 834–42, feb 2007. ISSN 1476-4687. doi: 10.1038/nature05659. URL http://www.ncbi.nlm.nih.gov/pubmed/17314969http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2405926.

Elaine Fuchs and Howard Green. Changes in keratin gene expression during terminal differentiation of the keratinocyte. *Cell*, 19(4):1033–1042, apr 1980. ISSN 00928674. doi: 10.1016/0092-8674(80)90094-X. URL http://www.ncbi.nlm.nih.gov/pubmed/6155214.

Emilios Gemenetzidis, Daniela Elena-Costea, Eric K Parkinson, Ahmad Waseem, Hong Wan, and Muy-Teck Teh. Induction of human epithelial stem/progenitor expansion by {FOXM1}. *Cancer Res.*, 70(22):9515–9526, nov 2010.

Arsham Ghahramani, Giacomo Donati, Nicholas M Luscombe, and Fiona M Watt. Epidermal Wnt signalling regulates transcriptome heterogeneity and proliferative fate in neighbouring cells. *Genome Biol.*, 19(1):3, jan 2018.

Eva Gó Mez-Orte, Beatriz Sá Enz-Narciso, Sergio Moreno, and Juan Cabello. Multiple functions of the noncanonical Wnt pathway. *Trends in Genetics*, 29:545–553, 2013.

doi: 10.1016/j.tig.2013.06.003. URL https://www.cell.com/trends/genetics/pdf/S0168-9525(13)00092-9.pdf.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. pages 2672–2680, 2014a. URL https://papers.nips.cc/paper/5423-generative-adversarial-nets.pdfhttp://papers.nips.cc/paper/5423-generative-adversarial-nets.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. jun 2014b.

Dominic Grün, Anna Lyubimova, Lennart Kester, Kay Wiebrands, Onur Basak, Nobuo Sasaki, Hans Clevers, and Alexander van Oudenaarden. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, 525(7568):251–255, sep 2015. ISSN 0028-0836. doi: 10.1038/nature14966. URL http://www.ncbi.nlm.nih.gov/pubmed/26287467http://www.nature.com/articles/nature14966.

Lei Gu, Sandra C Frommel, Christopher C Oakes, Ronald Simon, Katharina Grupp, Cristina Y Gerig, Dominik Bär, Mark D Robinson, Constance Baer, Melanie Weiss, Zuguang Gu, Matthieu Schapira, Ruprecht Kuner, Holger Sültmann, Maurizio Provenzano, ICGC Project on Early Onset Prostate Cancer, Marie-Laure Yaspo, Benedikt Brors, Jan Korbel, Thorsten Schlomm, Guido Sauter, Roland Eils, Christoph Plass, and Raffaella Santoro. {BAZ2A} ({TIP5}) is involved in epigenetic alterations

in prostate cancer and its overexpression predicts disease recurrence. *Nat. Genet.*, 47 (1):22–30, jan 2015.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved Training of Wasserstein {GANs}. *Advances in Neural Information Processing Systems 30*, pages 5763–5773, mar 2017.

Guoji Guo, Luca Pinello, Xiaoping Han, Shujing Lai, Li Shen, Ta-Wei Lin, Keyong Zou, Guo-Cheng Yuan, and Stuart H Orkin. {Serum-Based} Culture Conditions Provoke Gene Expression Variability in Mouse Embryonic Stem Cells as Revealed by {Single-Cell} Analysis. *Cell Rep.*, 14(4):956–965, feb 2016.

Minzhe Guo, Hui Wang, S. Steven Potter, Jeffrey A. Whitsett, and Yan Xu. SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis. *PLOS Computational Biology*, 11(11):e1004575, nov 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004575. URL http://dx.plos.org/10.1371/journal.pcbi.1004575.

Andrea Haegebarth and Hans Clevers. Wnt Signaling, Lgr5, and Stem Cells in the Intestine and Skin. *The American Journal of Pathology*, 174 (3):715–721, mar 2009. ISSN 00029440. doi: 10.2353/ajpath.2009.080758. URL http://www.ncbi.nlm.nih.gov/pubmed/19197002http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2665733http://linkinghub.elsevier.com/retrieve/pii/S0002944010609327.

Laleh Haghverdi, Aaron T L Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest

neighbors. *Nature Biotechnology*, 36(5):421–427, apr 2018. ISSN 1087-0156. doi: 10.1038/nbt.4091. URL http://www.nature.com/doifinder/10.1038/nbt.4091.

Tamar Hashimshony, Florian Wagner, Noa Sher, and Itai Yanai. {CEL-Seq}: single-cell {RNA-Seq} by multiplexed linear amplification. *Cell Rep.*, 2(3):666–673, sep 2012.

H Hennings and K A Holbrook. Calcium regulation of cell-cell contact and differentiation of epidermal cells in culture. An ultrastructural study. *Exp. Cell Res.*, 143(1): 127–142, jan 1983.

Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-Rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury. Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE SIGNAL PROCESSING MAGAZINE Digital Object Identifier*, 10 (2), 2012. doi: 10.1109/MSP.2012.2205597. URL https://www.cs.toronto.edu/{~}hinton/absps/DNN-2012-proof.pdf.

Katrin Hoffmeyer, Angelo Raggioli, Stefan Rudloff, Roman Anton, Andreas Hierholzer, Ignacio Del Valle, Kerstin Hein, Riana Vogt, and Rolf Kemler. Wnt/ -Catenin Signaling Regulates Telomerase in Stem Cells and Cancer Cells. *Science*, 336(6088):1549–54, jun 2012. ISSN 0036-8075. doi: 10.1126/science.1218370. URL http://www.ncbi.nlm.nih.gov/pubmed/22723415.

Ya-Chieh Hsu, Lishi Li, and Elaine Fuchs. Emerging interactions between skin stem cells and their niches. *Nat. Med.*, 20(8):847–856, aug 2014.

M Hu, D Krause, M Greaves, S Sharkis, M Dexter, C Heyworth, and T Enver. Multilineage gene expression precedes commitment in the hemopoietic system. *Genes Dev.*, 11(6):774–785, mar 1997.

J Huelsken, R Vogel, B Erdmann, G Cotsarelis, and W Birchmeier. beta-Catenin controls hair follicle morphogenesis and stem cell differentiation in the skin. *Cell*, 105 (4):533–45, may 2001. ISSN 0092-8674. URL http://www.ncbi.nlm.nih.gov/pubmed/11371349.

