A systems biology investigation into age-related changes in the maintenance of collagen and extracellular matrix in human skin.

Ciaran Welsh

Doctor of Philosophy

November 2018

The Institute for Cell & Molecular Biosciences

Abstract

As skin ages, its capacity for providing an effective barrier between an organism and the outside world diminishes. The dermal extracellular matrix (ECM) is a critical component of skin that provides structural integrity and nourishment to the avascular epidermis. The ability for the dermis to perform its function is dependent on its molecular composition, which is mostly a fibrous mesh of type 1 collagens, elastins, proteoglycans and glycoproteins. Fibroblasts are the resident caretakers of the dermis as they are responsible for its constant remodelling and turnover. With age the composition of the dermis changes. Collagen and elastin fibres gradually become fewer in number and disorganised while matrix metalloproteinases (MMPs) that degrade the ECM become more prevalent. A consequence of these changes is that the aged dermis has reduced mechanical and tensile strength compared to the young dermis. Transforming growth factor β (TGF-β) is a pleotropic cytokine that induces the synthesis and deposition of dermal ECM, amongst other functions. The molecular processes that lead the demise of skin integrity and the contribution of TGF-β to these processes are not well understood. The goal of this thesis is to investigate skin ageing using a systems biology approach to gain insight into the differences between the regulation of young and old fibroblasts by TGF-β.

This work describes the results of two high-throughput time series experiments. The first is an Affymetrix microarray experiment that was designed to understand how young fibroblasts transcriptionally respond to short term TGF-β treatment, with an emphasis on identifying transcriptional modes of cross-talk in the control of ECM synthesis. The second is a high-throughput quantitative polymerase chain reaction

(qPCR) experiment that measured 68 hand-selected genes involved in the production of ECM components or TGF-β signalling proteins. By comparing temporal data from neonatal, adult and senescent fibroblasts we provide insight into the transcriptional differences that exist between young and old fibroblasts.

The data collected using these high throughput methodologies were used in a combinatorial ordinary differential equation model selection problem that was designed to test the feasibility of four hypotheses concerning the fibroblast response to TGF-β in age and the role of connective tissue growth factor (CTGF) in TGF-β mediated collagen production. This work made use of two custom Python packages that are both available on Python package index (PyPI), PyCoTools (Welsh et., al. 2018) which is a toolbox for automating aspects of COPASI, a commonly used software package in systems modelling, and pytseries which provides a set of useful objects and methods for handling and manipulating time series data.

Preface

The work presented in the thesis was conducted as part of a collaboration between Newcastle University, Procter & Gamble (P&G) and Durham University. The overall aim is to further our understanding of skin ageing and to develop potential therapeutic targets for topical skin treatment. P&G are the funders of this effort and have access to state-of-the-art equipment for '-omic' scale experiments. Prof. Stefan Przyborski and his group at Durham University are experts in the field of skin ageing and tissue bioengineering while at Newcastle our expertise lies in data analysis and mechanistic modelling of biological networks.

In this thesis we describe two high-throughput experiments. The first in Chapter 4 was designed and conducted at P&G and the data was sent to me for analysis. The second in Chapter 5 was designed and analysed by me but the experiment was conducted by Dr Nicola Fullard at Durham University, who sent the samples to P&G for quantification using their high-throughput qPCR facilities. Though I did not perform the experiments myself, a limited description of the experimental procedures has been included for completeness. Other than the wet-lab experiments, all work conducted in this thesis is my own.

Acknowledgements

First and foremost I'd like to thank my supervisor for his guidance throughout my PhD. I would also like to thank Prof. Stefan Przyborski and his team, particularly Dr Nicola Fullard, at Durham University for performing the experimental work presented in Chapter 5. I would like to thank the organisers and funders of this project at P&G, Cincinnati, particularly Ryan Tasseff for constant sound advice throughout my PhD.

List of Figures

List of Tables

Contents

Chapter 1

Skin Ageing and TGF-β biology

1.1 Ageing: causes and mechanisms

Evolutionary theories of ageing attempt to understand why we age in terms of natural selection, where biological characteristics are selected in a population over generations. Characteristics that enhance an organism's capacity for survival are passed to progeny and are gradually incorporated into the species, while those which negatively impact survival eventually result in extinction. These concepts are known as adaptive and maladaptive evolution respectively. Alternatively, some biological characteristics are benign with respect to a species capacity for survival and do not impact survival in any meaningful way. These are non-adaptive characteristics.

A major argument in the study of ageing is whether ageing is an adaptive characteristic that specifically evolved for the benefit of the species as a whole or whether ageing is non-adaptive and occurs because of a by-product of life (Kowald & Kirkwood 2016). Early studies of ageing were conducted by Weismann et al. 1891 who suggested that ageing evolved to remove 'worn out' individuals from the population and that ageing was beneficial for the general population because it reduces competition for limited resources (Kirkwood & Cremer 1982). This argument is an adaptive view that suggests there are processes that actively drive the evolution of ageing. However, there are problems with the argument. Firstly, old age is rare in the wild due to the prevalence of

extrinsic mortality. Therefore, there is no evolutionary requirement to remove the elderly from the population. Moreover, this argument is circular because it suggests that old individuals are 'worn out' (i.e. frail), when its purpose is to explain the reason why old individuals become frail (Kirkwood 2005).

Medawar 1952 proposed the mutation accumulation theory of ageing which is essentially a non-adaptive view that ageing results from mutations accumulating over an evolutionary time scale (Medawar 1952). Medawar argued that if a gene or characteristic only manifests as deleterious after the age of sexual reproduction then the trait would still be passed to progeny. Thus Medawar made the assumption that selective pressures were negligible in the elderly because they are not of reproductive age.

In 1957 Williams proposed the antagonistic pleiotropy theory of ageing which builds on the idea that selective pressures decline with age (Williams 1957). He argued that animals in the wild do not often get old, since their life is full of dangers such as starvation and predation. Consequently, any characteristic that conveys a survival advantage when young is strongly selected by evolutionary pressures, even if that same characteristic eventually becomes harmful to the individual. Such characteristics will become established and result in the ageing phenotype.

Kirkwood 1977 refined these ideas further by proposing the disposable soma theory of ageing (Kirkwood 1977). Kirkwood differentiated between reproductive and somatic cellular processes and suggested that they are in competition for limited energy resources. He proposed that the amount of energy dedicated to either process presents a trade-off that has been optimised by evolution. Survival of the species relies on accurate reproduction since errors lead to an evolutionary disadvantage. But such accuracy requires energy expenditure, for example by genetic proof reading or removing erroneous proteins. However, not all energy can be spent on maintaining reproductive processes since an organism must be healthy enough to eventually procreate. Thus the disposable soma theory is similar in principle to the antagonistic pleiotropy theory, but here the antagonistic trait under examination is the amount of resource dedicated to reproduction, compared to maintenance of somatic or 'self' processes.

While evolutionary theories attempt to address why we age, they do not specifically examine ageing from a mechanistic perspective. Early studies into mechanisms of ageing assigned blame to a single cause such as the accumulation of damage from the action of free radicals (Harman 1972). A modern view is that ageing cannot be attributed to any one single cause but is a heterogeneous mix of contributing factors. Such a view has lead to the categorisation of a set of factors known as 'hallmarks' (Lopez-Otn et al. 2013, Aunan et al. 2016) in analogy to hallmarks of cancer (Hanahan & Weinberg 2011).

The integrity of an organism's genome is under constant threat from exogenous and endogenous factors, such as chemicals, pathogens, DNA replication errors or damaging by-products from the electron-transport chain. Most genetic lesions are repaired by an efficient DNA damage response (Pan et al. 2016). Though efficient, the DNA damage response is not perfect and when a lesion is not adequately repaired prior to proliferation, it may be incorporated into the genome. Occasionally, mutation combinations lead to diseases such as cancer, but more commonly they impact biological function in some small way or are benign. However, these mutations accumulate over the course of a life and so the stability of an organisms genome is an decreasing function of age. Although ageing itself is not a disease, the consequential genomic instability associated with ageing predisposes the elderly to diseases, such as cancer.

Most cells have a limited replicative potential (Hayflick & Moorhead 1961). During proliferation, cells must duplicate their DNA using molecular machinery known as DNA polymerases. This is a highly efficient process but with each successive cell division, non-coding chromosomal ends known as telomeres get progressively shorter. Eventually, after enough cell cycle iterations, telomeres become so short that they breach some gene-coding regions of DNA, which triggers a state of replicative cellular senescence and can no longer divide. Senescent cells undergo a number of phenotypic alterations, including changes in their secretory profile which can have paracrine effects on cells nearby (Nelson et al. 2012). This secretory profile is known as the age-associated secretory phenotype (Copp et al. 2010) whereby extraneous inflammatory cytokines and matrix metalloproteinases are secreted. The DNA damage response can activate senescence in response to multiple stimuli, with the aforementioned attrition in telomere

length being one mechanism. Other stimuli, such as DNA lesions (Di Micco et al. 2006), excessive exposure to reactive oxygen species (ROS) (Davalli et al. 2016), oncogene activation (Chandeck & Mooi 2010) and tumor suppressor gene activation (Qian & Chen 2010) are also capable of activating the DNA damage response and inducing cellular senescence. While senescent cells are usually removed by phagocytic immune processes, they accumulate with age either by enhanced production, decreased degradation or both. An increased prevalence of senescent cells exacerbates the chronic but low grade inflammation problem of the aged phenotype (Childs et al. 2015).

The term 'inflammaging' was coined in (Franceschi et al. 2000) to denote the observed increase in inflammation with age with concomitant reduction in the ability to cope with environmental stress. Inflammation is a complex series of protective biological responses that are stimulated by injury. The inflammatory response has phases: first blood flow to the injury is increased and immune cells are recruited to the area. If the damage is not severe, the inflammation may enter resolution phase after the injury is sufficiently dealt with. If the damage is more severe chronic inflammation may ensue, which is marked by macrophage and lymphocyte infiltration. Once the injury is eliminated, inflammation enters resolution phase where normality is restored. With age however, the capacity for inflammatory resolution is diminished and a persistent, low grade chronic state of inflammation drives disease progression and the ageing phenotype (Xia et al. 2016, Baylis et al. 2013)

Epigenetic changes are another hallmark of ageing (Lopez-Otn et al. 2013). DNA methylation patterns are shaped by the processes of adding and removing methyl groups into 5th position on cytosine residues. The amount of DNA methylation has been is thought to be a predictor of biological age and DNA hypermethylation is thought to predispose individuals diseases, by causing genomic instability (Jung & Pfeifer 2015).

Proteins are essential for all molecular processes and a considerable amount of energy goes into maintaining the processes which produce and degrade required proteins. The collection of processes which maintain a healthy proteome is referred to as proteostasis and involves mechanisms for monitoring protein folding for quality control and

maintaining correct protein localisation and concentrations (Roth & Balch 2011). Disruption of these processes can result in accumulation and aggregation of dysfunctional proteins and proteotoxic disruption of cellular processes (Lindner & Demarez 2009). With age, the amount of damaged proteins accumulate and result in diseases such as Alzheimer's, Parkinsons and Huntingtons disease (Saez & Vilchez 2014). Mechanistically, damaging agents such as ROS (Berlett & Stadtman 1997), impaired chaperone systems (Calderwood et al. 2009) and protein clearance mechanisms (i.e. autophagy) also play a role in proteostatic decline (Metcalf et al. 2012).

There is a large body of evidence indicating that DNA damage, cellular senescence, telomere attrition, inflammaging, epigenetic changes and loss of proteostasis are hallmark mechanisms of ageing. Other factors which are also considered hallmarks of ageing are: mitochondrial dysfunction, stem cell exhaustion, altered cellular communication, and deregulated nutrient sensing (Lopez-Otn et al. 2013, Aunan et al. 2016). These hallmarks are features of ageing in general. In the next sections we turn our attention to the structure, function and ageing of human skin.

1.2 Skin Ageing

Phenotypic changes in aged skin include dry, rough and itchy skin; uneven pigmentation; reduced capacity for healing and DNA repair; less and disorganised collagen and elastin networks; wrinkle formation; impaired dermal vasculature; reduced capacity for sensory perception, thermoregulation and immunosurveillance; impaired hair growth and sebaceous gland function; flattening of dermal papillae and reduced melanocyte concentrations; impaired skin barrier function and reduced cell turnover (Farage et al. 2009). Unlike other organs, the skin is exposed to high levels of extrinsic stress, particularly in sun-exposed areas because ultraviolet light represents the biggest risk factor for skin ageing. Other factors such as tobacco smoking, pollution and alcohol consumption also contribute towards skin ageing. Extrinsic skin ageing is superimposed

on intrinsic skin ageing and manifests as deeper, coarse wrinkle formation; dry leathery skin; uneven pigmentation; laxity and increased prevalence of telangiectasia (Farage et al. 2009, Sjerobabski-Masnec & Situm 2010).

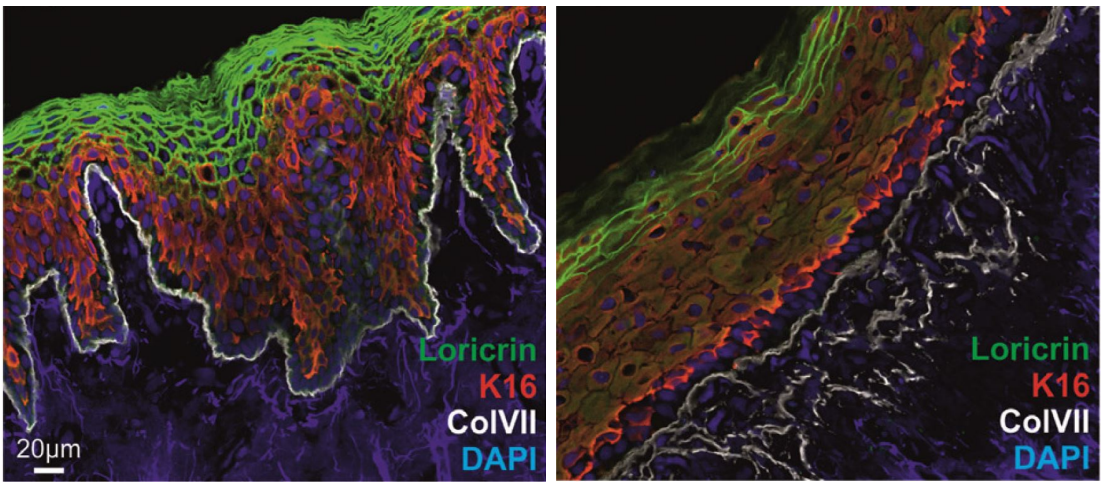
The integument system is a complex structure composed of its layers and appendages such as hair, nails and exocrine glands. All skin has the same tissue level structure: an outer epidermis which performs the barrier function, a thick dermis which provides strength and support for the epidermis and the subcutis which stores energy in the form of fat. The following sections describe the epidermis and the dermis with an emphasis on how they change with age.

1.2.1 Epidermis

The epidermis has functions beyond that of the barrier function and are performed by the various substructures within the epidermis. For example, sweat glands and hair follicles provide thermoregulation (Nadel et al. 1971) and controls water (Imokawa et al. 1989) and electrolyte (Jonsson & Benavente 1992) flux. Melanocytes produce melanin which assemble over keratinocyte nuclei and protect them from ultraviolet irradiation and the epidermis is a major anatomical site for immunosurveillance (Kupper & Fuhlbrigge 2004). The epidermis contains various mechanoreceptors that transduce sensations such as pain, temperature, pressure, touch, vibration and itchiness to the central nervous system (Khavkin & Ellis 2011).

The epidermis is a highly proliferative tissue and is separated into four morphologically distinct layers which are shown in Figure 1.1a: the stratum basale, the stratum spinosum, the stratum granulosum and the stratum corneum (Menon 2002). The epidermis is constantly in a state of proliferation and desquamation. The oldest cells are a layer of dead corneocytes on the skin surface that are constantly being shed at the same rate as newer cells replace them. This process, called keratinization, ensures the stratum corneum is optimally fit for its barrier and permeability functions.

The aged epidermis has several morphological characteristics which distinguish it from



(a) 27 year old epidermis. (b) 84 year old epidermis.

Figure 1.1: Immunofluorescence of epidermal tissue from (a) a 27 year old and (b) 84 year old foreskin tissue. Loricrin is a marker of the stratum corneum; kerarin 16 (K16) is a marker of live keratinocytes; COLVII is found within the basal lamina and DAPI stains nuclei. Images were produced by Andrea Trost at Salzburg University and are reproduced with permissions from (Breitenbach et al. 2015).

the young (Figure 1.1). The most obvious is the flattening of interdigitations between the dermal and epidermal compartments. In young tissue, these interdigitations increase the surface area and therefore adherence between the dermis and epidermis. In old tissue, interdigitation flattening causes loss of some of this strength. The basement membrane in young tissue is well-defined, with an obvious single-celled stratum basale layer from which keratinocytes proliferate. In old tissue however the basement membrane is clearly disorganised. In old tissue, keratinocytes of the stratum spinosum layer (stained red with keratin 16 in Figure 1.1) are larger and less densely packed than the young tissue. Moreover, old epidermis is thinner than in a young epidermis.

1.2.2 Dermis

The epidermis is separated from the dermis by a collagen IV rich basement membrane which is built and maintained by the keratinocytes of the stratum basale and papillary fibroblasts. The dermal-epidermal junction (DEJ) is vital for skin integrity as it regulates nutrient transport and enables paracrine communication between the dermis and epidermis (Burgeson & Christiano 1997). The dermis is situated directly beneath the DEJ and is a cellular sparse tissue consisting mostly of a fiberous extracellular

matrix (ECM). Components of the ECM include collagens, elastins, lamins, fibronectin, fibrulins, fibrillins, latent TGF-β complexes, matrillins, nidogens, tenascins, thrombospondins, vitronectin and an array of glycoproteins and proteoglycans (Lu et al. 2011).

The dermis has two morphologically distinct layers which are continuous but distinguished by anatomical location and by the local composition of the extracellular matrix (ECM). Dermal collagen is the major structural component of the dermis and is composed of three α chains. Type I collagen for instance is composed of two α1 and one α2 chain which together form anti-parallel fibrils (90nm in diameter) that are held together by interfibril cross-links. Collagen α chains are sequences of Gly-X-Y, where Gly is glycine and is responsible for collagens structure, and X and Y can be any amino acid but most commonly are proline and hydroxyproline (Fang et al. 2012). Occurrences of other amino acids are responsible for different subtypes of collagen. Dermal collagen is approximately 80% type I and 20% type III collagen (Figure 1.2) and together with elastins and other fibrillar components, form a structural mesh that is called the ECM. Most of the latter is in the upper or papillary dermis while type 1 collagen is ubiquitous but thicker in the reticular dermis (Meigel et al. 1977). Type 1 collagen has a long half-life of approximately 15 years (Verzijl et al. 2000).

Dermal fibroblasts are responsible for producing and maintaining ECM integrity. Collagens are produced as propeptides, the N- and C-terminal ends of which are proteolytically cleaved before assembly into collagen fibrils. Some of the lysine residues of procollagens are converted into hydroxylysine residues by lysyl hydroxylase before they are secreted. Once extracellular, amine groups of lysine and hydroxylysine residues are converted into aldehyde groups by lysyl oxidase (Tanzer 1973), the presence of which facilitates the spontaneous formation of collagen cross-linking that renders collagen molecules insoluble (Cole et al. 2018).

With age, both collagen and elastin networks become damaged (Figure 1.2 and Figure 1.3) particularly in photoexposed regions. Damaged ECM results in loss of strength and elasticity of the ECM and disorganisation of ECM components. Elastin

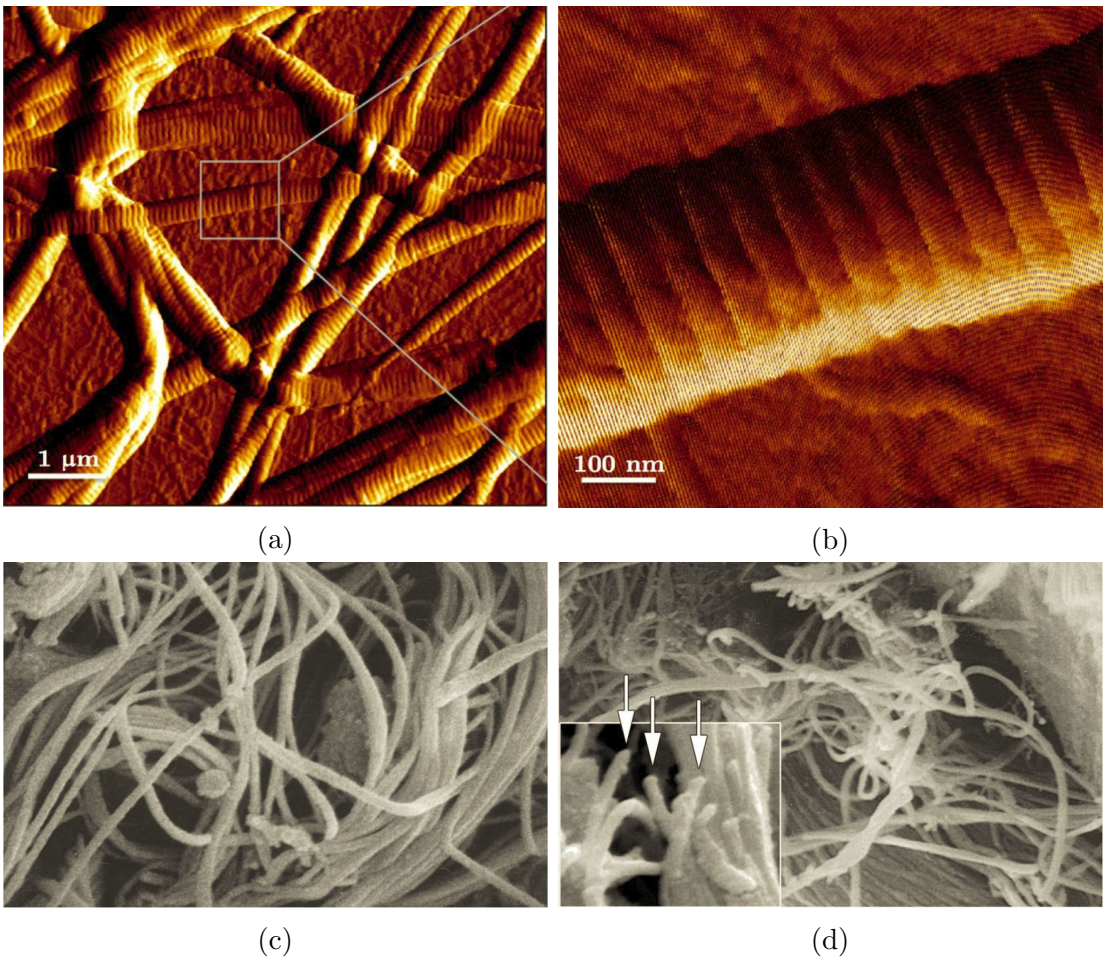
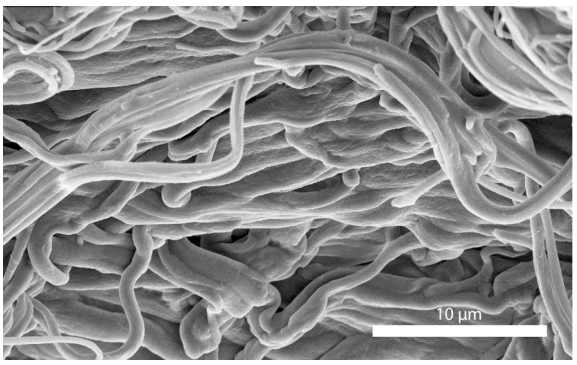


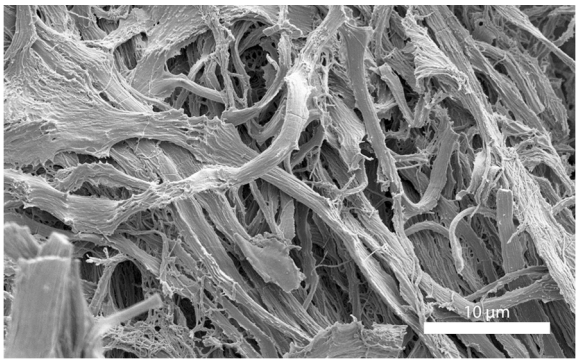
Figure 1.2: Structure of type 1 collagen. (a and b) Atom force microscopy (AFM) image of collagen type 1 at (a) micrometer scale and (b) nanometer scale. (c and d) scanning electron micrograph of collagen type 1 at x10000 magnification in (c) young and (d) old photodamaged tissue. (a and b) were produced by Meisam Asgari in Asgari et al. 2017 and reproduced under the creative commons licence (CC BY 4.0). (d and e) adapted from Cole et al. 2018 and reproduced with permission.

fibres which are thin and single stranded in young dermis, become beaded and loose connectivity with the epidermis in age (Seite et al. 2006). High levels of degradation and reduced replenishment by fibroblasts result in loss of collagen density and integrity (Quan et al. 2002, 2009, Quan & Fisher 2015). Since collagen has a long half-life, damaged or fragmented collagens accumulate and contribute towards the aged dermal phenotype (Figure 1.2) (Cole et al. 2018).

In healthy skin, fibroblasts perceive mechanical forces from external stimuli by integrins which tether fibroblasts to the ECM. Integrins adhere both to the ECM and to the fibroblast cytoskeleton, enabling the fibroblast to appropriately respond to disturbances in the ECM. This adhesion creates internal mechanical tension within the fibroblast. It



(a) 6 years old



(b) 90 years old

Figure 1.3: Electron micrograph of elastin network in dermis of (a) a 6 year old compared with that of (b) a 90 year old. These images were provided by Dr Christian Schmeltz and reproduced with his permission from (Ramamurthi & Kothapalli 2016), Chapter 1, Figure 1.4.

produces a stretched out morphology and is the normal healthy cellular state for fibroblasts (Varani et al. 2006). The reductions in quantity and quality of collagen in age reduces fibroblast adherence to the ECM and results in fibroblast collapse into a globular morphology with little cytoplasm (Quan & Fisher 2015, Cole et al. 2018). The resulting lack of tension changes how fibroblasts perceive the environment and leads to reduced ECM output. Consequently, a positive feedback of lower collagen production and lower mechanical tension results in a morphology called an 'age-associated dermal microenvironment' (AADM). The AADM not only affects collagen output but changes the composition of the ECM over time, as the fibroblasts adapt to the newer, less healthy dermal environment (Cole et al. 2018).

1.3 TGF-β Biology

Fibroblasts are receptive to extracellular signals that determine whether it should produce or degrade ECM components. TGF-β is a master regulator of the ECM that conveys the signal for fibroblasts to synthesize ECM components including collagens (Varga et al. 1987, Ignotz & Massague 1986), elastins (McGowan & McNamer 1990) and fibronectin (Ignotz & Massague 1986). It is for this reason that TGF-β is one of the main theme of this thesis. In the present section we introduce TGF-β biology, while in Chapter 4 we discuss TGF-β cross-talk and in Chapter 5 we discuss TGF-β biology in age.

The TGF-β superfamily is involved in a wide variety of cellular processes such as growth, differentiation, proliferation, cell cycle progression, adhesion, tissue development and repair processes. *De novo* TGF-β synthesis of all three TGF-β isoforms involves proteolytic cleavage by furin to produce the mature dimeric TGF-β molecule that is secreted into the extracellular matrix by fibroblasts in the form of the 290kDa large latent complex (LLC) (Miyazono et al. 1988). The LLC contains a 70kDa latency-associated peptide (LAP) which is non-covalently attached to the mature 25kDa TGF-β peptide, along with a 190kD latent TGF-β binding protein (LTBP) (Dallas et al. 1995).

Activation of TGF-β involves breaking the non-covalent bond that attaches TGF-β molecules to the LLC. Once broken, TGF-β is solubilized and free to engage cell surface receptors. TGF-β activation can occur by proteolytic mechanisms via enzymes such as plasmin (Lyons et al. 1990), MMP2 (Wang et al. 2006), thrombospondin (Schultz-Cherry & Murphy-Ullrich 1993) and MMP14 (Nguyen & Kang 2016) all of which cleave TGF-β precursors at the N-terminal Arg^{278} (Derynck et al. 1985). The LLC is attached to the ECM by integrins (such as $\alpha_5\beta_6$) and physical stimuli such as wounding can disturb the integrin-LLC bond and result in TGF-β release (Wipff et al. 2007).

TGF-β signalling is a well studied system and in recent years, advances in technology have enabled the observation of aspects of TGF-β signalling in higher resolution than

traditional methods in molecular biology (Zhang 2018). For instance, traditionally TGF-β was thought to bind to constitutively active homodimeric type 2 TGF-β receptors (R_2) leading to the resulting TGF-β-R_2 complex phosphorylating and binding to homodimeric type 1 receptors (R_1) and forming the mature ligand-receptor complex (LRC) (Cheifetz et al. 1988, Wrana et al. 1992). However, single molecule imaging has shown that R_2 exists primarily as monomers (Figure 1.4, 1) that undergo dimerization after TGF-β stimulation (Zhang et al. 2010). Moreover, TGF-β receptors are not stationary but are in constant flux to and from early endosomal and caveolae membrane compartments (Figure 1.4, 2) (Di Guglielmo et al. 2003). This flux occurs regardless of the presence of TGF-β and is an important source regulation for TGF-β signalling (Vilar et al. 2006). Traditionally, the activation and inactivation of TGF-β signalling was considered to be spatially segregated depending on the proteins which direct endocytosis. Receptor endocytosis can be directed by EEA1 into vesicles called early endosomes or by Cav1 into caveolae or caveosomes (Di Guglielmo et al. 2003). The latter was thought to be the anatomical site of LRC degradation by the proteasome, a process initiated by the canonical TGF-β negative feedback, Smad7 (Hayashi et al. 1997). For a long time this work shaped our understanding of TGF-β biology. More recently, however, real-time imaging studies revealed that caveolae and early endosomes can fuse cytoplasmically; were positive for both EEA1 and CAV1 (markers of early endosomes and caveolin 1 respectively) and capable of coordinating both positive and negative TGF-β signalling (Figure 1.4, 2) (He et al. 2015).

The core conduit of TGF-β signalling is the Smad second messenger system (Figure 1.4, 3). The established view is that receptor Smads (R-Smads), which include Smad1/2/3, are recruited to the activated LRC at early endosomes for phosphorylation of the carboxyl terminus of the R-Smad protein. This process is facilitated by accessory proteins such as Sara, which present R-Smads to the activated LRC (Penheiter et al. 2002). More recent evidence however, showed with live cell, single molecule imaging that R-Smad recruitment to the LRC is not ligand dependent but R-Smads are in constant association with R_1 (Li et al. 2016).

After phosphorylation, R-Smads can complex with Smad4 or Smad5 with a 2:1

stoichiometry (Inman et al. 2002). Once complexed, a nuclear import signal is exposed on the Smad4/5 protein resulting in nuclear translocation and Smad-mediated transcription (Figure 1.4, 4). Smad7 is under the transcriptional control of nuclear Smad complexes (Figure 1.4, 5). Smad7 biology is complex and although a lot of research has focused on Smad7, precisely how Smad7 controls the dynamics of Smad proteins is not well understood. Smad7 has several modes of feedback on the TGF-β system, as well as cross-talk with other networks (Yan et al. 2017). One of the best characterized mechanisms of the Smad7 negative feedback is that it binds to Smurf1/2 in the nucleus and transports it to the plasma membrane, where Smurf proteins can ubiquitinate R_1 and target it (along with Smad7) for degradation (Figure 1.4, 6) (Kavsak et al. 2000, Ebisawa et al. 2001, Suzuki et al. 2002). Moreover, the physical act of Smad7 binding to the R_1 is able to prevent R-Smads from accessing the receptor, thereby preventing phosphorylation and activation (Nakao et al. 1997, Hayashi et al. 1997, Souchelnytskyi et al. 1998). While at the receptor complex, Smad7 can additionally recruit proteins, such as GADD43-PP1c (Shi et al. 2004) or PP1α (Valdimarsdottir et al. 2006) which dephosphorylate and inactivate the ligand receptor complex. Smad7 can also bind to and sequester cytoplasmic Smad proteins (Figure 1.4, 7) (Yan et al. 2016). Smad7 is under intense regulation, for example by Arkadia which targets Smad7 for degradation (Koinuma et al. 2003, Liu et al. 2006). Smad7 stability is regulated by its acetylation status, as acetylation on Lys^{64} and Lys^{70} by p300, which prolongs the Smad7 half-life (Simonsson et al. 2005). Conversely, deacetylation, for instance by HDAC1 (Gronroos et al. 2002) and SIRT1 (Kume et al. 2007), diminish the Smad7 half-life.

Smad7 is not the only negative feedback that inhibits Smad signalling. In fact, there are many feedback mechanisms that exist at multiple levels of TGF-β signalling, including feedbacks on the activation of TGF-β ligands from the latent ECM reservoir; the activation and location status of TGF-β receptors; the Smads nuclear shuttling and transcriptional activity and by microRNA regulation (reviewed extensively in Yan et al. 2017).

Two well studied negative regulators of the TGF-β system are Ski and SNoN. These proteins are transcriptional co-repressors at Smad binding elements on DNA (Akiyoshi

et al. 1999, Stroschein et al. 1999). They are void of catalytic activity and cannot bind DNA directly but instead bind to nuclear complexed Smad proteins and inhibit their ability to bind DNA (Wu et al. 2002). While SNoN is a TGF-β responsive gene and therefore represents one of the many negative feedbacks of the TGF-β system (Akiyoshi et al. 1999), it is unclear as to whether Ski is a negative feedback or not. It has been shown that Ski protein was increased at low, but decreased at high concentrations of TGF-β treatment in fibroblasts (Liu et al. 2008), suggesting that Ski dynamics are highly non-linear.

Given the intricate connections between TGF-β signalling, the dermal ECM and general skin health, a better understanding of how TGF-β signals in young and old cells has the potential to facilitate pursuit of therapeutic or cosmetic intervention. In the following chapters we discuss the aims and objectives of this thesis; methods in modern systems biology; tools developed in this thesis for systems modelling and computational manipulation of time series data; a transcriptomic analysis of the early response to TGF-β in neonatal fibroblasts; a high-throughput qPCR analysis of transcriptional differences between neonatal, adult and senescent cell lines and a computational model examining the cross-talk between TGF-β and CTGF in neonatal and adult fibroblasts.

1.4 Aims and objectives

1. Investigate the basal differences between neonatal, senescent and adult fibroblasts.
2. Investigate the differences between neonatal, senescent and adult fibroblasts in response to TGF-β.
3. Investigate TGF-β cross-talk in the healthy young fibroblasts.
4. Explore means of unifying big data with mechanistic modelling.
5. Build a ODE model based on the data collected in this work that investigates:
   (a) The interaction between CTGF and TGF-β in young and old fibroblasts.

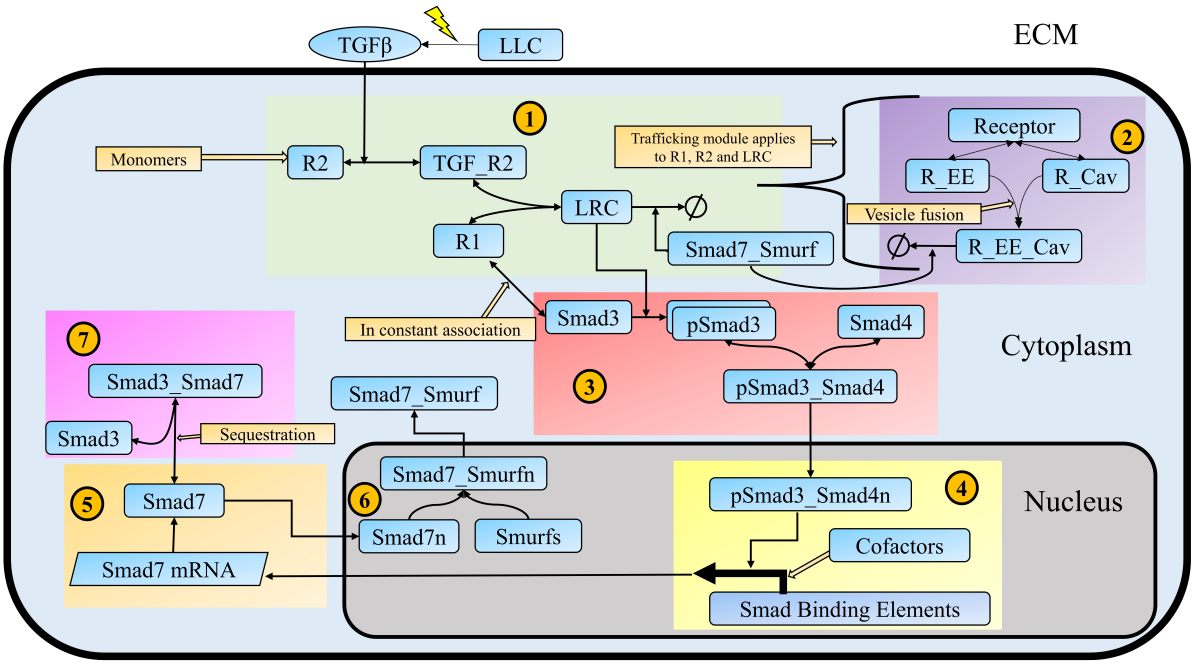


Figure 1.4: Schematic representation of TGF-β network. (1) TGF-β binds to R_2 monomers which dimerize before binding R_1. (2) R_1, R_2 and LRC all undergo internalisation to early early endosomes or caveosomes both of which can fuse cytoplasmically. (3) Smad3 (and Smad2) are phosphorylated by LRC and bind to Smad4. Smad3 is in a constant state of association and dissociation with R_1, irrespective of ligand. (4) Smad complexes transition into the cell nucleus where it can modulate genes containing Smad binding elements with the assistance of cofactors. (5) Transcription and translation of some genes, such as Smad7 can, influence TGF-β network dynamics. For example (6) newly synthesised cytoplasmic Smad7 translocates to the nucleus to associate with Smurf proteins, transition to the cell surface and inhibit LRC in caveosomes and in early endosome-caveolae fusion vesicles. (7) Smad7 can also sequester Smad3 cytoplasmically. LLC: large latent complex; R_1 and R_2: TGF-β type 1 and 2 receptors respectively; LRC: Ligand receptor complex, pSmad3: R_EE: R_1, R_2 or LRC in early endosomes; R_Cav: R1, R_2 or LRC in caveolae. Arrows represent biological processes such as biochemical reactions or endocytosis and double ended arrows represent reversible processes.

(b) Changes in TGF-β signalling in the aged fibroblast.

6. Develop an exhaustive combinatorial approach for model selection.

7. Develop tools that can be used in modelling biological systems.

8. Develop a computational means of handling time series data.

9. Experiment with clustering time series microarray data.

Chapter 2

A Systems Biology Approach

Reductionist science employs an approach where complex problems are broken down into more manageable parts, which are then analysed in isolation to determine their contribution to the system as a whole. This process has been highly successful in human history, particularly in physics and engineering. Newtons laws of gravity and planetary motion are examples of seminal discoveries that were grounded in reductionist science. Pioneers of modern biology have also based their research on a reductionist philosophy and seminal biological discoveries include determining the chemical structure and function of DNA (Watson & Crick 1953*b*,*a*) and sequencing of the human genome (Venter et al. 2001).

While reductionism is a highly successful and intuitive approach to science, its application to the study of biology has fundamental limitations. A full accounting of the parts of a biological system can no more describe how a biological system works than listing all the parts of an aeroplane can describe how the fully engineered product operates (Kitano 2002). Systems biologists are concerned with how biological function arises from the fully connected network of interacting components.

2.1 High-throughput experimentation

Advances in computer science and engineering have been instrumental in our ability to study biology from a systems perspective. Historically and in-line with a reductionist perspective, biological scientists have aimed to isolate one to a few objects for study. Examples of such experimental techniques include southern (Southern 1975), northern (Alwine et al. 1977) and western (Alwine et al. 1977) blots for detection of DNA, RNA and proteins in biological samples respectively. Based on the same principles as these experiments, engineers cooperating with biologists have built machines capable of enhancing the quantity and quality of measurements from the various levels of biological organisation.

DNA microarray technology, for example, is based on the concept of DNA hybridization and enables the parallel measurement of thousands of gene activities simultaneously, with only a small amount of sample (DeRisi et al. 1996). The complementary nature of DNA means that a prerequisite for DNA hybridisation is knowing the sequence of the gene being measured. Short probes that have a high affinity for a specific gene product are bound or synthesized onto a high density chip. For example, companies such as Affymetrix have built gene chips which are 'off-the-shelf' chips containing approximately 50,000 short DNA sequences that were synthesized on the surface of a chip using photolithography (Lockhart et al. 1996, Lockhart & Winzeler 2000). Transcripts in a sample are reverse transcribed to complementary DNA (cDNA) which are used as a template for *in vitro* transcription to produce complementary RNA (cRNA), which is then exposed to the microarray. The hybridization reaction is coupled to a fluorescence detection mechanism so that the amount of cRNA in a sample that is complementary to a probe is linearly proportional to the light emitted from that probe (Bolstad et al. 2005). The probe sequences are designed to encompass most of the known genes in a genome and therefore represents a global measurement of gene expression.

The prerequisite for using microarray technology is knowledge of the gene sequences under scrutiny. This is a limitation that precludes the discovery of unknown genes using

microarrays. RNA-seq on the other hand is an alternative method of transcriptome quantification that works by reverse transcribing the trascriptome and then sequencing the resulting cDNA using machines built (for example) by Illumina (Mardis 2008). The RNA-seq procedure directly and rapidly sequences all cDNA in a sample, maps them to the source genome and counts occurrences of transcripts from known or unknown genes and even non-coding RNAs. Sequencing technology is cheaper than it ever has been and superior to microarray in terms of resolution and sensitivity (Mortazavi et al. 2008, Metzker 2010). For this reason it is currently the best known method for transcriptional profiling a cell type under a given condition.

Both RNA-seq and microarray technologies have been coupled to chromatin immunoprecipitation (ChIP) experiments for identifying the DNA binding sites of a transcription factor. In ChIP, a transcription factor is cross-linked to the DNA that it is bound to, fragmented and extracted from the nucleus. Antibodies specific to that transcription factor are used to concentrate or *immunoprecipitate* DNA fragments. The chemical cross-links are then removed to release DNA fragments from proteins which are sequenced for quantification and identification (Visel et al. 2009).

While microarray, RNA-seq and ChIP-seq are all technologies focused on quantification of nucleic acids, proteomic methods aim to globally identify and quantify all the proteins present in a sample. Proteomics is inherently more complicated than transcriptomics because of the multidimensional nature of the proteome. Protein function depends not only on abundance but on post-translational modifications, cellular location, tissue of origin, solubility and interactions with other proteins, lipids or nucleic acids (Larance & Lamond 2015). Quantification of RNA relies on DNA complementarity but proteins do not have such a global complementary structures.

Instead, a large section of proteomic methods rely on the use of antibodies, by harnessing the antibody-antigen recognition property to probe for a protein. Analogous to the DNA microarray, probes (antibodies) are strategically placed and on a (typically) glass chip forming a protein dot matrix (Hu et al. 2011). After appropriate washing and blocking, the sample is incubated on the chip and the reaction of protein binding to

antibody is coupled to a fluorescence reporter reaction, which is then detected by a scanner (Chen et al. 2018).

Antibody based arrays are known as analytical protein arrays to distinguish them from functional protein arrays, where instead of antibodies, individually purified proteins are immobilised onto a chip to create the protein dot matrix. Analytical protein microarrays are typically used to detect proteins in a sample, for example in protein expression profiling or biomarker identification. In functional protein arrays on the other hand, the immobilized proteins are used to query biochemical properties of the immobilised proteins. For instance, functional protein arrays have been used to identify protein-protein, protein-lipid, protein-DNA, protein-RNA, protein-antibody and protein-small molecules interactions. Moreover, functional protein arrays can be used to identify substrates and enzymes in post-translational modification reactions (Hu et al. 2011, Sutandy et al. 2013).

Another, antibody based technology in proteomics is an extension of the western blotting procedure. Western blots are notorious for their vulnerability to sources of technical variation. Microfludic western blotting however is an machine which automates the western blotting procedure from start to finish and uses capillary tubes as a medium instead of gels. This method is more accurate because of greater precision is enabled by automating the procedure. However the limitation is that only a relatively small number of antibodies can be used at the same time, because antibodies with similar molecular weights will resolve at the same place in the capillaries and inhibit biological interpretation. Nevertheless, microfluidic western blotting is a promising means of quantifying proteins that solves many of the problems of the western blot (Nguyen et al. 2011).

2.2 Mass spectrometry proteomics

A major issue with antibody based proteomics is that antibodies can react with proteins other than the intended target. On a western blot, non-specific bindings become less of a problem because proteins are separated by size using gel electrophoresis. However, because this is not an option with protein microarrays, antibodies used on a protein microarray must be highly specific to their target. Antibodies are expensive and rarely specific enough to reliably detect the target with these methods, and this is a major barrier to their mainstream use.

Mass spectrometry based methods on the other hand do not rely on antibodies. In mass spectrometry, atoms are ionized in a vacuum by bombarding them with electrons, a process called *electrospray ionization*. Some of these collisions have enough energy to knock off some electrons and create a net positive charge. The ions are accelerated so that they have the same kinetic energy and then deflected using a magnet. The heavier the ions, the less they are deflected and the less positively charged the ions, the less they get deflected (Fenn et al. 1989, Walther & Mann 2010).

Proteins are too big and complicated to get a global mass/charge (m/z) measurement for the entire protein so they are first enzymatically digested into smaller peptides. The mixture of peptides produce a spectrum which is a plot of m/z ratios against the mass spectrometric signal (the ion current that is deflected). Peptide sequences can be ascertained by fragmenting the peptide along the carbon backbone using collision induced dissociation where particles are forced to collide with an inert gas such as helium which induces breaks. This process is called MS/MS or tandem MS and the resulting spectrum can be used to build up the sequence of the peptide. Because reading the MS-spectrum can be challenging, large databases exist that contain fragment spectra and peptide masses. These databases are routinely queried using the output from MS/MS to query fragments that already have an identity (Paizs & Suhai 2005, Sadygov et al. 2004, Walther & Mann 2010).

Tandem MS takes a complicated mix of proteins and produces an even more

complicated mix of peptide fragments. To reduce the complexity, proteins in a sample can be separated prior to the MS procedure. Multiple methods exist for doing so including: SDS gel electrophoresis followed by in-gel digestion (Shevchenko et al. 1996); digestion in solution (Link et al. 1999, Washburn et al. 2001) or filter-aided sample preparation which is a combination of both (Wisniewski et al. 2009). Moreover, MS has been coupled to high performance liquid chromatography (LC-MS/MS) for better separation of protein mixtures prior to MS (Petrovic & Barcel 2013).

High throughput experimental methods are at the heart of the top-down sub-paradigm of systems biology which aims to connect biological information in functional interaction networks. Where top-down approaches attempt to measure everything possible, they do not target anything specific. Their goal is to characterise a biological state globally and unbiasedly, and in doing so generate new testable hypotheses.

In the following subsections, several mathematical and bioinformatic techniques are described that are commonly used with high throughput data.

2.2.1 Principal Component Analysis

High throughput experiments are vulnerable to misinterpretation because of their high dimensionality and potential for systematic bias resulting from the number of steps in data acquisition, processing and interpretation. Quality control therefore aims to assess where data is reliable and which aspects need to be discarded. Principal component analysis (PCA) is a dimensionality reduction technique which aims to project p dimensional data onto lower dimensions q without losing the important variance in the data. PCA has a variety of applications, but in analysis of biological data, it usually serves to reduce data with many variables (i.e. repeats, time, cell type, strain) to lower dimensions ($q = 2$ or $q = 3$) that can be easily visualised. In the biological setting, p is usually the number of samples and each point on a PCA plot represents each one of the p samples. Colouring the points by various experimental factors enables assessment of the quality of data, because variables that should be similar (i.e. samples taken from a

particular condition) should cluster together while variables that should not be similar (i.e. replicates) should have no discernible pattern on the PCA plot.

PCA is a linear transformation of an original data set $X \in \mathbb{R}^{n \times p}$ into new uncorrelated data called principal components PC. The first principal component PC_1 is defined as the weighted sum of the variables with coefficients $a_{ij} \in A$ (Equation 2.1).

$$PC_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1j}X_j + a_{1p}X_p \tag{2.1}$$

The coefficients a_{ij} are chosen to maximize variance over all such linear combinations. Since this criteria can be satisfied by making the coefficients large, a restriction is placed on the coefficients so that the sum of each row is 1. The second principal component PC_2 is the linear combination of original variables which accounts for the maximal amount of remaining variance, with the restriction that PC_2 is uncorrelated (orthogonal) with PC_1. This process continues for all p variables.

To compute principal components, the original data are centred and the covariance or correlation matrix is computed. Since PCA is sensitive to scale data on different scales, the correlation matrix is used when the variables are very different (for example height in centimetres and weight in grams). The eigenvectors of the covariance or correlation matrix are the coefficients A and the corresponding eigenvalues represent the amount of total variance explained by the principal component. The eigenvectors are ordered by eigenvalues so that the eigenvector with the largest eigenvalue are the coefficients for PC_1 (Holland 2008).

Differential Expression with LIMMA

LIMMA is an R package designed to facilitate the statistical handling of complex experiments. It is not the only existing package for determining differential expression but because of its flexibility and stability, it is widely used. LIMMA fits a linear model to expression data for each gene and uses a concept called Empirical Bayes to exploit the parallel nature of microarray data and 'borrow information' across arrays to more

accurately identify genes that are differentially expressed genes. A detailed description of the LIMMA method and usage can be found in (Smyth 2004, 2005, Ritchie et al. 2015).

Part of the power of LIMMA is its flexibility to be used on any experimental design, including time series experiments. For time series experiments, there are two avenues of analysis that can be taken, depending on the nature of the experiment. Firstly, if the number of time points is small, it is possible to use LIMMA in the same way as for a non-time series experiment - by comparing individual time points to a control or to each other using the concept of contrasts. Contrasts are user defined comparisons of interest between experimental groups. Using contrasts, each gene is compared across the experimental conditions defined by the contrast using a moderated t-test, which is the same as a regular t-test in that it tests for equality between two means, but is moderated by using data from other arrays within the experiment. This moderation is what is referred to as 'borrowing information' from the ensemble of available genes. Multiple testing correction is then applied to p-value outputs. LIMMA also computes a moderated F-statistic which is analogous to the moderated t-statistic but for an ANOVA test. The F-statistic provides a measure of confidence for a gene being differentially expressed in any of the defined contrasts.

The second avenue for analysis of time series using LIMMA is better applied when more time points are available. The method relies on first fitting cubic splines with a predefined number of knots (i.e. fixed points). This number corresponds to the degrees of freedom *df* parameter. Then a moderated F-statistic is calculated between the time series and a control sample(s) (either a single sample or a time matched set of controls) to test whether there is a difference between groups.

2.2.2 Gene Ontology and Enrichment

The output from LIMMA is a list of genes that are differentially expressed between experimental conditions. Ontologies are formal representations of knowledge and a useful way of assigning functional behaviour to gene lists. In biology the most commonly

known ontology is the Gene Ontology (Ashburner et al. 2000) which is presented as a directed acyclic graph (DAG), where GO terms are nodes and parent-child relationship exist between nodes. Ontology trees are general at the top and become increasingly more specific with progression to leaf nodes at the bottom. Any node can have many parents and child nodes are essentially a subset of the parent concepts so that if a child node is annotated with a gene, the parents are implicitly also associated with that node.

In GO analysis, individual genes are assigned to GO terms. Such an assignment is known as an annotation and is performed either by a curator who assesses experimental evidence or computationally predicted. Genes are associated with as many GO terms as necessary but those which are assigned computationally are not as strongly assigned as those that are manually curated (Rhee et al. 2008).

Ontologies such as the GO are useful because they enable enrichment analyses where a reference gene list is queried with a list obtained from differential expression analysis to find genes that are under or over represented. This is done by comparing the number of times a GO term is observed in the reference set compared to what was observed experimentally (Rhee et al. 2008).

Many tools exist for enrichment analysis on multiple platforms. For example, in R, there is 'topGO' (Alexa & Rahnenfuhrer 2010) and 'GOstats' (Falcon & Gentleman 2006) while in Python there is 'goatools' (Tang et al. 2015, Klopfenstein et al. 2018). Moreover a web based service called DAVID (Huang da, Sherman & Lempicki 2009*a*,*b*, Huang da, Sherman, Zheng, Yang, Imamichi, Stephens & Lempicki 2009) is a useful tool because it integrates many ontologies, such as KEGG, Reactome or the gene ontology together in a single place.

2.3 Modelling in Systems Biology

While top-down systems biology concerns hypothesis generation, bottom-up systems biology is hypothesis-driven. Detailed (usually dynamic) information about the

components of a system are quantified and then pieced together in quantitative models to reproduce the behaviour of the system.

The types of mechanistic model that have been used in systems biology are broad. Ordinary differential equation (ODE) are familiar to system biologists and allow a description of system dynamics based on rate equations, usually following mass action kinetics or approximations that are designed reduce a system of equations, such as Michaelis Menten or Hill kinetics. ODE's have been used in physics and engineering for many years and methods for ODE analysis are well developed. Delayed differential equations (DDE) differ from ODEs in that they provide a mechanism for explicitly including a time delay so that the model can refer to a part of the model in the past. Partial differential equations (PDEs) on the other hand differ from ODEs in that they describe system dynamics in two dimensions, often space and time, rather than just in time.

These are all deterministic methods that describe the limiting behaviour of a system. Biological systems however are inherently stochastic and a lot of modelling effort in systems biology focuses on capturing this variation (Wilkinson 2006). Stochastic differential equations are a probabilistic representation of the underlying stochastic processes and incorporates a mechanism for modelling biological variability. Stochastic differential equations are similar to ODEs but with an added noise term (Gillespie 2000). Other stochastic simulation methods include the stochastic simulation algorithm (SSA) (Gillespie 1977) and the computationally simpler (but also exact) direct method (McCollum et al. 2006). These are numerical methods where each reaction has a stochastic rate constant which encodes the probability of the reaction occurring in a defined time interval.

These exact methods are computationally intensive, particularly when simulating distributions from medium to large models. Alternative approximate approaches have been developed to better handle the computational complexity of the exact solutions. For example, tau-leaping trades some accuracy for simulation speed by skipping the simulation of some time steps (Cao et al. 2006).

The biological and *in silico* consequences of stochasticity are more severe when the number of molecules in the system is low and in such situations the deterministic approximation is not suitable. On the other hand, simulating large molecule numbers using the SSA converges to the deterministic solution when copy numbers are large. Hybrid simulation methods can take advantage of this fact and use a combination of stochastic and deterministic integrators for enhanced simulation speed (Kiehl et al. 2004).

The aforementioned modelling approaches involve manually coding the system equations one by one. As an alternative, rule based modelling, such as BioNetGen (Blinov et al. 2004), enables the encoding of specific binding sites on proteins and how they interact with binding sites on other proteins. Then any two proteins that have these domains are able to interact. This approach is effective for building large models but despite tools such as BioNetFit (Thomas et al. 2016) that are capable of inferring the parameters of such models, parameter inference for rule based modelling can be challenging.

Partial differential equations represent one means of simulating biological processes in both time and space. Another way to simulate space is to make use of a computational lattice. Cellular automata are lattice based models where every element of the lattice represents a discrete state that change with time depending on model specific rules (Alber et al. 2003). Alternatively agent based models enable the abstraction of principles, concepts or ideas into autonomous computational objects called agents. The interacting agents are designed to reproduced some observed physical phenomenon. A huge variety of software exist for building agent based models (Abar et al. 2017). Some, such as NetLogo (Wilensky 2008), are frameworks designed to be user friendly while others such as BioCellion (Kang et al. 2014) and FLAME (Richmond et al. 2010) have sacrificed usability for sophisticated computational architectures that support scalability through parallel programming. Ecell (Tomita et al. 1999, Takahashi et al. 2003, 2004) on the other hand, strikes a balance between usability and speed by making effective use of Python for a user interface and C++ for computer efficiency. Ecell enables simulation at multiple scales by enabling the same front end model to be simulated using ODE, SSA, spatial Gillespie or as individual particles. In spatial Gillespie, a 3D cube is split

into sub-volumes, each of which is simulated using a separate realisation of the SSA and a set of additional rules which defines diffusion between sub-volumes. In Ecell particle simulations, a 3D grid of dodecahedrons represents space and each individual molecule is an agent capable of diffusion and reacting with other molecules, depending on the model topology.

2.4 ODE modelling in systems biology

Thus far we have discussed a range of high-throughput experiments and modelling frameworks that are available to the systems biologist. ODEs are by far the most commonly used modelling framework in systems biology as they are well-suited to predicting system dynamics and many software packages exist for their simulation and analysis. Moreover, many of the analyses that can be conducted on ODE models have already been well established by physicists and engineers. The purpose of this section is to formally introduce the ODE framework and discuss some of the ways it is possible to analyse them. Each of the topics covered below are large sub-fields of study in their own right and as such, the purpose of this section is to give a broad-brush flavour of the types of analysis that are applicable to ODE models, rather than a comprehensive review. Where possible, the reader is referred to appropriate literature for more details.

2.4.1 ODEs and the law of mass action

ODEs in systems biology have the general form shown in Equation 2.2

$$\dot{\vec{x}} = f_i(\vec{x}_{(t)}, \vec{u}_{(t)}, \vec{\theta}), \qquad i \in \{1, \cdots, |\vec{x}|\} \vec{x}_{(t_0)} \quad = \vec{x}_0 \tag{2.2}$$

where $\vec{x} \in \mathbb{R}^{|\vec{x}|}$, $\vec{u} \in \mathbb{R}^{|\vec{u}|}$ and $\vec{\theta} \in \mathbb{R}^{|\vec{\theta}|}$. $\dot{\vec{x}}$ states that the rate of change of model state variables $\vec{x}$ (i.e. components of biochemical networks) equals the solution of a function $f_i(\cdot)$ that depends on the state variables with time $\vec{x}_{(t)}$, a time dependent set of inputs

or perturbations $u(\vec{t})$ and a kinetic parameter set $\vec{\theta}$ (Tonsing et al. 2014).

The rate of change of a biochemical component x_i can be described by the sum of the individual rates of the reactions that x_i participates in. Often these rates are described by the law of mass action which states that the velocity v of a chemical reaction r at any given time proceeds proportionally to the product of the substrate concentrations $\vec{s_i} \in \vec{x_i}$ (Equation 2.3).

$$v \propto \prod_i s_i \tag{2.3}$$

Consequently, the concentrations can be multiplied by a constant θ to produce an equivalence relation that can be used to calculate the rate of a reaction at a particular time point (Equation 2.4).

$$v = \theta \prod_i s_i \tag{2.4}$$

When a mass action reaction has 0, 1 or 2 substrates, the mass action reaction is known as a zero, first or second order reaction respectively. This rule generalizes to any number of substrates, though its uncommon to use $> 3^{\text{rd}}$ order mass action reaction in an ODE system describing biochemical networks.

Mass action is the simplest rate law. However multiple mass action reactions can be approximated by making assumptions based on the behaviour of the chemical components under study. Michaelis-Menten kinetics for example approximates the reversible binding and unbinding of a substrate to an enzyme, its conversion to a product and release of product from the enzyme. Michaelis-Menten kinetics are based on two assumptions: firstly that the substrate concentrations $[s_i]$ are much bigger than enzyme concentrations $[E]$ and secondly that the binding rates occur much quicker than product conversion. Under these assumptions, the Michaelis-Menten rate law approximates the general behaviour of these four reactions. Michaelis-Menten kinetics follows a hyperbolic curve (with substrate concentration plotted against enzyme velocity) that limits the reaction rate v as substrate concentration increases. Biochemically, this upper limit represents the maximal velocity of the enzyme.

