Integrative study of gene expression and protein complexes

Kiana Toufighi

TESI DOCTORAL UPF / ANY 2014

THESIS DIRECTORS
Dr. Ben Lehner & Dr. Luis Serrano

DEPARTMENT
EMBL-CRG Systems Biology Unit
Centre for Genomic Regulation (CRG)

THESIS ABSTRACT

Over the last several decades, the emerging 'integrated' view of the cell has triumphed over the 'one gene/one protein/one function' paradigm. This is illustrated by the biologically opposite effects of key regulatory proteins in different cell types, in established versus primary cells, and *in vitro* versus *in vivo* situations. The persistent theme throughout this dissertation is the integration of a wide range of data sources for the purpose of understanding distinct cellular contexts. We first use circadian expression data from human epidermal stem cells to discover waves of transcripts expressed in tune with known clock genes and to show that time-of-day dependent responses to proliferation/differentiation cues is important for skin homeostasis. We then combine this expression data with information on protein structures and complexes to describe how protein-complex assembly is temporally regulated during differentiation. Lastly, we show that human protein complexes are composed of a stable 'core' and a plastic 'periphery' whose tissue-specific expression allows protein complexes to function in a context-dependent manner.

RESUMEN DE TESIS

En las últimas décadas, la emergente vista integrativa de la célula ha triunfado sobre el paradigma histórico : 'un gene/una proteína/una función'. Esto es ilustrado por los efectos biológicos opuestos de proteínas regulatorias clave en cultivos celulares inmortalizados frente a primarios e *in vitro* frente a *in vivo*. El tema persistente en este disertación es la integración de un amplio set de datos para estudiar los distintos contextos celulares. En primer lugar, utilizamos los datos de expresión génica obtenidos de células madre epidérmicas para descubrir las ondas de transcripción expresadas en sintonía con los genes conocidos de los ritmos circadianos. En este estudio demostramos que las respuestas de las células madres a las señales de proliferación/diferenciación dependen de hora del día y el tiempo circadiano es importante para la homeostasis de la piel. Posteriormente, combinamos estos datos de expresión con la información estructural de proteínas y complejos proteicos para describir la regulación temporal de complejos durante el proceso de diferenciación. Por último, mostramos que los complejos de proteínas humanos están compuestos de un 'núcleo' estable y una 'periferia' plástica cuya expresión específica de tejido celular permite que los complejos de proteínas funcionen de una manera dependiente del contexto.

PUBLICATIONS

Human Epidermal Stem Cell Function Is Regulated by Circadian Oscillations

Peggy Janich*, Kiana Toufighi*, Guiomar Solanas*, Nuno Miguel Luis, Susann Minkwitz, Luis Serrano#, Ben Lehner#, Salvador Aznar Benitah#

2013, Cell Stem Cell.

Dissecting the calcium-induced differentiation of human primary keratinocytes stem cells by integrative and structural network analyses

Kiana Toufighi, Jae-Seong Yang, Nuno Miguel Luis, Salvador Aznar Benitah#, Ben Lehner#, Luis Serrano#, Christina Kiel#

2014, In preparation.

Context-dependent plasticity in human protein complexes

Kiana Toufighi, Luis Serrano#, Ben Lehner#

2014, In preparation.

* Co-first authors.

Co-corresponding authors.

ACKNOWLEDGEMENTS

My decision to journey across the Atlantic to pursue graduate studies was a rather risky endeavour, but one which ultimately proved to be incredibly rewarding. Here, I would like to thank all of the people whose various contributions have made my time here as a doctoral student an enriching experience.

First of all, this work would not have been possible without the contribution of my supervisors Ben Lehner and Luis Serrano. Thank you both for giving me the opportunity to commence studies at a great institution and for providing guidance and support along the way on professional and at times personal matters. Your passion for and dedication to science has been awe-inspiring and I thank you for instilling in me great respect and appreciation for good science and good scientists.

I also want to thank Salvador Aznar-Benitah, officially our collaborator on two projects, unofficially my third supervisor. Thank you for your encouragement and your unwavering confidence in me even when we faced many difficulties with the project. During our lengthy discussions, I learned more about skin biology from you than from hours of reading. To this day, I still refer back to the detailed notes and hand-drawn diagrams you made me during our meetings.

I want to thank Christina Kiel with whom I have worked closely over the last year. Your dedication to work is exemplary and I appreciate your efforts and input in completing the final manuscript. Thank you Jae-Seong Yang, not only for your contribution to my work through SAPIN but also for being my seminar buddy, late-night

work companion, and brainstorming fellow. Thanks Peggy Janich for a very fruitful collaboration. I really enjoyed working with you and I admire you as a good scientist.

Over the years, I have met many incredible people, a subset of whom have made a lasting mark and whom I hope shall remain life-long friends. My surrogate family who took me in when I first arrived: Almer, Raik, Elena, Eli, Tony, and Jojo! Thanks Eli for being my Catalan sister. My informatics crew doubling as neighbours and friends: Peter, Erik, Anne, Marie, Javi, and Besray. We had some fun times together both inside the lab and out. Específicamente gracias Javi "Delgadito" por tu amistad. Te agradezco mucho por todo lo que me has enseñado en el trabajo y en la vida. He aprendido castellano sólo gracias a ti. Maria, Veronica, Hannah, Berni, and Carolina, you all contributed to a very nice atmosphere in the lab. Alejandro thanks for being an unforgettable friend with the perfect balance of nerdiness and charisma. To this day the phrase "no pain no glory" reminds me of you. Camilla you are truly awesome and the past four years would not have been the same without you. Thanks for many "AMAZING" experiences but more importantly thanks for bringing grappa into my life! Inna thanks for being so caring and for always having my back. For the rest of my life I will owe you for talking me out of what could have been a very stupid decision. In chronological order: Veronica, Emilia, Sarah, and Luisa, thanks for being wonderful housemates and dinner companions but more importantly friends — each in your own unique way.

Thank you Eldar for always being an incredibly good and understanding friend and for letting me pursue my dreams. I give my sincere thanks to my parents for supporting my at-times odd choices and for loving me unconditionally. Katty you are my best friend and Ali you have delivered whenever I have needed you.

Lastly I want to thank a very important person without whom this PhD dissertation would not have materialised. Toby, from running the domestic support unit to sitting on the one-seat scientific advisory board to minding the 24-hour emergency hotline despite the time difference, you ran a smooth operation through and through. I am not always sure I deserve your boundless love but I am sure as hell grateful for having it.

PREFACE

This dissertation covers several studies under the general topic of how computational analysis of spatiotemporal gene expression provides insight into the modular workings of the cell in various contexts. It is the objective of this dissertation to unify and present the findings of three separate manuscripts laid out over three distinct chapters. I have attempted to provide the necessary background to all three studies in the first chapter. The next three chapters can be read independently as they have been published or are in the process of being submitted for publication as separate scientific articles. Finally in the last two short chapters, a general overview of the results, a summary of all three end-of-chapter discussions, and concluding remarks have been presented.

GLOSSARY

Table 1: List of abbreviations

Abbreviation	Description
AF	affinity purification
AP2	apetala 2
ARNTL	aryl hydrocarbon receptor nuclear translocator-like
BMAL1	brain and muscle ARNT-like 1
BMP	bone morphogenetic protein
BrdU	bromodeoxyuridine
C/EBP	CCAAT/enhancer binding protein
CLOCK	circadian locomoter output cycles protein kaput
CRY1/2	cryptochrome 1/2
ECM	extracellular matrix
EGF	epidermal growth factor
FGF	fibroblast growth factor
GRHL3	grainyhead-like 3
HF	hair follicle
HT	high-throughput
HPLC	high-pressure liquid chromatogrphy
IEX-HPLC	ion exchange high-pressure liquid chromatography
IF	interfollicular
IFE	interfollicular epidermis
IRF6	interferon regulatory factor 6
KLF4	Kruppel-like factor 4
LC	liquid chromatography
MAPK	mitogen-activated protein kinase
MS	mass spectrometry
NF-κB	nuclear factor-κB

Table 1 – Continued

Abbreviation	**Description**
NR1D1/2	nuclear receptor subfamily 1, group D, member 1/2 (also Rev-Erbα)
PER1/2/3	period 1/2/3
PPI	protein-protein interaction
PTM	post-translational modification
Rev-Erbα	nuclear receptor subfamily 1, group D, member 1 (also NR1D1)
Rorα	RAR-related orphan receptor alpha
SC	stem cell
SCN	suprachiasmatic nucleus
SG	sebaceous gland
SHH	sonic hedgehog
TA	transit amplifying
TGFβ	transforming growth factor-beta
TTFL	transcriptional-translational feedback loop
UAS_G	GAL upstream activation site

LIST OF FIGURES

LIST OF TABLES

CONTENTS

CHAPTER

ONE

INTRODUCTION

Instead of using the usual recipe book analogy, if we take some poetic licence and think of the genome simply as a great piano with many thousands of keys, then each key can represent a gene and each note produced when a key is pressed can represent the expression of that gene. Much like pressing any one key for one quarter of a beat would produce a shorter note than pressing the same key for a full beat, the expression of a gene too can take the form of numerous isoforms both as transcripts and as proteins differing significantly in length and consequently secondary and tertiary structure. In this analogy a faulty key, one that sticks or sounds out of tune when struck, may represent a mutation, much like an unskilled pianist or one who has just suffered a hand injury impacting his ability to perform represents misregulation of any one gene or a number of genes. Both such scenarios can impact any musical passage played, at times perhaps ruining it altogether.

The part of this analogy most relevant to this work is not the characteristics of the individual parts that make the musical instrument, but rather the music itself. In the making of music, not every key has to be incorporated into every musical piece. In living organisms too, not every gene is expressed in every tissue and cell type nor in every process. Just like various genres, styles, and melodies are played by striking the keys of a piano in special combinations, mixing notes and chords in specific sequences with the right beat at the right tempo, various tissues too are established and maintained by the expression and regulation of specific

combinations of genes as are cellular processes. The underlying theme throughout the whole of this dissertation, evident to varying degrees in the three chapters that follow is how gene expression brings plasticity, variation, and context to the cell.

Across the three main chapters, we have consistently taken advantage of global maps of the cell in the form of mRNA expression and physical protein interaction maps to learn about distinct processes, cellular contexts, and tissues. We have utilized time-course mRNA expression data to observe that skin cells in culture have autonomous circadian rhythms (Chapter 2); we have learned that this circadian rhythm helps maintain homeostasis and there is evidence to suggest that differentiation happens in a time-of-day-dependent manner (Chapter 2); we looked at how protein complexes change their composition over the course of skin differentiation based on expression pattern of their member subunits (Chapter 3), and we studied the organization of human protein complexes and how they are composed of two distinct groups of subunits: a ubiquitously expressed core and a dynamic periphery, which — expressed in the right context — brings functional plasticity to the whole protein complex (Chapter 4).

1.1 Introduction to skin

The skin is the outermost organ that separates animals from their environment and protects them against environmental stresses such as water loss, pathogenic attacks, ultraviolet radiation, and thermal and mechanical injuries. Mammalian skin features an elaborate array of appendages like hair follicles, nails, sebaceous and sweat glands, blood vessels, and nerves, all of which are necessary for its various functions like thermal regulation, sensation, secretion, and absorption. To keep within the scope of this work we shall only introduce features of mammalian skin, specifically those pertaining to mice and men. We also note that, unless otherwise stated, all *in vivo* experiments discussed in this dissertation were conducted in mouse.

Mammalian skin comprises two contiguous layers: dermis and epidermis. The

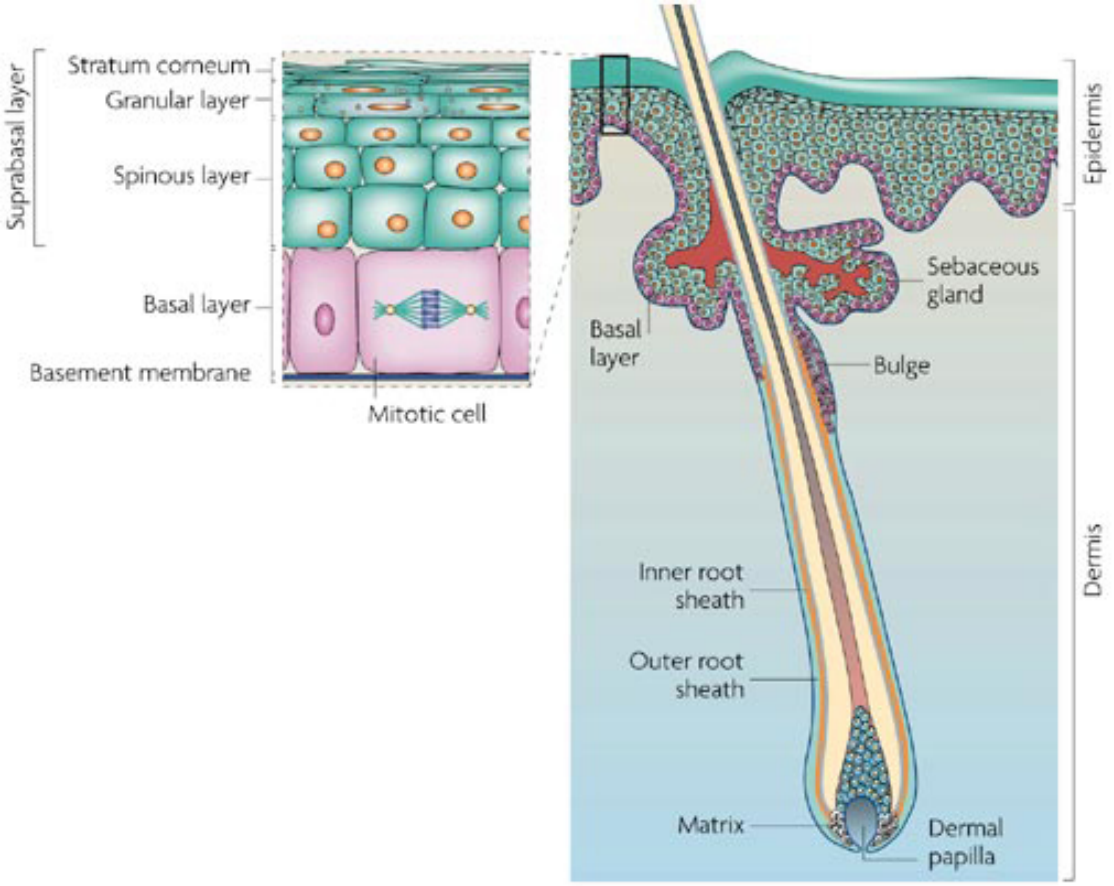

Figure 1.1: **Mammalian skin** : Cross-section of skin and close-up of epidermis (Jones and Simon, 2008).

dermis is the lower layer of the skin that is rich in connective tissue and contains collagen fibres, fibroblasts, blood vessels, and immune cells. The outermost layer of the skin, the epidermis, is a specialised multi-layered epithelium tissue, which is the defining component of the skin. Because it suffers direct, frequent, and at times damaging encounters with the environment, its need for repair and renewal is central to its organisation. The epidermis is nourished and supported by the subjacent dermis yet is separated from it by a thin sheet of fibres called the basement membrane (Figure 1.1). The epidermis adheres to this basement membrane, which not only serves as a boundary but also a growth-promoting platform, being rich in extracellular matrix proteins and growth factors (Paolo Dotto, 1999; Alberts *et al.*, 2002). The epidermis is composed of an inner layer of basal cells with proliferative

potential and overlying stratified layers of differentiating progeny, collectively called the suprabasal layer (Connelly *et al.*, 2011; Alberts *et al.*, 2002). Each layer of cells in the suprabasal layers, namely in order from bottom to top, spinous layer, granular layer, and cornified layer (or stratum corneum) is at a distinct point within the differentiation program. As cells move from the basal layer through the stratified suprabasal layer up to the squamous epithelium, they enlarge and terminally differentiate. In fact, the stage of terminal differentiation is correlated both with cell size and position (Clark & Coker, 1998; Watt & Green, 1982; Massagué & Gomis, 2006). It takes roughly two to three weeks for a basal cell to complete its migratory differentiation program, eventually terminating when it is shed off the surface of the skin (Figure 1.1).

In this manner, the epidermis continuously turns over while maintaining a constant number of cells through a process called homeostasis. Homeostasis in mammalian skin is driven by a population of well-characterized epidermal stem cells (SCs) which constantly self-renew in order to maintain the tissue. These epidermal SCs have two roles. First they have to replenish terminally differentiated cells, which are continually shed from the surface of the skin. Second they have to regenerate damaged or necrotic cells after injury.

Specialised appendages in the skin, such as hair follicles (HF) exist within pilosebaceous units along with the adjoining interfollicular epidermis (IFE) (Figure 1.2). The HF is an intricate structure with its own specialized subtypes of stem cells. To simplify, its most important components include a dermal papilla, which is a small nipple-like extension of the uppermost layer of the dermis into the epidermis, a cellular matrix above the dermal papilla, a hair shaft containing a hair fibre, a sebaceous gland, and a bulge region in the close lower proximity of the sebaceous gland (Figure 1.1). The IFE with its population of basal stem cells is also responsible for the maintenance of the skin barrier. In the next section we look at the stem cells within both HF and IFE.

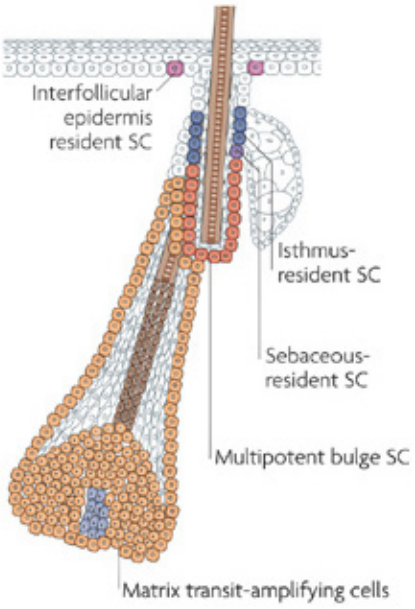

Figure 1.2: **Epidermal stem cells** : Schematic representation of the skin epidermis with the different resident SC compartments and transit-amplifying progeny identified. Bulge SCs are multipotent, residing in the permanent portion of the HF. IFE SCs are localised to the basal layer of the epidermis. Resident progenitors of the isthmus and SG reside in the outer root sheath that is above the bulge and below the SG. (adapted from Blanpain & Fuchs (2009)

1.1.1 Epidermal stem cells

Epidermal SCs are localized in specialized niches, which are sheltering microenvironments able to sequester the SCs from differentiation stimuli, apoptotic stimuli, or other signals that might challenge the SC reservoir (Koster, 2004; Moore & Lemischka, 2006; Mills *et al.*, 1999; Truong *et al.*, 2006; Senoo *et al.*, 2007). These SC niches also safeguard against excessive self-renewal, thus preventing cancer (Watt *et al.*, 2008; Moore & Lemischka, 2006). Epidermal stem cell niches can be divided into those localized to the basal region of the IFE, and those in HF, where they are specifically found in the sebaceous gland, the upper isthmus, and the bulge region (Rangarajan *et al.*, 2001; Watt *et al.*, 2006)(Figure 1.2). Throughout the life of the organism, epidermal SCs are normally quiescent and undifferentiated, but become proliferative and egress their niche, most of the time periodically to guarantee epidermal replenishment and hair growth or on rare occasions abruptly in response to

stress signals like wounding. We shall introduce both hair follicle and interfollicular SCs, although with greater emphasis on the IF SCs since our primary cell cultures originated from this population.

HF SCs are responsible for the periodic regeneration of the hair and sebaceous gland contained within the HF. Hair growth follows an intricate cycle of regeneration and degeneration during which hair grows until the SCs have exhausted their proliferative capacity, then stops and enters a destructive phase (catagen) characterized by high levels of apoptosis and tissue remodelling in the lower two-thirds of the HF. This is followed by a quiescent stage (telogen), which lasts one to two days, before a new growth stage (anagen) begins during which SCs in the bulge region of the HF become activated and migrate to the lower hair germ region where they form a large pool of highly proliferative transit amplifying (TA) matrix cells (Moriyama *et al.*, 2008; Blanpain & Fuchs, 2009) (Figure 1.3). Long-term pulse-chase experiments have revealed that SCs in the bulge region cycle very slowly and are relatively quiescent (Blanpain & Fuchs, 2006; Cotsarelis *et al.*, 1990; Braun, 2003). TA matrix progenitors in turn go through a few rapid rounds of cell division, then embark on seven concentric terminal differentiation programs to generate the entire mature hair follicle (Sen *et al.*, 2008; Blanpain & Fuchs, 2009; Frye *et al.*, 2007). These seven different cell lineages are distributed such that three are in the hair shaft and four in the inner root sheath (IRS) (Figure 4).

Initially, it was suggested that HF SCs in the bulge are also responsible for the replenishment of the IFE. However, it is now clear that under normal homeostatic conditions all epidermal SCs including those in the bulge (Frye *et al.*, 2007; Ito *et al.*, 2005) only form the differentiated lineages of their respective compartment. Only under challenging conditions, such as wounding, do epidermal SCs leave their compartment and give rise to the progeny in other compartments, thus transiently contributing to the regeneration of those tissues and demonstrating their own multipotency.

Interfollicular SCs are located in the basal layer of the epidermis in close contact with the underlying basement membrane. They express high levels of adhesion

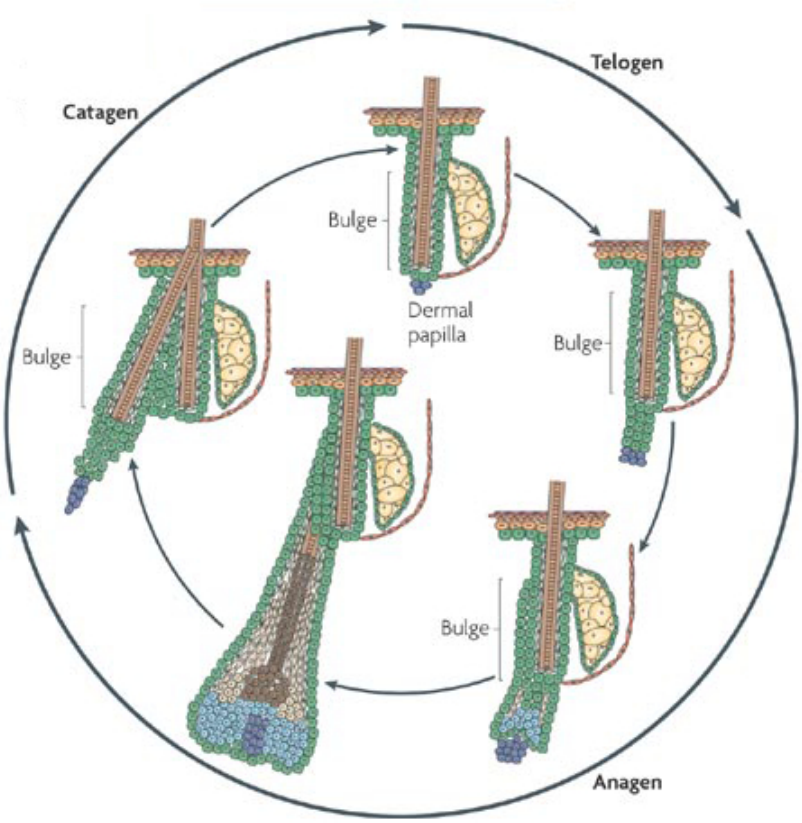

Figure 1.3: **Hair follicle cycle** : phases of hair follicle growth cycle (adapted from Blanpain & Fuchs (2009)

molecules, such as α6- and β1-integrins (Sen *et al.*, 2008; Jones *et al.*, 1995). Similar to HF SCs, basal IF SCs self-renew as well as routinely execute a program of terminal differentiation, dividing vertically, detaching from the basement membrane, and moving upward to the surface of the skin in a columnar fashion all the while increasing steadily in size.

There are two types of cell division taking part in epidermal homeostasis. In the first type, symmetric cell division, both daughter cells adopt identical fates. An example of this mode of division has been observed in early stages of embryogenesis, at 12.5 days after gastrulation, where over 75% of cell divisions are symmetric and parallel to the BM. At this stage, the developing embryo is kept protected by the rapid expansion of a single layer of epithelium through symmetric and lateral cell divisions

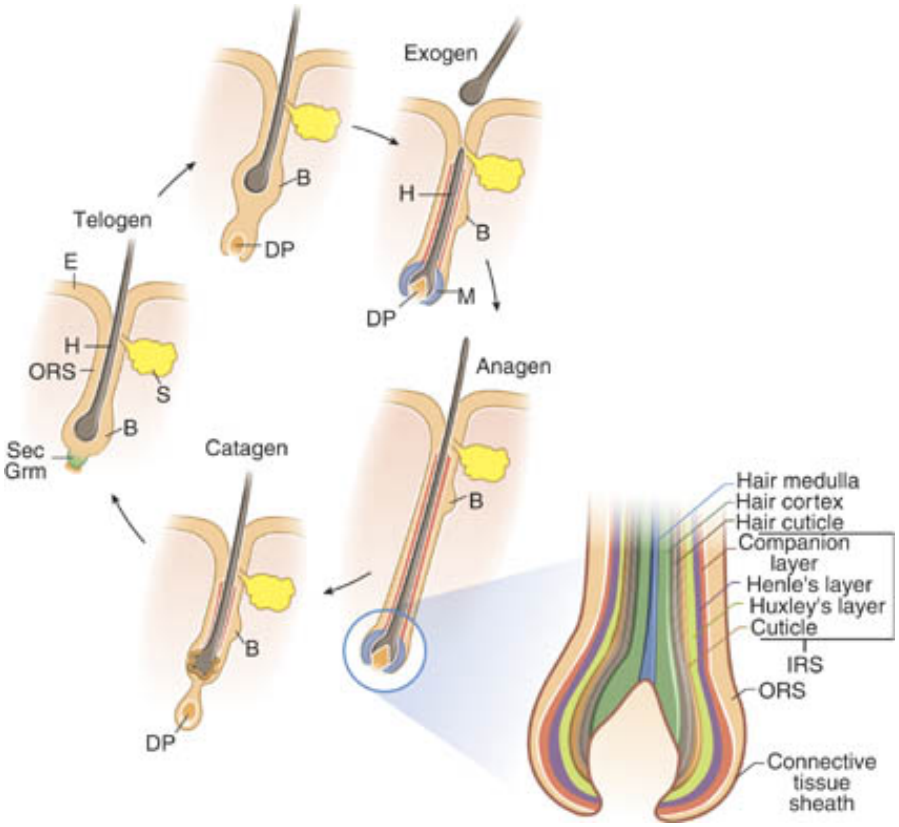

Figure 1.4: **Hair follicle anatomy** : seven distinct cell populations in the hair follicle along with growth cycle

— a mechanism by which daughter cells remain attached to the BM and continue to receive appropriate signals from basal extracellular matrix (Sen *et al.*, 2008; Lechler & Fuchs, 2005). This type of division, which is called symmetric self-renewal, can also be observed in adults, albeit with much lower frequency which presumably is sufficient to maintain SCs in the basal layer. Cells can also divide symmetrically but subsequently both enter differentiation. This type of division can be seen between embryonic day 13.5 and 15.5 whereby the suprabasal SCs are highly proliferative in order to keep up with the rapid increase in the size of the embryo (Blanpain & Fuchs, 2006; Lechler & Fuchs, 2005).

Asymmetric cell division is one of the two models proposed for the onset of stratification (Tumbar, 2004; Blanpain & Fuchs, 2009; Morris *et al.*, 2004) whereby the

mitotic spindle is perpendicular to the BM enabling the two daughter cells to have asymmetric cell fates, with the basal daughter cell remaining attached to the BM and retaining receptors for growth factors, and the suprabasal daughter cell being committed to differentiation (Blanpain & Fuchs, 2006; Lechler & Fuchs, 2005). There is evidence for this model in both embryogenesis and in normal adults where some 85% of mitoses happen asymmetrically and vertically (Candi *et al.*, 2005; Lechler & Fuchs, 2005). The second mode for asymmetric cell division involves a lateral division plane parallel to the BM. In this model, one of the daughter cells might receive an unequal quantity of a signal leading to the down-regulation of integrin expression and the subsequent detachment form the BM and the onset of differentiation (Candi *et al.*, 2005; Blanpain & Fuchs, 2009).