Saiful Islam, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian-Bing Fan, Peter Lönnerberg, and Sten Linnarsson. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome research*, 21 (7):1160–7, jul 2011. ISSN 1549-5469. doi: 10.1101/gr.110882.110. URL http://www.ncbi.nlm.nih.gov/pubmed/21543516http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3129258.

Mayumi Ito, Yaping Liu, Zaixin Yang, Jane Nguyen, Fan Liang, Rebecca J Morris, and George Cotsarelis. Stem cells in the hair follicle bulge contribute to wound repair but not to homeostasis of the epidermis. *Nature Medicine*, 11(12):1351–1354, dec 2005. ISSN 1078-8956. doi: 10.1038/nm1328. URL http://www.nature.com/articles/nm1328.

Justin J-L Wong, Amy Y M Au, William Ritchie, and John E J Rasko. Intron retention in mRNA: No longer nonsense; Intron retention in mRNA: No longer nonsense. doi:

10.1002/bies.201500117. URL https://onlinelibrary.wiley.com/doi/pdf/10.1002/bies.201500117.

Viljar Jaks, Nick Barker, Maria Kasper, Johan H van Es, Hugo J Snippert, Hans Clevers, and Rune Toftgård. Lgr5 marks cycling, yet long-lived, hair follicle stem cells. *Nature Genetics*, 40(11):1291–1299, nov 2008. ISSN 1061-4036. doi: 10.1038/ng.239. URL http://www.ncbi.nlm.nih.gov/pubmed/18849992http://www.nature.com/articles/ng.239.

Kim B Jensen, Ryan R Driskell, and Fiona M Watt. Assaying proliferation and differentiation capacity of stem cells using disaggregated adult mouse epidermis. *Nature Protocols*, 5(5):898–911, may 2010. ISSN 1754-2189. doi: 10.1038/nprot.2010.39. URL http://www.nature.com/articles/nprot.2010.39.

Eek-hoon Jho, Tong Zhang, Claire Domon, Choun-Ki Joo, Jean-Noel Freund, and Frank Costantini. Wnt/beta-catenin/Tcf signaling induces the transcription of Axin2, a negative regulator of the signaling pathway. *Molecular and cellular biology*, 22(4):1172–83, feb 2002. ISSN 0270-7306. URL http://www.ncbi.nlm.nih.gov/pubmed/11809808http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC134648.

P H Jones, S Harper, and F M Watt. Stem cell patterning and fate in human epidermis. *Cell*, 80(1):83–93, jan 1995. ISSN 0092-8674. URL http://www.ncbi.nlm.nih.gov/pubmed/7813021.

Philip H. Jones and Fiona M. Watt. Separation of human epidermal stem cells from transit amplifying cells on the basis of differences in integrin function and expression. *Cell*, 73(4):713–724, may 1993. ISSN 00928674. doi: 10.1016/0092-8674(93)90251-K. URL http://www.ncbi.nlm.nih.gov/pubmed/8500165https://www.sciencedirect.com/science/article/pii/009286749390251K.

Philip H. Jones, Benjamin D. Simons, and Fiona M. Watt. Sic Transit Gloria: Farewell to the Epidermal Transit Amplifying Cell? *Cell Stem Cell*, 1 (4):371–381, oct 2007. ISSN 19345909. doi: 10.1016/j.stem.2007.09.014. URL http://www.ncbi.nlm.nih.gov/pubmed/18371376http://linkinghub.elsevier.com/retrieve/pii/S1934590907001841.

Simon Joost, Amit Zeisel, Tina Jacob, Xiaoyan Sun, Gioele La Manno, Peter Lönnerberg, Sten Linnarsson, and Maria Kasper. Single-Cell Transcriptomics Reveals that Differentiation and Spatial Signatures Shape Epidermal and Hair Follicle Heterogeneity. *cels*, 3(3):221—-237.e9, sep 2016.

Satoshi Kakugawa, Paul F. Langton, Matthias Zebisch, Steven A. Howell, Tao-Hsin Chang, Yan Liu, Ten Feizi, Ganka Bineva, Nicola O'Reilly, Ambrosius P. Snijders, E. Yvonne Jones, and Jean-Paul Vincent. Notum deacylates Wnt proteins to suppress signalling activity. *Nature*, 519(7542):187–192, mar 2015. ISSN 0028-0836. doi: 10.1038/nature14259. URL http://www.nature.com/articles/nature14259.

Katannya Kapeli, Fernando J. Martinez, and Gene W. Yeo. Genetic mutations in RNA-binding proteins and their roles in ALS. *Human Genetics*,

136(9):1193–1214, sep 2017. ISSN 0340-6717. doi: 10.1007/s00439-017-1830-7. URL http://www.ncbi.nlm.nih.gov/pubmed/28762175http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5602095http://link.springer.com/10.1007/s00439-017-1830-7.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. oct 2017a. URL http://arxiv.org/abs/1710.10196.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive Growing of {GANs} for Improved Quality, Stability, and Variation. oct 2017b.

Kathryne Melissa Keays, Gregory P. Owens, Alanna M. Ritchie, Donald H. Gilden, and Mark P. Burgoon. Laser capture microdissection and single-cell RT-PCR without RNA purification. *Journal of Immunological Methods*, 302(1-2):90–98, jul 2005. ISSN 00221759. doi: 10.1016/j.jim.2005.04.018. URL http://www.ncbi.nlm.nih.gov/pubmed/16084216http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3279919http://linkinghub.elsevier.com/retrieve/pii/S0022175905001286.

Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7):740–742, jul 2014. ISSN 1548-7091. doi: 10.1038/nmeth.2967. URL http://www.nature.com/articles/nmeth.2967.

Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. {TopHat2}: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, 14 (4):R36, apr 2013. ISSN 1465-6914. doi: 10.1186/gb-2013-14-4-r36. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4053844{&}tool=pmcentrez{&}rendertype=abstract.

Nan-Hyung Kim, Soo-Hyun Choi, Tae Ryong Lee, Chang-Hoon Lee, and Ai-Young Lee. Cadherin 11, a miR-675 target, induces N-cadherin expression and epithelial-mesenchymal transition in melasma. *J. Invest. Dermatol.*, 134(12):2967–2976, dec 2014.

Vladimir Yu Kiselev, Kristina Kirschner, Michael T Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N Natarajan, Wolf Reik, Mauricio Barahona, Anthony R Green, and Martin Hemberg. SC3: consensus clustering of single-cell RNA-seq data. *Nature methods*, 14(5):483–486, may 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4236. URL http://www.ncbi.nlm.nih.gov/pubmed/28346451http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5410170.