Another simplification of mass action kinetics that is commonly used in modelling

biochemical systems is the Hill equation. Hill kinetics were originally designed to study cooperativity, a phenomenon where an protein binds multiple substrate and the affinity for subsequent substrate increases proportionally to the number of substrates already bound (Hill 1910). A typical example to explain cooperativity is haemoglobin that can bind four oxygen molecules. Once haemoglobin binds to a single oxygen molecule, its affinity for the second increases (Perutz 1989). Hill rate laws are similar to Michaelis-Menten rate laws but they have an additional exponent term h that controls the sensitivity of the transition between enzyme velocities. Specifically, a higher hill coefficient h in a plot of substrate concentration against enzyme velocity makes the transition from zero to maximal enzyme velocity very quick. Conversely, when $h = 1$ the Hill rate law converges to Michaelis-Menten kinetics. Hill kinetics are often used for ligand-receptor binding and for rates of transcription reactions. A detailed description of both Michaelis-Menten and Hill rate laws along with many other commonly used rate laws can be found in (Sauro 2011).

Systems of equations (such as in Equation 2.2) form a network where the nodes represent the state variables $\vec{x}$ and a directed edge is drawn from x_j to x_i if the rate of reaction f_i depends on x_j. These systems can be solved analytically in rare cases but most often they are numerically integrated using computational algorithms.

2.4.2 Sensitivity Analysis

The solution to Equation 2.2 is highly dependent on the parameter vector $\vec{\theta}$. One application of sensitivity analysis is to quantify how uncertainties in the parameter values affect model predictions (Arriola & Hyman 2009). A model prediction that changes very little with model parameters is more robust than one that changes very rapidly. The essence of a sensitivity analysis is to measure the change in state variables x_i when the input θ_j is changed by a small amount. Other uses for sensitivity analysis include optimal experimental design (Tonsing et al. 2014) or model reduction (Liu et al. 2004).

There is a large and diverse literature on sensitivity analysis which broadly divides the methods into local and global sensitivity analysis. Local sensitivity analysis enables the calculation of sensitivities at a particular point in parameter space and is computed by perturbing one parameter at a time. Global sensitivity analysis on the other hand quantifies the effect of changing all parameters simultaneously on model output.

The simplest method of calculating sensitivities is the finite difference approximation (Equation 2.5).

$$\frac{\partial x_i(t)}{\partial \theta_j} = \lim_{\Delta\theta_j \to 0} \frac{x_i(\theta_j + \Delta\theta_j) - x_i(\theta_j)}{\Delta\theta_j} \tag{2.5}$$

where:

$$x_i(\theta_j) = \text{model output with parameters } \theta_j$$

$$x_i(\theta_j + \Delta\theta_j) = \text{model output with perturbed parameter } \Delta\theta_j$$

While finite differences are the most commonly used method of sensitivity analysis, but they are computationally expensive for large models and depend on the choice of a perturbation factor $\Delta\theta$ (De Pauw & Vanrolleghem 2006, Zi et al. 2011). Finite approximations therefore require some trial and error to find the best value for $\Delta\theta$. Generally if the sensitivities change very little for sequential choices of $\Delta\theta$, the choice for $\Delta\theta$ is acceptable (Zi 2011).

An alternative to finite difference calculations is to solve the sensitivity differential equations directly (Rabitz et al. 1983). These also measure the difference in model output with respect to a change in parameters (one at a time) but the problem is solved by numerical integration of the ODE system and can be efficiently computed with tools such as CVODES in the SUNDIALS library (Serban & Hindmarsh 2003). The sensitivity equations are derived by differentiating Equation 2.2 with respect to parameter θ_j and are described by Equation 2.6.

$$\frac{\partial}{\partial t}\frac{dx_i}{d\theta_j} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial \theta_j} + \frac{\partial f}{\partial \theta_j} \tag{2.6}$$

where:

$$\frac{\partial f}{dx} = J = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_i} & \cdots & \frac{\partial f_1}{\partial x_{|\vec{x}|}} \\ \frac{\partial f_k}{\partial x_1} & \frac{\partial f_k}{\partial x_i} & \cdots & \frac{\partial f_k}{\partial x_{|\vec{x}|}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_{|\vec{f}|}}{\partial x_1} & \frac{\partial f_{|\vec{f}|}}{\partial x_i} & \cdots & \frac{\partial f_{|\vec{f}|}}{\partial x_{|\vec{x}|}} \end{bmatrix} \qquad \begin{array}{r} \text{for } i \in \vec{x}, \\ k \in \vec{f} \end{array}$$

is the Jacobian of the system,

$$\frac{\partial x}{\partial \theta_j} = \begin{bmatrix} \frac{\partial x_1}{\partial \theta_j} \\ \frac{\partial x_2}{\partial \theta_j} \\ \vdots \\ \frac{\partial x_{|\vec{x}|}}{\partial \theta_j} \end{bmatrix}$$

is a vector of sensitivities of state variable x_i to parameter θ_j and

$$\frac{\partial f}{\partial \theta_j} = \begin{bmatrix} \frac{\partial f_1}{\partial \theta_j} \\ \frac{\partial f_2}{\partial \theta_j} \\ \vdots \\ \frac{\partial f_{|\vec{f}|}}{\partial \theta_j} \end{bmatrix}$$

are equation derivatives f_k with respect to parameter θ_j.

The direct method avoids having to choose a variation size parameter $\Delta\theta$ and enables calculation of the sensitivity of all model variables to a particular parameter at once. The disadvantage is that the Jacobian matrix J is computationally expensive to calculate (Rabitz et al. 1983).

For non-linear ODE systems, local sensitivity analysis only applies to the parameter set for which it was computed. Biological systems however have a broad range of parameters and concentrations and so it is often desirable to characterise the changes in model output to extensive changes in parameter values. This is the concern of global sensitivity analysis which relies on sampling parameter space, computing local sensitivities and aggregating the results. The sampling conducted in global sensitivity analysis can be performed randomly (Monte Carlo sampling) or more intelligently with methods such as Latin hypercube sampling. A more detailed description of global sensitivity analysis in systems biology can be found in (Zi et al. 2011, Rabitz et al. 1983, Kent et al. 2013).

2.4.3 Model Calibration

Before an ODE model can make meaningful predictions, the parameters must be calibrated to experimental data. The parameters and initial values of an ODE model are usually unknown. Model calibration refers to the process of fitting a model to experimental data or the identification of model parameters given a data set. Model observables $\vec{y}$ are the subset of model variables that have been measured, usually at several time points $t \in \vec{T}$ in one or more experimental conditions $c \in \vec{C}$. These can also be any combination of model variables such as the sum or average of a set of model components. Model outputs $\vec{y}$ represent simulated observables. The experimental data is mapped to model outputs $\vec{y}$ by an observation function $\vec{g}$ that depends on model variables $\vec{x}$ and parameters $\vec{\theta}$ with measurement noise $\vec{\epsilon}(t)$ Equation 2.7 (Tonsing et al. 2014).

$$\vec{y} = g(\vec{x}(t), \vec{\theta}) + \vec{\epsilon}(t) \tag{2.7}$$

Parameter inference involves using optimization algorithms to optimize an objective function that measures the discrepancy between experimental data $\vec{y}$ and model simulations at all experimental conditions c and for all observables k and time points t. This objective function is known as the likelihood function which for normally distributed errors takes the form in Equation 2.8.

$$\ell = L(\vec{y}|\vec{\theta}, \dot{\vec{x}}) = \prod_{c=1}^{|\vec{C}|} \prod_{k=1}^{|\vec{y}|} \prod_{t=1}^{|\vec{t}|} \frac{1}{2\sqrt{\pi\sigma_{ckt}}} exp\Big(- \frac{[y_{ckt} - x_{ckt}(\vec{\theta}, t)]}{2\sigma_{ckt}} \Big)^2 \tag{2.8}$$

These products become easier to handle when the logarithm of the likelihood function (Equation 2.8) is used because they become sums rather than products (Equation 2.9).

$$log\ell = \sum_{c=1}^{|\vec{C}|} \sum_{k=1}^{|\vec{y}|} \sum_{t=1}^{|\vec{t}|} log_e\big[\frac{1}{2\sqrt{\pi\sigma_{ckt}}}\big] \Big(- \frac{[y_{ckt} - x_{ckt}(\vec{\theta}, t)]}{2\sigma_{ckt}} \Big)^2 \tag{2.9}$$

Multiplying both sides by -2 (Equation 2.10)

$$-2log\ell = \sum_{c=1}^{|\vec{C}|} \sum_{k=1}^{|\vec{y}|} \sum_{t=1}^{|\vec{t}|} (-2log_e\big[\frac{1}{2\sqrt{\pi\sigma_{ckt}}}\big]) \cdot -2 \Big(- \frac{[y_{ckt} - x_{ckt}(\vec{\theta}, t)]}{2\sigma_{ckt}} \Big)^2 \tag{2.10}$$

and making the assumption that the variance is known, that is, given by empirical data replicates, the first term of the sum becomes a constant and Equation 2.9 becomes Equation 2.10.

$$-2log\ell = \sum_{c=1}^{|\vec{C}|}\sum_{k=1}^{|\vec{y}|}\sum_{t=1}^{|\vec{t}|}\left(-\frac{[y_{ckt} - x_{ckt}(\vec{\theta}, t)]}{\sigma_{ckt}}\right)^2 = RSS(\vec{y}|\vec{\theta}, \dot{\vec{x}}) = RSS \tag{2.11}$$

Equation 2.11 is the residual sum of squares (RSS) objective function weighted by measurement errors. The likelihood function is equivalent to the residual sum of squares objective function under the assumption of additive Gaussian errors and known variance (Kreutz et al. 2012, Raue et al. 2009).

2.4.4 Model Validation

Following calibration, models should be validated to verify their predictive power. This is typically done by comparing model predictions with experimental data that was not used for model calibration. Whilst the term 'validation' is used, it is also misleading because validation implies the model is either correct or incorrect. However, models can only be falsified, because a model that behaves as expected in some ways would likely behave unexpectedly when tested in new ways. Model validation does not make the model hypothesis true but means that the model generates good, testable hypotheses. The purpose of validation is to decide whether the model is acceptable for its intended use (Rykiel 1996, Hasdemir et al. 2015).

A common method of model validation is known as the 'hold out' strategy, where a dataset is selected to be reserved for validation and is not used in calibration. A problem with the hold out strategy is that it is not obvious how to partition the data into the calibration (or training) and validation (or testing) datasets. As an alternative, cross validation is a statistical technique traditionally used for model selection that both enhances the rigour with which the model is validated and bypasses the partitioning problem. At its essence, cross validation is a resampling technique where calibration and validation datasets are randomly selected. Cross validation can be combined with the

'hold out' strategy, where the data that is set aside for validation is a single variable under a single condition. Resampling is conducted so that all combinations of test-train splits have an equal probability of being selected. An alternative is the stratified-random cross validation which is similar, but the test-train datasets are selected randomly, but with restrictions. These restrictions include the dimensions of the test and training sets or equal sampling from each experimental condition. Regardless of the methods chosen, the result is a set of parameter estimations each with a validation data set (Hasdemir et al. 2015, Klipp et al. 2009).

Validation errors can be evaluated using an analogue of the RSS to measure the distance of the model predictions to the validation data. These distances are summed up for each model and provide a metric for model quality based on how well they predict the validation data (Klipp et al. 2009).

2.4.5 Model Selection

Model selection is the process of choosing a 'best' model from a set of models, all of which represent an alternative model topology. This decision can be made using a variety of statistical tools, primarily from information theory (Burnham & Anderson 2003). A good model adheres to the principle of parsimony, whereby a model describes the observations well with as little complexity as possible. Generally, model selection criteria are based on the idea that a set of data has a certain fixed amount of information regarding the underlying process that they came from. Since modelling this data is an abstraction of reality, modelling the process causes some information in the data to be lost.

The goal in model selection is to find a model that has a perfect 1:1 mapping of the information contained in the data to a model of the information contained in the data (Burnham & Anderson 2003). This is an idealized and unrealistic goal since models are only abstractions. Model selection is about finding the model that looses as little information as possible in the transition from data to model. While the truth f is not

modellable, a model g of f should be a good approximation of f such that valid inferences can be made regarding f. While the notation f is used to represent this truth, there is no 'true' model in systems biology and f should be considered as simply the process that generated the data.

The Kullback-Leibler (KL) information (Kullback & Leibler 1951) (also known as the KL distance or divergence) quantifies the amount of information lost when approximating the reality f with a model g and the modeller should seek to minimize this distance (Burnham & Anderson 2003). The KL information is a fundamental quantity in information theory that is related to Shannon's entropy (Shannon 1948) and is the negative of Boltzmann's entropy (Burnham & Anderson 2003). The KL distance is the unidirectional distance from g to f and is described in Equation 2.12.

$$I(f,g) = \int f(x) log_e\left(\frac{f(x)}{g(x|\theta)}\right) dx \tag{2.12}$$

where f is the true probability distribution of the process being modelled, x is experimental data and $g(x|\theta)$ is a model of $f(x)$ parametrized with θ. The KL distance is always positive except when $f(x) = g(x|\theta)$, in which case $I(f,g) = 0$. Moreover, the KL distance is not a true 'distance' because $I(f,g) \neq I(g,f)$. An obvious shortcoming of the KL distance (Equation 2.12) is that both f and g must be known for its computation, which they are not. However, it is also possible to express the KL distance in relative form by rewriting Equation 2.12 as a difference of expectations with respect to the true distribution f. Algebraically, Equation 2.13 is derived by from Equation 2.12 using: 1) $log(\frac{a}{b}) = log(a) - log(b)$ and 2) $E(x) = \int_{-\inf}^{\inf} x f(x) dx$.

$$\begin{aligned} I(f,g) &= \int f(x) log_e[f(x)] dx - \int f(x) log_e[g(x|\theta)] dx \\ &= E_f[log_e(f(x))] - E_f[log_e(g(x|\theta))] \end{aligned} \tag{2.13}$$

The first expectation is a constant $E_f[log_e(f(x))] = C$ that depends only on the unknown true distribution f. Therefore rewriting Equation 2.12 as Equation 2.13 enables the computation of a relative distance between f and g, up to a constant defined

by C (Equation 2.14).

$$\begin{aligned} I(f,g) &= C - E_f[log_e(g(x|\theta))] \\ I(f,g) - C &= -E_f[log_e(g(x|\theta))] \end{aligned} \tag{2.14}$$

$I(f,g) - C$ is the directed distance from g to f and is characterised by $E_f[log_e(g(x|\theta))]$. If two models g_1 and g_2 are under examination and $I(f,g_1) < I(f,g_2)$, then $I(f,g_1) - C < I(f,g_2) - C$ and $-E_f[log_e(g_1(x|\theta))] < -E_f[log_e(g_2(x|\theta))]$. Therefore, if this quantity can be computed or estimated, one has a sound motive for selecting one model over another.

The Akaike information criteria (AIC) is based on the same principles as the KL distance and in fact, regards the KL distance as the fundamental quantity of interest concerning model selection. Akaike provided a means of estimating the KL by formalizing the relationship between Boltzmann entropy, the KL distance and maximum likelihood statistics. In other words, the AIC enables the estimation of the KL distance using the likelihood function at its maximum and therefore enables simultaneous model calibration and selection (Akaike 1973, 1974, 1985, 1994, Burnham & Anderson 2003). The AIC is the estimated expected relative KL information (Equation 2.15),

$$AIC = log\ell - K = constant - \hat{E}_{\hat{\theta}}[I(f,\hat{g})] \tag{2.15}$$

which was multiplied by -2 (Akaike 1973) so that the AIC is Equation 2.16

$$AIC = -2log\ell + 2K \tag{2.16}$$

Where K represents the number of free parameters in the model and enforces the principle of parsimony.

The AIC is a good approximation of the KL distance when enough data is available relative to the complexity of the model. This often is not the case and the AIC has been refined to encompass the case where less data is available (Hurvich & Tsai 1989, 1990, 1991, Hurvich & TSAI 1995*b*, Hurvich & Tsai 1995*a*, Burnham & Anderson 2003). The

corrected AIC, AIC_c is Equation 2.17

$$AIC_c = AIC + \frac{2K(K+1)}{(n-K-1)} \tag{2.17}$$

where n is the sample size. Equation 2.17 converges to the AIC for large n and therefore it is generally used instead of the AIC.

The RSS (Equation 2.11) is a special case of likelihood statistics where the errors in the experimental data are assumed to be Gaussian with constant variance. When these assumptions are valid Equation 2.18 can be used to compute the AIC.

$$AIC = n \cdot log\left(\frac{\sum \hat{\epsilon}_i^2}{n}\right) + 2K \tag{2.18}$$

where $\hat{\epsilon}_i^2$ represents the residuals or the differences between model and experimental values.

The AIC is not the only model selection criteria that exists. The Takeuchi Information Criteria (TIC), Bayesian Information Criteria (BIC), Quasi AIC (QAIC) and corrected QAIC (QAICc) are all alternative model selection criteria that are derived along with the AIC and AICc in Burnham & Anderson 2003.

2.4.6 Identifiability and Observability Analysis

A particular problem with model calibration is that of non-identifiability, where an optimization problem cannot uniquely identify parameter values because an infinite number of parameter sets all have the same objective function value. Identifiability issues are common in systems biology because of large network sizes relative to the availability of experimental data, which nearly always contains considerable uncertainty. Moreover, usually there is also a high degree of uncertainty in the model topology.

An identifiability analysis is concerned with determining whether it is possible to uniquely identify model parameters given the model and experimental data. A structural non-identifiability is a property of network topology and addresses the

question of whether an ideal noise-free dataset would be able to identify the parameters *a priori*. Structural identifiability is therefore a necessary condition for fully determining the model parameters (DiStefano III 2015). An example of a structural non-identifiability is a simple reversible state transition reaction that converges to equilibrium and only informed by a single observable. A change in the forwards rate constant can be mitigated by an equivalent change in the backwards parameter without affecting the objective function value. Consequently, an infinite number of potential solutions exist. A structural identifiability analysis requires working with the model equations directly, which is difficult for even relatively small models. Consequently a variety of methods use symbolic computation to analytically determine structural identifiability (Chis et al. 2011).

Even if a problem is fully structurally identifiable it may still be 'numerically' or 'practically' non-identifiable. A practical non-identifiability is assessed *a posterori* and arises from an insufficient quantity or quality of experimental data (DiStefano III 2015). The following sections describe two methods of assessing practical non-identifiability, the Fisher Information Matrix and profile likelihood methods.

2.4.7 Fisher information matrix based identifiability

The fisher information matrix (FIM) is a symmetric square matrix that measures the maximum amount of information a set of observables contain about a unknown parameter set (Erguler & Stumpf 2011). The rank of the FIM indicates the number of identifiable parameters. If the matrix is not invertible (i.e. singular) then the problem is non-identifiable and more data or topological changes should be considered.

The *observed* FIM is the FIM at a best parameter set and is the second derivative matrix of the objective function (Equation 2.8) with respect to parameters θ_i and θ_j Equation 2.19.

$$FIM(\hat{\theta}) = \frac{\partial^2 \ell}{\partial \hat{\theta_i} \partial \hat{\theta_j}}, \qquad i, j \in \{1, \cdots, |\vec{\theta}|\} \tag{2.19}$$

Practically, the FIM is often computed using the sensitivity of model variables to small

parameter perturbations (Equation 2.20).

$$S_{ij} = \begin{bmatrix} \frac{\partial x_1}{\partial \theta_1} + .. + \frac{\partial x_{|\vec{x}|}}{\partial \theta_1} \\ \vdots \\ \frac{\partial x_1}{\partial \theta_{|\vec{\theta}|}} + .. + \frac{\partial x_{|\vec{x}|}}{\partial \theta_{|\vec{\theta}|}} \end{bmatrix} \tag{2.20}$$

The observed FIM is then the model sensitivities multiplied by its own transpose (Equation 2.21).

$$FIM(\hat{\theta}) = S^T S \tag{2.21}$$

The correlation matrix between parameters C is the inverse of the observed FIM and this relation can be used to calculate approximate confidence intervals for parameter estimates (Equation 2.22) (Vanlier et al. 2013, Raue et al. 2009, Pawitan 2001, Tonsing et al. 2014).

$$\sigma_i^{\pm} = \hat{\theta}_i \pm \sqrt{\chi^2(\alpha, df) \cdot C_{ii}} \tag{2.22}$$

where $\chi^2(\alpha, df)$ is the χ^2 distribution with df degree of freedom which is the number of estimated parameters.

Profile likelihoods

FIM based confidence intervals are a good approximation of uncertainty when a large body of data are available that have low measurement error, however both of these conditions are often broken in systems biology. Profile likelihoods are an alternative method of identifiability analysis that relies on optimization. More specifically, a profile likelihood is a parameter scan of parameter estimations. Each parameter $\theta_i \in \vec{\theta}$ is fixed in turn to the best estimated value and systematically varied over the course of the scan. At each point of the scan all parameters except the parameter of interest are reoptimized. Formally, the profile likelihood for a single parameter is described in Equation 2.23.

$$\chi^2_{PL(\theta_i)} = \min_{\theta_j \neq \theta_i} [\ell(\theta_i)] \tag{2.23}$$

$\chi^2_{PL(\theta_i)}$ is a vector of values representing the path of the objective function as the parameter of interest θ_i is varied and the $\theta_i \neq \theta_j$ are reoptimized (Raue et al. 2009, Vanlier et al. 2013). At each point of the scan, the reoptimized model ($M[\theta_{reopt}]$) is a nested model of the original model ($M[\theta_{opt}]$) that contains the best estimated parameter values because they can be transformed into one another using linear constraints. The likelihood ratio of two nested models is approximately χ^2_{df} distributed with $df = |\vec{\theta}|$ and so comparing this ratio to the χ^2_{df} provides a confidence level for the profile likelihood (Vanlier et al. 2013) (Equation 2.24).

$$-2log\left(\frac{M[\theta_{reopt}]}{M[\theta_{opt}]}\right) \leq \chi^2_{1-\alpha} \tag{2.24}$$

where α denotes the desired significance level (Vanlier et al. 2013, Raue et al. 2009).

In addition to parameter profile likelihood, similar concepts have been applied to predictions and validations. In the same way that a parameter profile likelihood is a parameter scan of parameter estimations with the parameter of interest fixed over the duration of the scan, prediction and validation profile likelihoods follow the same concept, but the optimization is constrained by predictions or validations respectively, instead of the parameters. Analogous to a parameter profile likelihood, prediction and validation profile likelihoods are used to ascertain confidence intervals for model predictions and validations (Kreutz et al. 2012).

2.4.8 Sloppiness

The term sloppiness is used to describe what some believe to be a universal property of moderate to large ODE models that describe physical and biochemical phenomena (Gutenkunst 2007, Brown & Sethna 2003). This property arises from the geometry of the parameter space around an optimum point with respect to some experimental data. The Hessian matrix (Equation 2.25) is the second derivative matrix of the likelihood

function (Equation 2.8) with respect to the model parameters.

$$H_{\hat{\theta}ij} = \frac{\partial \ell}{\partial \theta_i \partial \theta_j} \tag{2.25}$$

When the Hessian matrix is evaluated at a best estimated parameter set $\hat{\theta}$, it is equivalent to the observed FIM (Equation 2.21), which is related to the sensitivities of model variables (under assumptions of normality) to parameter perturbations (Equation 2.20). The sensitivities are therefore used in the calculation of observed FIM and to determine the asymptotic confidence intervals (Equation 2.22) for a maximum likelihood estimation problem. These confidence intervals can be plotted as contours $RSS(\hat{\vec{\theta}}) = const$ around a best parameter set. Geometrically, the confidence intervals are ellipsoids, the borders of which are determined by comparison with the α quantile of the χ^2 distribution with $df = |\vec{\theta}|$ degrees of freedom (Tonsing et al. 2014).

The covariance matrix can also be analysed by eigenvalue decomposition: the width of the ith principal component of the multidimensional confidence ellipsoid is proportional to the square root of the eigenvalue λ_i^C. The corresponding eigenvector points in the direction of this principal axis. Principal components with small eigenvalues have large ellipsoids around the best parameter set and are called sloppy, while those with large eigenvalues have small ellipsoids and are called stiff (Tonsing et al. 2014, Brown & Sethna 2003, Gutenkunst 2007).

An eigenvalue spectrum separated by more than three orders of magnitude is known as a sloppy model. Some have proposed that sloppiness is a general property of models in systems biology, with 17 models that were analysed all having sloppy eigenvalue spectra (Gutenkunst 2007). The consequence is that model calibration is difficult even with a large amount of data. Tonsing et al. 2014 investigated the sources of sloppiness. They reasoned that the eigenvalue spectrum is determined only by the shape of the sensitivity matrix which in turn is determined by the experimental design. Starting from random, iid sampling, Tonsing et al. 2014 constructed and studied the eigenvalue spectra of different Hessian matrices, gradually introducing characteristics that make the Hessian matrix a closer approximation to a real model calibration problem. The rows if the

Hessian matrix index observations while the columns index model parameters. Tonsing et al. 2014 found that autocorrelations in time series data was a source of structure in the row direction of the Hessian matrix and that dissimilarities in the types of kinetic parameter (i.e. different units) for biochemical reactions were a source of structure in the column direction of the Hessian matrix. Moreover, since little or no information is provided from some observations about some parameters, real Hessian matrices are sparse. When they added sparsity to autocorrelations and parameter grouping by units they were able to reliably reproduce a sloppy eigenvalue spectra for simulated Hessian structures. The newly characterised structure of the Hessian matrix was exploited for optimal experimental design based on choosing experiments that minimize sloppiness.

2.4.9 Stability and bifurcation analysis

Stability analysis in the study of dynamical systems is used to describe the long term behaviour of a system. Namely if a system is stable it converges to a steady state and the net rate change for all model equations is 0. The steady state in this case is said to be an attractor. If a system is unstable, a model component tends towards infinity with time. Oscillating systems are known as marginally stable whereas unpredictable systems are known as chaotic (DiStefano III 2015).

Biological systems, like other types of physical system, are in constant steady state: opposing forces constantly counteract each other and produce the illusion of stationarity. Much of systems modelling in biology is involved in studying how the long term behaviour of a system changes when perturbed, whether it asymptotically tends towards the same steady state or to a qualitatively different steady state, for example in bistable systems.

A system of equations $\dot{\vec{x}}$ (Equation 2.2) is at steady state if the rate of change of all $x_i \in \vec{x}$ is 0 (Equation 2.26).

$$\dot{\vec{x}} = f(\vec{x}, \vec{u}, \vec{\theta}) = 0 \tag{2.26}$$

A solution to (Equation 2.26) is a multidimensional point in state space (i.e. one for each

model variable) that is also known as a critical or singular point. Finding critical points can be non-trivial and involves finding the nullclines on the phase portrait of the system. Assuming a 2 dimensional system, a phase portrait of the system can be produced by plotting the concentrations of model species against each other as the system evolves throughout time. If species a is on the x-axis, the x-nullcline is the line where $\dot{a} = 0$. Similarly, if species b is on the y-axis, the y-nullcline is the line where $\dot{b} = 0$. The steady state, fixed points or critical points of the system are where the nullclines intersect. This principal generalised to any number of dimensions, though it is more difficult analyse.

For non-linear systems such as those prevalent in systems biology, a steady state is evaluated by calculating the eigenvalues of the Jacobian matrix J (Equation 2.7). The eigenvalues $\vec{\lambda}$ are often complex numbers. If all real parts of $\vec{\lambda}$ are negative then the system is asymptotically stable while if at least 1 real part is positive, the system is unstable. If the system has mostly negative real parts but with some that are 0 then the system can be either stable or unstable (Wang et al. 2005).

Oscillating systems are prevalent in biology are can arise from situations such as a delayed negative feedback (Brsch & Schaber 2016) or the combination of a positive and negative feedback within one network (Tsai et al. 2008). A phase portrait of two species that participate in a stably oscillating system form a closed trajectory that is known as a limit cycle. Damped oscillations on the other hand tend to spiral in towards the stable steady state.

The stability of a non-linear system is dependent on model parameters $\vec{\theta}$. A bifurcation analysis reveals how the systems stability changes with model parameters. A bifurcation plot has parameter values on the x-axis and the solution to Equation 2.26 on the y-axis. As such there can be as many bifurcation plots as model parameters. If a system's qualitative behaviour changes at some value of parameter θ_j, it is a bifurcation point. Often there are no such points and the system always tends towards a predictable steady state solution that is dependent on the $\vec{\theta}$. Other times a system has regions of stability and regions of instability, depending on the values of $\vec{\theta}$. Oscillating systems usually have regions of parameter space that are non-oscillating, some that cause system

components to have damped oscillations and some that result in persistent oscillations. Each time the system qualitatively changes behaviour, there is a bifurcation point.

In larger models, traditional bifurcation analysis is of limited value because of their multidimensional nature. However, recent numerical methods enable global stability analysis and visualisation of multidimensional bifurcation diagrams. This method, called 'DYVIPAC' (dynamic visualisation based on parallel coordinates") essentially bootstraps a local stability analysis using Monte Carlo parameter sampling and numerically computes the stability of the system at each parameter set (Nguyen et al. 2015).

2.5 Conclusion

Systems biology is a modern approach to biological science that aims to integrate molecular biology with mathematics, statistics, engineering and computer science to revolutionise our understanding of biological processes. Systems biology emphasises the study of interactions between biological entities and how life emerges from these interactions. Systems biologists employ high throughput experiments for the identification of interaction networks and quantitative mechanistic modelling for a more detailed analyses of how lower level interactions can produce high level behaviour. Collectively these are all tools in the systems biologists arsenal that can be used to better our understanding of health and disease.

Chapter 3

Python packages, `PyCoTools` and `pytseries`, for systems modelling and analysis of time series data

Programming languages such as Python, Matlab and R are ideal for scientific computing because they offer paradigms such as object-oriented programming for code modularity and extendibility. Consequently, procedural tasks such as automating a workflow or iterating over an algorithm are readily performed. The Python programming language has been instrumental to the work presented in this thesis and two Python packages have been designed, developed and deployed throughout the course of the work.

The first is called PyCoTools (Welsh et al. 2018) and the full published article is presented in the appendix (Section 1.4). The purpose of PyCoTools is to facilitate systems modelling by automating the configuration of common tasks such as simulating or estimating the parameters of ODE models using COPASI (Hoops et al. 2006). COPASI is primarily a graphical user interface (GUI) based application for the development and analysis of ODE models. While the GUI interface allows COPASI to validate user input, it also slows the performance of common tasks, since it requires manual configuration every time a task is performed. COPASI also has application-programming interfaces (API) for several computing languages, but their

design closely mirrors the underlying C++ code and is not optimal for everyday use. Since COPASI encodes a model using XML, it is easy from a programming perspective to reproduce COPASI task configurations. Through combined use of the PyCoTools interface into COPASI, the Python programming language and other Python packages in systems modelling (Antimony Smith et al. 2009 and Tellurium Choi et al. 2016) it was possible to automate the configuration and execution of parameter estimations, time course simulations, model composition, model selection, profile likelihood calculations (Raue et al. 2009) and analysis of parameter estimation data via exploratory data analysis. The scope of PyCoTools was initially limited to model calibration but is in principle extendible to fully support all of COPASI features. This latter point is demonstrated by the recent addition of support for COPASIs sensitivity analysis task in PyCoTools.

The second Python package developed during this PhD is called `pytseries` and is the subject of discussion for the remainder of this chapter. `pytseries` was designed to facilitate the handling, analysis and manipulation of time series data.

3.1 Introduction

Time series data are repeated measurements of a variable at different points in time. For this reason, time series (or dynamic) data provides richer information than data collected at a single data point (static data). In molecular biology, time series data are often collected for mRNA and proteins in order to study the dynamics of cell signalling networks (Vizan et al. 2013).

In systems biology, time series data is of particular interest because they can be used to calibrate mechanistic models of biological systems (Raue et al. 2013). One of the major difficulties however is the volume of data required to adequately inform a mechanistic model. Traditional low throughput experiments such as western blots, ELISA or qPCR are time consuming, particularly for time series measurements and for the many

observables in a model. An alternative is to use high throughput experimental methods to measure many observables at once.

As mentioned, object oriented programming environments such as Matlab, R and Python are ideal for working with large datasets. Each language has their own data structures that serve as building blocks that users can piece together to perform useful analyses. For instance, Matlab has structs, vectors or matrices; R has data frames and tables and Python has add on packages such as pandas, numpy and scipy which have custom objects that make Python excellent for scientific computing.

While the existing data structures are perfectly adequate for handling timeseries data, they do not offer maximum usability because they are 'general purpose' and they cater for all types of data, not just time series data. In the remainder of this chapter we describe the `pytseries` Python package that was built for explicitly handling time series data and automating repetitive tasks. Specifically, methods are available for interpolation, normalisation, visualisation, computing the dynamic time warping distance (Salvador & Chan 2007) and for clustering time series data.

3.2 Methods

The `pytseries` package is implemented in Python 3.6 and the source code is available https://github.com/CiaranWelsh/pytseries. The package can be installed with the command '`pip install pytseries`' and the documentation is available at https://pytseries.readthedocs.io/en/latest/.

3.3 Results

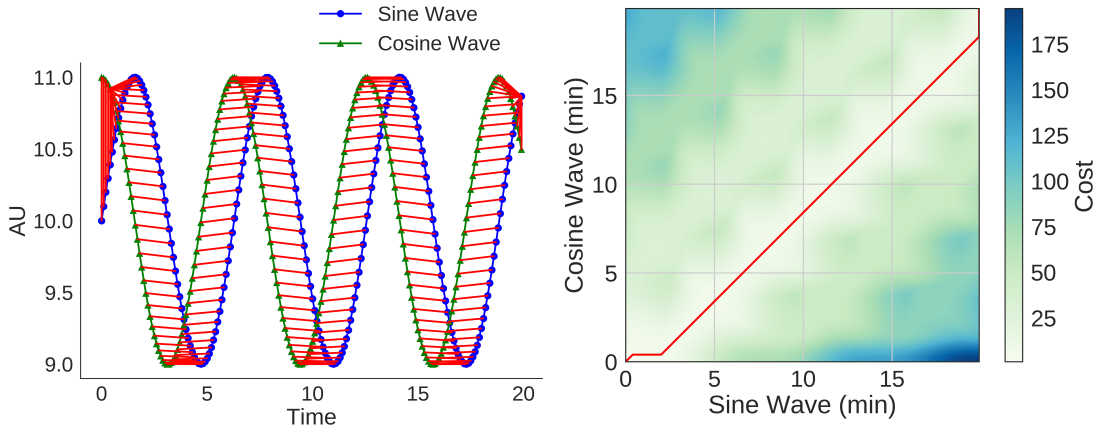
The `pytseries` package is written in Python and is based on well-established objects from Python's `pandas` library. Under the `pytseries` namespace there are three main modules. The `core` module contains two classes which are analogous to the `pandas.Series` and the `pandas.DataFrame` objects, but specifically for time series data. Both classes have `time` and `features` as properties and the time series data is stored under the `values` property. Data stored in these objects can be visualised, normalised, interpolated, differentiated and stored or retrieved from a database. Moreover, statistics such as means, standard deviation, coefficient of variance and median can be easily calculated.

The `dtw` module contains code for computing the dynamic time warping distance between two `TimeSeries` objects. Methods exist specifically for visualising the DTW distance between any X and Y.

The `clust` module contains objects for classifying time series data and expects `TimeSeriesGroup` objects as input. The `clust` module contains two classes, `FindSimilar` and `TimeSeriesKMeans`. The former simply computes the DTW matrix between all pairs of time series within a `TimeSeriesGroup` and returns those which are lower than a user provided threshold. The latter is a wrapper around the `tslearn.TimeSeriesKMeans` class https://github.com/rtavenar/tslearn. The `pytseries` implementation additionally adds methods for visualisation.

3.3.1 Dynamic Time Warping

Dynamic time warping is an algorithm developed in the late 1950's (Bellman & Kalaba 1958) and is most well-established for being applied to speech recognition (Rabiner et al. 1978). DTW finds the optimal 'warping' distance between two sequences $\vec{X} = \{x_1, \cdots, x_I\}$ and $\vec{Y} = \{y_1, \cdots, y_J\}$. The term 'warp' refers to stretching or contracting the time axis of $\vec{Y}$ so that $\vec{X}$ and $\vec{Y}$ are optimally aligned, or in other words,



(a) DTW plot

(b) Cost matrix plot

Figure 3.1: Dynamic time warping example using a sine and a cosine wave. (a) A plot of the sine and cosine wave where the connecting red bars indicate time points on the cosine wave that are mapped to the time points of the sine wave. (b) cost matrix C for the time warping presented in (a). The central red line indicates the optimal time warping path W.

the distance between them is minimal. In Figure 3.1, the time warping distance is represented by a series of red lines connecting the sine and cosine wave. Each data point (excluding the ends) on the cosine wave is moved along the x-axis to the point at which the euclidean distance between sine and cosine wave is minimal. This is performed for each time point and the new indices represent the optimum warping path $W \in \mathbb{R}^{k\times 2}$. The warping path between $\vec{X}$ and $\vec{Y}$ is

$$W = \{w_1, \cdots, w_k, w_K\}, \quad \max(I, J) \geq K < (I + J) \tag{3.1}$$

where I and J are the lengths of $\vec{X}$ and $\vec{Y}$ respectively and K is the length of the warping path. The k_{th} element of W is

$$w_k = (i, j) \tag{3.2}$$

where i and j index $\vec{X}$ and $\vec{Y}$ respectively. The qualifier in Equation 3.1 states that the size of the warp path K is greater than the bigger of $\vec{X}$ and $\vec{Y}$ but smaller than the sum of $\vec{X}$ and $\vec{Y}$. Three constraints are enforced with DTW:

1. The warp path must begin at the start of both $\vec{X}$ and $\vec{Y}$, i.e. $W_0 = (0, 0)$

2. The warp path must end at the end of both $\vec{X}$ and $\vec{Y}$, i.e. $W_K = (I, J)$

3. The warp path is monotonically increasing

Points (1) and (2) ensure every index is used while (3) ensures the path does not overlap itself (Salvador & Chan 2007). After alignment, the distance of the warp path W is computed as the sum of the individual distances indexed by the warp path (Equation 3.1)

$$Dist(W) = \sum_{k=1}^{K} Dist(W_{ki}, W_{kj}) \tag{3.3}$$

Where $Dist(W)$ is the distance for warp path W (typically Euclidean) and $Dist(W_{ki}, W_{kj})$ is the distance between two data points X_i and Y_j in the k_{th} element of the warp.

Computing DTW is computationally demanding. The brute force method is a non-viable option for even small sequences. To more efficiently compute the optimal path W, there exists a dynamic programming algorithm which is quadratic in its complexity and a fast DTW algorithm which is an approximation of the exact solution but linear in its complexity (Salvador & Chan 2007).

The dynamic programming method involves computing the solution to increasingly bigger subsections of the problem until the solution to the whole problem is found. The DTW algorithm first computes an I by J cost matrix C where the value at cell $C(i, j)$ is the calculated using Equation 3.3. The cost matrix C between a sine and cosine wave are shown in Figure 3.1. The value at each cell of the cost matrix is calculated as:

$$C(i, j) = Dist(i, j) + min(C(i-1, j), C(i, j-1), C(i-1, j-1)) \tag{3.4}$$

To find the warp path W, the algorithm starts at the end (highest time index) of both time series X_I and Y_J and finds the adjacent cell with the lowest cost. Because of the constraints, the warp path W can only go left, down or diagonally left and down. The warp path W is represented as a red line through the cost matrix in Figure 3.1b. The $Dist(W)$ is then the sum of these points.

3.3.2 Time Series Clustering

K-means is an unsupervised clustering algorithm that was not initially developed for time series data but has been adapted for this purpose using the DTW as a metric of similarity. The purpose of the K-means algorithm is to partition a set of time series ts into K groups.

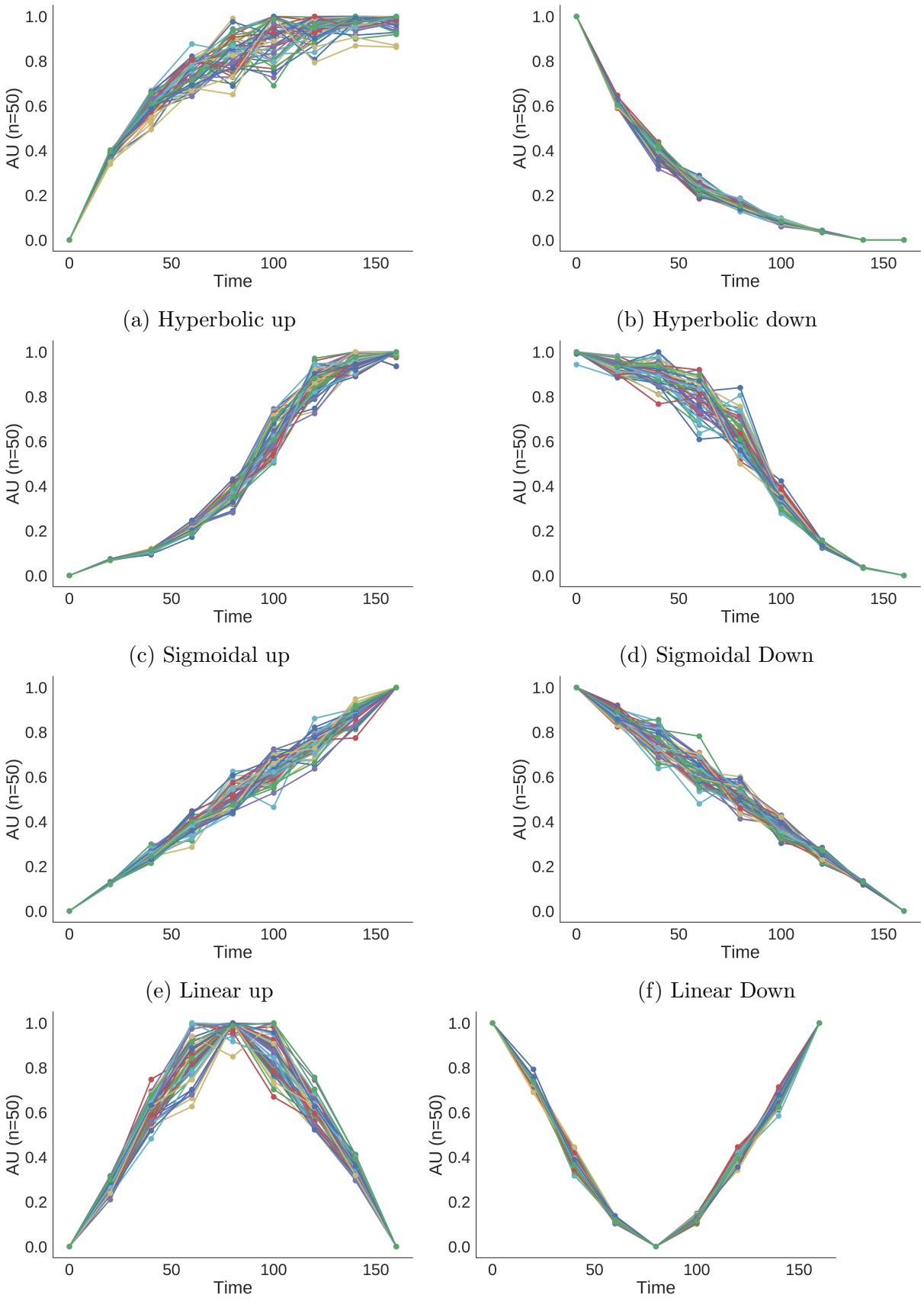
Briefly, in the K-means clustering algorithm, cluster prototypes (or centroids) $\mu_1, \cdots, \mu_k$ are first defined to be random vectors with the same dimensions as the time series data. Also defined is an indicator matrix r_{ij} which is 1 where the i^{th} time series belongs to cluster j and 0 everywhere else. The K-means algorithm clusters time series data by minimizing Equation 3.5, which is different from the canonical K-means algorithm because the DTW distance is used in place of typical distance functions such as the absolute difference or euclidean distance.

$$J(r, \mu) = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{k} r_{ij} DTW(C_j, x_i) \tag{3.5}$$

The first step for minimizing Equation 3.5 involves fixing the prototype vectors $\vec{\mu}$ while the assignments r are determined. For this, a DTW matrix is computed between each time series ts_i and each prototype C_j and assigned to the C_j with lowest DTW. Then, the indicator matrix r is fixed while the multidimensional average is computed for each cluster C_j at each time point. These are the new C_j vectors. This process is repeated until the clusters do not change significantly (Everitt 2011).

An important aspect of the K-means algorithm is that it is sensitive to initial assignments of C_j. The solution is to repeat the clustering multiple times from random initial conditions and to take the best (i.e. minimum $J(\cdot)$).

The K-means used in the `pytseries` package has been implemented by Romain Tavenard and is available in a Python package called `tslearn` under the BSD-2-clause licence. The source code is publicly available at the following link https://github.com/rtavenar/tslearn. `pytseries.clust` inherits from



(a) Hyperbolic up

(b) Hyperbolic down

(c) Sigmoidal up

(d) Sigmoidal Down

(e) Linear up

(f) Linear Down

(g) Transient up

(h) Transient down

Figure 3.2: Synthetic data before clustering. (a-h) 50 timeseries were simulated by adding noise to exemplar time series for each profile shapes so that they can be sorted by the K-means algorithm.

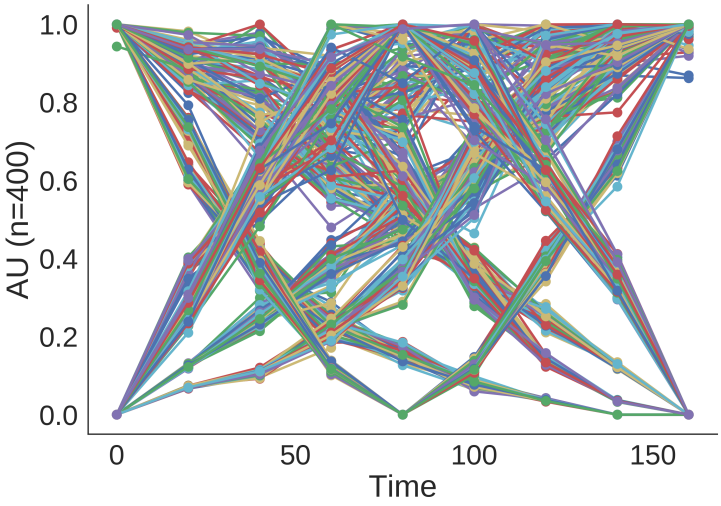


Figure 3.3: Visualization of all timeseries objects from the `core.TimeSeriesGroup` object.

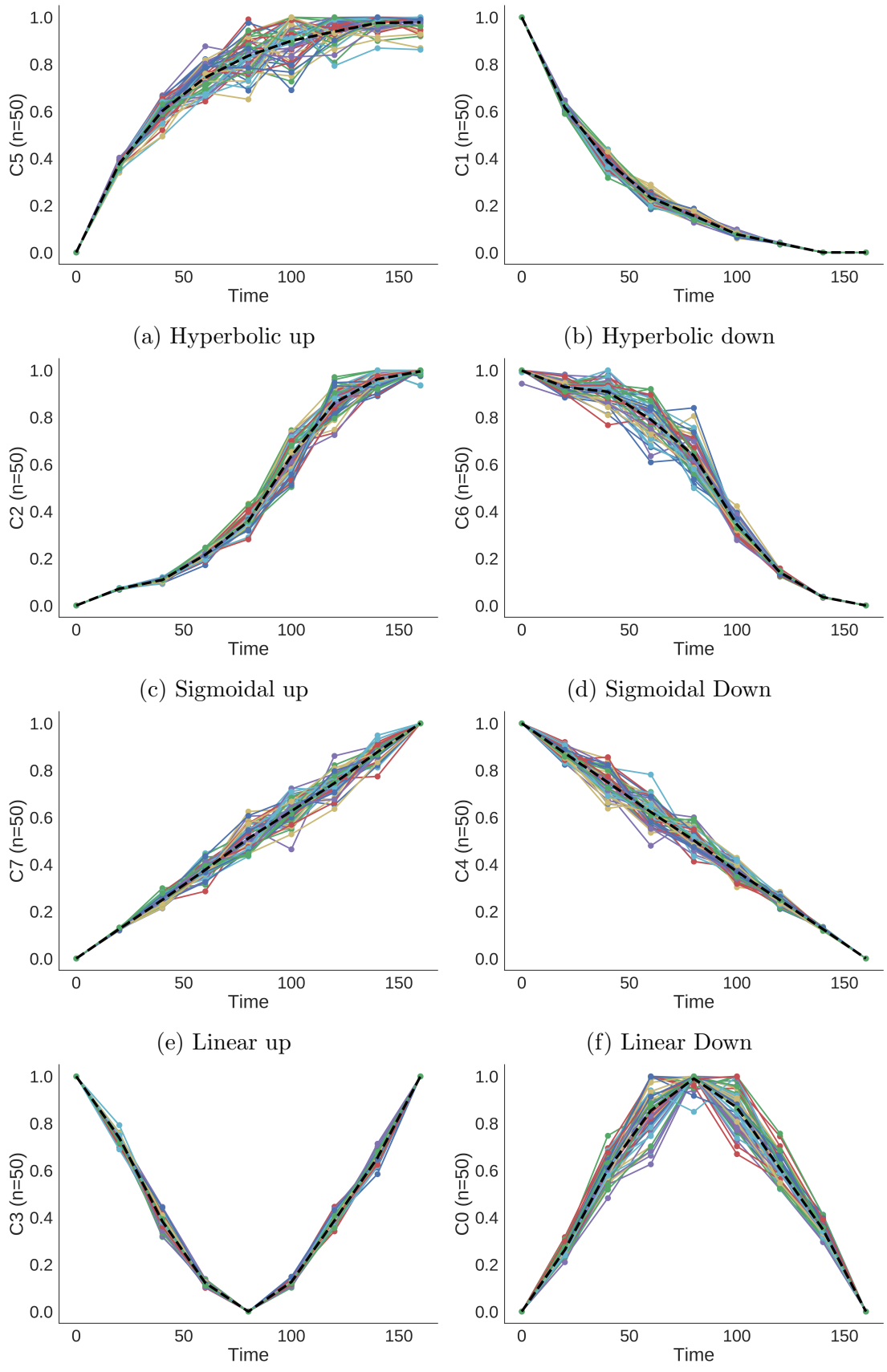
`tslearn.TimeSeriesKMeans` and implicitly converts to and from the expected the data structures used by `pytseries` and `tslearn`.

A Synthetic Example

In later chapters, this algorithm is applied a real dataset but here, a demonstration is provided using synthetic data.

To create the synthetic data, 8 time series with 9 time points over 180 minutes were created, each with a different shape. Then, 50 profiles were simulated from each profile by drawing from a 9-dimensional normal distribution $\mathcal{N}(ts_{ij}, 0.1 \cdot ts_{ij})$, one for each time point. Here ts_{ij} represents the j^{th} time point of the i^{th} time series. The resulting 400 time series are visualised in Figure 3.2 before concatenation into a single `TimeSeriesGroup` object and normalized so that the minimum and maximum values of each time series is 0 and 1 respectively (Figure 3.3).

The `clust.KMeansTimeSeries` algorithm was applied to the time series using $K = 8$ and 20 iterations from random starting conditions. As shown, in this ideal example where the number of clusters is known *a priori* and the profile shapes are clearly different, the clustering procedure can attain a 100% accuracy rate (Figure 3.4).



(a) Hyperbolic up

(b) Hyperbolic down

(c) Sigmoidal up

(d) Sigmoidal Down

(e) Linear up

(f) Linear Down

(g) Transient up

(h) Transient down

Figure 3.4: Clustering result for synthetic data. Each of the 50 time series from each group were correctly classified by the K-means algorithm. The dashed black line in each plot represents the cluster prototypes C_j

3.4 Discussion

This chapter discusses the `pytseries` package which is a Python implementation of data structures for handling time series data. The objects developed are based on known well-established objects from Pythons `pandas` library. The benefit of having objects designed specifically for time series data is that reusable methods can be built so that no time is wasted by reimplementing easily repeatable functions, such as developing publication quality figures of time series objects. Other features of the `pytseries` package include the ability to group time series objects and perform numerical operations such as interpolation or normalization on the collective group of time series and quickly computing various statistics regarding the time series group. The `pytseries` package implements computation of the DTW distance between two `TimeSeries` objects and enables the clustering time series data using a modification of the K-means algorithm.

The K-means algorithm was tested on synthetic data to test its efficacy in identifying similar groups of time series. The synthetic data were deliberately created so that 8 types of profile (Figure 3.2) could be grouped together (Figure 3.3) and sorted using the `TimeSeriesKMeans` class. The K-means algorithm was applied using the known number of clusters and the time series data were correctly sorted into their profiles of origin (Figure 3.4). While this example found the correct solution, the problem was purposefully designed to be easy for demonstrating its potential.

The object-oriented nature of the `pytseries` package means that the work is extensible. In addition to including other clustering algorithms from the `tslearn` package, there are a number of other analyses that could be implemented to work with the `pytseries` objects. Examples include gene ontology enrichment and network inference algorithms such as Aracne (Margolin et al. 2006). Collectively, the `pytseries` package enhances productivity when working with time series data in Python.

Chapter 4

A genome-wide analysis of early TGF-β signalling in neonatal fibroblasts

4.1 Introduction

The pleotropic nature of the TGF-β system is well established (Fleisch et al. 2006), though it is still unclear how TGF-β signalling can accommodate such a diverse set of contextual biological actions (Zhang 2018). As discussed in Chapter 1, TGF-β binds to its receptors on target cells and triggers a set of signalling events that lead to the activation of the Smad second messenger system. It is well understood that Smads play a key role in TGF-β signalling, but it is also clear that a huge amount of communication exists between TGF-β and other signalling pathways including Erk1/2 MAPK, PI3K/Akt and Rho-like GTPases (Zhang 2009). This section expands on the introduction given in Chapter 1 to discuss cross-talk in the TGF-β system.

While many of the details remain unknown, research to date has provided an insight into some of the connections that exist between TGF-β and other signalling pathways. Activated TGF-β receptors in specialised Cav1 positive regions of cell membrane, called caveolae, are targeted for degradation by Smad7 (Di Guglielmo et al. 2003). However,

activated TGF-β receptors can also activate Erk1/2 through a signal that is initiated by TGF-β receptors in caveolae compartments (Mulder 2000, Muthusamy et al. 2015). In canonical Erk1/2 activation, EGF binds to its receptors and induces their autophosphorylation on tyrosine residues. This event causes two SH2 domain containing adaptor proteins, Shc and Grb2, to bind the receptor and recruit Sos, a guanine exchange factor (GEF) for Ras (Diaz et al. 1997, Suzuki et al. 2007). Ras can then activate a third tier MAPK, Raf, which initiates a phosphorylation cascade that activates Mek and Erk (Minden et al. 1994, Lake et al. 2016). TGF-β signalling is able to connect with this process by activating and recruiting ShcA to type 1 receptors (Lee et al. 2007). As a consequence, Ras is 'loaded' with GTP leading to Raf, Mek and Erk activation (Zhang 2009).

TGF-β is also able to activate the other two MAPK pathways, Jnk and p38. After cells are stimulated with TGF-β, type 1 receptors interact with Traf6, a protein typically considered as a signalling component of NF-κB activation (Deng et al. 2000). This event leads to Traf6 autoubiquitination and then to activation of Tak1, a kinase well known for its involvements in Jnk and p38 activation by intermediate activation of second tier MAPKs, MK3 and MKK6 (Yamashita et al. 2008, Sorrentino et al. 2008). Moreover, because Tak1 can phosphorylate and reduce the stability of SNoN (Kajino et al. 2007), a negative regulator of Smad signalling, Tak1 positively influences Smad signalling at the same time as activating Erk1/2-mediated transcription.

Traf6 is an important protein in the transduction of TGF-β signals because it is involved not only in the activation of MAPKs but also PI3K. Traf6 polyubiquitinates the p85 subunit of PI3K and promotes an association between TGF-β type I receptors and PI3K (Hamidi et al. 2017). Moreover, Akt phosphorylation on both Ser473 and Thr308 has been measured in fibroblasts treated with TGF-β (Wilkes et al. 2005). An interesting aspect of TGF-β mediated PI3K and Erk activation is that PI3K and Erk signalling have long been shown to have numerous lines of cross-talk independently from TGF-β (Mendoza et al. 2011, Ursini-Siegel et al. 2012).

The evidence that PI3K, Erk and TGF-β are in close communication raises the question

as to whether they should be considered separate entities or whether they are better perceived as multiple aspects of the same signalling pathway. Such a view point is given weight by evidence that the regulation of TGF-β responsive genes, such as COL1A1 and COL1A2, do not solely depend on Smads but also on Erk (Bhogal & Bona 2008) and PI3K (Wilkes et al. 2005). For instance, PI3K acting on signals from TGF-β induces activation of Pak2 and c-Abl (Wilkes & Leof 2006) which then activates PKC-δ (Yuan et al. 1998, Sun et al. 2000). PKC-δ is involved in disinhibition of COL1A2 transcription by Fli1 (Jinnin et al. 2005), a inhibitory transcription factor for the synthesis of both COL1A2 and CTGF (Jinnin et al. 2005, Asano & Trojanowska 2013, Nakerakanti et al. 2006, Czuwara-Ladykowska et al. 2001). While Fli1 inhibits, Ets1 stimulates ECM synthesis, is activated by TGF-β and operates at the same promoter site as Fli1. Therefore, TGF-β via PI3K, Pak2, c-Abl and PKC-δ prepares the type 1 collagen promoter for transcription (Asano et al. 2007, 2009) while Smad3 (Chen et al. 1999), Sp1 (Greenwel et al. 1997), AP1 (Ponticos et al. 2015, Ponticos, Harvey, Ikeda, Abraham & Bou-Gharios 2009) and Ets1 (Czuwara-Ladykowska et al. 2002) form a transcriptional activation complex on gene promoters leading to the transcription of collagen precursors. Thus, while Smads have an essential role in collagen production, they are not the only elements that require consideration because a coordinated system of signalling cascades collaborate in fine tuning the response.

TGF-β cross-talk extends beyond Erk and PI3K as interactions with various other signalling pathways have also been documented. For instance, Jak1 and Stat3 signalling (Tang et al. 2017, Su et al. 2017) is important in the integration of profibrotic signals in the fibroblast (Chakraborty et al. 2017). Jak1 is a constitutive TGF-β receptor type I binding protein and is necessary for TGF-β mediated activation of Stat3. Further, Stat3 is a binding partner of Smad3 and this interaction attenuates Smad3 signalling (Wang et al. 2016). Other modes of TGF-β cross-talk include TGF-β-activated NF-κB via tak1 (Gingery et al. 2008, Freudlsperger et al. 2013); regulation of GSK3-β, an important event in fibroblast differentiation (Caraci et al. 2008, Baarsma et al. 2013) and interactions with Snail (Vincent et al. 2009), Rho GTPases (Atfi et al. 1997, Edlund et al. 2002), Wnt (DiRenzo et al. 2016) and Hedgehog signalling pathways (Javelaud

et al. 2012, Kluppel & Wrana 2005).