Once cells begin their migratory journey from the basal to the suprabasal layer at the onset of differentiation, they undergo a number of biochemical and morphological changes. They switch off the expression of genes encoding keratin 5 and keratin 14 (KRT5 and KRT14) which are markers of epithelial cells in the basal layer and function as cytoskeletal filaments to protect cells from mechanical stress (Candi *et al.*, 2005; Fuchs & Green, 1980). At the same time, migrating cells switch on the expression of keratin 1 and keratin 10 (KRT1 and KRT10) in order to form an expansive cytoskeletal filament to bolster cell-cell junctions and make the body surface resistant to mechanical stress. As the migrating cells pass from the spinous layer to the granular layer (Figure 2) they continue to increase in size; they also lose their organelles (Woelfle *et al.*, 2004; Lavker & Matoltsy, 1970; Dodd *et al.*, 2005) leaving keratins as the main structural proteins in the cytoplasm. These eventually become cross-linked, forming a scaffold (Dunlap, 1999; Rice & Green, 1977). At the stage of the stratum corneum the cells are simply cellular skeletons without any metabolic activity. Human and mouse IFE are similar in many aspects concerning their morphology and biochemical activity. Yet unlike human IFE where SCs remain quiescent for long stretches, murine IFE is maintained by stem cells that divide on a daily basis. The progeny of the SCs then undergo several rounds of division before they detach from the basement membrane and begin their differentiation route into the suprabasal layers (Jones & Simons, 2006).

1.1.2 Epidermal SCs in culture

Over the years, *in vivo* analyses of epithelial tissues, under normal conditions and in loss-/gain-of-function mutants, have provided significant insights into the maintenance of epithelial homeostasis. However, understanding the detailed molecular mechanisms regulating epidermal homeostasis and differentiation also requires biochemical *in vitro* studies. In this regard, studies involving cultured epithelial cells have served as an important complementary approach. In 1975 two influential studies described the establishment of an immortal epidermal cell line derived from a mouse teratoma as well as cultivation of primary human keratinocytes in culture albeit with restricted lifetime of 20-50 cell generations (Rheinwald & Green, 1975a,b). Despite this early success, it took many attempts for normal human skin cells to be transformed into immortalized cell lines (Boukamp *et al.*, 1988). This was an important step. Cell lines originating from cancers, though informative, were deemed limited in their usefulness because of significant differences that existed in growth regulation linked to cell line immortality (Paolo Dotto, 1999). Since those early days, giant leaps have been made and the development of epithelial cell cultures has reached such heights that they can now be used to produce grafts that regenerate an epidermis over a full-thickness wound (Coolen *et al.*, 2007). For basic research, both primary and immortalized cell lines, derived from both mouse and human and originating from both healthy and tumorous tissue, are readily available (Balsalobre *et al.*, 1998; Coolen *et al.*, 2007; Balsalobre, 2000). However, we will limit ourselves to the discussion of human primary keratinocytes, as they were the mainstay of this work and all experiments were conducted on primary keratinocyte cultures isolated from human subjects.

As previously mentioned, one feature of the epidermis is the polarized pattern of epithelial growth and differentiation, where a single layer of proliferating keratinocytes is localized at the base underneath multiple differentiating layers. In culture, both mouse and human primary keratinocytes exhibit many properties of basal stem cells in morphology and biochemistry — continuously proliferating and exhibiting appropriate markers of stemness — when kept at low calcium concentrations (0.05 mM) so long as they remain attached to the culture dish or underlying

matrix (Paolo Dotto, 1999). It has been observed however, that even at low calcium concentrations, a minority of cells spontaneously detach from the culture dish and express many if not all markers associated with the suprabasal layers and become terminally differentiated (Paolo Dotto, 1999). On the other hand, differentiation of keratinocytes attached to the underlying fibroblast feeder cells or the culture plate can be induced by the addition of calcium (0.12 to 2 mM) (Hennings *et al.*, 1980). In fact, addition of calcium to primary keratinocyte cultures elicits a relatively complete differentiation program, inducing not only biochemical markers but also many of the structural changes that occur *in vivo*. It has been reported that, *in vivo*, there exists an increased gradient of calcium concentrations from the basal to the upper epidermal layers. For example, experiments featuring ion-capture cytochemistry have shown that extracellular calcium concentrations are significantly increased in the mid- to upper layers (Menon *et al.*, 1992) and a steady increase of calcium toward the cornified layer has been observed by particle probe methods (Forslind *et al.*, 1997). In summary, calcium induction of cultured epithelial cells serves as a robust model for the study of skin differentiation *in vitro* and calcium-induced signalling is likely also relevant *in vivo*, although it need not be the critical trigger for differentiation.

1.1.3 Relevant pathways to skin homeostasis and differentiation

Morphological details of epidermal stratification are well characterized in both human and mouse, but the molecular mechanisms, which orchestrate these processes, have only begun to emerge. A number of pathways have been shown to be required for skin homeostasis and for determining stem cell fate. Almost all have been characterized utilizing mouse genetics. Although some of the markers used to identify stem cells are different in mouse versus human skin, common signalling pathways appear to control epithelial stem cell maintenance, activation, lineage determination, and differentiation in both animals. In this section, evidence gathered from the study of both human and murine cell culture as well as *in vivo* mouse genetics will be presented. We discuss signalling pathways pertaining to both quiescence and differentiation together in one section since very often the these signals acts as

an ‘on’ or ‘off’ switch in deciding cell fate.

To start off, pathways required for maintenance/quiescence, activation, and fate determination of IF and HF stem cells during both embryogenesis and adulthood are Wnt/β-catenin, bone morphogenetic protein (BMP), transforming growth factor-beta (TGFβ), fibroblast growth factor (FGF), sonic hedgehog (SHH), epidermal growth factor (EGF), mitogen-activated protein kinase (MAPK), nuclear factor-κB (NF-κB), and Notch signalling pathway among others (Fuchs, 2007). In addition, a number of transcription factors including p63 (homologous to p53), the apetala 2 (AP2) family, CCAAT/enhancer binding protein (C/EBP), Kruppel-like factor 4 (KLF4), interferon regulatory factor 6 (IRF6), and grainyhead-like 3 (GRHL3) have been identified (Blanpain & Fuchs, 2009). For the sake of brevity, we will attempt to introduce only those pathways, which are relevant to this work.

Wnt/β-catenin Signalling Pathway

Wnt/β-catenin signalling is well implicated in promoting hair follicle (HF) formation during embryogenesis and differentiation in the adult animal (Blanpain & Fuchs, 2009). Over two decades of study on the canonical Wnt pathway have revealed its central logic, whereby β-catenin is the key mediator to the transcriptional regulation of hundreds of Wnt target genes (Logan & Nusse, 2004; Clevers, 2006; MacDonald *et al.*, 2009). In brief, in the absence of a Wnt signal — the default ‘off’ state — β-catenin is continuously degraded by the action of the Axin complex, composed of Axin, the tumour suppressor adenomatous polyposis coli gene product (APC), casein kinase 1 (CK1), and glycogen synthase kinase 3 (GSK3). CK1 and GSK3 sequentially phosphorylate β-catenin, marking it for recognition by an E3 ubiquitin ligase subunit, which leads to its ubiquitination and proteasomal degradation (MacDonald *et al.*, 2009). The continual degradation of β-catenin prevents it from reaching the nucleus, which means Wnt target genes remain in ‘off’ state through the DNA-binding of T cell factor/lymphoid enhancer factor (TCF/LEF) family of proteins (MacDonald *et al.*, 2009). The pathway is switched ‘on’ upon the binding of the Wnt ligand to Frizzled (FZD), a transmembrane receptor, along with its corecep-

tor, low-density lipoprotein receptor-related protein 6 (LRP6), or a close homologue LRP5. The assembly of these three proteins recruits another, Dishevelled (DVL) and leads to the phosphorylation and activation of LRP6. This in turn recruits the Axin complex to the receptors and leads to the inhibition of Axin complex-mediated phosphorylation and degradation of β-catenin. Stabilised β-catenin proteins accumulate in the cytoplasm, shuttle to the nucleus, and form complexes with TCF/LEF thereby derepressing Wnt target genes (MacDonald *et al.*, 2009; Clevers, 2006).

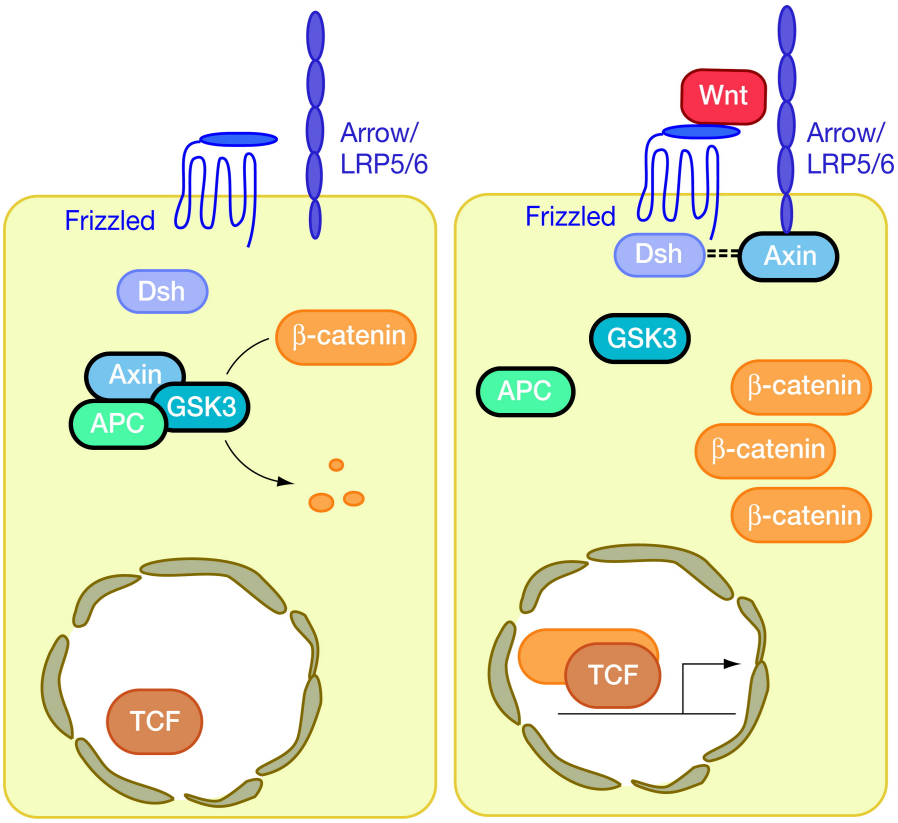

Figure 1.5: **Canonical Wnt signalling pathway** : in cells not exposed to a Wnt signal (left panel), β-catenin is degraded through interactions with Axin, APC, and the protein kinase GSK-3. Wnt proteins (right panel) bind to the cell surface Frizzled/LRP receptor complex transducing a signal to Dishevelled (Dsh) and to Axin, which may directly interact (dashed lines). Consequently, the degradation of β-catenin is inhibited resulting in its accumulation in the cytoplasm and nucleus. β-catenin then interacts with TCF to control transcription. Negative regulators outlined in black, positive regulator outlined in colour (from Logan & Nusse (2004)).

Wnts are conserved in all metazoans. In mammals, the complexity and the specificity required to maintain various tissues and/or execute distinct developmental programs is achieved through a catalogue of 19 Wnt ligands (MacDonald *et al.*, 2009). Mouse genetics have revealed the expression of several Wnt ligands at distinct developmental stages in mammalian skin, among them Wnt3, Wnt3a, Wnt4, Wnt5a, Wnt7a, Wnt10a, and Wnt10b (Huelsken *et al.*, 2001). In mouse embryonic skin, for instance, Wnt/β-catenin signals the formation of the first HFs as it was shown that the skin-specific ablation of β-catenin or the overexpression of Wnt inhibitor Dickkopf1 (Dkk1) blocks the formation of hair placodes (Huelsken *et al.*, 2001; Andl *et al.*, 2002). Conversely, it was demonstrated that overexpression of β-catenin, which mimics an activated Wnt pathway, in the skin of transgenic mice induces de novo hair follicles (Gat *et al.*, 1998), and yields skin- and hair-derived tumours (Chan *et al.*, 1999). One of the studies incorporating a β-catenin conditional knock-out also showed that β-catenin is required for HF SC niche specification in adult mice. In its absence, stem cells adopt an epidermal fate (Huelsken *et al.*, 2001). The same study also demonstrated that a hyperproliferative IFE phenotype resulted from conditional β-catenin knock-out. This seeming contradiction — the opposite effect of β-catenin deletion on epidermal SCs compared to HF SCs — remained controversial until recently, when the role of the Wnt/β-catenin pathway in the two SC compartments was teased apart by two independent groups (Choi *et al.*, 2013; Lim *et al.*, 2013). In one of the studies, selectively deleting β-catenin or Wntless (Wls), a gene required for Wnt ligands secretion, or even overexpressing Dkk1 to block the Wnt signal, indicated that Wnt signalling is not directly required for follicular SC maintenance, but is required for activation of the secondary germ cells (Figure 1.5) to regenerate the follicle. Both groups showed, however, that Dkk1-mediated suppression of Wnt signalling reduced proliferation in epidermis rather than promoted it. Furthermore, deletion of β-catenin in hairless skin also caused decreased proliferation, suggesting that the hyperproliferative phenotypes might result from inflammation caused by follicular degeneration as a form of response to a repair defect. Thus the groups come to the same conclusion that Wnt signalling promotes proliferation of stem cells in both compartments, similar to what has been observed in other epithelia (Choi *et al.*, 2013; Lim *et al.*, 2013). Finally, it has been shown that after severe wounding a Wnt-dependent signal can

induce de novo HF formation from the IFE similar to early stages of embryogenesis (Ito *et al.*, 2007). This highlights the multipotency of IFE SCs and how they can adopt HF fate in response to Wnt signals during physiopathological conditions.

TGFβ Signalling Pathway

The transforming growth factor beta (TGFβ) superfamily of cytokines is ubiquitous, multifunctional, and essential to survival. They are involved in many cellular functions in both the adult organism and the developing embryo, from cell division and apoptosis to adhesion and immune responses (Clark & Coker, 1998; Massagué & Gomis, 2006). The mammalian TGFβ isoforms (TGFβ1, TGFβ2, and TGFβ3) are secreted as latent precursors and signal through type I and type II TGFβ receptors (Figure 1.6). The type II receptor, to which a whole host of TGFβ ligands including activins, inhibins, bone morphogenic proteins (BMPs), and the TGFβ proteins themselves bind, associates with type I receptor to form a complex in which the former receptor phosphorylates the latter. A third membrane-anchored protein, known as type III receptor, helps this process by capturing and presenting the TGFβ activator to the signalling receptors I and II (Massagué & Gomis, 2006). The signal then proceeds via a series of intracellular SMAD proteins, which form a complex after activation, relocate to the nucleus, and associate with other DNA-binding transcription factors, as well as co-activators or co-repressors to regulate gene expression through the inhibition or activation of hundreds of target genes. Which genes are targeted is dependent on which type of ligand initiated the signal, the stage of development, and/or the cell type where the signal is being transduced (Annes, 2003).

The role of TGFβ signalling is implicated in skin homeostasis by its dual function as inhibitor of epithelial cell growth and activator of fibroblast proliferation and protein synthesis (Buschke *et al.*, 2011; Derynck *et al.*, 2001). Similarly, TGFβ is involved in controlling the composition of the ECM (Watabe & Miyazono, 2009). Several TGFβ signalling components are highly expressed in HF including TGFβ2, LTBPs and activated SMAD2 (Tumbar, 2004). In several mouse skin models, it has been shown that depending on time of expression TGFβ can act both as a tumour suppressor

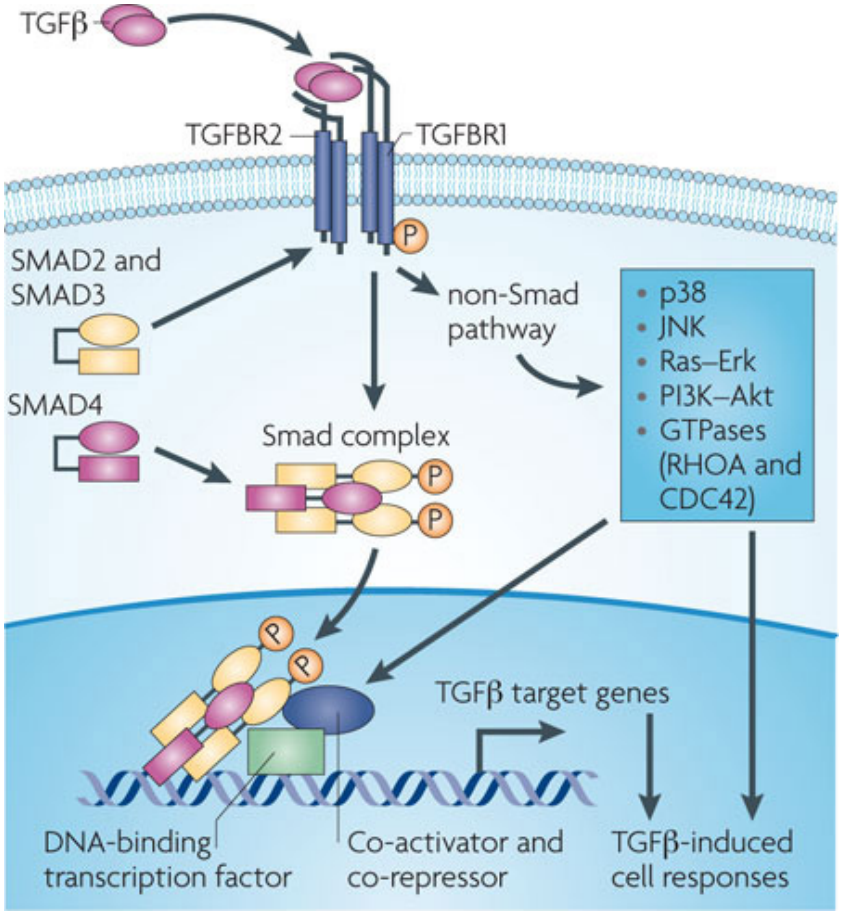

Figure 1.6: **TGFβ signalling pathway** : TGFβ signalling is transduced through Smad and non-Smad pathways. TGFβ ligand binds to receptors TGFBR2 and TGFBR1. TGFBR2 phosphorylates (P) TGFBR1, which subsequently phosphorylates and activates SMAD2 and SMAD3. Activated SMAD2 and SMAD3 form a Smad complex with SMAD4 and translocate into the nucleus. In the nucleus, the Smad complex interacts with other DNA-binding transcription factors, and co-activators and co-repressors, binds to the promoter regions of TGFβ target genes and regulates the transcription of target genes. TGFβ stimulation also activates other signalling cascades in addition to the Smad pathway. TGFβ receptors activate p38, JNK, Ras–Erk, PI3K–Akt, and small GTPases such as RHOA and CDC42 (from Ikushima & Miyazono (2010)).

and tumour promoter (Bierie & Moses, 2006; Oshimori & Fuchs, 2012). In one mouse skin model, it has been shown that TGFβ1 expression is rapidly induced in suprabasal keratinocytes *in vivo*, in response to hyperplasia, implicating signalling molecule in the regulation of epidermal homeostasis (Derynck *et al.*, 2001). Over-expression of TGFβ1 or TGFβ receptor type II in the skin of transgenic mice also

provided evidence for tumour suppressor activity and resistance to TPA-induced hyperplasia (Cui *et al.*, 1995; Wang *et al.*, 1999; Derynck *et al.*, 2001). Conversely, expression of a dominant negative TGFβ receptor type II eliminates TGFβ induced growth arrest, leading to epidermal hyperplasia (Wang *et al.*, 1999; Derynck *et al.*, 2001). In another study, it was shown that conditional expression of TGFβ1 in the interfollicular epidermis of mice at later stages promoted a rapid progression to metastasis (Weeks *et al.*, 2001). Altogether these studies strongly suggest that the TGFβ signalling pathway, next to BMP signalling, is one of the pathways that transmit quiescent and growth signals to epidermal SCs and their niche at various stages of development and adult life.

BMP Signalling Pathway

As already mentioned, BMPs belong to the same superfamily as TGFβs and function in tissue morphogenesis, homeostasis, and cancer by regulating diverse biological processes like proliferation, apoptosis, differentiation, and ECM production. Skin epithelial cells express receptors for both BMPs and TGFβs. Whereas the TGFβ pathway mediates its signal by phosphorylating SMAD2/3, BMP signalling operates through the phosphorylation of SMAD1/5/8. Each of these SMAD proteins joins SMAD4 to form a bipartite complex and modulate transcription on a large-scale. Although the general inhibitory effects of BMP are well documented, its role in the skin is only emerging. There is accumulating evidence supporting the view that the equilibrium between secreted BMP proteins and their inhibitors like noggin and gremlin plays an important role in HF morphogenesis during development but also activates HF SCs in adults. Loss- and gain-of-function studies in adult mice suggest that BMP signalling stimulates quiescence in bulge SCs. The cyclic pattern of BMP2 and BMP4 expression in the dermis fits with the observation that bulge SCs are in one of two states, resistant or responsive to activation (Plikus *et al.*, 2008).

Notch Signalling Pathway

When it comes to changing from epidermal growth to differentiation, several signalling pathways — some working in sequence, some in parallel — control the switch. Theses can merge into a common or at least partially overlapping downstream pathway or alternatively their cascade may proceed in parallel. The triggering signals can be biochemical, like induction by calcium or TPA in cell culture (Hennings *et al.*, 1980; Paolo Dotto, 1999), they can be physical cues initiated by the ECM through the BM (Connelly *et al.*, 2011) or they could be molecular like the transcription regulator p63 which works in conjunction with the canonical Notch pathway (Clark & Coker, 1998; Massagué & Gomis, 2006). Loss- and gain-of-function studies in vertebrates have revealed that p63 is essential for the proper stratification of skin and for maintaining the renewal potential of the various SCs in their distinct niches (Koster, 2004; Mills *et al.*, 1999; Truong *et al.*, 2006; Senoo *et al.*, 2007). Canonical Notch signalling is also crucial for the early switch from growth to differentiation when basal stem cells commit to spinous fate (Watt *et al.*, 2008). Notch functions broadly in specifying cell fates during differentiation and morphogenesis by modulating the transcription of many target genes. To summarize, Notch is a transmembrane protein with one extracellular and one intracellular binding domain. The Notch signalling pathway gets activated upon binding of a ligand to Notch's extracellular domain and the proteolytic cleavage and release of its intracellular domain (NICD) into the cytoplasm. Most of the effects of the NICD have been attributed to its ability to bind the transcriptional repressor RBP-J, resulting in the transcription of a number of target genes, most notably Hes and Hey, which are normally suppressed in the absence of Notch activity. Excessive Notch signalling induces basal stem cells to commit to a spinous fate (Rangarajan *et al.*, 2001). It has further been shown in mice, that loss of Hes1, an important Notch target in skin epidermis, alters differentiation (Moriyama *et al.*, 2008). Furthermore, in another experiment the conditional ablation of RBPJ, a DNA-binding protein that forms a heterodimer transcription factor with the Notch intracellular domain to relay the Notch signal to the nucleus, also blocks specification of spinous cell fate (Blanpain *et al.*, 2006).

Epigenetic Regulation

In recent years, several studies have attempted to understand and characterize the epigenetic switches that orchestrate the transition between SC proliferation and differentiation via changing transcription of many genes (Sen *et al.*, 2008; Frye *et al.*, 2007). In one study in mouse it has been shown that Myc is required in the epidermis for the stem cells to egress their SC niche and begin their proliferation and eventually switch to terminal differentiation. MYC regulates the transition of quiescent SCs to TA cells by inducing global histone modifications typically associated with active chromatin state and permissive for transcription factor binding (Frye *et al.*, 2007). Furthermore, another study has shown that epigenetic derepression of lineage-defining genes, specifically the removal of the H3K27me3 mark, is required for the proper commitment of epidermal SCs into the suprabasal and later granular epidermal layers (Sen *et al.*, 2008). Furthermore, this study showed that overexpression of specific histone demethylase known to interact with H3K27me3 caused premature activation of terminal differentiation in cultured human epidermal SCs (Sen *et al.*, 2008). Hence, epigenetic mechanisms are important, but remain to be comprehensively characterized.

1.1.4 Markers of stemness and differentiation

Access to markers that allow the identification of various populations of stem cells as well as their progression through differentiation has facilitated research in this field. As previously mentioned, some markers used to identify stem cells are different in mouse versus human skin. In this section, we shall introduce markers that are common to both animals in addition to human specific stem cell markers.

Slow-cycling SCs were identified in the bulge region of the HF through label retention assays (Blanpain & Fuchs, 2006). Earlier on it was thought these label-retaining cells in the bulge are *bona fide* SCs based since they were the longest-lived cells within the epidermis based on the result of lineage tracing experiments (Tumbar, 2004; Morris *et al.*, 2004). However, it is not known that quiescence is not the hallmark of SCs and IFE SCs continuously and repeatedly proliferate in order to

maintain the skin barrier (Blanpain & Fuchs, 2006). Since label retention assays would not work on IFE SCs — the label would dilute too rapidly — several different SC specific proteins have been identified and can be used as markers to differentiate between the distinct populations of epidermal SCs (Figure 1.7).

As was mentioned, the epidermis is separated from the dermis by the basement membrane (also known as basal lamina). The basal layer is characterized by its high expression of α6-integrin and β1-integrin, allowing it to adhere to the BM. In the basal layer, keratins 5 and 14 (KRT5 and KRT14) play the role of structural proteins. They assemble into 10-nm keratin intermediate filaments which, along with microtubules (tubulin) and microfilaments (actin), form the cytoskeleton of epithelial cells. When basal SCs enter the spinous layer, they stop dividing and synthesize a new set of proteins characteristic of cornification (Candi *et al.*, 2005).

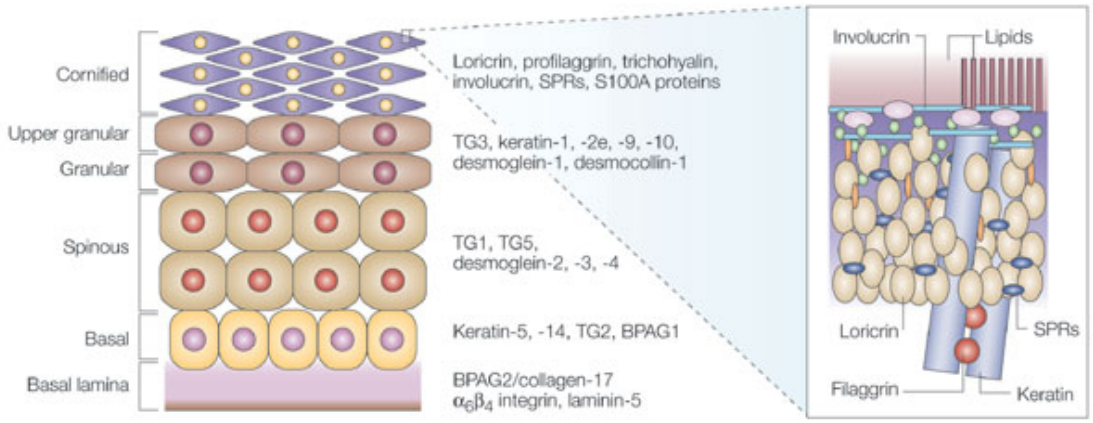

Figure 1.7: **Markers of stem cell populations** : Proteins particular to distinct locations in the epidermis during skin differentiation can be seen which enable apoptosis only in the basal layer and cornification only in the supra- basal layers. At the molecular level, the cornified envelope is formed by proteins that are highly crosslinked by transglutaminases, with specific lipids on the outside, to guarantee specific physical properties. BPAG denotes bullosus pemphigoid antigen, SPR denotes small proline-rich proteins, and TG is transglutaminase (from Candi *et al.* (2005)).

At the molecular level, the cross-linking of proteins with transglutaminases (TG1, TG3 and TG5) forms the impermeable cornified envelope. Concurrently, a series of other structural proteins, including involucrin (IVL), loricrin (LOR), trichohyalin

(THH) and small proline-rich proteins (SPRs), are synthesized and subsequently cross-linked by the aforementioned transglutaminases (TGs) just beneath the plasma membrane. These proteins get expressed in an increased gradient with the expression peaking in the cornified layer and that is why they are utilized as markers for differentiation (Candi *et al.*, 2005). Other proteins present in the cornified envelope include the S100 family of Ca_{2+}-binding proteins, which transmit Ca_{2+}-dependent cell-regulatory signals. S100A7, S100A10 and S100A11 are expressed in the basal and spinous layers and are substrates of TG1 and TG2, while S100A7 is present in the cytosol and in the endoplasmic reticulum. When the intracellular Ca_{2+} concentration increases, cytoplasmic S100A7 redistributes to α-actinin- and paxillin-containing peripheral complexes (Candi *et al.*, 2005).