Dennis D. Knutson. Ultrastructural Observations in Acne Vulgaris: the Normal Sebaceous Follicle and Acne Lesions. *Journal of Investigative Dermatology*, 62(3):288–307, mar 1974. ISSN 0022-202X. doi: 10.1111/1523-1747.EP12676804. URL https://www.sciencedirect.com/science/article/pii/S0022202X15442411.

Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C Marioni, and Sarah A Teichmann. The technology and biology of single-cell RNA se-

quencing. *Molecular cell*, 58(4):610–20, may 2015a. ISSN 1097-4164. doi: 10.1016/j.molcel.2015.04.005. URL http://www.ncbi.nlm.nih.gov/pubmed/26000846.

Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Jason C H Tsang, Tomislav Ilicic, Johan Henriksson, Kedar N Natarajan, Alex C Tuck, Xuefei Gao, Marc Bühler, Pentao Liu, John C Marioni, and Sarah A Teichmann. Single Cell {RNA-Sequencing} of Pluripotent States Unlocks Modular Transcriptional Variation. *Cell Stem Cell*, 17(4): 471–485, oct 2015b.

Kai Kretzschmar and Fiona M Watt. Markers of epidermal stem cell subpopulations in adult mammalian skin. *Cold Spring Harbor perspectives in medicine*, 4(10), jul 2014. ISSN 2157-1422. doi: 10.1101/cshperspect.a013631. URL http://www.ncbi.nlm.nih.gov/pubmed/24993676http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4200210.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. URL http://www.cs.toronto.edu/{~}fritz/absps/imagenet.pdf.

Maxim V Kuleshov, Matthew R Jones, Andrew D Rouillard, Nicolas F Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L Jenkins, Kathleen M Jagodnik, Alexander Lachmann, Michael G McDermott, Caroline D Monteiro, Gregory W Gundersen, and Avi Ma'ayan. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.*, 44(W1):W90—-7, jul 2016.

Arne Kusserow, Kevin Pang, Carsten Sturm, Martina Hrouda, Jan Lentfer, Heiko A. Schmidt, Ulrich Technau, Arndt von Haeseler, Bert Hobmayer, Mark Q. Martindale, and Thomas W. Holstein. Unexpected complexity of the Wnt gene family in a sea anemone. *Nature*, 433(7022):156–160, jan 2005. ISSN 0028-0836. doi: 10.1038/nature03158. URL http://www.nature.com/doifinder/10.1038/nature03158.

Magdalini Kypriotou, Marcel Huber, and Daniel Hohl. The human epidermal differentiation complex: cornified envelope precursors, {S100} proteins and the 'fused genes' family. *Exp. Dermatol.*, 21(9):643–649, sep 2012.

Alexander Lachmann, Huilei Xu, Jayanth Krishnan, Seth I Berger, Amin R Mazloom, and Avi Ma'ayan. {ChEA}: transcription factor regulation inferred from integrating genome-wide {ChIP-X} experiments. *Bioinformatics*, 26(19):2438–2444, oct 2010.

Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9(4):357–359, mar 2012.

R M Lavker and T T Sun. Heterogeneity in epidermal basal keratinocytes: morphological and functional correlations. *Science (New York, N.Y.)*, 215(4537):1239–41, mar 1982. ISSN 0036-8075. doi: 10.1126/SCIENCE.7058342. URL http://www.ncbi.nlm.nih.gov/pubmed/7058342.

Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. sep 2016. URL http://arxiv.org/abs/1609.04802.

Hee Kyu Lee, Yong Seok Choi, Young Ae Park, and Sunjoo Jeong. Modulation of oncogenic transcription and alternative splicing by beta-catenin and an RNA aptamer in colon cancer cells. *Cancer research*, 66(21):10560–6, nov 2006. ISSN 0008-5472. doi: 10.1158/0008-5472.CAN-06-2526. URL http://www.ncbi.nlm.nih.gov/pubmed/17079480.

Jacob H. Levine, Erin F. Simonds, Sean C. Bendall, Kara L. Davis, El-ad D. Amir, Michelle D. Tadmor, Oren Litvin, Harris G. Fienberg, Astraea Jager, Eli R. Zunder, Rachel Finck, Amanda L. Gedman, Ina Radtke, James R. Downing, Dana Pe'er, and Garry P. Nolan. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*, 162(1):184–197, jul 2015. ISSN 0092-8674. doi: 10.1016/J.CELL.2015.05.047. URL https://www.sciencedirect.com/science/article/pii/S0092867415006376.

Yang Liao, Gordon K. Smyth, and Wei Shi. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30 (7):923–930, apr 2014. ISSN 14602059. doi: 10.1093/bioinformatics/btt656.

Wen-Hui Hui Lien, Xingyi Guo, Lisa Polak, Lee N. Lawton, Richard A. Young, Deyou Zheng, and Elaine Fuchs. Genome-wide maps of histone modifications unwind in vivo chromatin states of the hair follicle lineage. *Cell Stem Cell*, 9(3):219–232, sep 2011. ISSN 19345909. doi: 10.1016/j.stem.2011.07.015. URL http://dx.doi.org/10.1016/j.stem.2011.07.015.

Xinhong Lim and Roel Nusse. Wnt signaling in skin development, homeostasis,

and disease. *Cold Spring Harb. Perspect. Biol.*, 5(2), feb 2013. ISSN 1943-0264. doi: 10.1101/cshperspect.a008029. URL http://www.ncbi.nlm.nih.gov/pubmed/23209129http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3552514.

Xinhong Lim, Si Hui Tan, Winston Lian Chye Koh, Rosanna Man Wah Chau, Kelley S. Yan, Calvin J. Kuo, Renée van Amerongen, Allon Moshe Klein, and Roel Nusse. Interfollicular epidermal stem cells self-renew via autocrine Wnt signaling. *Science*, 342(6163):1226–1230, dec 2013. ISSN 0036-8075. doi: 10.1126/science.1239730. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1239730.

Chieh Lin, Siddhartha Jain, Hannah Kim, and Ziv Bar-Joseph. Using neural networks for reducing the dimensions of single-cell {RNA-Seq} data. *Nucleic Acids Res.*, 45(17): e156, sep 2017.

T D Littlewood, D C Hancock, P S Danielian, M G Parker, and G I Evan. A modified oestrogen receptor ligand-binding domain as an improved switch for the regulation of heterologous proteins. *Nucleic acids research*, 23(10):1686–1690, 1995. ISSN 0305-1048. doi: 10.1093/nar/23.10.1686.