In this chapter a genome wide analysis of the immediate fibroblast response to TGF-β was conducted. The aim was to identify genes that respond to TGF-β stimulation and influence the activity of other signalling pathways. Differential expression analysis revealed 443 individual probes encompassing 126 genes that were differentially regulated within the first 3h of TGF-β stimulation in neonatal fibroblasts. A pathway analysis was then conducted using DAVID (Huang et al. 2008), an integrated platform for enrichment pathway analysis. Both KEGG and (Kanehisa et al. 2016, Kanehisa & Goto 2000) and Reactome (Fabregat et al. 2018) pathway analyses were conducted and provides support for TGF-β mediated regulation of genes involved in TGF-β, PI3K, Erk and FOXO pathways and weak support for TGF-β mediated regulation of HIF and TNF-α signalling. Lastly, the differentially expressed genes were classified based on their dynamic profile using the time series K-means algorithm that was described in Chapter 3.

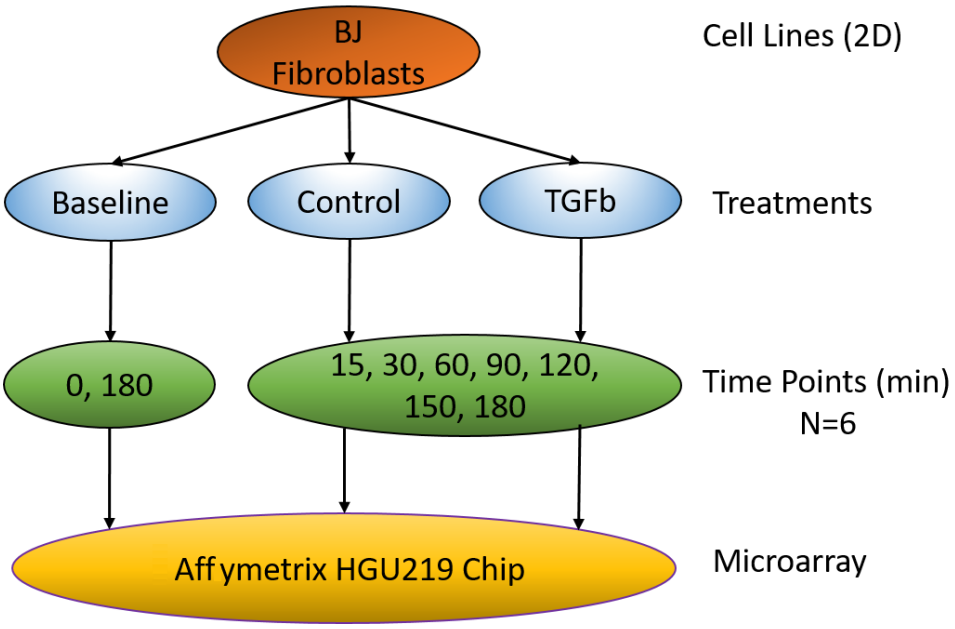
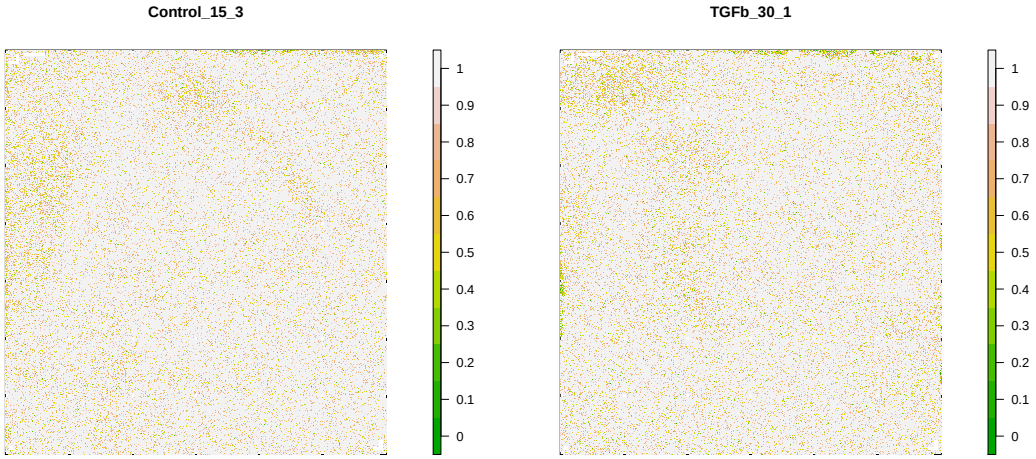


Figure 4.1: Experiment design. BJ neonatal fibroblasts were treated with baseline, control or 10ng ml^{-1} TGF-β1 for the indicated time points. After RNA extraction, samples were quantified using Affymetrix microarray technology using the HGU219 gene chip.



(a) 15 minute time point, repeat 3 for control treatment

(b) 30 minute time point, repeat 1 for TGF-β treatment

Figure 4.2: Representative examples of probe level models for microarray data. Colour bars represent the normalised intensity of the raw probe.

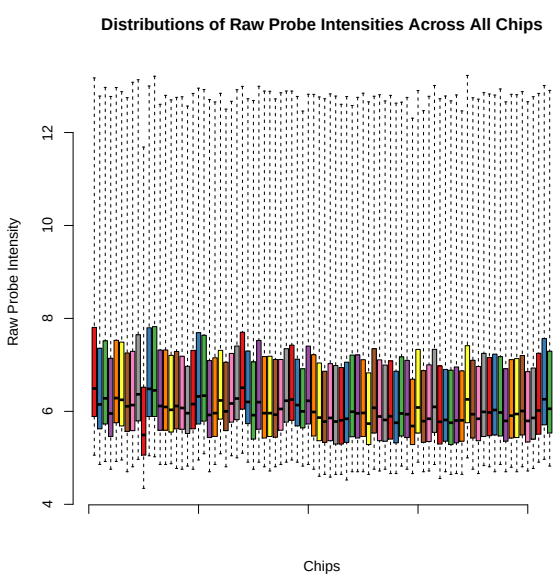
4.2 Methods

4.2.1 Cell Culture

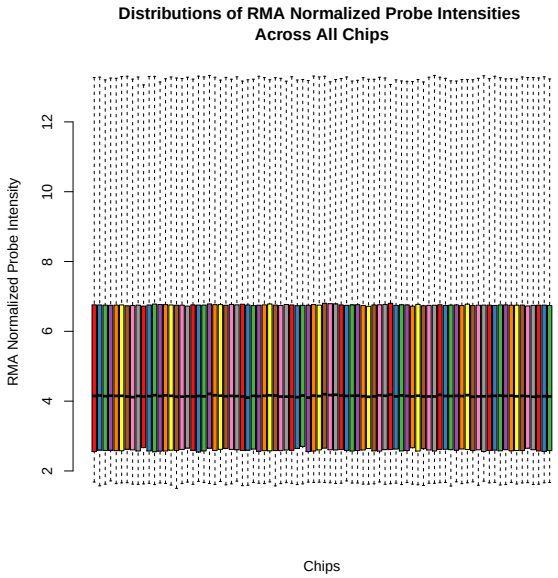
Neonatal BJ fibroblasts (CRL-2522, ATCC, Manassas, VA, lot 62341989, P-6) were cultured using standard tissue culture techniques at P&G, Cincinnati. Cells were seeded at a density of $25,000\text{cm}^2$ into 12-well plates using 'complete' EMEM which contains 10% fetal bovine serum (FBS, ATCC) in Modified Dulbecco Eagle Medium (EMEM, ATCC). After overnight incubation, media was removed and replaced with 1 ml of complete EMEM.

4.2.2 Microarray Experiment Design

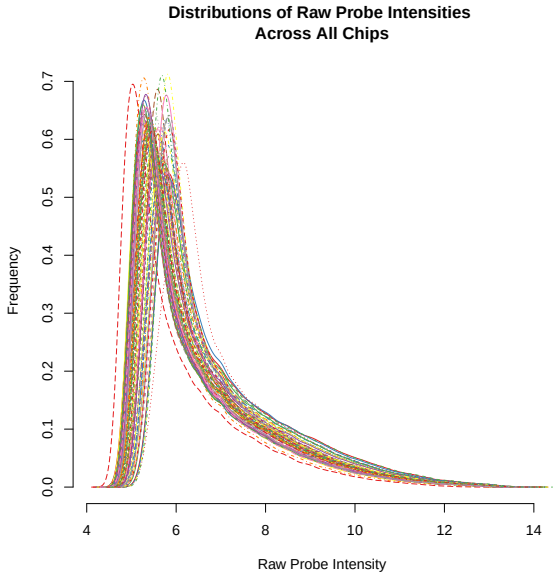
After an overnight incubation, cells were treated in one of three ways: 1) Baseline cells were not manipulated in any way; 2) Control cells had 1 ml of 0.1% BSA in HCl (1:1000) added to the existing media and 3) Treated cells had 1 ml of media containing 20ng ml^{-1} TGF-β to make a final in-well concentration of 10ng ml^{-1}. At the designated



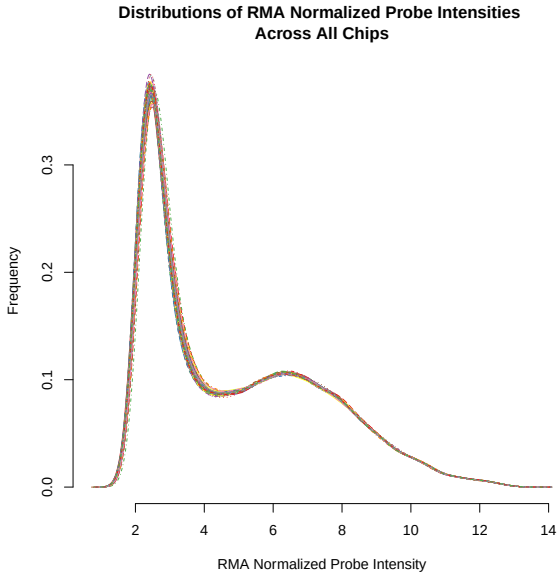
(a) Box plot before normalization



(b) Box plot after normalization



(c) Histogram before normalization



(d) Histogram after normalization

Figure 4.3: Comparison of probe intensity distributions before and after normalization.

times (0 or 180 for baseline and 15, 30, 60, 120, 150, 180 minutes post treatment for control or treated groups Figure 4.1), conditioned media was aspirated and cells were lysed using 1ml RLT Lysis Buffer (Qiagen, Hilden, Germany). The contents were mixed, transferred to Eppendorf tubes and frozen at -20°C for storage. All 16 conditions (7 time points * 2 treatments + 2 baseline) were repeated 6 times. Samples were then processed on an Affymetrix GeneTitan microarray platform using the human genome U219 chip (HGU219).

4.3 Results

4.3.1 Data Preprocessing and Quality Control

Microarray data is subject to a lot of variation which can arise from both biological and technical origins. For this reason, microarray data must be subject to rigorous quality control to identify and remove potential outliers prior to inferences. This section describes the use of several routine measures of quality control that was conducted on tehe microarray data prior to differential expression.

Probe Level Model

Raw Affymetrix microarray data are stored in 'CEL' files and represent raw probe intensities. A probe level model reconstructs an image of the chip from the corresponding CEL files and any irregularities that occurred during the experimental process, such as trapped air bubbles or scratches are visible (Bolstad et al. 2005).

To construct the probe level model, the R package package `AffyPLM` was used with default parameters for each of the 96 microarray chips in this experiment (Bolstad et al. 2005). Figure 4.2 shows a representative example of the images produced by the probe level model. These images show a uniform distribution of very low intensity probes.

Since they are void of obvious experimental artefacts they were deemed good enough quality to be taken to the next stage of preprocessing.

Background Correction, Normalization and Summarisation

The preprocessing of an Affymetrix microarray experiment involves background correction, normalization and summarisation. In this analysis, the 'robust multiarray averaging' (RMA) method was used as implemented in the R package, `rma` (Irizarry et al. 2003). As shown in Figure 4.3, the normalization procedure was successful in aligning the data distributions from each chip.

Principal Component Analysis

Principal component analysis (PCA) was used identify potential outlier samples and to provide an overview of the data. In Figure 4.4, all control samples cluster together, along with the early time points from the TGF-β treated samples. Later time points from the control samples cluster separately from the earlier time points which indicates that the fibroblasts are not inert but active, even under the control condition, thereby justifying the time matched controls. The TGF-β treated samples showed progressive clustering of sequential time points away from the control samples. Figure 4.4 does not show any obvious outliers indicating the experiment was conducted effectively and the resulting data is of high quality.

Table 4.1: Differential expression comparison for different degrees of freedom

df	adj.P.val < 0.001	F-statistic > 100
3	6956	527
4	6689	443
5	6454	360

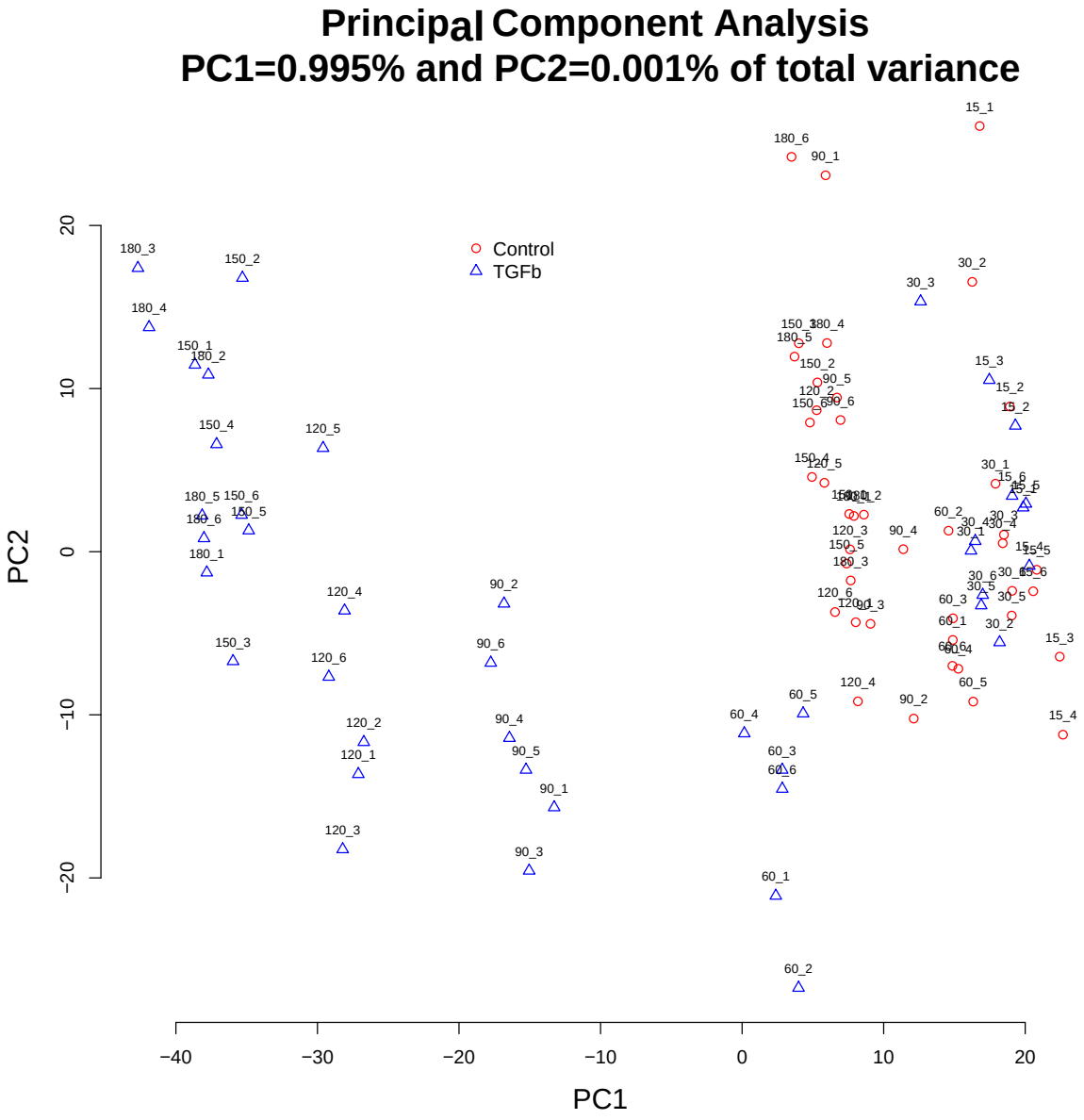


Figure 4.4: PCA on normalized microarray experiment. Each spot represents a single sample on a gene chip. Red circles indicate control samples while blue triangles are TGF-β treated samples. Labels are time in minutes and replicate separated by underscore.

4.3.2 Differential Expression

The Affymetrix HGU219 chip contains 49,386 probes. 24,739 were non-responsive across all conditions and therefore removed prior to subsequent analysis. The remaining 24,647 genes were subject to differential expression analysis using LIMMA.

Using LIMMA, a moderated F-test (Smyth 2004) was conducted to test whether each probe is significantly different in the TGF-β time series compared to the control time series. This process uses a basis-spline for interpolation and requires a degrees of freedom *df* parameter. Since there is no fixed way to choose the *df* parameter, a *df* of 3, 4 and 5 were tested (Table 4.1). Higher values for *df* resulted in a lower number of differentially expressed genes. To strike a balance between reducing the number of genes for subsequent analysis and not loosing potentially interesting differentially expressed genes, a degrees of freedom parameter of 4 was chosen for subsequent analysis. A second parameter required for differential expression analysis is a p-value cut-off. The p-value represents the probability that the null-hypothesis, i.e. that the probe under examination is the same under both the control and the TGF-β condition, holds true. The p-values were adjusted for multiple testing using the Benjimini-Hochberg correction but, as shown in Table 4.1, a p-value cut-off of 0.001 still resulted in too many genes being differentially expressed. Therefore a more stringent cut-off of an F-statistic >100 was chosen which corresponds to a much lower p-value.

Overall, 443 probes corresponding to 126 genes were significantly differential expressed when exposed to TGF-β over time, compared to control). The official gene symbols and the corresponding protein names are described in Tables 4.2 to 4.6. These tables are linked to the official GeneCards entry via the hyperlink in the 'Protein' column (Safran et al. 2010). The time series data corresponding to the genes in these tables are presented in Figures 4.5 to 4.13.

Table 4.2: Index of genes that are differentially expressed in TGF-β treated fibroblasts compared to controls. 1 of 5.

Gene Symbol	Protein
ABHD17C	Abhydrolase domain containing 17C
ADAM19	ADAM metallopeptidase domain 19
ADAMTS1	ADAM metallopeptidase with thrombospondin type 1 motif 1
ADM	Adrenomedullin
AHRR	Arylhydrocarbon receptor repressor
AMIGO2	Adhesion molecule with Ig like domain 2
ANGPTL4	Angiopoietin like 4
ARHGAP29	Rho GTPase activating protein 29
ARID5B	ATrich interaction domain 5B
ARRDC4	Arrestin domain containing 4
AUTS2	Autism susceptibility candidate 2
AVPI1	Arginine vasopressin induced 1
BCL11A	Bcell CLL/lymphoma 11A
BCL6	Bcell CLL/lymphoma 6
BCOR	BCL6 corepressor
BDNF	Brain derived neurotrophic factor
BHLHE40	Basic helixloophelix family member e40
BMP4	Bone morphogenetic protein 4
CDKN2B	Cyclin dependent kinase inhibitor 2B
CEBPD	CCAAT/enhancer binding protein delta
CITED2	Cbp/p300 interacting transactivator with Glu/Asp rich carboxy-terminal domain 2
CLDN4	Claudin 4
CNKSR3	CNKSR family member 3
COL16A1	Collagen type XVI alpha 1 chain
COL27A1	Collagen type XXVII alpha 1 chain
COL7A1	Collagen type VII alpha 1 chain
COMP	Cartilage oligomeric matrix protein
CTGF	Connective tissue growth factor
CTPS1	CTP synthase 1
DACT1	Dishevelled binding antagonist of beta catenin 1

Table 4.3: Index of genes that are differentially expressed in TGF-β treated fibroblasts compared to controls. 2 of 5.

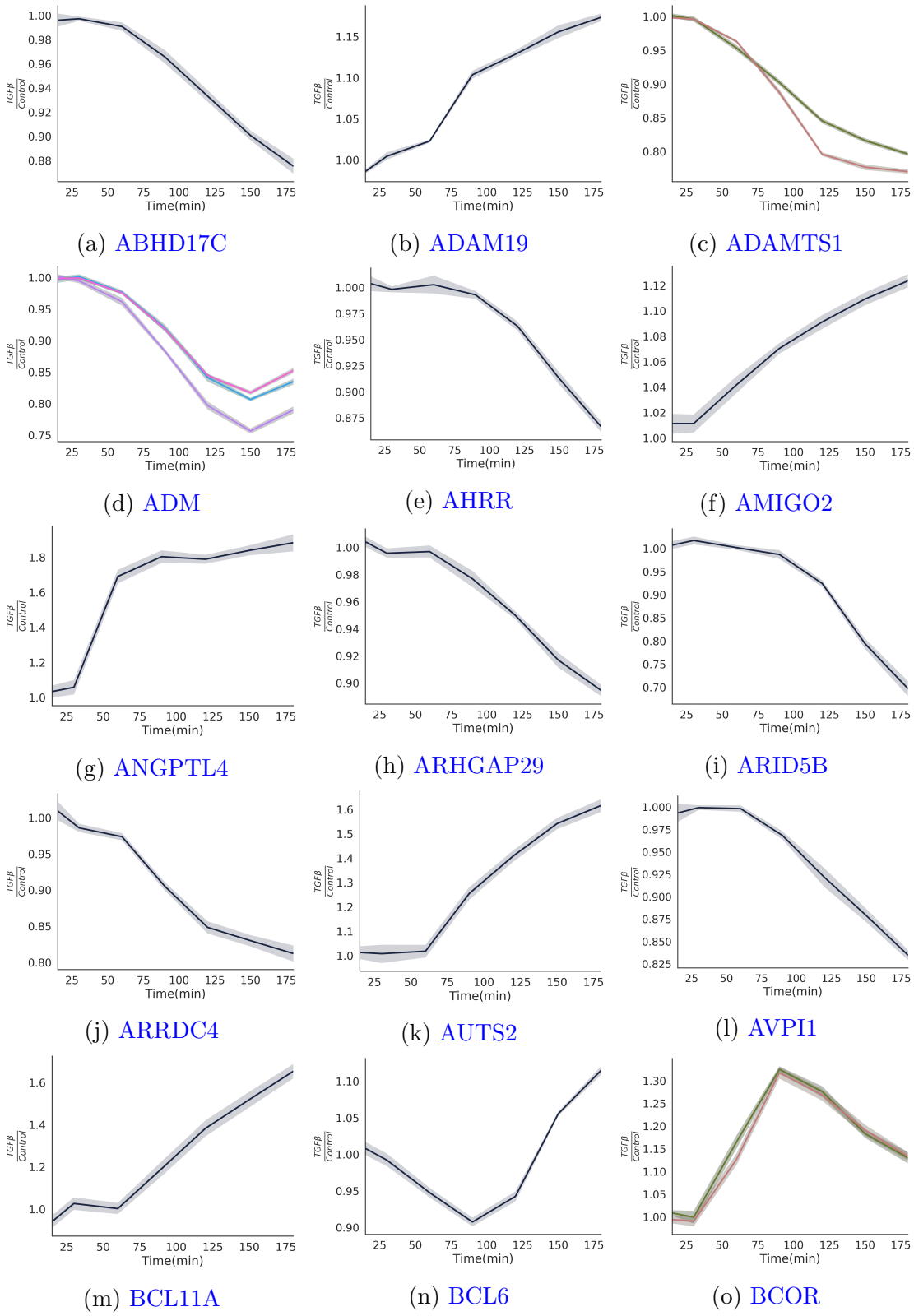
Gene Symbol	Protein
DUSP2	Dual specificity phosphatase 2
DUSP5	Dual specificity phosphatase 5
DUSP6	Dual specificity phosphatase 6
EDN1	Endothelin 1
EGR2	Early growth response 2
ELMSAN1	ELM2 and Myb/SANT domain containing 1
ENPP1	Ectonucleotide pyrophosphatase/phosphodiesterase 1
ESM1	Endothelial cell specific molecule 1
FAM46B	Family with sequence similarity 46 member B
FN1	Fibronectin 1
FOSB	FosB protooncogene, AP1 transcription factor subunit
FOXP1	Forkhead box P1
FZD7	Frizzled class receptor 7
GADD45B	Growth arrest and DNA damage inducible beta
GAL	Galanin and GMAP prepropeptide
GATA2	GATA binding protein 2
GCNT1	Glucosaminyl (Nacetyl) transferase 1, core 2
GEM	GTP binding protein overexpressed in skeletal muscle
GLI2	GLI family zinc finger 2
HBEGF	Heparin binding EGF like growth factor
HES1	Hes family bHLH transcription factor 1
HMOX1	Heme oxygenase 1
HOXA13	Homeobox A13
ID1	Inhibitor of DNA binding 1, HLH protein
ID2	Inhibitor of DNA binding 2, HLH protein
ID3	Inhibitor of DNA binding 3, HLH protein
IER3	Immediate early response 3
IFIT2	Interferon induced protein with tetratricopeptide repeats 2
IL11	Interleukin 11
IL6	Interleukin 6

Table 4.4: Index of genes that are differentially expressed in TGF-β treated fibroblasts compared to controls. 3 of 5.

Gene Symbol	Protein
JUNB	JunB protooncogene, AP1 transcription factor subunit
KANK4	KN motif and ankyrin repeat domains 4
KCTD11	Potassium channel tetramerization domain containing 11
KLF5	Kruppel like factor 5
KRTAP1-5	Keratin associated protein 15
LARP6	La ribonucleoprotein domain family member 6
LIF	Leukemia inhibitory factor
LMCD1	LIM and cysteine rich domains 1
LTBP2	Latent transforming growth factor beta binding protein 2
MURC	Muscle related coiledcoil protein
MYO10	Myosin X
NABP1	Nucleic acid binding protein 1
NEDD9	Neural precursor cell expressed, developmentally downregulated 9
NET1	Neuroepithelial cell transforming 1
NGF	Nerve growth factor
NIPAL4	NIPA like domain containing 4
NR2F2	Nuclear receptor subfamily 2 group F member 2
NUAK1	NUAK family kinase 1
OSR2	Oddskipped related transciption factor 2
PCDH18	Protocadherin 18
PDGFA	Platelet derived growth factor subunit A
PFKFB3	6phosphofructo2kinase/fructose2,6biphosphatase 3
PFKFB4	6phosphofructo2kinase/fructose2,6biphosphatase 4
PGM2L1	Phosphoglucomutase 2 like 1
PMEPA1	Prostate transmembrane protein, androgen induced 1
PRICKLE1	Prickle planar cell polarity protein 1
PRICKLE2	Prickle planar cell polarity protein 2
PRR5L	Proline rich 5 like
PTX3	Pentraxin 3
PXDC1	PX domain containing 1

Table 4.5: Index of genes that are differentially expressed in TGF-β treated fibroblasts compared to controls. 4 of 5.

Gene Symbol	Protein
RGS4	Regulator of Gprotein signaling 4
RHOB	Ras homolog family member B
RNF144B	Ring finger protein 144B
S1PR5	Sphingosine1phosphate receptor 5
SERPINE1	Serpin family E member 1
SERTAD2	SERTA domain containing 2
SGK223	Homolog of rat pragma of Rnd2
SH3PXD2A	SH3 and PX domains 2A
SIX1	SIX homeobox 1
SKIL	SKI like protooncogene
SLC19A2	Solute carrier family 19 member 2
SMAD7	SMAD family member 7
SOCS2	Suppressor of cytokine signaling 2
SOX4	SRYbox 4
STC2	Stanniocalcin 2
STK38L	Serine/threonine kinase 38 like
SUSD6	Sushi domain containing 6
TBX5	Tbox 5
TFAP2A	Transcription factor AP2 alpha
TFAP2C	Transcription factor AP2 gamma
TGFBI	Transforming growth factor beta induced
TNFAIP6	TNF alpha induced protein 6
TOB1	Transducer of ERBB2, 1
TPM1	Tropomyosin 1 (alpha)
TRIB1	Tribbles pseudokinase 1
TSHZ3	Teashirt zinc finger homeobox 3
TSPAN13	Tetraspanin 13
TSPAN2	Tetraspanin 2
TUFT1	Tuftelin 1
TWIST1	Twist family bHLH transcription factor 1



(a) ABHD17C (b) ADAM19 (c) ADAMTS1

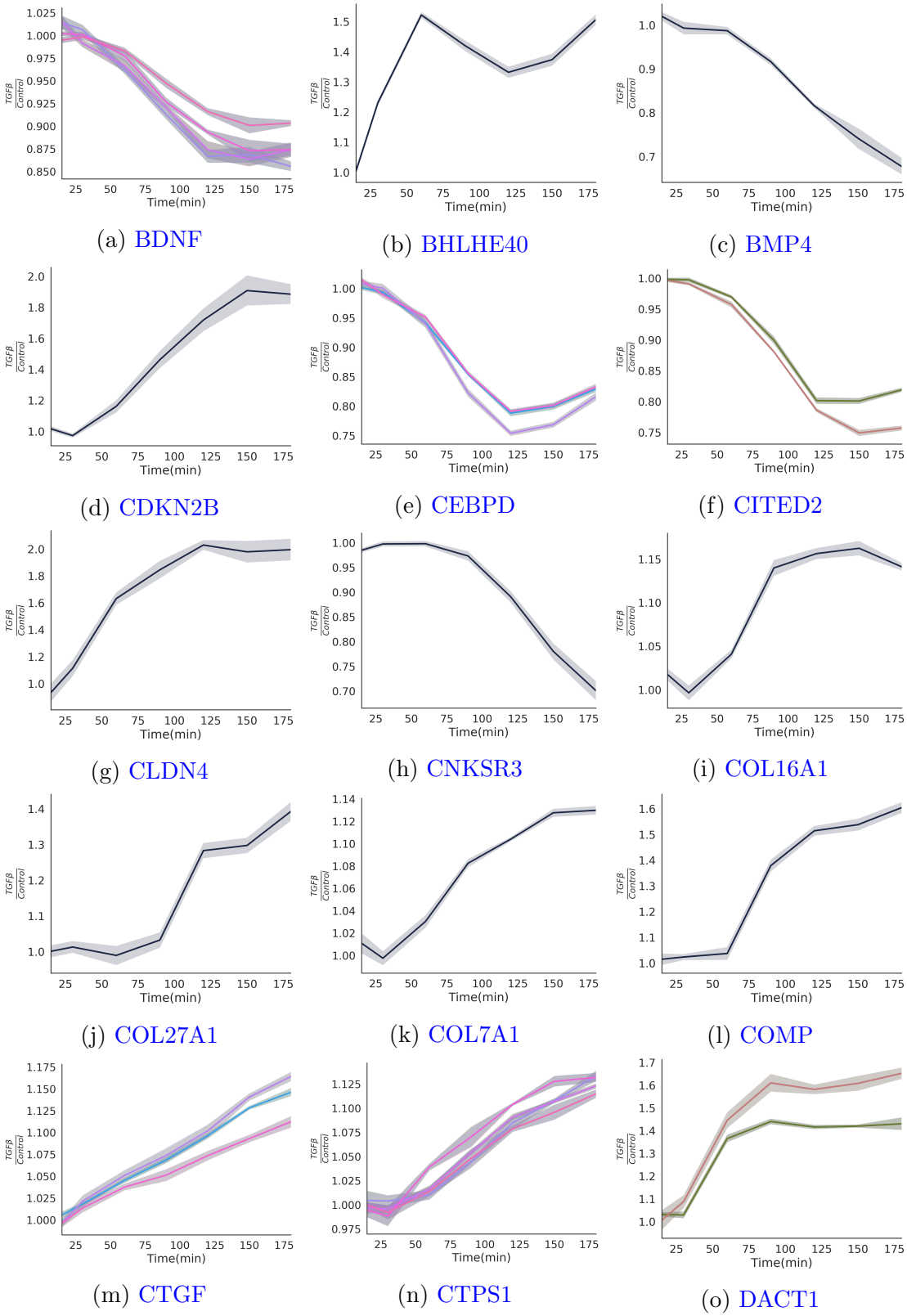
(d) ADM (e) AHRR (f) AMIGO2

(g) ANGPTL4 (h) ARHGAP29 (i) ARID5B

(j) ARRDC4 (k) AUTS2 (l) AVPI1

(m) BCL11A (n) BCL6 (o) BCOR

Figure 4.5: Dynamic profiles of genes that are differentially expressed in TGF-β treatment compared to control. Number 1 of 9.



(a) BDNF (b) BHLHE40 (c) BMP4

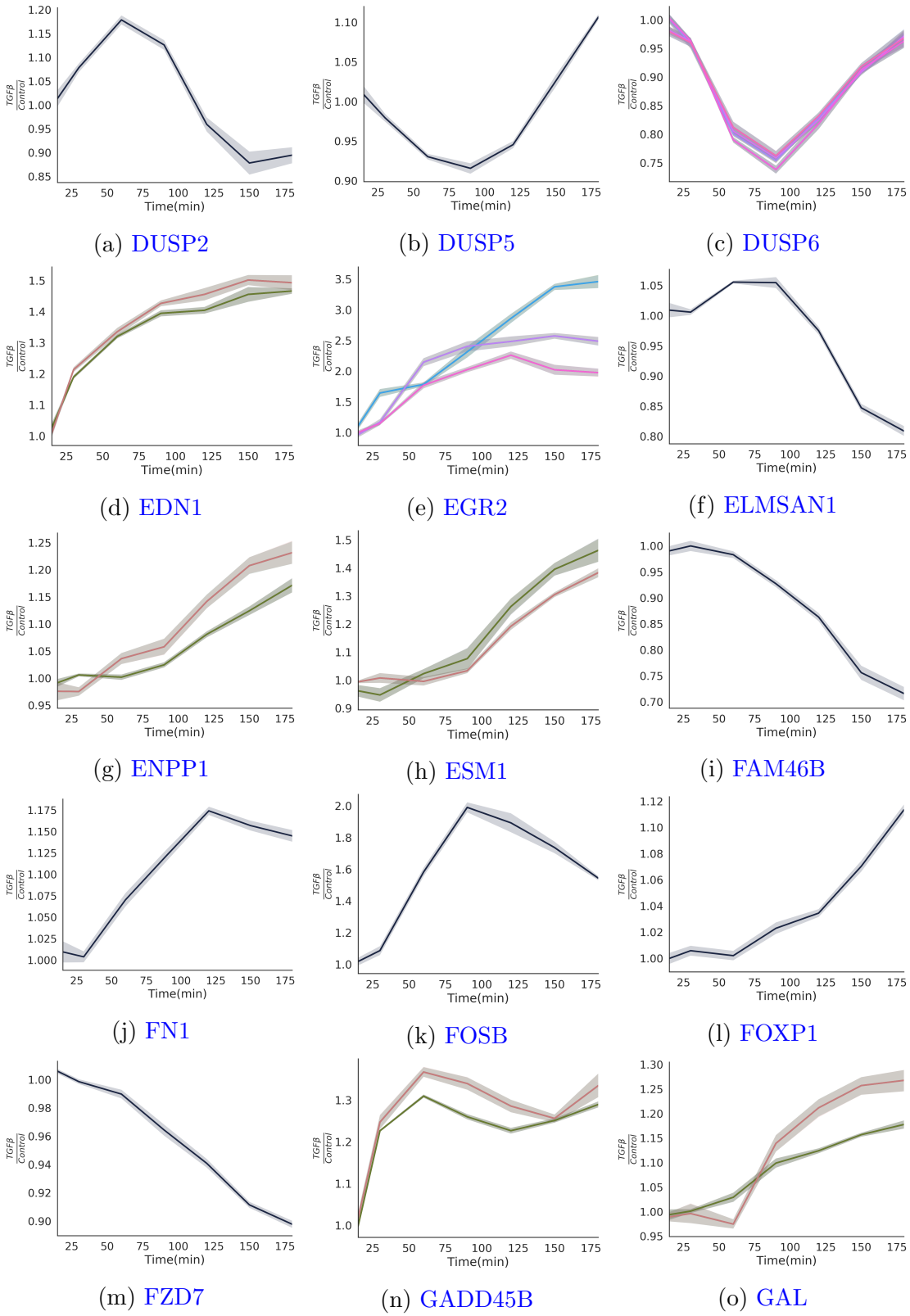
(d) CDKN2B (e) CEBPD (f) CITED2

(g) CLDN4 (h) CNKSR3 (i) COL16A1

(j) COL27A1 (k) COL7A1 (l) COMP

(m) CTGF (n) CTPS1 (o) DACT1

Figure 4.6: Dynamic profiles of genes that are differentially expressed in TGF-β treatment compared to control. Number 2 of 9.



(a) DUSP2 (b) DUSP5 (c) DUSP6

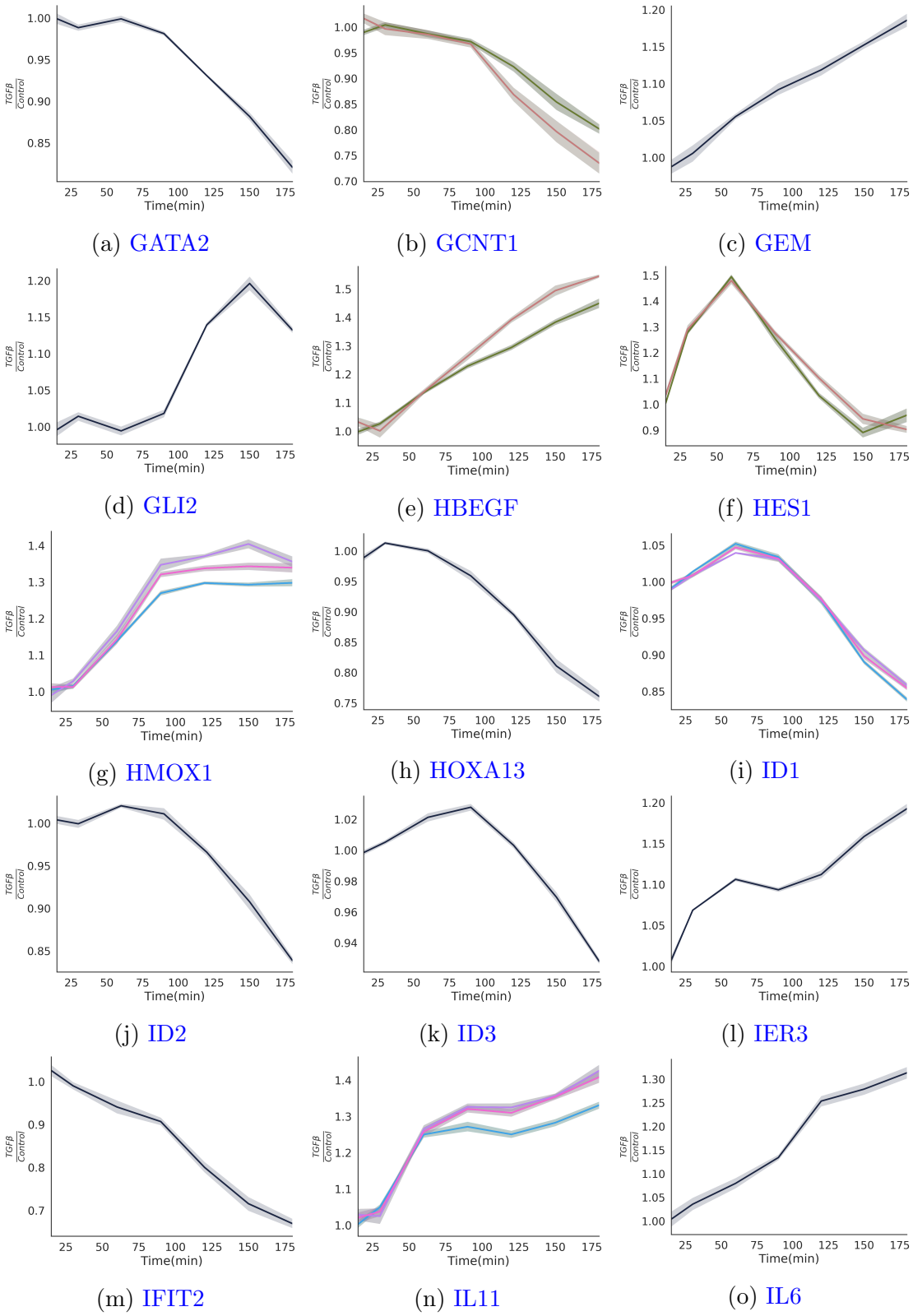
(d) EDN1 (e) EGR2 (f) ELMSAN1

(g) ENPP1 (h) ESM1 (i) FAM46B

(j) FN1 (k) FOSB (l) FOXP1

(m) FZD7 (n) GADD45B (o) GAL

Figure 4.7: Dynamic profiles of genes that are differentially expressed in TGF-β treatment compared to control. Number 3 of 9.



(a) GATA2 (b) GCNT1 (c) GEM

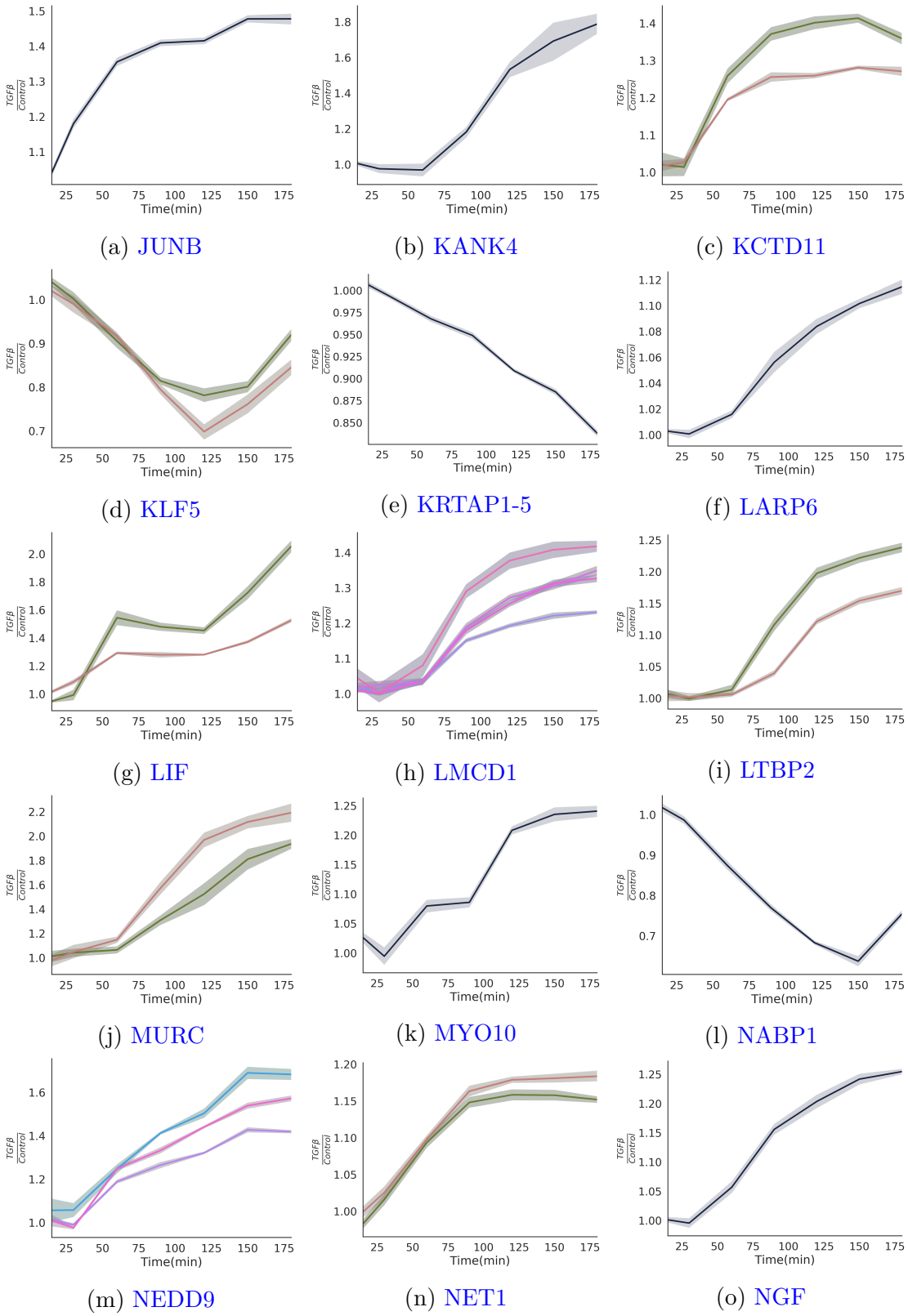
(d) GLI2 (e) HBEGF (f) HES1

(g) HMOX1 (h) HOXA13 (i) ID1

(j) ID2 (k) ID3 (l) IER3

(m) IFIT2 (n) IL11 (o) IL6

Figure 4.8: Dynamic profiles of genes that are differentially expressed in TGF-β treatment compared to control. Number 4 of 9.



(a) JUNB (b) KANK4 (c) KCTD11

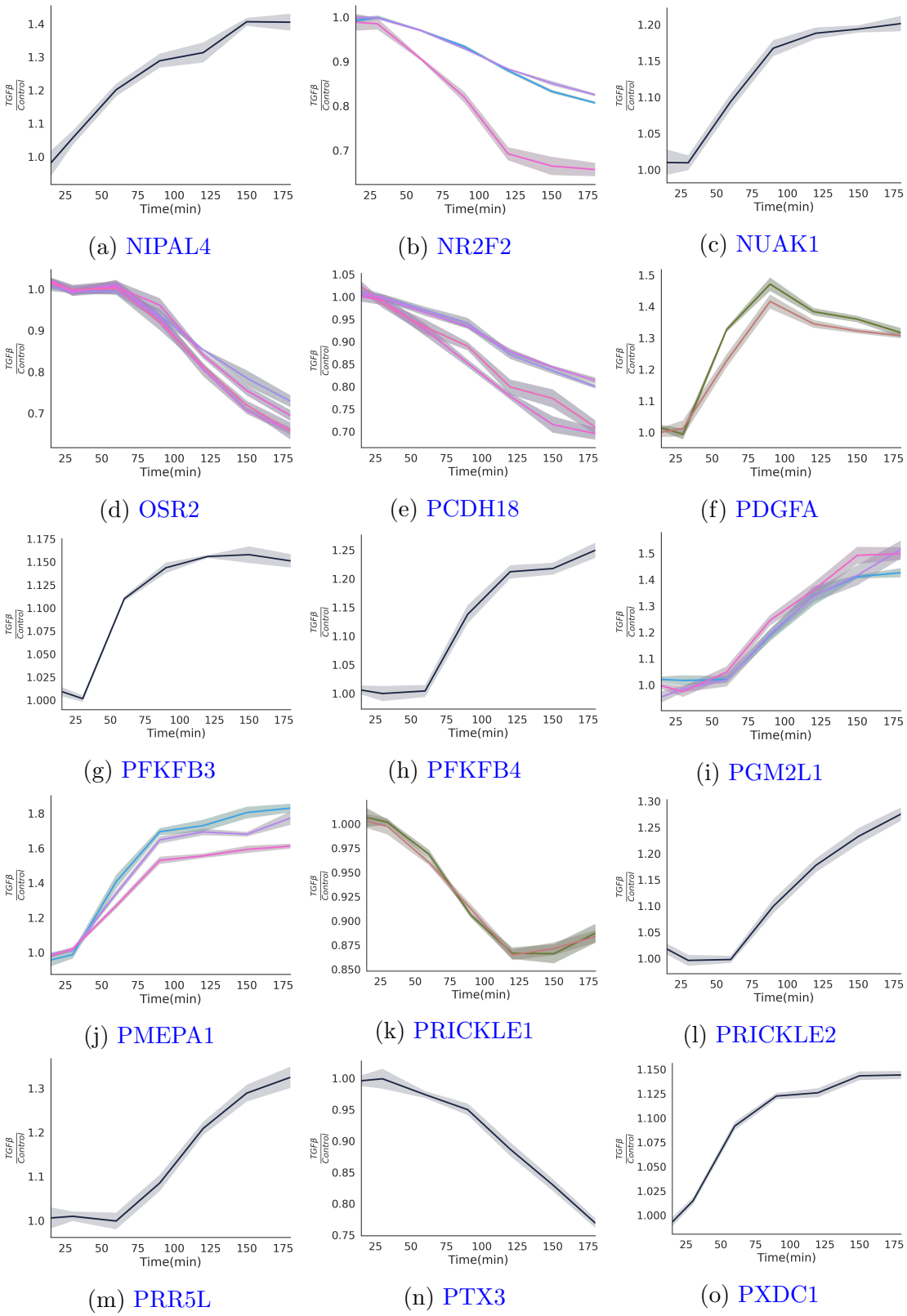
(d) KLF5 (e) KRTAP1-5 (f) LARP6

(g) LIF (h) LMCD1 (i) LTBP2

(j) MURC (k) MYO10 (l) NABP1

(m) NEDD9 (n) NET1 (o) NGF

Figure 4.9: Dynamic profiles of genes that are differentially expressed in TGF-β treatment compared to control. Number 5 of 9.



(a) NIPAL4 (b) NR2F2 (c) NUAK1

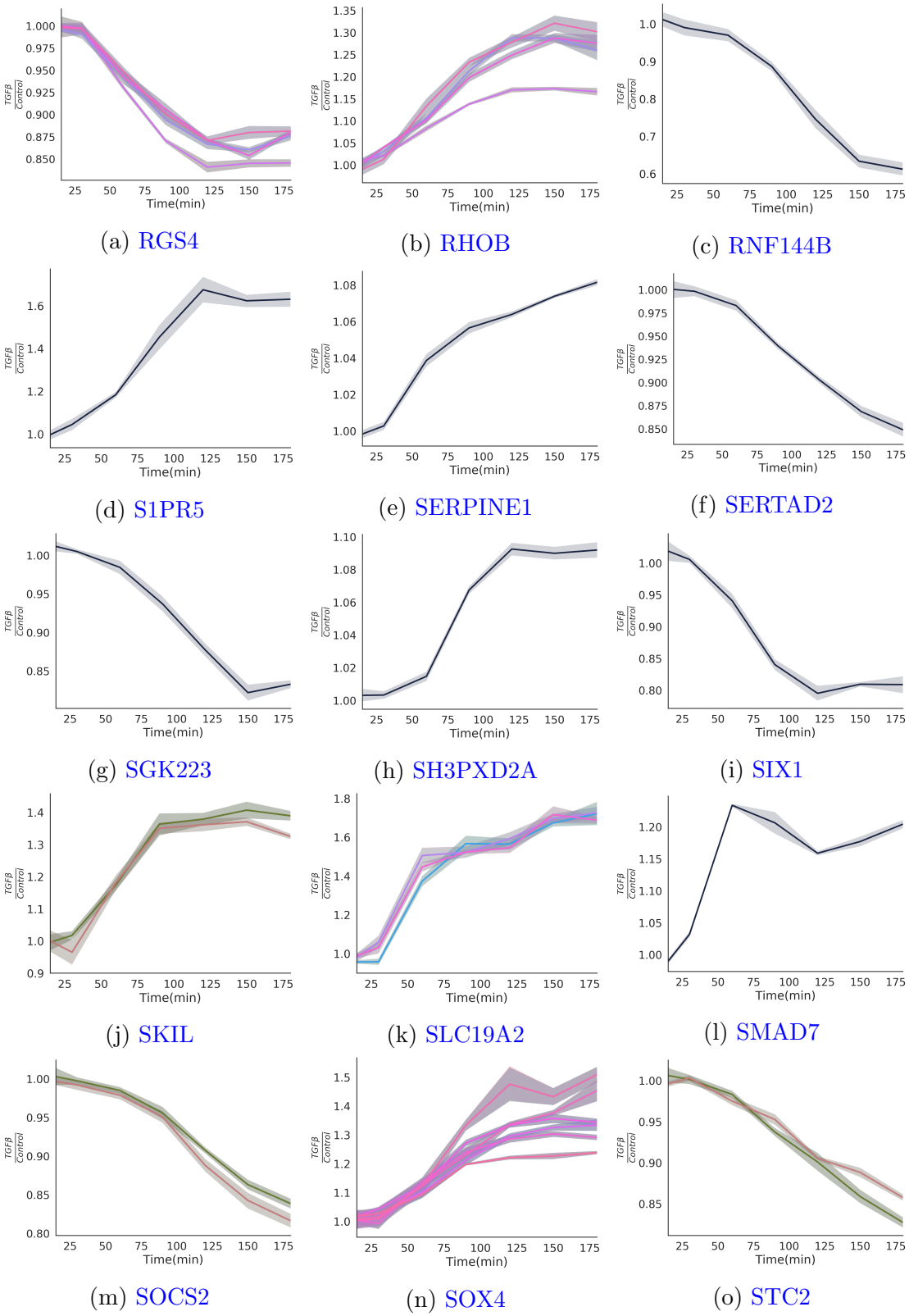
(d) OSR2 (e) PCDH18 (f) PDGFA

(g) PFKFB3 (h) PFKFB4 (i) PGM2L1

(j) PMEPA1 (k) PRICKLE1 (l) PRICKLE2

(m) PRR5L (n) PTX3 (o) PXDC1

Figure 4.10: Dynamic profiles of genes that are differentially expressed in TGF-β treatment compared to control. Number 6 of 9.



(a) RGS4 (b) RHOB (c) RNF144B

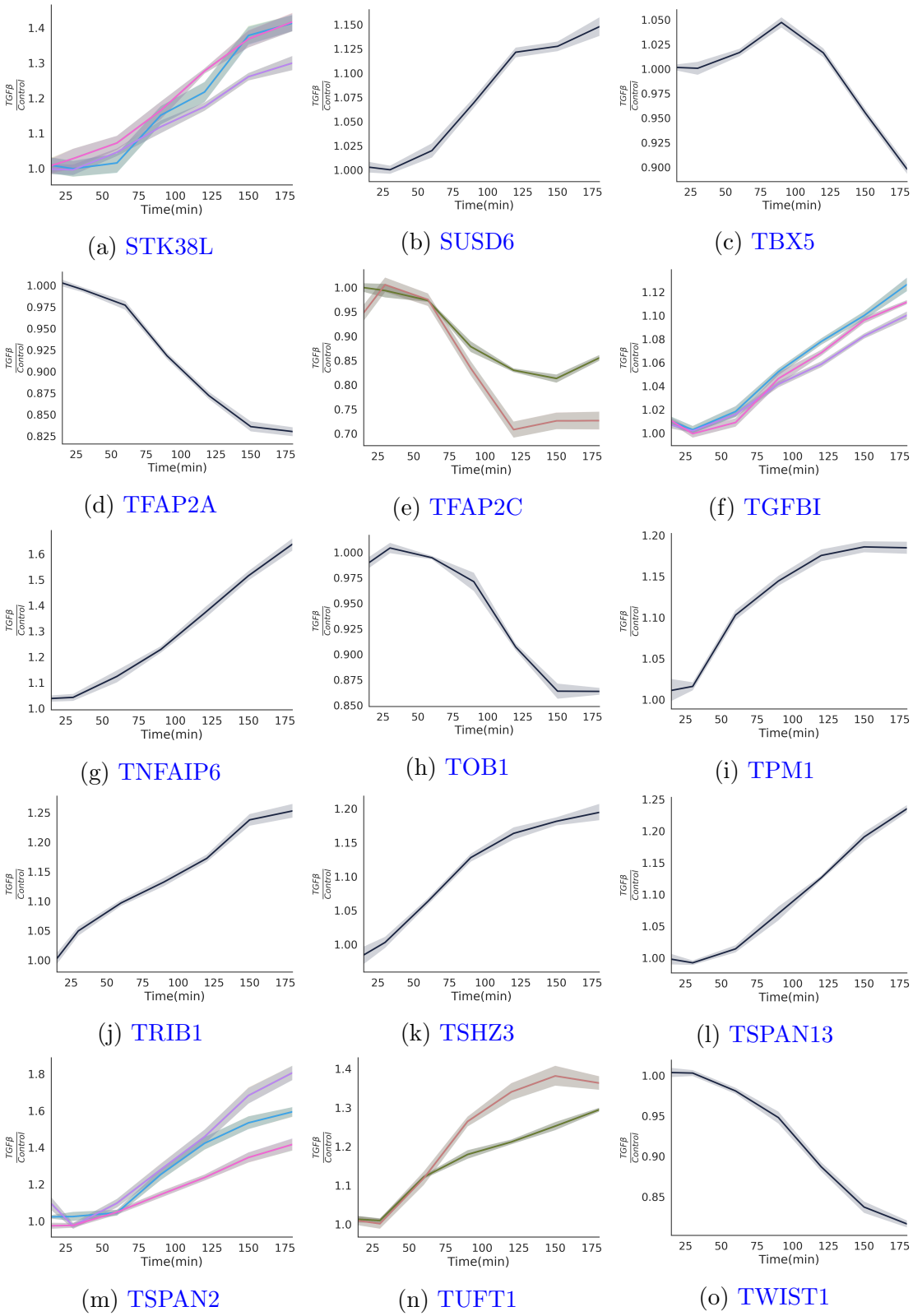
(d) S1PR5 (e) SERPINE1 (f) SERTAD2

(g) SGK223 (h) SH3PXD2A (i) SIX1

(j) SKIL (k) SLC19A2 (l) SMAD7

(m) SOCS2 (n) SOX4 (o) STC2

Figure 4.11: Dynamic profiles of genes that are differentially expressed in TGF-β treatment compared to control. Number 7 of 9.



(a) STK38L (b) SUSD6 (c) TBX5

(d) TFAP2A (e) TFAP2C (f) TGFBI

(g) TNFAIP6 (h) TOB1 (i) TPM1

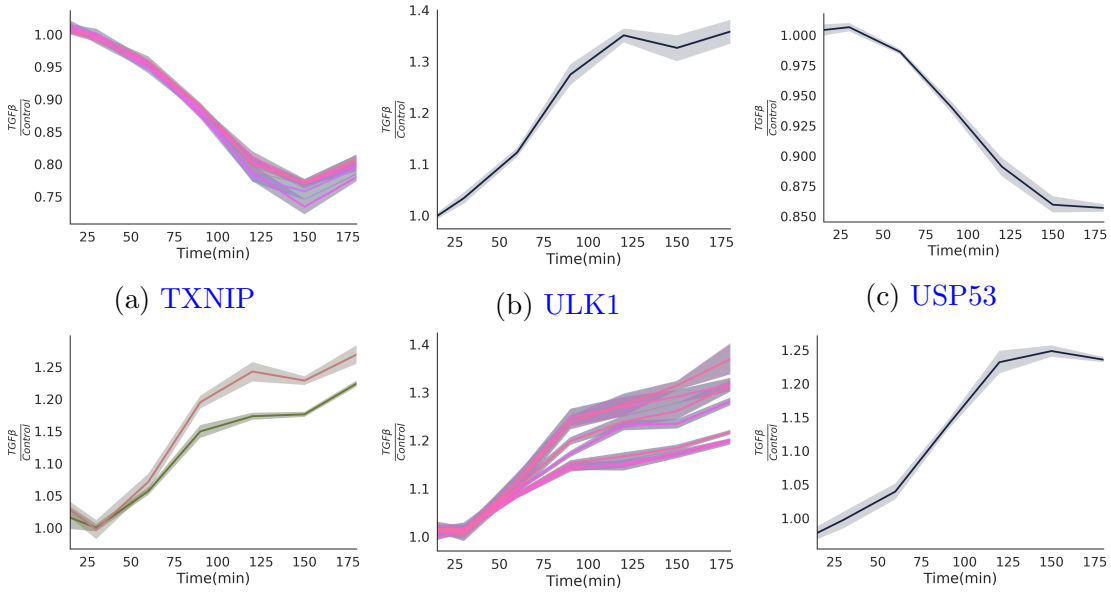
(j) TRIB1 (k) TSHZ3 (l) TSPAN13

(m) TSPAN2 (n) TUFT1 (o) TWIST1

Figure 4.12: Dynamic profiles of genes that are differentially expressed in TGF-β treatment compared to control. Number 8 of 9.