1.2 Introduction to circadian rhythms

Since the advent of life on earth, terrestrial animals have evolved under predictable day-night cycles and almost all known light-sensitive organisms possess an intrinsic clock mechanism. There are great benefits for organisms that can anticipate cyclical changes in their environment and fine-tune their physiological functions and metabolic processes accordingly, affording them selective advantages as demonstrated experimentally (Woelfle *et al.*, 2004; Dodd *et al.*, 2005). This intrinsic clock mechanism, most commonly referred to as circadian rhythm (from the Latin circa diem, 'about a day'), has a period length of approximately 24 hours (Dunlap, 1999). In mammals, many aspects of physiology follow circadian rhythms, including sleep-wake cycles, hormone production, blood pressure, renal function, body temperature, and food intake and metabolism (Pittendrigh, 1993; Dibner *et al.*, 2010; Bass, 2012).

Circadian rhythms are entrained by several environmental Zeitgeber (time giver) signals such as feeding, temperature, and light. Light, the most dominant synchronizer, stimulates specialized photoreceptive neuronal ganglion cells in the retina of the eye. These cells in turn transmit the stimulus via the optic nerve to the central circadian pacemaker, the suprachiasmatic nucleus (SCN) (Panda, 2007).

In human and mouse, the SCN is composed of approximately 10,000 to 20,000 cells in the hypothalamus. The SCN in turn transmits the rhythmic information through neuronal and hormonal signals to all cells of the body while simultaneously integrating information from peripheral organs to generate consistent rhythms in the animal. Almost all peripheral tissues, including liver, heart, lungs, kidney and skin, possess their own peripheral clock so they can modulate tissue-specific gene expression, coordinated by transcription factors and chromatin remodellers, in a circadian manner (Asher *et al.*, 2008, 2010; Duong *et al.*, 2011; DiTacchio *et al.*, 2011). Numerous studies have profiled the transcriptome of various organisms to show that as much as 15% of all transcripts are rhythmically expressed throughout the day depending on the tissue analysed (Akhtar *et al.*, 2002; Ceriani *et al.*, 2002; Panda *et al.*, 2002; Storch *et al.*, 2002; McCarthy *et al.*, 2007; Doherty, 2010; Koike *et al.*, 2012; Menet *et al.*, 2012). Circadian rhythms are self-sustained, meaning that oscillations can persist even in the absence of external cues or Zeitgeber. For instance, when mice are kept under constant darkness for weeks or even months, they maintain their circadian rhythms albeit at times with slightly shorter periods (Dibner *et al.*, 2010). They are also cell-autonomous, since even cells in culture that have been propagated for years possess robust circadian oscillations which become detectable at a population level when they are synchronized by serum or dexamethasone shock (Balsalobre *et al.*, 1998; Balsalobre, 2000). On the molecular level circadian rhythms are controlled by transcriptional-translational feedback loops of several positive and negative regulators (Brown *et al.*, 2012).

Curiously, although as discussed, day-night cycles have remained stable since the emergence of life, core clock genes are not universally conserved. Although all investigated organism possessing an intrinsic clock, reveal a common model of transcriptional-translational feedback loop (TTFL), not many have homologous TTFL components. Even between human and fly — where they are numerous highly conserved biochemical processes — there are few homologous core clock genes namely the Period genes. In cyanobacterium, for instance the molecular clock is composed of KaiA, KaiB, and KaiC proteins only, which are not homologues of mammalian or fly clock components. The quest for a common source of evolutionary innovation has been the central question of a number of recent studies which

have proposed a transcription-independent universal circadian clock based on the oxidation state of peroxiredoxins (Edgar *et al.*, 2012; O'Neill & Reddy, 2012; O'Neill *et al.*, 2012).

1.2.1 Mammalian clock from mice to men

The underlying mechanism driving circadian rhythms at the molecular level is an interconnected series of negative feedback loops common to a wide array of light-sensitive organisms from cyanobacteria to mammals. Indeed, circadian oscillations are fine-tuned on multiple levels of transcriptional and translational regulation. The circadian pathways in human and mouse are extremely highly conserved. The main positive regulators are transcriptional factors BMAL1 (also known as ARNTL) and CLOCK, which are encoded by Brain and muscle ARNT-like 1 (BMAL1), and the Circadian locomoter output cycles protein kaput (CLOCK) gene. It has been shown that CLOCK is not essential and that CLOCK-deficient mice continue to exhibit more or less robust behavioural and molecular rhythms because a close homolog of CLOCK, NPAS2 encoded by Neuronal PAS domain-containing protein 2 (NPAS2) gene, is able to functionally act in its stead in the SCN to regulate circadian rhythmicity (DeBruyne *et al.*, 2007). These transcription factors, which are basic-helix-loop-helix, heterodimerize and bind to E-box elements in promoters of target genes to induce their expression. Among clock-controlled target genes are Cryptochrome 1 and 2 (CRY1/2), Period 1-3 (PER1, PER2 and PER3), RORα and REV-ERBα (NR1D1), which are an integral part of the circadian core machinery. As PER proteins accumulate in the cytoplasm, they are phosphorylated by CKIε and GSK3, which target them for ubiquitin-mediated protein degradation. At the same time, as CRY protein levels increase in the cytoplasm, they form stable complexes with PER proteins and translocate to the nucleus. Once in the nucleus, CRY/PER complexes repress BMAL1/CLOCK activity, resulting in the repression of their own transcription. The circadian core is further stabilized by opposing functions of RORα(activator) and REV-ERBα(repressor). Both transcription factors, which bind to RRE-elements in the BMAL1 promoter with RORα activating the transcription of BMAL1 and REV-ERBα inhibiting it (Gallego & Virshup, 2007; Sahar & Sassone-

Corsi, 2009). Aside from phosphorylation, other post-translational modifications (PTMs), including sumoylation and acetylation, have been shown to regulate the activity of clock proteins (Cardone, 2005; Hirayama *et al.*, 2007). Lastly, several chromatin-remodelling enzymes are associated with BMAL1/CLOCK complexes to modulate gene transcription, e.g. SIRT1, JARID1a and SIN3A (Asher *et al.*, 2008; DiTacchio *et al.*, 2011; Duong *et al.*, 2011). It has also been shown that the circadian master regulator CLOCK itself is a histone acetyltransferase (Doi *et al.*, 2006).

1.2.2 Detecting circadian gene expression profile

As already discussed, in several organisms circadian clock genes were found to be transcription factors. Because of this, many researchers turned to microarray and, more recently RNA-sequencing technology to survey the global regulation of gene expression as a function of time. Over the years, a number of algorithmic approaches have been used to detect genes with circadian expression. In this section, I will introduce the most prominent methods. Detecting circadian patterns in large data sets is a specialized problem, which requires specific and powerful statistical tests to discriminate between real cycling genes against a backdrop of noisy genes. Furthermore, precise and statistically reliable measures of various attributes of gene rhythms such as period length, phase, and amplitude are necessary. Approaches widely applied to this problem include Fourier analysis (Wichert *et al.*, 2003), a technique borrowed from signal processing, curve-fitting (Straume, 2004), autocorrelation (Levine *et al.*, 2002), and a recent nonparametric statistical method (Michael E Hughes, 2007). In a recent benchmarking study, it was demonstrated that the more high-resolution a time course data set is, the more powerful the detection mechanism with performance peaking for all of the above mentioned studies when samples were taken every hour over the course of at least two daily cycles (Michael E Hughes, 2007). Polynomial curve fitting algorithms can also be applied especially in cases where temporal samples are sparse and only peak detection methods can be employed (Lack & Lushington, 1996).

1.2.3 Circadian rhythms in the skin

The earliest experiments to link diurnal patterns to cellular processes in the epidermis did so by demonstrating that DNA synthesis and mitosis show circadian fluctuations suggesting a cyclic mode of epidermal cell proliferation in human and mouse (Schell *et al.*, 1981b,a, 1983). Epidermal SCs of the bulge and the IFE are heterogeneous in their circadian clock activity and this heterogeneity creates two distinct stem cell states resulting in different responses to activation and dormancy cues (Janich *et al.*, 2011). These SC populations go through active and quiescent stages in a circadian manner because several epidermal stem cell and homeostasis genes including some member of the Wnt and TGFβ pathways are directly under the control of the clock (Janich *et al.*, 2011). Furthermore, in the same study it has been shown that clock deficiency leads to decreased responsiveness, accumulation of dormant stem cells and increased tissue aging (Janich *et al.*, 2011). It has been shown that mouse IFE basal SCs peak in proliferative activity at night, while accumulation of ROS as a result of metabolic activity happens during the day in an antiphase manner to proliferation (Geyfman *et al.*, 2012). Furthermore, the progression rate of skin squamous tumours differs depending on exposure time to UVB radiation since skin is more susceptible to radiation at night when the cells are more proliferative (Gaddameedhi *et al.*, 2011; Geyfman *et al.*, 2012). Another recent work, has used microarrays to survey changes in human suction-blister epidermis obtained at three time point during the day along to identify a circadian transcription factor, Krüppel-like factor 9 (KLF9) as a candidate regulator of keratinocyte proliferation/differentiation. Gain- and loss-of- function experiments showed strong antiproliferative effects of Klf9. Putative Klf9 target genes include proliferation/differentiation markers that also show circadian expression *in vivo*, suggesting that Klf9 affects keratinocyte proliferation/differentiation by controlling the expression of target genes in a daytime-dependent manner.

1.3 Introduction to protein physical interactions

Much of cellular biology in the twentieth century has revolved around a reductionist approach. An approach that has ventured to decipher the behaviour of cells by deciphering the behaviour of the individual molecules that form them. In the form of genetics this approach has been immensely successful. Some of the most ground-breaking discoveries of the 19th and 20th centuries were made using classical genetics: the mechanism of inheritance (Mendel, 1865), the linear arrangement of genes along chromosomes (Sturtevant, 1913), that one gene somehow regulates the synthesis of one protein (Tatum & Beadle, 1942), that mutations can be artificially induced using radiation (Muller, 1927, 1928), that DNA contains genetic information (Avery *et al.*, 1944), and that the genetic code links the sequence of DNA to the sequence and ultimately three-dimensional structure of proteins (Nirenberg & Leder, 1964; Bernfield & Nirenberg, 1965).

In the past two decades, the vast amounts of data that have been generated have enabled us to study systems as well as individual components. The 'omics' revolution has re-highlighted the complexity that the chemistry of life entails. To understand the mechanism behind a process for example, one needs to understand the signal that triggers the process, the post-translation modifications that activate or deactivate various components involved in the process, and the interactions between the components. Furthermore, gaining a deep understanding of a cellular process, even a ubiquitous one, in a particular context, in one cell-type or during one specific stage of development, does not necessarily translate to all other contexts. Although comprehensive and informative, 'omics' data, still need to rely on the 'awesome power of genetics' (the favourite phrase of a former professor of mine) as a complementary approach.

One of the ways researchers in systems biology have dealt with the daunting task of organising, analysing, and understanding 'omics' data, has been to represent them as a simplified network of components (nodes) and relationships (edges). Components can be macromolecules like genes, mRNAs, proteins, or protein complexes and relationships can be physical, biochemical, or genetic interactions; they be

conceptual like functional annotations, or they can represent statistical metrics like similarity in phenotypic or transcription patterns (Vidal *et al.*, 2011). One type of network often used to simplify and study the cell are protein-protein interaction (PPI) networks. In the next few sections, I introduce how PPI networks are detected experimentally or obtained through literature curation and what approaches have been used for their analysis in various model organisms.

1.3.1 Detecting protein-protein interactions

Systematic and unbiased detection of protein-protein interactions applied at the scale of whole genomes or proteomes has been underway in a number of model organisms. Two methodologies in particular are currently in wide usage for large-scale mapping. Mapping of binary interactions (detecting direct interaction between two proteins) is carried out by various flavours of the yeast two-hybrid (Y2H) system, while mapping of membership in protein complexes (detecting direct and indirect associations between several proteins) is carried out by affinity- or immuno-purification or liquid-chromatography followed by mass spectrometry (IP-MS, AP-MS and/or LC-MS).

Yeast Two-Hybrid

Devised 25 years ago (Fields & Song, 1989), the yeast two-hybrid (Y2H) methodology takes advantage of properties of the yeast protein GAL4, a transcriptional activator of enzymes required for galactose metabolism. GAL4 has two essential but distinct domains: an N-terminal DNA-binding domain (BD) that binds the promoter of genes to be transcribed in a specific region called GAL upstream activation site (UAS_G), and a C-terminal activation domain (AD) which binds other factors involved in transcription initiation. Since the DNA-binding and the transcriptional activation functions of GAL4 are separable, the two domains can be put into two separate hybrid proteins. For example, the GAL4 DNA-binding domain can be fused to protein X while its activating domain can be put in protein Y. If proteins X and Y form a protein-protein interaction, then the two domains will reconstitute when

they come into close proximity of each other and the transcription of any gene under the control of a UAS_G can occur (Figure 8). To scale up this methodology to high-throughput (HT), proteins from any source can be cloned into expression plasmids, transfected into yeast cells, and tested against each other as bait-prey pairs. If chimeric proteins X and Y — bait and prey — do indeed physically interact, the transcription of a reporter gene will get activated (Figure 1.8).

This method, albeit powerful is prone to both type I and type II errors. If the fusion of BD or AD to their respective chimeric proteins blocks the interaction surface and impedes physical association of bait and prey, the result is a false negative. Other false negative errors stem from the fact that certain interactions may not take place in yeast, the typical host organism for Y2H. For example, a bacterial or mammalian protein may need a particular chaperone to fold properly or a PTM before interacting with its target and if these are unavailable in yeast, no interaction occurs. More importantly, a source of false positive error for Y2H is that although an interaction is detected *in vitro*, it may never occur *in vivo* due to differences in sub-cellular localization and tissue- or time-specific expression. These false positives, referred to as biological false positives, are nearly impossible to identify using interaction assays alone. Technical false positives however, can occur in any experimental system. Since the early days of Y2H, the rate of technical false positives has decreased substantially by improvements such as using low copy-number plasmid vectors and retesting interaction pairs (Cusick, 2005). Another source of false positives are strong auto-activators, baits (BD-X) that turn on transcription even before they interact with their prey (AD-Y). These can be discarded easily by checking for reporter gene expression before the prey is added. More challenging are latent or weak auto-activators that arise due to accumulation of mutations in the bait during propagation of bait containing yeast cells. Auto-activators appear as promiscuous baits with many interaction partners, often lacking any common functional annotation. Computational methods can be employed to remove this class of erroneous hits from the final data set by imposing a cut-off on the total number of interactions allowed. Finally, the Y2H system can only be used to study soluble proteins and hence cannot be used to detect interactions between insoluble integral membrane proteins.

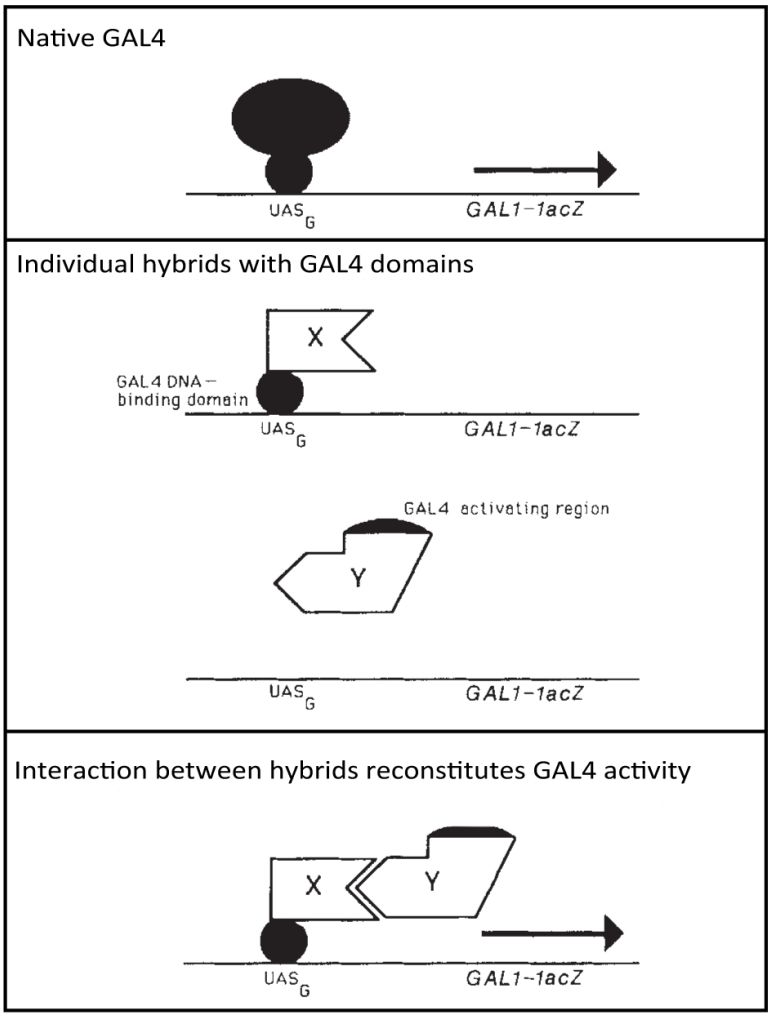

Figure 1.8: **Model of transcriptional activation by reconstitution of GAL4 activity** : The native GAL4 protein contains both DNA-binding and active regions and induces GAL1-lacZ transcription (top panel). Hybrids containing either only the DNA-binding domain (upper part of middle panel) or only activating region (lower part of middle panel) are incapable of inducing transcription. A protein-protein interaction between X and Y brings the GAL4 domains into each others vicinity and results in transcriptional activity (of often a reporter gene) (from Fields & Song (1989)).

Early genome-wide Y2H studies contained a significant number of technical false positives, yet the high proportion of false positives did not necessarily diminish their impact and despite these limitations, since the year 2000, Y2H has been used widely

and applied to whole proteomes to systematically test all pairwise combinations of proteins in several model organisms (Rain *et al.*, 2001; Li, 2004; Uetz *et al.*, 2000; Ito *et al.*, 2001; Stelzl *et al.*, 2005; Rual *et al.*, 2005; Giot, 2003). These HT studies are estimated to have uncovered only a relatively moderate portion of the interactome — the complete collection of all physical protein-protein interactions that can occur within a cell — in each organism. For instance, the first two published HT Y2H yeast studies (Uetz *et al.*, 2000; Ito *et al.*, 2001) together discovered just under 5,000 PPIs, a number that is thought to be a mere 10-15% of the estimated total possible interactions (von Mering *et al.*, 2002) and even though both studies used the same 6,000 ORFs as baits, there was only 15% overlap in detected PPIs. Naturally this poor overlap has raised concerns that Y2H data are noisy (von Mering *et al.*, 2002). Even more alarming, the two individual studies had less than 13% overlap with a set of high-confidence PPIs curated from single-gene biochemical studies (Costanzo *et al.*, 2000). As stated, there are several sources of type I and type II errors in Y2H experiments, which can partly explain the low overlap between data sets. Another explanation, placing a positive spin on the low overlap between, for instance the two yeast mentioned studies, is that neither study had reached saturation and so different protein interactions may have been sampled (von Mering *et al.*, 2002).

Tandem Affinity-/Immuno-Purification and Mass Spectrometry

While Y2H detects direct binary protein-protein interactions, affinity purification methods identify components of stable complexes. The basic strategy involves purification of a protein complex using an affinity tag placed on one of the components of the complex to pull down the entire complex and then characterizing all components by mass spectrometry. This method was proposed as a generic procedure to purify proteins expressed at their natural levels and was given the name tandem affinity purification (TAP) tag (Rigaut *et al.*, 1999). The procedure entails fusing two affinity tags — usually a peptide or small protein — to a target protein and then introducing the construct to a host cell. There are various flavours of TAP tag but the original method consisted of an Immunoglobin G (IgG) binding domain of the bacterium Staphylococcus aureus protein A and a calmodulin binding peptide (CBP)

separated by a TEV protease cleavage site (Rigaut *et al.*, 1999). The target protein with the TAP tag first binds to beads or columns coated with the antibody IgG. The TAP tag is then broken apart by an enzyme, TEV protease, which recognizes the TEV protease cleavage site and the target protein and its interaction partners are removed by washing with a solvent. In the second step, the material eluted from the IgG beads are incubated with calmodulin-coated beads and calcium. After several wash steps, to remove contaminants like the TEV protease that may have lingered behind from the first affinity selection, the bound material can be released using EGTA (Figure 1.9). After the target protein has been washed through two affinity columns, it can be examined for binding partners using mass spectrometry. The advantage of using two tags significantly reduces non-specific background, as compared to a single tag approach (Puig *et al.*, 2001).

One obvious drawback of TAP tag is that tagging may disrupt complex formation (von Mering *et al.*, 2002). Another experimental consideration to be made is that retrieved peptides from MS cannot always be mapped back uniquely to their correct proteins because many proteins are highly similar with only slight differences arising from polymorphisms, alternative splicing and PTMs (Schlüter *et al.*, 2009). Furthermore, with this strategy, a protein that is known to be a member of one complex may be purified with another complex. This may be a valid case of a protein involved in multiple functions or intracellular communication between complexes, or it might represent a contaminant (Cusick, 2005). As more territory is covered in the interactome, more components shared by multiple complexes with differing functions will be discovered (Krause *et al.*, 2004) further augmenting the challenge of assigning function based on co-purification strategies. Another pitfall is that assignment of a component to a particular complex often relies on experimental stringency and arbitrary thresholds. For example, in three large-scale studies employing a purification strategy (Ho *et al.*, 2002b; Gavin *et al.*, 2006; Krogan *et al.*, 2006), to be discussed in the next section, three distinct interaction confidence measures were devised, thresholds on those confidence measures were selected empirically, and the resulting PPI networks were clustered into protein complexes using different methods, each custom-optimized. In short, the bespoke analysis pipelines in many HT co-purification studies introduce variability to the resulting PPI network.

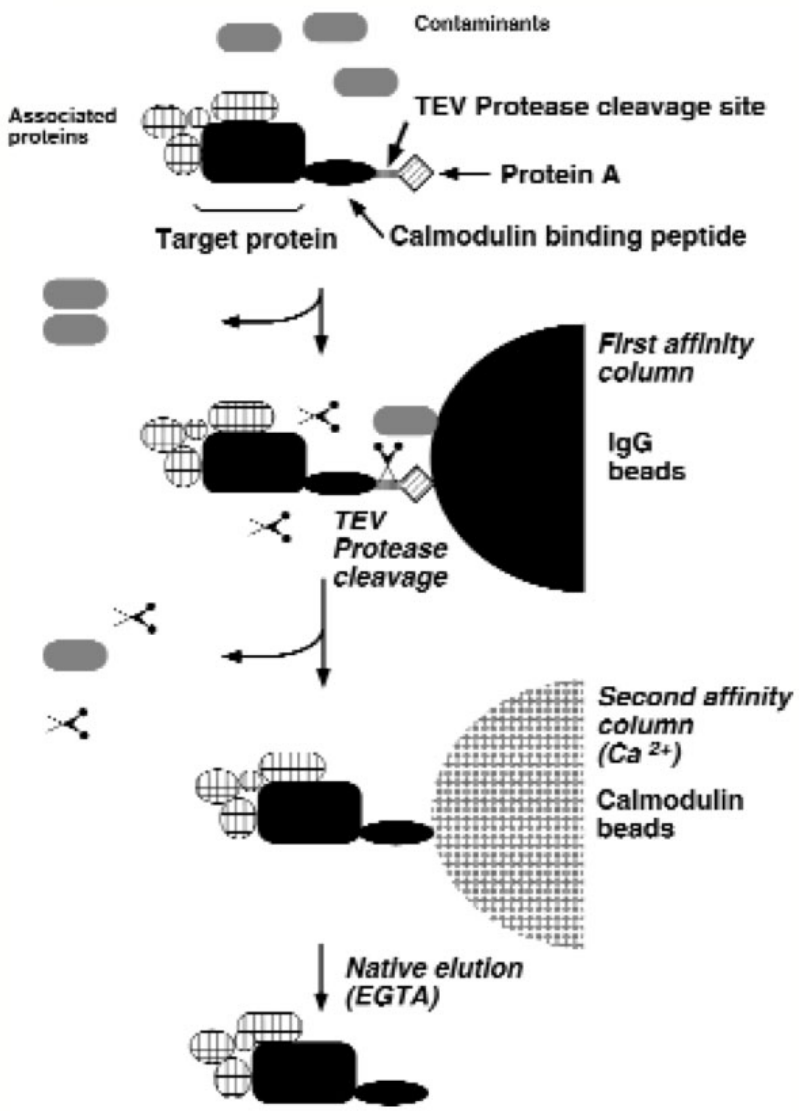

Figure 1.9: **TAP tag methodology** : Overview of the TAP procedure with two wash steps and the Immunoglobin G (IgG) binding domain and antibody IgG (from Rigaut *et al.* (1999)).

A very closely related strategy to AP-MS is immuno-precipitation/purification followed by MS (IP-MS or CoIP-MS), which works through an antibody that targets a known endogenous protein, believed to be a member of a larger complex. By targeting this known member with a high-affinity antibody one could pull down

the entire protein complex out of solution provided the protein subunits are tightly associated with one another. The major difficulty with immunoprecipitation and its main difference with AP is that for each single protein to be isolated endogenously a highly specific antibody has to be generated. To circumvent this obstacle, many groups have engineered tags to introduce into their protein of interest as explained above. However, while the use of TAP tagging makes pull-down experiments quite convenient it has raised some concerns regarding the biological relevance of this approach as the exogenous tag may hinder native interactions or introduce new ones.

In spite of these limitations, both methods, and AP-MS in particular, have been widely applied to map protein complexes in yeast (Ho *et al.*, 2002a; Gavin *et al.*, 2006; Krogan *et al.*, 2006), fly (Guruharsha *et al.*, 2011), and Arabidopsis (Arabidopsis Interactome Mapping Consortium *et al.*, 2011). In human too, AP-MS has been employed to characterize the composition of protein complexes in experiments utilizing anywhere between dozens to hundreds of unique baits (Ewing *et al.*, 2007; Sowa *et al.*, 2009; Behrends *et al.*, 2010; Hutchins *et al.*, 2010). Likewise, immunoprecipitation has been applied to identify endogenous complexes from human cell lines (Malovannaya *et al.*, 2011). Nonetheless, the considerable successes achieved in the comprehensive identification of protein complexes in model organisms listed above remain unparalleled in mammalian cells mainly due to limited availability of high quality antibodies. This has hindered efforts in scaling up AP-MS or IP-MS experiments and validating in an unbiased and systematic manner all PPIs in human cells.

Liquid chromatography and Mass Spectrometry

Another approach complementary to AP/IP-MS, which does not rely on antibodies, is liquid chromatography followed by MS. In brief, liquid chromatography refers to a broad range of techniques that rely on chemical properties of molecules such as their mass and polarity in order to separate or fractionate them out of a mixed solution as a function of how fast/slow they move through two phases: a stationary

phase and a mobile phase (Lescuyer *et al.*, 2004). In the present day, next generation liquid chromatography known as high performance liquid chromatography (HPLC) is widely used whereby very high pressure and very small packing particles (in a column) are used for fractionation of samples. In this technique, the sample is forced by a liquid at high pressure (mobile phase) through a column packed with irregularly shaped and spherical particles (stationary phase) designed specifically to accomplish particular types of separations. The components of the sample move through the column at different velocities based on the specifics of their physical interaction with the stationary phase. The velocity of each component depends on its chemical properties, the properties of the stationary phase and the composition of the mobile phase. The time at which a specific component emerges from the column is known as retention time. Differences in retention time allow for the various components of a sample to be collected separately as they elute the column and are key to the success of the technique.

Perhaps the most flexible and widely used method for protein fractionation in biochemical studies is ion exchange HPLC (IEX-HPLC). This method is suitable for the separation of complex biological samples such as whole cell lysates prior to mass spectrometry as it maintains good resolution. In IEX-HPLC, the resolution of a mixture of proteins is achieved based on the differential retention of distinct proteins to the charged surfaces of the stationary phase, due to differences in the surface charge properties of the proteins themselves (Havugimana *et al.*, 2007). Effective separation depends on the sequential elution of the proteins bound to the stationary phase by way of application of a salt gradient to the mobile phase (Havugimana *et al.*, 2007). Individual fractions can then be collected in a timed manner as they elute from the column and subsequently subjected to basic LC-MS procedures for protein identification. In a recent study, a combination of several biochemical fractionation technologies, primarily IEX-HPCL, was used to comprehensively identify $\sim$ 14,000 endogenous human PPIs grouped into 622 putative protein complexes in cytoplasmic and nuclear extracts (Havugimana 2012).