Cristina Lo Celso, David M Prowse, and Fiona M Watt. Transient activation of beta-catenin signalling in adult mouse epidermis is sufficient to induce new hair follicles but continuous activation is required to maintain hair follicle tumours. *Development*, 131(8):1787–1799, apr 2004. ISSN 0950-1991. doi: 10.1242/dev.01052.

Cristina Lo Celso, Melanie A. Berta, Kristin M. Braun, Michaela Frye, Stephen Lyle, Christos C. Zouboulis, and Fiona M. Watt. Characterization of bipotential epidermal progenitors derived from human sebaceous gland: contrasting roles of c-Myc and beta-catenin. *Stem Cells*, 26(5):1241–1252, may 2008. ISSN 1549-4918. doi: 10.1634/stemcells.2007-0651. URL http://www.ncbi.nlm.nih.gov/pubmed/18308950http://doi.wiley.com/10.1634/stemcells.2007-0651.

Vanessa Lopez-Pajares, Kun Qu, Jiajing Zhang, Dan E Webster, Brook C Barajas, Zurab Siprashvili, Brian J Zarnegar, Lisa D Boxer, Eon J Rios, Shiying Tao, Markus Kretz, and Paul A Khavari. A {LncRNA-MAF:MAFB} transcription factor network regulates epidermal differentiation. *Dev. Cell*, 32(6):693–706, mar 2015.

Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12):550, 2014. ISSN 1474-760X. doi: 10.1186/s13059-014-0550-8. URL http://www.ncbi.nlm.nih.gov/pubmed/25516281http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4302049.

Raphaelle Luisier, Giulia E. Tyzack, Claire E. Hall, Jamie S. Mitchell, Helen Devine, Doaa M. Taha, Bilal Malik, Ione Meyer, Linda Greensmith, Jia Newcombe, Jernej Ule, Nicholas M. Luscombe, and Rickie Patani. Intron retention and nuclear loss of SFPQ are molecular hallmarks of ALS. *Nature Communications*, 9(1):2010, dec 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-04373-8. URL http://www.nature.com/articles/s41467-018-04373-8.

Bryan T MacDonald, Keiko Tamai, and Xi He. Wnt/beta-catenin signaling: components, mechanisms, and diseases. *Developmental cell*, 17(1):9–26, jul 2009. ISSN 1878-1551. doi: 10.1016/j.devcel.2009.06.016. URL http://www.ncbi.nlm.nih.gov/pubmed/19619488http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2861485.

Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, John J Trombetta, David A Weitz, Joshua R Sanes, Alex K Shalek, Aviv Regev, and Steven A McCarroll. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214, may 2015. ISSN 1097-4172. doi: 10.1016/j.cell.2015.05.002. URL http://www.ncbi.nlm.nih.gov/pubmed/26000488http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4481139.

Ilaria Malanchi, Hector Peinado, Deepika Kassen, Thomas Hussenet, Daniel Metzger, Pierre Chambon, Marcel Huber, Daniel Hohl, Amparo Cano, Walter Birchmeier, and Joerg Huelsken. Cutaneous cancer stem cell maintenance is dependent on β-catenin signalling. *Nature*, 452(7187):650–653, apr 2008. ISSN 0028-0836. doi: 10.1038/nature06835. URL http://www.ncbi.nlm.nih.gov/pubmed/18385740http://www.nature.com/articles/nature06835.

L N Marekov and P M Steinert. Ceramides are bound to structural proteins of the human foreskin epidermal cornified cell envelope. *The Journal of biological chemistry*,

273(28):17763–70, jul 1998. ISSN 0021-9258. doi: 10.1074/JBC.273.28.17763. URL http://www.ncbi.nlm.nih.gov/pubmed/9651377.

Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12, may 2011. ISSN 2226-6089. doi: 10.14806/ej.17.1.200. URL http://journal.embnet.org/index.php/embnetjournal/article/view/200/479.

Celia Pilar Martinez-Jimenez, Nils Eling, Hung-Chang Chen, Catalina A Vallejos, Aleksandra A Kolodziejczyk, Frances Connor, Lovorka Stojic, Timothy F Rayner, Michael J T Stubbington, Sarah A Teichmann, Maike de la Roche, John C Marioni, and Duncan T Odom. Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science (New York, N.Y.)*, 355 (6332):1433–1436, mar 2017. ISSN 1095-9203. doi: 10.1126/science.aah4115. URL http://www.ncbi.nlm.nih.gov/pubmed/28360329http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5405862.

Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133, dec 1943. ISSN 0007-4985. doi: 10.1007/BF02478259. URL http://link.springer.com/10.1007/BF02478259.

J. A. McGrath and J. Uitto. *Anatomy and Organization of Human Skin*. Wiley-Blackwell, Oxford, UK, may 2010. ISBN 9781444317633. doi: 10.1002/9781444317633.ch3. URL http://doi.wiley.com/10.1002/9781444317633.ch3.

B. J. Merrill. Wnt Pathway Regulation of Embryonic Stem Cell Self-Renewal. *Cold Spring Harbor Perspectives in Biology*, 4(9):a007971–a007971, sep 2012. ISSN 1943-0264. doi: 10.1101/cshperspect.a007971. URL http://www.ncbi.nlm.nih.gov/pubmed/22952393http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3428775http://cshperspectives.cshlp.org/lookup/doi/10.1101/cshperspect.a007971.

Joëlle Michaud, Ken M Simpson, Robert Escher, Karine Buchet-Poyau, Tim Beissbarth, Catherine Carmichael, Matthew E Ritchie, Frédéric Schütz, Ping Cannon, Marjorie Liu, Xiaofeng Shen, Yoshiaki Ito, Wendy H Raskind, Marshall S Horwitz, Motomi Osato, David R Turner, Terence P Speed, Maria Kavallaris, Gordon K Smyth, and Hamish S Scott. Integrative analysis of {RUNX1} downstream pathways and target genes. *BMC Genomics*, 9:363, jul 2008.

Ajay Mishra, Bénédicte Oulès, Angela Oliveira Pisco, Tony Ly, Kifayathullah Liakath-Ali, Gernot Walko, Priyalakshmi Viswanathan, Matthieu Tihy, Jagdeesh Nijjher, Sara-Jane Dunn, Angus I Lamond, and Fiona M Watt. A protein phosphatase network controls the temporal and spatial dynamics of differentiation commitment in human epidermis. *Elife*, 6, oct 2017.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral Normalization for Generative Adversarial Networks. In *{ICML} Workshop on Implicit Models*, aug 2017. URL http://arxiv.org/abs/1802.05957.