Table 4.6: Index of genes that are differentially expressed in TGF-β treated fibroblasts compared to controls. Number 5 of 5.

Gene Symbol	Protein
TXNIP	Thioredoxin interacting protein
ULK1	Unc51 like autophagy activating kinase 1
USP53	Ubiquitin specific peptidase 53
VDR	Vitamin D (1,25 dihydroxyvitamin D3) receptor
VEGFA	Vascular endothelial growth factor A
ZNF365	Zinc finger protein 365



(a) TXNIP (b) ULK1 (c) USP53

(d) VDR (e) VEGFA (f) ZNF365

Figure 4.13: Dynamic profiles of genes that are differentially expressed in TGF-β treatment compared to control. Number 9 of 9.

4.3.3 Pathway Analysis

Differentially expressed genes were subject to a pathway enrichment analysis, using the DAVID (Dennis et al. 2003) interface to KEGG and Reactome. Both the Reactome and KEGG based analyses identified components of the TGF-β signalling system as significantly enriched in the 126 differentially expressed genes. KEGG also identified genes involved in PI3K, FOXO, TNF-α, Hippo, MAPK and HIF signalling (Table 4.7). After correction for multiple hypothesis testing, only PI3K, FOXO and TNF-α were significantly enriched. In addition to those shown in the KEGG analysis, the Reactome enrichment analysis found pathways leading to MAPK1/3 activation, negative regulation of MAPK and signalling by BMP as enriched terms in the list of differentially expressed genes (Table 4.8).

Table 4.7: KEGG enrichment results from DAVID (version 6.8).

Name	Gene Symbol	PValue	Benjamini	FDR
hsa04350:TGF-beta signaling pathway	BMP4, CDKN2B, ID2, SMAD7, ID1, ID3	4.67E-04	0.0513	0.529
hsa04390:Hippo signaling pathway	BMP4, ID2, ID1, CTGF, SERPINE1, GLI2, FZD7	0.001	0.0396	1.21
hsa04066:HIF-1 signaling pathway	IL6, PFKFB3, EDN1, VEGFA, SERPINE1	0.007	0.18	7.66
hsa04010:MAPK signaling pathway	DUSP5, BDNF, DUSP2, PDGFA, GADD45B, NGF, DUSP6	0.0139	0.272	14.74
hsa04668:TNF signaling pathway	LIF, IL6, EDN1, JUNB	0.0495	0.512	43.83
hsa04151:PI3K-Akt signaling pathway	IL6, PDGFA, COL27A1, COMP, VEGFA, FN1, NGF	0.051	0.482	44.8

4.3.4 Time Series Clustering

Classifying genes using ontologies is one method of grouping genes. As an alternative a time series clustering algorithm was used to group genes based on their dynamic profiles. This algorithm is part of the `pytseries` package and was discussed at length in Chapter 3. Prior to use, the data was linearly interpolated so that the algorithm has 30

Table 4.8: Reactome enrichment results from DAVID (version 6.8).

Name	Gene Symbol	PValue	Benjamini	FDR
SMAD2/SMAD3:SMAD4 heterotrimer regulates transcription	CDKN2B, SMAD7, SERPINE1, JUNB	0.0019	0.228	2.167
Transcriptional regulation of white adipocyte differentiation	KLF5, EGR2, CEBPD, NR2F2, ANGPTL4	0.0032	0.197	3.657
RAF-independent MAPK1/3 activation	DUSP5, DUSP2, DUSP6	0.0043	0.182	4.986
Integrin cell surface interactions	COL7A1, COMP, COL16A1, FN1	0.0285	0.634	28.904
Negative regulation of MAPK pathway	DUSP5, DUSP2, DUSP6	0.0345	0.623	33.875
Molecules associated with elastic fibres	BMP4, LTBP2, FN1	0.0325	0.628	33.9
Regulation of gene expression by Hypoxia-inducible Factor	VEGFA, CITED2	0.0745	0.834	59.866
Collagen biosynthesis and modifying enzymes	COL7A1, COL27A1, COL16A1	0.0942	0.821	68.835

time points to work with, rather than 7. The data were normalized using the 'minmax' method (Equation 4.1) prior to clustering

$$ts_{minmax} = \frac{ts_t - min(ts)}{max(ts) - min(ts)} \tag{4.1}$$

where t indexes time, ts is the time series data before normalisation and ts_{minmax} is the normalised time series. The time series K-means algorithm requires the number of clusters be chosen before running the algorithm. To decide on this parameter, the clustering procedure was repeated 30 times for increasing values of K (Figure 4.14). The steepest change in gradient occurs at $K = 7$ clusters and so this was the value of the K parameter used in subsequent analysis. This method is known as the 'elbow' method for choosing K and represents a rule of thumb when using the K-means algorithm (Bholowalia & Kumar 2014).

The clusters resulting from applying the DTW K-means algorithm to the current dataset are shown in Figure 4.15. The DTW K-means clustering algorithm has performed reasonably well with regards to its ability to group time series objects that

are similar in shape. In general the clustering procedure has found a transiently increasing group Figure 4.15a, a hyperbolic decreasing curve Figure 4.15b, two sigmoidal increasing curves Figure 4.15c and Figure 4.15c, a transiently decreasing curve Figure 4.15e, a hyperbolic increasing curve Figure 4.15f and a sigmoidal decreasing curve Figure 4.15g. However, despite the clusters being qualitatively similar, the gene lists they produced were seemingly random. Moreover, individual enrichment analyses on the separate clusters did not provide any more biological interpretation than when the analysis was conducted on the whole list. For this reason the individual enrichment analyses were not included in this chapter.

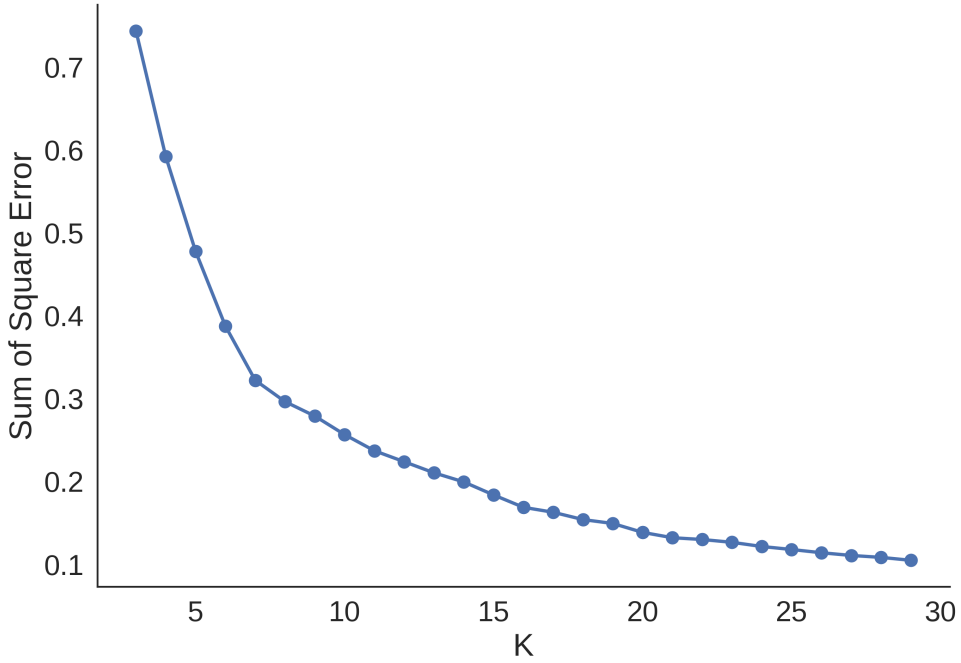
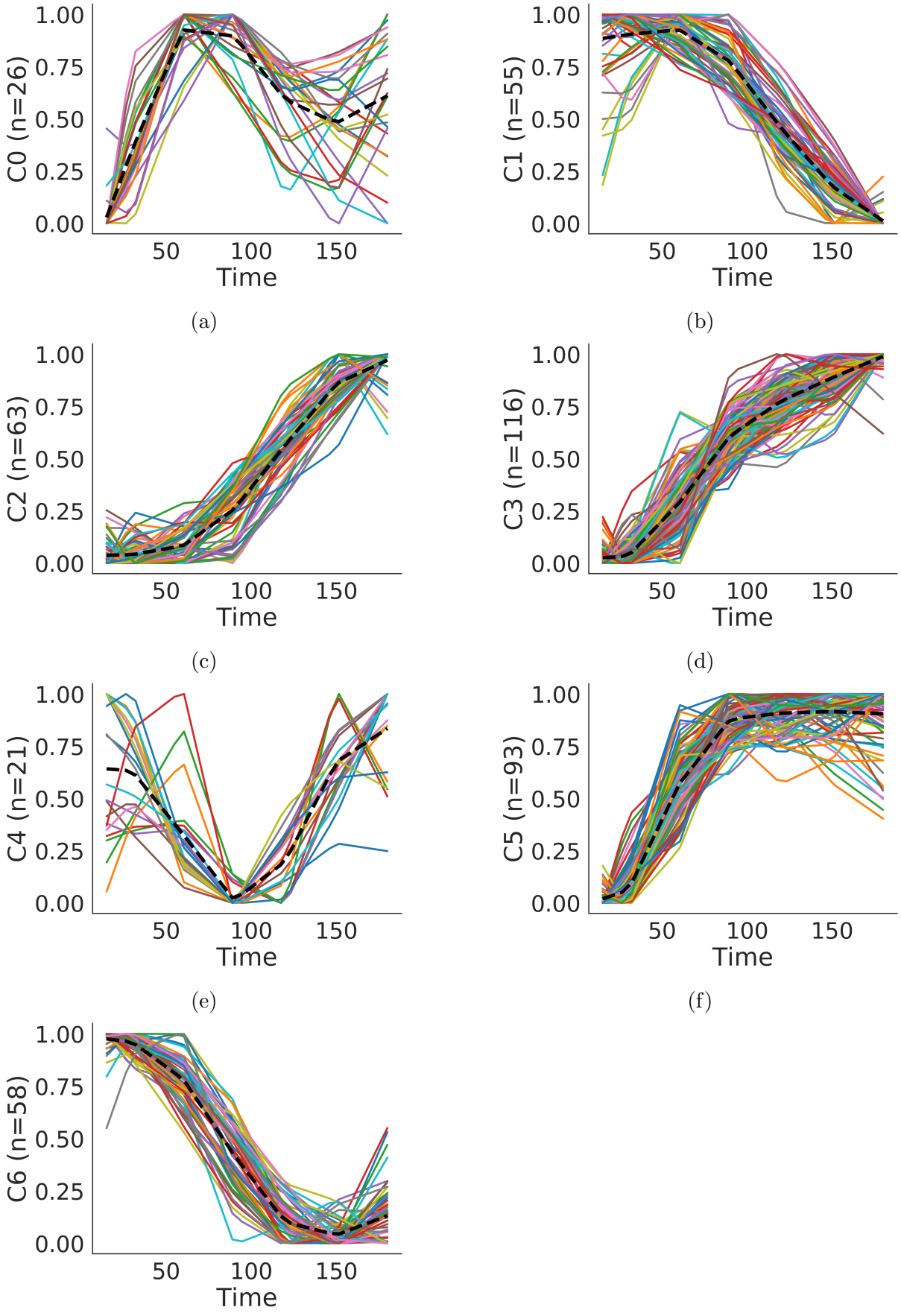


Figure 4.14: The 'elbow' method. The sum of squared errors is plotted as a function of K. K is chosen as elbow of the characteristic metaphorical arm produced by this plot.

4.4 Discussion

TGF-β signalling is complicated by its close communications with a diverse set of proteins that are known to also function in other pathways. This ability for cross-talk is thought to confer diverse biological functions to TGF-β signalling, depending on cellular context (Cellire et al. 2011, Zhang 2018). Examples such as in collagen transcription



(a) (b) (c) (d) (e) (f) (g)

Figure 4.15: Clustering of differentially expressed genes. Normalized and interpolated time series data were clustered using the K-means algorithm with $k = 7$. The non-normalized data for the resulting clusters are shown (a-g).

provide evidence that TGF-β does not only use Smad second messengers, but has a large variety of other proteins available for cooperation in the regulation of TGF-β mediated transcription (Quan et al. 2010, Asano et al. 2009, Makino et al. 2017).

In this chapter, the fibroblast transcriptional response to TGF-β was measured using an Affymetrix microarray platform. Affymetrix microarray technology enables massive parallel measurements of gene expression patterns in a sample. Because most of the human genome is included on an Affymetrix gene chip, this technology enables a global analysis of a cell's expression profile in response to a certain condition when compared to an appropriate control (Lockhart et al. 1996, Lockhart & Winzeler 2000).

Using the Affymetrix GeneTitan technology, we have profiled the early fibroblast response to TGF-β over time and identified 443 probes corresponding to 127 genes that were differentially expressed in TGF-β treated fibroblasts over time, compared to the controls (Tables 4.2 to 4.2 and Figures 4.5 to 4.5).

Many of the genes that responded to TGF-β are already known to respond to TGF-β. For instance, PAI1 is well known to be induced by TGF-β, a process thought to involve Mek/Erk (Kutz et al. 2001) and HIFα (Ueno et al. 2011), both of which were flagged by the enrichment analysis (Table 4.7). Moreover, RhoB (Figure 4.11b) is reportedly induced by TGF-β and this process involves Mek/Erk and Smad3 (Vasilaki et al. 2010). SKIL (Deheuninck & Luo 2009) and Smad7 (Nakao et al. 1997, Kavsak et al. 2000) are known negative feedback regulators of the Smad system. Together with the quality control (Figure 4.4), these consistencies with the literature provide confidence and offer good quality insight into the young fibroblast response to TGF-β.

Despite some ties with literature, deriving biological understanding from long lists of genes is difficult using a literature search alone and so two alternative methods were used in attempt to gain insight into which pathways transcriptionally respond to TGF-β. The first was a pathway enrichment analysis on the list of differentially expressed genes in order to ascertain whether known pathways were overrepresented in the list. This analysis was conducted twice with two different databases, KEGG (Table 4.7) and Reactome (Table 4.8). As expected, both analyses identified Smad

signalling as overrepresented. The KEGG analysis additionally identified Hippo, HIF1, MAPK, TNF-α and PI3K signalling though, only Hippo signalling remained overrepresented on statistical grounds after multiple hypothesis correction. The Reactome analysis identified a differentiation pathway; MAPK activation and negative regulation; HIF1 signalling and two terms relating to ECM integrity. These results corroborate the promiscuous nature of the TGF-β system in its capacity for cross-talk with other signalling pathways.

One of the limitations of enrichment analysis is that they operate under the assumption that a term is important only when a certain number of genes are annotated to that term. When there are many genes are annotated to a term this assumption is reasonable, since a term annotated with many genes is quite likely overrepresented within a gene list. However, when few genes are annotated to a term, it is unknown what impact the products of such genes have on a biological process. For instance, control points in biological networks (such as feedbacks) may have more of an impact than other proteins, such as a structural component of the cytoskeleton. Moreover, the information contained in databases are not necessarily complete. These limitations make it important not to unequivocally rely on the results. Therefore the information presented in Table 4.7 and Table 4.8 should be considered carefully and used as a starting point for directing further investigation. To demonstrate this point, the collagen biosynthesis Reactome term was not significantly enriched in the list of differentially expressed genes, but this does not necessarily mean that fibroblasts are not producing proteins that are involved in collagen biosynthesis. In fact, since fibroblasts respond to TGF-β by producing collagen (Varani et al. 2006), it is highly likely that fibroblasts produce proteins that facilitate the production of collagen.

Another, more fundamental, limitation to consider is that transcriptional data can only provide a limited amount of information regarding signalling cross talk because the signalling occurs at a different level of biological organisation. Transcript measurements can to a limited extent be used as a surrogate for understanding what is present at the protein level, but it is not known whether there is a one-to-one correspondence between amounts of mRNA and protein (Liu et al. 2016). Moreover, transcript data does not

provide any information regarding post-translational modifications which are essential for understanding the control of biochemical signalling pathways. Therefore, proteomic experiments equivalent to the experiment presented here would be beneficial for understanding how TGF-β interacts with other signalling systems.

Since the pathway enrichment analysis does not make use of the dynamic aspect of the data, an alternative approach was devised to cluster the data based on the shape of the dynamic profiles. Specifically, the differentially expressed genes were grouped using the time series K-means algorithm that was discussed in Chapter 3. This algorithm is an adaptation of the K-means algorithm that uses the dynamic time warping algorithm to find the optimum alignment of the time index prior to applying the K-means clustering algorithm. The underlying assumption with this approach is that genes with a similar dynamic profile are coregulated and may perform a similar function. Identifying genes which behave similarly may therefore help identify the components of cellular function.

The purpose of the time series K-means algorithm is to identify groups of similar time series. In this respect the algorithm has performed well but surprisingly, the clusters were less amenable to biological interpretation than anticipated. The question arises of why the algorithm did not provide biological insight. The first reason is that perhaps the underlying assumption about coregulated genes having a similar profile is wrong. Another scenario is that the K-means algorithm may require some tuning before it performs well with a dataset such as this. The tuning parameters include the choice of the number of clusters K; whether to normalise the time series data prior to clustering; if so which method is best and whether to interpolate the time series prior to clustering and if so, how many time points is best? These issues have not been resolved in this thesis and present challenges for future work

This chapter has discussed the analysis of the a genome wide dataset collected from neonatal fibroblasts in response to TGF-β. An emphasis was placed on identifying TGF-β responsive proteins that are also known to be involved in other pathways. Collectively, 126 genes responded differently to TGF-β in comparison to control and a pathway enrichment analysis was used to identify which of these genes are known to

involved in other signalling pathways. Extrapolating from the example provided by TGF-β, an interesting notion arises regarding the nature of how biochemical networks are perceived. Historically, biochemists have considered and classified different signalling pathways as separate entities. However it is important to remember that signalling pathways are artificial constructs that we use to help us comprehend the complexity of biochemical networks. On some level, this may have hindered our thinking since a more contemporary viewpoint, exemplified by the fibroblast response to TGF-β, is that all biochemical pathways are connected to a degree.

Chapter 5

Neonatal, senescent and adult fibroblast transcriptional dynamics in response to TGF-β

5.1 Introduction

Skin provides a protective barrier between an organism and the outside world. Structurally, skin is a complex and heterogeneous tissue composed of the epidermal and dermal compartments. The epidermis performs the barrier function while the dermis provides structural integrity and nutrients to the epidermis. When skin ages a variety of changes occur in its composition including wrinkling and laxity, atrophy, reduced elasticity and reduced mechanical strength (see Chapter 1).

A major aspect of skin ageing is how collagen networks change over time. It is well established that aged skin has reduced collagen content (Varani et al. 2006) but aged skin is also highly fragmented and disorganised. The damaged state of the aged dermis predisposes the it more damage by reducing the ability of fibroblasts to adhere to the ECM. The reduced mechanical tension makes fibroblasts become globular and detached from the ECM and this impacts their ability to maintain the dermal ECM and feeds a detrimental cycle into impaired skin integrity (Cole et al. 2018).

The critical aspect of this process is reduced collagen levels. Healthy fibroblast physiology relies on being able to adhere to components of the ECM such as collagens and fibronectin. Research to date indicates that reduced collagen levels are the combined result of decreased collagen production (Varani et al. 2006) and increased collagen degradation by matrix metalloproteases (MMPs) (Quan et al. 2009).

TGF-β is well-known to be a primary inducer of collagen production (Varga et al. 1987). Therefore, it is prudent to consider processes managed by TGF-β as candidates for those which change with age. Changes in TGF-β signalling has been associated with both fibrotic disorders (Asano et al. 2005) and skin ageing (Quan et al. 2004), which in some respects can be considered opposites. Moreover, TGF-β has been shown to induce the expression of a host of other ECM proteins and inhibit the production of others, such as MMPs (Qin et al. 2017).

In this chapter, a high throughput time series experiment was designed to provide insight into the differences between neonatal and adult fibroblasts (56yo+), and the differences between their response to TGF-β. Moreover since replicative cellular senescence is another important aspect of ageing, the same investigation was repeated with irradiation induced-senescent fibroblasts. Collectively, the activities of 68 genes were measured over 96h in response to 5ng mL^{-1} TGF-β or negative control in 3 donors apiece for neonatal, irradiation-induced senescent and adult fibroblasts. By comparing how these fibroblasts behave under these various conditions, we provide insight into the differences between young and old fibroblast biology. This information contributes towards our general understanding of skin ageing.

5.2 Methods

5.2.1 Cell Culture

Three independent human neonatal dermal fibroblasts (HDFn) labelled A (Caucasian male, catalogue number: C-004-5C, lot number: #1366434), B (Caucasian male, catalogue number: C-004-5C, lot number: #1366356) and C (Caucasian male, catalogue number: C-004-5C, lot number: #1206197); three irradiation-induced senescent cell lines (D, E and, F) which are the same as A, B and C respectively but irradiated with 20Gy ten days prior to seeding, and three adult cell lines G (55 years old Caucasian female, catalogue number: C-013-5C, lot number: #1528526), H (60 year old, Caucasian male, catalogue number: A11634, lot number: #1090465) and I (65 year old, Caucasian female, catalogue number: A11636, lot number: #200910-901) were purchased from Life Technologies and seeded into standard tissue culture treated 12-well dishes at a density of 10,000 cells/cm^2 in 3.5ml M106 medium (ThermoFisher Scientific, catalogue number: M106500) supplemented with LSSG (ThermoFisher Scientific, catalogue number: S00310) and at 27°C, 5%CO_2 for 4 days. Senescent cells were seeded at higher density of 65,000 cells/cm^2.

5.2.2 Treatment Protocol

Cells from each cell line were serum starved 24h prior to treatment by removing LSGS supplementation from media. Following 24h of incubation at 37°C and 5% CO_2, cells were assigned one of three treatments: baseline, control or TGF-β. Baseline samples were not treated in any way prior to harvesting at experiment start (0h) and end (96h). TGF-β and control samples were treated with media containing 5ng mL^{-1} TGF-β reconstituted in 10mM citric acid or 0.1% BSA in 10mM citric acid respectively. In control and TGF-β groups, cells were harvested at 0.5, 1, 2, 3, 4, 8, 12, 24, 48, 72, 96 hours post treatment. All 216 conditions were repeated 6 times resulting in 1296 individual samples. Samples were shipped to Procter and Gamble (P&G), Cincinnati for

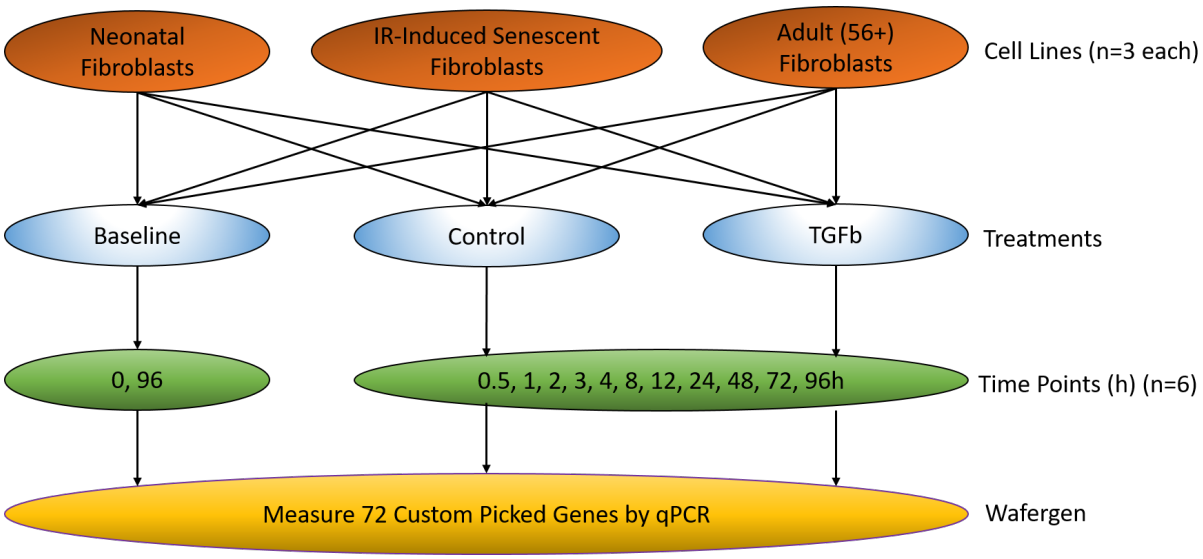


Figure 5.1: High-throughput qPCR experimental design. Three neonatal, irradiation-induced senescent neonatal and adult fibroblasts were treated with TGF-β1, BSA or left untreated (baseline) for indicated times. All samples were processed by high throughput qPCR technology by WaferGen Biosystems enabling the measurement of 72 genes in each of the 1296 individual samples.

quantification on a high throughput PCR Smart chip platform by WaferGen. The experiment design is described in Figure 5.1.

5.2.3 Quality Control

Principal component analysis (PCA) was conducted in Python using scikit-learn's implementation of PCA on the the raw C_T values. PCA cannot handle missing data so any gene which had > 5 missing data points were excluded from the PCA and any profiles with $0 > x > 6$ missing profiles were imputed using the median of the other time points for that gene. However, the number of data points that required imputation was low.

5.2.4 Normalization

The $2^{-\Delta\Delta C_T}$ method of qPCR normalization is a relative technique described by Livak & Schmittgen 2001 and has been used extensively by molecular biologists. The method relies on normalization of a measurement to a housekeeping gene and a calibrator

sample. The following describes how the equation can be derived.

Exponential amplification of mRNA in a polymerase chain reaction can be described as:

$$X_n = X_0 \cdot (1 + E_X)^n \tag{5.1}$$

where:

X_n = Number of target molecules at cycle n

X_0 = Initial number of target molecules

E_X = Amplification efficiency of target amplification

n = Number of Cycles

The cycle threshold C_T for target X, $C_{T,X}$ is the `number of cycles` it takes for fluorescence emitted from target X to reach a predefined threshold T. Therefore, substituting $C_{T,X}$ into Equation 5.1 gives

$$X_T = X_0 \cdot (1 + E_X)^{C_{T,X}} = K_X \tag{5.2}$$

where K_X is a constant. The $2^{-\Delta\Delta C_T}$ method of PCR quantification is a relative method that depends on the parallel measurement of a reference gene in the same sample. The reference gene is assumed to be constant throughout the experiment and should be chosen accordingly. Similar to Equation 5.2, the number of cycles it takes for the fluorescence emitted from the reference probe to reach threshold T can be described by

$$R_T = R_0 \cdot (1 + E_R)^{C_{T,R}} = K_R \tag{5.3}$$

where the variables are analogous to the variables in Equation 5.2. By taking the ratio

$$\frac{X_T}{R_T} = \frac{X_0 \cdot (1 + E_x)^{C_{T,X}}}{R_0 \cdot (1 + E_R)^{C_{T,R}}} = \frac{K_X}{K_R} = K \tag{5.4}$$

and assuming the PCR amplification efficiencies are equal for target X_T and reference R_T

$$E_X = E_R = E \tag{5.5}$$

Equation 5.4 can be simplified to

$$\frac{X_T}{R_T} = \frac{X_0}{R_0} \cdot (1+E)^{C_{T,X}-C_{T,R}} = K \tag{5.6}$$

Substituting $X_N = \frac{X_0}{R_0}$ and $C_{T,X} - C_{T,R} = \Delta C_T$ into Equation 5.6 gives

$$X_N \cdot (1+E)^{\Delta C_T} = K \tag{5.7}$$

and rearranging gives

$$X_N = K \cdot (1+E)^{-\Delta C_T} \tag{5.8}$$

The ratio $K = \frac{K_X}{K_R}$ is an unknown quantity but by dividing Equation 5.8 for a sample of interest $X_{N,q}$ by a 'calibrator sample' $X_{N,cb}$ and assuming a PCR efficiency of 1, which is reasonable for well optimized PCR conditions (Livak & Schmittgen 2001), the K's cancel out.

$$\frac{X_{N,q}}{X_{N,cb}} = \frac{K \cdot (1+E)^{-\Delta C_{T,q}}}{K \cdot (1+E)^{-\Delta C_{T,cb}}} = 2^{-(\Delta C_{T,q} - \Delta C_{T,cb})} = 2^{-\Delta\Delta C_T} \tag{5.9}$$

PPIA, the gene that encodes for Peptidylprolyl Isomerase A, was determined to be stationary throughout the experimental conditions and was used as a reference gene for all samples. Multiple choices exist for the calibrating normalisation. Instead of choosing between them, which highlights one feature but loses information in another dimension of the data, the data were displayed without performing this secondary normalisation (Equation 5.9). Without the calibration samples, the $2^{-\Delta C_T}$ values (Equation 5.8) are still proportional to the amount of the target in sample.

5.2.5 Differential Expression

The R package, LIMMA (Smyth 2005, Ritchie et al. 2015) was used to bootstrap a differential expression analysis. Specifically, LIMMA was configured using a spline matrix for time series following the instructions in the LIMMA user guide and 4 degrees of freedom. The differential expression statistic was bootstrapped to ascertain

confidence intervals for each gene being differentially expressed based on the data.

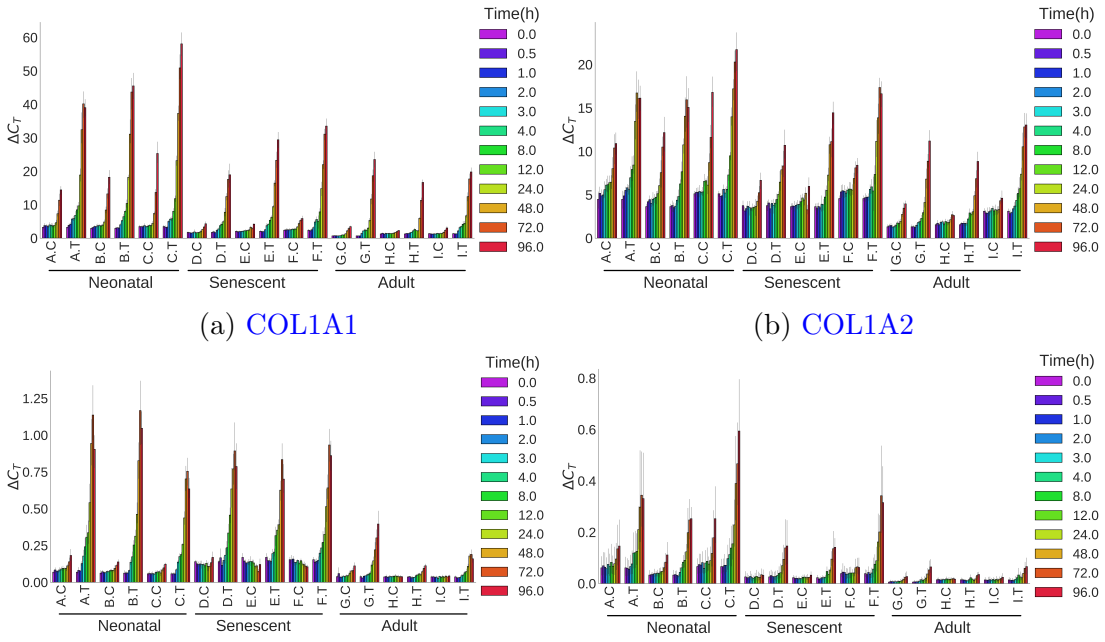
To perform the bootstrap, all samples were first normalized for the reference gene using $2^{-\Delta C_T}$ (Equation 5.8). Since useful information can be gained from both the TGF-β time series experiment and basal differences in cell types, a separate differential expression analysis was conducted for both time series and baseline samples, though the bootstrapping procedure differed slightly between them. The first element that was randomized was done so for both baseline and time series analyses and was the ordering of the 6 replicates. In the baseline analysis, the 96h time point was calibrated to the 0h time point so that the data represented the amount of gene g produced in 96h without stimulation. In the time series analysis, both TGF-β and control time series were calibrated to a baseline sample $B \in \{0h, 96h, \frac{0h+96h}{2}\}$ (Equation 5.9), the choice of which was randomly sampled. Then the TGF-β samples were normalized to their time matched controls. In both baseline and time series analyses, one adult or senescent cell line and one neonatal cell line was randomly sampled for input into LIMMA which was used to compute differential expression statistics. In the baseline analysis, LIMMA was used to contrast the amount of gene transcript in the senescent or adult cell lines compared to the neonatals. A corrected p-value for a moderated t-statistic (Smyth 2004) was computed for each gene representing the probability of the null hypothesis (that gene activity in senescent or adult cell lines are the same as the neonatal cell lines) being true. In the time series analysis, a moderated F-statistic (Smyth 2004) was computed to compare the differences in the amount of TGF-β-induced gene expression between cell lines over time. In both cases, a Bonferroni corrected p-value <0.05 was considered differentially expressed. This process was repeated 10,000 times in both the baseline and time series analysis whilst randomizing the elements that could be randomized. The proportion of times a gene was differentially expressed was recorded.

5.3 Results

5.3.1 Dynamic Measurements of Gene Activity in Neonatal, Adult and Senescent Fibroblasts.

High throughput qPCR was used to quantify gene expression levels over time in basal or TGF-β stimulated neonatal, adult and senescent fibroblasts. The experimental design (Figure 5.1) enabled the measurement of 72 genes on the high throughput qPCR platform. To select these genes, an extensive literature search was conducted to identify components of the dermal ECM or aspects of TGF-β signalling that may be different between fibroblasts in these different states. These genes are presented in Tables 5.1 to 5.2 with links to the relevant GeneCards entry (Safran et al. 2010) or to the appropriate graph (Figures 5.2 to 5.14).

Collagens



(a) COL1A1 (b) COL1A2

(c) COL4A1 (d) COL5A1

Figure 5.2: Time series measurements for collagen transcripts in neonatal, senescent and adult cell lines treated with 5ng mL^{-1} TGF-β or negative control. 0h time point corresponds to the baseline 0h control. Labels are cell line followed by control C or TGF-β T separated by a full stop.

Table 5.1: Table of measured genes and corresponding protein names. Table 1 of 2.

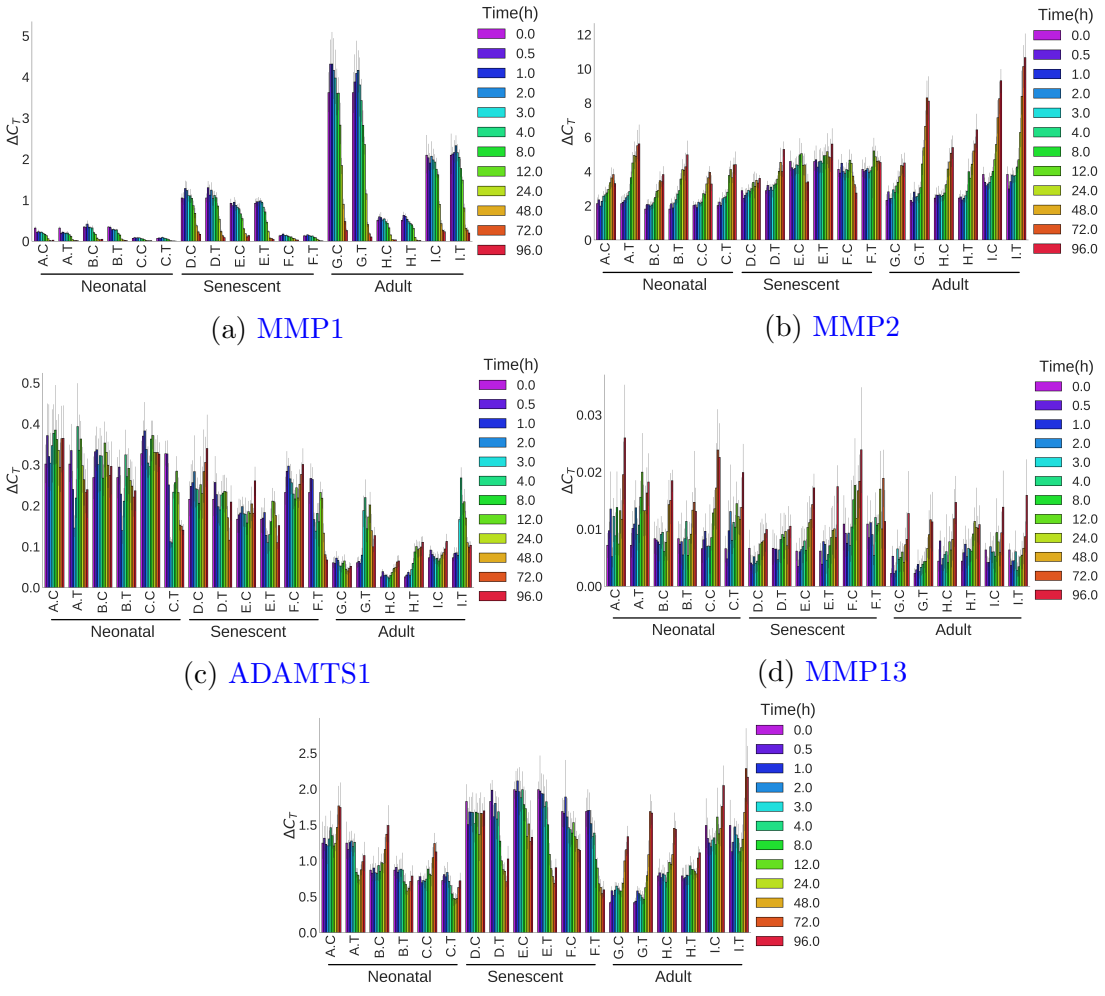
Gene Symbol	Protein
ACTA2	αSMA
ADAMTS1	ADAM Metallopeptidase Domain 10
ATP6AP1	ATPase H+ Transporting Accessory Protein 1
BGN	Biglycan
BHLHE40	Basic Helix-Loop-Helix Family Member E40
CAV1	Caveolin 1
CDKN2A	Cyclin Dependent Kinase Inhibitor 2A
CDKN2B	Cyclin Dependent Kinase Inhibitor 2B
COL1A1	Collagen Type I Alpha 1 Chain
COL1A2	Collagen Type I Alpha 2 Chain
COL4A1	Collagen Type IV Alpha 1 Chain
COL5A1	Collagen Type V Alpha 1 Chain
CTGF	Connective Tissue Growth Factor
DCN	Decorin
EGR2	Early Growth Response 2
ELN	Elastin
ENG	Endoglin
ETS1	ETS Proto-Oncogene 1
FBLN1	Fibulin 1
FBN1	Fibrillin 1
FGF2	Fibroblast Growth Factor 2
FN1	Fibronectin 1
GADD45B	Growth Arrest And DNA Damage Inducible Beta
HAS2	Hyaluronan Synthase 2
ID1	Inhibitor Of Differentiation 1
IL1A	Interleukin 1 Alpha
IL1B	Interleukin 1 Beta
IL6	Interleukin 6
ITGA1	Integrin Subunit Alpha 1
ITGA2	Integrin Subunit Alpha 2
JUN	Jun Proto-Oncogene
JUNB	JunB Proto-Oncogene

Table 5.2: Table of measured genes and corresponding protein names. Table 2 of 2.

Gene Symbol	Protein
LARP6	La Ribonucleoprotein Domain Family Member 6
LOXL1	Lysyl Oxidase Like 1
LOXL2	Lysyl Oxidase Like 2
LTBP2	Latent Transforming Growth Factor Beta Binding Protein 2
MMP1	Matrix Metallopeptidase 1
MMP13	Matrix Metallopeptidase 13
MMP14	Matrix Metallopeptidase 14
MMP2	Matrix Metallopeptidase 2
NOX4	NADPH Oxidase 4
PDGFA	Platelet Derived Growth Factor Subunit A
PMEPA1	Prostate Transmembrane Protein, Androgen Induced 1
PPIA	Peptidylprolyl Isomerase A
PPP3CA	Protein Phosphatase 3 Catalytic Subunit Alpha
PTEN	Phosphatase And Tensin Homolog
RARA	Retinoic Acid Receptor Alpha
RARG	Retinoic Acid Receptor Gamma
RHOB	Ras Homolog Family Member B
SERPINE1	Serpin Family E Member 1
SERPINE2	Serpin Family E Member 2
SKI	SKI Proto-Oncogene
SKIL	SKI Like Proto-Oncogene
SMAD3	SMAD Family Member 3
SMAD7	SMAD Family Member 7
SPARC	Secreted Protein Acidic And Cysteine Rich
TGFB1	Transforming Growth Factor Beta 1
TGFBR1	Transforming Growth Factor Beta Receptor 1
TGFBR2	Transforming Growth Factor Beta Receptor 2
THBS1	Thrombospondin 1
THBS2	Thrombospondin 2
TIMP1	TIMP Metallopeptidase Inhibitor 1
TIMP3	TIMP Metallopeptidase Inhibitor 3
TP53BP1	Tumor Protein P53 Binding Protein 1
VCAN	Versican
VEGFA	Vascular Endothelial Growth Factor A
VIM	Vimentin

Fibroblasts produced collagens both basally with time in culture and in response to TGF-β (Figure 5.2). Neonatal cell lines all produced more collagen in response to TGF-β compared to control. Both senescent and adult fibroblasts produced less COL1A1 (Figure 5.2a), COL1A2 (Figure 5.2b) and COL5A1 (Figure 5.2d) compared to neonatal cells under basal conditions. Moreover, COL4A1 (Figure 5.2c) was produced to similarly in neonatal and senescent cell lines. Each of the measured collagens exhibited a reduced response to TGF-β in the senescent and adult cell lines compared to neonatal except the COL4A1 in senescent cell lines, which were expressed to a similar degree.

MMPs



(a) MMP1

(b) MMP2

(c) ADAMTS1

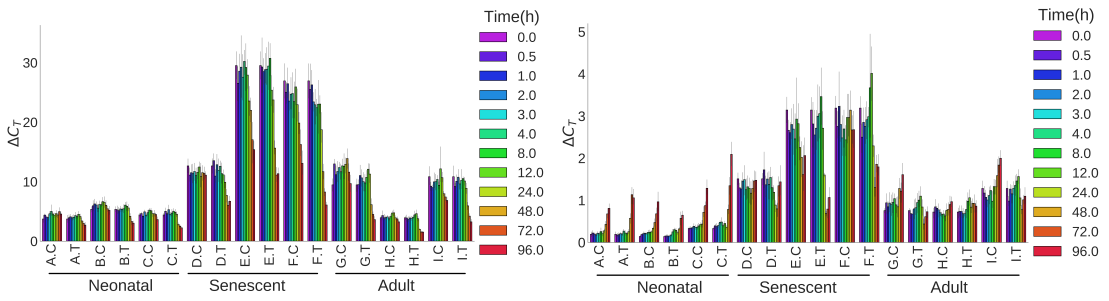
(d) MMP13

(e) MMP14

Figure 5.3: Time series measurements for MMP transcripts in neonatal, senescent and adult cell lines treated with 5ng mL^{-1} TGF-β or negative control. 0h time point corresponds to the baseline 0h control. Labels are cell line followed by control C or TGF-β T separated by a full stop.

MMP1 was upregulated in adult compared to neonatal fibroblasts, although there was a considerable degree of heterogeneity in that cell lines G and I produced more MMP1 than cell line H (Figure 5.3a). Senescent fibroblasts displayed a similar behaviour to adults, though to a lesser degree. With time in culture, the amount of MMP1 produced was markedly decreased in all the cell lines, and this behaviour was TGF-β independent. In contrast, MMP2 was upregulated in response to TGF-β in both neonatal and adult cell lines. Adult cell lines produced more MMP2 than the neonatal cell lines (Figure 5.3b). ADAMTS1 production was lower in adult cell lines both basally and induced by TGF-β. However, under TGF-β stimulation, adult cell lines produced more ADAMTS1, a response that was not observed in neonatal or senescent cells. MMP13 (Figure 5.3e) and MMP14 (Figure 5.3d) were also measured but MMP13 appeared to be detected at only background levels and the MMP14 data is difficult to interpret.

TIMPs

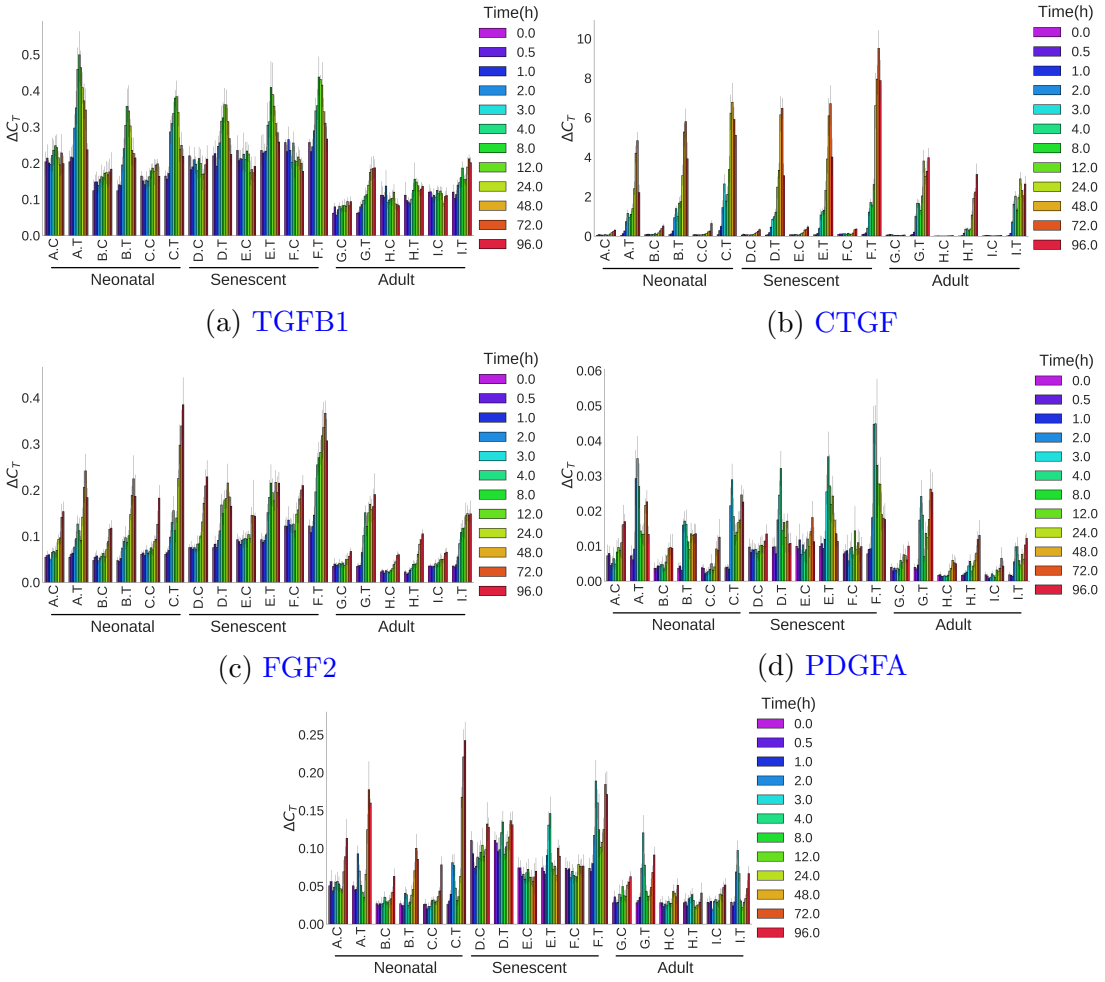


(a) TIMP1 (b) TIMP3

Figure 5.4: Time series measurements for TIMP transcripts in neonatal, senescent and adult cell lines treated with 5ng mL^{-1} TGF-β or negative control. 0h time point corresponds to the baseline 0h control. Labels are cell line followed by control *C* or TGF-β *T* separated by a full stop.

The amount of TIMP1 (Figure 5.4a) and TIMP3 (Figure 5.4b) produced in each cell line did not depend on TGF-β. The amount of TIMP1 produced by cells under the control condition declined over the course of the 96h. The amount of TIMP3 on the other hand increased over that time frame, except in the senescent cell lines which displayed heterogeneous behaviour.

Growth factors



(a) TGFB1 (b) CTGF

(c) FGF2 (d) PDGFA

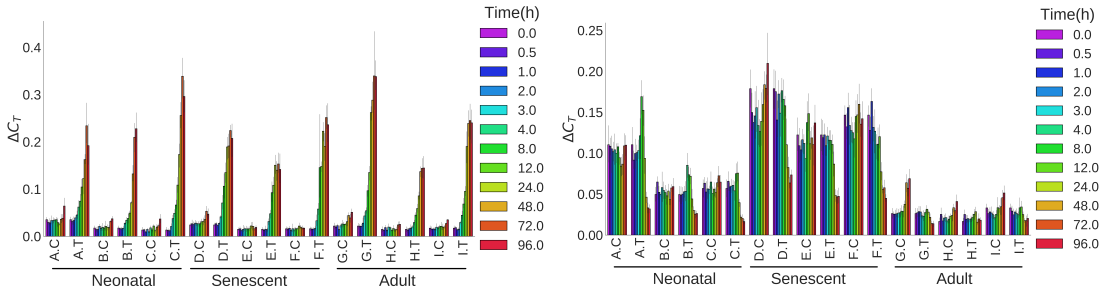
(e) VEGFA

Figure 5.5: Time series measurements for growth factors transcripts in neonatal, senescent and adult cell lines treated with 5ng mL^{-1} TGF-β or negative control. 0h time point corresponds to the baseline 0h control. Labels are cell line followed by control *C* or TGF-β *T* separated by a full stop.

TGF-β induced its own production in all cell types (Figure 5.5a). The amount produced was lower in adult cells than in neonatal while senescence did not impair the fibroblasts ability to produce TGF-β in response to TGF-β. CTGF was produced at low levels under basal conditions and the amount produced was reduced in adult fibroblasts (Figure 5.5b). TGF-β induced a dramatic increase in CTGF production. The CTGF profile has a characteristic dip at around 4h post TGF-β stimulation. The amount of CTGF produced in the adult cell lines was lower than in the neonatal cell lines. Fibroblasts produced VEGFA when cultured *in vitro* and the amount dramatically increased 24-48h after seeding (Figure 5.5e). FGF2 was produced gradually over time

and its production was enhanced by TGF-β stimulation. Adult fibroblasts produced less FGF2 than neonatal fibroblasts, but still responded to TGF-β. Senescent fibroblasts did not respond differently to TGF-β compared to neonatal fibroblasts. PDFGA production on the other hand was enhanced by TGF-β treatment in all cell types, though this response was blunted in TGF-β stimulation of adult compared to neonatal cell lines.

Integrins



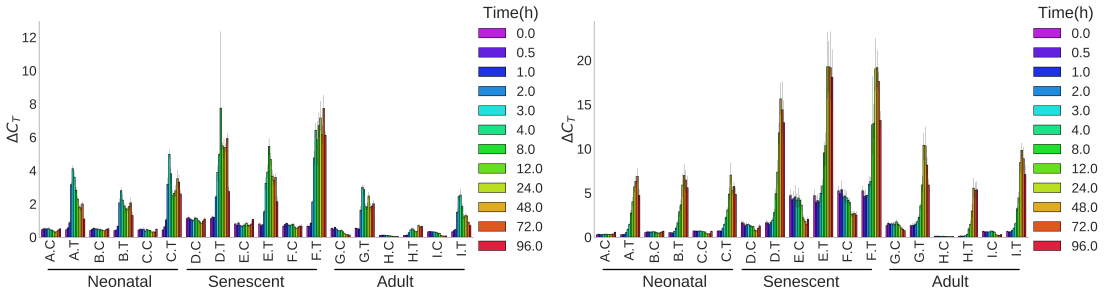
(a) ITGA1 (b) ITGA2

Figure 5.6: Time series measurements for integrin transcripts in neonatal, senescent and adult cell lines treated with 5ng mL^{-1} TGF-β or negative control. 0h time point corresponds to the baseline 0h control. Labels are cell line followed by control *C* or TGF-β *T* separated by a full stop.

ITGA1 (Figure 5.6a) was not differentially expressed in age or senescence compared to neonatal cell lines but responded to TGF-β with a dramatic increase in its production. ITGA2 expression on the other hand (Figure 5.6a) was reduced in adult cell lines under the control condition but enhanced in senescent cell lines. ITGA2 was unresponsive to TGF-β.

Serpines

SERPINE1 (Figure 5.7a) and SERPINE2 (Figure 5.7a) both respond to TGF-β with increased expression. SERPINE1 has a characteristic dual peak at 3h and 72h post TGF-β stimulation. SERPINE2 on the other hand takes longer to respond, peaking between 48h and 72h after addition of TGF-β. Senescence enhances the production of both SERPINE1 and SERPINE2, particularly in response to TGF-β. Adult cells display some heterogeneity regarding SERPINE's. Specifically, cell line G produced

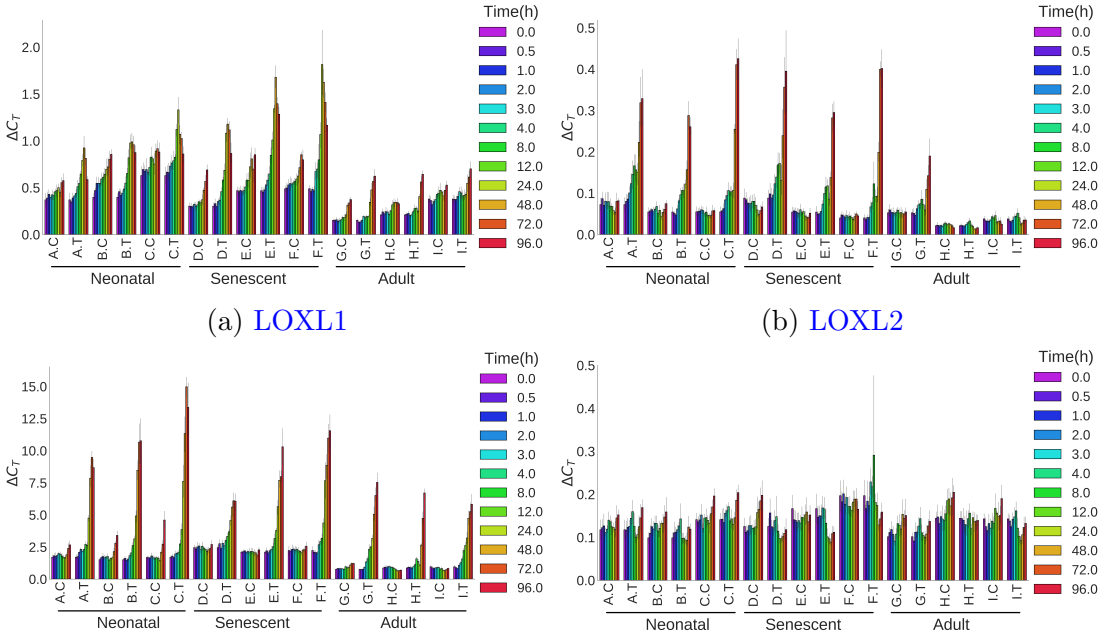


(a) SERPINE1 (b) SERPINE2

Figure 5.7: Time series measurements for serpine transcripts in neonatal, senescent and adult cell lines treated with 5ng mL^{-1} TGF-β or negative control. 0h time point corresponds to the baseline 0h control. Labels are cell line followed by control *C* or TGF-β *T* separated by a full stop.

more SERPINE2 than neonatal controls while cell line H produced less.

Collagen processing



(a) LOXL1 (b) LOXL2

(c) SPARC (d) LARP6

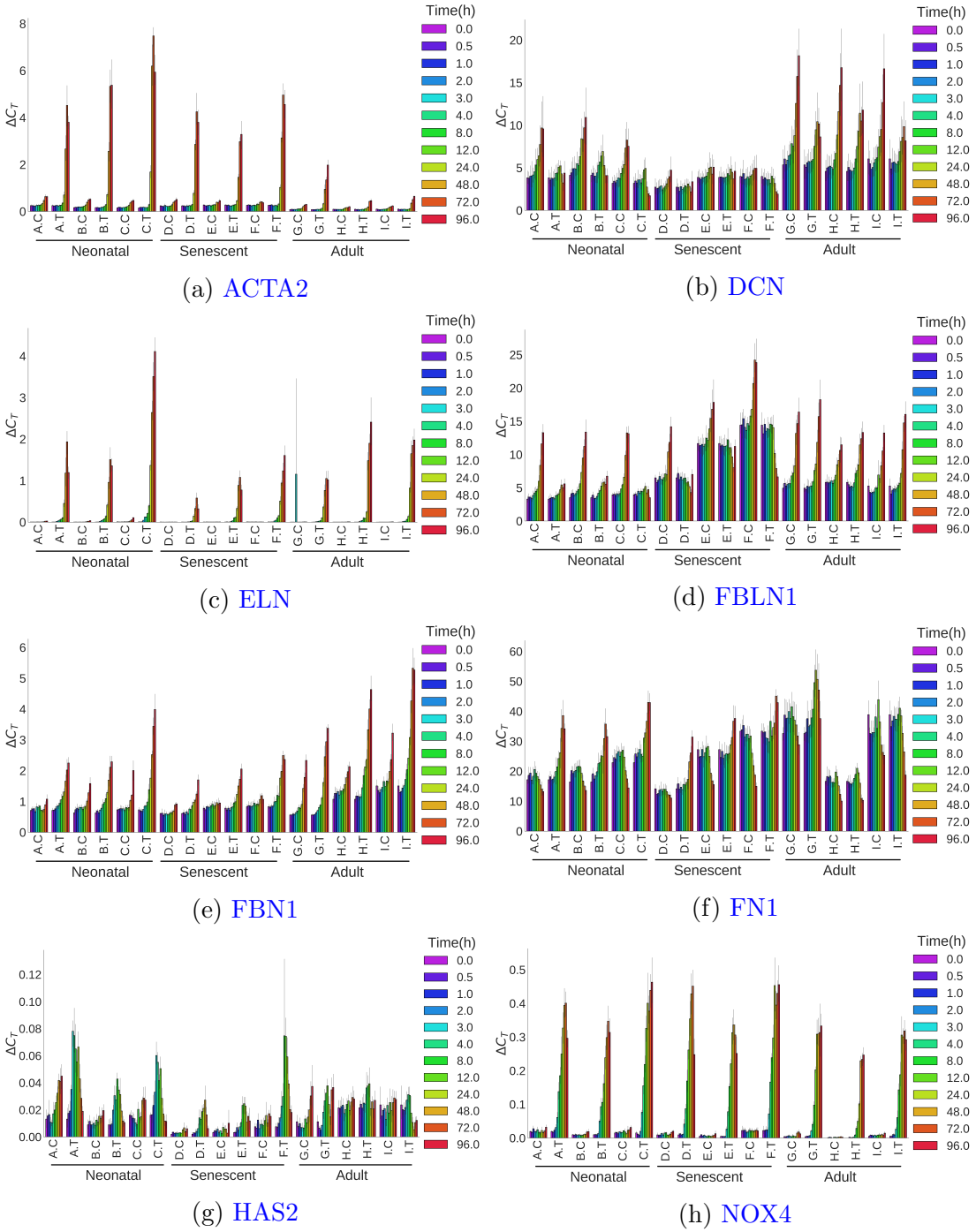
Figure 5.8: Time series measurements for serpine transcripts in neonatal, senescent and adult cell lines treated with 5ng mL^{-1} TGF-β or negative control. 0h time point corresponds to the baseline 0h control. Labels are cell line followed by control *C* or TGF-β *T* separated by a full stop.