Literature curation

In addition to the experimental platforms discussed, compilation or curation of existing data published in scientific studies is another way of mapping protein-protein interactions. Curating the literature to gather PPIs from low-throughput studies is a complementary approach to Y2H and AP/MS (Vidal *et al.*, 2011). Literature-curated interaction maps are advantageous as they use already available information. However, they are limited by the variable quality of individual studies and the lack of reporting of negative results (Cusick *et al.*, 2009). Of databases providing protein complex annotations discovered by low-throughput affinity-based methods, CORUM for human (Ruepp *et al.*, 2008, 2010) and CYGD for yeast (Guldener, 2004) are noteworthy. There are numerous databases dedicated to the compilation of PPIs detected by HT methods in various organisms, which we will discuss in the next section.

1.3.2 Mapping protein complexes through data integration

Despite massive efforts in experimentally detecting PPIs and complex memberships, none of the experimentally derived data sets of protein physical interactions are truly comprehensive. For this reason, many attempts have been made to integrate data from numerous sources, combining Y2H, AP-MS, other affinity-based methods as well as low-throughput data in order to get a more complete picture. Various data repositories including BIND (Bader, 2003), MIPS (Pagel *et al.*, 2005), DIP (Salwinski, 2004), BioGRID (Stark *et al.*, 2006), HPRD (Mishra, 2006), MINT (Chatr-aryamontri *et al.*, 2007), IntAct (Kerrien *et al.*, 2007), APID2NET (Hernandez-Toro *et al.*, 2007), MPIDB (Goll *et al.*, 2008), PINA (Wu *et al.*, 2008), STRING (Franceschini *et al.*, 2012), and HIPPIE (Schaefer *et al.*, 2012) are all devoted to routinely integrating and organizing data and providing confidence scores for each interaction. These scores usually take into account the number and type of experimental evidence and typically give less weight to high-throughput experiments like Y2H than reproduced affinity-based assays.

In parallel, computational prediction efforts have utilized many sources of infor-

mation including conserved protein and DNA sequences, domain co-occurrence, phylogenetic profiles, co-expression, functional annotations, co-localization or homologous interactions in other species to complete the PPI network of individual organisms (Cusick, 2005). These computational prediction methods are efficient to implement and usually generate networks with large numbers of nodes and edges. However, as they rely on indirect information, they only serve as a complementary source to experimentally verified data (Vidal *et al.*, 2011). And so systematic HT detection of PPIs shall continue until the interactomes of important model organisms have been completely mapped.

1.3.3 Making biological discoveries through integration of protein interaction data with other data sources

Aside from data integration to obtain a more complete interactome, making sense of the interactome network, incomplete as it may be, is a major challenge in systems biology. This task requires applying scalable data mining techniques and incorporating even more data sources in order to gain new insights. Discovering functional modules often relies on mathematical techniques. Strategies such as hierarchical and k-means clustering, factorization, and principle components analysis are just a few of the standard methods employed. In addition, investigating the properties of the network itself — connectivity, degree distribution, cliques, and hubs — has enabled important biological insights like the observation that many biological networks have power-law node degree distribution (Barabási & Oltvai, 2004; Vidal, 2001). In this section, I will give an overview of integrative computational approaches to the study of protein complexes which were either directly detected in AP/MS experiments, curated from literature, or indirectly extracted from HT binary PPI data.

One of the first studies that demonstrated the utility of integrating PPI and transcriptome data for the purpose of studying transcriptional regulation, accomplished the task in yeast by extracting subnetworks based on the coherent expression patterns of their genes (Ideker *et al.*, 2002). The study first provided a proof of princi-

ple experiment involving the integration of PPI data and gene expression changes measured in response to a single perturbation — deletion of GAL80 — compared to wild type to show that the extracted differentially expressed subnetworks have high overlap with well-studied galactose-induction regulatory pathways described in the literature. Further, the authors applied this methodology to large networks with expression data across multiple perturbations to identify subnetworks with significant changes over particular subsets of conditions (Ideker *et al.*, 2002). In a related study, high-confidence yeast PPIs and microarray datasets were combined and again represented as a graph, with the distinction that edges bore the weight of expression dissimilarity or distance between genes (Chen & Yuan, 2006). With a new clustering algorithm called betweenness partitioning, the work identified and cross-validated dozens of functional modules that conferred similar deletion phenotypes (Chen & Yuan, 2006).

Around the same time another integrative yeast study was published that went beyond just extracting functional modules, by exploring a biologically relevant hypothesis (de Lichtenberg *et al.*, 2005). This work combined time-course gene expression data measured during the yeast cell cycle with protein complexes inferred from PPI data to demonstrate that many protein complexes become active at specific points during the cell cycle as a consequence of the dynamic expression of one or a few of their subunits. In addition to gene expression data, this study has also incorporated information on subcellular localization in order to filter out interactions between proteins annotated to incompatible compartments. The resulting network was suggestive of just-in-time complex assembly where tight regulation of one or a few complex components regulates the activity of the whole complex rather than just-in-time synthesis as thought previously.

Data integration into PPI networks proved to be so useful that several large-scale studies also employed it to validate the quality of their data and provide preliminary analyses. For example, additional data sources were integrated with the output of HT AP/MS experiments (Ho *et al.*, 2002b; Gavin *et al.*, 2006; Krogan *et al.*, 2006). In one paper (Gavin *et al.*, 2006), genome-wide AP/MS assays were conducted using individual protein baits in exponentially growing yeast. Using clustering, pro-

tein complexes were identified as those proteins that tend to co-purify together. Furthermore, protein complexes were sub-classified into cores, proteins that always co-purify together in all purifications of a protein complex, and attachments, proteins that only sometimes co-purify with a protein complex. This study demonstrated that even in a single snap-shot during the yeast cell cycle, protein complexes could assemble differently due to the presence or absence of varying attachments. This partitioning in the data was studied by integration of many more data sources including information on subcellular localization, protein three-dimensional structures, Y2H-inferred binary interactions, and gene expression data to show that the perpetually present cores are more likely to be co-localized, physically interacting, and co-expressed at the same time during the cell cycle (Gavin *et al.*, 2006). It has to be noted, however, that similar ideas on the distinction between core and attachment subunits by way of organizing protein complex subunits into sub-modules were published much earlier in a small-scale theoretical study (Dezso *et al.*, 2003).

In a second study published the same year, affinity purification tagging was used to purify 4000 distinctly tagged proteins, followed by two mass spectrometry methods in parallel to increase coverage and confidence (Krogan *et al.*, 2006). Comparable to the first study, interaction scores were assigned and a Markov clustering (MCL) algorithm (Enright *et al.*, 2002) was used to identify protein complexes from the generated PPI network while additional data like hand-curated MIPS complexes and Gene Ontology annotations were used for cross-validation. Unlike in the first study, no subclasses were defined within protein complexes. However a similar partitioning was seen in the data whereby 40% of protein complexes contained subunits shared between several complexes.

In another similar study in human, a comprehensive PPI network was obtained by integrating data from a number of different databases. This network was then combined with expression data measured over a set of 79 human tissues. One interesting finding of this study was that the most tissue-specific proteins, only expressed in a narrow subset of tissues, interact with constitutively expressed proteins and, conversely, nearly all constitutively expressed 'housekeeping' proteins have some interactions with very tissue-specific proteins (Bossi & Lehner, 2009).

This study suggests that the action of proteins involved in core cellular processes can be modified to function in specific niches as a result of their interactions with specialized proteins.

A recent study has managed to comprehensively catalogue human protein complexes, where data integration is carried out at the same time as experimentation (Havugimana *et al.*, 2012). The method has used massive co-fractionation followed by mass spectrometry to isolate and identify soluble protein complexes. In this work, experimental data were integrated with several lines of supporting evidence using machine-learning methods in order to build a high-confidence human PPI network and protein complexes were then derived from the network using the cluster growth algorithm ClusterONE (Nepusz *et al.*, 2012). In addition, various data sources were incorporated to analyse protein complexes and to show their subunits are enriched for biological processes, transcriptional regulatory motifs, pathological processes, and post-translational modifications compared to all proteins captured. Furthermore, many of the complexes were found to have subunits which either had RNAi-induced phenotypes in human cell culture or were orthologous to genes associated with mutant phenotypes in mouse, yeast, or worm. Furthermore, subunits of the predicted human protein complexes were also much more likely to have links to diseases as gathered from UniProt, the Genetic Association Database (GAD), and Online Mendelian Inheritance in Man (OMIM) (Havugimana *et al.*, 2012).

In recent years, protein interaction networks have been combined with other data to understand human pathology. In one study, a network of over 8,000 orthologous proteins in humans, rats, and mice was manually curated from the literature and combined with gene expression changes in the blood in response to an inflammatory stimulus (Calvano *et al.*, 2005). The findings of the work provided details into the regulation of global white blood cells in the framework of the immune system. Two other groups have also shown that changes in protein interaction networks can be a predictor of breast cancer prognosis (Chuang *et al.*, 2007; Taylor *et al.*, 2009a). In one study, gene expression data from two cohorts of breast cancer patients — those whose cancers metastasized and those whose did not — were combined with PPI data in order to discover differentially expressed or co-expressed subnetworks

and use classifiers to predict disease progression (Chuang *et al.*, 2007). A second study too, incorporated the same two patient data sets used by Chuang with human protein interaction data in order to predict a binary favourable/unfavourable outcome for each patient sample. However, unlike the previous study, instead of extracting clusters, the researchers used the global properties of the interaction network, specifically hubs and patterns of co-expression between them and their interacting partners (Taylor *et al.*, 2009b).

1.3.4 The interactome in 3-D

The existence of hubs with numerous interaction partners within PPI networks suggests that all interactions cannot occur simultaneously. Looking at PPIs in terms of their three-dimensional structures can be informative in teasing out which interactions can occur simultaneously and which ones are mutually exclusive. The first study to incorporate 3D structure information into PPI analysis did so by exploring the intersection between all yeast protein complexes with resolved 3D structures and yeast interactions identified by Y2H (Aloy & Russell, 2002). Given a known 3D complex structure and homologous sequences for each one of the interacting proteins, the proposed method in the study ranked all possible interactions between homologues in the same species based on empirical potentials of the known structures. This method was applied to over 2,500 protein interactions in yeast and although a small number could be mapped onto the set of interacting complexes with known 3D structures and even fewer possible interactions could be inferred, this study opened the door for similar strategies to be used. Not long after, another study in yeast attempted to characterize proteins competing for access to the same structural interfaces (Kim *et al.*, 2006). This study integrated domain information from Pfam and structural information from PDB with a high-confidence yeast interaction network (Kim *et al.*, 2006). The work specifically focused on hubs, defined as proteins with five or more interactions, and — by looking at which domains supported each interaction — classified hubs into multi-interface or single-interface. The major finding of this study was that multi-interface hubs, capable of supporting multiple interactions simultaneously are overwhelmingly essential compared to

single-interface hubs (65% compared to 32%) and they evolve at a slower rate (measured by Ka/Ks). Furthermore, it was observed that multi-interface hubs and their interacting neighbours tend to be co-expressed at a higher frequency (25%) compared to single-interface hubs and their interaction partners (17%) — a difference that is statistically significant. This is an intuitive observation since it is reasonable that the interaction partners of single-interface hubs are not co-expressed, because if they were they would have to compete for binding to the same single interface (Kim *et al.*, 2006). These findings for the first time provided a structural explanation of expression dynamics for hubs and their interacting partners.

Today the idea that inter-protein competition is an important consideration in PPIs is accepted (Stein *et al.*, 2011). In recent years several studies have incorporated structural information specifically focusing on interacting interfaces to determine how binary interactions take place within experimentally characterized protein complexes or signalling pathways (Kiel *et al.*, 2011; Yang *et al.*, 2012; Kiel *et al.*, 2013). These studies have elucidated whether proteins compete with each other and form mutually exclusive interactions or bind to different interfaces of the same protein in compatible fashion and have simplified these categories by applying logic gates to each, namely OR to the former and AND to the latter. In one study in human, integration of proteomics data and protein structural information together with literature mining was added to the signal transduction network important to the function of rhodopsin (Kiel *et al.*, 2011). To study this pathway crucial to human vision, proteins were superimposed with their interacting domains and then classified into mutually exclusive or mutually compatible interactions using structural data via a platform called SAPIN (Yang *et al.*, 2012). This structurally annotated pathway enabled the researchers to outline the order of events during signalling by specifically working out which interactions are mutually exclusive due to a common binding interface (Kiel *et al.*, 2011).

CHAPTER

TWO

HUMAN EPIDERMAL STEM CELL FUNCTION IS REGULATED BY CIRCADIAN OSCILLATIONS

Peggy Janich[5]*, Kiana Toufighi[1,2,4]*, Guiomar Solanas[1,2]*, Nuno Miguel Luis[6], Susann Minkwitz[7], Luis Serrano[1,2,3,4]#, Ben Lehner[1,2,3,4]#, Salvador Aznar Benitah[1,2,3]#

1 Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain.
2 Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain.
3 Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluís Companys 23, 08010 Barcelona, Spain.
4 EMBL-CRG Systems Biology Unit, CRG.
5 Center for Integrative Genomics, Faculty of Biology and Medicine, University of Lausanne, Lausanne, Switzerland.
6 Buck Institute for Research on Aging, 8001 Redwood Boulevard, Novato, CA 94945, USA.
7 Julius-Wolff-Institut, Charité Campus Virchow Klinikum, Augustenburger Platz 1, D-13353 Berlin, Germany.

* Co-first authors.
# Co-corresponding authors.

Janich P, Toufighi K, Solanas G, Luis NM, Minkwitz S, Serrano L, Lehner B, Benitah SA. Human epidermal stem cell function is regulated by circadian oscillations. Cell Stem Cell. 2013 Dec 5;13(6):745-53. doi:10.1016/j.stem.2013.09.004

CHAPTER

THREE

DISSECTING THE CALCIUM-INDUCED DIFFERENTIATION OF HUMAN PRIMARY KERATINOCYTES STEM CELLS BY INTEGRATIVE AND STRUCTURAL NETWORK ANALYSES

Kiana Toufighi[1,2], Jae-Seong Yang[1,2], Nuno Miguel Luis[3], Salvador Aznar Benitah[1,2,4]#, Ben Lehner[1,2,5]#, Luis Serrano[1,2,5]#, Christina Kiel[1,2]#

1 EMBL Systems Biology Research Unit and Centre for Genomic Regulation, Dr. Aiguader Street 88, 08003 Barcelona, Spain.
2 Universitat Pompeu Fabra, 08003 Barcelona, Spain.
3 Buck Institute for Research on Aging, 8001 Redwood Boulevard, Novato, CA 94945, USA.
4 Institute for Research in Biomedicine, Parc Científic de Barcelona, Carrer Baldiri Reixac 10, 08028 Barcelona, Spain. 5Institució Catalana de Recerca i Estudis Avançats, Passeig Lluís Companys 23, 08010 Barcelona, Spain.

Co-corresponding authors.

Toufighi K, Yang JS, Luis NM, Aznar Benitah S, Lehner B, Serrano L, Kiel C. Dissecting the calcium-induced differentiation of human primary keratinocytes stem cells by integrative and structural network analyses. PLoS Comput Biol. 2015 May 6;11(5):e1004256. doi: 10.1371/journal.pcbi.1004256

CHAPTER

FOUR

CONTEXT-DEPENDENT PLASTICITY IN HUMAN PROTEIN COMPLEXES

Kiana Toufighi[1,2], Luis Serrano[1,2,3]#, Ben Lehner[1,2,3]#

1 EMBL Systems Biology Research Unit and Centre for Genomic Regulation, Dr. Aiguader Street 88, 08003 Barcelona, Spain.
2 Universitat Pompeu Fabra, 08003 Barcelona, Spain.
3 Institució Catalana de Recerca i Estudis Avançats, Passeig Lluís Companys 23, 08010 Barcelona, Spain.

Co-corresponding authors.

4.1 Abstract

The composition of human protein complexes is not rigid but is likely to differ depending on, for example, the cell type, cell cycle, and physiological conditions. We reasoned that this variation has partly been captured in current protein complex annotations because many protein complexes have been studied by independent research groups under different conditions. Hence, as the compendium of annotated human protein complexes steadily increases in size, we can exploit this experimental variation as a feature to learn about the context-specific behaviour of proteins and protein complexes. Here, we classify human protein complex subunits into those in the 'core' (subunits shared by multiple annotated complexes) and those in the 'periphery' (subunits exclusive to one or only few complexes) and demonstrate quantitatively that the two classes show differing evolutionary and disease-related attributes and tissue-specific expression patterns. Our results indicate that at the core of human protein complexes there exists a class of ubiquitously expressed subunits which are more likely to be co-expressed, are more evolutionarily conserved, and are more likely to be essential than their peripheral counterparts. Members of the peripheral class, in turn, have higher disorder content and are more likely to belong to the same gene family. From this, a picture emerges in which the function of a stable protein complex core is modified by the attachment or detachment of periphery proteins that allow protein complexes to function in a plastic, context-dependent manner.

4.2 Introduction

Many cellular processes are performed by macromolecular machines that consist of protein complexes built from one to several dozen individual polypeptides, whose functions range from transfer and processing of information to building molecular structures to mediating metabolic reactions (Bray, 1995). The interactions between and within these molecular machines are at the heart of the modular nature of the cell and are crucial to its form and function (Rain *et al.*, 2001; Alberts, 1998).

In the last decade and a half, considerable efforts have been made to comprehensively map the network of protein-protein interactions using mainly two general methodologies: a) yeast-two-hybrid (Y2H) assays and b) affinity-/immunopurification or liquid-chromatography followed by mass spectrometry. While the latter class of methods is capable of isolating stable protein complexes in their entirety, the former system surveys binary protein interactions and thus additional computational steps are required to predict protein complexes from binary interactome data. Irrespective of technical differences, both approaches have been used in an array of model organisms, including bacteria (Rain *et al.*, 2001; Butland *et al.*, 2005; Shimoda *et al.*, 2008; Kuhner *et al.*, 2009), yeast (Giot, 2003; Uetz *et al.*, 2000; Ito *et al.*, 2001; Ho *et al.*, 2002a; Gavin *et al.*, 2006; Krogan *et al.*, 2006), fly (Giot, 2003; Guruharsha *et al.*, 2011), worm (Walhout *et al.*, 2000; Li, 2004), higher plants (Arabidopsis Interactome Mapping Consortium *et al.*, 2011) and more recently human (Stelzl *et al.*, 2005; Rual *et al.*, 2005; Havugimana *et al.*, 2012). Together these data have provided insights into the organization of protein complexes, their evolutionary conserved topologies (Fraser, 2005; Kim *et al.*, 2006), their role in disease (Vidal *et al.*, 2011), and how cellular functions are modularized within the global interactome. In addition, they have enabled more rapid functional annotation of unknown proteins through guilt-by-association approaches (Oliver, 2000).

Notably, a number of studies have implemented the idea of organizing protein complex subunits into sub-modules. This was performed first in a small-scale theoretical study (Dezso *et al.*, 2003) and later on applied to large-scale data by Gavin *et al.* (Ho *et al.*, 2002b) who performed genome-wide affinity purifications using individual protein baits in exponentially growing yeast. An index was defined to quantify the likelihood of interaction between proteins and then clustering was employed to identify protein complexes. Even in the specific context that was considered — a single snapshot of the yeast proteome averaged over cell cycle phases — considerable heterogeneity in protein-complex composition was observed. To study this heterogeneity, proteins in complexes were classified into two types: 'core' components always co-purifying together in all purifications, and 'attachments', only sometimes co-purifying with the cores. In addition, another subclass of attachment proteins, which were always present together in multiple complexes, was defined as

'modules'. This natural partitioning in the data was studied by integration of many more data sources, revealing that proteins within cores tend to be co-expressed during the cell cycle and sporulation; that within cores and modules, proteins are more likely to be co-expressed, to be co-localized to the same cellular compartments, and to be annotated to the same function; that proteins within distinct cores and modules are more likely to be present or absent together as orthologs in distant species; and that proteins within cores and modules are most likely to be in direct physical contact as inferred by 3D structures and Y2H data (Gavin *et al.*, 2006).

In a second study published at the same time, Krogan also used affinity purification tagging to purify proteins and then used two mass spectrometry methods in parallel to increase coverage and confidence (Krogan *et al.*, 2006). They also assigned interaction scores and used clustering to identify protein complexes. Although unlike in the Gavin study, no subclasses were defined within protein complexes, a similar partitioning was observed in the data whereby out of 547 identified protein complexes, about 40% contained shared subunits, which participated in multiple complexes. Precisely identifying biologically relevant modules from protein interaction networks is not a trivial task, yet in recent years, several approaches, inspired by these yeast data sets, have attempted to formulate and discover these modules, specifically core and periphery structures within them (Palla *et al.*, 2005; Derényi *et al.*, 2005; Luo *et al.*, 2009).

Of the $>$ 20,000 proteins encoded in the human genome, about a fifth are currently annotated as members of protein complexes in the curated public database CORUM (Ruepp *et al.*, 2010). Data from large-scale high-throughput studies have deliberately been excluded from CORUM, thus making it a source of high-confidence information on whole complexes reported in individual biochemical studies. Although far from complete, this moderately sized subset exhibits interesting structured relationships. An example is the one-to-many membership of proteins to protein complexes (Ruepp *et al.*, 2010), which arises in part because many related protein complexes have been investigated by several laboratories independently. In part, differences in reported complex composition may stem from variation in technical procedures such as the use of more or less stringent wash steps or from the use of

different antibodies. However — and more interestingly — they may also disclose real biological variation that exists in the composition of complexes as they regulate specific processes at different times or in different places. Much like the interactome has been shown to be dynamic, exhibiting temporal (de Lichtenberg *et al.*, 2005), condition- or tissue-specific changes (Bossi & Lehner, 2009), the complexome or the compendium of protein complexes is also likely to be changing in composition in response to varying cellular states.

We hypothesized that we could exploit the variation in subunit compositions identified in different studies to learn about the compositional plasticity of human protein complexes. Complexes reported in the CORUM database derive from a wide range of experimental set-ups, cell and tissue types and physiological conditions, and therefore constitute a rich source of information to investigate how the interplay between invariant protein-complex cores and dynamic protein-complex periphery, an interplay which makes molecular machines compositionally plastic by multiplying functionality.

4.3 Results

4.3.1 Human protein complexes have highly overlapping subunits

As noted by others (Ruepp *et al.*, 2008; Havugimana *et al.*, 2012), one of the characteristics of human protein complexes is that many have a high degree of protein subunit overlap (Figure 4.1A). This can be seen in a global map of protein complexes, where visually distinguishing one protein complex from the other is often difficult because of the high degree of connectivity (Figure 4.1A). We reasoned that there is likely to be valuable information in this overlap between the subunits reported for different protein complexes that can be used to provide insights into cellular organization.

For each protein complex pair *i, j*, we calculated the corresponding Simpson coeffi-

cient $s_{i,j}$, which quantifies the overlap in protein members between the complexes on a scale of 0 to 1. Protein complexes and their pairwise Simpson coefficients were represented as a network with protein complexes connected if their Simpson coefficient is greater than 0.5 (see Figure 4.1B-C, Figure S1A, and Methods). To capture cliques — complete or fully connected sub-networks — of overlapping human protein complexes, we used a clique-finding method (Derényi *et al.*, 2005). Only considering dimers or larger complexes defines a set of 394 cliques encompassing 722 protein complexes and 1,364 unique proteins. Of these, 420 proteins appear in at least two different complexes within the same clique at least once ('core' subunits) and the remaining 944 never appear in more than one protein complex of any one clique ('periphery' subunits). Henceforth we refer to this classification as 'set 1'. In order to show our results are robust to how cliques are defined we also analysed a second set, filtering out complexes with less than three subunits to yield 335 cliques encompassing 616 protein complexes and 1,329 unique proteins, 387 of which are classified as core subunits and the remaining 942 as periphery subunits (referred to as 'set 2').

4.3.2 Core proteins are more essential and more evolutionarily conserved than periphery proteins

We then compared various properties of core and periphery subunits to see how they differ. First, we assessed the essentiality of core and periphery proteins. We took advantage of a recent study (Marcotte *et al.*, 2012) in which systematic loss-of-function screens were conducted in 72 breast, ovarian, and pancreatic cancer cell lines using an shRNA library targeting 15,000 human genes, resulting in the identification of almost 300 highly essential human genes (in at least 50% of the 72 cell lines their loss-of-function phenotype scored as lethal). In the absence of a gold standard set of human essential genes, we asked whether there is a difference in the proportion of core proteins that were represented in this data set compared to that of periphery proteins. Whereas 17% of core subunits (n = 420) from set 1 scored as essential in this assay, only 10% of peripheral subunits (n = 944) did (Figure 4.2A, P = 5e-04, Fisher's exact test.). Further, the proportion of core genes that have

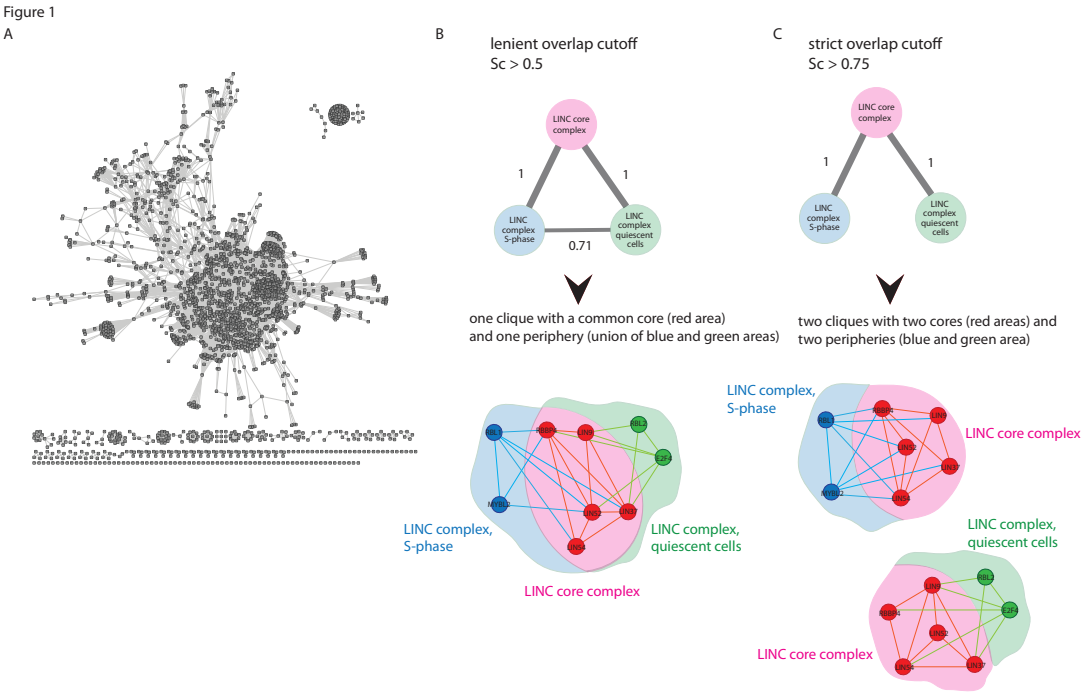

Figure 4.1: **Illustration of high connectivity among CORUM protein complexes** : (A) Network representation of all human protein complexes in CORUM where each protein complex must at least be dimer. There are 2,316 proteins represented as nodes in this network. With the exception of several dozen isolated protein complexes at the bottom, the rest of the network is highly connected indicating the extent to which protein subunits belong to and are shared by multiple protein complexes. (B, C) "LINC complex, S-phase" (blue, ID5593) and "LINC complex, quiescent cells" (green, ID5596), purified independently in two separate studies, share a common set of proteins (red) defined as the "core" in this study. Incidentally, this core itself is annotated as an independent "LINC core complex" (red, ID5589) in CORUM. Simpson coefficients (Sc) are defined based on the magnitude of overlap between any two complexes (see Methods), resulting in a simple network with protein complexes as nodes connected by Sc-weighted edges. Depending on the Sc cut-off, some edges may be pruned, a more lenient cut-off (B) leads to the identification of one clique whereas a stricter cut-off (C) yields two distinct cliques.

been implicated in heritable disorders with high-penetrance phenotypes is higher in the core set, 23% (n = 420) compared to 17% (n = 944) in the peripheral protein subunits (Figure 4.2B, P = 0.008, Fisher's exact test). These enrichments are very similar when considering protein complexes in set 2 (Figure 4.2A-B, P = 5e-04, P = 0.009, Fisher's exact test). We also compared evolutionarily sequence conservation of core versus periphery subunits by computing a per protein conservation

score for one-to-one orthologs between human and other species. We observed that core genes tend to be somewhat more conserved. Compared to mouse, core subunits (n = 372) have a mean conservation score of 0.91 versus 0.89 for periphery subunits (n = 825), (P = 0.0043). In chicken, the mean conservation score of core subunits (n = 324) was 0.78 compared to a mean of 0.73 of periphery subunits (n = 710), (P = 1e-04). In fly, the mean conservation score of core subunits (n = 194) was 0.49 compared to a mean of 0.46 of periphery subunits (n = 339), (P = 0.038) (Figure S3A).