R Molinuevo, A Freije, I de Pedro, S W Stoll, J T Elder, and A Gandarillas. {FOXM1}

allows human keratinocytes to bypass the oncogene-induced differentiation checkpoint in response to gain of {MYC} or loss of p53. *Oncogene*, 36(7):956–965, feb 2017.

R T Moon, J L Christian, R M Campbell, L L McGrew, A A DeMarais, M Torres, CHENG-JUNG Lai, D J Olson, and G M Kelly. Dissecting Wnt signalling pathways and Wnt-sensitive developmental processes through transient misexpression analyses in embryos of Xenopus laevis. *Development (Cambridge, England). Supplement*, pages 85–94, 1992.

Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7):621–628, jul 2008. ISSN 1548-7091. doi: 10.1038/nmeth.1226. URL http://www.ncbi.nlm.nih.gov/pubmed/18516045http://www.nature.com/articles/nmeth.1226.

Felix B Müller, Marcel Huber, Tamar Kinaciyan, Ingrid Hausser, Christina Schaffrath, Thomas Krieg, Daniel Hohl, Bernhard P Korge, and Meral J Arin. A human keratin 10 knockout causes recessive epidermolytic hyperkeratosis. *Hum. Mol. Genet.*, 15(7):1133–1141, apr 2006.

Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods*, 12(5):453–457, may 2015.

Catherin Niemann. Differentiation of the sebaceous gland. *Dermato-endocrinology*, 1(2):64–7, mar 2009. ISSN 1938-1980. URL http://www.ncbi.nlm.nih.gov/pubmed/20224685http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2835892.

Catherin Niemann, David M Owens, Jörg Hülsken, Walter Birchmeier, and Fiona M Watt. Expression of DeltaNLef1 in mouse epidermis results in differentiation of hair follicles into squamous epidermal cysts and formation of skin tumours. *Development (Cambridge, England)*, 129(1):95–109, jan 2002. ISSN 0950-1991. URL http://www.ncbi.nlm.nih.gov/pubmed/11782404.

E J O'Keefe, R A Briggaman, and B Herman. Calcium-induced assembly of adherens junctions in keratinocytes. *J. Cell Biol.*, 105(2):807–817, aug 1987.

Philip Owens, Hisham Bazzi, Erin Engelking, Gangwen Han, Angela M. Christiano, and Xiao-Jing Wang. Smad4-dependent desmoglein-4 expression contributes to hair follicle integrity. *Developmental Biology*, 322(1):156–166, oct 2008. ISSN 0012-1606. doi: 10.1016/J.YDBIO.2008.07.020. URL https://www.sciencedirect.com/science/article/pii/S0012160608010804?via{%}3Dihub.

Mahalia E Page, Patrick Lombard, Felicia Ng, Berthold Göttgens, and Kim B Jensen. The epidermis comprises autonomous compartments maintained by distinct stem cell populations. *Cell Stem Cell*, 13(4):471–482, oct 2013.

Mark Peifer, Li-Mei Pai, and Michael Casey. Phosphorylation of the Drosophila Adherens Junction Protein Armadillo: Roles for Wingless Signal and Zeste-white 3 Ki-

nase. *Developmental Biology*, 166(2):543–556, dec 1994. ISSN 00121606. doi: 10.1006/dbio.1994.1336. URL http://www.ncbi.nlm.nih.gov/pubmed/7529201http://linkinghub.elsevier.com/retrieve/pii/S0012160684713364.

Christian P. Petersen and Peter W. Reddien. Wnt Signaling and the Polarity of the Primary Body Axis. *Cell*, 139(6):1056–1068, dec 2009. ISSN 00928674. doi: 10.1016/j.cell.2009.11.035. URL http://www.ncbi.nlm.nih.gov/pubmed/20005801http://linkinghub.elsevier.com/retrieve/pii/S0092867409014937.

Viktor Petukhov, Jimin Guo, Ninib Baryawno, Nicolas Severe, David Scadden, Maria G. Samsonova, and Peter V. Kharchenko. Accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments. *bioRxiv*, page 171496, sep 2017. doi: 10.1101/171496. URL https://www.biorxiv.org/content/early/2017/09/13/171496.

Christina Philippeos, Stephanie B. Telerman, Bénédicte Oulès, Angela O. Pisco, Tanya J. Shaw, Raul Elgueta, Giovanna Lombardi, Ryan R. Driskell, Mark Soldin, Magnus D. Lynch, and Fiona M. Watt. Spatial and Single-Cell Transcriptional Profiling Identifies Functionally Distinct Human Dermal Fibroblast Subpopulations. *Journal of Investigative Dermatology*, 138(4):811–825, apr 2018. ISSN 0022202X. doi: 10.1016/j.jid.2018.01.016. URL http://www.ncbi.nlm.nih.gov/pubmed/29391249http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5869055https://linkinghub.elsevier.com/retrieve/pii/S0022202X18300368.

Simone Picelli, Omid R Faridani, Åsa K Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols*, 9(1):171–181, jan 2014. ISSN 1754-2189. doi: 10.1038/nprot.2014.006. URL http://www.nature.com/doifinder/10.1038/nprot.2014.006.

A M Pierce, I B Gimenez-Conti, R Schneider-Broussard, L A Martinez, C J Conti, and D G Johnson. Increased {E2F1} activity induces skin tumors in mice heterozygous and nullizygous for p53. *Proc. Natl. Acad. Sci. U. S. A.*, 95(15):8858–8863, jul 1998.

Emma Pierson and Christopher Yau. {ZIFA}: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.*, 16:241, nov 2015.

C S Potten. Cell replacement in epidermis (keratopoiesis) via discrete units of proliferation. *International review of cytology*, 69:271–318, 1981. ISSN 0074-7696. URL http://www.ncbi.nlm.nih.gov/pubmed/6163744.

B C Powell, E A Passmore, A Nesci, and S M Dunn. The Notch signalling pathway in hair growth. *Mech. Dev.*, 78(1-2):189–192, nov 1998.

Guo-Jun Qi. Loss-Sensitive Generative Adversarial Networks on Lipschitz Densities. jan 2017. URL http://arxiv.org/abs/1701.06264.

W Qiao, A G Li, P Owens, X Xu, X-J Wang, and C-X Deng. Hair follicle defects and squamous cell carcinoma formation in Smad4 conditional knockout mouse skin. *Oncogene*, 25(2):207–217, jan 2006. ISSN 0950-9232. doi: 10.1038/sj.onc.1209029. URL http://www.nature.com/articles/1209029.