LOXL1 (Figure 5.8a) was weakly produced in response to TGF-β in neonatal cells. In adult cells, the control and TGF-β induced levels of LOXL1 were reduced compared to neonatal cell lines. LOXL2 (Figure 5.8b) on the other hand has a more pronounced

response to TGF-β that was reduced in adult but not senescent fibroblasts. SPARC (Figure 5.8c) responded to TGF-β similarly to LOXL1 while LARP6 (Figure 5.8d) did not transcriptionally respond to TGF-β and was unaffected by age or senescence.

ECM components

ACTA2 (Figure 5.9a) was produced by fibroblasts over time without TGF-β stimulation. However, TGF-β stimulated a strong response from ACTA2, which was blunted in senescent and considerably reduced in adult fibroblasts. DCN (Figure 5.9b) on the other hand was produced by fibroblasts and TGF-β attenuates this response. Senescent fibroblasts were less able to produce DCN in the absence of TGF-β while adult cells produced more. ELN (Figure 5.9c) was present at a very low quantities in control cells and TGF-β induced a strong transcriptional response that was blunted by both age and senescence. Like DCN, FBLN1 (Figure 5.9d) was produced by fibroblasts and TGF-β treatment attenuated its production. FBN1 (Figure 5.9e) was produced both with and without TGF-β stimulation but was unaffected by age or senescence. The amount of FN1 (Figure 5.9f) produced by fibroblasts declined over 96h in the absence of TGF-β while TGF-β treatment induced its production. HAS2 (Figure 5.9g) was produced both with and without TGF-β but this response was lost in adult cell lines. Small amounts of NOX4 (Figure 5.9h) were produced in control neonatal fibroblasts and age reduced the amount produced compared to neonatal or senescent cell lines. TGF-β induced transcription of NOX4 in all cell lines and this response was similar across the cell lines. THBS1 (Figure 5.10a) production was induced by TGF-β and age attenuated this response in cell line G but not H and I. The response to TGF-β was more pronounced in senescence compared to neonatal fibroblasts. VCAN (Figure 5.10c) was produced under basal conditions. TGF-β caused an increase in the amount of VCAN produced but also initiated a negative feedback that induced a decline in VCAN at 24h. This response was similar in all three conditions though basal levels of VCAN were lower in adult cell lines. VIM (Figure 5.10d) was present in large quantities in all three cell lines. TGF-β induced a reduction in the amount of VIM mRNA in both senescence



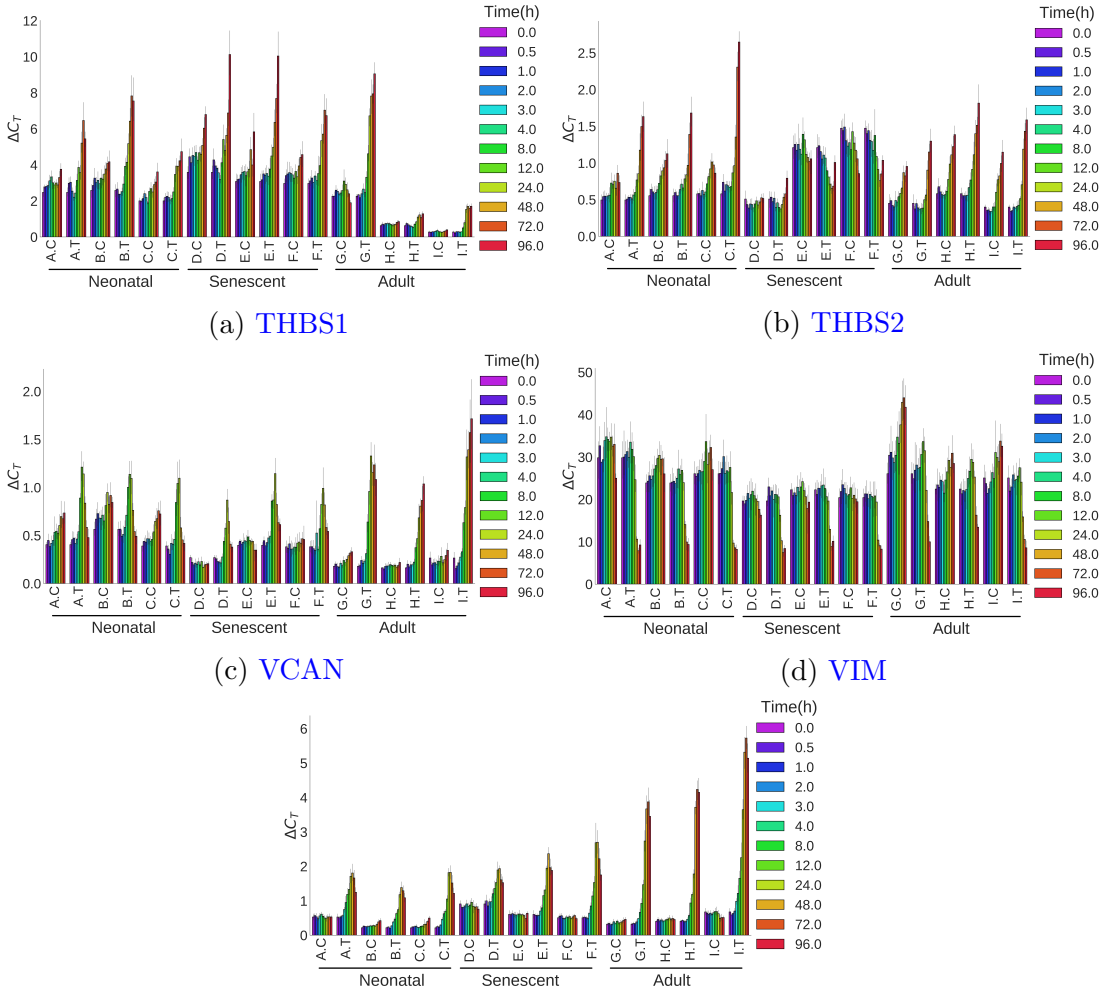
(a) ACTA2 (b) DCN

(c) ELN (d) FBLN1

(e) FBN1 (f) FN1

(g) HAS2 (h) NOX4

Figure 5.9: Time series measurements for ECM components transcripts in neonatal, senescent and adult cell lines treated with 5ng mL^{-1} TGF-β or negative control. 0h time point corresponds to the baseline 0h control. Labels are cell line followed by control (C) or TGF-β (T) separated by a full stop.



(a) THBS1

(b) THBS2

(c) VCAN

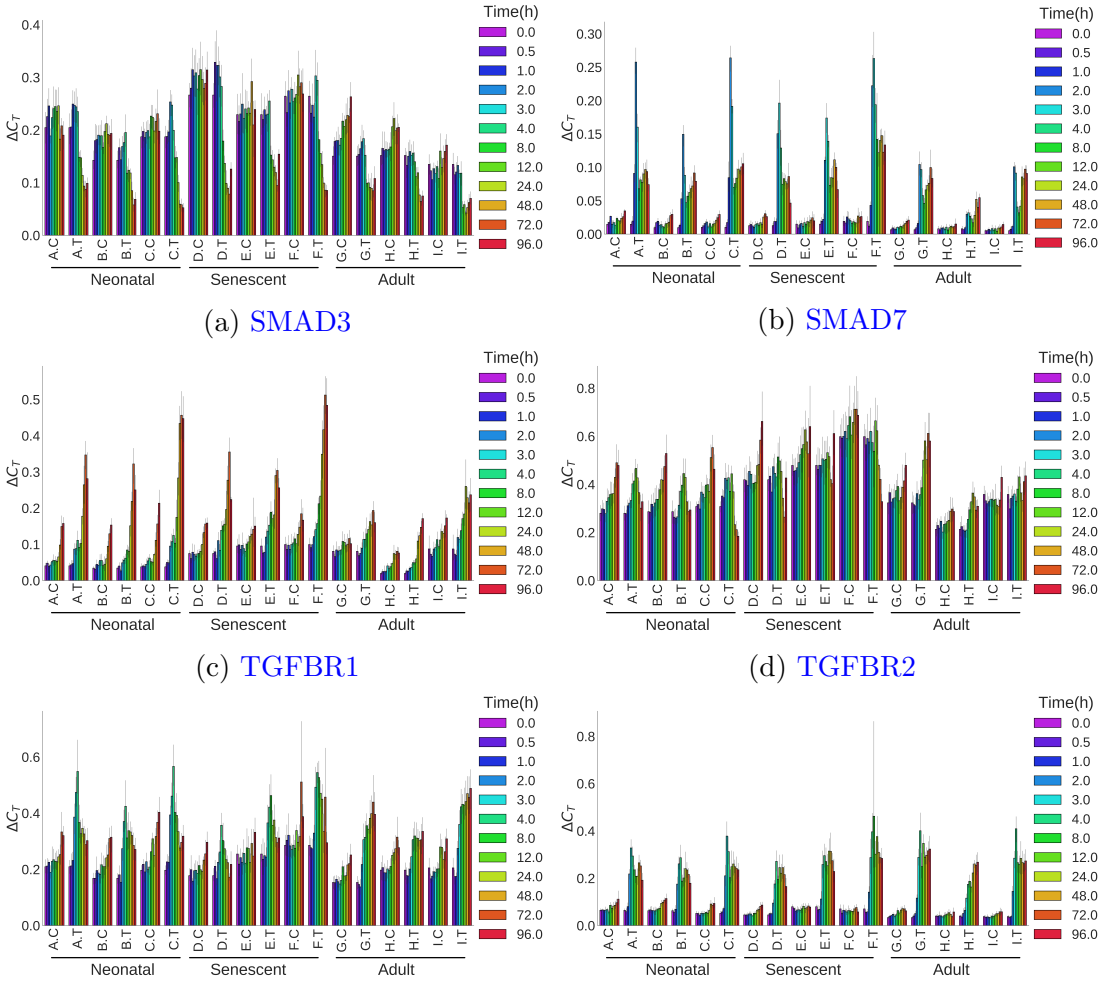
(d) VIM

(e) LTBP2

Figure 5.10: Time series measurements for ECM components transcripts in neonatal, senescent and adult cell lines treated with 5ng mL^{-}1 TGF-β or negative control. 0h time point corresponds to the baseline 0h control. Labels are cell line followed by control C or TGF-β T separated by a full stop.

and adult cell lines. Fibroblasts produce constant amount of LTBP2 (Figure 5.10e) under basal conditions while TGF-β stimulation induced LTBP2 production. More LTBP2 was produced in adult cell lines compared to neonatal or senescent cells.

TGF-β signalling



(a) SMAD3 (b) SMAD7

(c) TGFBR1 (d) TGFBR2

(e) SKI (f) SKIL

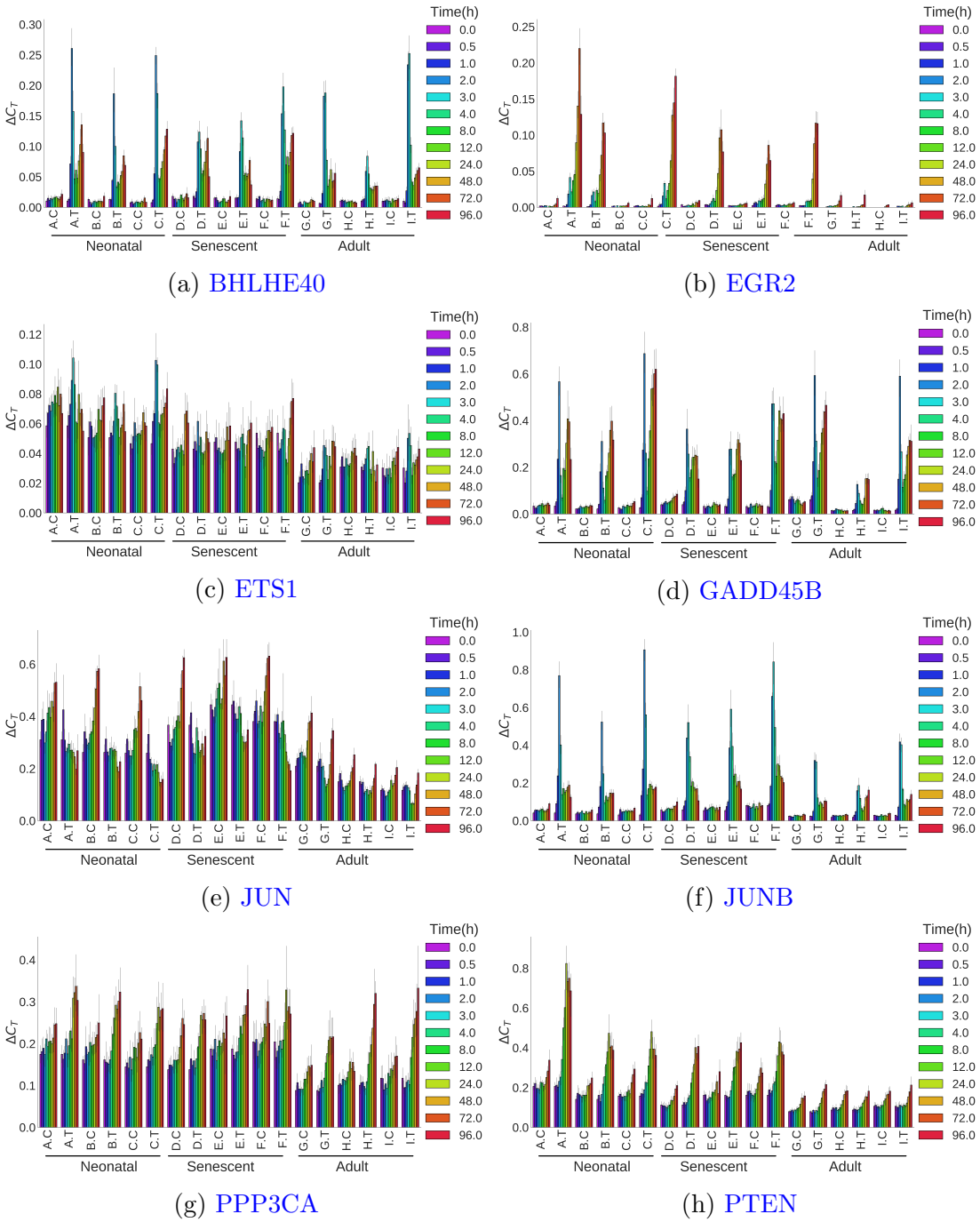
Figure 5.11: Time series measurements for ECM components transcripts in neonatal, senescent and adult cell lines treated with 5ng mL^{-1} TGF-β or negative control. 0h time point corresponds to the baseline 0h control. Labels are cell line followed by control (C) or TGF-β (T) separated by a full stop.

TGF-β reduced the amount of SMAD3 produced at 12h after stimulation and this response was observed in all cell lines (Figure 5.11a). SMAD7 (Figure 5.11b) was not significantly produced without TGF-β stimulation and control levels were similar between cell lines. TGF-β induced a characteristic strong first peak of SMAD7 production at 2h and then a second weaker peak at 72h post-stimulation. The SMAD7

response was not affected by senescence but was blunted in age. TGFBR1 (Figure 5.11c) was produced without TGF-β stimulation in all cell lines and TGF-β induced its production which peaked at 72h post stimulation. Adult cell lines had a blunted TGFBR1 response to TGF-β compared to neonatal and senescent cell lines. TGFBR2 was produced in the control condition (Figure 5.11d). TGF-β did not impact on the magnitude of TGFBR2 produced but production was attenuated at late time points after TGF-β stimulation. The amount of TGFBR2 was reduced in adult cell lines. SKI (Figure 5.11e) was produced under basal conditions. TGF-β stimulation enhanced SKI production, which peaked at 4h. The response in age was not significantly different between cell lines.

Other signalling

BHLHE40 was produced without TGF-β stimulation in neonatal fibroblasts (Figure 5.12a). TGF-β enhances the production of BHLHE40 and showed a characteristic dual peak, first at 1h and second at 72h post TGF-β stimulation. EGR2 was produced at low levels in the fibroblast and TGF-β increased the amount that was produced (Figure 5.12b). EGR2 also displayed a dual peak in response to TGF-β, first at 4h and the second, much larger, at 72h. Adult cell lines had a blunted EGR2 response to TGF-β compared to neonatal. Two of the control time series for EGR2 under the adult condition (G and I) were below the levels of detection and are not shown in the graph. ETS1 did not change over time or in response to TGF-β (Figure 5.12c). Adult cell lines produce less ETS1 than neonatal cell lines. TGF-β induced a double peak response from GADD45B (Figure 5.12d), first a strong response at 2h and a second at 48-72h post stimulation. The expression pattern of GADD45B did not change between cell lines. JUN production was inhibited by TGF-β in neonatal and senescent cell lines (Figure 5.12e). Adult cell lines produced less JUN than neonatal or senescent cell lines and TGF-β did not impact the JUN induction profile. TGF-β induced JUNB production in all cell lines (Figure 5.12f). The amount of JUNB in adult cells was lower in both basal and TGF-β induced conditions while senescent cells have a



(a) BHLHE40 (b) EGR2

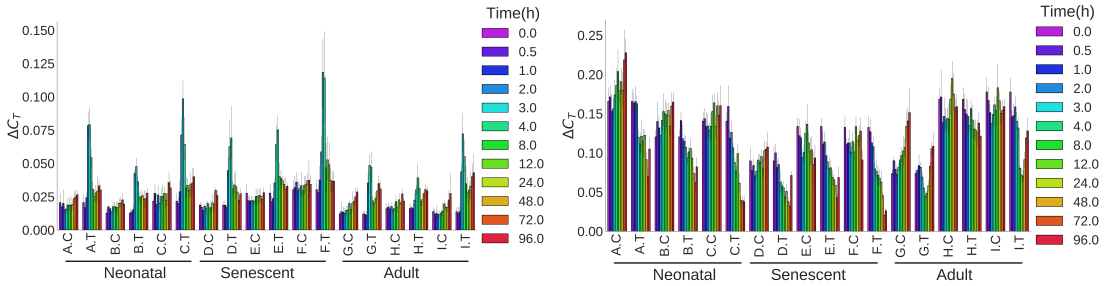
(c) ETS1 (d) GADD45B

(e) JUN (f) JUNB

(g) PPP3CA (h) PTEN

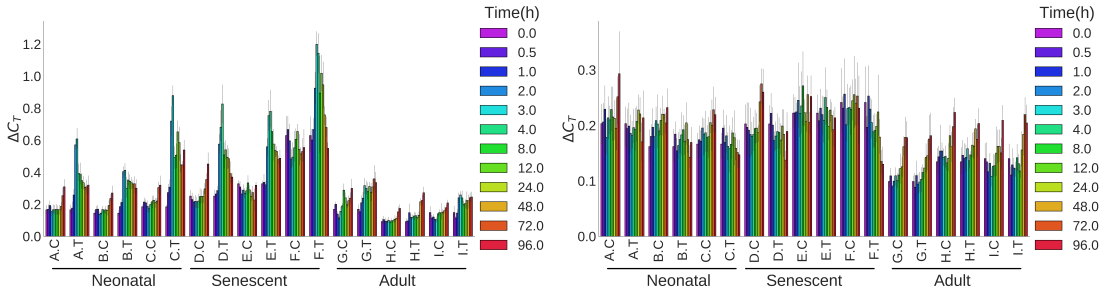
Figure 5.12: Time series measurements for ECM components transcripts in neonatal, senescent and adult cell lines treated with 5ng mL^{-1} TGF-β or negative control. 0h time point corresponds to the baseline 0h control. Labels are cell line followed by control (C) or TGF-β (T) separated by a full stop.

similar response as neonatal. Levels of PPP3CA in all cell lines increased over time and TGF-β enhanced this response (Figure 5.12g). Adult cell lines produced less PPP3CA than neonatal or senescent cells under basal conditions but not in the presence of TGF-β. PTEN was produced over time in TGF-β stimulated and unstimulated conditions. Neonatal and senescent cell lines produced similar PTEN profiles in neonatal and senescent conditions (Figure 5.12h). Adult cell lines produce less PTEN and have a blunted response to TGF-β compared to neonatal cell lines. RARA was



(a) RARA

(b) RARG



(c) RHOB

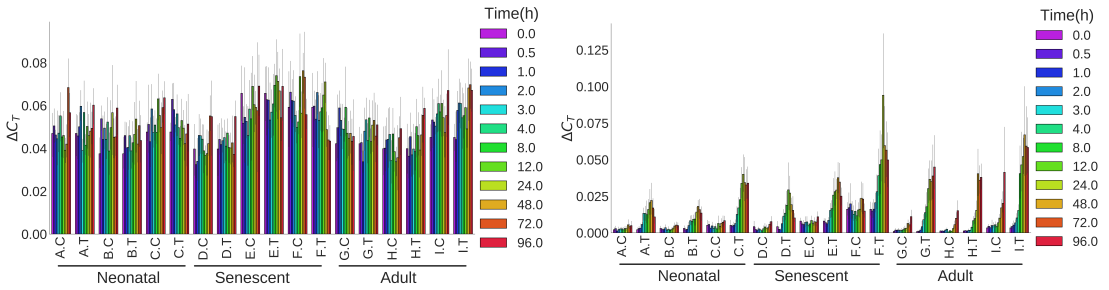
(d) TP53BP1

Figure 5.13: Time series measurements for other signalling component transcripts in neonatal, senescent and adult cell lines treated with 5ng mL^{-1} TGF-β or negative control. 0h time point corresponds to the baseline 0h control. Labels are cell line followed by control (C) or TGF-β (T) separated by a full stop.

produced in response to TGF-β with a peak at 2-3h post stimulation (Figure 5.13a). Senescent cell lines showed a similar response to neonatal cell lines but adult cell lines had a blunted response. TGF-β inhibited RARG production in neonatal and senescent fibroblasts (Figure 5.13b). In contrast, in adult cells RARG levels increased after an initial decline in response to TGF-β. RHOB was produced by fibroblasts in the absence of TGF-β stimulation, though it required 72-96h in culture. TGF-β induced RHOB production that peaked at 2-3h post stimulation. In adult cells the RHOB response was blunted. TP53BP1 did not respond to TGF-β in fibroblasts. In adult cell lines the amount of TP53BP1 was increased at late time points in culture but this behaviour was

absent in neonatal and senescent cell lines (Figure 5.13d).

Senescence Markers



(a) CDKN2A (b) CDKN2B

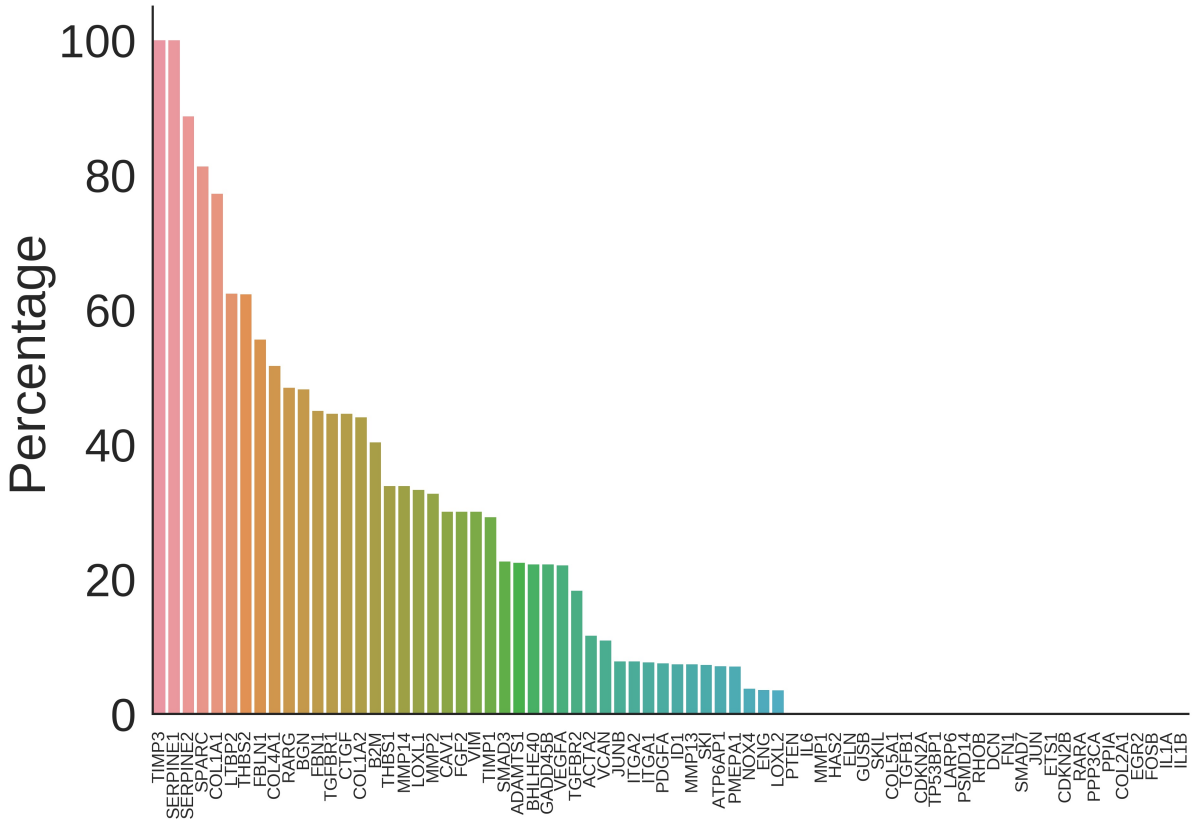
Figure 5.14: Time series measurements for senescence marker transcripts in neonatal, senescent and adult cell lines treated with 5ng mL^{-1} TGF-β or negative control. 0h time point corresponds to the baseline 0h control. Labels are cell line followed by control (C) or TGF-β (T) separated by a full stop.

CDKN2A did not respond to TGF-β stimulation in any cell line (Figure 5.14a).

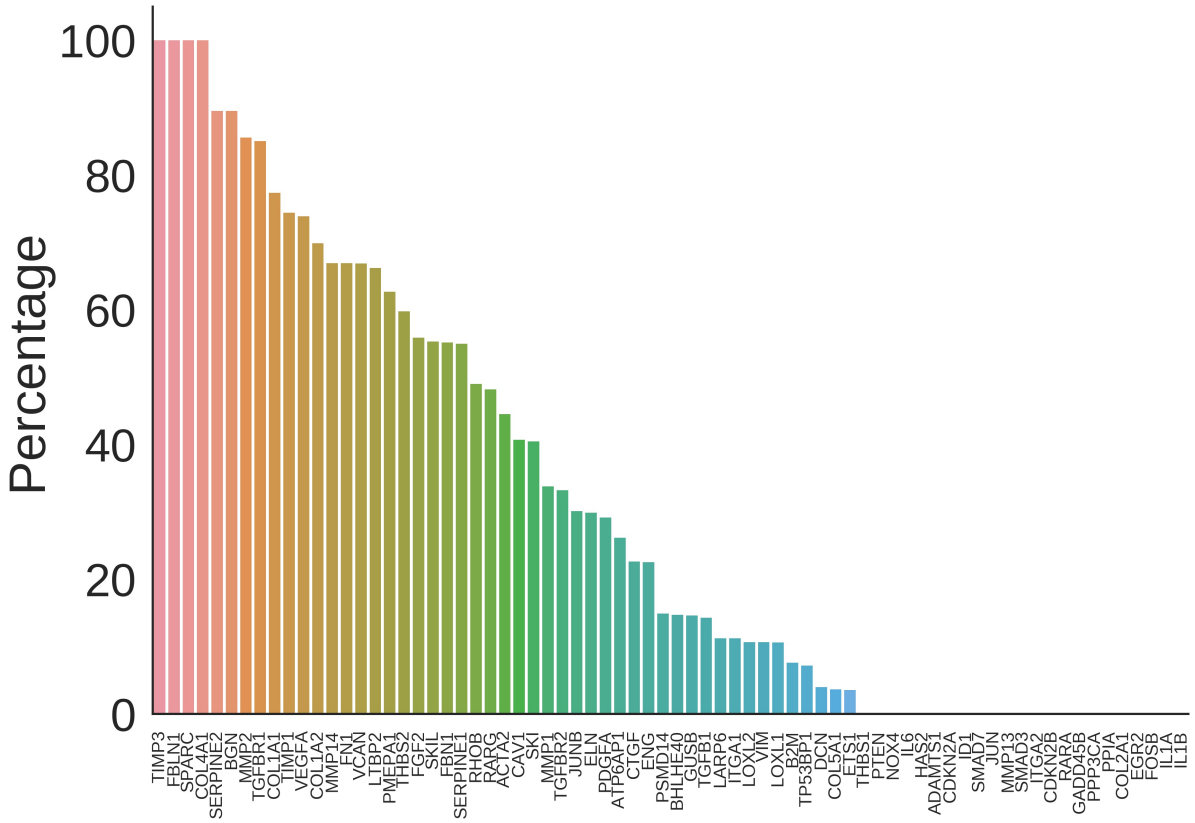
CDKN2B on the other hand was produced in response to TGF-β in all cell lines, but most notably in the adult cell lines (Figure 5.14a). /

Differential Expression with LIMMA Resampling

To provide statistical support for genes that are different between groups, a differential expression analysis was conducted using LIMMA. In Figures 5.2 to 5.14, the data were presented as $2^{-\Delta C_T}$ (i.e. no calibrating normalisation) rather than $2^{-\Delta\Delta C_T}$ because it better displays the available information without highlighting one dimension at the cost of reducing information in another. For differential expression analysis, $2^{-\Delta\Delta C_T}$ values were used to highlight two different questions: 1) is the amount of each gene produced over 96h under the baseline condition different between senescent or adult fibroblasts compared with neonatal and 2) is the response to TGF-β different over time in adult and senescent cells compared to neonatal. The two analyses, referred to as the baseline and time series analysis respectively, were bootstrapped 10,000 times (as described in the methods) in order to exploit the variability in the data to attain confidence intervals (Figure 5.15 and Figure 5.16 respectively). This procedure was coded manually using R

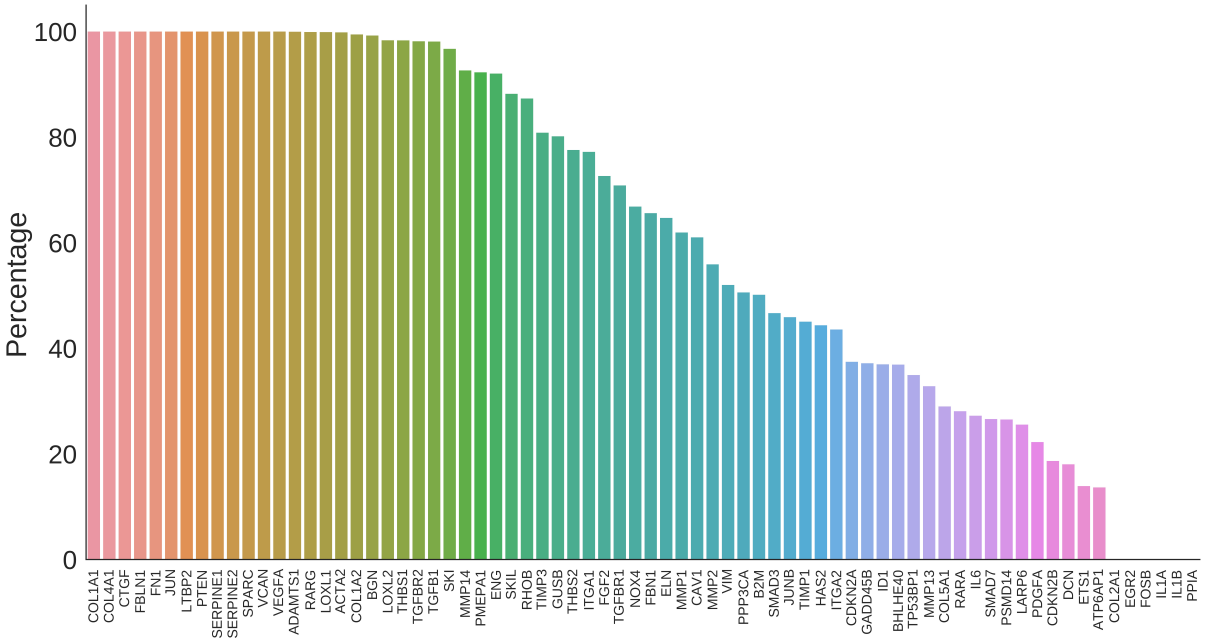


(a) Adult versus neonatal

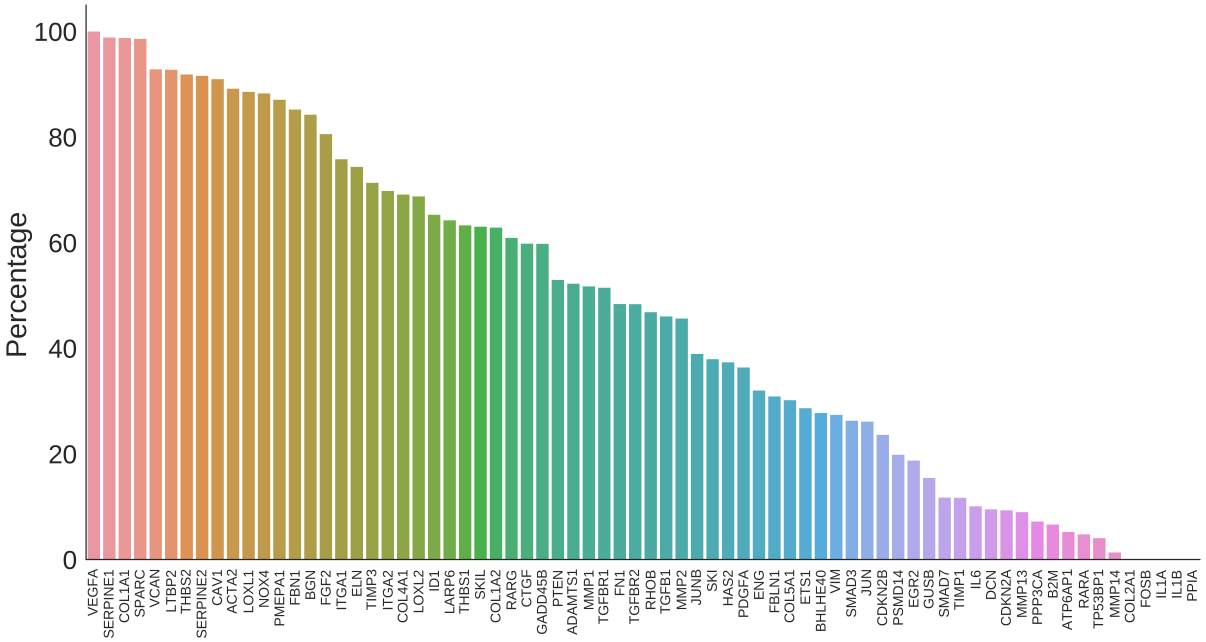


(b) Senescent versus neonatal

Figure 5.15: Baseline data at 96h was normalized to baseline data at 0h in each cell line. Then baseline data from (a) adult and (b) senescent cell lines were compared with neonatal cells lines for differential expression using LIMMA (p-value=<0.05, Bonferroni corrected). The process was bootstrapped $1e^4$ times as described in the methods and the percentage of times a gene is found differentially expressed is shown.



(a) Adult versus neonatal



(b) Senescent versus neonatal

Figure 5.16: Time series data were normalized to the baselines and TGF-βsamples were normalized to control samples at the same time point to get fold change at each time point. Basis-splines (4 degrees of freedom) were fit to all fold change profiles before (a) adult and (b) senescent cell lines were compared against neonatal for differential expression (p-value > 0.05, Bonferroni corrected). The process was bootstrapped $1e^4$ times using random parameters as described in the methods. Shown are the percentages each gene was determined to be differentially expressed in (a) adult and (b) senescent cell lines compared to neonatal.

and is described in the methods section. This approach to differential expression provides the percentage of times that a gene was differentially expressed out of all the ways that the analysis can be conducted. The percentages can be interpreted as a probability that a gene is differentially expressed based on the available data.

5.3.2 Quality Control

This dataset has 72 genes and 1296 individual samples. PCA was used to reduce the dimensionality of the data and as a measure of quality control for the dataset. Four of genes were discarded due to anomalies leaving 68 for the subsequent analysis. The scree plot in Figure 5.17g shows the amount of variance explained in the top 10 principal components and that the majority of the variance is explained by the first two. Therefore, the first two principal components were visualised as a 2D scatter plot coloured by various experimental factors (Figure 5.17). These graphs give a good indication that the experimental data is good quality. Collectively, (Figures 5.17a to 5.17f) show that individual cell lines and cell types cluster together; that time in culture is an important determinant of expression levels (Figure 5.17c) and that TGF-β induced time dependent transcription changes in all cell lines (Figure 5.17d). Biological replicates do not cluster on the PCA (Figure 5.17f) indicating that experimental bias has not been introduced by labelling sample replicates.

5.4 Discussion

The dermis is a fibrous mesh of collagens and elastins with a large number of support proteins that make a fully operational dermis. With age the composition of the dermis gradually deteriorates and function is impaired. Many of the physical characteristics of skin ageing such as wrinkling and dermal thinning can be attributed to these changes (Cole et al. 2018). TGF-β is a master regulator of ECM biosynthetic processes and so

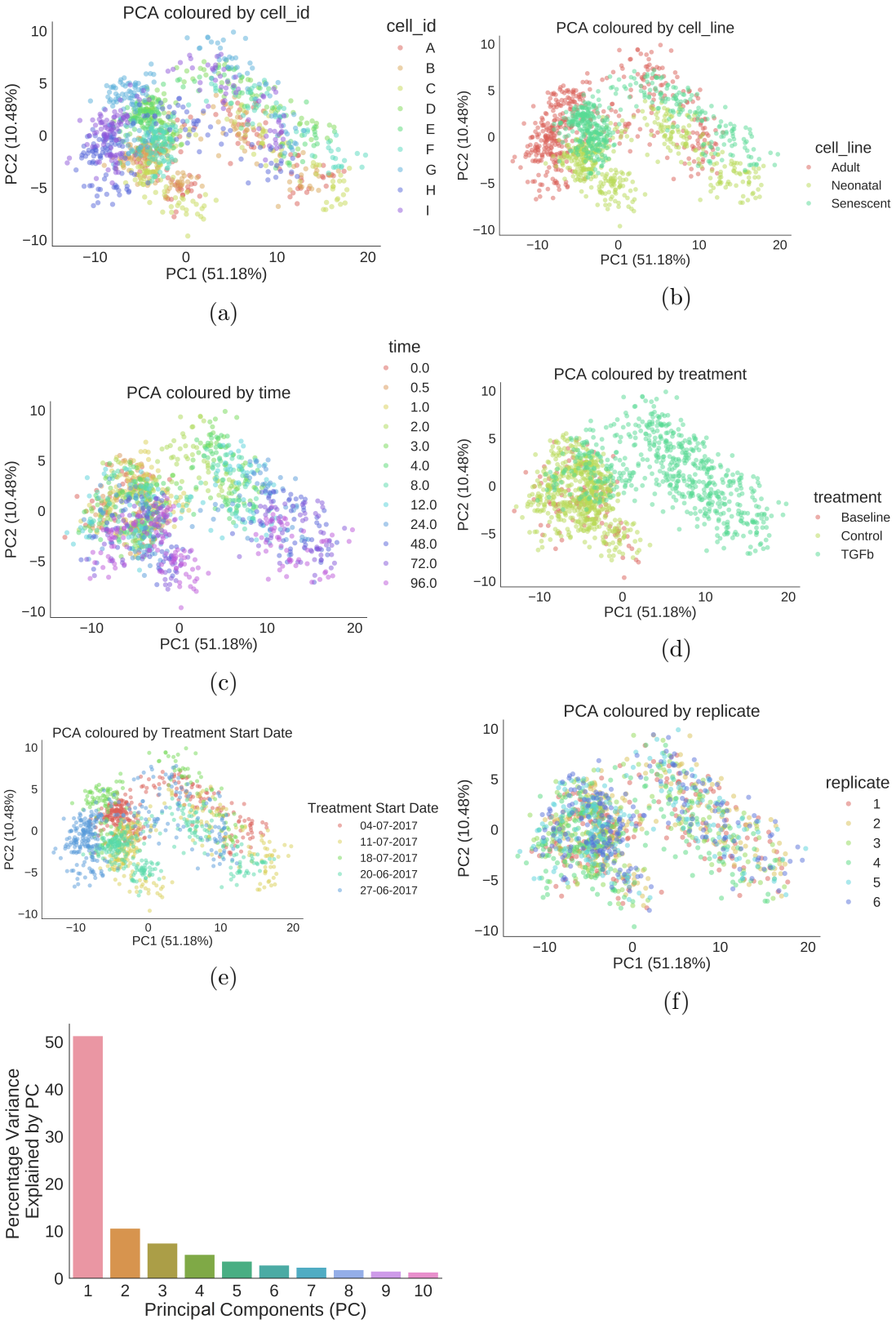


Figure 5.17: (a-g) PCA on raw C_T data coloured by the indicated experimental factors. (f) Scree plot showing percentage of variance explained by the PCA.

the purpose of this chapter was to treat neonatal, senescent and adult fibroblasts with TGF-β and to compare the activities of genes known to be involved in TGF-β signalling or ECM integrity. The following sections provide a discussion on various aspects of the data described in this chapter.

ECM biosynthesis

The data presented in this chapter support the notion that the aged dermis has a pro-catabolic environment relative to the young. This concept is not novel and much of the data presented here confirms what is already known. For instance in Varani et al. 2006, type 1 procollagen levels were 3 times lower in aged (80+) compared to young (18-29) individuals (15ng/mm^2 compared to 5ng/mm^2 respectively). The data in Figure 5.2 are consistent with these observations as less collagen production was observed in adult compared to neonatal cell lines. This occurred in both basal and TGF-β induced conditions. In addition to type 1 collagens, we have also provided evidence for impaired type IV and V collagens in age (Figure 5.2c and Figure 5.2d). The data for several genes had a similar interpretation as collagen in that the older cells produced less in both the control and TGF-β induced conditions. These genes include TGFB1 (Figure 5.5a), CTGF (Figure 5.5b), VEGFA (Figure 5.5e), ITGA2 (Figure 5.6b), LOXL1(Figure 5.8a), LOXL2 (Figure 5.8b), ACTA2 (Figure 5.9a), THBS1 (Figure 5.10a), SMAD7 (Figure 5.11b) and JUNB (Figure 5.12f).

CTGF is an important regulator of fibrotic signalling pathways and is a secondary responder to TGF-β (Quan et al. 2002, Wahab, Weston & Mason 2005, Ponticos, Holmes, Shi-wen, Leoni, Khan, Rajkumar, Hoyles, Bou-Gharios, Black, Denton, Abraham, Leask & Lindahl 2009). Evidence that blocking CTGF signalling was able to reduce the amount of collagen produced by sclerotic fibroblasts emphasises the correlation between CTGF and collagen levels (Makino et al. 2017, Sonnylal et al. 2010). Systemic sclerosis and skin ageing are both characterised by aberrant collagen deposition. In systemic sclerosis, too much collagen is produced while in skin ageing, too little is produced. It has been found that CTGF levels correlate with collagen levels

(Quan et al. 2010) and the data shown in Figure 5.5b supports this hypothesis. Collectively these data indicate that CTGF is a regulator of collagen production and imbalances in CTGF regulation have adverse consequences. Moreover, CTGF is a potential therapeutic target for modulating collagen homeostasis.

Reduced collagen levels with age may result from both reduced production and increased degradation. MMPs are proteases in the extracellular matrix, some of which can degrade collagen. It has been shown that the aged dermis produces higher levels of MMPs compared to younger individuals. Specifically, the aged dermis found higher levels of MMP1, MMP3, MMP9, MMP10, MMP11, MMP23, MMP24, MMP27 and MMP28 in the aged dermis compared to young (Qin et al. 2017, Fisher et al. 2009). In agreement with Qin et al. 2017, Figure 5.3 shows that MMP1 but not MMP2, MMP13 or MMP14 show enhances expression in aged tissue compared to young. In addition, Figure 5.3 also suggests that a degree of heterogeneity exists regarding MMP1 levels in age.

The MMP1 response to TGF-β (Figure 5.3a) is a surprising result that distinguishes this experiment from other reported studies. White et al. 2000 characterised a TGF-β inhibitory element (TIE) in the MMP1 promoter that is involved in constitutive MMP1 repression. Further, both White et al. 2000 and Edwards et al. 1996 observed that TGF-β inhibited PMA-induced MMP1 production, while (Yuan & Varga 2001) observed that TGF-β inhibited IL-1β production. These data point towards TGF-β mediated transcriptional repression of MMP1, which is an attractive idea because it is intuitive from a resource allocation point of view that if TGF-β is directing anabolic processes, than catabolic processes should be inhibited. This experiment has been cited for showing evidence that TGF-β stimulation inhibits MMP1 (for example in Leivonen et al. 2013). However, contrary to this hypothesis, MMP1 did not positively or negatively respond to TGF-β stimulation in any of the age or senescent cell lines and this result is robust across all 9 cell lines under study (Figure 5.3a). A hypothesis to explain this apparent discrepancy is that in (White et al. 2000, Edwards et al. 1996, Yuan & Varga 2001), TGF-β only inhibited an increase in MMP1 production that was induced by either TPA or IL-1β. On the other hand, we have measured the MMP1 response to TGF-β directly, without prior stimulation with a positive regulator of

MMP1. Thus a plausible conclusion consistent with all the evidence is that TGF-β inhibits induced MMP1, but not basal MMP1 production. Another interesting aspect of this result is that without the time matched controls, the measured response from MMP1 would appear to be down regulated by TGF-β, which would have confounded our understanding. The data in Figure 5.3a emphasises the importance of time matched controls when studying the dynamics of biological systems.

The dynamic profile of MMP1 is also interesting because in all cell lines and treatments, the amount of MMP1 started high and was reduced only as a consequence of time. The severity of the reduction in MMP1 expression was much greater in adult and senescent cells compared to neonatals due to a larger initial amount of MMP1. It is not clear why this has occurred. A hypothesis is that this is an artefact of the experimental protocol since the fibroblasts are cultured *in vitro* which is outside of normal *in vivo* conditions.

While MMPs degrade the ECM, TIMPs are inhibitors of MMPs and therefore inhibit the degradation of the ECM. TIMPs are often reported to be upregulated by TGF-β (Lin et al. 2017, Park et al. 2015, Kwak et al. 2006, Leivonen et al. 2013). However, the data in this experiment do not support TIMPs as transcriptionally upregulated by TGF-β in fibroblasts (Figure 5.4a). In fact, the opposite was found in that TGF-β induced a reduction in the amount of TIMPs being produced, though it took >48h to do so. The TIMP3 response to TGF-β was only increased in the TGF-β group of cell line C (Figure 5.4b). TIMP2 was additionally measured in a preliminary experiment but was omitted from the main experiment because is was unresponsive to TGF-β. Assuming the integrity of the data, these results cast doubt on whether TIMP1 is normally expressed in healthy dermal fibroblasts in response to TGF-β. This result requires further experimental validation.

α-SMA

TGF-β induces proliferation of fibroblasts and their differentiation to myofibroblasts (Liu et al. 2016, Negmadjanov et al. 2015). Normally fibroblasts are in a quiescent state, directing the normal homeostasis of dermal tissues. Under physiological responses such

as wound repair, fibroblasts undergo differentiation and change their phenotype to an 'active' myofibroblast state which display characteristics of smooth muscle cells. α-SMA is a marker for myofibroblasts (Zanotti et al. 2010, Evans et al. 2003) which facilitates contraction of a wound (Darby & Hewitson 2007). Figure 5.9a therefore indicates that TGF-β induces differentiation of fibroblasts to myofibroblasts and that this process takes a couple of days. Note that this time frame is consistent with how long it takes for the largest quantities of collagen and a few other proteins to be produced in response to persistent TGF-β stimulation (Figure 5.2). Assuming α-SMA is an accurate marker for myofibroblasts, Figure 5.9a shows that adult fibroblasts ability to differentiate is severely impaired compared to neonatal and damage induced senescent fibroblasts. The implications of this hypothesis are profound. It means that in older cells, when the dermis needs to produce large quantities of ECM, that it is unable to do so with the same vigour as young cells. Therefore, lack of differentiability may mechanistically be related to why wound healing takes longer in the elderly.

TGF-β signalling

Smad3 is a prototypical effector of TGF-β signalling and essential for the transcription of type 1 collagens Runyan et al. 2003. Purohit et al. 2016 observed reduced Smad3 levels in both age and senescent fibroblasts and showed using silencing RNA experiments that reduced Smad3 could account for the observed reduction in type I procollagen in the aged dermis. While it is certainly feasible that less Smad3 in adult cells would cause reduced collagen production, Figure 5.11a only provides weak support for this hypothesis since the differences between adult and neonatal cell lines are not so profound. However, just because this data does not provide good support for a transcriptional mechanism of Smad3 decline in age does not preclude the possibility that Smad3 protein levels are reduced in age because of an alternative mechanism, such as enhanced Smad3 degradation. Therefore the role of Smad3 in ageing fibroblasts is still an open question, but based on this data it is unlikely to involve a transcriptional mechanism.

Another interesting aspect about the Smad3 data is that 8-12h post stimulation by

TGF-β, Smad3 mRNA levels are reduced to what appears to be a new stable steady state (Figure 5.11a). This observation suggests the existence of a late acting negative feedback in the TGF-β response to persistent TGF-β stimulation. To our knowledge, this insight into Smad signalling is novel. It is noteworthy that the timing of the drop in Smad3 mRNA levels directly precedes the incline in α-SMA and so a hypothesis is that this drop in Smad3 mRNA occurs before or during fibroblast differentiation.

While Smad3 is the most important effector Smad for ECM regulation (Li et al. 2003), Smad7 is the most important negative feedback of the Smad system (Hayashi et al. 1997, Nakao et al. 1997, Yan et al. 2016, von Gersdorff et al. 2000, Ebisawa et al. 2001, Hanyu et al. 2001, Pulaski et al. 2001, Suzuki et al. 2002, Shi et al. 2004, Zhang et al. 2007). Smad7 both negatively regulates Smad signalling and represents a mechanism of cross-talk with other signalling pathways (Yan & Chen 2011). A considerable portion of the work that studied Smad7 has been conducted on keratinocyte (HaCaT) cell lines and indicate that Smad7 levels peak at approximately 1h post-TGF-β stimulation (Denissova et al. 2000). Figure 5.11b provides evidence that in fibroblasts, Smad7 mRNA levels peaks at 1-3h post-TGF-β stimulation. In neonatal and senescent cell lines, a second peak in Smad7 production was observed 48-72h post-TGF-β stimulation. In adult cell lines the magnitude of the first peak is markedly reduced compared to neonatal cell lines. It is not clear what the biological purpose of this second peak in Smad7 production is but given that Smad7 inhibits the Smad signalling pathway, its likely that canonical Smad signalling is not active at these later time points. It is also unclear what the biological implications of reduced Smad7 production in age has on the TGF-β biology. It may be that Smad7 production is lower in adult cells because Smad signalling is impaired and requires less inhibition.

Smad7 is the prototypical negative feedback of the Smad system. Ski and SNoN on the other hand, are two structurally related negative regulators of Smad proteins that operate by physically binding to nuclear Smad complexes (Akiyoshi et al. 1999). Mechanistically, they interact with phosphorylated Smad2/3 and Smad4 and sequester them in the cytoplasm (Kokura et al. 2003). In the nucleus, Ski and SNoN inhibit Smad transcriptional activity (Wu et al. 2002, Stroschein et al. 1999, Luo 2004, Deheuninck &

Luo 2009, Yan et al. 2017). Ski and SNoN can occupy Smad binding elements on gene promoters and inhibit transcription by recruiting corepressors (Yan et al. 2017). On stimulation by TGF-β, SNoN interacts with Smurf2 leading to its proteasomal degradation before once again being transcribed a 2 hours later (Stroschein et al. 1999, Bonni et al. 2001). The present study supports the notion that both Ski (Figure 5.11e) and SNoN (SKIL) (Figure 5.11f) are negative feedback regulators of the TGF-β system. In line with Stroschein et al. 1999, Ski and SNoN took 2h to be induced in TGF-β stimulated fibroblasts of all ages. Since neither of these datasets show a decrease in transcription <2h, they do not directly support the negative regulation of Ski and SNoN at early time points but neither do they conflict with this idea. Figure 5.11e provides weak support for loss of Ski abundance with age.

AP1

As discussed in detail in Chapter 4, TGF-β has many lines of cross-talk with other signalling pathways. AP1 transcription factors are heterodimers composed of a Jun and Fos protein and are known Smad interacting partners (Zhang et al. 1998). AP1 dimers have been implicated as part of the antagonistic relationship between pro-inflammatory TNF-α and anti-inflammatory TGF-β signalling (Verrecchia et al. 2000). Specifically, c-Jun has been shown to act as a transcriptional co-repressor for Smad signalling (Atfi et al. 1997, Wendling et al. 2003, Chung et al. 1996) but at the same time is involved in upregulation of proteins involved in ECM degradation including MMPs (Benbow & Brinckerhoff 1997). c-Jun is also an effector of increased MMP levels in context of photoageing (Shi & Ruan 2013). Conversely, JunB containing AP1 complexes are transcriptional co-activators as well as a transcriptional target of Smads (Ponticos, Harvey, Ikeda, Abraham & Bou-Gharios 2009). The antagonistic relationship between c-Jun and JunB containing AP1 complexes is also reflected in collagen regulation, since c-Jun inhibits and JunB activates COL1A2 transcription (Ponticos et al. 2015, Ponticos, Holmes, Shi-wen, Leoni, Khan, Rajkumar, Hoyles, Bou-Gharios, Black, Denton, Abraham, Leask & Lindahl 2009).

Figure 5.12e and Figure 5.12f show the transcriptional response of c-Jun and JunB respectively in fibroblasts. In neonatal cells, c-Jun is progressively upregulated under basal conditions but transcriptionally inhibited by TGF-β. An important and interesting finding is that c-Jun levels are not only reduced in age, but the ability of TGF-β to antagonize c-Jun is abrogated (Figure 5.12e). JunB was confirmed to be transcriptionally responsive to TGF-β with a sharp peak occurring at 2h and subsequent inhibition 3-4h after TGF-β stimulation Figure 5.12f. Although Ponticos et al. 2015 implicates JunB in the production of collagen, its specific role is not clear. Given the disparate time frame of how JunB compared to COL1A1/2 respond to TGF-β, it is unlikely that JunB is a major factor directly mediating collagen production but more likely that JunB plays a role in setting the cellular state prior to efficient collagen production. The main finding regarding JunB is that in all adult cell lines tested, the magnitude of the JunB response to TGF-β was blunted (Figure 5.12f), which is consistent with the observed reduction in adult cells to produce collagen (Figure 5.2) and with the idea that adult fibroblast are less able to prepare for an efficient anabolic response to TGF-β. It is noteworthy that the JunB and Smad7 (Figure 5.11b) dynamic profiles are very similar, suggesting a potential coregulation.

5.4.1 PI3K

TGF-β communicates with signalling nodes such as PI3K (Wilkes et al. 2005, 2003, Wilkes & Leof 2006) and Rho GTPases (Atfi et al. 1997). PTEN (Figure 5.12h) production is enhanced by TGF-β treatment and this response is blunted or abrogated in adult cell lines. Since PTEN is a negative regulator of PI3K signalling, these data suggest that TGF-β induces a late activating negative feedback to terminate PI3K signalling. Similarly, RhoB (Figure 5.13c) responded to TGF-β in neonatal but not in adult cell lines suggesting a reduced response in adult cell lines.

Data preprocessing and uncertainty

In Livak & Schmittgen 2001 the $2^{-\Delta\Delta C_T}$ method was presented for normalisation of qPCR data. As described in the methods, in the $2^{-\Delta\Delta C_T}$ method of normalisation, a signal of interest is divided by a reference gene from the same sample and then by a control sample (known as a calibrator), which is normalised to the same reference gene. In this experiment however, the data were presented as $2^{-\Delta C_T}$ values instead of $2^{-\Delta\Delta C_T}$ because the alternative was to decide on a calibrating sample, which highlights one aspect of the data whilst impeding the interpretation of another. For example, one option for the calibrating normalization was to divide the TGF-β treated samples by the time matched controls. The problem with this is that because the adult cell lines had less of some genes in the basal condition, the corresponding increase in response to TGF-β was more than in the neonatal cell lines, even though the neonatal cell lines produced more. The data presented in this way can be misleading as it highlights the effect of TGF-β at the cost of being able to adequately compare the control cells. As an alternative to using the time matched controls, it is also possible to use the neonatal cell line as calibrator for the senescent and adult cell lines. This however highlights the differences between cell type at the cost of clarity in how the neonatal cells respond over time. Additionally, since there are three cell lines for each neonatal, senescent and adult groups, by pairing (say) adult cell line G with neonatal cell line A, information is lost regarding the potential pairing between cell line G with cell lines B and C. To circumvent the need to choose a normalisation or to perform multiple normalisations the data were presented as $2^{\Delta C_T}$ values. This enables optimal interpretability of the data as it leaves the choice of which aspect of the experiment to look at to the reader.

While the data in Figures 5.2 to 5.14 were presented as $2^{-\Delta C_T}$ values, the differential expression analysis used $2^{-\Delta\Delta C_T}$ values to reflect the question being asked. In the baseline analysis, the 96h were calibrated to the 0h baseline samples so that the data being compared between cell lines represented the amount of mRNA produced over 96h. In the time series analysis, the concern was whether adult or senescent cells responded differently to TGF-β compared to the neonatal cell lines.

Instead of making an arbitrary choice for the parameters of the differential expression analysis, the entire process was repeated 10,000 times with random parameters so that the proportion of times a gene is considered differentially expressed can be computed. One aspect that was randomised is the pairing between replicates. The rational is that it is only luck that pairs the a particular TGF-β sample to a corresponding control and it is no less valid to pair that sample with another sample, provided it is from the same experimental condition. Another aspect that was randomized was the choice of baseline sample for the calibrating normalisation of the time series analysis, since it is equally valid to choose the 0h, 96h or the average of the two. Moreover, instead of pairing a particular neonatal cell line with a particular adult or senescent cell line, since all combinations of pairing are valid, this choice was randomized throughout the resampling procedure. The end result is that the variability inherent in the data was exploited to ascertain confidence intervals concerning whether a gene is differentially expressed in the baseline (Figure 5.15) or time series (Figure 5.16) analysis.

Limitations

A limitation of this work is that only the mRNA level of biological organisation was measured, whereas protein function occurs at the protein level. Since there is not necessarily a one-to-one correspondence between mRNA and protein (Liu et al. 2016), it would be illuminating to perform some parallel proteomic and phosphoproteomic experiments to provide a more comprehensive understanding of the underlying biological system. Moreover, some of the results (i.e. MMP1 Figure 5.3a) would benefit from experimental validation using an alternative method (i.e. ELISA, western blots, perturbation studies).

Another limitation of this work is that irradiation-induced senescence was used as a model for replicative senescence. It has been shown that there strong similarities between the replicative and irradiation-induced senescence (Marthandan et al. 2016), but there still may be important differences which should be considered when drawing conclusions about replicative senescence from a irradiation-induced senescence model.

Other limitations include a potential issue with TGF-β stability over the 96 hours in culture. Some of the genes (i.e. Figure 5.2a) appeared to be down regulated at the later time point but it is not clear whether this down regulation is a result of a negative feedback or a lack of TGF-β stimulation. Since the stability of TGF-β is unknown, it is possible that after 96h, the available TGF-β may have been depleted. Moreover, throughout the experiment the cells were starved of nutrients. This was necessary because serum used in tissue culture contains TGF-β and so to remove the potential confounding variable, the cells were treated with TGF-β in media without serum. It is not known whether these fibroblasts behave differently as a result of not having these nutrients but it possible that the responses observed are not fully representative of the underlying biology.