Finally, we computed the percentage of different groups (namely cores, peripheries, clique complexes, non-clique complexes, and all complexes), which were present in their entirety in at least 200 other species as defined by prokaryotic and eukaryotic clusters of orthologous groups, COGs and KOGs. We found that out of our set of cliques, even when we controlled for set size, 14% of cores were present all together compared to only 7% (P = 0.01, Fisher's exact test) of peripheries (Figure 4.2C). Notably, the level of cross-species conservation, among cores is significantly higher than among protein complexes of the same size that do not participate in cliques — 14% compared to 3% (P = 3.17e-07, Fisher's exact test). This suggests that highly utilised core components that are present in several protein complexes are more highly conserved than protein complexes of the same size, which do not partake in protein community structures and are presumably involved in specialized functions (Figure 4.2C). Further, the observation that individual cores are also more likely to be totally present in a given species than entire protein complexes (P = 1.09e-05) implies a close functional interdependence of these components over higher levels of organization like the complex itself (Figure 4.2C).

4.3.3 Periphery proteins are more likely to be paralogs of each other

The proportion of protein subunits in the periphery that are duplicates of one another — as defined by EnsemblCompara GeneTrees (Vilella *et al.*, 2009) — is 27% compared with 20% of core subunits, a difference of 26% for set 1 (P = 0.0034, Fisher's

Figure 2

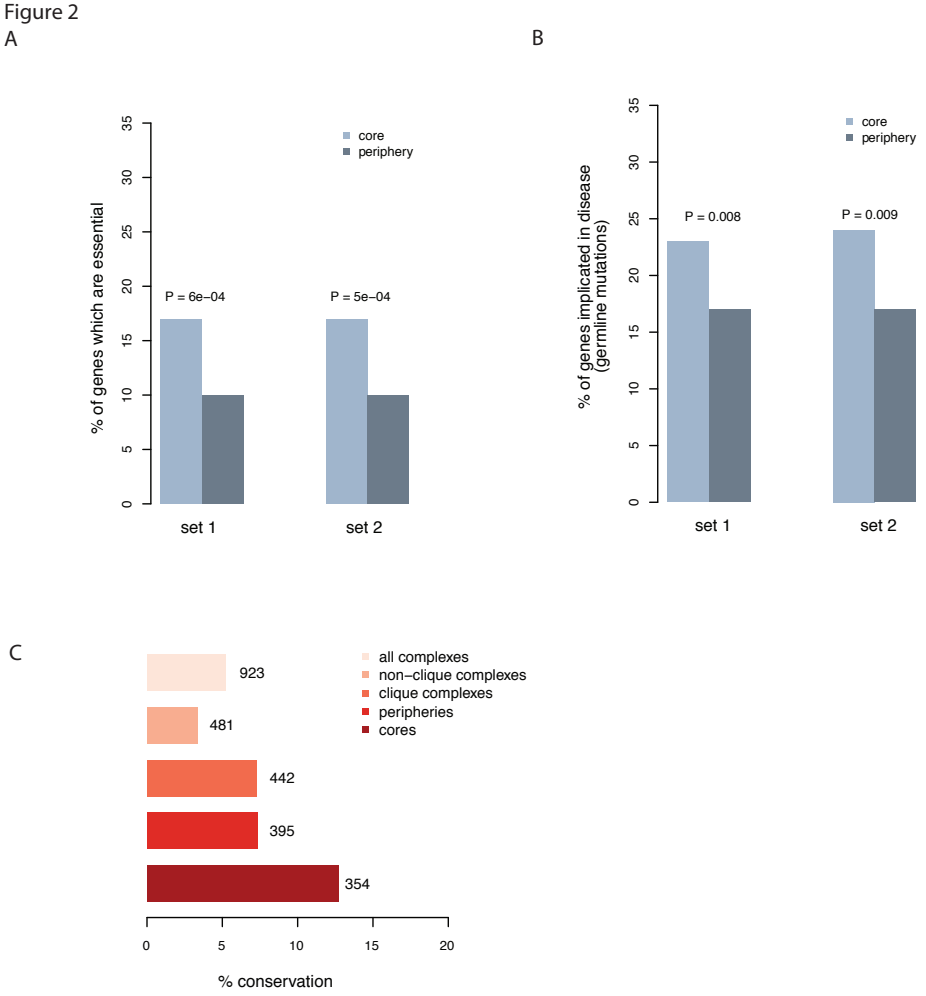

Figure 4.2: **Core subunits are more likely to be essential and more conserved than periphery subunits** : (A) A higher proportion of core compared to periphery subunits are essential: 17% (n = 420) compared with 10% (n = 944) when considering set 1 (P = 5e-04), and 17% (n = 387) versus 10% (n = 942) when considering set 2 (P = 6e-04). (B) The proportion of core subunits implicated in heritable diseases is higher than that of periphery subunits: 23% (n = 420) compared with 17% (n = 944), (P = 0.008) in set 1. Likewise, core and periphery subunits derived from set 2 show a difference in enrichment for genes involved in disorders caused by germ line mutations: 24% (n = 387) compared with 17% (n = 942), (P = 0.009). (C) Among clique components, core components are most likely to be present all together in other species as defined by COGs and KOGs than others. For example 14% of cores are present together as COGs and KOGs compared to 7% of peripheries (P = 0.01) and 5% of whole complexes (P = 1.09e-05). The numbers next to the bars are effect sizes. Statistical significance for all of the above tested using Fisher's exact test.

exact test) and 28% compared with 18%, a difference of 36% for set 2 (P = 1e-04, Fisher's exact test) (Figure 4.3A). However there is no difference between core and periphery proteins subunits in the size of the paralagous gene families to which the proteins belong (Figure S3B). In addition, there is no significant difference in the age of the duplications when comparing the two classes (Figure 4.3B). We also found no significant difference in protein-family size — as defined by Interpro (Hunter *et al.*, 2011) — between the two groups (see methods, Figure S3C). This implies that the enrichment of paralogs we observe in the periphery class is not because the paralogs come from larger or older families. Rather, there is higher diversity of small groups of paralogs — mostly pairs, triples and quadruplets. Furthermore, peripheral paralogs for the most part (~75%) map to the same protein complexes defined by the same co-purification experiment. It has been proposed that gene duplications are one mechanism by which protein complexes frequently evolve (Pereira-Leal *et al.*, 2007; Finnigan *et al.*, 2012), with frequent conservation of protein interactions following duplication (Pereira-Leal, 2005). Our results suggest that gene duplication has made a quantitatively larger contribution to the evolution of the peripheries of human protein complexes.

4.3.4 Clique components, in particular cores, are highly co-expressed across tissues

We obtained genome-wide expression data measured in 16 distinct human tissues using RNA sequencing technology (Bradley *et al.*, 2012) and quantified the extent to which core and peripheral subunits differ in their mRNA expression . First we observed that the 357 expressed core genes have a higher average mRNA abundance compared to the 755 expressed periphery genes (means of 64 and 43 and medians of 16 and 13 RPKM for the core and periphery genes, respectively, P = 2.7e-10, Wilcoxon rank sum test. RPKM: reads per kilobase per million). We then compared the breadth of expression of the two classes and found that both are widely expressed — both core and periphery genes are present in an average of 15 tissues out the total of 16 where a gene's presence is defined as an mRNA abundance of $>$ 1 RPKM (Figure S4A, no significant difference in distributions, P = 0.06, Wilcoxon

Figure 3
A B

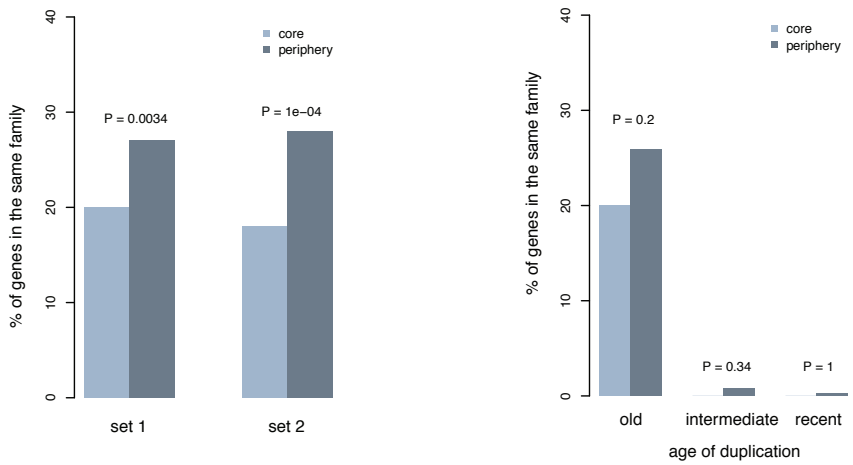

Figure 4.3: **Periphery genes are more likely to be composed of paralogous pairs or groups** : (A) In set 1, only 20% of core genes ($n = 420$) were duplicates of each other compared with 27% ($n = 944$) of periphery genes, a 26% change, $P = 0.003$. For set 2 again 18% ($n = 387$) of core proteins subunits were paralogous to one another compared with 28% ($n = 942$) of periphery proteins, a 36% change, $P =$ 1e-04. Statistical significance tested using Fisher's exact test. (B) Time of duplication of paralagous core genes and paralagous periphery genes are not significantly different from one another; 20% of core genes ($n = 420$) have old time of duplication compared with 26% of periphery genes ($n = 944$), ($P = 0.2$), 0% of core paralaogs ($n = 420$) have intermediate age of duplication compared with 1% periphery paralogs ($n = 944$), ($P = 0.34$), and 0% of core paralogs ($n = 420$) were duplicated recently compared with 0.27% of periphery paralogs ($n = 944$) ($P = 1$). None of these differences are significant. Statistical significance for both A and B tested using Fisher's exact test.

rank sum test).

Next, we used the Pearson correlation coefficient (PCC) of co-expression of genes within defined components of cliques to identify whether their associations are context-specific (that is, genes within each component are not always co-expressed) or universal (that is, genes within each component are always co-expressed). Core genes in our set of 394 cliques revealed a PCC distribution skewed toward the positive (mean=0.56) and significantly shifted in comparison to that of periphery

genes (mean=0.46, P = 5.03e-34, Wilcoxon rank sum test). A similar discord is observed between cores and peripheries from the same protein complexes — cores are the same as clique cores but peripheries are sets of subunits that together with their cores make up one unique protein complex. Within protein complexes too core genes show higher average co-expression as measured by PCC (mean = 0.56) compared to periphery subunits (mean=0.49, P = 3.3e-23, Wilcoxon rank sum test). Indeed, the PCC distribution of core genes is binomial with the majority of core genes centring on 0.82, which indicates a tight co-regulation for many core genes (Figure 4.4A). Further, when we analysed protein complexes separately, we found a similar heterogeneity in expression to peripheral genes, while also significantly different from core genes (mean=0.46, P = 1.3e-25, Wilcoxon rank sum test) (Figure 4.4A). This differential expression across tissues, in protein complexes as a whole and specifically in their peripheral components contrasted to the tight co-expression of their core components, supports the notion that context-specific regulation of periphery genes can modify the function of protein complexes through association with the stably expressed core (Figure S4D).

Finally we compared core and peripheral subunits within each clique, separating them into paralogs and non-paralogs. Core proteins are more positively correlated in expression across the tissues than periphery proteins independent of whether they have duplicates or not (Figure S4B, P = 0.09 for paralogous subunits, P = 2.5e-43 for non-paralogous subunits).

4.3.5 Periphery proteins tend to be larger and more disordered

We also contrasted the physical and chemical properties of the two classes of subunits. We found that on average periphery proteins tend to be larger than core proteins with the 917 core proteins having a mean amino acid length of 755 compared to an average sequence length of 676 for the 412 core proteins (P = 0.004) (Figure 4.5A). Periphery proteins have on average a very slight enrichment in the number of Pfam domains (Alberts 2002); 917 periphery proteins have on average 2.3 Pfam domains compared to 1.9 of the 412 core proteins (P = 0.046) (Figure

Figure 4

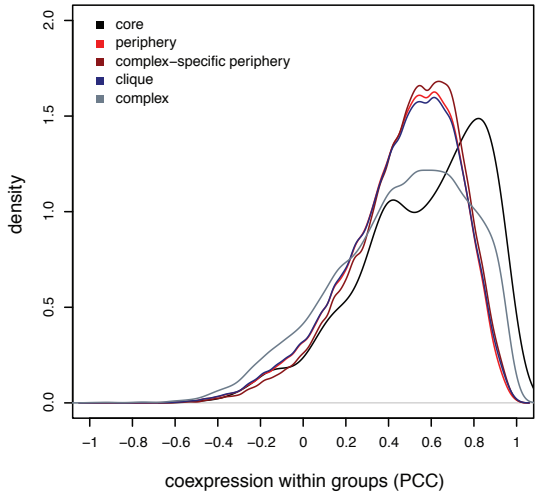

Figure 4.4: **Core and periphery protein subunits display similar yet distinct patterns of expression across human tissues** : (A) Comparison of expression correlation between various components within 394 cliques derived from complexes composed of one or more subunits; mRNA abundance cutoff > 1 RPMK across 16 human tissues. Expressed core genes (n = 399) show significantly higher expression correlation amongst themselves (mean=0.56) than expressed periphery genes (mean=0.46, P = 5.03e-34) or expressed periphery genes separated into groups corresponding to protein complexes from which they originated (mean=0.49, P = 3.3e-23) or the entire set of clique genes (mean=0.46, P = 1.9e-33) or the entire set of protein complexes (mean=0.46, P = 1.3e-25). Statistical significance tested using Wilcoxon rank sum test.

4.5B). We then looked at disorder content to not only classify the two groups further but to also understand whether there could be a difference in the number and type of physical interactions each class is capable of. Given that there is a strong positive correlation between disorder content and protein size (Figure 4.5C) we compared normalized intrinsic protein disorder between the two classes. We found periphery subunits tend to be more disordered (22% compared to 19% mean disorder content, P = 0.002). The higher disorder content in periphery genes is suggestive of a scenario in which periphery proteins are more often involved in lower affinity or transient interactions. Furthermore, recent evidence suggests that intrinsic disorder is necessary for proper complex assembly and correlates with protein-complex size (Alberts 2002). Thus peripheral proteins, as a consequence of having higher

intrinsic disorder, might be more flexible to form heterologous interfaces with each other and other core proteins during complex assembly.

Given that disordered regions are more likely to undergo post-translational modifications (Russell & Gibson, 2008), we also examined phosphorylation, ubiquitination, sumoylation, acetylation, methylation, and O-linked glycosylation using two public databases (Beltrao *et al.*, 2012; Hornbeck *et al.*, 2011), but found no detectable difference in the distributions of the number of modifications the proteins in each group undergo (Figure S5, S6). Furthermore, we looked at protein half-life data (Schwanhäusser *et al.*, 2011) and confirmed that there is no difference in protein turnover rate given the result of the PTM analysis (Figure S6D).

4.4 Discussion

Many biological functions are carried out by the activity of highly interacting cellular components, often referred to as functional modules. In the present study, we investigated the properties of one type of module: protein complexes identified in small-scale experiments. Our results are consistent with a model where many protein complexes have an invariant core interacting with variable periphery subunits. The present study confirms much of the existing knowledge pertaining to modular sub-organisation of protein complexes in yeast while it also highlights the utility of such model in a multicellular organism through the integration of tissue-specific expression data.

Our results broadly mirror prior bioinformatics analyses of protein complexes in yeast, which found that protein subunits of stable protein complexes tend to be more highly conserved and twice as likely to be essential than proteins involved in more transient temporary interactions (Dezso *et al.*, 2003; Kim *et al.*, 2006). In addition, protein complex cores as an entity have remained conserved together to a greater extent than peripheries or even specialised protein complexes of comparable size, which are not part of protein complex cliques. It has also been shown in yeast that protein complex cores — defined as complex subcomponents consistently

co-purified across multiple experiments employing different protein baits — were co-expressed, co-localized, and more likely to physically interact (Ho *et al.*, 2002b). Our study demonstrates many of the same principles in human, specifically that sub-modules within protein complexes exhibit different levels of heterogeneity in expression across various tissues, with core components displaying tight co-regulation in expression and periphery components providing tissue-specific variability to the protein complex as a whole, thus providing context-dependent flexibility. We also show that proteins in the periphery tend to have more intrinsic disorder than proteins in the core. This may be because peripheral proteins engage in a higher number of transient interactions. Intrinsic disorder has been linked to complex size as it is thought that structural disorder is necessary for proper protein complex assembly (Hegyi *et al.*, 2007) since protein flexibility — predicted using a simple ratio of a protein's solvent-reachable surface area to what is expected given its molecular weight and related to protein intrinsic disorder — is required for complex assembly and especially crucial to large and cyclic protein complexes (Marsh & Teichmann, 2014).

Given that the number (~400) of core subunits is small, a plastic exchange of attachment proteins in the periphery provides a modular and efficient scheme for diversifying function in a temporally or spatially specific manner according to the physiological needs of the cell. Consistent with this view, we found the periphery to be more diverse in its catalogue of paralogous proteins. In other words, since the periphery shows a higher proportion of paralogous families and since in general there is no large difference between paralogous family sizes in core and periphery, the implication here is that the periphery combines a greater diversity of paralogous families. This supports the notion that periphery subunits may have emerged through gene duplications consistent with previous studies on evolution of protein complexes (Pereira-Leal, 2005; Pereira-Leal *et al.*, 2007; Finnigan *et al.*, 2012). Another interesting finding surrounding paralagous proteins in the periphery is that they are no more likely to be co-expressed than any pair of non-paralogous periphery proteins from the same clique. In fact ~75% of paralagous proteins in the periphery co-exist together in the same protein complexes. Therefore, since paralogous proteins often join together to assemble a single protein complex, their expression variability across tissues is no greater than non-paralogs. Taken together,

our results suggest that human protein complexes are highly plastic and that this plasticity is provided through the flexible attachment of periphery proteins to a stable core, diversifying function in a temporally or spatially specific manner according to the physiological needs of the cell.

4.5 Methods

4.5.1 Protein complex network generation and clique finding

A reference set of complexes was obtained from the CORUM database (Ruepp *et al.*, 2008) of curated mammalian protein complexes. As of the latest release on (February 2012) there are 1,331 human complexes in total, 1287 with at least two subunits, 931 with at least three. All analyses were performed in parallel on the latter two sets. Human complexes consisting of at least two or at least three distinct protein subunits (i.e. set of 1,287 and 931 complexes) were then examined in a pairwise all-by-all manner for their magnitude of overlap with each other. Overlap was computed using Simpson coefficient for any two complexes c_i and c_j:

$$SC = \frac{c_i \bigcap c_j}{min(|c_i|, |c_j|)}$$

where the normalization factor is the size of the smaller complex. We then created a network of protein complexes between which edges are drawn only when a predefined level of overlap (shared subunits) is met. In our study, we imposed a cutoff of 0.5 on the Simpson coefficient and only connected protein complex nodes that meeting it. This cutoff — employed in another study as well (Havugimana *et al.*, 2012) — is optimal when analysed as a function clique compactness. With strict cutoff of $SC > 0.8$, protein complex overlap is so high that very little alternative subunits exist resulting in small or non-existent peripheries, while lax SC cutoffs result in virtually no common core to analyse (Figure S1).

We then used a described method, CFinder (Derényi *et al.*, 2005) to recover fully connected subgraphs or cliques — that is communities where all protein complex nodes have at least 50% overlap with one another. We achieve this by extracting

parts of the aforementioned protein complexes network that are highly connected with edges meeting the overlap threshold. These connected subgraphs called clusters, modules, communities or cliques do not have a universally-accepted definition, yet their very presence is a hallmark of the hierarchical nature of real networks (Barabasi & Albert, 1999; Albert & Barabási, 2002; Dorogovtsev & Mendes, 2002) . Although clique finding is a NP-complete problem meaning that it has an efficiently verifiable solution but not necessarily an efficient solution (Cook, 1971), in real networks, which often tend to be sparse or less dense than their theoretical counterparts, numerous efficient solutions have been proposed (Derényi *et al.*, 2005; Blatt *et al.*, 1996; Girvan & Newman, 2002; Radicchi *et al.*, 2004; Newman, 2004). Utilizing a recent powerful method (Lechler & Fuchs, 2005), we extract the most highly overlapping cliques of protein complexes — that is communities where all protein complex nodes have at least 50% overlap with each other. Once we have identified all such protein complex nodes, we can partition them into core subunits (i.e. subunits common to all nodes in a particular clique) and periphery subunits (i.e. subunits exclusive to one node in a particular clique).

4.5.2 Conservation analysis

To study conservation using identity scores, selected set of orthologous sequences, as provided by Ensembl (Flicek *et al.*, 2012), were aligned using three different programs: MUSCLE v3.8 (Edgar, 2004), MAFFT v6.712b (Katoh & Toh, 2008), and DiAlign-TX (Lassmann *et al.*, 2009). Alignments were performed in forward and reverse direction (i.e using the Head or Tail approach (Landan & Graur, 2007)), and the six resulting alignments were combined using M-Coffee (Wallace, 2006). Resulting alignments were used to compute identity scores of orthologs to human reference proteins by using trimAl v1.4 (Capella-Gutierrez *et al.*, 2009). We also calculated the percentage of different classes of core, periphery, etc. that were present together in at least 200 out of the 417 species represented in the clusters of orthologous groups, COG/KOG data sets (Tatusov, 1997), downloaded from the STRING database (Franceschini *et al.*, 2012). Since cores seldom contain more than 2-3 genes (Figure S2A), while peripheries are much larger (Figure S2B), the

probability of all peripheral genes within a clique to have been conserved together as a group will be lower. To control for this, we look at individual core/periphery within protein complexes and analyse one set of attachment proteins within the set of peripheral subunits. We additionally control for size by comparing core and peripheries with similar size distributions.

4.5.3 Protein family size

We used two data sets of protein family definitions from Uniprot (Magrane & Consortium, 2011) and Ensembl (Flicek *et al.*, 2012). The first database assigns protein family membership using Interpro (Hunter *et al.*, 2011) while the latter relies on EnsemblCompara GeneTrees (Vilella *et al.*, 2009), derived gene families across all metazoan sequences. We use these two sets to check for differences between core and periphery genes in terms of size of protein families and found no significant difference between their size distributions. Using this alternate set of protein family annotations, which complements one-to-one paralogs, we found that on average core paralogs come from families of 3.5 genes and periphery paralogs come from families of 3.7 genes (median=3) for both groups). In fact when we check the distribution of family sizes between core and periphery subunits we observe that paralogs for the most part come from families of the same size with triplicates, quadruplicates and quintuplicates more or less evenly representing 60% of duplicate families (Figure S3B). In fact when we check the distribution of family sizes between core and periphery subunits we observe that paralogs for the most part come from families of the same size with triplicates, quadruplicates and quintuplicates more or less evenly representing 60% of duplicate families.

4.5.4 Additional data sets

A set of essential human genes was obtained from this study (Marcotte *et al.*, 2012). Information on protein disorder content was used in a previous study (Vavouri *et al.*, 2009) originally identified using Globprot (Linding, 2003) and DisEMBL (Beltrao & Serrano, 2005). Information on mRNA and protein half-lives was obtained from a

large-scale pulse-chase study in mouse (Schwanhäusser *et al.*, 2011). Complete data sets on post-translation modifications including acetylation, methylation, o-linked glycosylation, phosphorylation, sumoylation, and ubiquitination came from PhosphoSitePlus database (Hornbeck *et al.*, 2011) in addition to a data set of primarily acetylation, phosphorylation, and ubiquitination provided by Pedro Beltrao from PTMfunc database (Beltrao *et al.*, 2012). Protein domain information was downloaded from Pfam (Punta *et al.*, 2012). RNA-seq data obtained from 16 various human tissues by the Illumina Human BodyMap 2.0 project (Bradley *et al.*, 2012) (HBM) (www.illumina.com; ArrayExpress ID: E-MTAB-513) served as the basis for the analysis of expression correlation across tissues. Disease-causing mutations came from OMIM (Hamosh, 2004) and appropriate keyword filters were applied to obtain a gene list whose germline mutations lead to phenotypes with high penetrance (Mendelian disorders).

4.5.5 Statistical analyses

All statistical tests and plots were performed in R version 3.0.2.

4.6 Acknowledgments

We would like to acknowledge Salvador Capella for conducting high-quality sequence alignments to generate ortholog identity scores. We thank Tobias Warnecke for discussions and comments on the manuscript. This work has been funded by the PROSPECTS, grant agreement number HEALTH-F4-2008-201648, to L.S. from the European Union. Research in the lab of B.L. is funded by the European Research Council (ERC), MINECO Plan Nacional grant BFU2011-26206, ERASysBio+ ERANET project EUI2009-04059 GRAPPLE, the EMBO Young Investigator Program, EU Framework 7 project 277899 4DCellFate, and the EMBL/CRG Systems Biology Program. K.T. is funded by a La Caixa Ph.D. fellowship as well as ERC.

Figure 5

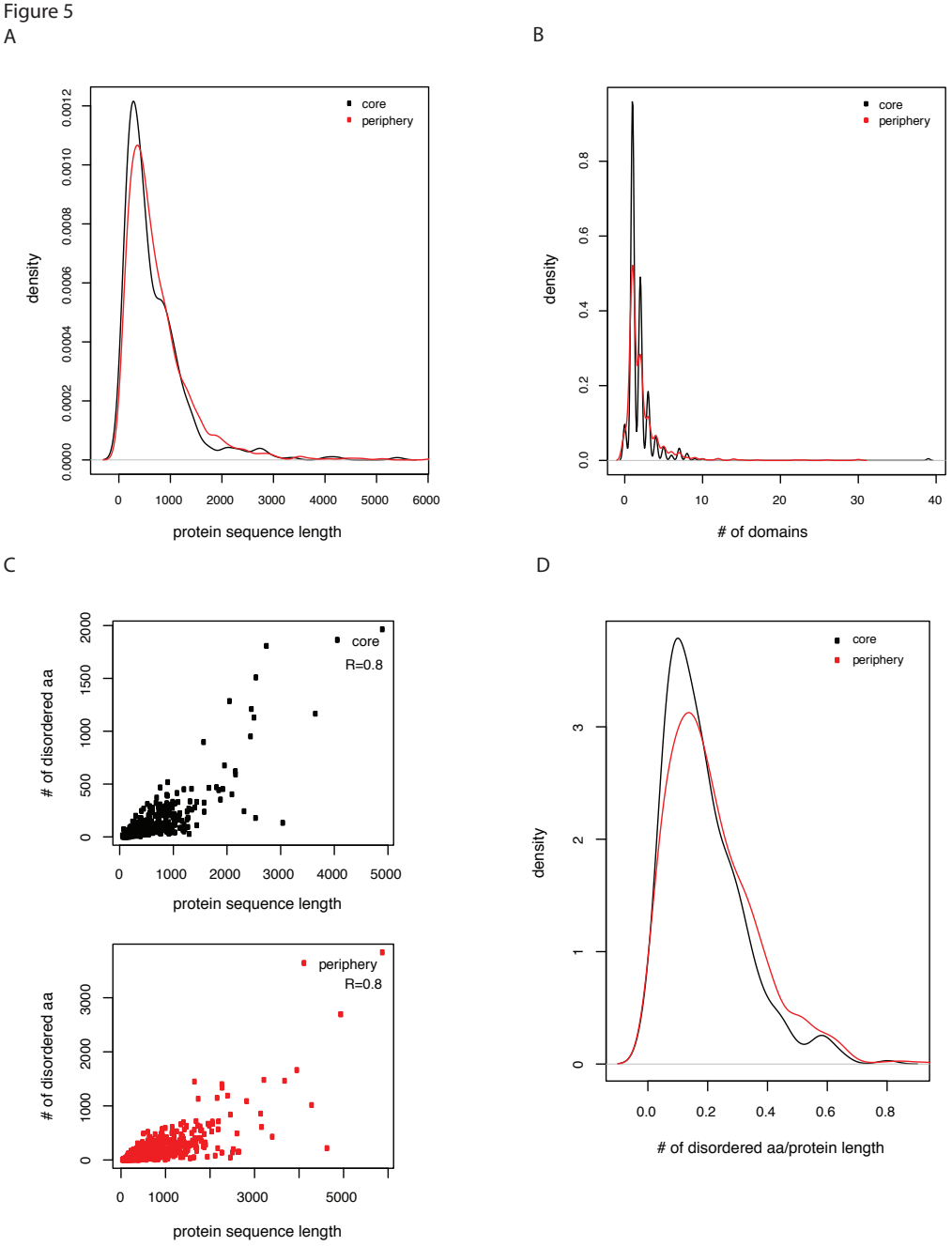

Figure 4.5: **Periphery subunits are larger, possess more interaction domains and have more disordered content** : (A) Sequence length distribution of 412 core proteins with mean of 676 is lower than sequence length of 917 periphery proteins with mean of 755 amino acids ($P = 0.004$). (B) The average number of Pfam domains for 412 core proteins, 1.9 is lower than the average of 2.3 domains for 917 periphery proteins ($P = 0.046$). (C) In both core and periphery classes, the longer protein sequences exhibit higher number of disordered regions (D) Density plots of disordered amino acid counts normalized by total sequence length in core ($n = 405$, black line) and periphery ($n = 888$, red line) populations. The mean proportion of disordered content to protein sequence length for core proteins is 0.19 and for periphery is 0.22. These two populations have a real shift in distribution ($P = 0.002$). Statistical significance tested using Wilcoxon rank sum test in A, B and D.