Xiaojie Qiu, Qi Mao, Ying Tang, Li Wang, Raghav Chawla, Hannah A Pliner, and Cole Trapnell. Reversed graph embedding resolves complex single-cell trajectories. *Nature methods*, 14(10):979–982, oct 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4402. URL http://www.ncbi.nlm.nih.gov/pubmed/28825705http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5764547.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. nov 2015. URL http://arxiv.org/abs/1511.06434.

Bushra Raj, Daniel E Wagner, Aaron McKenna, Shristi Pandey, Allon M Klein, Jay Shendure, James A Gagnon, and Alexander F Schier. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nature Biotechnology*, 36(5):442–450, mar 2018. ISSN 1087-0156. doi: 10.1038/nbt.4103. URL http://www.nature.com/doifinder/10.1038/nbt.4103.

Daniel Ramsköld, Shujun Luo, Yu-Chieh Wang, Robin Li, Qiaolin Deng, Omid R Faridani, Gregory A Daniels, Irina Khrebtukova, Jeanne F Loring, Louise C Laurent, Gary P Schroth, and Rickard Sandberg. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology*, 30(8):777–782, aug 2012. ISSN 1087-0156. doi: 10.1038/nbt.2282. URL http://www.ncbi.nlm.nih.gov/pubmed/22820318http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3467340http://www.nature.com/articles/nbt.2282.

S Reddy, T Andl, a Bagasra, M M Lu, D J Epstein, E E Morrisey, and S E Millar. Characterization of Wnt gene expression in developing and postnatal hair follicles and identification of Wnt5a as a target of Sonic hedgehog in hair follicle morphogenesis. *Mechanisms of development*, 107(1-2):69–82, 2001. ISSN 09254773. doi: 10.1016/S0925-4773(01)00452-X.

Scott Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning What and Where to Draw. oct 2016a. URL http://arxiv.org/abs/1610.02454.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative Adversarial Text to Image Synthesis, jun 2016b. ISSN 1938-7228. URL http://proceedings.mlr.press/v48/reed16.html.

R H Rice and H Green. Presence in human epidermal cells of a soluble protein precursor of the cross-linked envelope: activation of the cross-linking by calcium ions. *Cell*, 18(3):681–94, nov 1979. ISSN 0092-8674. URL http://www.ncbi.nlm.nih.gov/pubmed/42494.

M. R. Romero, J. M. Caroll, and F. M. Watt. Analysis of cultured keratinocytes from a transgenic mouse model of psoriasis: effects of suprabasal integrin expression on keratinocyte adhesion, proliferation and terminal differentiation. *Experimental Dermatology*, 8(1):53–67, feb 1999. ISSN 0906-6705. doi: 10.1111/j.1600-0625.1999.tb00348.x. URL http://doi.wiley.com/10.1111/j.1600-0625.1999.tb00348.x.

Natali Romero-Barrios, Maria Florencia Legascue, Moussa Benhamed, Federico Ariel, and Martin Crespi. Splicing regulation by long noncoding RNAs. *Nucleic Acids Research*, 46(5):2169–2184, mar 2018. ISSN 0305-1048. doi: 10.1093/nar/gky095. URL https://academic.oup.com/nar/article/46/5/2169/4841659.

Panteleimon Rompolas, Kailin R Mesa, Kyogo Kawaguchi, Sangbum Park, David Gonzalez, Samara Brown, Jonathan Boucher, Allon M Klein, and Valentina Greco. Spatiotemporal coordination of stem cell commitment during epidermal homeostasis. *Science (New York, N.Y.)*, 352(6292):1471–4, jun 2016. ISSN 1095-9203. doi: 10.1126/science.aaf7012. URL http://www.ncbi.nlm.nih.gov/pubmed/27229141http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4958018.

Amit Roshan, Kasumi Murai, Joanna Fowler, Benjamin D Simons, Varvara Nikolaidou-Neokosmidou, and Philip H Jones. Human keratinocytes have two interconvertible modes of proliferation. *Nat. Cell Biol.*, 18(2):145–156, feb 2016.

Orit Rozenblatt-Rosen, Michael J T Stubbington, Aviv Regev, and Sarah A Teichmann. The Human Cell Atlas: from vision to reality. *Nature*, 550(7677):451–453, oct 2017.

C Ruhrberg, M A Hajibagheri, D A Parry, and F M Watt. Periplakin, a novel component of cornified envelopes and desmosomes that belongs to the plakin family and forms complexes with envoplakin. *The Journal of cell biology*, 139(7):1835–49, dec 1997. ISSN 0021-9525. URL http://www.ncbi.nlm.nih.gov/pubmed/9412476http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2132639.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 1986. ISSN 00280836. doi: 10.1038/323533a0.

Andreas Sagner, Zachary B. Gaber, Julien Delile, Jennifer H. Kong, David L. Rousso, Caroline A. Pearson, Steven E. Weicksel, Manuela Melchionda, S. Neda Mousavy Gharavy, James Briscoe, and Bennett G. Novitch. Olig2 and Hes regulatory dynamics during motor neuron differentiation revealed by single cell transcriptomics. *PLOS Biology*, 16(2):e2003127, feb 2018. ISSN 1545-7885. doi: 10.1371/journal.pbio.2003127. URL http://dx.plos.org/10.1371/journal.pbio.2003127.

Reyhaneh Salehi-Tabar, Loan Nguyen-Yamamoto, Luz E Tavera-Mendoza, Thomas Quail, Vassil Dimitrov, Beum-Soo An, Leon Glass, David Goltzman, and John H White. Vitamin {D} receptor as a master regulator of the {c-MYC/MXD1} network. *Proc. Natl. Acad. Sci. U. S. A.*, 109(46):18827–18832, nov 2012.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs. URL https://papers.nips.cc/paper/6125-improved-techniques-for-training-gans.pdf.

M Salto-Tellez, B K Peh, K Ito, S H Tan, P Y Chong, H C Han, K Tada, W Y Ong, R Soong, D C Voon, and Y Ito. RUNX3 protein is overexpressed in human basal cell carcinomas. *Oncogene*, 25(58):7646–7649, dec 2006. ISSN 0950-9232. doi: 10.1038/sj.onc.1209739. URL http://www.ncbi.nlm.nih.gov/pubmed/16767156http://www.nature.com/articles/1209739.

Raffaella Santoro, Junwei Li, and Ingrid Grummt. The nucleolar remodeling complex {NoRC} mediates heterochromatin formation and silencing of ribosomal gene transcription. *Nat. Genet.*, 32(3):393–396, nov 2002.