Finally, a critical aspect of fibroblast biology is that they normally exist under mechanical tension. Young healthy fibroblasts adhere to the ECM via integrin connections to ECM components such as collagen and fibronectin (Cole et al. 2018). In age this process is affected because of damaged collagen and less ECM. In this experiment the cells were cultured in a 2D-monolayer which are under completely different mechanical tensions than they would be in a 3D environment. An understanding of how these components respond in 3D would be useful for elucidating potential confounding factors derived from 2D *in vivo* culture and for understanding how fibroblasts respond in a more biologically realistic environment.

Conclusion

In this chapter we have discussed the results of a high throughput qPCR experiment that was designed to study transcriptional differences in neonatal, senescent and adult fibroblasts in response to TGF-β or a control. Collectively several useful insights have been discussed regarding how fibroblasts differ with age and in response to TGF-β.

Chapter 6

A systems modelling investigation into TGF-β cross-talk with CTGF in neonatal and adult fibroblasts

6.1 Introduction

TGF-β is involved in diverse biological contexts (Zhang 2018). Such diversity is possible because of the extensive interconnectivity between TGF-β and other signalling systems (Zhang 2009). Dysregulated TGF-β signalling is a feature of numerous disease states, including various cancers (Hamidi et al. 2017, Freudlsperger et al. 2013) and fibrotic disorders (Morris et al. 2011, Samuel et al. 2010). Moreover, changes in TGF-β signalling with age has been associated with skin ageing (Purohit et al. 2016, Fisher et al. 2016).

Loss and disorganisation of dermal collagen fibrils is a prominent feature of skin ageing (Quan & Fisher 2015, Cole et al. 2018). TGF-β is a known positive regulator of collagen production and its functional decline with age contributes towards loss of collagen with age (Quan et al. 2004, Purohit et al. 2016). Experimental evidence suggests that Smad3 levels are correlated with collagen levels in fibroblasts and that old fibroblasts have less Smad3 than young fibroblats (Purohit et al. 2016). Fisher et al. 2016 found that

reduced fibroblast spreading in age was associated with reduced levels of TGF-β receptor type II, which exacerbates the aged dermal phenotype. Moreover, the environment in which fibroblasts reside influences its ECM producing abilities. Reduced collagen levels and increased collagen fragmentation with age are associated with reduced fibroblast adherence to the ECM, which changes its morphology and responsiveness to extracellular stimuli (Cole et al. 2018).

CTGF is produced in response to TGF-β and plays a significant role in TGF-β induced collagen production (Quan et al. 2010, 2002, Grotendorst 1997, Chapter 5). CTGF levels are increased in fibrotic diseases such as systemic sclerosis (Makino et al. 2017) and reduced in aged dermis (Quan et al. 2010). The pathways which transduce CTGF signalling are not well understood. CTGF can bind to both integrin (Weston et al. 2003, Asano et al. 2005, Nakerakanti et al. 2011) and TrkA (Wahab, Schaefer, Weston, Yiannikouris, Wright, Babelova, Schaefer & Mason 2005) receptors and physically bind both BMP4 and TGF-β in the extracellular space (Abreu et al. 2002). The interaction between CTGF and TGF-β potentiates TGF-β signalling while the corresponding interaction with BMP4 is inhibitory to BMP4 signalling (Abreu et al. 2002).

A large body of evidence indicates that Erk1/2 signalling is involved in mediating the synergy between TGF-β and CTGF (Arnott et al. 2008, Cheng et al. 2015, Leask et al. 2003). Erk1/2 is directly activated by TGF-β stimulation. Erk1/2 phosphorylates Ets1 (Haines et al. 2011, Plotnik et al. 2014) a known activator of both collagen (Asano et al. 2009, Czuwara-Ladykowska et al. 2002) and CTGF (Nakerakanti et al. 2006, 2011) production. Moreover, Erk1/2 can phosphorylate Smad3 in the linker region which is a different site to the Smad activating phosphorylation and increases Smad3 stability (Hough et al. 2012).

In addition to being produced by TGF-β, CTGF is capable of modulating TGF-β signalling by inhibiting Smad7, the canonical Smad negative regulator (Hayashi et al. 1997). Tieg1 is a negative regulator of Smad7 transcription and its expression is induced by CTGF (Hu et al. 2017, Wahab, Weston & Mason 2005).

These data together provide evidence that CTGF modulates TGF-β signalling; that

CTGF and TGF-β cooperate in the production of collagen; and that the mechanisms controlling TGF-β and CTGF interactions may be modified in the aged dermis compared to young. This chapter discusses the development of a mechanistic ODE model of TGF-β signalling that is modulated by CTGF and differences between young and old fibroblasts. While currently incomplete, the model has been developed as far as possible with the available data.

6.2 Methods

6.2.1 Experimental Data

This model was informed using experimental data collected in Chapter 5. The scope was limited to four genes from cell lines C and I to represent neonatal and adult cell lines respectively. These were chosen as they had the largest difference in collagen quantities. The measurements used were COL1A1, COL1A2, Smad7 and CTGF mRNA from both TGF-β treated and control conditions. The baseline 0h time point was used for both control and treated time series and the 6 replicates were averaged prior to estimation.

Experimental data for the model calibration was configured using four individual tab separated files containing details of the four different conditions and the relevant independent variables. Each file contained measurements of the four genes and an independent variable column containing the starting values information for that condition. The starting values were set to empirical values under the baseline 0h condition.

An independent variable was used in each of the four files to indicate whether the data was collected from adult or neonatal cell lines. Namely, the 'Adult' variable was set to 1 if the data was from adult cell lines or 0 otherwise. The 'Adult' variable was then multiplied by an appropriate rate law to toggle an appropriate reaction on and off. Another indicator variable was used to denote whether the data was collected from

TGF-β or the control treatment and was fixed throughout simulations.

6.2.2 Model Construction

The base model and four hypotheses were constructed in separate modules using the antimony model definition language (Smith et al. 2009). All combinations of hypotheses were constructed using Python's `itertools` library and basic string concatenation. Tellurium (Choi et al. 2018) was used to convert the models into SBML (Hucka et al. 2003) which was then imported into COPASI (Hoops et al. 2006). This process was handled automatically with PyCoTools (Welsh et al. 2018).

6.2.3 Model Calibration

All 16 models were calibrated in exactly the same way. The reactions represented with a red arrow in Figure 6.1 were fixed to the same values throughout calibration of each model while the black arrows were estimated between the boundaries of $1{\cdot}10^{-12}$ and $1{\cdot}10^{4}$. The fixed values were chosen based either on literature or on the (Zi & Klipp 2007) model. The estimation boundaries were based on preliminary estimations using the base model which indicated that some of the reactions required a low reaction rate. Basal transcription parameters were constrained to be lower than induced transcription parameters. The residual sum of squares (RSS) objective function was weighted by using the mean squared and weights were normalized for each experiment. All configuration was conducted automatically using PyCoTools (Welsh et al. 2018), except that of constraints which was performed manually for the 16 models using COPASI.

6.2.4 Model Selection

Models were selected based on the Akaike information criteria (AIC) corrected for small sample sizes (AICc) (Hurvich & Tsai 1993). The AICc was calculated using

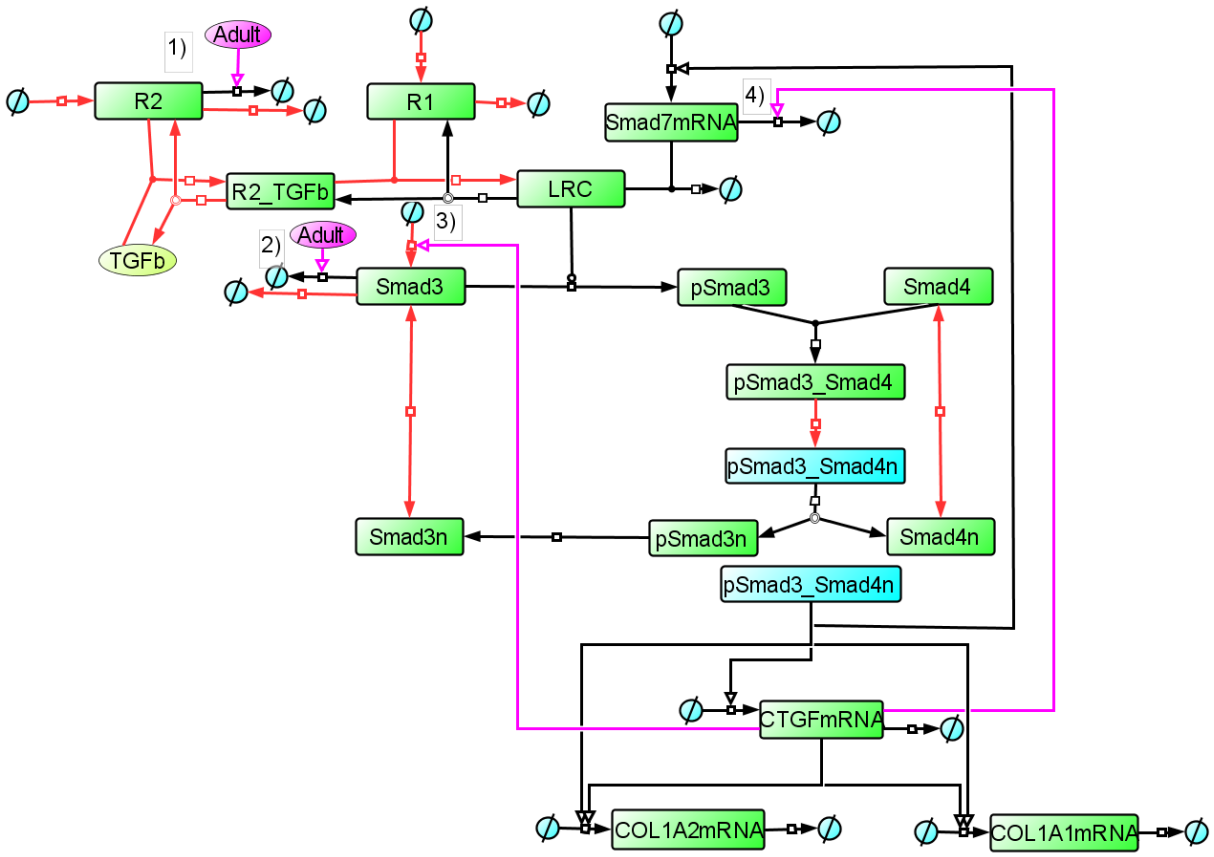


Figure 6.1: Graphical depiction of ODE model network. All reactions are mass action. The 'Adult' component is an indicator variable. Red edges indicate fixed parameters. Black edges indicate estimated parameters. Pink edges are alternative hypotheses.

Equation 6.1

$$AICc = -2ln\left(\frac{RSS}{n}\right) + 2K + \left(2K\frac{K+1}{n-K-1}\right) \tag{6.1}$$

where:

RSS = Residual sum of squares objective function value

n = Number of data points used throughout entire calibration

K = Number of model parameters

6.3 Results

6.3.1 Model Construction

The reaction schema depicted in Figure 6.1 is mostly an abstraction of core TGF-β signalling and is comprised entirely of mass action reactions (model equations in Section B.1. Because the network is largely unobserved, an attempt was made at fixing as many parameters as possible prior to estimation. Moreover, translation reactions were not included. Instead, mRNA products were used under the assumption that the protein levels followed the same trajectory as the transcript levels.

Receptors

Type 2 receptors have been shown to bind reversibly to TGF-β with a forward rate constant of 1.38 nmol L^{-1} min^{-1} and a backwards rate of 0.009 nmol min^{-1} (De Crescenzo et al. 2003). The backwards rate was increased in the model to 0.5 nmol min^{-1} to control the levels of available $R2\_TGFb$. This parameter was not estimated because this part of the network was unobserved. The rate constants for the $R2\_TGFb$ complex binding to $R1$ are unknown. Since this is a forwards and backwards reaction, it is likely a structurally non-identifiable reaction and so the forwards rate was fixed to 1 nmol min^{-1} L^{-1} while the backwards rate was estimated.

The rates of $R1$ and $R2$ production and degradation were set to values so that the steady state levels were 30 and 45 nmol L^{-1} when the 'Adult' and 'TGFb' indicator variables were set to 0. The degradation reactions were set to 0.0278 min^{-1} as in Vilar et al. 2006.

Earlier models of TGF-β signalling placed importance on TGF-β receptor internalization on the dynamics of TGF-β signalling (Vilar et al. 2006, Zi & Klipp 2007). This decision was based on data from Di Guglielmo et al. 2003 describing TGF-β receptor compartmentalisation into early endosomes or caveosomes as a determinant of whether TGF-β receptors phosphorylate downstream Smads or are degraded by signals

initiated by Smad7, respectively. This feature was abstracted out of the current model because these vesicles can fuse intracellularly and display both types of behaviour (He et al. 2015, Di Guglielmo et al. 2003). Furthermore, receptor stoichiometry was not taken into account.

Smad3 binds constitutively to type 1 TGF-β receptors (Li et al. 2016). However, including these reactions and the rate constants measured by Li et al. introduced undue complexity given the scope of the model and so Smad3 phosphorylation was abstracted to a simple mass action reaction catalysed by the ligand-receptor complex (LRC). The LRC was subjected to TGF-β dependent degradation by $Smad7$ which is produced by nuclear Smads.

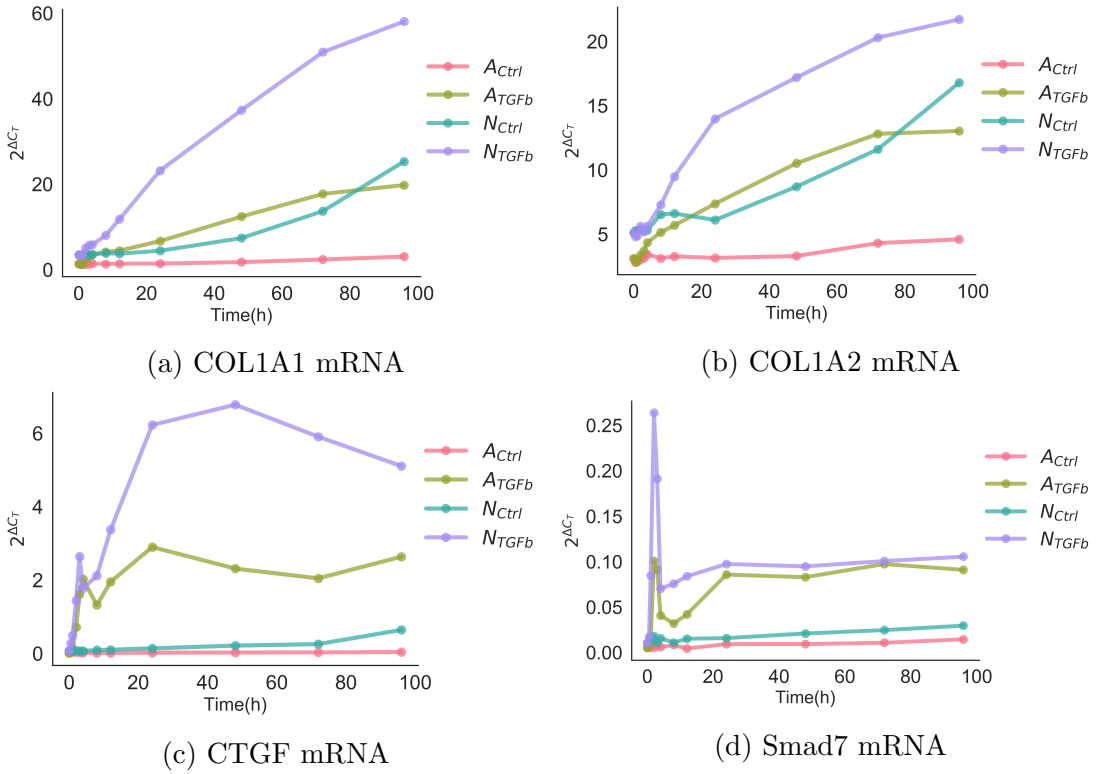
Smad Shuttling and Transcription

Total levels of $Smad4$ were constant throughout simulations. The initial levels of $Smad4$ were set to steady state with the $Adult$ and $TGFb$ indicator variables set to 0. Care was taken so that the amount of $Smad4$ in the system was not a limiting factor for $pSmad3\_Smad4$ formation and transition to the nucleus.

Production and degradation reactions were included for Smad3 so that they can be modulated by age and CTGF. These two reactions were not estimated because $Smad3$ is not an observable and the resulting estimations would likely be poor. Instead they were set to steady state when $Adult$ and $TGFb$ indicator variables were set to 0.

The $Smad3$ and $pSmad3\_Smad4$ import rate constants were set to 0.16 nmol min^{-1}. The $Smad3n$ and $Smad4n$ export rate constants were set to 1 nmol $min^{-1}$and 0.5 nmol min^{-1}respectively. The $Smad4$ import rate constant was set to 0.08 min^{-1}. These parameters were taken from Zi & Klipp 2007. Although Smad3 mRNA levels were observed to decline in Chapter 5 this mechanism was not modelled.

$pSmad3$ binds and unbinds $Smad4$ in the cytoplasm and nuclear compartments respectively. $pSmad3$ was assumed to require Smad4 for nuclear import and is not directly exported from the nucleus but dephosphorylated with mass action kinetics and



(a) COL1A1 mRNA

(b) COL1A2 mRNA

(c) CTGF mRNA

(d) Smad7 mRNA

Figure 6.2: Experimental data showing the trajectory of (a) COL1A1, (b) COL1A2, (c) CTGF mRNA and (d) Smad7 mRNA in response to TGF-β or control in adult (A) or neonatal (N) cell lines. These data represent the averages of the data presented in Chapter 5.

then exported as non-phosphorylated *Smad*3.

*Smad*3_*Smad*4*n* induces transcription reactions, all of which were modelled as the sum of basal constant production rates and Smad induced rates. CTGF was assumed to induce collagen production independently of Smads and the degradation rates of COL1A1 and COL1A2 were assumed the be the same. Basal rates were constrained so that they are lower than their respective induced production rates.

Compartment Volumes

The model contained three compartments, Cytoplasm, Nucleus and Medium. The nuclear compartment was assumed to be $1{\cdot}10^{-12}$ L^3. Since the ratio of nuclear to cytoplasmic volume is not known, the cytoplasm was assumed to be three times larger than the nucleus. The naming convention used was that model components with n appended to the name denote model variables which reside in the nuclear compartment

and all other model components are in the cytosolic compartment, except $TGFb$ which was in the medium. The medium volume was set to $2.5 \cdot 10{-}8$ which is an approximation of the amount of volume 'visible' to a single cell. However, since the amount of $TGFb$ was fixed throughout simulations, this detail is only relevant for future extensions of the model that encompass other ECM components. Care was taken to scale multi-compartment reactions by the correct compartment volume.

6.3.2 Extension Hypotheses

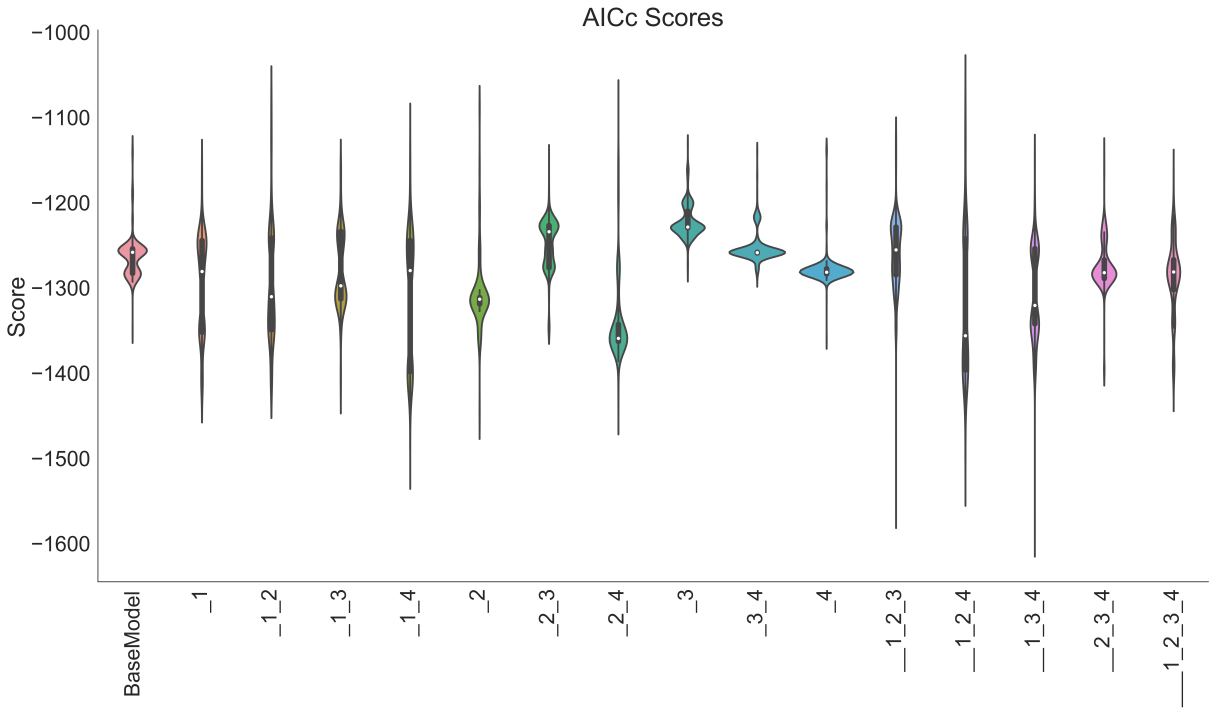
So far, we have only described the base model which is used in all other models, exactly as described above. Four additional extension hypotheses were proposed, all based on literature. In Figure 6.1, these four hypotheses are represented by pink arrows. Two of these hypotheses involve putative influences of age on collagen output while the other two involve potential modes of TGF-β modulation by CTGF.

Hypotheses 1 and 2

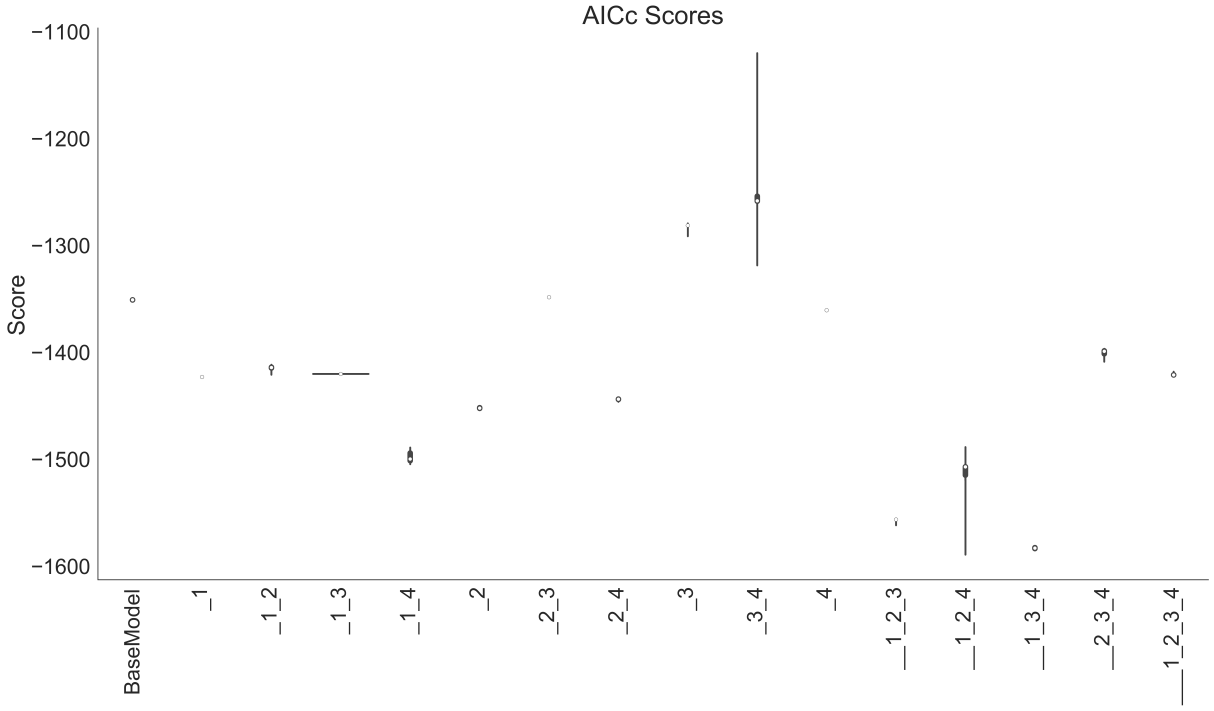
Levels of both TGF-β type 2 receptors (Fisher et al. 2016) and Smad3 (Purohit et al. 2016) have been implicated in the observed reduction in collagen levels. These mechanisms are introduced when the *Adult* indicator variable is set to 1, with additional the reactions representing the increased degradation of $R2$ ($H1$) and $Smad3$ ($H2$). These are hypotheses 1 ($H1$) and 2 ($H2$) respectively. Note that increased degradation was used rather than decreased production so that mass action rate laws can be used and to prevent the need for more complicated rate laws such as hill equations that would require greater data demands than available to parametrize.

Hypothesis 3

CTGF reportedly activates Erk1/2 (Arnott et al. 2008, Cheng et al. 2015, Leask et al. 2003) which in turn has been shown to phosphorylate Smad3 (Hough et al. 2012) in the



(a)



(b)

Figure 6.3: Model selection criteria for (a) fit 1 and (b) fit 2. Violin plots display the median of a distribution as a central point, the interquartile range as a thick central bar, the 95% confidence interval as the thin central bar and the thickness of the violin represents a kernel density estimation of the frequency of fits with the observed AICc score. The x-axis represents each of the model combinations which are labelled according to which hypotheses are contained in that model, delimited by an underscore.

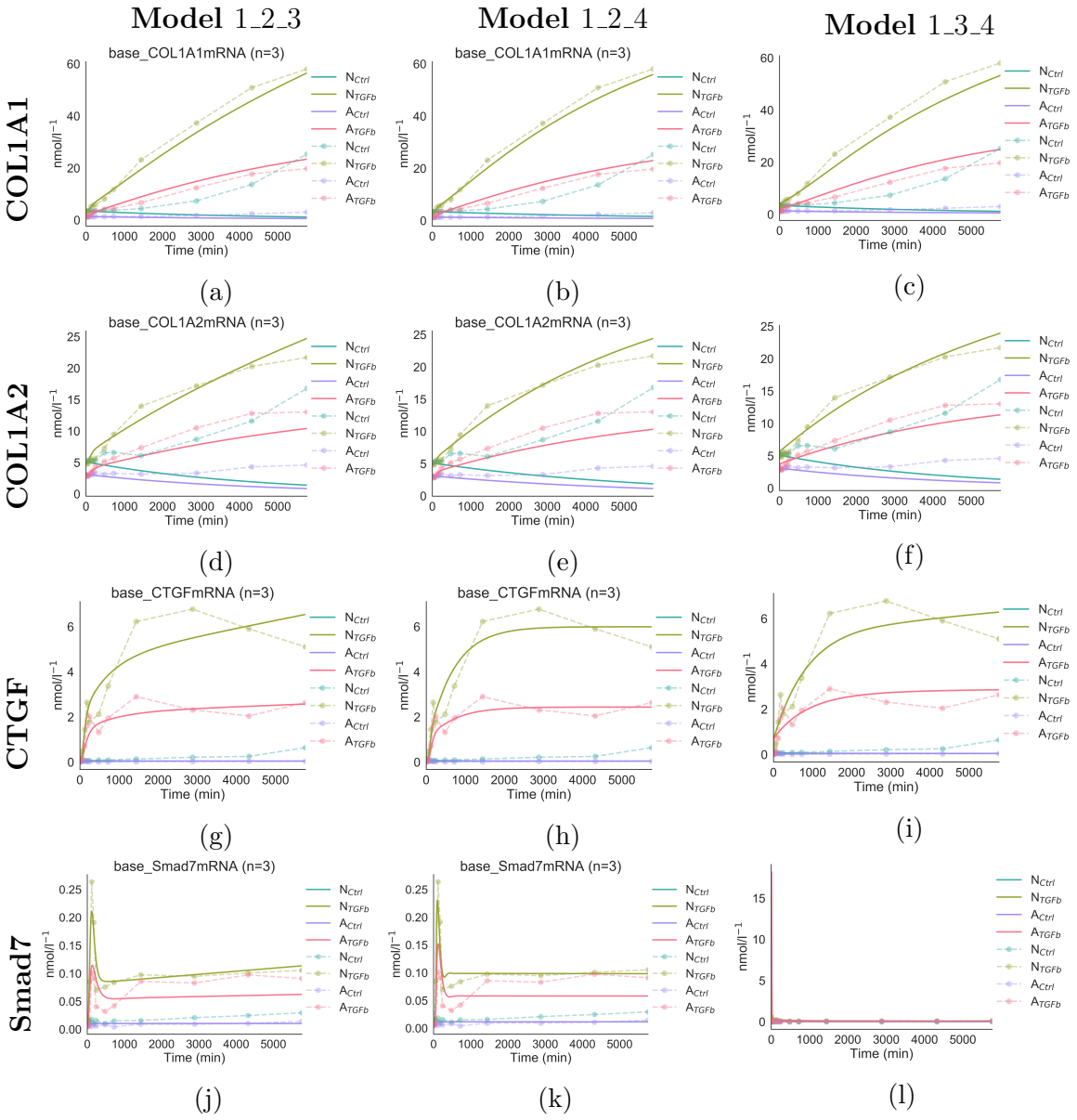


Figure 6.4: Profiles for best fitting parameter sets in models 1_2_3 and 1_2_4 and 1_3_4.

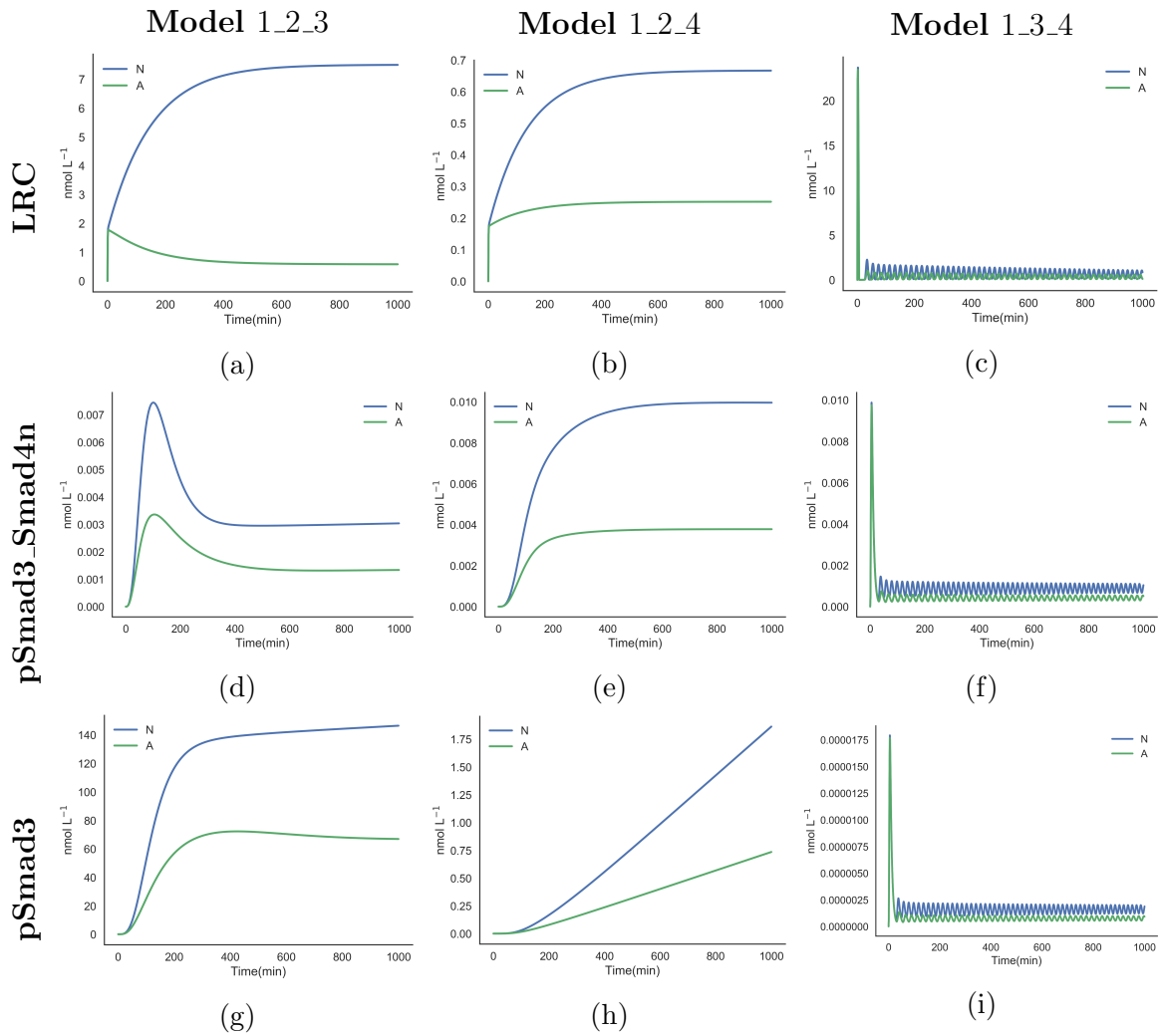


Figure 6.5: Simulation comparison of other profiles in models 1_2_3 and 1_2_4 and 1_3_4.

linker region. Linker Smad3 phosphorylation enhances Smad3 stability so CTGF activation of Erk1/2 may lead to enhanced Smad3 stability. The additional intermediary reactions were omitted due to lack of experimental data.

Hypothesis 4

Smad7 is the major negative regulatory Smad as it is produced in response to TGF-β and terminates Smad signalling. CTGF has been shown to upregulate an inhibitor of Smad7 transcription, Tieg1 (Hu et al. 2017, Wahab, Weston & Mason 2005) which is hypothesis number 4 ($H4$). Again, the intermediate reactions with Tieg1 were omitted because of a lack of data.

6.3.3 Model Calibration

Model calibration was conducted using COPASI and configured using PyCoTools as described in the methods. The data used in the calibration has been replotted from Chapter 5 in Figure 6.2 for convenience. These data represent averages of 6 repeats and the errors are not displayed because they were not used for estimation purposes. Model calibration was conducted in two stages, denoted 'Fit1' and 'Fit2'. The RSS was minimized using the particle swarm algorithm as implemented in COPASI (Hoops et al. 2006). For Fit1, 400 parameter estimations (swarm size=400, iteration limit=3000) were conducted per model with randomized initial conditions. The best parameter sets from Fit1 were inserted into their respective models and 200 more parameter estimations swarm (size=200, iteration limit=2000) were conducted per model without randomizing the initial conditions.

The distribution of AICc scores are displayed in Figure 6.3. Of the 15 additional model hypotheses, only models 3 and 3_4 made the model fits worse. These two models only contained $H3$ and $H4$ both of which pertain to interactions between CTGF and TGF-β, rather than a mechanism to explain the data from old fibroblasts. Three of the four '3 tier' combination hypotheses were the best performing of all the tested models. Notably, model 2_3_4 which lacked repression of $R2$ in age performed considerably worse than model 1_2_3, model 1_2_4 or model 1_3_4, providing evidence that $H1$ is more important for fitting the age dimension of this dataset than $H2$.

Figure 6.4 shows the best fitting profiles for models 1_2_3, 1_2_4 and 1_3_4. All three models were equally capable of fitting COL1A1 and COL1A2 profiles under the TGF-β condition while none of the models were capable of fitting the basal condition, regardless of age group. The CTGF profile dynamics cannot be completely captured by any of these model topologies. The Smad7 profile fits model 1_2_3 and 1_2_4 equally but for model 1_3_4 Smad7 is over-produced at early time points which is masking this graphs interoperability.

Since model 1_3_4 technically has the best AIC scores and yet appears to have the worst

fit in Figure 6.4, the models with their best fitting parameter sets were investigated further by simulating a time series from each and comparing the non-observed profiles. Figure 6.5 shows this comparison for $pSmad3$, $pSmad3\_Smad4$ and LRC which clearly demonstrates that model 1_3_4 has found an over-fitting sustained oscillatory parameter set. The profiles from the other two models are considerably different except the $pSmad3\_Smad4$ profile.

6.4 Discussion

With age, skin becomes progressively thinner and less elastic. These properties make old skin less able to perform its functions, more prone to diseases and less aesthetically pleasing compared with younger skin. It is therefore desirable from both a therapeutic and cosmetic perspective to find new and reliable means of manipulating the biochemical signalling networks that regulate skin.

TGF-β is a major regulatory cytokine in the deposition and degradation of ECM components. TGF-β is known to interact with a variety of other biochemical signalling systems which confounds our understanding of how TGF-β operates. In this chapter, we began the process of building a mechanistic model which describes TGF-β induced collagen production in young and old fibroblasts and the involvement of CTGF in these processes.

In Chapter 5 the results of a high throughput experiment was discussed, where we measured the dynamics of many genes pertinent to ECM biology in neonatal and adult fibroblasts treated with TGF-β or control (Figure 6.2). In this chapter we take some of these observations together with knowledge from literature and proposed a mechanistic model which abstracts our current understanding of canonical TGF-β signalling and proposes hypotheses for 1) how TGF-β signalling is modified with age and 2) how CTGF and TGF-β synergize in the production of type I collagens (Figure 6.1).

Four hypotheses were proposed, all of which are substantiated by literature. Two

hypotheses, $H1$ and $H2$ relate to how TGF-β signalling is different in aged fibroblasts while the other two, $H3$ and $H4$, relate to how CTGF modulates TGF-β signalling in age. To discriminate between these hypotheses, a series of models were systematically constructed that such that all possible combinations of model hypothesis were represented. Each model was calibrated under exactly the same conditions so that the models differed only in topology. The AICc Equation 6.1 was calculated for each parameter set of each model and the results displayed using violin plots which is a hybrid of a box plot and kernel density estimation plot (Figure 6.3).

The results show that models without one of the ageing hypotheses ($H1$ and $H2$) perform worst in fitting the experimental data in Figure 6.2. This is intuitive because without an ageing hypothesis the model would be unlikely to fit the adult data. Three model hypotheses, model 1_2_3, 1_3_4 and 1_3_4 outperformed the rest in fitting this data. Comparing these models with model 2_3_4 suggests that is $Smad3$ down-regulation in age is less important than $R2$ down regulation in age.

Since these three model hypotheses had similar model selection scores, they were selected for further analysis. Experimental versus simulated time series plots were generated using the best fitting parameter sets for each model (Figure 6.4). They suggest that models 1_2_3 and 1_2_4 are virtually indistinguishable in terms of comparison with available experimental data. Model 1_3_4 was similar to the other two regarding collagen and CTGF production but very different regarding the Smad7 profile. This was surprising because model 1_3_4 had the lowest AICc score of all models (Figure 6.3). On closer inspection, Figure 6.5 shows that this model has found an oscillatory parameter and displays over-fitting behaviour.

While the profiles of model observables were very similar between the three best models (Figure 6.4), the unobserved model components displayed very different behaviour (Figure 6.5). Interestingly, even without data, model 1_2_3 displayed a $pSmad3\_Smad4n$ profile that is coherent with what might be expected, based on literature for phosphorylation of Smad2 (Vizan et al. 2013). Therefore, while both model 1_2_4 and 1_2_3 fit the data similarly, model 1_2_3 holds better accordance with the literature. For

this reason, model 1_2_3 is the best of these models in terms of reproducing the observed data.

Some aspects of this modelling effort have worked well but it is clear that the model requires further development in order to better capture the behaviour of all the observables under all conditions. One aspect of all of these models that require more attention is their ability to fit the data from the basal condition. The fact that none of these models captured basal behaviour suggests that basal transcription is more complicated than simple constant production, as it is in the model. It may be prudent to measure the amount of residual TGF-β that is produced by fibroblasts as it is possible that autocrine TGF-β production contributes towards basal behaviour.

In addition to the basal data not fitting the models, the models were unable to fully capture the CTGF profile. The CTGF has a trough 8-12h post TGF-β stimulation Figure 6.2c. The origin of this negative regulation is unknown and none of the models are capable of reproducing this behaviour. It may be that this negative regulation occurs during a time that the cell is in a state where it may or may not differentiate into myofibroblasts. This is a speculation based on the overlapping timings of increases in CTGF and α-SMA production (Figure 5.9a), a marker for myofibroblasts.

There are several limitations with making conclusions based on the current model. First and foremost is that the model is not informed by any protein level data. Because of this multiple assumptions were made in attempt to reduce the size of the parameter space. For example, the initial concentration parameters and several kinetic rate constants were assigned to values used in previous TGF-β models (Zi & Klipp 2007) while others were guessed based only on intuition. This lack of data means that many model parameters are not identifiable. An identifiability analysis was not conducted for these models because they have more serious problems than identifiability, namely topological inaccuracies. The topology requires adjusting prior to repeating the calibration process and then, when the model behaviour adequately matches the experimental data, an identifiability analysis becomes more relevant. In the process of resolving the topological issues, new data should be produced and this data will

contribute towards resolving the identifiability issues.

Another limitation of this model is that the model variables representing mRNA were used in reactions as a surrogate for their corresponding proteins. This means the time that it takes for transcription is neglected and this may have an impact on network dynamics. For example, $H3$ and $H4$ are implemented as gross simplifications of two plausible hypotheses based on literature evidence. However, since both hypotheses have intermediary reactions (involving activation of Erk1/2 for $H3$ and production of Tieg1 for $H4$) the signal should be delayed compared to how it is currently modelled.

Another factor which may impact network dynamics is that the Smad shuttling system was designed to have minimal complexity, despite the fact that other models (such as (Schmierer et al. 2008) who's model was dedicated to this issue) used a slightly more complicated topology. The primary goal of both of these simplifications was to reduce the size of the network to make it more suitable for the available data. However, the result is that the proposed mechanisms may not qualitatively represent the dynamics of interest. Rectifying this is a matter for future considerations.

In this chapter a modelling strategy has been devised to exhaustively test combinations of model hypothesis. Four biological relevant hypotheses have been proposed that attempt to explain some of the differences that exist between adult and neonatal fibroblasts. Based on the current modelling efforts, the best model hypothesis was model 1_2_3 as this most closely resembled the experimental data and appeared to be behaving in accordance with intuitive notions for how the Smad system should behave. However, there are many limitations that must be considered when drawing biological conclusions and it is highly likely that more data, particularly protein level measurements, would help reduce these uncertainties. In reality, the data presented in this chapter represents a single turn of the systems biology cycle and the model requires refinement prior to being used to make predictions. Of note, the modelling strategy used in this chapter is generalisable to any systems modelling effort and represents an excellent way of testing multiple topological hypotheses.

Chapter 7

Conclusion

With age the ability for skin to perform its functions becomes diminished. As in ageing in general, the causative factors in skin ageing are complex and not well understood. The dermal ECM is vital for skin integrity and plays a structural and nourishing support roles for the avascular epidermis. One of the phenotypes of skin ageing is that skin becomes thinner with age. At the molecular level, the amount and integrity of type 1 collagen is impaired and this is a major contributing factor towards skin thinning (Cole et al. 2018). Impaired collagen homeostasis can be driven by two potential sources, reduced production and enhanced degradation. Literature reports that both mechanisms are superimposed (Quan et al. 2013, 2002) although the root causes of the change are largely unknown.

Dermal fibroblasts are the caretakers of the dermis: they reside in, produce, maintain and degrade the dermal ECM. Fibroblasts are receptive to their environment. They dynamically adhere to the ECM using transmembrane integrins which provide the fibroblast with information about the state of the dermal ECM. Fibroblasts are also receptive to other signalling molecules, some of which such as TGF-β, are anabolic in nature and induce ECM synthesis from fibroblasts. Others, such as TNF-α are catabolic and induce the production of negative regulators of the ECM. Together TGF-β and TNF-α provide a mechanistic basis for fine tuning the composition of the ECM.

The focus in this work has been on investigating TGF-β in its ability to induce fibroblast

transcription. Two separate but related issues have been addressed: 1) how does TGF-β communicate with other signalling pathways to achieve the observed transcriptional profile and 2) what differences exist between neonatal, senescent and adult fibroblasts in their response to TGF-β? Both of these issues are complex and as expected from such ambitious goals, nether question has been solved. Instead this thesis contributes towards the overall goals of understanding TGF-β cross-talk and skin ageing.

The first question was addressed in Chapter 4 which presents an unbiased, global and dynamic transcriptional profile of the young fibroblast response to TGF-β. Most studies that look at TGF-β cross-talk use targeted experimental techniques from molecular biology to isolate an aspect of TGF-β biology (Wilkes et al. 2005, Freudlsperger et al. 2013, Ji et al. 2014, Ponticos, Harvey, Ikeda, Abraham & Bou-Gharios 2009). Instead of picking out a single aspect of TGF-β cross-talk, an Affymetrix microarray was used to parallelise mRNA quantification at the transcriptome level to (in principle) identify all genes that respond to TGF-β within 3h of stimulation. The analysis identified many genes that were expected to be regulated by TGF-β including CTGF, FN1, ID1, JUNB, PDGFA, RHOB, SERPINE1, SKIL and TGFBI (Chapter 4). The list of differentially expressed genes was fed into a pathway analysis using both KEGG and Reactome databases with the intention of identifying pathways that are overrepresented in the gene list. Indeed, genes that are known to be part of TGF-β, Erk MAPK, PI3K, HIPPO, HIF1 and TNF-α pathways were found in this list. However, as discussed in Chapter 4 there are likely other genes regulated by TGF-β that participate in the regulation of other pathways.

The primary goal of this experiment was to identify some of the boundaries (i.e. cross-talk) of Smad signalling in neonatal fibroblasts. The boundaries of a signalling pathway are not true boundaries but a conceptual mechanism for comprehending the complexity of biochemical signalling. In reality biochemical pathways, exemplified by TGF-β, Erk1/2 and PI3K, are an inseparable network of interactions. Pathways may interact directly or indirectly via any number of nodes. It is conceivable that the overall action any single signalling molecule (such as a growth factor) resonate throughout the cell and coordinate the actions of numerous cellular signalling pathways - whether that

action is negative regulation to make way for the activity of another set of proteins or positive regulation to facilitate the production of a set of proteins for a required phenotype.

There are two types of interface between signalling pathways: direct protein-protein interactions such as binding, post-translational modifications or degradation, and transcriptional interfaces where new proteins are introduced into a system. While the work presented in Chapter 4 was a global study of the transcriptional response to TGF-β, the methods employed were unable to study the protein-protein interface. Given that ultimately, biochemical signalling occurs at the protein level of biological organisation, an upper limit exists in the amount of information can be ascertained about communication between signal transduction cascades using microarray technology. Future studies that attempt to identify how signal cascades collaborate in their response to TGF-β (or indeed any other signalling molecule) may benefit from a yeast-2-hybrid experiment for studying protein-protein interactions (Albers et al. 2005) or a proteomic based approach to directly observe and quantify the biochemical components that are involved in the interactions.

In Chapter 5, instead of using full transcriptomic based approaches, a set of genes were hand selected for investigation using high-throughput qPCR based technology. In doing so, some of the depth of the microarray was traded breath so that the activities of the chosen genes could be measured in a larger number of experimental conditions. The aim was to choose genes known or likely to be interesting with regards to ECM integrity and TGF-β signalling and see how their transcriptional dynamics change with age or senescence.

In this regard the experiment was highly successful and multiple lines of evidence support the literature regarding changes which exist between the young and old dermal fibroblast. Namely, it has been known for a long time that collagen levels (particularly type 1) are reduced in aged dermal tissue (Varani et al. 2006). In Chapter 5 a comparison was made between neonatal and adult fibroblasts' ability to produce collagen, both in the presence and absence of TGF-β, a known stimulator of type 1

collagen production (Varga et al. 1987). As in Varani et al. 2006, the results in Chapter 5 suggest that as we age, dermal fibroblasts produce less collagen, both basally and in response to TGF-β. Also evident from the literature is that aged cells reside in an ever increasing inflammatory environment (Licastro et al. 2005) and that MMPs are produced in response to inflammatory cytokines such as TNF-α and IL-6 (Du et al. 2016). MMP1 is the primary MMP responsible for the proteolytic processing of type I collagen into collagen fragments which can then be processes by other proteases (Fligiel et al. 2003). In Chapter 5, aged cells also showed an increase in the levels of MMP1, suggesting that increases in collagen degradation as well as decreases in its production are contributing factors towards skin thinning with age.

The experiment in Chapter 5 employed time matched controls. In the study of biology, time series experiments are often conducted without time matched controls, likely because of the extra experimental effort and cost involved. However, cells are not inert and the longer they spend in culture the less adequate the 0 time point becomes as a negative control. Cellular processes are constantly occurring, even in the absence of treatment which may confound our interpretation of an experiment if not properly controlled. In Chapter 5 the MMP1 data (Figure 5.3a) provides a strong argument for time matched controls because without them the data would suggest that MMP1 levels are inhibited by a TGF-β induced mechanism. The time matched controls enabled us to see that this negative regulation of MMP1 occurs regardless of the presence of TGF-β.

The MMP1 data from Chapter 5 call into question the claims that TGF-β negatively regulates MMP1 transcription and raises the question that if our data are to be trusted, why have other studies stated otherwise? We discussed the possibility that TGF-β inhibits MMP1 production from fibroblasts stimulated with TGF-β or IL6 but not under basal conditions. The data in Figure 5.3a also raises the question of why should MMP1 levels be down regulated simply from time in culture? This question is difficult to answer, but it is possible that the cells in culture are under abnormal growth conditions compared to *in vivo* (i.e. no 3D environment or ECM) conditions and they respond by producing and secreting ECM, and by down regulation of ECM degrading proteins.

Another finding in Chapter 5 was that the levels of Smad3 mRNA drop at 12h post stimulation by TGF-β (Figure 5.11a). It is unclear what the purpose of this dynamic is, but it is interesting to note that it occurs at approximately the same time as α-SMA, a marker for differentiated fibroblasts, is produced (Figure 5.9a). Also occurring around this time is the biggest increase in COL1A1 and other ECM components such as CTGF in response to TGF-β. Collectively this information suggests that a persistent large dose of TGF-β induces fibroblasts to differentiate into myofibroblasts, which are more productive in terms of ECM output. However, given Smad3 levels are down regulated, it is plausible that ECM regulation in myofibroblasts is less dependent on Smad3. This raises the question of what, if not Smad3, is the primary regulator of collagen production?

In Chapter 6 a small subset of the data from Chapter 5 was used together with an extensive literature review to formulate an ODE model to explain the observed reduction in collagen levels in adult compared to neonatal fibroblasts. Four mechanisms were proposed, two regarding the influence of CTGF on type 1 collagen production and two regarding how these interactions change with age. A combinatorial model selection strategy was used to build a set of models that exhaustively tests the 'hypothesis-space'. The results suggest that the 'best model' of those tested was the model that contained three mechanisms: 1) Smad3 reduction in age; 2) TGF-β type 2 receptors reduced in age and 3) CTGF-mediated enhanced Smad3 stability. Despite coming to the conclusion that this model was the best at reproducing the data of the models tested, further testing of alternative topologies would likely be informative to capture the behaviour of the basal and the CTGF data. These issues require addressing and will be followed up in future work.

Part of the issue with this modelling effort was that many parts of the network are unobserved. The two ways to handle this issue is to 1) reduce the model complexity or 2) to increase the number of observables. As part of the modelling process, the model complexity was reduced as much as possible by minimising the size of the network and reusing parameters from similar models. The implications of these simplifications are not known but it is highly likely that collecting more data to inform the model would be

beneficial for parameter identification and generating hypotheses concerning topology.

The Python programming language has been instrumental for this work. Two Python packages have been designed, developed, documented and deployed on the Python package index (PyPI) for others to use. The first package, `PyCoTools` (Welsh et al. 2018 and Section 1.4) is an alternative interface to COPASI which is a widely used modelling toolbox for systems biologists (Hoops et al. 2006). PyCoTools was used extensively in the modelling strategy presented in Chapter 6 to automatically handle the implementation and configuration of the 16 models. Without PyCoTools this strategy would be much more difficult (to the point of being non-viable), because of how time consuming and error prone it is to build and configure all these models manually.

The other Python package, `pytseries`, was discussed in Chapter 3. Its primary focus was to provide a set of classes to simplify the handling of time series data, by automating repetitive tasks such as numerical computation between different time series. Because of the object-oriented (i.e. modular) nature of this package, it is readily extensible and this property was exploited to build a clustering module using the base classes. Specifically, the dynamic time warping and the K-means algorithm were combined to cluster time series based on profile similarities. While the algorithm performed well in initial tests with perfectly separable data, the algorithm as employed in Chapter 4 on time series microarray data generally performed poorly. In principal the `pytseries` module can be further extended to implement alternative clustering strategies that may perform better in application to microarray data, as well as any number of other bioinformatic applications, such as enrichments analysis or network inference.

Overall, the work presented in this thesis offers some progress in addressing two challenges. Firstly by measuring the fibroblast transcriptome in response to TGF-β over time we gain insight into which signalling cascades are transcriptionally modulated by TGF-β. We have also provided an insight into the dynamic profiles of several transcripts that are important in skin ageing, both in response to TGF-β and how the response changes with age. We have made progress in the development of a systems biology workflow that unifies high throughput experimentation with mechanistic

modelling and developed a means of combinatorially building SBML models. Using this method we have developed a computational investigation into the roles of CTGF on TGF-β in age. Lastly we have discussed the development of two Python packages, both of which have utility in systems biology investigations: PyCoTools and pytseries. .

Bibliography

Abar, S., Theodoropoulos, G. K., Lemarinier, P. & OHare, G. M. P. (2017), 'Agent based modelling and simulation tools: A review of the state-of-art software', *Computer Science Review* **24**, 13–33.
URL: http://www.sciencedirect.com/science/article/pii/S1574013716301198

Abreu, J. G., Ketpura, N. I., Reversade, B. & De Robertis, E. M. (2002), 'Connective-tissue growth factor (ctgf) modulates cell signalling by bmp and tgf-', *Nature cell biology* **4**(8), 599–604.
URL: https://www.nature.com/articles/ncb826.pdf

Akaike, H. (1973), 'Maximum likelihood identification of gaussian autoregressive moving average models', *Biometrika* **60**(2), 255–265.

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE transactions on automatic control* **19**(6), 716–723.
URL: https://ieeexplore.ieee.org/ielx5/9/24140/01100705.pdf?tp=&arnumber=1100705&isnumber=24140

Akaike, H. (1985), *Prediction and entropy*, Springer, pp. 387–410.

Akaike, H. (1994), *Implications of informational point of view on the development of statistical science*, Springer, pp. 421–432.

Akiyoshi, S., Inoue, H., Hanai, J.-i., Kusanagi, K., Nemoto, N., Miyazono, K. & Kawabata, M. (1999), 'c-ski acts as a transcriptional co-repressor in transforming growth factor- signaling through interaction with smads', *Journal of Biological Chemistry* **274**(49), 35269–35277.
URL: http://www.jbc.org/content/274/49/35269.full.pdf

Alber, M. S., Kiskowski, M. A., Glazier, J. A. & Jiang, Y. (2003), *On cellular automaton approaches to modeling biological cells*, Springer, pp. 1–39.

Albers, M., Kranz, H., Kober, I., Kaiser, C., Klink, M., Suckow, J., Kern, R. & Koegl, M. (2005), 'Automated yeast two-hybrid screening for nuclear receptor-interacting proteins', *Molecular &*

Cellular Proteomics **4**(2), 205–213.
URL: http://www.mcponline.org/content/4/2/205.full.pdf

Alexa, A. & Rahnenfuhrer, J. (2010), 'topgo: enrichment analysis for gene ontology', *R package version* **2**(0).

Alwine, J. C., Kemp, D. J. & Stark, G. R. (1977), 'Method for detection of specific rnas in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with dna probes', *Proceedings of the National Academy of Sciences of the United States of America* **74**(12), 5350–5354.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC431715/

Arnott, J. A., Zhang, X., Sanjay, A., Owen, T. A., Smock, S. L., Rehman, S., DeLong, W. G., Safadi, F. F. & Popoff, S. N. (2008), 'Molecular requirements for induction of ctgf expression by tgf-1 in primary osteoblasts', *Bone* **42**(5), 871–885.
URL: https://www.ncbi.nlm.nih.gov/pubmed/18314002

Arriola, L. & Hyman, J. M. (2009), *Sensitivity analysis for uncertainty quantification in mathematical models*, Springer, pp. 195–247.
URL: https://www.researchgate.net/publication/226918007_Sensitivity_Analysis_for_Uncertainty_Quantification_in_Mathematical_Models

Asano, Y., Czuwara, J. & Trojanowska, M. (2007), 'Transforming growth factor-beta regulates dna binding activity of transcription factor fli1 by p300/creb-binding protein-associated factor-dependent acetylation', *J Biol Chem* **282**(48), 34672–83.
URL: https://www.ncbi.nlm.nih.gov/pubmed/17884818http://www.jbc.org/content/282/48/34672.full.pdfhttp://www.jbc.org/content/282/48/34672.full.pdf

Asano, Y., Ihn, H., Yamane, K., Jinnin, M., Mimura, Y. & Tamaki, K. (2005), 'Increased expression of integrin alpha(v)beta3 contributes to the establishment of autocrine tgf-beta signaling in scleroderma fibroblasts', *J Immunol* **175**(11), 7708–18.
URL: http://www.jimmunol.org/content/jimmunol/175/11/7708.full.pdf

Asano, Y., Markiewicz, M., Kubo, M., Szalai, G., Watson, D. K. & Trojanowska, M. (2009), 'Transcription factor fli1 regulates collagen fibrillogenesis in mouse skin', *Mol Cell Biol* **29**(2), 425–34.
URL: http://mcb.asm.org/content/29/2/425.full

Asano, Y. & Trojanowska, M. (2013), 'Fli1 represses transcription of the human alpha2(i) collagen gene by recruitment of the hdac1/p300 complex', *PLoS One* **8**(9), e74930.
URL: https://www.ncbi.nlm.nih.gov/pubmed/24058639

Asgari, M., Latifi, N., Heris, H. K., Vali, H. & Mongeau, L. (2017), 'In vitro fibrillogenesis of tropocollagen type iii in collagen type i affects its relative fibrillar topology and mechanics', *Scientific*

reports **7**(1), 1392.
URL:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5431193/pdf/41598_2017_Article_1476.pdf

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S. & Eppig, J. T. (2000), 'Gene ontology: tool for the unification of biology', *Nature genetics* **25**(1), 25.
URL: http://www.nature.com/articles/ng0500_25.pdf

Atfi, A., Djelloul, S., Chastre, E., Davis, R. & Gespach, C. (1997), 'Evidence for a role of rho-like gtpases and stress-activated protein kinase/c-jun n-terminal kinase (sapk/jnk) in transforming growth factor -mediated signaling', *Journal of Biological Chemistry* **272**(3), 1429–1432.
URL: http://www.jbc.org/content/272/3/1429.abstracthttp://www.jbc.org/content/272/3/1429.full.pdfhttp://www.jbc.org/content/272/3/1429.full.pdf

Aunan, J. R., Watson, M. M., Hagland, H. R. & Sreide, K. (2016), 'Molecular and biological hallmarks of ageing', *British Journal of Surgery* **103**(2).
URL: https://www.ncbi.nlm.nih.gov/pubmed/26771470

Baarsma, H. A., Engelbertink, L. H. J. M., Hees, L. J., Menzen, M. H., Meurs, H., Timens, W., Postma, D. S., Kerstjens, H. A. M. & Gosens, R. (2013), 'Glycogen synthase kinase-3 (gsk-3) regulates tgf-1-induced differentiation of pulmonary fibroblasts', *British journal of pharmacology* **169**(3), 590–603.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3682707/pdf/bph0169-0590.pdf

Baylis, D., Bartlett, D. B., Patel, H. P. & Roberts, H. C. (2013), 'Understanding how we age: insights into inflammaging', *Longev Healthspan* **2**(1), 8.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3922951/pdf/2046-2395-2-8.pdf

Bellman, R. & Kalaba, R. (1958), 'On adaptive control processes', *IRE Transactions on Automatic Control* **4**(2), 1–9.
URL: https://ieeexplore.ieee.org/ielx5/8163/24269/01104847.pdf?tp=&arnumber=1104847&isnumber=24269

Benbow, U. & Brinckerhoff, C. E. (1997), 'The ap-1 site and mmp gene regulation: what is all the fuss about?', *Matrix biology* **15**(8-9), 519–526.
URL: https://www.sciencedirect.com/science/article/pii/S0945053X97900263

Berlett, B. S. & Stadtman, E. R. (1997), 'Protein oxidation in aging, disease, and oxidative stress', *J Biol Chem* **272**(33), 20313–6.
URL: http://www.jbc.org/content/272/33/20313.full.pdf

Bhogal, R. K. & Bona, C. A. (2008), 'Regulatory effect of extracellular signal-regulated kinases (erk) on type i collagen synthesis in human dermal fibroblasts stimulated by il-4 and il-13', *Int Rev Immunol* **27**(6), 472–96.
URL: https://www.tandfonline.com/doi/pdf/10.1080/08830180802430974?needAccess=true

Bholowalia, P. & Kumar, A. (2014), 'Ebk-means: A clustering technique based on elbow method and k-means in wsn', *International Journal of Computer Applications* **105**(9).