4.7 Supplementary Information

4.7.1 Supplementary Figures

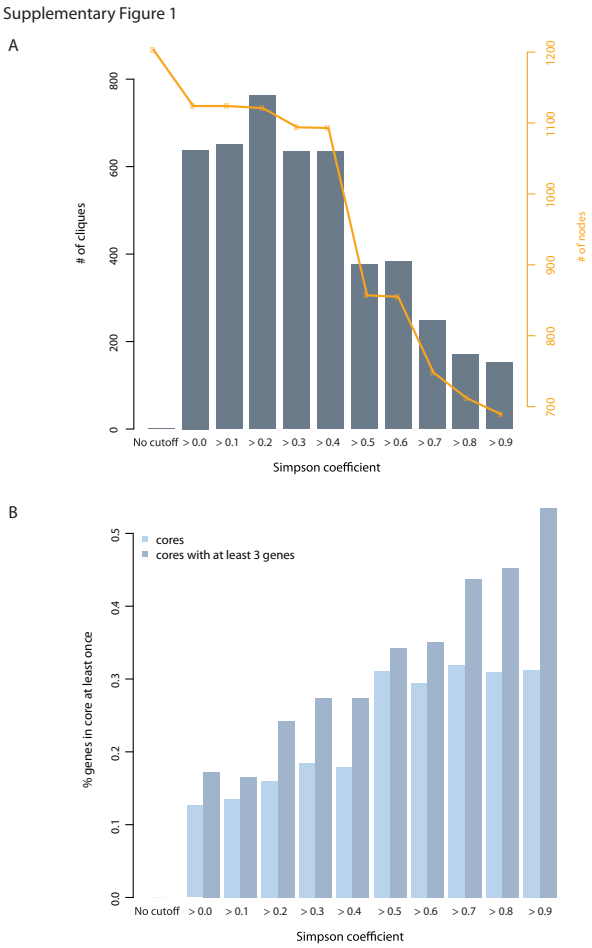

Figure S1: **Criteria used for selection of Simpson's coefficient cutoff** : (A) Unbiased cutoffof Simpson coefficient (SC) was selected by examining the effect of a representative set of cutoffs in range of 0 and 1 (lower and upper bounds of SC) on clique finding. 'No cutoff' served as a negative control — in a fully connected network no clique should be found. The orange y-axis display the total number of nodes (protein complexes) as a function ofi ncreasing SC cutoff— the more edges get pruned since their weight does not meet the SC cutoffthe less nodes remain in the large network. Isolated nodes are not taken into consideration as cliques are found only as fully connected subgraphs of the larger network. (B) This plot is a proxy for compactness. The stricter the Simpson coefficient (SC), the more genes end up in the core. For example, when SC > 0.9, although fewer cliques are captured, their overlap is so high that nearly 50% of the gene in the entire data set are in the shared cores. At our selected cutoffof SC > 0.5 we have 387 core genes and 942 periphery genes (41% of 1330 genes).

Supplementary Figure 2

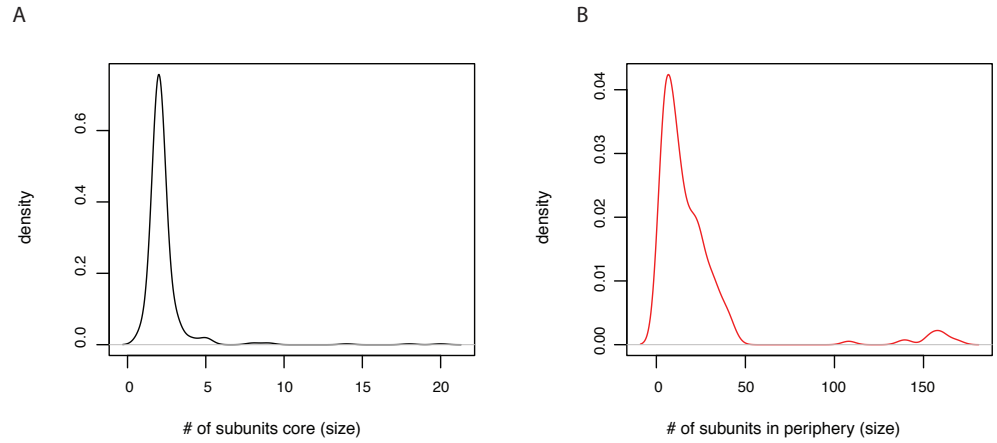

Figure S2: **Clique peripheries are larger than clique cores** : Probability density plots depict the component size distribution (total number of subunits) of (A) clique cores derived from set 1 and (B) clique peripheries derived from set 1.

Supplementary Figure 3

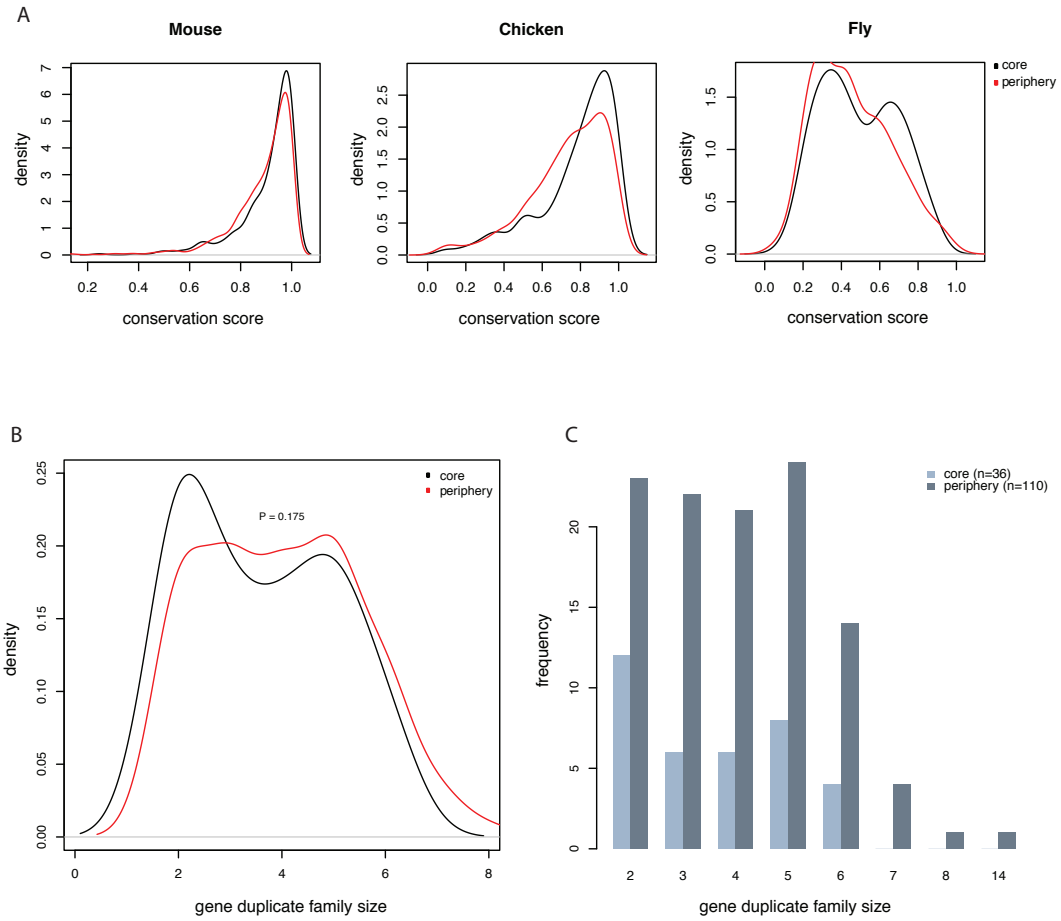

Figure S3: **Gene duplicates in both groups of core and periphery subunits come from families with same size distributions** : (A) Probability density plots depict the conservation of core and periphery genes as measured by comparative analysis with their respective one-to-one orthologs in mouse, chicken, and fly. Compared to mouse, core subunits (n = 372) mean conservation score is 0.91 versus 0.89 for periphery subunits (n = 825), P = 0.0043. In chicken, mean conservation score of core subunits (n = 324) was 0.78 compared to a mean of 0.73 of periphery subunits (n = 710), P = 1e-04. In fly, mean conservation score of core subunits (n = 194) was 0.49 compared to a mean of 0.46 of periphery subunits (n = 339), P = 0.038. Statistical significance tested using Wilcoxon rank sum test. (B) Density plots of gene duplicate family sizes for core subunits (black) and periphery subunits (red) reveal that the average family size for 36 paralagous gene families in the core, 3.6 (median = 3.5) is not significantly different from the average family size of 4.1 (median = 4) for 110 paralagous gene families in the periphery (P = 0.175). Statistical significance tested using Wilcoxon rank sum test. (C) Side-by-side comparison of core and periphery duplicate gene family sizes displayed as a percentage of total family counts.

Supplementary Figure 4

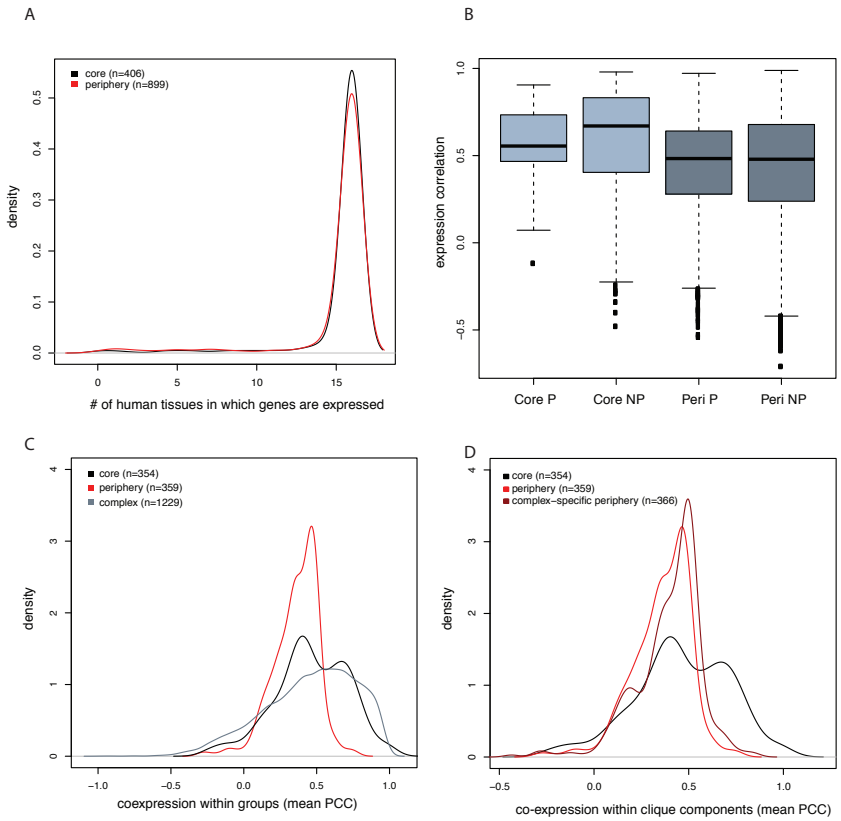

Figure S4: **Core and periphery subunits are co-expressed across human tissues but to differing extents** : (A) Breadth of gene expression across 16 human tissues show that 406 core genes are expressed on average in 15.4 tissues compared to 899 periphery genes which are expressed on average in 15.1 tissues (P = 0.06). (B) Four groups of protein subunits: CP or core paralogs (n = 15), CNP or core non-paralogs (n = 736), PP or periphery paralogs (n = 2, 522), and PNP or periphery non-paralogs (n = 22, 494) are compared in terms of the distribution of the Pearson correlation coefficients of their mRNA expression. CNP ranks first with median 0.67 (mean=0.59), CP has second highest median of 0.55 (mean=0.54), then PP and PNP tied for lowest values both with median 0.48 (mean=0.44). Six way comparison of groups reveals pairs of distributions are not significantly shifted ($P_{CP,CNP}$ = 0.3, $P_{CP,PP}$ = 0.09, $P_{CP,PNP}$ = 0.17, $P_{CNP,PP}$ = 2.1e-44, $P_{CNP,PNP}$ = 2.5e-43, $P_{PP,PNP}$ = 0.2). (C) Density plots of mean expression correlation of cores (n = 354) and peripheries (n = 359) as above against a backdrop of complexes (n = 1, 229) with a mean = 0.34. There is a real shift in distribution of mean PCC between core components and whole complexes (P = 1.3e-13) but not between periphery components and whole complexes (P = 0.42). (D) Density plots of mean expression correlation of subunits within each component of 394 cliques derived from complexes composed of one or more subunits; mRNA abundance cutoff > 1 RPMK. Clique components are cores (n = 354), peripheries (n = 359), and peripheries split into source protein complexes (n = 366). On average cores are more highly correlated in expression (mean=0.46) than peripheries (mean=0.36), (P = 2.1e-10) and peripheries split into individual protein complexes (mean=0.4), (P = 8.5e-05). Statistical significance tested using Wilcoxon rank sum test for all.

Supplementary Figure 5

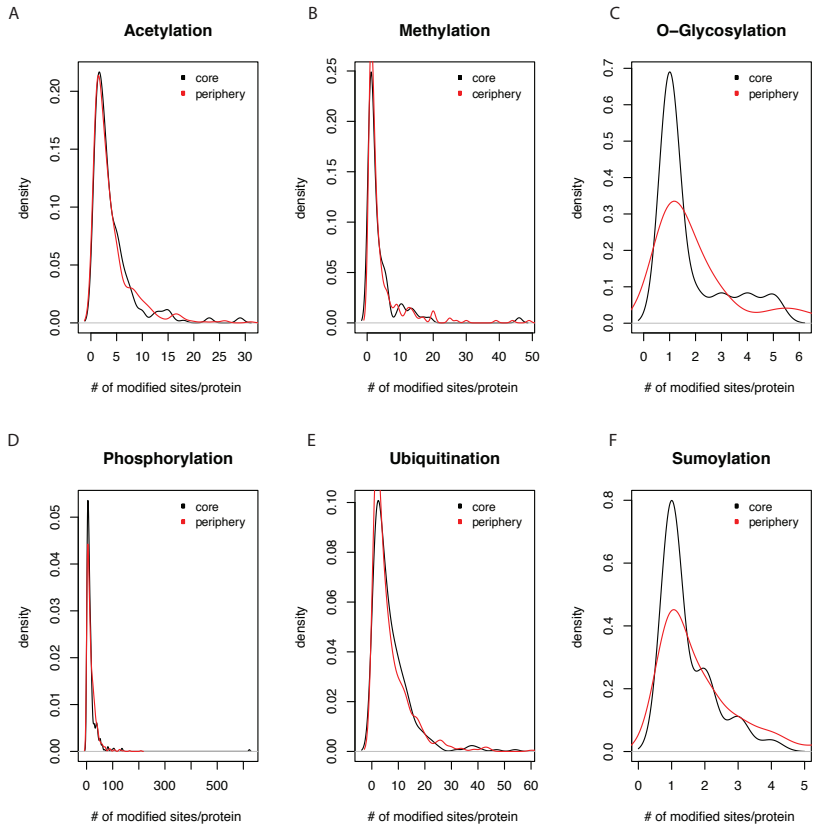

Figure S5: **Posttranslastional modifications (data from PhosphoSitePlus) affect core and periphery genes to equal degrees** : Density plots of total modifications per protein (A) Average number of acetylated residues of core proteins (n = 256), 4.1 (median=3) compared to 4.3 (median=3) of periphery proteins (n = 575), is not significantly different (P = 0.75) (B) Average number of methylated residues of core proteins (n = 88), 4 (median=2) compared to 4.5 (median=2) of periphery proteins (n = 250), is not significantly different (P = 0.44) (C) Average number of O-glycosylated residues of core proteins (n = 13), 1.8 (median=1) compared to 3.6 (median=1) of periphery proteins (n = 21), is not significantly different (P = 0.36) (D) Average number of phosphorylated residues of core proteins (n = 406), 16.8 (median=9), compared to 17.3 (median=11) of periphery proteins (n = 885), is not significantly different (P = 0.06) (E) Average number of ubiquitinated residues of core proteins (n = 350), 7 (median=5) compared to 7.4 (median=4) of periphery proteins (n = 706), is not significantly different (P = 0.27) (F) Average number of sumoylated residues of core proteins (n = 33), 1.5 (median=1) compared to 2.1 (median=1) of periphery proteins (n = 78), is not significantly different (P = 0.09).

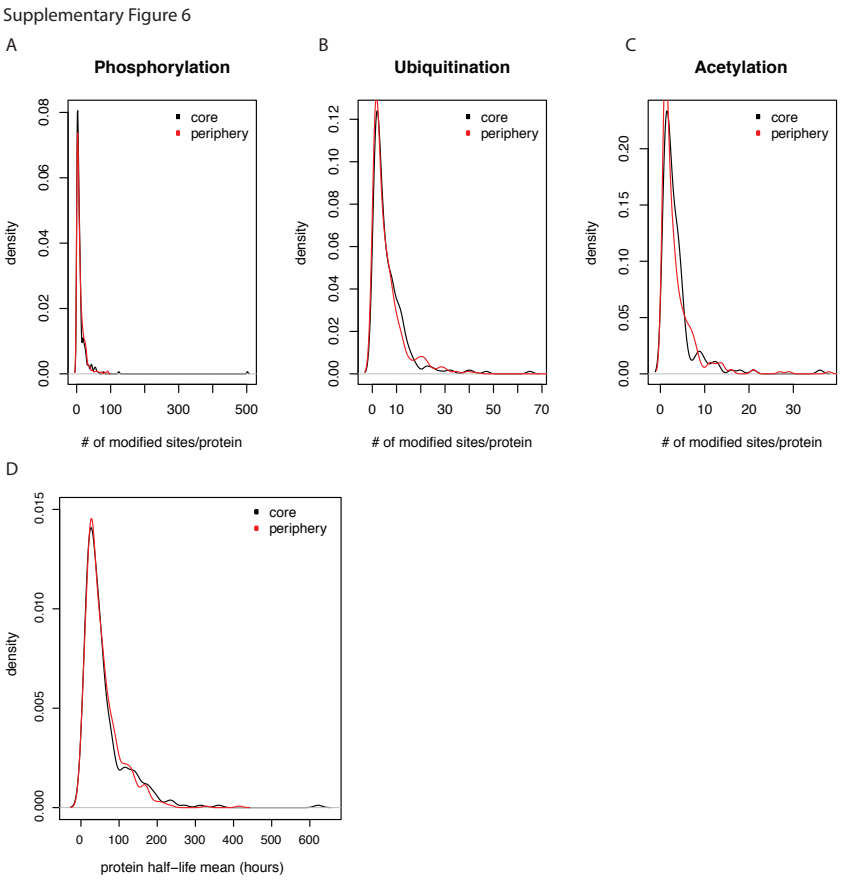

Figure S6: **Posttranslational modifications (data from PTMfunc) affect core and periphery genes to equal degrees** : (A) Average number of phosphorylated residues of core proteins (n = 298), 11.8 (median=6) compared to 10.6 (median=6) of periphery proteins (n = 624), is not significantly different (P = 0.52) (B) Average number of ubiquitinated residues of core proteins (n = 260), 6.5 (median=4) compared to 6.4 (median=4) of periphery proteins (n = 502), is not significantly different (P = 0.13) (C) Average number of acetylated residues of core proteins (n = 166), 3.6 (median=2), compared to 3.9 (median=2) of periphery proteins (n = 338), P = 0.32. (D) Average half-life in hours of core proteins (n = 309) is 62 hours (median=41) compared to 56 hours (median= 42) for periphery proteins (n = 612), is not significantly different (P = 0.83). Statistical significance tested using Wilcoxon rank sum test for all.

CHAPTER

FIVE

DISCUSSION

In section 2.1 of the Results, we analyze time course transcription data sampled from PHK cultures both in their stem cell state and upon calcium-induced differentiation. Using these data we show that epidermal cells in culture have free-running circadian rhythms. We further identify cohorts of genes that are in tune with the circadian expression of core clock genes and, by focusing on individual candidate pathways as identified from these genes sets, namely calcium and TGFβ, we show experimentally that differentiation happens more efficient at specific times during the day.

The induction of differentiation by calcium in PHK cultures is not a perfect model for studying epidermal differentiation. One major reason is that the synchronization method used — the short serum shock given prior to the time-course RNA collection — can also synchronize the cell cycle, which is often approximately 24 hours in cultured cells. This can confound globally measured gene expression changes, in particular those associated with the cell cycle with those linked to the circadian clock. A better model for circadian control of human epidermal homeostasis would be the study of primary intact epidermis collected at different times of the day by employing a punch biopsy method. This would eliminate the need for serum shock, as the cells would be in sync with the central pacemaker upon collection. In the future, this type of study can be used to complement experiments on PHK cultures.

Furthermore, the design of the microarray experiment, which involved time course of RNA collection at 5-hour intervals for 45 hours, was not ideal for the study of circadian rhythms. The experiment was originally designed to survey global transcriptional changes during skin differentiation, with the circadian question not being a primary focus. The experiment therefore did not cover at least two complete circadian cycles and the time points were not chosen to be factors of 24. This made the task of identifying circadian transcripts quite challenging. A recent study has compared a number of mainstream methods used to detect periodicity in data and shown quantitatively that higher sampling resolution provides more statistical power for any one of these methods (Hughes *et al.*, 2010). In the future, employing a strategy whereby the gene expression time course is more densely sampled would be desirable, especially when the accurate detection of circadian rhythms is the main objective of the study.

Another limitation of our data, which made the task of identifying circadian or even dynamically regulated transcripts quite challenging was the fact that the magnitude of change for many genes tended to be very restricted. In fact, numerous markers of differentiation and/or signalling genes, well characterized for their role in skin homeostasis and differentiation had mRNA fold changes hovering around or below 2, an oft-used cut-off to identify genes with significant changes in expression. In the future, to obtain a more accurate count of transcripts globally, performing RNA sequencing in place of microarrays would be desirable.

We also performed extensive benchmarking of a number of highly cited methods for the identification of circadian patterns against our own data, simulated/synthetic data, as well as legacy data sets (Storch *et al.*, 2002). Specifically we tested COSOP (Straume, 2004), Fourier transform followed by Fisher's G test (Wichert *et al.*, 2003), and JTK_Cycle (Hughes *et al.*, 2010). Furthermore, we devised two additional ad hoc methods based on autocorrelation and cosine curve fitting. All five methods performed equally well in detecting the core clock genes BMAL1, PER1-3, NR1D1-2, and CRY1/2 as circadian in our data as well as previously published expression data (Storch *et al.*, 2002). We simulated synthetic data composed of 512 positive controls with circadian profiles and 512 negative controls with non-circadian profiles

complete with varying period lengths for the positive hits, varying amounts of noise, and a few outliers to test all five methods. We found that at cutoff of p-value < 0.05, the true positive rate ranged from 29% to 56% while specificity ranged from 71% to 96%. In particular, for example JTK_Cycle performed the worst with TPR of 29% and specificity of 71% while auto-correlation had the highest sensitivity of 56% and Fisher's G test had the highest specificity of 96%. When we tested the same set of methods on published data in which circadian genes had been identified (Storch *et al.*, 2002), all methods fared slightly worse, although by an almost equal margin. At the same p-value cutoff of 0.05 the range of true positive rate was between 27% to 45% with Fisher's G test out-performing the others.

The challenge arose when these methods were tested on our data. As mentioned, we were limited by unconventional time sampling and the fact that our data did not extend over two circadian cycles. While benchmarking highly cited methods, we found that even genes known to be under the control of clock genes in mammalian epidermis as reported previously (Janich *et al.*, 2011) could not be detected in a statistically sound manner when compared to a background of random genes and as a result we decided to devise our own method instead of applying one that has been described in the literature. To this end we devised a custom method in order to compare the expression pattern of individual genes to those of known clock components by subdividing the expression time course into six overlapping 15h time windows (i.e. 5-20h, 10-25h, 15-30h, 20-35h, 25-40h, 30-45h) and for each gene fitting a polynomial corresponding to a time window with four time points. This simple method enabled us to describe genes in terms of peaks and troughs and match them against clock genes with the same sequence of peaks and troughs and to overcome the limitations imposed by limited time resolution.

In section 2.2 we looked at how protein complexes change their composition over the course of skin differentiation based on expression pattern of their member sub-units by combining protein complex information with our time course expression data. One of the first steps in our study involved the characterization of global changes in transcriptional regulation. For this, it was necessary to distinguish dynamic from non-dynamic genes, followed by unsupervised clustering to segregate

genes in an unbiased manner based on their temporal regulation. Many genes in our data showed a limited dynamic range of expression. For example, less than 8% of genes had differential expression across the time course with a magnitude of change above 4 folds. To give this a frame of reference, the number of genes that are differentially expressed with a fold change > 4 throughout our skin differentiation data set compared to a randomly sampled data set with equal dimensions from an expression atlas of 79 different human tissues (Su *et al.*, 2004) is far lower (8% compared to 21%). A comprehensive comparison of six microarray platforms has shown cross-platform correlation to be very high between Agilent microarray (the platform used in our study) and other oligonucleotide microarray technologies (Yauk, 2004) in variability and sensitivity and the detection of differential gene expression. In fact validation of a handful of genes which, were of interest to our study (e.g. circadian genes, markers of differentiation) by RT-qPCR corroborated the range of change we detect with our microarray platform. Therefore, this limitation in dynamic range may be due to the experimental conditions or a reflection of physiological state of the cells. Yet this characteristic of the data made the task of clustering to identify various trends a bit more challenging.

Further, another limitation of k-means clustering is selecting an optimal number of clusters or k. We employed a wide range of metrics such as silhouette index, Dunn index, and F- test, which measures the ratio of the between-group variance to the total variance to choose the optimal number of clusters. Most of these measures resulted in $k < 10$. One limitation of employing a small k value, is that subtle trends such as circadian expression of known genes are not picked up or rather they are grouped together with more general trends within our 8 clusters. Hence our k-mean clustering strategy where $k=8$ was more successful in identifying various general trends in expression over the course of differentiation.

In this section, another question of interest to our research was whether paralogous genes compete with one another to bind to the same interfaces. However one major limitation for this type of study is that we only have mRNA expression data and hence we can only establish correlations between structural information and gene homology. In the future, it would be desirable to quantify protein abundances

in order to distinguish competing from non-competing interactions. Despite these limitations, we were able to establish a number of statistically sound correlations. Through using three-dimensional protein structures between interacting partners for a common hub, we identified that mutually exclusive interactions (MEI) are enriched in dynamic genes and by contrast compatible interactions (COI) are enriched for non-dynamic genes.

Section 2.3 was a purely integrative study. The main limitation that comes to mind is the quality and the availability of data sources that we had to rely on. For example, subcellular localization data was hard to come by for mammals and the two data sets that we used covered a very narrow range of organelles (i.e. nucleus, cytoplasm, nucleolus). Nonetheless, the collective data and our analyses in this study were indicative of a picture in which the function of a stable protein complex core is modified by the attachment or detachment of periphery proteins that allow protein complexes to function in a plastic, context-dependent manner.