Jürgen Schweizer, Mitsuru Kinjo, Gerhard Fürstenberger, and Hermelita Winter. Sequential expression of mRNA-encoded keratin sets in neonatal mouse epidermis: Basal cells with properties of terminally differentiating cells. *Cell*, 37(1):159–170, may 1984. ISSN 00928674. doi: 10.1016/0092-8674(84)90311-8. URL https://www.sciencedirect.com/science/article/pii/0092867484903118.

Alex K. Shalek, Rahul Satija, Joe Shuga, John J. Trombetta, Dave Gennert, Diana Lu, Peilin Chen, Rona S. Gertner, Jellert T. Gaublomme, Nir Yosef, Schraga Schwartz, Brian Fowler, Suzanne Weaver, Jing Wang, Xiaohui Wang, Ruihua Ding, Raktima Raychowdhury, Nir Friedman, Nir Hacohen, Hongkun Park, Andrew P. May, and Aviv Regev. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, 510(7505):363–369, jun 2014a. ISSN 0028-0836. doi: 10.1038/nature13437. URL http://www.nature.com/articles/nature13437.

Alex K Shalek, Rahul Satija, Joe Shuga, John J Trombetta, Dave Gennert, Diana Lu, Peilin Chen, Rona S Gertner, Jellert T Gaublomme, Nir Yosef, Schraga Schwartz, Brian Fowler, Suzanne Weaver, Jing Wang, Xiaohui Wang, Ruihua Ding, Raktima Raychowdhury, Nir Friedman, Nir Hacohen, Hongkun Park, Andrew P May, and Aviv Regev. Single-cell {RNA-seq} reveals dynamic paracrine control of cellular variation. *Nature*, 510(7505):363–369, jun 2014b.

Yan Shi, Bin Shu, Ronghua Yang, Yingbin Xu, Bangrong Xing, Jian Liu, Lei Chen, Shaohai Qi, Xusheng Liu, Peng Wang, Jinming Tang, and Julin Xie. Wnt and Notch signaling pathway involved in wound healing by targeting c-Myc and Hes1 separately. *Stem Cell Res. Ther.*, 6:120, jun 2015.

Violeta Silva-Vargas, Cristina Lo Celso, Adam Giangreco, Tyler Ofstad, David M Prowse, Kristin M Braun, and Fiona M Watt. {$"β$-Catenin} and Hedgehog Signal Strength Can Specify Number and Location of Hair Follicles in Adult Epidermis without Recruitment of Bulge Stem Cells. *Dev. Cell*, 9(1):121–131, 2005.

Timothy Sterne-Weiler, Robert J. Weatheritt, Andrew Best, Kevin C. H. Ha, and Benjamin J. Blencowe. Whippet: an efficient method for the detection and quantification of alternative splicing reveals extensive transcriptomic complexity. *bioRxiv*, page 158519, jul 2017. doi: 10.1101/158519. URL https://www.biorxiv.org/content/early/2017/07/03/158519.

Tim J. Stevens, David Lando, Srinjan Basu, Liam P. Atkinson, Yang Cao, Steven F. Lee, Martin Leeb, Kai J. Wohlfahrt, Wayne Boucher, Aoife O'Shaughnessy-Kirwan, Julie Cramard, Andre J. Faure, Meryem Ralser, Enrique Blanco, Lluis Morey, Miriam Sansó, Matthieu G. S. Palayret, Ben Lehner, Luciano Di Croce, Anton Wutz, Brian Hendrich, Dave Klenerman, and Ernest D. Laue. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, 544(7648):59–64, mar 2017. ISSN 0028-0836. doi: 10.1038/nature21429. URL http://www.nature.com/doifinder/10.1038/nature21429.

Hikaru Takeda, Stephen Lyle, Alexander J F Lazar, Christos C Zouboulis, Ian Smyth, and Fiona M Watt. Human sebaceous tumors harbor inactivating mutations in LEF1. *Nature Medicine*, 12(4):395–397, apr 2006. ISSN 1078-8956. doi: 10.1038/nm1386. URL http://www.nature.com/articles/nm1386.

David W M Tan, Kim B Jensen, Matthew W B Trotter, John T Connelly, Simon Broad, and Fiona M Watt. Single-cell gene expression profiling reveals functional heterogeneity of undifferentiated human epidermal cells. *Development*, 140(7):1433–1444, apr 2013.

Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, Kaiqin Lao, and M Azim Surani. {mRNA-Seq} whole-transcriptome analysis of a single cell. *Nat. Methods*, 6(5):377–382, may 2009. ISSN 1548-7091. doi: 10.1038/nmeth.1315. URL http://www.ncbi.nlm.nih.gov/pubmed/19349980http://www.nature.com/articles/nmeth.1315.

Luyi Tian, Shian Su, Xueyi Dong, Daniela Amann-Zalcenstein, Christine Biben, Azadeh Seidi, Douglas J Hilton, Shalin H. Naik, and Matthew E. Ritchie. scPipe: a flexible R/Bioconductor preprocessing pipeline for single-cell RNA-sequencing data. *bioRxiv*, page 175927, mar 2018. doi: 10.1101/175927. URL https://www.biorxiv.org/content/early/2018/03/01/175927.

Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn.

The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, 32(4):381–386, apr 2014. ISSN 1546-1696. doi: 10.1038/nbt.2859. URL http://www.nature.com.docelec.univ-lyon1.fr/nbt/journal/v32/n4/full/nbt.2859.html.

T. Tumbar, Geraldine Guasch, Valentina Greco, Cedric Blanpain, William E Lowry, Michael Rendl, and Elaine Fuchs. Defining the Epithelial Stem Cell Niche in Skin. *Science*, 303(5656):359–363, jan 2004. ISSN 0036-8075. doi: 10.1126/science.1092436. URL http://www.ncbi.nlm.nih.gov/pubmed/14671312http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2405920http://www.sciencemag.org/cgi/doi/10.1126/science.1092436.

Mathias Uhlén, Linn Fagerberg, Björn M Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, IngMarie Olsson, Karolina Edlund, Emma Lundberg, Sanjay Navani, Cristina Al-Khalili Szigyarto, Jacob Odeberg, Dijana Djureinovic, Jenny Ottosson Takanen, Sophia Hober, Tove Alm, Per-Henrik Edqvist, Holger Berling, Hanna Tegel, Jan Mulder, Johan Rockberg, Peter Nilsson, Jochen M Schwenk, Marica Hamsten, Kalle von Feilitzen, Mattias Forsberg, Lukas Persson, Fredric Johansson, Martin Zwahlen, Gunnar von Heijne, Jens Nielsen, and Fredrik Pontén. Proteomics. Tissue-based map of the human proteome. *Science (New York, N.Y.)*, 347(6220):1260419, jan 2015. ISSN 1095-9203. doi: 10.1126/science.1260419. URL http://www.ncbi.nlm.nih.gov/pubmed/25613900.