Blinov, M. L., Faeder, J. R., Goldstein, B. & Hlavacek, W. S. (2004), 'Bionetgen: software for rule-based modeling of signal transduction based on the interactions of molecular domains', *Bioinformatics* **20**(17), 3289–91.
URL: https://www.ncbi.nlm.nih.gov/pubmed/15217809

Bolstad, B. M., Collin, F., Brettschneider, J., Simpson, K., Cope, L., Irizarry, R. A. & Speed, T. P. (2005), *Quality assessment of Affymetrix GeneChip data*, Springer, pp. 33–47.

Bonni, S., Wang, H. R., Causing, C. G., Kavsak, P., Stroschein, S. L., Luo, K. & Wrana, J. L. (2001), 'Tgf-beta induces assembly of a smad2-smurf2 ubiquitin ligase complex that targets snon for degradation', *Nat Cell Biol* **3**(6), 587–95.
URL: https://www.nature.com/articles/ncb0601_587.pdf

Breitenbach, J. S., Rinnerthaler, M., Trost, A., Weber, M., Klausegger, A., Gruber, C., Bruckner, D., Reitsamer, H. A., Bauer, J. W. & Breitenbach, M. (2015), 'Transcriptome and ultrastructural changes in dystrophic epidermolysis bullosa resemble skin aging', *Aging (Albany NY)* **7**(6), 389.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4505166/

Brown, K. S. & Sethna, J. P. (2003), 'Statistical mechanical approaches to models with many poorly known parameters', *Physical review E* **68**(2), 021904.
URL: https://www.ncbi.nlm.nih.gov/pubmed/14525003

Brsch, A. & Schaber, J. (2016), 'How time delay and network design shape response patterns in biochemical negative feedback systems', *BMC systems biology* **10**(1), 82–82.
URL: https://www.ncbi.nlm.nih.gov/pubmed/27558510https://www.ncbi.nlm.nih.gov/pmc/PMC4995745/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4995745/pdf/12918_2016_Article_325.pdf

Burgeson, R. E. & Christiano, A. M. (1997), 'The dermal-epidermal junction', *Current Opinion in Cell Biology* **9**(5), 651–658.
URL: https://www.ncbi.nlm.nih.gov/pubmed/9330868

Burnham, K. P. & Anderson, D. R. (2003), *Model selection and multimodel inference: a practical information-theoretic approach*, Springer Science & Business Media.
URL: https://www.springer.com/gb/book/9780387953649

Calderwood, S. K., Murshid, A. & Prince, T. (2009), 'The shock of aging: Molecular chaperones and the heat shock response in longevity and aging a mini-review', *Gerontology* **55**(5), 550–558.
URL: https://www.ncbi.nlm.nih.gov/pubmed/19546513

Cao, Y., Gillespie, D. T. & Petzold, L. R. (2006), 'Efficient step size selection for the tau-leaping simulation method', *The Journal of chemical physics* **124**(4), 044109.
URL: https://aip.scitation.org/doi/pdf/10.1063/1.2159468

Caraci, F., Gili, E., Calafiore, M., Failla, M., La Rosa, C., Crimi, N., Sortino, M. A., Nicoletti, F., Copani, A. & Vancheri, C. (2008), 'Tgf-beta1 targets the gsk-3beta/beta-catenin pathway via erk activation in the transition of human lung fibroblasts into myofibroblasts', *Pharmacol Res* **57**(4), 274–82.
URL: https://ac.els-cdn.com/S1043661808000236/1-s2.0-S1043661808000236-main.pdf?_tid=cd18baef-ec37-4b81-a79a-304f87958a74&acdnat=1532171634_7f7edf06faabf20d1829845a0e9ae60e

Cellire, G., Fengos, G., Herv, M. & Iber, D. (2011), 'plasticity of tgf- signaling', *BMC Systems Biology* **5**, 184–184.
URL: https://arxiv.org/abs/1111.0201

Chakraborty, D., umov, B., Mallano, T., Chen, C.-W., Distler, A., Bergmann, C., Ludolph, I., Horch, R. E., Gelse, K. & Ramming, A. (2017), 'Activation of stat3 integrates common profibrotic pathways to promote fibroblast activation and tissue fibrosis', *Nature communications* **8**(1), 1130.
URL:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5654983/pdf/41467_2017_Article_1236.pdf

Chandeck, C. & Mooi, W. J. (2010), 'Oncogene-induced cellular senescence', *Adv Anat Pathol* **17**(1), 42–8.

Cheifetz, S., Bassols, A., Stanley, K., Ohta, M., Greenberger, J. & Massague, J. (1988), 'Heterodimeric transforming growth factor beta. biological properties and interaction with three types of cell surface receptors', *J Biol Chem* **263**(22), 10783–9.
URL: http://www.jbc.org/content/263/22/10783.full.pdf

Chen, S.-J., Yuan, W., Mori, Y., Levenson, A., Varga, J. & Trojanowska, M. (1999), 'Stimulation of type i collagen transcription in human skin fibroblasts by tgf-: Involvement of smad 3', *Journal of*

Investigative Dermatology **112**(1), 49–57.
URL: https://www.ncbi.nlm.nih.gov/pubmed/9886263

Chen, Z., Dodig-Crnkovi, T., Schwenk, J. M. & Tao, S.-c. (2018), 'Current applications of antibody microarrays', *Clinical Proteomics* **15**, 7.
URL: https://www.ncbi.nlm.nih.gov/pubmed/29507545

Cheng, J.-C., Chang, H.-M., Fang, L., Sun, Y.-P. & Leung, P. C. K. (2015), 'Tgf-1 up-regulates connective tissue growth factor expression in human granulosa cells through smad and erk1/2 signaling pathways', *PloS one* **10**(5), e0126532–e0126532.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4425519/pdf/pone.0126532.pdf

Childs, B. G., Durik, M., Baker, D. J. & van Deursen, J. M. (2015), 'Cellular senescence in aging and age-related disease: from mechanisms to therapy', *Nature medicine* **21**(12), 1424–1435.
URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4748967/

Chis, O.-T., Banga, J. R. & Balsa-Canto, E. (2011), 'Structural identifiability of systems biology models: a critical comparison of methods', *PloS one* **6**(11), e27755.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3222653/pdf/pone.0027755.pdf

Choi, E., Kim, W., Joo, S. K., Park, S., Park, J. H., Kang, Y. K., Jin, S.-Y. & Chang, M. S. (2018), 'Expression patterns of stat3, erk and estrogen-receptor are associated with development and histologic severity of hepatic steatosis: a retrospective study', *Diagnostic pathology* **13**(1), 23.
URL:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5883355/pdf/13000_2018_Article_698.pdf

Choi, K., Medley, J. K., Cannistra, C., Konig, M., Smith, L., Stocking, K. & Sauro, H. M. (2016), 'Tellurium: a python based modeling and reproducibility platform for systems biology', *bioRxiv* p. 054601.

Chung, K.-Y., Agarwal, A., Uitto, J. & Mauviel, A. (1996), 'An ap-1 binding sequence is essential for regulation of the human 2 (i) collagen (col1a2) promoter activity by transforming growth factor', *Journal of Biological Chemistry* **271**(6), 3272–3278.
URL: http://www.jbc.org/content/271/6/3272.full.pdf

Cole, M. A., Quan, T., Voorhees, J. J. & Fisher, G. J. (2018), 'Extracellular matrix regulation of fibroblast function: redefining our perspective on skin aging', *J Cell Commun Signal* **12**(1), 35–43.
URL:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5842211/pdf/12079_2018_Article_459.pdf

Copp, J.-P., Desprez, P.-Y., Krtolica, A. & Campisi, J. (2010), 'The senescence-associated secretory phenotype: The dark side of tumor suppression', *Annual review of pathology* **5**, 99–118.

URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4166495/https://www.annualreviews.org/doi/pdf/10.1146/annurev-pathol-121808-102144https://www.annualreviews.org/doi/pdf/10.1146/annurev-pathol-121808-102144

Czuwara-Ladykowska, J., Sementchenko, V. I., Watson, D. K. & Trojanowska, M. (2002), 'Ets1 is an effector of the transforming growth factor beta (tgf-beta) signaling pathway and an antagonist of the profibrotic effects of tgf-beta', *J Biol Chem* **277**(23), 20399–408.
URL: http://www.jbc.org/content/277/23/20399.full.pdf

Czuwara-Ladykowska, J., Shirasaki, F., Jackers, P., Watson, D. K. & Trojanowska, M. (2001), 'Fli-1 inhibits collagen type i production in dermal fibroblasts via an sp1-dependent pathway', *J Biol Chem* **276**(24), 20839–48.
URL: http://www.jbc.org/content/276/24/20839.full.pdf

Dallas, S. L., Miyazono, K., Skerry, T. M., Mundy, G. R. & Bonewald, L. F. (1995), 'Dual role for the latent transforming growth factor-beta binding protein in storage of latent tgf-beta in the extracellular matrix and as a structural matrix protein', *The Journal of cell biology* **131**(2), 539–549.
URL: http://jcb.rupress.org/content/jcb/131/2/539.full.pdf

Darby, I. A. & Hewitson, T. D. (2007), 'Fibroblast differentiation in wound healing and fibrosis', *Int Rev Cytol* **257**, 143–79.

Davalli, P., Mitic, T., Caporali, A., Lauriola, A. & D'Arca, D. (2016), 'Ros, cell senescence, and novel molecular mechanisms in aging and age-related diseases', *Oxid Med Cell Longev* **2016**, 3565127.

De Crescenzo, G., Litowski, J. R., Hodges, R. S. & O'Connor-McCourt, M. D. (2003), 'Real-time monitoring of the interactions of two-stranded de novo designed coiled-coils: effect of chain length on the kinetic and thermodynamic constants of binding', *Biochemistry* **42**(6), 1754–63.
URL: https://www.ncbi.nlm.nih.gov/pubmed/12578390https://pubs.acs.org/doi/pdfplus/10.1021/bi0268450

De Pauw, D. J. W. & Vanrolleghem, P. A. (2006), 'Avoiding the finite difference sensitivity analysis deathtrap by using the complex-step derivative approximation technique'.

Deheuninck, J. & Luo, K. (2009), 'Ski and snon, potent negative regulators of tgf-beta signaling', *Cell Res* **19**(1), 47–57.
URL: https://www.nature.com/articles/cr2008324.pdf

Deng, L., Wang, C., Spencer, E., Yang, L., Braun, A., You, J., Slaughter, C., Pickart, C. & Chen, Z. J. (2000), 'Activation of the ib kinase complex by traf6 requires a dimeric ubiquitin-conjugating enzyme complex and a unique polyubiquitin chain', *Cell* **103**(2), 351–361.
URL: https://www.ncbi.nlm.nih.gov/pubmed/11057907

Denissova, N. G., Pouponnot, C., Long, J., He, D. & Liu, F. (2000), 'Transforming growth factor beta -inducible independent binding of smad to the smad7 promoter', *Proc Natl Acad Sci U S A* **97**(12), 6397–402.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC18614/pdf/pq006397.pdf

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C. & Lempicki, R. A. (2003), 'David: database for annotation, visualization, and integrated discovery', *Genome biology* **4**(9), R60.

DeRisi, J., Penland, L., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A. & Trent, J. M. (1996), 'Use of a cdna microarray to analyse gene expression', *Nat. genet* **14**, 457–460.
URL: https://www.nature.com/articles/ng1296-457.pdf

Derynck, R., Jarrett, J. A., Chen, E. Y., Eaton, D. H., Bell, J. R., Assoian, R. K., Roberts, A. B., Sporn, M. B. & Goeddel, D. V. (1985), 'Human transforming growth factor- complementary dna sequence and expression in normal and transformed cells', *Nature* **316**(6030), 701.

Di Guglielmo, G. M., Le Roy, C., Goodfellow, A. F. & Wrana, J. L. (2003), 'Distinct endocytic pathways regulate tgf-beta receptor signalling and turnover', *Nat Cell Biol* **5**(5), 410–21.
URL: https://www.nature.com/articles/ncb975.pdf

Di Micco, R., Fumagalli, M., Cicalese, A., Piccinin, S., Gasparini, P., Luise, C., Schurra, C., Nuciforo, P. G., Bensimon, A. & Maestro, R. (2006), 'Oncogene-induced senescence is a dna damage response triggered by dna hyper-replication', *Nature* **444**(7119), 638.
URL: https://www.nature.com/articles/nature05327.pdf

Diaz, B., Barnard, D., Filson, A., MacDonald, S., King, A. & Marshall, M. (1997), 'Phosphorylation of raf-1 serine 338-serine 339 is an essential regulatory event for ras-dependent activation and biological signaling', *Mol Cell Biol* **17**(8), 4509–16.
URL: http://mcb.asm.org/content/17/8/4509.full.pdf

DiRenzo, D. M., Chaudhary, M. A., Shi, X., Franco, S. R., Zent, J., Wang, K., Guo, L.-W. & Kent, K. C. (2016), 'A crosstalk between tgf-/smad3 and wnt/-catenin pathways promotes vascular smooth muscle cell proliferation', *Cellular Signalling* **28**(5), 498–505.
URL: https://www.ncbi.nlm.nih.gov/pubmed/26912210

DiStefano III, J. (2015), *Dynamic systems biology modeling and simulation*, Academic Press.

Du, G., Liu, C., Li, X., Chen, W., He, R., Wang, X., Feng, P. & Lan, W. (2016), 'Induction of matrix metalloproteinase-1 by tumor necrosis factor- is mediated by interleukin-6 in cultured fibroblasts of keratoconus', *Experimental Biology and Medicine* **241**(18), 2033–2041.
URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5102130/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5102130/pdf/10.1177_1535370216650940.pdf

Ebisawa, T., Fukuchi, M., Murakami, G., Chiba, T., Tanaka, K., Imamura, T. & Miyazono, K. (2001), 'Smurf1 interacts with transforming growth factor- type i receptor through smad7 and induces receptor degradation', *Journal of Biological Chemistry* **276**(16), 12477–12480.
URL: http://www.jbc.org/content/276/16/12477.full.pdf

Edlund, S., Landstrm, M., Heldin, C.-H. & Aspenstrm, P. (2002), 'Transforming growth factor-induced mobilization of actin cytoskeleton requires signaling by small gtpases cdc42 and rhoa', *Molecular biology of the cell* **13**(3), 902–914.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC99608/pdf/mk0302000902.pdf

Edwards, D. R., Leco, K. J., Beaudry, P. P., Atadja, P. W., Veillette, C. & Riabowol, K. T. (1996), 'Differential effects of transforming growth factor-1 on the expression of matrix metalloproteinases and tissue inhibitors of metalloproteinases in young and old human fibroblasts', *Experimental Gerontology* **31**(1), 207–223.
URL: https://www.ncbi.nlm.nih.gov/pubmed/8706790

Erguler, K. & Stumpf, M. P. H. (2011), 'Practical limits for reverse engineering of dynamical systems: a statistical analysis of sensitivity and parameter inferability in systems biology models', *Molecular BioSystems* **7**(5), 1593–1602.
URL: http://pubs.rsc.org/en/content/articlepdf/2011/mb/c0mb00107dhttps://pubs.rsc.org/en/content/articlepdf/2011/mb/c0mb00107d

Evans, R. A., Tian, Y. a. C., Steadman, R. & Phillips, A. O. (2003), 'Tgf-1-mediated fibroblastmyofibroblast terminal differentiationthe role of smad proteins', *Experimental Cell Research* **282**(2), 90–100.
URL: https://www.ncbi.nlm.nih.gov/pubmed/12531695

Everitt, B. S. (2011), 'Cluster analysis', *Wiley* .

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., Milacic, M., Roca, C. D., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Viteri, G., Weiser, J., Wu, G., Stein, L., Hermjakob, H. & D'Eustachio, P. (2018), 'The reactome pathway knowledgebase', *Nucleic Acids Res* **46**(D1), D649–d655.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5753187/pdf/gkx1132.pdf

Falcon, S. & Gentleman, R. (2006), 'Using gostats to test gene lists for go term association', *Bioinformatics* **23**(2), 257–258.
URL: https://www.ncbi.nlm.nih.gov/pubmed/17098774

Fang, M., Goldstein, E. L., Turner, A. S., Les, C. M., Orr, B. G., Fisher, G. J., Welch, K. B., Rothman, E. D. & Banaszak Holl, M. M. (2012), 'Type i collagen d-spacing in fibril bundles of dermis, tendon,

and bone: bridging between nano-and micro-level tissue hierarchy', *ACS nano* **6**(11), 9503–9514.
URL: https://pubs.acs.org/doi/pdfplus/10.1021/nn302483x

Farage, M. A., Miller, K. W. & Maibach, H. I. (2009), *Textbook of aging skin*, Springer Science & Business Media.

Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. (1989), 'Electrospray ionization for mass spectrometry of large biomolecules', *Science* **246**(4926), 64–71.
URL: http://science.sciencemag.org/content/sci/246/4926/64.full.pdf

Fisher, G. J., Quan, T., Purohit, T., Shao, Y., Cho, M. K., He, T., Varani, J., Kang, S. & Voorhees, J. J. (2009), 'Collagen fragmentation promotes oxidative stress and elevates matrix metalloproteinase-1 in fibroblasts in aged human skin', *The American journal of pathology* **174**(1), 101–114.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2631323/pdf/JPATH174000101.pdf

Fisher, G. J., Shao, Y., He, T., Qin, Z., Perry, D., Voorhees, J. J. & Quan, T. (2016), 'Reduction of fibroblast size/mechanical force down-regulates tgf-beta type ii receptor: implications for human skin aging', *Aging Cell* **15**(1), 67–76.
URL: https://www.ncbi.nlm.nih.gov/pubmed/26780887

Fleisch, M. C., Maxwell, C. A. & Barcellos-Hoff, M. H. (2006), 'The pleiotropic roles of transforming growth factor beta in homeostasis and carcinogenesis of endocrine organs', *Endocr Relat Cancer* **13**(2), 379–400.
URL: http://erc.endocrinology-journals.org/content/13/2/379.full.pdf

Fligiel, S. E. G., Varani, J., Datta, S. C., Kang, S., Fisher, G. J. & Voorhees, J. J. (2003), 'Collagen degradation in aged/photodamaged skin in vivo and after exposure to matrix metalloproteinase-1 in vitro', *Journal of Investigative Dermatology* **120**(5), 842–848.
URL: https://www.ncbi.nlm.nih.gov/pubmed/12713591

Franceschi, C., Bonafe, M., Valensin, S., Olivieri, F., De Luca, M., Ottaviani, E. & De Benedictis, G. (2000), 'Inflamm-aging. an evolutionary perspective on immunosenescence', *Ann N Y Acad Sci* **908**, 244–54.
URL:
https://nyaspubs.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1749-6632.2000.tb06651.x

Freudlsperger, C., Bian, Y., Contag Wise, S., Burnett, J., Coupar, J., Yang, X., Chen, Z. & Van Waes, C. (2013), 'Tgf-beta and nf-kappab signal pathway cross-talk is mediated through tak1 and smad7 in a subset of head and neck cancers', *Oncogene* **32**(12), 1549–59.
URL: https://www.nature.com/articles/onc2012171.pdf

Gillespie, D. T. (1977), 'Exact stochastic simulation of coupled chemical reactions', *The journal of physical chemistry* **81**(25), 2340–2361.
URL: https://pubs.acs.org/doi/pdfplus/10.1021/j100540a008

Gillespie, D. T. (2000), 'The chemical langevin equation', *The Journal of Chemical Physics* **113**(1), 297–306.
URL: https://aip.scitation.org/doi/pdf/10.1063/1.481811

Gingery, A., Bradley, E. W., Pederson, L., Ruan, M., Horwood, N. J. & Oursler, M. J. (2008), 'Tgf-beta coordinately activates tak1/mek/akt/nfkb and smad pathways to promote osteoclast survival', *Exp Cell Res* **314**(15), 2725–38.
URL: https://www.ncbi.nlm.nih.gov/pubmed/18586026

Greenwel, P., Inagaki, Y., Hu, W., Walsh, M. & Ramirez, F. (1997), 'Sp1 is required for the early response of 2 (i) collagen to transforming growth factor-1', *Journal of Biological Chemistry* **272**(32), 19738–19745.
URL: http://www.jbc.org/content/272/32/19738.full.pdf

Gronroos, E., Hellman, U., Heldin, C. H. & Ericsson, J. (2002), 'Control of smad7 stability by competition between acetylation and ubiquitination', *Mol Cell* **10**(3), 483–93.
URL: https://ac.els-cdn.com/S1097276502006391/1-s2.0-S1097276502006391-main.pdf?_tid=875d61be-9f9d-4e92-93bb-c7c1258c31bf&acdnat=1532289578_31dc8a704faf2313a9bb09656593d4cb

Grotendorst, G. R. (1997), 'Connective tissue growth factor: a mediator of tgf-beta action on fibroblasts', *Cytokine Growth Factor Rev* **8**(3), 171–9.
URL: https://ac.els-cdn.com/S1359610197000105/1-s2.0-S1359610197000105-main.pdf?_tid=7face8f5-b535-4d6a-8337-4a194ab7753f&acdnat=1533381390_e09c6b2488c309f7cf817adf5bf8fdae

Gutenkunst, R. (2007), 'Sloppiness, modeling, and evolution in biochemical networks'.

Haines, P., Samuel, G. H., Cohen, H., Trojanowska, M. & Bujor, A. M. (2011), 'Caveolin-1 is a negative regulator of mmp-1 gene expression in human dermal fibroblasts via inhibition of erk1/2/ets1 signaling pathway', *J Dermatol Sci* **64**(3), 210–6.
URL: https://www.ncbi.nlm.nih.gov/pubmed/21925842

Hamidi, A., Song, J., Thakur, N., Itoh, S., Marcusson, A., Bergh, A., Heldin, C.-H. & Landstrm, M. (2017), 'Tgf- promotes pi3k-akt signaling and prostate cancer cell migration through the traf6-mediated ubiquitylation of p85', *Sci. Signal.* **10**(486), eaal4186.
URL: http://stke.sciencemag.org/content/sigtrans/10/486/eaal4186.full.pdf

Hanahan, D. & Weinberg, R. A. (2011), 'Hallmarks of cancer: the next generation', *Cell* **144**(5), 646–74.
URL: https://www.ncbi.nlm.nih.gov/pubmed/21376230

Hanyu, A., Ishidou, Y., Ebisawa, T., Shimanuki, T., Imamura, T. & Miyazono, K. (2001), 'The n domain of smad7 is essential for specific inhibition of transforming growth factor-beta signaling', *J Cell Biol* **155**(6), 1017–27.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2150897/pdf/0106023.pdf

Harman, D. (1972), 'The biologic clock: the mitochondria?', *Journal of the American Geriatrics Society* **20**(4), 145–147.
URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1532-5415.1972.tb00787.x

Hasdemir, D., Hoefsloot, H. C. J. & Smilde, A. K. (2015), 'Validation and selection of ode based systems biology models: how to arrive at more reliable decisions', *BMC Systems Biology* **9**(1), 32.
URL: https://doi.org/10.1186/s12918-015-0180-0https://bmcsystbiol.biomedcentral.com/track/pdf/10.1186/s12918-015-0180-0https://bmcsystbiol.biomedcentral.com/track/pdf/10.1186/s12918-015-0180-0

Hayashi, H., Abdollah, S., Qiu, Y., Cai, J., Xu, Y.-Y., Grinnell, B. W., Richardson, M. A., Topper, J. N., Gimbrone Jr, M. A. & Wrana, J. L. (1997), 'The mad-related protein smad7 associates with the tgf receptor and functions as an antagonist of tgf signaling', *Cell* **89**(7), 1165–1173.
URL: https://www.sciencedirect.com/science/article/pii/S0092867400803037?via%3Dihubhttps://ac.els-cdn.com/S0092867400803037/1-s2.0-S0092867400803037-main.pdf?_tid=cdeb312f-1f84-4469-8d22-ac3f95bb4004&acdnat=1532419028_e333803ce1687aaf7fa60a55764f59ff

Hayflick, L. & Moorhead, P. S. (1961), 'The serial cultivation of human diploid cell strains', *Experimental Cell Research* **25**(3), 585–621.
URL: http://www.sciencedirect.com/science/article/pii/0014482761901926https://ac.els-cdn.com/0014482761901926/1-s2.0-0014482761901926-main.pdf?_tid=0ae53d19-fb93-4bea-8e37-4957d1bc63b1&acdnat=1533471483_cb5452cf439e9e2b3bff6288712407ad

He, K., Yan, X., Li, N., Dang, S., Xu, L., Zhao, B., Li, Z., Lv, Z., Fang, X. & Zhang, Y. (2015), 'Internalization of the tgf- type i receptor into caveolin-1 and eea1 double-positive early endosomes', *Cell research* **25**(6), 738.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4456627/pdf/cr201560a.pdf

Hill, A. V. (1910), 'The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves', *j. physiol.* **40**, 4–7.

Holland, S. M. (2008), 'Principal components analysis (pca)', *Department of Geology, University of Georgia, Athens, GA* pp. 30602–2501.

Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P. & Kummer, U. (2006), 'Copasi - a complex pathway simulator', *Bioinformatics* **22**(24), 3067–3074.
URL: https://www.ncbi.nlm.nih.gov/pubmed/17032683

Hough, C., Radu, M. & Dore, J. J. (2012), 'Tgf-beta induced erk phosphorylation of smad linker region regulates smad signaling', *PLoS One* **7**(8), e42513.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3412844/pdf/pone.0042513.pdf

Hu, S., Xie, Z., Qian, J., Blackshaw, S. & Zhu, H. (2011), 'Functional protein microarray technology', *Wiley interdisciplinary reviews. Systems biology and medicine* **3**(3), 255–268.
URL:
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3044218/https://onlinelibrary.wiley.com/doi/pdf/10.1002/wsbm.118https://onlinelibrary.wiley.com/doi/pdf/10.1002/wsbm.118

Hu, Z. C., Shi, F., Liu, P., Zhang, J., Guo, D., Cao, X. L., Chen, C. F., Qu, S. Q., Zhu, J. Y. & Tang, B. (2017), 'Tieg1 represses smad7-mediated activation of tgf-beta1/smad signaling in keloid pathogenesis', *J Invest Dermatol* **137**(5), 1051–1059.
URL: https://www.ncbi.nlm.nih.gov/pubmed/28108300https://ac.els-cdn.com/S0022202X17300167/1-s2.0-S0022202X17300167-main.pdf?_tid=0d3ed695-e126-4c3e-a343-8eddc6a792b8&acdnat=1532419075_64c42ef7ff2967075afc2eabe3d8af75

Huang, D. W., Sherman, B. T. & Lempicki, R. A. (2008), 'Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists', *Nucleic acids research* **37**(1), 1–13.

Huang da, W., Sherman, B. T. & Lempicki, R. A. (2009*a*), 'Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists', *Nucleic Acids Res* **37**(1), 1–13.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2615629/pdf/gkn923.pdf

Huang da, W., Sherman, B. T. & Lempicki, R. A. (2009*b*), 'Systematic and integrative analysis of large gene lists using david bioinformatics resources', *Nat Protoc* **4**(1), 44–57.
URL: http://www.nature.com/articles/nprot.2008.211

Huang da, W., Sherman, B. T., Zheng, X., Yang, J., Imamichi, T., Stephens, R. & Lempicki, R. A. (2009), 'Extracting biological meaning from large gene lists with david', *Curr Protoc Bioinformatics* **Chapter 13**, Unit 13.11.
URL: https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/0471250953.bi1311s27

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D. & Cornish-Bowden, A. (2003), 'The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models', *Bioinformatics* **19**(4), 524–531.

Hurvich, C. M. & Tsai, C. (1993), 'A corrected akaike information criterion for vector autoregressive model selection', *Journal of time series analysis* **14**(3), 271–279.

Hurvich, C. M. & Tsai, C.-L. (1989), 'Regression and time series model selection in small samples', *Biometrika* **76**(2), 297–307.

Hurvich, C. M. & Tsai, C.-L. (1990), 'Model selection for least absolute deviations regression in small samples', *Statistics & Probability Letters* **9**(3), 259–265.
URL: http://www.sciencedirect.com/science/article/pii/016771529090065Fhttps://ac.els-cdn.com/016771529090065F/1-s2.0-016771529090065F-main.pdf?_tid=3a126ded-20ef-4457-bdb3-2829a99e06fa&acdnat=1533060671_22863083b3e7cd920f4a8008dc996d53

Hurvich, C. M. & Tsai, C.-L. (1991), 'Bias of the corrected aic criterion for underfitted regression and time series models', *Biometrika* **78**(3), 499–509.
URL: https://academic.oup.com/biomet/article/78/3/499/255872

Hurvich, C. M. & Tsai, C.-L. (1995*a*), 'Model selection for extended quasi-likelihood models in small samples', *Biometrics* pp. 1077–1084.

Hurvich, C. M. & TSAI, C.-L. (1995*b*), 'Relative rates of convergence for efficient model selection criteria in linear regression', *Biometrika* **82**(2), 418–425.

Ignotz, R. A. & Massague, J. (1986), 'Transforming growth factor-beta stimulates the expression of fibronectin and collagen and their incorporation into the extracellular matrix', *Journal of Biological Chemistry* **261**(9), 4337–4345.
URL: http://www.jbc.org/content/261/9/4337.full.pdf

Imokawa, G., Akasaki, S., Minematsu, Y. & Kawai, M. (1989), 'Importance of intercellular lipids in water-retention properties of the stratum corneum: induction and recovery study of surfactant dry skin', *Archives of dermatological research* **281**(1), 45–51.
URL: https://link.springer.com/content/pdf/10.1007%2FBF00424272.pdf

Inman, G. J., Nicolas, F. J. & Hill, C. S. (2002), 'Nucleocytoplasmic shuttling of smads 2, 3, and 4 permits sensing of tgf-beta receptor activity', *Mol Cell* **10**(2), 283–94.
URL: https://ac.els-cdn.com/S1097276502005853/1-s2.0-S1097276502005853-main.pdf?

tid=a3bba29b-5aaf-407c-8c4b-92ba9870b553&acdnat=1533475125
30f867ba1b4f5778110921a534dfa6f0

Irizarry, R. A., Hobbs, B., Collin, F., BeazerBarclay, Y. D., Antonellis, K. J., Scherf, U. & Speed, T. P. (2003), 'Exploration, normalization, and summaries of high density oligonucleotide array probe level data', *Biostatistics* **4**(2), 249–264.
URL: https://www.ncbi.nlm.nih.gov/pubmed/12925520

Javelaud, D., Pierrat, M.-J. & Mauviel, A. (2012), 'Crosstalk between tgf- and hedgehog signaling in cancer', *FEBS Letters* **586**(14), 2016–2025.
URL: http://www.sciencedirect.com/science/article/pii/S0014579312003754https://febs.onlinelibrary.wiley.com/doi/pdf/10.1016/j.febslet.2012.05.011https://ac.els-cdn.com/S0014579312003754/1-s2.0-S0014579312003754-main.pdf?_tid=5c7bfd10-72c0-4f28-9220-c15a57c9a97d&acdnat=1532289600_90b8eeb9bbd9d4eced797b1fb623ceec

Ji, H., Tang, H., Lin, H., Mao, J., Gao, L., Liu, J. & Wu, T. (2014), 'Rho/rock crosstalks with transforming growth factor/smad pathway participates in lung fibroblastmyofibroblast differentiation', *Biomedical reports* **2**(6), 787–792.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4179758/pdf/br-02-06-0787.pdf

Jinnin, M., Ihn, H., Yamane, K., Mimura, Y., Asano, Y. & Tamaki, K. (2005), '2 (i) collagen gene regulation by protein kinase c signaling in human dermal fibroblasts', *Nucleic acids research* **33**(4), 1337–1351.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC552962/pdf/gki275.pdf

Jonsson, G. & Benavente, J. (1992), 'Determination of some transport coefficients for the skin and porous layer of a composite membrane', *Journal of Membrane Science* **69**(1), 29–42.
URL: http://www.sciencedirect.com/science/article/pii/037673889280165Ghttps://ac.els-cdn.com/037673889280165G/1-s2.0-037673889280165G-main.pdf?_tid=3b5d3278-1568-11e8-8a0e-00000aab0f02&acdnat=1519039985_7de0dd59e450cdbbf2edc83500c50c98

Jung, M. & Pfeifer, G. P. (2015), 'Aging and dna methylation', *BMC Biology* **13**(1), 7.
URL: https://doi.org/10.1186/s12915-015-0118-4https://bmcbiol.biomedcentral.com/track/pdf/10.1186/s12915-015-0118-4

Kajino, T., Omori, E., Ishii, S., Matsumoto, K. & Ninomiya-Tsuji, J. (2007), 'Tak1 mapk kinase kinase mediates transforming growth factor- signaling by targeting snon oncoprotein for degradation', *Journal of Biological Chemistry* **282**(13), 9475–9481.
URL: http://www.jbc.org/content/282/13/9475.full.pdf

Kanehisa, M. & Goto, S. (2000), 'Kegg: kyoto encyclopedia of genes and genomes', *Nucleic Acids Res* **28**(1), 27–30.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC102409/pdf/gkd027.pdf

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. (2016), 'Kegg as a reference resource for gene and protein annotation', *Nucleic Acids Research* **44**(D1), D457–D462.
URL: http://dx.doi.org/10.1093/nar/gkv1070https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702792/pdf/gkv1070.pdfhttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702792/pdf/gkv1070.pdf

Kang, S., Kahan, S., McDermott, J., Flann, N. & Shmulevich, I. (2014), 'Biocellion: accelerating computer simulation of multicellular biological system models', *Bioinformatics* **30**(21), 3101–3108.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4609016/pdf/btu498.pdf

Kavsak, P., Rasmussen, R. K., Causing, C. G., Bonni, S., Zhu, H., Thomsen, G. H. & Wrana, J. L. (2000), 'Smad7 binds to smurf2 to form an e3 ubiquitin ligase that targets the tgf receptor for degradation', *Molecular cell* **6**(6), 1365–1375.
URL: https://www.sciencedirect.com/science/article/pii/S1097276500001349?via%3Dihubhttps://ac.els-cdn.com/S1097276500001349/1-s2.0-S1097276500001349-main.pdf?_tid=a35c94e4-6ed5-415b-8fc8-b763dbcf2ccd&acdnat=1532419243_222c1519358fa4d1c1f16f3010c5cbff

Kent, E., Neumann, S., Kummer, U. & Mendes, P. (2013), 'What can we learn from global sensitivity analysis of biochemical systems?', *PLOS ONE* **8**(11), e79244.
URL: https://doi.org/10.1371/journal.pone.0079244https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3828278/pdf/pone.0079244.pdfhttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3828278/pdf/pone.0079244.pdf

Khavkin, J. & Ellis, D. A. F. (2011), 'Aging skin: histology, physiology, and pathology', *Facial Plastic Surgery Clinics* **19**(2), 229–234.
URL: https://www.facialplastic.theclinics.com/article/S1064-7406(11)00004-6/fulltexthttps://www.sciencedirect.com/science/article/pii/S1064740611000046?via%3Dihub

Kiehl, T. R., Mattheyses, R. M. & Simmons, M. K. (2004), 'Hybrid simulation of cellular behavior', *Bioinformatics* **20**(3), 316–322.
URL: https://academic.oup.com/bioinformatics/article/20/3/316/186005

Kirkwood, T. B. L. (1977), 'Evolution of ageing', *Nature* **270**(5635), 301.

Kirkwood, T. B. L. (2005), 'Understanding the odd science of aging', *Cell* **120**(4), 437–447.
URL: http://www.sciencedirect.com/science/article/pii/S0092867405001017https://www.sciencedirect.com/science/article/pii/S0092867405001017?via%3Dihubhttps://ac.els-cdn.com/S0092867405001017/1-s2.0-S0092867405001017-main.pdf?_tid=dc0dd9ba-c094-484a-99e0-06a4c06ffbc1&acdnat=1532419412_e1c5e45a430d8c76435b35ed86321b77

Kirkwood, T. B. L. & Cremer, T. (1982), 'Cytogerontology since 1881: a reappraisal of august weismann and a review of modern progress', *Human genetics* **60**(2), 101–121.
URL: https://link.springer.com/content/pdf/10.1007%2FBF00569695.pdf

Kitano, H. (2002), 'Systems biology: a brief overview', *Science* **295**(5560), 1662–1664.
URL: http://science.sciencemag.org/content/sci/295/5560/1662.full.pdf

Klipp, E., Liebermeister, W., Wierling, C., Kowald, A., Lehrach, H. & Herwig, R. (2009), *Systems Biology: A Textbook*, Vol. 77, Wiley.

Klopfenstein, D., Zhang, L., Pedersen, B. S., Ramrez, F., Vesztrocy, A. W., Naldi, A., Mungall, C. J., Yunes, J. M., Botvinnik, O. & Weigel, M. (2018), 'Goatools: A python library for gene ontology analyses', *Scientific reports* **8**(1), 10872.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6052049/pdf/41598_2018_Article_28948.pdf

Kluppel, M. & Wrana, J. L. (2005), 'Turning it up a notch: cross-talk between tgf beta and notch signaling', *Bioessays* **27**(2), 115–8.
URL: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bies.20187

Koinuma, D., Shinozaki, M., Komuro, A., Goto, K., Saitoh, M., Hanyu, A., Ebina, M., Nukiwa, T., Miyazawa, K., Imamura, T. & Miyazono, K. (2003), 'Arkadia amplifies tgf-beta superfamily signalling through degradation of smad7', *Embo j* **22**(24), 6458–70.
URL: http://emboj.embopress.org/content/embojnl/22/24/6458.full.pdf

Kokura, K., Kim, H., Shinagawa, T., Khan, M. M., Nomura, T. & Ishii, S. (2003), 'The ski-binding protein c184m negatively regulates tumor growth factor- signaling by sequestering the smad proteins in the cytoplasm', *Journal of Biological Chemistry* **278**(22), 20133–20139.
URL: http://www.jbc.org/content/278/22/20133.abstracthttp://www.jbc.org/content/278/22/20133.full.pdf

Kowald, A. & Kirkwood, T. B. (2016), 'Can aging be programmed? a critical literature review', *Aging Cell* .

Kreutz, C., Raue, A. & Timmer, J. (2012), 'Likelihood based observability analysis and confidence intervals for predictions of dynamic models', *BMC Systems Biology* **6**(1), 120.
URL: https://bmcsystbiol.biomedcentral.com/track/pdf/10.1186/1752-0509-6-120

Kullback, S. & Leibler, R. A. (1951), 'On information and sufficiency', *The annals of mathematical statistics* **22**(1), 79–86.
URL: https://projecteuclid.org/euclid.aoms/1177729694

Kume, S., Haneda, M., Kanasaki, K., Sugimoto, T., Araki, S., Isshiki, K., Isono, M., Uzu, T., Guarente, L., Kashiwagi, A. & Koya, D. (2007), 'Sirt1 inhibits transforming growth factor beta-induced apoptosis in glomerular mesangial cells via smad7 deacetylation', *J Biol Chem* **282**(1), 151–8.
URL: http://www.jbc.org/content/282/1/151.full.pdf

Kupper, T. S. & Fuhlbrigge, R. C. (2004), 'Immune surveillance in the skin: mechanisms and clinical consequences', *Nat Rev Immunol* **4**(3), 211–22.
URL: https://www.nature.com/articles/nri1310.pdf

Kutz, S. M., Hordines, J., McKeown-Longo, P. J. & Higgins, P. J. (2001), 'Tgf-beta1-induced pai-1 gene expression requires mek activity and cell-to-substrate adhesion', *J Cell Sci* **114**(Pt 21), 3905–14.
URL: http://jcs.biologists.org/content/joces/114/21/3905.full.pdf

Kwak, H. J., Park, M. J., Cho, H., Park, C. M., Moon, S. I., Lee, H. C., Park, I. C., Kim, M. S., Rhee, C. H. & Hong, S. I. (2006), 'Transforming growth factor-beta1 induces tissue inhibitor of metalloproteinase-1 expression via activation of extracellular signal-regulated kinase and sp1 in human fibrosarcoma cells', *Mol Cancer Res* **4**(3), 209–20.
URL: http://mcr.aacrjournals.org/content/molcanres/4/3/209.full.pdf

Lake, D., Corra, S. A. L. & Mller, J. (2016), 'Negative feedback regulation of the erk1/2 mapk pathway', *Cellular and Molecular Life Sciences* **73**(23), 4397–4413.
URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5075022/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5075022/pdf/18_2016_Article_2297.pdfhttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5075022/pdf/18_2016_Article_2297.pdf

Larance, M. & Lamond, A. I. (2015), 'Multidimensional proteomics for cell biology', *Nature Reviews Molecular Cell Biology* **16**(5), 269.
URL: https://www.nature.com/articles/nrm3970.pdf

Leask, A., Holmes, A., Black, C. M. & Abraham, D. J. (2003), 'Connective tissue growth factor gene regulation requirements for its induction by transforming growth factor-2 in fibroblasts', *Journal of Biological Chemistry* **278**(15), 13008–13015.
URL: http://www.jbc.org/content/278/15/13008.full.pdf

Lee, M. K., Pardoux, C., Hall, M. C., Lee, P. S., Warburton, D., Qing, J., Smith, S. M. & Derynck, R. (2007), 'Tgf- activates erk map kinase signalling through direct phosphorylation of shca', *The EMBO journal* **26**(17), 3957–3967.
URL: http://emboj.embopress.org/content/embojnl/26/17/3957.full.pdf

Leivonen, S. K., Lazaridis, K., Decock, J., Chantry, A., Edwards, D. R. & Kahari, V. M. (2013), 'Tgf-beta-elicited induction of tissue inhibitor of metalloproteinases (timp)-3 expression in fibroblasts involves complex interplay between smad3, p38alpha, and erk1/2', *PLoS One* **8**(2), e57474.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3585359/pdf/pone.0057474.pdf

Li, J. H., Huang, X. R., Zhu, H. J., Johnson, R. & Lan, H. Y. (2003), 'Role of tgf-beta signaling in extracellular matrix production under high glucose conditions', *Kidney Int* **63**(6), 2010–9.
URL: https://ac.els-cdn.com/S0085253815491185/1-s2.0-S0085253815491185-main.pdf?_tid=74fefd75-7662-4d19-a9a1-5b8b496efb48&acdnat=1533299132_4c220c12050161a38a86d734e50af23f

Li, N., Yang, Y., He, K., Zhang, F., Zhao, L., Zhou, W., Yuan, J., Liang, W. & Fang, X. (2016), 'Single-molecule imaging reveals the activation dynamics of intracellular protein smad3 on cell membrane', *Scientific reports* **6**, 33469.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5027577/pdf/srep33469.pdf

Licastro, F., Candore, G., Lio, D., Porcellini, E., Colonna-Romano, G., Franceschi, C. & Caruso, C. (2005), 'Innate immunity and inflammation in ageing: a key for understanding age-related diseases', *Immunity & ageing : I & A* **2**, 8–8.
URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1166571/

Lin, P.-S., Chang, H.-H., Yeh, C.-Y., Chang, M.-C., Chan, C.-P., Kuo, H.-Y., Liu, H.-C., Liao, W.-C., Jeng, P.-Y., Yeung, S.-Y. & Jeng, J.-H. (2017), 'Transforming growth factor beta 1 increases collagen content, and stimulates procollagen i and tissue inhibitor of metalloproteinase-1 production of dental pulp cells: Role of mek/erk and activin receptor-like kinase-5/smad signaling', *Journal of the Formosan Medical Association* **116**(5), 351–358.
URL: https://www.ncbi.nlm.nih.gov/pubmed/27720345

Lindner, A. B. & Demarez, A. (2009), 'Protein aggregation as a paradigm of aging', *Biochimica et Biophysica Acta (BBA) - General Subjects* **1790**(10), 980–996.
URL: https://www.sciencedirect.com/science/article/pii/S0304416509001718

Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M. & Yates, J. R. (1999), 'Direct analysis of protein complexes using mass spectrometry', *Nature biotechnology* **17**(7), 676.
URL: https://www.nature.com/articles/nbt0799_676.pdf

Liu, G., Swihart, M. T. & Neelamegham, S. (2004), 'Sensitivity, principal component and flux analysis applied to signal transduction: the case of epidermal growth factor mediated signaling', *Bioinformatics* **21**(7), 1194–1202.
URL: https://www.ncbi.nlm.nih.gov/pubmed/15531606

Liu, W., Rui, H., Wang, J., Lin, S., He, Y., Chen, M., Li, Q., Ye, Z., Zhang, S. & Chan, S. C. (2006), 'Axin is a scaffold protein in tgf signaling that promotes degradation of smad7 by arkadia', *The EMBO journal* **25**(8), 1646–1658.
URL: http://emboj.embopress.org/content/embojnl/25/8/1646.full.pdf

Liu, X., Li, P., Liu, P., Xiong, R., Zhang, E., Chen, X., Gu, D., Zhao, Y., Wang, Z. & Zhou, Y. (2008), 'The essential role for c-ski in mediating tgf-beta1-induced bi-directional effects on skin fibroblast proliferation through a feedback loop', *Biochem J* **409**(1), 289–97.
URL: http://www.biochemj.org/content/ppbiochemj/409/1/289.full.pdf

Liu, Y., Beyer, A. & Aebersold, R. (2016), 'On the dependency of cellular protein levels on mrna abundance', *Cell* **165**(3), 535–50.
URL: https://ac.els-cdn.com/S0092867416302707/1-s2.0-S0092867416302707-main.pdf?_tid=520f5722-3689-4ae0-b1f7-d804de2e06e8&acdnat=1533143569_696ac56e9b7abaeb865a07b5d8475248

Livak, K. J. & Schmittgen, T. D. (2001), 'Analysis of relative gene expression data using real-time quantitative pcr and the 2(-delta delta c(t)) method', *Methods* **25**(4), 402–8.
URL: https://www.ncbi.nlm.nih.gov/pubmed/11846609

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M. & Norton, H. (1996), 'Expression monitoring by hybridization to high-density oligonucleotide arrays', *Nature biotechnology* **14**(13), 1675–1680.
URL: https://www.nature.com/articles/nbt1296-1675.pdf

Lockhart, D. J. & Winzeler, E. A. (2000), 'Genomics, gene expression and dna arrays', *Nature* **405**(6788), 827–36.
URL: http://www.nature.com/articles/35015701.pdf

Lopez-Otn, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. (2013), 'The hallmarks of aging', *Cell* **153**(6), 1194–1217.
URL: https://www.cell.com/abstract/S0092-8674(13)00645-4

Lu, P., Takai, K., Weaver, V. M. & Werb, Z. (2011), 'Extracellular matrix degradation and remodeling in development and disease', *Cold Spring Harb Perspect Biol* **3**(12).
URL: http://cshperspectives.cshlp.org/content/3/12/a005058.full.pdf

Luo, K. (2004), 'Ski and snon: negative regulators of tgf- signaling', *Current opinion in genetics & development* **14**(1), 65–70.
URL: https://ac.els-cdn.com/S0959437X03001692/1-s2.0-S0959437X03001692-main.pdf?_tid=f5ac39e0-e9b9-4046-bee1-9bf241d0c9ab&acdnat=1533302913_98521bec6205f8f925fb392e520c402c

Lyons, R. M., Gentry, L. E., Purchio, A. F. & Moses, H. L. (1990), 'Mechanism of activation of latent recombinant transforming growth factor beta 1 by plasmin', *The Journal of cell biology* **110**(4), 1361–1367.
URL: http://jcb.rupress.org/content/jcb/110/4/1361.full.pdf

Makino, K., Makino, T., Stawski, L., Lipson, K. E., Leask, A. & Trojanowska, M. (2017), 'Anti-connective tissue growth factor (ctgf/ccn2) monoclonal antibody attenuates skin fibrosis in mice models of systemic sclerosis', *Arthritis Res Ther* **19**(1), 134.
URL: https://www.ncbi.nlm.nih.gov/pubmed/28610597https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5470189/pdf/13075_2017_Article_1356.pdfhttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5470189/pdf/13075_2017_Article_1356.pdf

Mardis, E. R. (2008), 'Next-generation dna sequencing methods', *Annu Rev Genomics Hum Genet* **9**, 387–402.
URL: https://www.annualreviews.org/doi/pdf/10.1146/annurev.genom.9.081307.164359

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D. & Califano, A. (2006), 'Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context', *BMC Bioinformatics* **7**(Suppl 1), S7–S7.
URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1810318/https://bmcbioinformatics.biomedcentral.com/track/pdf/10.1186/1471-2105-7-S1-S7

Marthandan, S., Baumgart, M., Priebe, S., Groth, M., Schaer, J., Kaether, C., Guthke, R., Cellerino, A., Platzer, M., Diekmann, S. & Hemmerich, P. (2016), 'Conserved senescence associated genes and pathways in primary human fibroblasts detected by rna-seq', *PLOS ONE* **11**(5), e0154531.
URL: https://doi.org/10.1371/journal.pone.0154531https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4854426/pdf/pone.0154531.pdf

McCollum, J. M., Peterson, G. D., Cox, C. D., Simpson, M. L. & Samatova, N. F. (2006), 'The sorting direct method for stochastic simulation of biochemical systems with varying reaction execution behavior', *Comput Biol Chem* **30**(1), 39–49.
URL: https://www.sciencedirect.com/science/article/pii/S1476927105001088

McGowan, S. E. & McNamer, R. (1990), 'Transforming growth factor-beta increases elastin production by neonatal rat lung fibroblasts', *Am J Respir Cell Mol Biol* **3**(4), 369–76.

Medawar, P. B. (1952), 'An unsolved problem in biology'.

Meigel, W. N., Gay, S. & Weber, L. (1977), 'Dermal architecture and collagen type distribution', *Arch Dermatol Res* **259**(1), 1–10.

Mendoza, M. C., Er, E. E. & Blenis, J. (2011), 'The ras-erk and pi3k-mtor pathways: cross-talk and compensation', *Trends Biochem Sci* **36**(6), 320–8.
URL: https://www.ncbi.nlm.nih.gov/pubmed/21531565

Menon, G. K. (2002), 'New insights into skin structure: scratching the surface', *Adv Drug Deliv Rev* **54 Suppl 1**, S3–17.
URL: https://www.ncbi.nlm.nih.gov/pubmed/12460712

Metcalf, D. J., Garca-Arencibia, M., Hochfeld, W. E. & Rubinsztein, D. C. (2012), 'Autophagy and misfolded proteins in neurodegeneration', *Experimental Neurology* **238**(1), 22–28.
URL: https://www.ncbi.nlm.nih.gov/pubmed/21095248

Metzker, M. L. (2010), 'Sequencing technologies - the next generation', *Nat Rev Genet* **11**(1), 31–46.
URL: https://www.nature.com/articles/nrg2626.pdf

Minden, A., Lin, A., McMahon, M., Lange-Carter, C., Drijard, B., Davis, R. J., Johnson, G. L. & Karin, M. (1994), 'Differential activation of erk and jnk mitogen-activated protein kinases by raf-1 and mekk', *Science* **266**(5191), 1719–1723.
URL: http://science.sciencemag.org/content/sci/266/5191/1719.full.pdf

Miyazono, K., Hellman, U., Wernstedt, C. & Heldin, C. H. (1988), 'Latent high molecular weight complex of transforming growth factor beta 1. purification from human platelets and structural characterization', *Journal of Biological Chemistry* **263**(13), 6407–6415.
URL: http://www.jbc.org/content/263/13/6407.full.pdf

Morris, E., Chrobak, I., Bujor, A., Hant, F., Mummery, C., Ten Dijke, P. & Trojanowska, M. (2011), 'Endoglin promotes tgf-beta/smad1 signaling in scleroderma fibroblasts', *J Cell Physiol* **226**(12), 3340–8.
URL: http://onlinelibrary.wiley.com/store/10.1002/jcp.22690/asset/22690_ftp.pdf?v=1&t=jchs1ue1&s=2dd62ff0abcfb1817fc4f8ca77f188a1cc71284ehttps://onlinelibrary.wiley.com/doi/pdf/10.1002/jcp.22690

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. (2008), 'Mapping and quantifying mammalian transcriptomes by rna-seq', *Nat Methods* **5**(7), 621–8.
URL: http://www.nature.com/articles/nmeth.1226

Mulder, K. M. (2000), 'Role of ras and mapks in tgf signaling', *Cytokine & growth factor reviews* **11**(1-2), 23–35.
URL: https://www.ncbi.nlm.nih.gov/pubmed/10708950

Muthusamy, B. P., Budi, E. H., Katsuno, Y., Lee, M. K., Smith, S. M., Mirza, A. M., Akhurst, R. J. & Derynck, R. (2015), 'Shca protects against epithelialmesenchymal transition through compartmentalized inhibition of tgf--induced smad activation', *PLoS biology* **13**(12), e1002325.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4682977/pdf/pbio.1002325.pdf

Nadel, E. T. H. A. N. R., Bullard, R. O. B. E. R. T. W. & Stolwijk, J. A. (1971), 'Importance of skin temperature in the regulation of sweating', *Journal of applied physiology* **31**(1), 80–87.
URL: https://www.physiology.org/doi/pdf/10.1152/jappl.1971.31.1.80

Nakao, A., Afrakhte, M., Morn, A., Nakayama, T., Christian, J. L., Heuchel, R., Itoh, S., Kawabata, M., Heldin, N.-E. & Heldin, C.-H. (1997), 'Identification of smad7, a tgf-inducible antagonist of tgf-signalling', *Nature* **389**(6651), 631.
URL: https://www.nature.com/articles/39369.pdf

Nakerakanti, S. S., Bujor, A. M. & Trojanowska, M. (2011), 'Ccn2 is required for the tgf- induced activation of smad1 - erk1/2 signaling network', *PLOS ONE* **6**(7), e21911.
URL: https://doi.org/10.1371/journal.pone.0021911https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3132735/pdf/pone.0021911.pdfhttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3132735/pdf/pone.0021911.pdf

Nakerakanti, S. S., Kapanadze, B., Yamasaki, M., Markiewicz, M. & Trojanowska, M. (2006), 'Fli1 and ets1 have distinct roles in connective tissue growth factor/ccn2 gene regulation and induction of the profibrotic gene program', *Journal of Biological Chemistry* **281**(35), 25259–25269.
URL: http://www.jbc.org/content/281/35/25259.abstracthttp://www.jbc.org/content/281/35/25259.full.pdfhttp://www.jbc.org/content/281/35/25259.full.pdf

Negmadjanov, U., Godic, Z., Rizvi, F., Emelyanova, L., Ross, G., Richards, J., Holmuhamedov, E. L. & Jahangir, A. (2015), 'Tgf-beta1-mediated differentiation of fibroblasts is associated with increased mitochondrial content and cellular respiration', *PLoS One* **10**(4), e0123046.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4388650/pdf/pone.0123046.pdf

Nelson, G., Wordsworth, J., Wang, C., Jurk, D., Lawless, C., Martin-Ruiz, C. & von Zglinicki, T. (2012), 'A senescent cell bystander effect: senescence-induced senescence', *Aging Cell* **11**(2), 345–9.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3488292/pdf/acel0011-0345.pdf

Nguyen, L. K., Degasperi, A., Cotter, P. & Kholodenko, B. N. (2015), 'Dyvipac: an integrated analysis and visualisation framework to probe multi-dimensional biological networks', *Sci Rep* **5**, 12569.

URL: https://www.ncbi.nlm.nih.gov/pubmed/26220783https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4518224/pdf/srep12569.pdf

Nguyen, M. D. & Kang, K. A. (2016), 'Mmp-14 triggered fluorescence contrast agent', *Adv Exp Med Biol* **923**, 413–419.
URL: https://link.springer.com/content/pdf/10.1007%2F978-3-319-38810-6_54.pdf

Nguyen, U., Squaglia, N., Boge, A. & Fung, P. A. (2011), 'The simple western: a gel-free, blot-free, hands-free western blotting reinvention', *Nature Methods* **8**(11).

Paizs, B. & Suhai, S. (2005), 'Fragmentation pathways of protonated peptides', *Mass Spectrom Rev* **24**(4), 508–48.
URL: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mas.20024

Pan, M.-R., Li, K., Lin, S.-Y. & Hung, W.-C. (2016), 'Connecting the dots: From dna damage and repair to aging', *International Journal of Molecular Sciences* **17**(5), 685.
URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4881511/https://res.mdpi.com/def50200b6ae1d53f475c25507454a00e9534cfab5cfbca90b17cb99f5701cc530970a0a318325a2e41aedeb82cedcbbee385filename=&attachment=1

Park, S.-A., Kim, M.-J., Park, S.-Y., Kim, J.-S., Lim, W., Nam, J.-S. & Yhong Sheen, Y. (2015), 'Timp-1 mediates tgf--dependent crosstalk between hepatic stellate and cancer cells via fak signaling', *Scientific Reports* **5**, 16492.
URL: http://dx.doi.org/10.1038/srep16492https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4637930/pdf/srep16492.pdfhttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4637930/pdf/srep16492.pdf

Pawitan, Y. (2001), *In all likelihood: statistical modelling and inference using likelihood*, Oxford University Press.