CHAPTER

SIX

SUMMARY OF SCIENTIFIC FINDINGS

1. Core clock transcripts oscillate in an autonomous circadian manner in epidermal stem cells of cultured primary human keratinocytes.
2. The oscillation of core clock genes was maintained when PHK were induced by calcium to differentiate, although the amplitude of some clock transcripts in the negative limb of the clock showed an increase.
3. The successive oscillations (peaks and troughs) of the core clock genes subdivide the day into at least five temporal intervals, in effect segregating epidermal stem cell functions since each temporal interval defines a different functional category in both stem cells and differentiating cells.
4. Genes in tune with peaks pertaining to NR1D1/2 and PER1-3, genes known to peak during late-night to early morning hours, are enriched for pathways related to differentiation, whereas genes synchronized to peaks of clock genes CRY1 and BMAL1, known to culminate in the afternoon and evening hours, correspond to pathways inducing DNA replication, UV protection, and cell division.
5. Epidermal stem cells respond more efficiently to differentiation cues, specifically TGFβand calcium, in a time-of-day-dependent manner.
6. Disruption of circadian clock function through the overexpression of PER1 and PER2, or upon knockdown of CRY1 and CRY2 leads to spontaneous differentiation of PHKs.

7. Analysis of gene expression during skin differentiation reveals dynamically changing proteins in complex with non-dynamically expressed proteins in almost two thirds of complexes.
8. Structural analyses identifies that mutually exclusive surface interactions are enriched in dynamic genes.

CHAPTER

SEVEN

BIBLIOGRAPHY

Akashi, M., Soma, H., Yamamoto, T., Tsugitomi, A., Yamashita, S., Yamamoto, T., Nishida, E., Yasuda, A., Liao, J.K. & Node, K. (2010). Noninvasive method for assessing the human circadian clock using hair follicle cells. *Proceedings of the National Academy of Sciences of the United States of America*.

Akhtar, R.A., Reddy, A.B., Maywood, E.S., Clayton, J.D., King, V.M., Smith, A.G., Gant, T.W., Hastings, M.H. & Kyriacou, C.P. (2002). Circadian cycling of the mouse liver transcriptome, as revealed by cDNA microarray, is driven by the suprachiasmatic nucleus. *Current Biology*.

Akiyama, M., Minami, Y., Nakajima, T., Moriya, T. & Shibata, S. (2001). Calcium and pituitary adenylate cyclase-activating polypeptide induced expression of circadian clock gene mPer1 in the mouse cerebellar granule cell culture. *Journal of neurochemistry*.

Albert, R. & Barabási, A.L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*.

Alberts, B. (1998). The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. & Walter, P. (2002). *Molecular Biology of the Cell*. Garland Science.

Aloy, P. & Russell, R.B. (2002). Interrogating protein interaction networks through structural biology. *Proceedings of the National Academy of Sciences of the United States of America*.

Andl, T., Reddy, S.T., Gaddapara, T. & Millar, S.E. (2002). WNT signals are required for the initiation of hair follicle development. *Developmental Cell*.

Andreoli, J.M., Jang, S.I., Chung, E., Coticchia, C.M., Steinert, P.M. & Markova, N.G. (1997). The expression of a novel, epithelium-specific ets transcription factor is restricted to the most differentiated layers in the epidermis. *Nucleic Acids Research*.

Annes, J.P. (2003). Making sense of latent TGFbeta activation. *Journal of cell science*.

Arabidopsis Interactome Mapping Consortium, Dreze, M., Carvunis, A.R., Charloteaux, B., Galli, M., Pevzner, S.J., Tasan, M., Ahn, Y.Y., Balumuri, P., Barabasi, A.L., Bautista, V., Braun, P., Byrdsong, D., Chen, H., Chesnut, J.D., Cusick, M.E., Dangl, J.L., de los Reyes, C., Dricot, A., Duarte, M., Ecker, J.R., Fan, C., Gai, L., Gebreab, F., Ghoshal, G., Gilles, P., Gutierrez, B.J., Hao, T., Hill, D.E., Kim, C.J., Kim, R.C., Lurin, C., MacWilliams, A., Matrubutham, U., Milenkovic, T., Mirchandani, J., Monachello, D., Moore, J., Mukhtar, M.S., Olivares, E., Patnaik, S., Poulin, M.M., Przulj, N., Quan, R., Rabello, S., Ramaswamy, G., Reichert, P., Rietman, E.A., Rolland, T., Romero, V., Roth, F.P., Santhanam, B., Schmitz, R.J., Shinn, P., Spooner, W., Stein, J., Swamilingiah, G.M., Tam, S., Vandenhaute, J., Vidal, M., Waaijers, S., Ware, D., Weiner, E.M., Wu, S. & Yazaki, J. (2011). Evidence for Network Evolution in an Arabidopsis Interactome Map. *Science*.

Asher, G., Gatfield, D., Stratmann, M., Reinke, H., Dibner, C., Kreppel, F., Mostoslavsky, R., Alt, F.W. & Schibler, U. (2008). SIRT1 Regulates Circadian Clock Gene Expression through PER2 Deacetylation. *Cell*.

Asher, G., Reinke, H., Altmeyer, M., Gutierrez-Arcelus, M., Hottiger, M.O. & Schibler, U. (2010). Poly(ADP-Ribose) Polymerase 1 Participates in the Phase Entrainment of Circadian Clocks to Feeding. *Cell*.

Avery, O.T., MacLeod, C.M. & McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Induction of

transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III. *Molecular medicine (Cambridge, Mass.)*.

Bader, G.D. (2003). BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Research*.

Balsalobre, A. (2000). Resetting of Circadian Time in Peripheral Tissues by Glucocorticoid Signaling. *Science*.

Balsalobre, A., Damiola, F. & Schibler, U. (1998). A serum shock induces circadian gene expression in mammalian tissue culture cells. *Cell*.

Barabasi, A. & Albert, R. (1999). Emergence of scaling in random networks. *Science*.

Barabási, A.L. & Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. *Nature reviews Genetics*.

Bass, J. (2012). Circadian topology of metabolism. *Nature*.

Behrends, C., Sowa, M.E., Gygi, S.P. & Harper, J.W. (2010). Network organization of the human autophagy system. *Nature*.

Beltrao, P. & Serrano, L. (2005). Comparative genomics and disorder prediction identify biologically relevant SH3 protein interactions. *PLoS computational biology*.

Beltrao, P., Albanèse, V., Kenner, L.R., Swaney, D.L., Burlingame, A., Villén, J., Lim, W.A., Fraser, J.S., Frydman, J. & Krogan, N.J. (2012). Systematic Functional Prioritization of Protein Posttranslational Modifications. *Cell*.

Bernfield, R. & Nirenberg, M. (1965). RNA codewords and protein synthesis. The nucleotide sequence of multiple codewords for Phenylalanine, Serine, Leucine, and Proline. *Science*.

Bierie, B. & Moses, H.L. (2006). Tumour microenvironment: TGFβ: the molecular Jekyll and Hyde of cancer. *Nature Reviews Cancer*.

Bikle, D.D., Ng, D., Tu, C.L., Oda, Y. & Xie, Z. (2001). Calcium- and vitamin D-regulated keratinocyte differentiation. *Molecular and cellular endocrinology*.

Blanco, S., Kurowski, A., Nichols, J., Watt, F.M., Benitah, S.A. & Frye, M. (2011). The RNA–Methyltransferase Misu (NSun2) Poises Epidermal Stem Cells to Differentiate. *PLoS genetics*.

Blanpain, C. & Fuchs, E. (2006). Epidermal stem cells of the skin. *Annual Review of Cell and Developmental Biology*.

Blanpain, C. & Fuchs, E. (2009). Epidermal homeostasis: a balancing act of stem cells in the skin. *Nature Reviews Molecular Cell Biology*.

Blanpain, C., Lowry, W.E., Pasolli, H.A. & Fuchs, E. (2006). Canonical notch signaling functions as a commitment switch in the epidermal lineage. *Genes & Development*.

Blatt, M., Wiseman, S. & Domany, E. (1996). Superparamagnetic clustering of data. *Physical review letters*.

Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* (*Oxford, England*).

Bossi, A. & Lehner, B. (2009). Tissue specificity and the human protein interaction network. *Molecular Systems Biology*.

Boukamp, P., Petrussevska, R.T., Breitkreutz, D., Hornung, J., Markham, A. & Fusenig, N.E. (1988). Normal keratinization in a spontaneously immortalized aneuploid human keratinocyte cell line. *The Journal of cell biology*.

Bradley, R.K., Merkin, J., Lambert, N.J. & Burge, C.B. (2012). Alternative Splicing of RNA Triplets Is Often Regulated and Accelerates Proteome Evolution. *PLoS biology*.

Braun, K.M. (2003). Manipulation of stem cell proliferation and lineage commitment: visualisation of label-retaining cells in wholemounts of mouse epidermis. *Development*.

Bray, D. (1995). Protein molecules as computational elements in living cells. *Nature*.

Brissette, J.L., Kumar, N.M., Gilula, N.B., Hall, J.E. & Dotto, G.P. (1994). Switch in gap junction protein expression is associated with selective changes in junctional permeability during keratinocyte differentiation. *Proceedings of the National Academy of Sciences of the United States of America*.

Brown, S.A., Kowalska, E. & Dallmann, R. (2012). (Re)inventing the Circadian Feedback Loop. *Developmental Cell*.

Brown, W.R. (1991). A review and mathematical analysis of circadian rhythms in cell proliferation in mouse, rat, and human epidermis. *The Journal of investigative dermatology*.

Bunger, M.K., Wilsbacher, L.D., Moran, S.M., Clendenin, C., Radcliffe, L.A., Hogenesch, J.B., Simon, M.C., Takahashi, J.S. & Bradfield, C.A. (2000). Mop3 is an essential component of the master circadian pacemaker in mammals. *Cell*.

Buschke, S., Stark, H.J., Cerezo, A., Pratzel-Wunder, S., Boehnke, K., Kollar, J., Langbein, L., Heldin, C.H. & Boukamp, P. (2011). A decisive function of transforming growth factor- /Smad signaling in tissue morphogenesis and differentiation of human HaCaT keratinocytes. *Molecular biology of the cell*.

Butland, G., Peregrín-Alvarez, J.M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J. & Emili, A. (2005). Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature*.

Calvano, S.E., Xiao, W., Richards, D.R., Felciano, R.M., Baker, H.V., Cho, R.J., Chen, R.O., Brownstein, B.H., Cobb, J.P., Tschoeke, S.K., Miller-Graziano, C., Moldawer, L.L., Mindrinos, M.N., Davis, R.W., Tompkins, R.G., Lowry, S.F., Large Scale Collab Res Program, I. & to Injury, H.R. (2005). A network-based analysis of systemic inflammation in humans. *Nature Cell Biology*.

Candi, E., Schmidt, R. & Melino, G. (2005). The cornified envelope: a model of cell death in the skin. *Nature Reviews Molecular Cell Biology*.

Capella-Gutierrez, S., Silla-Martinez, J.M. & Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* (*Oxford, England*).

Cardone, L. (2005). Circadian Clock Control by SUMOylation of BMAL1. *Science*.

Ceriani, M.F., Hogenesch, J.B., Yanovsky, M., Panda, S., Straume, M. & Kay, S.A. (2002). Genome-Wide Expression Analysis in DrosophilaReveals Genes Controlling Circadian Behavior. *The Journal of neuroscience*.

Chan, E.F., Gat, U., McNiff, J.M. & Fuchs, E. (1999). A common human skin tumour is caused by activating mutations in beta-catenin. *Nature Genetics*.

Chatr-aryamontri, A., Ceol, A., Palazzi, L.M., Nardelli, G., Schneider, M.V., Castagnoli, L. & Cesareni, G. (2007). MINT: the Molecular INTeraction database. *Nucleic Acids Research*.

Chen, J. & Yuan, B. (2006). Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics* (*Oxford, England*).

Cheng, X., Jin, J., Hu, L., Shen, D., Dong, X.p., Samie, M.A., Knoff, J., Eisinger, B., Liu, M.l., Huang, S.M., Caterina, M.J., Dempsey, P., Michael, L.E., Dlugosz, A.A., Andrews, N.C., Clapham, D.E. & Xu, H. (2010). TRP Channel Regulates EGFR Signaling in Hair Morphogenesis and Skin Barrier Formation. *Cell*.

Choi, Y.S., Zhang, Y., Xu, M., Yang, Y., Ito, M., Peng, T., Cui, Z., Nagy, A., Hadjantonakis, A.K., Lang, R.A., Cotsarelis, G., Andl, T., Morrisey, E.E. & Millar, S.E. (2013). Distinct Functions for Wnt/b-Catenin in Hair Follicle Stem Cell Proliferation and Survival and Interfollicular Epidermal Homeostasis. *Stem Cell*.

Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D. & Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Molecular Systems Biology*.

Clark, D.A. & Coker, R. (1998). Transforming growth factor-beta (TGF-beta). *International Journal of Biochemistry and Cell Biology*.

Clevers, H. (2006). Wnt/β-Catenin Signaling in Development and Disease. *Cell*.

Connelly, J.T., Mishra, A., Gautrot, J.E. & Watt, F.M. (2011). Shape-Induced Terminal Differentiation of Human Epidermal Stem Cells Requires p38 and Is Regulated by Histone Acetylation. *PLoS ONE*.

Cook, S.A. (1971). The complexity of theorem-proving procedures.

Coolen, N.A., Verkerk, M., Reijnen, L., Vlig, M., van den Bogaerdt, A.J., Breetveld, M., Gibbs, S., Middelkoop, E. & Ulrich, M.M.W. (2007). Culture of keratinocytes for transplantation without the need of feeder layer cells. *Cell transplantation*.

Costanzo, M.C., Hogan, J.D., Cusick, M.E., Davis, B.P., Fancher, A.M., Hodges, P.E., Kondu, P., Lengieza, C., Lew-Smith, J.E., Lingner, C., Roberg-Perez, K.J., Tillberg, M., Brooks, J.E. & Garrels, J.I. (2000). The yeast proteome database (YPD) and Caenorhabditis elegans proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Research*.

Cotsarelis, G., Sun, T.T. & Lavker, R.M. (1990). Label-retaining cells reside in the bulge area of pilosebaceous unit: implications for follicular stem cells, hair cycle, and skin carcinogenesis. *Cell*.

Cui, W., Fowlis, D.J., Cousins, F.M., Duffie, E., Bryson, S., Balmain, A. & Akhurst, R.J. (1995). Concerted action of TGF-beta 1 and its type II receptor in control of epidermal homeostasis in transgenic mice. *Genes & Development*.

Cusick, M.E. (2005). Interactome: gateway into systems biology. *Human Molecular Genetics*.

Cusick, M.E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A.R., Simonis, N., Rual, J.F., Borick, H., Braun, P., Dreze, M., Vandenhaute, J., Galli, M., Yazaki, J., Hill, D.E., Ecker, J.R., Roth, F.P. & Vidal, M. (2009). Literature-curated protein interaction datasets. *Nature methods*.

de Lichtenberg, U., Jensen, L.J., Brunak, S. & Bork, P. (2005). Dynamic complex formation during the yeast cell cycle. *Science*.

DeBruyne, J.P., Weaver, D.R. & Reppert, S.M. (2007). CLOCK and NPAS2 have overlapping roles in the suprachiasmatic circadian clock. *Nature neuroscience*.

Derényi, I., Palla, G. & Vicsek, T. (2005). Clique percolation in random networks. *Physical review letters*.

Derynck, R., Akhurst, R.J. & Balmain, A. (2001). TGF-beta signaling in tumor suppression and cancer progression. *Nature Genetics*.

Dezso, Z., Oltvai, Z.N. & Barabási, A.L. (2003). Bioinformatics analysis of experimentally determined protein complexes in the yeast Saccharomyces cerevisiae. *Genome Research*.

Dibner, C., Schibler, U. & Albrecht, U. (2010). The Mammalian Circadian Timing System: Organization and Coordination of Central and Peripheral Clocks. *Annual review of physiology*.

Diss, G., Dubé, A.K., Boutin, J., Gagnon-Arsenault, I. & Landry, C.R. (2013). A Systematic Approach for the Genetic Dissection of Protein Complexes in Living Cells. *CellReports*.

DiTacchio, L., Le, H.D., Vollmers, C., Hatori, M., Witcher, M., Secombe, J. & Panda, S. (2011). Histone Lysine Demethylase JARID1a Activates CLOCK-BMAL1 and Influences the Circadian Clock. *Science*.

Dodd, A.N., Salathia, N., Hall, A., Kévei, E., Tóth, R., Nagy, F., Hibberd, J.M., Millar, A.J. & Webb, A.A.R. (2005). Plant circadian clocks increase photosynthesis, growth, survival, and competitive advantage. *Science*.

Doherty, C. (2010). Circadian control of global gene expression patterns. *Annual review of genetics*.

Doi, M., Hirayama, J. & Sassone-Corsi, P. (2006). Circadian Regulator CLOCK Is a Histone Acetyltransferase. *Cell*.

Dorogovtsev, S.N. & Mendes, J.F. (2002). Evolution of networks. *Advances in physics*.

Dunlap, J.C. (1999). Molecular Bases for Circadian Clocks. *Cell*.

Duong, H.A., Robles, M.S., Knutti, D. & Weitz, C.J. (2011). A Molecular Mechanism for Circadian Clock Negative Feedback. *Science*.

Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*.

Edgar, R.S., Green, E.W., Zhao, Y., Van, O.G., Olmedo, M. & Qin, X. (2012). Peroxiredoxins are conserved markers of circadian rhythms. *Nature*.

Enright, A.J., Van Dongen, S. & Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*.

Ewing, R.M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., McBroom-Cerajewski, L., Robinson, M.D., O'Connor, L., Li, M., Taylor, R., Dharsee, M., Ho, Y., Heilbut, A., Moore, L., Zhang, S., Ornatsky, O., Bukhman, Y.V., Ethier, M., Sheng, Y., Vasilescu, J., Abu-Farha, M., Lambert, J.P., Duewel, H.S., Stewart, I.I., Kuehl, B., Hogue, K., Colwill, K., Gladwish, K., Muskat, B., Kinach, R., Adams, S.L., Moran, M.F., Morin, G.B., Topaloglou, T. & Figeys, D. (2007). Large-scale mapping of human protein–protein interactions by mass spectrometry. *Molecular Systems Biology*.

Fields, S. & Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature*.

Finnigan, G.C., Hanson-Smith, V., Stevens, T.H. & Thornton, J.W. (2012). Evolution of increased complexity in a molecular machine. *Nature*.

Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Garcia-Giron, C., Gordon, L., Hourlier, T., Hunt, S., Juettemann, T., Kahari, A.K., Keenan, S., Komorowska, M., Kulesha, E., Longden, I., Maurel, T., McLaren, W.M., Muffato, M., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H.S., Ritchie, G.R.S., Ruffier, M., Schuster, M., Sheppard, D., Sobral, D., Taylor, K., Thormann, A., Trevanion, S., White, S., Wilder, S.P., Aken, B.L., Birney, E., Cunningham, F., Dunham, I., Harrow, J., Herrero, J., Hubbard, T.J.P., Johnson, N., Kinsella, R., Parker, A., Spudich, G., Yates, A., Zadissa, A. & Searle, S.M.J. (2012). Ensembl 2013. *Nucleic Acids Research*.

Forslind, B., Lindberg, M., Roomans, G.M., Pallon, J. & Werner-Linde, Y. (1997). Aspects on the physiology of human skin: studies using particle probe analysis. *Microscopy research and technique*.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C. & Jensen, L.J. (2012). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*.

Fraser, H.B. (2005). Modularity and evolutionary constraint on proteins. *Nature Genetics*.

Frye, M., Fisher, A.G. & Watt, F.M. (2007). Epidermal Stem Cells Are Defined by Global Histone Modifications that Are Altered by Myc-Induced Differentiation. *PLoS ONE*.

Fuchs, E. (2007). Scratching the surface of skin development. *Nature*.

Fuchs, E. & Green, H. (1980). Changes in keratin gene expression during terminal differentiation of the keratinocyte. *Cell*.

Gaddameedhi, S., Selby, C.P., Kaufmann, W.K., Smart, R.C. & Sancar, A. (2011). Control of skin cancer by the circadian rhythm. *Proceedings of the National Academy of Sciences of the United States of America*.

Gallego, M. & Virshup, D.M. (2007). Post-translational modifications regulate the ticking of the circadian clock. *Nature Reviews Molecular Cell Biology*.

Gandarillas, A. & Watt, F.M. (1997). c-Myc promotes differentiation of human epidermal stem cells. *Genes & Development*.

Gat, U., DasGupta, R., Degenstein, L. & Fuchs, E. (1998). De Novo hair follicle morphogenesis and hair tumors in mice expressing a truncated beta-catenin in skin. *Cell*.

Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dümpelfeld, B., Edelmann, A., Heurtier, M.A., Hoffman, V., Hoefert,

C., Klein, K., Hudak, M., Michon, A.M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J.M., Kuster, B., Bork, P., Russell, R.B. & Superti-Furga, G. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature*.

Gehring, J. (2013). The proteinProfiles package.

Geyfman, M. & Anderson, B. (2010). Clock genes, hair growth and aging. *Aging*.

Geyfman, M., Kumar, V., Liu, Q., Ruiz, R., Gordon, W., Espitia, F., Cam, E., Millar, S.E., Smyth, P. & Ihler, A. (2012). Brain and muscle Arnt-like protein-1 (BMAL1) controls circadian cell proliferation and susceptibility to UVB-induced DNA damage in the epidermis. *Proceedings of the National Academy of Sciences of the United States of America*.

Giot, L. (2003). A Protein Interaction Map of Drosophila melanogaster. *Science*.

Girvan, M. & Newman, M.E.J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*.

Goll, J., Rajagopala, S.V., Shiau, S.C., Wu, H., Lamb, B.T. & Uetz, P. (2008). MPIDB: the microbial protein interaction database. *Bioinformatics* (*Oxford, England*).

Guldener, U. (2004). CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Research*.

Guruharsha, K.G., Rual, J.F., Zhai, B., Mintseris, J., Vaidya, P., Vaidya, N., Beekman, C., Wong, C., Rhee, D.Y., Cenaj, O., McKillip, E., Shah, S., Stapleton, M., Wan, K.H., Yu, C., Parsa, B., Carlson, J.W., Chen, X., Kapadia, B., VijayRaghavan, K., Gygi, S.P., Celniker, S.E., Obar, R.A. & Artavanis-Tsakonas, S. (2011). A Protein Complex Network of Drosophila melanogaster. *Cell*.

Hamosh, A. (2004). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*.

Hansson, J., Rafiee, M.R., Reiland, S., Polo, J.M., Gehring, J., Okawa, S., Huber, W., Hochedlinger, K. & Krijgsveld, J. (2012). Highly Coordinated Proteome Dynamics during Reprogrammingof Somatic Cells to Pluripotency. *CellReports*.

Havugimana, P.C., Wong, P. & Emili, A. (2007). Improved proteomic discovery by sample pre-fractionation using dual-column ion-exchange high performance liquid chromatography. *Journal of Chromatography B*.

Havugimana, P.C., Hart, G.T., Nepusz, T., Yang, H., Turinsky, A.L., Li, Z., Wang, P.I., Boutz, D.R., Fong, V., Phanse, S., Babu, M., Craig, S.A., Hu, P., Wan, C., Vlasblom, J., Dar, V.u.N., Bezginov, A., Clark, G.W., Wu, G.C., Wodak, S.J., Tillier, E.R.M., Paccanaro, A., Marcotte, E.M. & Emili, A. (2012). A census of human soluble protein complexes. *Cell*.

Hegyi, H., Schad, E. & Tompa, P. (2007). Structural disorder promotes assembly of protein complexes. *BMC Structural Biology*.

Hennings, H., Michael, D., Cheng, C., Steinert, P., Holbrook, K. & Yuspa, S.H. (1980). Calcium regulation of growth and differentiation of mouse epidermal cells in culture. *Cell*.

Hernandez-Toro, J., Prieto, C. & De Las Rivas, J. (2007). APID2NET: unified interactome graphic analyzer. *Bioinformatics* (*Oxford, England*).

Hirayama, J., Sahar, S., Grimaldi, B., Tamaru, T., Takamatsu, K., Nakahata, Y. & Sassone-Corsi, P. (2007). CLOCK-mediated acetylation of BMAL1 controls circadian function. *Nature*.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K. & Boutilier, K. (2002a). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A.R., Sassi, H., Nielsen, P.A., Rasmussen, K.J., Andersen, J.R., Johansen, L.E., Hansen, L.H., Jespersen, H., Podtelejnikov, A., Nielsen,

E., Crawford, J., Poulsen, V., Srensen, B.D., Matthiesen, J., Hendrickson, R.C., Gleeson, F., Pawson, T., Moran, M.F., Durocher, D., Mann, M., Hogue, C.W.V., Figeys, D. & Tyers, M. (2002b). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*.

Honma, M., Stubbs, M., Collins, I., Workman, P., Aherne, W. & Watt, F.M. (2006). Identification of Novel Keratinocyte Differentiation Modulating Compounds by High-Throughput Screening. *Journal of Biomolecular Screening*.

Hornbeck, P.V., Kornhauser, J.M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V. & Sullivan, M. (2011). PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Research*.

Huang, D.W., Sherman, B.T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M.W., Lane, H.C. & Lempicki, R.A. (2007). DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Research*.

Huang, D.W., Sherman, B.T. & Lempicki, R.A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*.

Huelsken, J., Vogel, R., Erdmann, B., Cotsarelis, G. & Birchmeier, W. (2001). beta-Catenin controls hair follicle morphogenesis and stem cell differentiation in the skin. *Cell*.

Hughes, M.E., Hogenesch, J.B. & Kornacker, K. (2010). JTK CYCLE: An Efficient Non-parametric Algorithm for Detecting Rhythmic Components in Genome-Scale Data Sets. *Journal of Biological Rhythms*.

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T.K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S., de Castro, E., Coggill, P., Corbett, M., Das, U., Daugherty, L., Duquenne, L., Finn, R.D., Fraser, M., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., McMenamin, C., Mi, H., Mutowo-Muellenet, P., Mulder, N., Natale, D., Orengo, C., Pesseat, S., Punta, M., Quinn, A.F., Rivoire,

C., Sangrador-Vegas, A., Selengut, J.D., Sigrist, C.J.A., Scheremetjew, M., Tate, J., Thimmajanarthanan, M., Thomas, P.D., Wu, C.H., Yeats, C. & Yong, S.Y. (2011). InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Research*.

Hutchins, J.R.A., Toyoda, Y., Hegemann, B., Poser, I., Heriche, J.K., Sykora, M.M., Augsburg, M., Hudecz, O., Buschhorn, B.A., Bulkescher, J., Conrad, C., Comartin, D., Schleiffer, A., Sarov, M., Pozniakovsky, A., Slabicki, M.M., Schloissnig, S., Steinmacher, I., Leuschner, M., Ssykor, A., Lawo, S., Pelletier, L., Stark, H., Nasmyth, K., Ellenberg, J., Durbin, R., Buchholz, F., Mechtler, K., Hyman, A.A. & Peters, J.M. (2010). Systematic Analysis of Human Protein Complexes Identifies Chromosome Segregation Proteins. *Science*.

Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A.F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* (*Oxford, England*).

Ikushima, H. & Miyazono, K. (2010). signalling: a complex webin cancer progression. *Nature Reviews*.

Ito, M., Liu, Y., Yang, Z., Nguyen, J., Liang, F., Morris, R.J. & Cotsarelis, G. (2005). Stem cells in the hair follicle bulge contribute to wound repair but not to homeostasis of the epidermis. *Nature Medicine*.

Ito, M., Yang, Z., Andl, T., Cui, C., Kim, N., Millar, S.E. & Cotsarelis, G. (2007). Wnt-dependent de novo hair follicle regeneration in adult mouse skin after wounding. *Nature*.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*.

Janich, P., Pascual, G., Merlos-Suárez, A., Batlle, E., Ripperger, J., Albrecht, U., Obrietan, K., Croce, L.D. & Benitah, S.A. (2011). The circadian molecular clock creates epidermal stem cell heterogeneity. *Nature*.

Janich, P., Toufighi, K., Solanas, G., Luis, N.M., Minkwitz, S., Serrano, L., Lehner, B. & Benitah, S.A. (2013). Human Epidermal Stem Cell Function Is Regulated by Circadian Oscillations. *Cell stem cell*.

Jensen, K.B., Driskell, R.R. & Watt, F.M. (2010). Assaying proliferation and differentiation capacity of stem cells using disaggregated adult mouse epidermis. *Nature Protocols*.

Jones, P. & Simons, B.D. (2006). Epidermal homeostasis: do committed progenitors work while stem cells sleep? *The Journal of investigative dermatology*.

Jones, P.H., Harper, S. & Watt, F.M. (1995). Stem cell patterning and fate in human epidermis. *Cell*.

Jouffe, C., Cretenet, G., Symul, L., Martin, E., Atger, F., Naef, F. & Gachon, F. (2013). The Circadian Clock Coordinates Ribosome Biogenesis. *PLoS biology*.

Katoh, K. & Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Briefings in bioinformatics*.

Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., Kohler, C., Khadake, J., Leroy, C., Liban, A., Lieftink, C., Montecchi-Palazzi, L., Orchard, S., Risse, J., Robbe, K., Roechert, B., Thorneycroft, D., Zhang, Y., Apweiler, R. & Hermjakob, H. (2007). IntAct–open source resource for molecular interaction data. *Nucleic Acids Research*.

Kiel, C., Beltrao, P. & Serrano, L. (2008). Analyzing Protein Interaction Networks Using Structural Information. *Annual Review of Biochemistry*.

Kiel, C., Vogt, A., Campagna, A., Chatr-aryamontri, A., Lange, M.S.d., Beer, M., Bolz, S., Mack, A.F., Kinkl, N., Cesareni, G., Serrano, L. & Ueffing, M. (2011). Structural and functional protein network analyses predict novel signaling functions for rhodopsin. *Molecular Systems Biology*.

Kiel, C., Verschueren, E., Yang, J.S. & Serrano, L. (2013). Integration of Protein Abundance and Structure Data Reveals Competition in the ErbB Signaling Network. *Science Signaling*.

Kim, P.M., Lu, L.J., Xia, Y. & Gerstein, M.B. (2006). Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*.

Koike, N., Yoo, S.H., Huang, H.C., Kumar, V., Lee, C., Kim, T.K. & Takahashi, J.S. (2012). Transcriptional architecture and chromatin landscape of the core circadian clock in mammals. *Science*.

Kolly, C., Suter, M.M. & Müller, E.J. (2005). Proliferation, cell cycle exit, and onset of terminal differentiation in cultured keratinocytes: pre-programmed pathways in control of C-Myc and Notch1 prevail over extracellular calcium signals. *The Journal of investigative dermatology*.

Koster, M.I. (2004). p63 is the molecular switch for initiation of an epithelial stratification program. *Genes & Development*.

Krause, R., von Mering, C., Bork, P. & Dandekar, T. (2004). Shared components of protein complexes?versatile building blocks or biochemical artefacts? *BioEssays*.

Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., Punna, T., Peregrín-Alvarez, J.M., Shales, M., Zhang, X., Davey, M., Robinson, M.D., Paccanaro, A., Bray, J.E., Sheung, A., Beattie, B., Richards, D.P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M.M., Vlasblom, J., Wu, S., Orsi, C., Collins, S.R., Chandran, S., Haw, R., Rilstone, J.J., Gandi, K., Thompson, N.J., Musso, G., St Onge, P., Ghanny, S., Lam, M.H.Y., Butland, G., Altaf-Ul, A.M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J.S., Ingles, C.J., Hughes, T.R., Parkinson, J., Gerstein, M., Wodak, S.J., Emili, A. & Greenblatt, J.F. (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*.

Kuhner, S., van Noort, V., Betts, M.J., Leo-Macias, A., Batisse, C., Rode, M., Yamada, T., Maier, T., Bader, S., Beltran-Alvarez, P., Castano-Diez, D., Chen, W.H., Devos, D., Guell, M., Norambuena, T., Racke, I., Rybin, V., Schmidt, A., Yus, E., Aebersold, R., Herrmann, R., Bottcher, B., Frangakis, A.S., Russell, R.B., Serrano, L., Bork, P. & Gavin, A.C. (2009). Proteome Organization in a Genome-Reduced Bacterium. *Science*.

Lack, L.C. & Lushington, K. (1996). The rhythms of human sleep propensity and core body temperature. *Journal of sleep research*.

Landan, G. & Graur, D. (2007). Heads or Tails: A Simple Reliability Check for Multiple Sequence Alignments. *Molecular Biology and Evolution*.

Lassmann, T., Frings, O. & Sonnhammer, E.L.L. (2009). Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic Acids Research*.

Lavker, R.M. & Matoltsy, A.G. (1970). Formation of horny cells: The fate of cell organelles and differentiation products in ruminal epithelium. *The Journal of cell biology*.

Lechler, T. & Fuchs, E. (2005). Asymmetric cell divisions promote stratification and differentiation of mammalian skin. *Nature*.

Lescuyer, P., Hochstrasser, D.F. & Sanchez, J.C. (2004). Comprehensive proteome analysis by chromatographic protein prefractionation. *ELECTROPHORESIS*.

Levine, J.D., Funes, P., Dowse, H.B. & Hall, J.C. (2002). BMC Neuroscience | Full text | Signal analysis of behavioral and molecular cycles. *BMC Neuroscience*.

Li, S. (2004). A Map of the Interactome Network of the Metazoan C. elegans. *Science*.

Lichti, U., Anders, J. & Yuspa, S.H. (2008). Isolation and short-term culture of primary keratinocytes, hair follicle populations and dermal cells from newborn mice and keratinocytes from adult mice for in vitro analysis and for grafting to immunodeficient mice. *Nature Protocols*.

Lim, X., Tan, S.H., Koh, W.L.C., Chau, R.M.W., Yan, K.S., Kuo, C.J., van Amerongen, R., Klein, A.M. & Nusse, R. (2013). Interfollicular Epidermal Stem Cells Self-Renew via Autocrine Wnt Signaling. *Science*.

Lin, K.K., Kumar, V., Geyfman, M., Chudova, D., Ihler, A.T., Smyth, P., Paus, R., Takahashi, J.S. & Andersen, B. (2009). Circadian Clock Genes Contribute to the Regulation of Hair Follicle Cycling. *PLoS genetics*.

Linding, R. (2003). GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Research*.

Logan, C.Y. & Nusse, R. (2004). The Wnt signalling pathway in development and disease. *Annual Review of Cell and Developmental Biology*.

Lucas, D., Battista, M., Shi, P.A., Isola, L. & Frenette, P.S. (2008). Mobilized Hematopoietic Stem Cell Yield Depends on Species-Specific Circadian Timing. *Cell stem cell*.

Luo, F., Li, B., Wan, X.F. & Scheuermann, R.H. (2009). Core and periphery structures in protein interaction networks. *BMC bioinformatics*.

MacDonald, B.T., Tamai, K. & He, X. (2009). Wnt/b-Catenin Signaling:Components, Mechanisms, and Diseases. *Developmental Cell*.

Magrane, M. & Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database*.

Malovannaya, A., Lanz, R.B., Jung, S.Y., Bulynko, Y., Le, N.T., Chan, D.W., Ding, C., Shi, Y., Yucer, N., Krenciute, G., Kim, B.J., Li, C., Chen, R., Li, W., Wang, Y., O'Malley, B.W. & Qin, J. (2011). Analysis of the Human Endogenous Coregulator Complexome. *Cell*.

Marcotte, R., Brown, K.R., Suarez, F., Sayad, A., Karamboulas, K., Krzyzanowski, P.M., Sircoulomb, F., Medrano, M., Fedyshyn, Y., Koh, J.L.Y., van Dyk, D., Fedyshyn, B., Luhova, M., Brito, G.C., Vizeacoumar, F.J., Vizeacoumar, F.S., Datti, A., Kasimer, D., Buzina, A., Mero, P., Misquitta, C., Normand, J., Haider, M., Ketela, T., Wrana, J.L., Rottapel, R., Neel, B.G. & Moffat, J. (2012). Essential Gene Profiles in Breast, Pancreatic, and Ovarian Cancer Cells. *Cancer Discovery*.

Marsh, J.A. & Teichmann, S.A. (2014). Protein Flexibility Facilitates Quaternary Structure Assembly and Evolution. *PLoS biology*.

Massagué, J. & Gomis, R.R. (2006). The logic of TGFβ signaling. *FEBS Letters*.

Masse, I., Barbollat-Boutrand, L., Molina, M., Berthier-Vergnes, O., Joly-Tonetti, N., Martin, M.T., de Fromentel, C.C., Kanitakis, J. & Lamartine, J. (2012). Functional

interplay between p63 and p53 controls RUNX1 function in the transition from proliferation to differentiation in human keratinocytes. *Cell Death and Disease*.

McCarthy, J.J., Andrews, J.L., McDearmon, E.L., Campbell, K.S., Barber, B.K., Miller, B.H., Walker, J.R., Hogenesch, J.B., Takahashi, J.S. & Esser, K.A. (2007). Identification of the circadian transcriptome in adult mouse skeletal muscle. *Physiological Genomics*.

McMahon, A., Butovich, I.A., Mata, N.L., Klein, M., Ritter, R., Richardson, J., Birch, D.G., Edwards, A.O. & Kedzierski, W. (2007). Retinal pathology and skin barrier defect in mice carrying a Stargardt disease-3 mutation in elongase of very long chain fatty acids-4. *Molecular vision*.

Mendel, G. (1865). Experiments in plant hybridization (1865). *Read at the February*.

Méndez-Ferrer, S., Lucas, D., Battista, M. & Frenette, P.S. (2008). Haematopoietic stem cell release is regulated by circadian oscillations. *Nature*.

Mendoza-Parra, M.A., Walia, M., Sankar, M. & Gronemeyer, H. (2011). Dissecting the retinoid-induced differentiation of F9 embryonal stem cells by integrative genomics. *Molecular Systems Biology*.

Menet, J.S., Rodriguez, J., Abruzzi, K.C. & Rosbash, M. (2012). Nascent-Seq reveals novel features of mouse circadian transcriptional regulation. *eLife*.

Menon, G.K., Elias, P.M., Lee, S.H. & Feingold, K.R. (1992). Localization of calcium in murine epidermis following disruption and repair of the permeability barrier. *Cell and Tissue Research*.

Michael E Hughes, L.D.K.H.S.R.P.S.P.J.H. (2007). High resolution time course analysis of gene expression from the liver and pituitary. *Cold Spring Harbor symposia on quantitative biology*.

Mills, A.A., Zheng, B., Wang, X.J., Vogel, H., Roop, D.R. & Bradley, A. (1999). p63 is a p53 homologue required for limb and epidermal morphogenesis. *Nature*.

Mishra, G.R. (2006). Human protein reference database–2006 update. *Nucleic Acids Research*.

Moore, K.A. & Lemischka, I.R. (2006). Stem cells and their niches. *Science*.

Moriyama, M., Durham, A.D., Moriyama, H., Hasegawa, K., Nishikawa, S.I., Radtke, F. & Osawa, M. (2008). Multiple Roles of Notch Signaling in the Regulation of Epidermal Development. *Developmental Cell*.

Morris, R.J., Liu, Y., Marles, L., Yang, Z., Trempus, C., Li, S., Lin, J.S., Sawicki, J.A. & Cotsarelis, G. (2004). Capturing and profiling adult hair follicle stem cells. *Nature Biotechnology*.

Mosca, R., Céol, A. & Aloy, P. (2012). Interactome3D: adding structural details to protein networks. *Nature methods*.

Muller, H.J. (1927). Artificial transmutation of the gene. *Science*.

Muller, H.J. (1928). The Production of Mutations by X-Rays. *Proceedings of the National Academy of Sciences of the United States of America*.

Nepusz, T., Yu, H. & Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. *Nature methods*.

Newman, M.E.J. (2004). Fast algorithm for detecting community structure in networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*.

Nirenberg, M. & Leder, P. (1964). RNA codewords and protein synthesis. The effect of trinucleotides upon the binding of SRNA to ribosomes. *Science*.

Nomura, K., Takeuchi, Y., Yamaguchi, S., Okamura, H. & Fukunaga, K. (2003). Involvement of calcium/calmodulin-dependent protein kinase II in the induction of mPer1. *Journal of neuroscience research*.

Nomura, K., Takeuchi, Y. & Fukunaga, K. (2006). MAP kinase additively activates the mouse Per1 gene promoter with CaM kinase II. *Brain Research*.

Novershtern, N., Subramanian, A., Lawton, L.N., Mak, R.H., Haining, W.N., McConkey, M.E., Habib, N., Yosef, N., Chang, C.Y., Shay, T., Frampton, G.M., Drake, A.C.B., Leskov, I., Nilsson, B., Preffer, F., Dombkowski, D., Evans, J.W., Liefeld, T., Smutko, J.S., Chen, J., Friedman, N., Young, R.A., Golub, T.R., Regev, A. & Ebert, B.L. (2011).

Densely Interconnected Transcriptional Circuits Control Cell States in Human Hematopoiesis. *Cell*.

Oettgen, P., Alani, R.M., Barcinski, M.A., Brown, L., Akbarali, Y., Boltax, J., Kunsch, C., Munger, K. & Libermann, T.A. (1997). Isolation and characterization of a novel epithelium-specific transcription factor, ESE-1, a member of the ets family. *Molecular and Cellular Biology*.

Ogi, T., Limsirichaikul, S., Overmeer, R.M., Volker, M., Takenaka, K., Cloney, R., Nakazawa, Y., Niimi, A., Miki, Y., Jaspers, N.G., Mullenders, L.H.F., Yamashita, S., Fousteri, M.I. & Lehmann, A.R. (2010). Three DNA Polymerases, Recruited by Different Mechanisms, Carry Out NER Repair Synthesis in Human Cells. *Molecular cell*.

O'Leary, M.N., Schreiber, K.H., Zhang, Y., Duc, A.C.E., Rao, S., Hale, J.S., Academia, E.C., Shah, S.R., Morton, J.F., Holstein, C.A., Martin, D.B., Kaeberlein, M., Ladiges, W.C., Fink, P.J., MacKay, V.L., Wiest, D.L. & Kennedy, B.K. (2013). The Ribosomal Protein Rpl22 Controls Ribosome Composition by Directly Repressing Expression of Its Own Paralog, Rpl22l1. *PLoS genetics*.

Oliver, S. (2000). Guilt-by-association goes global. *Nature*.

O'Neill, J.S. & Reddy, A.B. (2012). Circadian clocks in human red blood cells. *Nature*.

O'Neill, J.S., van Ooijen, G., Dixon, L.E., Troein, C., Corellou, F., Bouget, F.Y., Reddy, A.B. & Millar, A.J. (2012). Circadian rhythms persist without transcription in a eukaryote. *Nature*.

Oshimori, N. & Fuchs, E. (2012). Paracrine TGF-b Signaling Counterbalances BMP-Mediated Repressionin Hair Follicle Stem Cell Activation. *Stem Cell*.

Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stumpflen, V., Mewes, H.W., Ruepp, A. & Frishman, D. (2005). The MIPS mammalian protein-protein interaction database. *Bioinformatics (Oxford, England)*.

Palla, G., Derényi, I., Farkas, I. & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*.

Panda, S. (2007). Multiple Photopigments Entrain the Mammalian Circadian Oscillator. *Neuron*.

Panda, S., Antoch, M.P., Miller, B.H., Su, A.I., Schook, A.B., Straume, M., Schultz, P.G., Kay, S.A., Takahashi, J.S. & Hogenesch, J.B. (2002). Coordinated Transcription of Key Pathways in the Mouse by the Circadian Clock. *Cell*.

Paolo Dotto, G. (1999). Signal Transduction Pathways Controlling the Switch Between Keratinocyte Growth and Differentiation. *Critical Reviews in Oral Biology & Medicine*.

Paragh, G., Schling, P., Ugocsai, P., Kel, A.E., Liebisch, G., Heimerl, S., Moehle, C., Schiemann, Y., Wegmann, M., Farwick, M., Wikonkál, N.M., Mandl, J., Langmann, T. & Schmitz, G. (2008). Novel sphingolipid derivatives promote keratinocyte differentiation. *Experimental Dermatology*.

Pereira-Leal, J.B. (2005). Novel specificities emerge by stepwise duplication of functional modules. *Genome Research*.

Pereira-Leal, J.B., Levy, E.D., Kamp, C. & Teichmann, S.A. (2007). Evolution of protein complexes by duplication of homomeric interactions. *Genome biology*.

Pittendrigh, C.S. (1993). Temporal organization: reflections of a Darwinian clock-watcher. *Annual review of physiology*.

Plikus, M.V., Mayer, J.A., de la Cruz, D., Baker, R.E., Maini, P.K., Maxson, R. & Chuong, C.M. (2008). Cyclic dermal BMP signalling regulates stem cell activation during hair regeneration. *Nature*.

Prudhomme, W., Daley, G.Q., Zandstra, P. & Lauffenburger, D.A. (2004). Multivariate proteomic analysis of murine embryonic stem cell self-renewal versus differentiation signaling. *Proceedings of the National Academy of Sciences of the United States of America*.

Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M. & Séraphin, B. (2001). The Tandem Affinity Purification (TAP) Method: A General Procedure of Protein Complex Purification. *Methods*.

Pulimeno, P., Paschoud, S. & Citi, S. (2011). A Role for ZO-1 and PLEKHA7 in Recruiting Paracingulin to Tight and Adherens Junctions of Epithelial Cells. *Journal of Biological Chemistry*.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E.L.L., Eddy, S.R., Bateman, A. & Finn, R.D. (2012). The Pfam protein families database. *Nucleic Acids Research*.

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V. & Parisi, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*.

Rain, J.C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schächter, V., Chemama, Y., Labigne, A. & Legrain, P. (2001). The protein-protein interaction map of Helicobacter pylori. *Nature*.

Ranganathan, P., Agrawal, A., Bhushan, R., Chavalmane, A.K., Kalathur, R., Takahashi, T. & Kondaiah, P. (2007). Expression profiling of genes regulated by TGF-beta: Differential regulation in normal and tumour cells. *BMC Genomics*.

Rangarajan, A., Talora, C., Okuyama, R., Nicolas, M., Mammucari, C., Oh, H., Aster, J.C., Krishna, S., Metzger, D. & Chambon, P. (2001). Notch signaling is a direct determinant of keratinocyte growth arrest and entry into differentiation. *The EMBO Journal*.

Rheinwald, J.G. & Green, H. (1975a). Formation of a keratinizing epithelium in culture by a cloned cell line derived from a teratoma. *Cell*.

Rheinwald, J.G. & Green, H. (1975b). Serial cultivation of strains of human epidermal keratinocytes: the formation of keratinizing colonies from single cells. *Cell*.

Rice, R.H. & Green, H. (1977). The cornified envelope of terminally differentiated human epidermal keratinocytes consists of cross-linked protein. *Cell*.

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M. & Séraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology*.

Ritchie, M.E., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A. & Smyth, G.K. (2007). A comparison of background correction methods for two-colour microarrays. *Bioinformatics* (*Oxford, England*).

Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D.S., Zhang, L.V., Wong, S.L., Franklin, G., Li, S., Albala, J.S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R.S., Vandenhaute, J., Zoghbi, H.Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M.E., Hill, D.E., Roth, F.P. & Vidal, M. (2005). Towards a proteome-scale map of the human protein–protein interaction network. *Nature*.

Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Stransky, M., Waegele, B., Schmidt, T., Doudieu, O.N., Stümpflen, V. & Mewes, H.W. (2008). CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Research*.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C. & Mewes, H.W. (2010). CORUM: the comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Research*.

Russell, R.B. & Gibson, T.J. (2008). A careful disorderliness in the proteome: Sites for interaction and targets for future therapies. *FEBS Letters*.

Sahar, S. & Sassone-Corsi, P. (2009). Metabolism and cancer:the circadian clock connection. *Nature Reviews*.

Salwinski, L. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research*.

Sando, G.N., Zhu, H., Weis, J.M., Richman, J.T., Wertz, P.W. & Madison, K.C. (2003). Caveolin expression and localization in human keratinocytes suggest a role in lamellar granule biogenesis. *The Journal of investigative dermatology*.

Santa Maria, S.R., Gangavarapu, V., Johnson, R.E., Prakash, L. & Prakash, S. (2007). Requirement of Nse1, a Subunit of the Smc5-Smc6 Complex, for Rad52-Dependent Postreplication Repair of UV-Damaged DNA in Saccharomyces cerevisiae. *Molecular and Cellular Biology*.

Schaefer, M.H., Fontaine, J.F., Vinayagam, A., Porras, P., Wanker, E.E. & Andrade-Navarro, M.A. (2012). HIPPIE: Integrating Protein Interaction Networks with Experiment Based Quality Scores. *PLoS ONE*.

Schell, H., Hornstein, O.P., Egdmann, W. & Schwarz, W. (1981a). Evidence of diurnal variation of human epidermal cell proliferation. II. Duration of epidermal DNA synthesis. *Archives of dermatological research*.

Schell, H., Schwarz, W., Hornstein, O.P., Bernlochner, W. & Weghorn, C. (1981b). Evidence of diurnal variation of human epidermal cell proliferation. I. Epidermal 3H-labeling index and serum cortisol rhythm. *Archives of dermatological research*.

Schell, H., Hornstein, O.P. & Schwarz, W. (1983). Circadian rhythm of DNA-labeling index in normal human epidermis. *The Journal of investigative dermatology*.

Schlüter, H., Apweiler, R., Holzhütter, H.G. & Jungblut, P.R. (2009). Finding one's way in proteomics: a protein species nomenclature. *Chemistry Central Journal*.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. & Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*.

Sen, G.L., Webster, D.E., Barragan, D.I., Chang, H.Y. & Khavari, P.A. (2008). Control of differentiation in a self-renewing mammalian tissue by the histone demethylase JMJD3. *Genes & Development*.

Sen, G.L., Reuter, J.A., Webster, D.E., Zhu, L. & Khavari, P.A. (2010). DNMT1 maintains progenitor function in self-renewing somatic tissue. *Nature*.

Senoo, M., Pinto, F., Crum, C.P. & McKeon, F. (2007). p63 Is Essential for the Proliferative Potential of Stem Cells in Stratified Epithelia. *Cell*.

Sertic, S., Pizzi, S., Cloney, R., Lehmann, A.R., Marini, F., Plevani, P. & Muzi-Falconi, M. (2011). Human exonuclease 1 connects nucleotide excision repair (NER) processing with checkpoint activation in response to UV irradiation. *Proceedings of the National Academy of Sciences of the United States of America*.

Shen, L., Qu, X., Ma, Y., Zheng, J., Chu, D., Liu, B., Li, X., Wang, M., Xu, C., Liu, N., Yao, L. & Zhang, J. (2014). Tumor suppressor NDRG2 tips the balance of oncogenic TGF-b via EMT inhibition in colorectal cancer. *Oncogenesis*.

Shimoda, Y., Shinpo, S., Kohara, M., Nakamura, Y., Tabata, S. & Sato, S. (2008). A Large Scale Analysis of Protein-Protein Interactions in the Nitrogen-fixing Bacterium Mesorhizobium loti. *DNA Research*.

Smith, S.A., Ray, D., Cook, K.B., Mallory, M.J., Hughes, T.R. & Lynch, K.W. (2013). Paralogs hnRNP L and hnRNP LL Exhibit Overlapping but Distinct RNA Binding Constraints. *PLoS ONE*.

Sowa, M.E., Bennett, E.J., Gygi, S.P. & Harper, J.W. (2009). Defining the Human Deubiquitinating Enzyme Interaction Landscape. *Cell*.

Spörl, F., Korge, S., Jürchott, K., Wunderskirchner, M., Schellenberg, K., Heins, S., Specht, A., Stoll, C., Klemz, R., Maier, B., Wenck, H., Schrader, A., Kunz, D., Blatt, T. & Kramer, A. (2012). Krüppel-like factor 9 is a circadian transcription factor in human epidermis that controls proliferation of keratinocytes. *Proceedings of the National Academy of Sciences of the United States of America*.

Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A. & Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*.

Stein, A., Mosca, R. & Aloy, P. (2011). Three-dimensional modeling of protein interactions and complexes is going 'omics. *Current Opinion in Structural Biology*.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C.,

Bock, N., Kietzmann, S., Goedde, A., Toksöz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H. & Wanker, E.E. (2005). A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome. *Cell*.

Storch, K.F., Lipan, O., Leykin, I., Viswanathan, N., Davis, F.C., Wong, W.H. & Weitz, C.J. (2002). Extensive and divergent circadian gene expression in liver and heart. *Nature*.

Straume, M. (2004). DNA microarray time series analysis: automated statistical assessment of circadian rhythms in gene expression patterning. *Methods in enzymology*.

Sturtevant, A.H. (1913). The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *Journal of experimental zoology*.

Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M.P., Walker, J.R. & Hogenesch, J.B. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*.

Tatum, E.L. & Beadle, G.W. (1942). Genetic Control of Biochemical Reactions in Neurospora: An "Aminobenzoicless" Mutant. *Proceedings of the National Academy of Sciences of the United States of America*.

Tatusov, R.L. (1997). A Genomic Perspective on Protein Families. *Science*.

Taylor, I.W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., Bull, S., Pawson, T., Morris, Q. & Wrana, J.L. (2009a). Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature Biotechnology*.

Taylor, J.M., Street, T.L., Hao, L., Copley, R., Taylor, M.S., Hayden, P.J., Stolper, G., Mott, R., Hein, J., Moffatt, M.F. & Cookson, W.O.C.M. (2009b). Dynamic and physical clustering of gene expression during epidermal barrier formation in differentiating keratinocytes. *PLoS ONE*.

Truong, A.B., Kretz, M., Ridky, T.W., Kimmel, R. & Khavari, P.A. (2006). p63 regulates proliferation and differentiation of developmentally mature keratinocytes. *Genes & Development*.

Tumbar, T. (2004). Defining the Epithelial Stem Cell Niche in Skin. *Science*.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. & Rothberg, J.M. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*.

van de Peppel, J., Kemmeren, P., van Bakel, H., Radonjic, M., van Leenen, D. & Holstege, F.C.P. (2003). Monitoring global messenger RNA changes in externally controlled microarray experiments. *EMBO reports*.

Vavouri, T., Semple, J.I., Garcia-Verdugo, R. & Lehner, B. (2009). Intrinsic Protein Disorder and Interaction Promiscuity Are Widely Associated with Dosage Sensitivity. *Cell*.

Vidal, M. (2001). A biological atlas of functional maps. *Cell*.

Vidal, M., Cusick, M.E. & Barabási, A.L. (2011). Interactome Networks and Human Disease. *Cell*.

Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R. & Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research*.

von Kriegsheim, A., Baiocchi, D., Birtwistle, M., Sumpton, D., Bienvenut, W., Morrice, N., Yamada, K., Lamond, A., Kalna, G., Orton, R., Gilbert, D. & Kolch, W. (2009). Cell fate decisions are specified by the dynamic ERK interactome. *Nature Cell Biology*.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. & Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*.

Walhout, A.J., Sordella, R., Lu, X., Hartley, J.L., Temple, G.F., Brasch, M.A., Thierry-Mieg, N. & Vidal, M. (2000). Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science*.

Wallace, I.M. (2006). M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Research*.

Wang, W., Barnaby, J.Y., Tada, Y., Li, H., Tör, M., Caldelari, D., Lee, D.u., Fu, X.D. & Dong, X. (2012). Timing of plant immune responses by a central circadian regulator. *Nature*.

Wang, X.J., Liefer, K.M., Tsai, S., O'Malley, B.W. & Roop, D.R. (1999). Development of gene-switch transgenic mice that inducibly express transforming growth factor beta1 in the epidermis. *Proceedings of the National Academy of Sciences of the United States of America*.

Watabe, T. & Miyazono, K. (2009). Roles of TGF-β family signaling in stem cell renewal and differentiation. *Cell Research*.

Watanabe, M., Hida, A., Kitamura, S., Enomoto, M., Ohsawa, Y., Katayose, Y., Nozaki, K., Moriguchi, Y., Aritake, S., Higuchi, S., Tamura, M., Kato, M. & Mishima, K. (2012). Biochemical and Biophysical Research Communications. *Biochemical and Biophysical Research Communications*.

Watt, F.M. & Green, H. (1982). Stratification and terminal differentiation of cultured epidermal cells. *Nature*.

Watt, F.M., Celso, C.L. & Silva-Vargas, V. (2006). Epidermal stem cells: an update. *Current Opinion in Genetics & Development*.

Watt, F.M., Estrach, S. & Ambler, C.A. (2008). Epidermal Notch signalling: differentiation, cancer and adhesion. *Current Opinion in Cell Biology*.

Weeks, B.H., He, W., Olson, K.L. & Wang, X.J. (2001). Inducible expression of transforming growth factor beta1 in papillomas causes rapid metastasis. *Cancer research*.

Wichert, S., Fokianos, K. & Strimmer, K. (2003). Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics* (*Oxford, England*).

Woelfle, M.A., Ouyang, Y., Phanvijhitsiri, K. & Johnson, C.H. (2004). The Adaptive Value of Circadian Clocks. *Current Biology*.

Wu, J., Vallenius, T., Ovaska, K., Westermarck, J., Mäkelä, T.P. & Hautaniemi, S. (2008). Integrated network analysis platform for protein-protein interactions. *Nature methods*.

Yang, J.S., Campagna, A., Delgado, J., Vanhee, P., Serrano, L. & Kiel, C. (2012). SAPIN: a framework for the structural analysis of protein interaction networks. *Bioinformatics* (*Oxford, England*).

Yauk, C.L. (2004). Comprehensive comparison of six microarray technologies. *Nucleic Acids Research*.