J. v. Neumann. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, 100(1):

295–320, dec 1928. ISSN 0025-5831. doi: 10.1007/BF01448847. URL http://link.springer.com/10.1007/BF01448847.

Catalina A. Vallejos, John C. Marioni, and Sylvia Richardson. BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. *PLOS Computational Biology*, 11(6):e1004333, jun 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004333. URL http://dx.plos.org/10.1371/journal.pcbi.1004333.

Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008a. ISSN ISSN 1533-7928. URL http://www.jmlr.org/papers/v9/vandermaaten08a.html.

Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using {t-SNE}. *J. Mach. Learn. Res.*, 9(Nov):2579–2605, 2008b.

Lars Velten, Simon F Haas, Simon Raffel, Sandra Blaszkiewicz, Saiful Islam, Bianca P Hennig, Christoph Hirche, Christoph Lutz, Eike C Buss, Daniel Nowak, Tobias Boch, Wolf-Karsten Hofmann, Anthony D Ho, Wolfgang Huber, Andreas Trumpp, Marieke A G Essers, and Lars M Steinmetz. Human haematopoietic stem cell lineage commitment is a continuous process. *Nat. Cell Biol.*, 19(4):271–281, apr 2017.

Bo Wang, Junjie Zhu, Emma Pierson, Daniele Ramazzotti, and Serafim Batzoglou. Visualization and analysis of single-cell {RNA-seq} data by kernel-based similarity learning. *Nat. Methods*, 14(4):414–416, apr 2017.

F M Watt and C A Collins. Role of beta-catenin in epidermal stem cell expansion, lineage selection, and cancer. *Cold Spring Harb. Symp. Quant. Biol.*, 73:503–512, nov 2008.

F M Watt and B L Hogan. Out of Eden: stem cells and their niches. *Science*, 287(5457): 1427–1430, feb 2000.

Fiona M. Watt and Kim B. Jensen. Epidermal stem cell diversity and quiescence. *EMBO Molecular Medicine*, 1(5):260–267, aug 2009. ISSN 17574676. doi: 10.1002/emmm.200900033. URL http://www.ncbi.nlm.nih.gov/pubmed/20049729http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2850061http://embomolmed.embopress.org/cgi/doi/10.1002/emmm.200900033.

Fiona M Watt, B Simon, and David M Prowse. Cultivation and Retroviral Infection of Human Epidermal Keratinocytes. In *Cell Biology*, pages 133–138. dec 2006.

Hanseul Yang, Rene C Adam, Yejing Ge, Zhong L Hua, and Elaine Fuchs. {Epithelial-Mesenchymal} Micro-niches Govern Stem Cell Lineage Choices. *Cell*, 169(3):483—-496.e13, apr 2017.

Leilei Yang, Lijuan Wang, and Xiao Yang. Disruption of Smad4 in Mouse Epidermis Leads to Depletion of Follicle Stem Cells. *Molecular Biology of the Cell*, 20 (3):882–890, feb 2009. ISSN 1059-1524. doi: 10.1091/mbc.e08-07-0731. URL http://www.molbiolcell.org/doi/10.1091/mbc.e08-07-0731.

Evgeny Zamyatin and Andrey Filchenkov. Learning to Generate Chairs with Generative Adversarial Nets. may 2017.

Ying V. Zhang, Janice Cheong, Nichita Ciapurin, David J. McDermitt, and Tudorita Tumbar. Distinct self-renewal and differentiation phases in the niche of infrequently dividing hair follicle stem cells. *Cell Stem Cell*, 5(3):267–278, sep 2009. ISSN 19345909. doi: 10.1016/j.stem.2009.06.004. URL http://dx.doi.org/10.1016/j.stem.2009.06.004.

Grace X Y Zheng, Billy T Lau, Michael Schnall-Levin, Mirna Jarosz, John M Bell, Christopher M Hindson, Sofia Kyriazopoulou-Panagiotopoulou, Donald A Masquelier, Landon Merrill, Jessica M Terry, Patrice A Mudivarti, Paul W Wyatt, Rajiv Bharadwaj, Anthony J Makarewicz, Yuan Li, Phillip Belgrader, Andrew D Price, Adam J Lowe, Patrick Marks, Gerard M Vurens, Paul Hardenbol, Luz Montesclaros, Melissa Luo, Lawrence Greenfield, Alexander Wong, David E Birch, Steven W Short, Keith P Bjornson, Pranav Patel, Erik S Hopmans, Christina Wood, Sukhvinder Kaur, Glenn K Lockwood, David Stafford, Joshua P Delaney, Indira Wu, Heather S Ordonez, Susan M Grimes, Stephanie Greer, Josephine Y Lee, Kamila Belhocine, Kristina M Giorda, William H Heaton, Geoffrey P McDermott, Zachary W Bent, Francesca Meschi, Nikola O Kondov, Ryan Wilson, Jorge A Bernate, Shawn Gauby, Alex Kindwall, Clara Bermejo, Adrian N Fehr, Adrian Chan, Serge Saxonov, Kevin D Ness, Benjamin J Hindson, and Hanlee P Ji. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnology*, 34 (3):303–311, mar 2016. ISSN 1087-0156. doi: 10.1038/nbt.3432. URL http://www.

nature.com/articles/nbt.3432.

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

a J Zhu and F M Watt. Beta-Catenin Signalling Modulates Proliferative Potential of Human Epidermal Keratinocytes Independently of Intercellular Adhesion. *Development (Cambridge, England)*, 126(10):2285–2298, 1999. ISSN 0950-1991.

Jun-Yan Zhu and Taesung Park. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks Monet Photos. URL https://arxiv.org/pdf/1703.10593.pdf.

Rapolas Zilionis, Juozas Nainys, Adrian Veres, Virginia Savova, David Zemmour, Allon M Klein, and Linas Mazutis. Single-cell barcoding and sequencing using droplet microfluidics. *Nat. Protoc.*, 12(1):44–73, jan 2017.

Christos C Zouboulis. Acne and sebaceous gland function. *Clinics in dermatology*, 22(5):360–6, sep 2004. ISSN 0738-081X. doi: 10.1016/j.clindermatol.2004.03.004. URL http://www.ncbi.nlm.nih.gov/pubmed/15556719.

Justina žurauskienė and Christopher Yau. pcaReduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinformatics*, 17(1):140, dec 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-0984-y. URL http://www.biomedcentral.com/1471-2105/17/140.