Penheiter, S. G., Mitchell, H., Garamszegi, N., Edens, M., Dore, J. J., J. & Leof, E. B. (2002), 'Internalization-dependent and -independent requirements for transforming growth factor beta receptor signaling via the smad pathway', *Mol Cell Biol* **22**(13), 4750–9.
URL: https://www.ncbi.nlm.nih.gov/pubmed/12052882https://www.ncbi.nlm.nih.gov/pmc/articles/PMC133902/pdf/1514.pdfhttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC133902/pdf/1514.pdf

Perutz, M. F. (1989), 'Mechanisms of cooperativity and allosteric regulation in proteins', *Quarterly Reviews of Biophysics* **22**(2), 139–237.
URL: https://www.cambridge.org/core/article/mechanisms-of-cooperativity-and-allosteric-regulation-in-proteins/

D27C634EDA98D62A21D3A08D19EDF114https:
//www.cambridge.org/core/services/aop-cambridge-core/content/view/
D27C634EDA98D62A21D3A08D19EDF114/S0033583500003826a.pdf/
div-class-title-mechanisms-of-cooperativity-and-allosteric-regulation-in-proteins-div.
pdf

Petrovic, M. & Barcel, D. (2013), 'Liquid chromatographytandem mass spectrometry', *Analytical and Bioanalytical Chemistry* **405**(18), 5857–5858.
URL: https://doi.org/10.1007/s00216-013-7018-7https:
//link.springer.com/content/pdf/10.1007%2Fs00216-013-7018-7.pdf

Plotnik, J. P., Budka, J. A., Ferris, M. W. & Hollenhorst, P. C. (2014), 'Ets1 is a genome-wide effector of ras/erk signaling in epithelial cells', *Nucleic Acids Res* **42**(19), 11928–40.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4231772/pdf/gku929.pdf

Ponticos, M., Harvey, C., Ikeda, T., Abraham, D. & Bou-Gharios, G. (2009), 'Junb mediates enhancer/promoter activity of col1a2 following tgf-beta induction', *Nucleic Acids Res* **37**(16), 5378–89.
URL: https://academic.oup.com/nar/article/37/16/5378/2410013

Ponticos, M., Holmes, A. M., Shi-wen, X., Leoni, P., Khan, K., Rajkumar, V. S., Hoyles, R. K., Bou-Gharios, G., Black, C. M., Denton, C. P., Abraham, D. J., Leask, A. & Lindahl, G. E. (2009), 'Pivotal role of connective tissue growth factor in lung fibrosis: Mapk-dependent transcriptional activation of type i collagen', *Arthritis Rheum* **60**(7), 2142–55.
URL: http://onlinelibrary.wiley.com/store/10.1002/art.24620/asset/24620_ftp.pdf?v=
1&t=jchs2szg&s=6f131dcfb5ee8f91009282ab6b166496583761c4https:
//onlinelibrary.wiley.com/doi/pdf/10.1002/art.24620

Ponticos, M., Papaioannou, I., Xu, S., Holmes, A. M., Khan, K., Denton, C. P., Bou-Gharios, G. & Abraham, D. J. (2015), 'Failed degradation of junb contributes to overproduction of type i collagen and development of dermal fibrosis in patients with systemic sclerosis', *Arthritis & Rheumatology* **67**(1), 243–253.
URL: http://dx.doi.org/10.1002/art.38897http:
//onlinelibrary.wiley.com/store/10.1002/art.38897/asset/art38897.pdf?v=1&t=i6dl573d&
s=9922d387fec47a2be7271fe36a067fdd56e7899bhttps:
//onlinelibrary.wiley.com/doi/pdf/10.1002/art.38897

Pulaski, L., Landstrom, M., Heldin, C. H. & Souchelnytskyi, S. (2001), 'Phosphorylation of smad7 at ser-249 does not interfere with its inhibitory role in transforming growth factor-beta-dependent

signaling but affects smad7-dependent transcriptional activation', *J Biol Chem* **276**(17), 14344–9.
URL: http://www.jbc.org/content/276/17/14344.full.pdf

Purohit, T., He, T., Qin, Z., Li, T., Fisher, G. J., Yan, Y., Voorhees, J. J. & Quan, T. (2016), 'Smad3-dependent regulation of type i collagen in human dermal fibroblasts: Impact on human skin connective tissue aging', *J Dermatol Sci* **83**(1), 80–3.
URL:
https://www.ncbi.nlm.nih.gov/pubmed/27132061https://www.sciencedirect.com/science/article/pii/S0923181116300597?via%3Dihubhttps://ac.els-cdn.com/S0923181116300597/1-s2.0-S0923181116300597-main.pdf?_tid=a43cbe72-5445-44bc-994e-979be2f88456&acdnat=1532173928_c91ab86536e90f0f29e21c2bc16ed499

Qian, Y. & Chen, X. (2010), 'Tumor suppression by p53: making cells senescent', *Histol Histopathol* **25**(4), 515–26.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2841029/pdf/nihms184271.pdf

Qin, Z., Balimunkwe, R. M. & Quan, T. (2017), 'Agerelated reduction of dermal fibroblast size upregulates multiple matrix metalloproteinases as observed in aged human skin in vivo', *British Journal of Dermatology* .

Quan, T. & Fisher, G. J. (2015), 'Role of age-associated alterations of the dermal extracellular matrix microenvironment in human skin aging: a mini-review', *Gerontology* **61**(5), 427–434.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4524793/pdf/nihms654343.pdf

Quan, T., He, T., Kang, S., Voorhees, J. J. & Fisher, G. J. (2002), 'Connective tissue growth factor: expression in human skin in vivo and inhibition by ultraviolet irradiation', *J Invest Dermatol* **118**(3), 402–8.
URL: http://www.jidonline.org/article/S0022-202X(15)41581-7/pdfhttps://ac.els-cdn.com/S0022202X15415817/1-s2.0-S0022202X15415817-main.pdf?_tid=47a61641-7f97-460c-9e6b-301d275da20e&acdnat=1532173936_8d66d32e7180c21dca6e9b6391a91d9b

Quan, T., He, T., Kang, S., Voorhees, J. J. & Fisher, G. J. (2004), 'Solar ultraviolet irradiation reduces collagen in photoaged human skin by blocking transforming growth factor-beta type ii receptor/smad signaling', *Am J Pathol* **165**(3), 741–51.
URL: https://www.ncbi.nlm.nih.gov/pubmed/15331399https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1618600/pdf/JPATH165000741.pdfhttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC1618600/pdf/JPATH165000741.pdf

Quan, T., Little, E., Quan, H., Qin, Z., Voorhees, J. J. & Fisher, G. J. (2013), 'Elevated matrix metalloproteinases and collagen fragmentation in photodamaged human skin: impact of altered

extracellular matrix microenvironment on dermal fibroblast function', *J Invest Dermatol* **133**(5), 1362–6.
URL: https://www.ncbi.nlm.nih.gov/pubmed/23466932https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3637921/pdf/nihms429894.pdfhttps://ac.els-cdn.com/S0022202X15362370/1-s2.0-S0022202X15362370-main.pdf?_tid=28caea56-6f06-48e6-b57e-4f96eb247ef7&acdnat=1532173949_593724a104cf23ead46fd7e38212de17

Quan, T., Qin, Z., Xia, W., Shao, Y., Voorhees, J. J. & Fisher, G. J. (2009), 'Matrix-degrading metalloproteinases in photoaging', *J Investig Dermatol Symp Proc* **14**(1), 20–4.
URL: https://www.ncbi.nlm.nih.gov/pubmed/19675548https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2909639/pdf/nihms201357.pdfhttps://ac.els-cdn.com/S1087002415305050/1-s2.0-S1087002415305050-main.pdf?_tid=eecfddb4-fde2-4f76-a465-c7620c74239c&acdnat=1533054634_dab43d6059b5bf45279cbbb8be2fb91f

Quan, T., Shao, Y., He, T., Voorhees, J. J. & Fisher, G. J. (2010), 'Reduced expression of connective tissue growth factor (ctgf/ccn2) mediates collagen loss in chronologically aged human skin', *J Invest Dermatol* **130**(2), 415–24.
URL: https://ac.els-cdn.com/S0022202X15346984/1-s2.0-S0022202X15346984-main.pdf?_tid=df9cd622-3caf-49f3-8a26-9dd1da9778bb&acdnat=1532173955_92dfb0d9f0580c9e06efb8709f580525

Rabiner, L., Rosenberg, A. & Levinson, S. (1978), 'Considerations in dynamic time warping algorithms for discrete word recognition', *IEEE Transactions on Acoustics, Speech, and Signal Processing* **26**(6), 575–582.
URL: https://ieeexplore.ieee.org/ielx6/29/26135/01163164.pdf?tp=&arnumber=1163164&isnumber=26135

Rabitz, H., Kramer, M. & Dacol, D. (1983), 'Sensitivity analysis in chemical kinetics', *Annual review of physical chemistry* **34**(1), 419–461.
URL: https://www.annualreviews.org/doi/pdf/10.1146/annurev.pc.34.100183.002223

Ramamurthi, A. & Kothapalli, C. (2016), *Elastic Fiber Matrices: Biomimetic Approaches to Regeneration and Repair*, CRC Press LLC.
URL: https://books.google.co.uk/books?id=qnQ1jgEACAAJ

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmuller, U. & Timmer, J. (2009), 'Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood', *Bioinformatics* **25**(15), 1923–9.
URL: https://www.ncbi.nlm.nih.gov/pubmed/19505944

Raue, A., Kreutz, C., Theis, F. J. & Timmer, J. (2013), 'Joining forces of bayesian and frequentist methodology: a study for inference in the presence of non-identifiability', *Philos Trans A Math Phys Eng Sci* **371**(1984), 20110544.
URL: https://www.ncbi.nlm.nih.gov/pubmed/23277602http://rsta.royalsocietypublishing.org/content/roypta/371/1984/20110544.full.pdf

Rhee, S. Y., Wood, V., Dolinski, K. & Draghici, S. (2008), 'Use and misuse of the gene ontology annotations', *Nature reviews. Genetics* **9**(7), 509–509.
URL: https://www.nature.com/articles/nrg2363.pdf

Richmond, P., Walker, D., Coakley, S. & Romano, D. (2010), 'High performance cellular level agent-based simulation with flame for the gpu', *Briefings in bioinformatics* **11**(3), 334–347.
URL: https://www.ncbi.nlm.nih.gov/pubmed/20123941

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. & Smyth, G. K. (2015), 'limma powers differential expression analyses for rna-sequencing and microarray studies', *Nucleic acids research* **43**(7), e47–e47.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4402510/pdf/gkv007.pdf

Roth, D. M. & Balch, W. E. (2011), 'Modeling general proteostasis: proteome balance in health and disease', *Current opinion in cell biology* **23**(2), 126–134.
URL: https://www.ncbi.nlm.nih.gov/pubmed/21131189

Runyan, C. E., Schnaper, H. W. & Poncelet, A. C. (2003), 'Smad3 and pkcdelta mediate tgf-beta1-induced collagen i expression in human mesangial cells', *Am J Physiol Renal Physiol* **285**(3), F413–22.
URL: https://www.physiology.org/doi/pdf/10.1152/ajprenal.00082.2003

Rykiel, E. J. (1996), 'Testing ecological models: the meaning of validation', *Ecological Modelling* **90**(3), 229–244.
URL: https://www.sciencedirect.com/science/article/pii/0304380095001522

Sadygov, R. G., Cociorva, D. & Yates, J. R., r. (2004), 'Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book', *Nat Methods* **1**(3), 195–202.
URL: https://www.nature.com/articles/nmeth725.pdf

Saez, I. & Vilchez, D. (2014), 'The mechanistic links between proteasome activity, aging and age-related diseases', *Current Genomics* **15**(1), 38–51.
URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3958958/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3958958/pdf/CG-15-38.pdfhttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3958958/pdf/CG-15-38.pdf

Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., Nativ, N., Bahir, I., Doniger, T., Krug, H., Sirota-Madi, A., Olender, T., Golan, Y., Stelzer, G., Harel, A. & Lancet, D. (2010), 'Genecards version 3: the human gene integrator', *Database: The Journal of Biological Databases and Curation* **2010**, baq020.
URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938269/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938269/pdf/baq020.pdfhttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938269/pdf/baq020.pdf

Salvador, S. & Chan, P. (2007), 'Fastdtw: Toward accurate dynamic time', *Warping in Linear Time and Space* .

Samuel, G. H., Bujor, A. M., Nakerakanti, S. S., Hant, F. N. & Trojanowska, M. (2010), 'Autocrine transforming growth factor signaling regulates extracellular signal-regulated kinase 1/2 phosphorylation via modulation of protein phosphatase 2a expression in scleroderma fibroblasts', *Fibrogenesis & tissue repair* **3**(1), 25.
URL: https://fibrogenesis.biomedcentral.com/track/pdf/10.1186/1755-1536-3-25?site=fibrogenesis.biomedcentral.comhttps://fibrogenesis.biomedcentral.com/track/pdf/10.1186/1755-1536-3-25

Sauro, H. M. (2011), *Enzyme kinetics for systems biology*, Future Skill Software.

Schmierer, B., Tournier, A. L., Bates, P. A. & Hill, C. S. (2008), 'Mathematical modeling identifies smad nucleocytoplasmic shuttling as a dynamic signal-interpreting system', *Proceedings of the National Academy of Sciences* **105**(18), 6608–6613.
URL: http://www.pnas.org/content/pnas/105/18/6608.full.pdf

Schultz-Cherry, S. & Murphy-Ullrich, J. E. (1993), 'Thrombospondin causes activation of latent transforming growth factor-beta secreted by endothelial cells by a novel mechanism', *J Cell Biol* **122**(4), 923–32.
URL: http://jcb.rupress.org/content/jcb/122/4/923.full.pdf

Seite, S., Zucchi, H., Septier, D., IgondjoTchen, S., Senni, K. & Godeau, G. (2006), 'Elastin changes during chronological and photoageing: the important role of lysozyme', *Journal of the European Academy of Dermatology and Venereology* **20**(8), 980–987.
URL: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-3083.2006.01706.x

Serban, R. & Hindmarsh, A. C. (2003), Cvodes: An ode solver with sensitivity analysis capabilities, Report, Technical Report UCRL-JP-200039, Lawrence Livermore National Laboratory.

Shannon, C. E. (1948), 'A mathematical theory of communication', *Bell System Technical Journal* **27**(3), 379–423.

URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1948.tb01338.xhttps://ieeexplore.ieee.org/ielx7/6731005/6773023/06773024.pdf?tp=&arnumber=6773024&isnumber=6773023

Shevchenko, A., Wilm, M., Vorm, O. & Mann, M. (1996), 'Mass spectrometric sequencing of proteins from silver-stained polyacrylamide gels', *Analytical Chemistry* **68**(5), 850–858.
URL: https://doi.org/10.1021/ac950914hhttps://pubs.acs.org/doi/pdfplus/10.1021/ac950914hhttps://pubs.acs.org/doi/pdfplus/10.1021/ac950914h

Shi, L. Q. & Ruan, C. L. (2013), 'Expression and significance of mmp-7, c-jun and c-fos in rats skin photoaging', *Asian Pac J Trop Med* **6**(10), 768–70.
URL: https://ac.els-cdn.com/S1995764513601352/1-s2.0-S1995764513601352-main.pdf?_tid=0c31ee45-8585-4af2-b5a9-7a2810eb86e2&acdnat=1532178411_b01949c5ab0fc2035b621c7077b201a1

Shi, W., Sun, C., He, B., Xiong, W., Shi, X., Yao, D. & Cao, X. (2004), 'Gadd34-pp1c recruited by smad7 dephosphorylates tgfbeta type i receptor', *J Cell Biol* **164**(2), 291–300.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2172339/pdf/200307151.pdf

Simonsson, M., Heldin, C. H., Ericsson, J. & Gronroos, E. (2005), 'The balance between acetylation and deacetylation controls smad7 stability', *J Biol Chem* **280**(23), 21797–803.
URL: http://www.jbc.org/content/280/23/21797.full.pdf

Sjerobabski-Masnec, I. & Situm, M. (2010), 'Skin aging', *Acta Clin Croat* **49**(4), 515–8.

Smith, L. P., Bergmann, F. T., Chandran, D. & Sauro, H. M. (2009), 'Antimony: a modular model definition language', *Bioinformatics* **25**(18), 2452–2454.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2735663/pdf/btp401.pdf

Smyth, G. K. (2004), 'Linear models and empirical bayes methods for assessing differential expression in microarray experiments', *Statistical applications in genetics and molecular biology* **3**(1), 1–25.

Smyth, G. K. (2005), *Limma: linear models for microarray data*, Springer, pp. 397–420.

Sonnylal, S., ShiWen, X., Leoni, P., Naff, K., Van Pelt, C. S., Nakamura, H., Leask, A., Abraham, D., BouGharios, G. & de Crombrugghe, B. (2010), 'Selective expression of connective tissue growth factor in fibroblasts in vivo promotes systemic tissue fibrosis', *Arthritis & Rheumatology* **62**(5), 1523–1532.

Sorrentino, A., Thakur, N., Grimsby, S., Marcusson, A., von Bulow, V., Schuster, N., Zhang, S., Heldin, C. H. & Landstrom, M. (2008), 'The type i tgf-beta receptor engages traf6 to activate tak1 in a receptor kinase-independent manner', *Nat Cell Biol* **10**(10), 1199–207.
URL: https://www.nature.com/articles/ncb1780.pdf

Souchelnytskyi, S., Nakayama, T., Nakao, A., Morn, A., Heldin, C.-H., Christian, J. L. & Ten Dijke, P. (1998), 'Physical and functional interaction of murine andxenopus smad7 with bone morphogenetic protein receptors and transforming growth factor- receptors', *Journal of Biological Chemistry* **273**(39), 25364–25370.
URL: http://www.jbc.org/content/273/39/25364.full.pdf

Southern, E. M. (1975), 'Detection of specific sequences among dna fragments separated by gel electrophoresis', *J mol biol* **98**(3), 503–517.
URL: https://www.sciencedirect.com/science/article/pii/S0022283675800830

Stroschein, S. L., Wang, W., Zhou, S., Zhou, Q. & Luo, K. (1999), 'Negative feedback regulation of tgf-signaling by the snon oncoprotein', *Science* **286**(5440), 771–774.
URL: http://science.sciencemag.org/content/sci/286/5440/771.full.pdf

Su, S.-a., Yang, D., Wu, Y., Xie, Y., Zhu, W., Cai, Z., Shen, J., Fu, Z., Wang, Y. & Jia, L. (2017), 'Ephrinb2 regulates cardiac fibrosis through modulating the interaction of stat3 and tgf-/smad3 signalingnovelty and significance', *Circulation research* **121**(6), 617–627.
URL: http://circresaha.smart01.highwire.org/content/circresaha/121/6/617.full.pdf

Sun, X., Wu, F., Datta, R., Kharbanda, S. & Kufe, D. (2000), 'Interaction between protein kinase c delta and the c-abl tyrosine kinase in the cellular response to oxidative stress', *J Biol Chem* **275**(11), 7470–3.
URL: http://www.jbc.org/content/275/11/7470.full.pdf

Sutandy, F. X., Qian, J., Chen, C. S. & Zhu, H. (2013), 'Overview of protein microarrays', *Curr Protoc Protein Sci* **Chapter 27**, Unit 27.1.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3680110/pdf/nihms465562.pdf

Suzuki, C., Murakami, G., Fukuchi, M., Shimanuki, T., Shikauchi, Y., Imamura, T. & Miyazono, K. (2002), 'Smurf1 regulates the inhibitory activity of smad7 by targeting smad7 to the plasma membrane', *Journal of Biological Chemistry* **277**(42), 39919–39925.
URL: http://www.jbc.org/content/277/42/39919.full.pdf

Suzuki, K., Wilkes, M. C., Garamszegi, N., Edens, M. & Leof, E. B. (2007), 'Transforming growth factor beta signaling via ras in mesenchymal cells requires p21-activated kinase 2 for extracellular signal-regulated kinase-dependent transcriptional responses', *Cancer Res* **67**(8), 3673–82.
URL: https://www.ncbi.nlm.nih.gov/pubmed/17440079http://cancerres.aacrjournals.org/content/canres/67/8/3673.full.pdfhttp://cancerres.aacrjournals.org/content/canres/67/8/3673.full.pdf

Takahashi, K., Ishikawa, N., Sadamoto, Y., Sasamoto, H., Ohta, S., Shiozawa, A., Miyoshi, F., Naito, Y., Nakayama, Y. & Tomita, M. (2003), 'E-cell 2: Multi-platform e-cell simulation system', *Bioinformatics* **19**(13), 1727–1729.

Takahashi, K., Kaizu, K., Hu, B. & Tomita, M. (2004), 'A multi-algorithm, multi-timescale method for cell simulation', *Bioinformatics* **20**(4), 538–46.
URL: https://www.ncbi.nlm.nih.gov/pubmed/14990450

Tang, H., Klopfenstein, D., Pedersen, B., Flick, P., Sato, K., Ramirez, F., Yunes, J. & Mungall, C. (2015), 'Goatools: tools for gene ontology', *Zenodo. doi* **10**.

Tang, L.-Y., Heller, M., Meng, Z., Yu, L.-R., Tang, Y., Zhou, M. & Zhang, Y. E. (2017), 'Transforming growth factor- (tgf-) directly activates the jak1-stat3 axis to induce hepatic fibrosis in coordination with the smad pathway', *Journal of Biological Chemistry* **292**(10), 4302–4312.
URL: http://www.jbc.org/content/292/10/4302.full.pdfhttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5354477/pdf/zbc01017004302.pdf

Tanzer, M. L. (1973), 'Cross-linking of collagen', *Science* **180**(4086), 561–566.
URL: http://science.sciencemag.org/content/sci/180/4086/561.full.pdf

Thomas, B. R., Chylek, L. A., Colvin, J., Sirimulla, S., Clayton, A. H., Hlavacek, W. S. & Posner, R. G. (2016), 'Bionetfit: a fitting tool compatible with bionetgen, nfsim and distributed computing environments', *Bioinformatics* **32**(5), 798–800.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4907397/pdf/btv655.pdf

Tomita, M., Hashimoto, K., Takahashi, K., Shimizu, T. S., Matsuzaki, Y., Miyoshi, F., Saito, K., Tanida, S., Yugi, K., Venter, J. C. & Hutchison, C. A., r. (1999), 'E-cell: software environment for whole-cell simulation', *Bioinformatics* **15**(1), 72–84.

Tonsing, C., Timmer, J. & Kreutz, C. (2014), 'Cause and cure of sloppiness in ordinary differential equation models', *Physical Review E* **90**(2), 023303.
URL: https://link.aps.org/doi/10.1103/PhysRevE.90.023303https://journals.aps.org/pre/abstract/10.1103/PhysRevE.90.023303

Tsai, T. Y.-C., Choi, Y. S., Ma, W., Pomerening, J. R., Tang, C. & Ferrell, J. E. (2008), 'Robust, tunable biological oscillations from interlinked positive and negative feedback loops', *Science* **321**(5885), 126–129.
URL: http://science.sciencemag.org/content/sci/321/5885/126.full.pdf

Ueno, M., Maeno, T., Nomura, M., Aoyagi-Ikeda, K., Matsui, H., Hara, K., Tanaka, T., Iso, T., Suga, T. & Kurabayashi, M. (2011), 'Hypoxia-inducible factor-1alpha mediates tgf-beta-induced pai-1

production in alveolar macrophages in pulmonary fibrosis', *Am J Physiol Lung Cell Mol Physiol* **300**(5), L740–52.
URL: https://www.physiology.org/doi/pdf/10.1152/ajplung.00146.2010

Ursini-Siegel, J., Hardy, W. R., Zheng, Y., Ling, C., Zuo, D., Zhang, C., Podmore, L., Pawson, T. & Muller, W. J. (2012), 'The shca sh2 domain engages a 14-3-3/pi3 k signaling complex and promotes breast cancer cell survival', *Oncogene* **31**(48), 5038–5044.
URL: https://www.nature.com/articles/onc20124.pdf

Valdimarsdottir, G., Goumans, M.-J., Itoh, F., Itoh, S., Heldin, C.-H. & ten Dijke, P. (2006), 'Smad7 and protein phosphatase 1 are critical determinants in the duration of tgf-/alk1 signaling in endothelial cells', *BMC cell biology* **7**(1), 16.
URL: https://bmccellbiol.biomedcentral.com/track/pdf/10.1186/1471-2121-7-16

Vanlier, J., Tiemann, C. A., Hilbers, P. A. J. & van Riel, N. A. W. (2013), 'Parameter uncertainty in biochemical models described by ordinary differential equations', *Mathematical Biosciences* **246**(2), 305–314.
URL: http://www.sciencedirect.com/science/article/pii/S0025556413000783https://ac.els-cdn.com/S0025556413000783/1-s2.0-S0025556413000783-main.pdf?_tid=f14bcfbf-429c-4078-9748-4ab285c28f01&acdnat=1529398543_20468f90e13cfaba7cc7335c07ff7699https://ac.els-cdn.com/S0025556413000783/1-s2.0-S0025556413000783-main.pdf?_tid=663bb1a8-1a41-4ab5-87b1-d3c0e084c56a&acdnat=1533060593_9c066a9470eb0a1f7ed7ff49bd88f70f

Varani, J., Dame, M. K., Rittie, L., Fligiel, S. E. G., Kang, S., Fisher, G. J. & Voorhees, J. J. (2006), 'Decreased collagen production in chronologically aged skin : Roles of age-dependent alteration in fibroblast function and defective mechanical stimulation', *The American Journal of Pathology* **168**(6), 1861–1868.
URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1606623/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1606623/pdf/JPATH168001861.pdfhttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC1606623/pdf/JPATH168001861.pdf

Varga, J., Rosenbloom, J. & Jimenez, S. A. (1987), 'Transforming growth factor beta (tgf beta) causes a persistent increase in steady-state amounts of type i and type iii collagen and fibronectin mrnas in normal human dermal fibroblasts', *Biochem J* **247**(3), 597–604.
URL:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1148454/pdf/biochemj00244-0105.pdf

Vasilaki, E., Papadimitriou, E., Tajadura, V., Ridley, A. J., Stournaras, C. & Kardassis, D. (2010), 'Transcriptional regulation of the small gtpase rhob gene by tgf-induced signaling pathways', *The*

FASEB Journal **24**(3), 891–905.
URL: http://www.fasebj.org/content/24/3/891.full.pdf

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A. & Holt, R. A. (2001), 'The sequence of the human genome', *science* **291**(5507), 1304–1351.
URL: http://science.sciencemag.org/content/sci/291/5507/1304.full.pdf

Verrecchia, F., Pessah, M., Atfi, A. & Mauviel, A. (2000), 'Tumor necrosis factor-alpha inhibits transforming growth factor-beta /smad signaling in human dermal fibroblasts via ap-1 activation', *J Biol Chem* **275**(39), 30226–31.
URL: https://www.ncbi.nlm.nih.gov/pubmed/10903323http://www.jbc.org/content/275/39/30226.full.pdfhttp://www.jbc.org/content/275/39/30226.full.pdf

Verzijl, N., DeGroot, J., Thorpe, S. R., Bank, R. A., Shaw, J. N., Lyons, T. J., Bijlsma, J. W., Lafeber, F. P., Baynes, J. W. & TeKoppele, J. M. (2000), 'Effect of collagen turnover on the accumulation of advanced glycation end products', *J Biol Chem* **275**(50), 39027–31.
URL: http://www.jbc.org/content/275/50/39027.full.pdf

Vilar, J. M., Jansen, R. & Sander, C. (2006), 'Signal processing in the tgf-beta superfamily ligand-receptor network', *PLoS Comput Biol* **2**(1), e3.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1356091/pdf/pcbi.0020003.pdf

Vincent, T., Neve, E. P. A., Johnson, J. R., Kukalev, A., Rojo, F., Albanell, J., Pietras, K., Virtanen, I., Philipson, L. & Leopold, P. L. (2009), 'A snail1smad3/4 transcriptional repressor complex promotes tgf- mediated epithelialmesenchymal transition', *Nature cell biology* **11**(8), 943.
URL: https://www.nature.com/articles/ncb1905.pdf

Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C. & Chen, F. (2009), 'Chip-seq accurately predicts tissue-specific activity of enhancers', *Nature* **457**(7231), 854.
URL: https://www.nature.com/articles/nature07730.pdf

Vizan, P., Miller, D. S., Gori, I., Das, D., Schmierer, B. & Hill, C. S. (2013), 'Controlling long-term signaling: receptor dynamics determine attenuation and refractory behavior of the tgf-beta pathway', *Sci Signal* **6**(305), ra106.
URL: http://stke.sciencemag.org/content/sigtrans/6/305/ra106.full.pdf

von Gersdorff, G., Susztak, K., Rezvani, F., Bitzer, M., Liang, D. & Bottinger, E. P. (2000), 'Smad3 and smad4 mediate transcriptional activation of the human smad7 promoter by transforming growth

factor beta', *J Biol Chem* **275**(15), 11320–6.
URL: http://www.jbc.org/content/275/15/11320.full.pdf

Wahab, N. A., Schaefer, L., Weston, B. S., Yiannikouris, O., Wright, A., Babelova, A., Schaefer, R. & Mason, R. M. (2005), 'Glomerular expression of thrombospondin-1, transforming growth factor beta and connective tissue growth factor at different stages of diabetic nephropathy and their interdependent roles in mesangial response to diabetic stimuli', *Diabetologia* **48**(12), 2650–60.
URL: https://www.ncbi.nlm.nih.gov/pubmed/16270194https://link.springer.com/content/pdf/10.1007%2Fs00125-005-0006-5.pdfhttps://link.springer.com/content/pdf/10.1007%2Fs00125-005-0006-5.pdf

Wahab, N. A., Weston, B. S. & Mason, R. M. (2005), 'Modulation of the tgfbeta/smad signaling pathway in mesangial cells by ctgf/ccn2', *Exp Cell Res* **307**(2), 305–14.
URL: https://www.ncbi.nlm.nih.gov/pubmed/15950619

Walther, T. C. & Mann, M. (2010), 'Mass spectrometrybased proteomics in cell biology', *The Journal of cell biology* **190**(4), 491–500.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2928005/pdf/JCB_201004052.pdf

Wang, G., Yu, Y., Sun, C., Liu, T., Liang, T., Zhan, L., Lin, X. & Feng, X. H. (2016), 'Stat3 selectively interacts with smad3 to antagonize tgf-beta signalling', *Oncogene* **35**(33), 4388–98.
URL: https://www.nature.com/articles/onc2015446.pdf

Wang, L., Zhang, Z. G., Zhang, R. L., Gregg, S. R., Hozeska-Solgot, A., LeTourneau, Y., Wang, Y. & Chopp, M. (2006), 'Matrix metalloproteinase 2 (mmp2) and mmp9 secreted by erythropoietin-activated endothelial cells promote neural progenitor cell migration', *J Neurosci* **26**(22), 5996–6003.
URL: http://www.jneurosci.org/content/jneuro/26/22/5996.full.pdf

Wang, X., Bi, Z., Chu, W. & Wan, Y. (2005), 'Il-1 receptor antagonist attenuates map kinase/ap-1 activation and mmp1 expression in uva-irradiated human fibroblasts induced by culture medium from uvb-irradiated human skin keratinocytes', *Int J Mol Med* **16**(6), 1117–24.
URL: http://www.spandidos-publications.com/ijmm/16/6/1117https://www.spandidos-publications.com/ijmm/16/6/1117https://www.spandidos-publications.com/ijmm/16/6/1117

Washburn, M. P., Wolters, D. & Yates III, J. R. (2001), 'Large-scale analysis of the yeast proteome by multidimensional protein identification technology', *Nature biotechnology* **19**(3), 242.
URL: https://www.nature.com/articles/nbt0301_242.pdf

Watson, J. D. & Crick, F. H. C. (1953*a*), 'Genetical implications of the structure of deoxyribonucleic acid', *Nature* **171**(4361), 964–967.
URL: http://www.nature.com/articles/171964a0.pdf

Watson, J. D. & Crick, F. H. C. (1953*b*), 'Molecular structure of nucleic acids', *Nature* **171**(4356), 737–738.

Weismann, A., Poulton, E. B. & Shipley, A. E. (1891), *Essays upon heredity and kindred biological problems*, Vol. 1, Clarendon press.

Welsh, C. M., Fullard, N., Proctor, C. J., Martinez-Guimera, A., Isfort, R. J., Bascom, C. C., Tasseff, R., Przyborski, S. A. & Shanley, D. P. (2018), 'Pycotools: A python toolbox for copasi', *Bioinformatics* **1**, 9.
URL: https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty409/5001390

Wendling, J., Marchand, A., Mauviel, A. & Verrecchia, F. (2003), '5-fluorouracil blocks transforming growth factor-beta-induced alpha 2 type i collagen gene (col1a2) expression in human fibroblasts via c-jun nh2-terminal kinase/activator protein-1 activation', *Mol Pharmacol* **64**(3), 707–13.
URL: https://www.ncbi.nlm.nih.gov/pubmed/12920208http://molpharm.aspetjournals.org/content/64/3/707.longhttp://molpharm.aspetjournals.org/content/molpharm/64/3/707.full.pdfhttp://molpharm.aspetjournals.org/content/molpharm/64/3/707.full.pdf

Weston, B. S., Wahab, N. A. & Mason, R. M. (2003), 'Ctgf mediates tgf-beta-induced fibronectin matrix deposition by upregulating active alpha5beta1 integrin in human mesangial cells', *J Am Soc Nephrol* **14**(3), 601–10.
URL: https://www.ncbi.nlm.nih.gov/pubmed/12595495http://jasn.asnjournals.org/content/14/3/601.full.pdfhttp://jasn.asnjournals.org/content/14/3/601.full.pdf

White, L. A., Mitchell, T. I. & Brinckerhoff, C. E. (2000), 'Transforming growth factor inhibitory element in the rabbit matrix metalloproteinase-1 (collagenase-1) gene functions as a repressor of constitutive transcription', *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression* **1490**(3), 259–268.
URL: http://europepmc.org/abstract/med/10684971

Wilensky, U. (2008), 'Netlogo 4.0. 4'.

Wilkes, M. C. & Leof, E. B. (2006), 'Transforming growth factor beta activation of c-abl is independent of receptor internalization and regulated by phosphatidylinositol 3-kinase and pak2 in mesenchymal cultures', *J Biol Chem* **281**(38), 27846–54.

URL: https://www.ncbi.nlm.nih.gov/pubmed/16867995http://www.jbc.org/content/281/38/27846.full.pdfhttp://www.jbc.org/content/281/38/27846.full.pdf

Wilkes, M. C., Mitchell, H., Penheiter, S. G., Dore, J. J., Suzuki, K., Edens, M., Sharma, D. K., Pagano, R. E. & Leof, E. B. (2005), 'Transforming growth factor-beta activation of phosphatidylinositol 3-kinase is independent of smad2 and smad3 and regulates fibroblast responses via p21-activated kinase-2', *Cancer Res* **65**(22), 10431–40.
URL: https://www.ncbi.nlm.nih.gov/pubmed/16288034http://cancerres.aacrjournals.org/content/canres/65/22/10431.full.pdfhttp://cancerres.aacrjournals.org/content/canres/65/22/10431.full.pdf

Wilkes, M. C., Murphy, S. J., Garamszegi, N. & Leof, E. B. (2003), 'Cell-type-specific activation of pak2 by transforming growth factor beta independent of smad2 and smad3', *Mol Cell Biol* **23**(23), 8878–89.
URL: https://www.ncbi.nlm.nih.gov/pubmed/14612425https://www.ncbi.nlm.nih.gov/pmc/articles/PMC262664/pdf/0682.pdfhttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC262664/pdf/0682.pdf

Wilkinson, D. J. (2006), *Stochastic modelling for systems biology*, Chapman and Hall/CRC.

Williams, G. C. (1957), 'Pleiotropy, natural selection, and the evolution of senescence', *Evolution* **11**(4), 398–411.
URL: http://www.jstor.org/stable/2406060

Wipff, P.-J., Rifkin, D. B., Meister, J.-J. & Hinz, B. (2007), 'Myofibroblast contraction activates latent tgf-1 from the extracellular matrix', *The Journal of cell biology* **179**(6), 1311–1323.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2140013/pdf/jcb1791311.pdf

Wisniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. (2009), 'Universal sample preparation method for proteome analysis', *Nat Methods* **6**(5), 359–62.
URL: http://www.nature.com/articles/nmeth.1322

Wrana, J. L., Attisano, L., Crcamo, J., Zentella, A., Doody, J., Laiho, M., Wang, X.-F. & Massague, J. (1992), 'Tgf signals through a heteromeric protein kinase receptor complex', *Cell* **71**(6), 1003–1014.
URL: https://www.sciencedirect.com/science/article/pii/009286749290395S?via%3Dihubhttps://ac.els-cdn.com/009286749290395S/1-s2.0-009286749290395S-main.pdf?_tid=861d8918-0051-4f18-b0cd-3788778a227c&acdnat=1533115128_0c9a2a2ba744fc966a5d8ac526206e9e

Wu, J.-W., Krawitz, A. R., Chai, J., Li, W., Zhang, F., Luo, K. & Shi, Y. (2002), 'Structural mechanism of smad4 recognition by the nuclear oncoprotein ski: insights on ski-mediated repression of tgf- signaling', *Cell* **111**(3), 357–367.
URL: https://ac.els-cdn.com/S0092867402010061/1-s2.0-S0092867402010061-main.pdf?_tid=c4e29be2-c372-4711-be9f-9a0921365cca&acdnat=1533302879_e2c527b6fa60fc703ac6e4f31670ab95

Xia, S., Zhang, X., Zheng, S., Khanabdali, R., Kalionis, B., Wu, J., Wan, W. & Tai, X. (2016), 'An update on inflamm-aging: Mechanisms, prevention, and treatment', *Journal of Immunology Research* **2016**, 8426874.
URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4963991/

Yamashita, M., Fatyol, K., Jin, C., Wang, X., Liu, Z. & Zhang, Y. E. (2008), 'Traf6 mediates smad-independent activation of jnk and p38 by tgf-beta', *Mol Cell* **31**(6), 918–24.
URL: https://ac.els-cdn.com/S1097276508006175/1-s2.0-S1097276508006175-main.pdf?_tid=af4cdab8-5358-48e9-b293-dc613a0da94e&acdnat=1531820997_d4c27f1753e2d2187013bda2a8fbb4dbhttps://ac.els-cdn.com/S1097276508006175/1-s2.0-S1097276508006175-main.pdf?_tid=fff21614-2aa5-4f79-a174-fc1d157dfb27&acdnat=1533115174_1cc54796b9a4f305a6d867a8d0ffee0e

Yan, X. & Chen, Y. G. (2011), 'Smad7: not only a regulator, but also a cross-talk mediator of tgf-beta signalling', *Biochem J* **434**(1), 1–10.
URL: http://www.biochemj.org/content/434/1/1.longhttp://www.biochemj.org/content/ppbiochemj/434/1/1.full.pdfhttp://www.biochemj.org/content/ppbiochemj/434/1/1.full.pdf

Yan, X., Liao, H., Cheng, M., Shi, X., Lin, X., Feng, X. H. & Chen, Y. G. (2016), 'Smad7 protein interacts with receptor-regulated smads (r-smads) to inhibit transforming growth factor-beta (tgf-beta)/smad signaling', *J Biol Chem* **291**(1), 382–92.
URL: https://www.ncbi.nlm.nih.gov/pubmed/26555259https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4697173/pdf/zbc382.pdf

Yan, X., Xiong, X. & Chen, Y. G. (2017), 'Feedback regulation of tgf-beta signaling', *Acta Biochim Biophys Sin (Shanghai)* .
URL: https://www.ncbi.nlm.nih.gov/pubmed/29228156

Yuan, W. & Varga, J. (2001), 'Transforming growth factor-beta repression of matrix metalloproteinase-1 in dermal fibroblasts involves smad3', *J Biol Chem* **276**(42), 38502–10.
URL: http://www.jbc.org/content/276/42/38502.full.pdf

Yuan, Z. M., Utsugisawa, T., Ishiko, T., Nakada, S., Huang, Y., Kharbanda, S., Weichselbaum, R. & Kufe, D. (1998), 'Activation of protein kinase c delta by the c-abl tyrosine kinase in response to ionizing radiation', *Oncogene* **16**(13), 1643–8.
URL: https://www.nature.com/articles/1201698.pdf

Zanotti, S., Gibertini, S. & Mora, M. (2010), 'Altered production of extra-cellular matrix components by muscle-derived duchenne muscular dystrophy fibroblasts before and after tgf-1 treatment', *Cell and tissue research* **339**(2), 397–410.
URL: https://link.springer.com/content/pdf/10.1007%2Fs00441-009-0889-4.pdf

Zhang, S., Fei, T., Zhang, L., Zhang, R., Chen, F., Ning, Y., Han, Y., Feng, X. H., Meng, A. & Chen, Y. G. (2007), 'Smad7 antagonizes transforming growth factor beta signaling in the nucleus by interfering with functional smad-dna complex formation', *Mol Cell Biol* **27**(12), 4488–99.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1900056/pdf/1636-06.pdf

Zhang, W., Yuan, J., Yang, Y., Xu, L., Wang, Q., Zuo, W., Fang, X. & Chen, Y. G. (2010), 'Monomeric type i and type iii transforming growth factor-beta receptors and their dimerization revealed by single-molecule imaging', *Cell Res* **20**(11), 1216–23.
URL: https://www.nature.com/articles/cr2010105.pdf

Zhang, Y. E. (2009), 'Non-smad pathways in tgf- signaling', *Cell research* **19**(1), 128.
URL: https://www.nature.com/articles/cr2008328.pdf

Zhang, Y. E. (2018), 'Mechanistic insight into contextual tgf- signaling', *Current opinion in cell biology* **51**, 1–7.
URL:
https://www.sciencedirect.com/science/article/pii/S0955067417301230?via%3Dihubhttps://ac.els-cdn.com/S0955067417301230/1-s2.0-S0955067417301230-main.pdf?_tid=3daa8b92-584c-4cc1-934b-e0b5e59b73a2&acdnat=1533115274_6d599fb723169a92c88608f8970b6abb

Zhang, Y., Feng, X.-H. & Derynck, R. (1998), 'Smad3 and smad4 cooperate with c-jun/c-fos to mediate tgf--induced transcription', *Nature* **394**(6696), 909.
URL: https://www.nature.com/articles/29814.pdf

Zi, Z. (2011), 'Sensitivity analysis approaches applied to systems biology models', *IET systems biology* **5**(6), 336–346.

Zi, Z., Feng, Z., Chapnick, D. A., Dahl, M., Deng, D., Klipp, E., Moustakas, A. & Liu, X. (2011), 'Quantitative analysis of transient and sustained transforming growth factor signaling dynamics',

Molecular systems biology **7**(1), 492.

URL: http://msb.embopress.org/content/msb/7/1/492.full.pdf

Zi, Z. & Klipp, E. (2007), 'Constraint-based modeling and kinetic analysis of the smad dependent tgf-signaling pathway', *PloS one* **2**(9), e936.

URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1978528/pdf/pone.0000936.pdf

Appendix A

PyCoTools: A Python toolbox for COPASI

Bioinformatics, 2018, 1–9
doi: 10.1093/bioinformatics/bty409
Advance Access Publication Date: 22 May 2018
Original Paper

OXFORD

Systems biology

PyCoTools: a Python toolbox for COPASI

Ciaran M. Welsh[1], Nicola Fullard[3], Carole J. Proctor[2], Alvaro Martinez-Guimera[1], Robert J. Isfort[4], Charles C. Bascom[4], Ryan Tasseff[4], Stefan A. Przyborski[3,*] and Daryl P. Shanley[1,*]

[1]Institute for Cell and Molecular Biosciences, Newcastle University, Newcastle NE1 7RU, UK, [2]Institute of Cellular Medicine, Newcastle University, Newcastle NE1 7RU, UK, [3]Department of Biosciences, Durham University, Durham DH1 3LE, UK and [4]The Proctor & Gamble Company, Cincinnati, OH 45202, USA

*To whom correspondence should be addressed.
Associate Editor: Jonathan Wren

Received on February 2, 2018; revised on April 23, 2018; editorial decision on May 12, 2018; accepted on May 18, 2018

Abstract

Motivation: COPASI is an open source software package for constructing, simulating and analyzing dynamic models of biochemical networks. COPASI is primarily intended to be used with a graphical user interface but often it is desirable to be able to access COPASI features programmatically, with a high level interface.

Results: PyCoTools is a Python package aimed at providing a high level interface to COPASI tasks with an emphasis on model calibration. PyCoTools enables the construction of COPASI models and the execution of a subset of COPASI tasks including time courses, parameter scans and parameter estimations. Additional 'composite' tasks which use COPASI tasks as building blocks are available for increasing parameter estimation throughput, performing identifiability analysis and performing model selection. PyCoTools supports exploratory data analysis on parameter estimation data to assist with troubleshooting model calibrations. We demonstrate PyCoTools by posing a model selection problem designed to show case PyCoTools within a realistic scenario. The aim of the model selection problem is to test the feasibility of three alternative hypotheses in explaining experimental data derived from neonatal dermal fibroblasts in response to TGF-β over time. PyCoTools is used to critically analyze the parameter estimations and propose strategies for model improvement.

Availability and implementation: PyCoTools can be downloaded from the Python Package Index (PyPI) using the command 'pip install pycotools' or directly from GitHub (https://github.com/CiaranWelsh/pycotools). Documentation at http://pycotools.readthedocs.io.

Contact: stefan.przyborski@durham.ac.uk or daryl.shanley@newcastle.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In biology, systems modelling is used to reproduce the dynamics of a biochemical network of molecular interactions with a mathematical model. It has proved particularly useful in the study of cell signalling systems such as NF-κB (Adamson *et al.*, 2016; Ashall *et al.*, 2009; Nelson *et al.*, 2004), mTOR (Dalle Pezze *et al.*, 2012, 2016), p53 (Purvis *et al.*, 2012; Sun *et al.*, 2011) and TGF-β (Schmierer *et al.*, 2008; Vilar *et al.*, 2006; Wang *et al.*, 2014; Zi and Klipp, 2007; Zi *et al.*, 2014). In these studies, the essential biological relationships are represented by a series of ordinary differential equations (ODE) to generate a model. Hypotheses can then be tested by performing *in-silico* experiments. Before ODE models can be used to make meaningful predictions they must first be calibrated to experimental data.

Model calibration is a notoriously difficult problem typically due to the size and complexity of the systems involved and a lack of appropriate experimental data. ODE models are prevalent in systems biology because they are well-suited for predicting system dynamics and because many computational tools have been developed

explicitly for the construction, simulation and analysis of biological networks. Among these tools are Data2Dynamics (Raue *et al.*, 2015), Systems Biology Workbench (Sauro *et al.*, 2003), AMIGO (Balsa-Canto and Banga, 2011), SBpipe (Dalle Pezze and Le Novère, 2017), libRoadRunner (Sauro *et al.*, 2013; Somogyi *et al.*, 2015), Antimony (Smith *et al.*, 2009), Tellurium (Choi *et al.*, 2016), Ecell (Takahashi *et al.*, 2003), PyDsTool (http://www2.gsu.edu/~matrhc/PyDSTool.htm), PySCeS (Olivier, 2005), ABC-SysBio (Liepe *et al.*, 2010), Condor Copasi (Kent *et al.*, 2012) and COPASI (Hoops *et al.*, 2006).

COPASI is a widely used tool in modelling biological systems because it supports a variety of modelling applications including deterministic, stochastic and hybrid model solvers, parameter estimation, optimization, parameter scans, steady state analysis, local sensitivity analysis and metabolic control analysis. COPASI has a graphical user interface (GUI) which makes the tool accessible to non-expert programmers and mathematicians, but also has a command line interface for batch processing and an application programming interface (API) for several programming languages. These APIs have been used for integrating the COPASI framework with custom software, for example in JigCell Run Manager (Palmisano *et al.*, 2015), CellDesigner (Matsuoka *et al.*, 2014), ManyCell (Dada and Mendes, 2012) and ModelMage (Flöttmann *et al.*, 2008).

The Python programming language is useful for scientific computing because of its concise syntax and the availability of open source toolboxes such as pandas (https://pandas.pydata.org/), numpy (http://www.numpy.org/), scipy (http://www.scipy.org/), sklearn (Pedregosa *et al.*, 2011) and matplotlib (Hunter, 2007), which together provide a series of well-documented, easy-to-use, high-level tools for interacting with and manipulating numerical data. Development of further tools in Python is enabled by the Python Package Index (PyPI) where code can be made freely available to other developers. As a result, Python has an extensive publicly available code base for scientific computing that competes well with other commercial and non-commercial environments such as Matlab and R.

Here we present PyCoTools, an open-source Python package which provides a high level interface to COPASI tasks with an emphasis on model calibration. COPASI tasks are integrated with the Python environment to provide additional features which are non-native to COPASI. Features include: the construction of COPASI models with Antimony (Smith *et al.*, 2009); the automation of repeat parameter estimation configurations, chaser parameter estimations and parameter estimations for multiple models (e.g. model selection); automation of the profile likelihood method of identifiability analysis (Raue *et al.*, 2013; Schaber, 2012) with visualization facilities which are flexible enough to support model reduction (Maiwald *et al.*, 2016); visualization of time courses from ensembles of parameter sets and multiple ways of visualizing parameter estimation data. We demonstrate PyCoTools by defining a model selection problem to introduce a known negative feedback into a previously published model of TGF-β signalling (Zi and Klipp, 2007) using new data.

2 Materials and methods

2.1 Experimental

2.1.1 Cell lines and treatment

Neonatal human dermal fibroblasts (HDFn, Life Technologies, C-004-5C) were cultured as per manufacturer guidelines in M106 (Life Technologies M-106-500) supplemented with LSGS (Life Technologies S-003-10). HDFn were seeded at a density of 10 000 cells/cm^2 into 12 well plates (Greiner 665180) in 4 ml complete M106 and cultured for 3 days. Media was aspirated, cells washed twice with DPBS and replaced with 4 ml M106 without LSGS and cells were serum starved for 24 h. HDFn were treated with 5 ng ml^{-1} TGF-β1 (Life Technologies, PHG9211) in M106 media without LSGS for 0, 1, 2, 4, 8, 12 h. To harvest, media was aspirated, cells were washed twice in DPBS and then lysed in 350 µl RLT buffer (Qiagen 79216).

2.1.2 High-throughput qPCR

Lysates were snap frozen in liquid nitrogen and stored -80°C prior to quantification. Cell lystes were thawed at 4°C and then RNA was isolated using the Biomek FxP and the RNAdvance Tissue Isolation kit (Beckman Coulter, p/n A32646). The resulting RNA was quantified using the Nandrop 8000 (Nanodrop, ND-8000). cDNA was generated using 500 ng of TotalRNA and Applied Biosystems High Capacity cDNA with Reverse Transcription kit (Applied Biosystems p/n 4368814). cDNA, assays and dilutions of Applied Biosystems Taqman Fast Advanced MasterMix (Applied Biosystems, p/n 4444965) were plated onto a Wafergen MyDesign SmartChip (TakaraBio, p/n 640036) using the Wafergen Nanodispenser. The chip was then loaded into the SmartChip cycler and qPCR performed using the following conditions: hold Stage 50°C for 2 min, 95°C for 10 min, PCR Stage 95°C for 15 s and 60°C for 1 min. After 40 cycles the reaction was stopped and the data was exported for analysis.

Prior to use for fitting, cycle threshold C_T values were normalized using the $2^{-\Delta\Delta C_T}$ method of quantitative PCR normalization to the geometric mean of four reference genes (B2M, PPIA, GAPDH, ACTB) per sample (Livak and Schmittgen, 2001).

2.2 Computational

2.2.1 PyCoTools availability and installation

PyCoTools was developed partially on Windows 7 and partially on Ubuntu 16.04.2 with the Anaconda distribution of Python 2.7 and COPASI version's 4.19.158 and 4.21.166. PyCoTools can be installed with 'pip', Python's native package manager using the command 'pip install pycotools'. PyCoTools can also be downloaded directly from source at https://github.com/CiaranWelsh/pycotools. More detailed instructions on installation and PyCoTools usage can be found in the PyCoTools documentation (http://pycotools.readthedocs.io).

2.2.2 Definition of the model selection problem

All models were built by downloading the Zi and Klipp (2007) model from BioModels (ID: BIOMD0000000163) and modifying it as appropriate using the COPASI user interface for each model. The models are available in the supplementary content as SBML files. Model selection was performed by calibrating each model to the same experimental data and then evaluating model selection criteria. The Ski mRNA and Smad7 mRNA profiles were measured whilst protein level data were derived by assuming that Smad7 and Ski protein appear 30 min after the mRNA and at 100 times the magnitude. Since the experimental data units are arbitrary and the Zi and Klipp (2007) model simulates in nanomoles per litre, the experimental data were mapped to the model via an observation function (Equation 1).

$$X_{Obs_{(t)}} = \frac{X_{(t)}}{X_{SF}} \tag{1}$$

where:

$X_{Obs_{(t)}}$ = A mapping between experimental and simulated data

$X_{(t)}$ = Amount of model species X at time t

X_{SF} = Scale factor for species X = 100

$X \in \{\text{Smad7mRNA}, \text{SkimRNA}, \text{Smad7Protein}, \text{SkiProtein}\}$

All scale factors were set to 100 which is a reasonable value to ensure new profiles were of the same order of magnitude as the

original. The initial concentration of Smad7 and Ski protein were set to 100 times that of the corresponding mRNA and all new kinetic parameters were estimated. All parameters from the original Zi *et al.* (2014) model were fixed at the published values, including initial concentration parameters. Initial concentrations of Smad7 mRNA and Ski mRNA were set using Equation 2:

$$X_{(t0)} = X_{(\mu,t0)} \cdot X_{SF} \tag{2}$$

where:

$X_{(t0)}$ = Initial amount of species X in the model

$X_{(\mu,t0)}$ = Empirical average of species X at t = 0 in arbitrary units

All parameters were estimated between the boundaries of $1e^{-7}$ and $1e^{4}$. Three hundred parameter estimations were performed per model using COPASI's stochastic genetic algorithm with a population size of 300 over 500 generations and starting from random values. The residual sum of squares (RSS) objective function was weighted using the standard deviation of the 6 data replicates. All parameter estimations were configured and run simultaneously using PyCoTools 'tasks.MultiModelFit' class on a computer cluster running the Sun-Grid Engine job scheduling software. The estimations can optionally be configured to run on a single machine.

2.2.3 An idealized model selection problem

In addition to the main model selection demonstration, another idealized model selection demonstration has been provided in the supplementary content. The purpose of this alternative demonstration is to provide an example with short execution times that parallels the main model selection problem and provides code that users can run themselves. Specifically, in this alternative model selection problem we create three models (a negative feedback motif, a positive feedback motif and a feed-forward motif) using the Antimony interface. Analogous to the main problem defined above, we then perform model selection using synthetic experimental data from the negative feedback topology, visualize the results and run an identifiability analysis.

3 Results

3.1 Overview of PyCoTools facilities and architecture

PyCoTools provides COPASI users with a means of efficiently configuring and running COPASI tasks from a Python environment. The PyCoTools package is comprised of three main modules: 'model', 'tasks' and 'viz'.

The 'Model' object under the 'model' module plays a central role in PyCoTools by using Python's 'lxml' library to extract model information from the COPASI XML and store it in Python classes. Manipulating XML was chosen because of its widespread use in systems biology and because well documented tools exist for its manipulation. The information extracted is subsequently available as 'Model' attributes. The 'Model' enables users to add, remove and change model components and acts as a central entity that can be modified and configured by other PyCoTools classes. As an alternative means of building models, the 'model' module provides an interface to and from the SBML model definition language, Antimony (Smith *et al.*, 2009). PyCoTools wraps functions from Tellurium (Choi *et al.*, 2016) and command line COPASI to convert between Antimony, SBML and COPASI models, thereby facilitating the transition between environments.

The 'tasks' module uses the 'Model' class extensively to configure COPASI tasks. Supported tasks include deterministic, stochastic or hybrid time courses, arbitrary dimensional parameter scans or repeat tasks, and parameter estimations. Additionally, tasks are provided which are not available in COPASI within a single function. Specifically, PyCoTools automates the configuration of 'repeat parameter estimations' and increases the rate by which parameter estimations can be run. This is achieved by automatically configuring COPASI's repeat parameter estimation feature and running model replicates simultaneously. A queueing system is introduced to prevent overuse of limited computational resources. PyCoTools supports the configuration and running of 'chaser estimations' where parameter estimates from a global algorithm are inserted into the model and driven to a minimum with a local algorithm. Other tasks supported by PyCoTools include model selection and the calculation of profile likelihoods for assessing a identifiability status of a model (Raue *et al.*, 2009; Schaber, 2012).

The 'viz' module [the concept of which takes inspiration from the Ecell software by Takahashi *et al.* (2003)] contains all PyCoTools visualization facilities. The aim of the 'viz' module is to produce publication quality figures of time courses, parameter estimations, profile likelihoods and model selection. The 'viz' module also provides a host of exploratory data analysis tools for analyzing repeat parameter estimation data. These tools and their usage are described next.

3.1.1 Tools for analysis of repeat parameter estimation data

Repeat parameter estimation data can be visualized in multiple ways and this information can be used to diagnose problems and direct modelling efforts. The tools provided in PyCoTools collectively allow one to gauge uncertainty in model predictions or parameter estimates, assess the performance of algorithms used for optimization, visualize distributions of parameters and visualize putative relationships between parameters.

Usually the first item of interest after a parameter estimation is to visualize simulated predictions against empirical data. PyCoTools extends the basic 'simulated versus experimental time course plot' to calculate and display confidence intervals for each profile. This is achieved by inserting parameter sets into the model in turn, simulating a time course and aggregating the results by bootstrapping an estimator (e.g. the mean) of the users choice. By visualizing predictions from several parameter sets, uncertainty is propagated from parameter estimates to model predictions. The 'ensemble time course' thus emphasizes model strengths and weaknesses, highlighting regions of confidence and those which require attention.

While ensemble time courses are used to inform our confidence on model predictions, profile likelihoods are used to inform our confidence on parameter values. Briefly, a profile likelihood is a parameter scan of parameter estimations, starting from a best parameter set. Each parameter is fixed in turn and its value is systematically varied over the course of the scan. The remaining parameters are re-optimized at each point of the scan and the objective function value traces a path through parameter space. The shape of this profile is then compared to a confidence threshold based on the likelihood ratio statistic (Raue *et al.*, 2009).

A profile likelihood typically has one of three interpretations. If the profile does not exceed the threshold in one or both directions and is not flat, the parameter is practically non-identifiable. In this case, the trajectory of the other model components over the profile may be used to direct model reduction strategies (Maiwald *et al.*, 2016). If a profile is completely flat the parameter is structurally non-identifiable, which means the parameter is algebraically related to another. To resolve structural non-identifiabilities, one can fix

one of the parameters in a relationship to an arbitrary value. Of note, one must be cautious about using profile likelihoods to render a parameter structurally non-identifiable because the profile likelihood method only samples the parameter space. It is possible that the profile appears flat but only on the scale of the sampled profile. Therefore, structurally non-identifiable parameters should be further investigated to determine any relationships which might exist. Finally, if the profile exceeds this threshold in both directions the parameter is identifiable and the parameter values at which the profile exceeds the threshold are the upper and lower confidence boundaries for the parameter (Raue *et al.*, 2009). Ideally, for precise model predictions, every estimated parameter in a defined parameter estimation problem should be identifiable. In reality, limited data and overly complex model structures often lead to identifiability issues.

Maiwald *et al.* (2016) extended the usefulness of profile likelihood from assessing identifiability to model reduction. A practical non-identifiability exists because the optimization does not have enough data to inform model parameters, or put another way, the model is too complex for the data. Viewing the paths traced by other parameters in a profile likelihood analysis (e.g. putting the trajectory of another parameter on the y-axis rather than the objective function value) provides information about the relationship between the parameter of interest on the x-axis and the parameter on the y-axis. Identifying this relationship enables steps to be taken to resolve the problem by fixing parameters or replacing non-identifiable species or parameters with algebraic equations. Profile likelihoods are therefore useful in a data-driven approach to iteratively refine an optimization problem, fixing parameters where possible and modifying the topology as necessary until the model fits the experimental data.

Profile likelihood calculations are a computationally intense task and to be useful, it is required that the starting parameter set is optimal, or at least very close to optimal, with respect to the data. It is therefore prudent to assess this condition before conducting a profile likelihood analysis. The performance of an optimization problem can be evaluated by plotting the sorted objective function value [i.e. residual sum of squares (RSS) or likelihood] for each parameter estimation iteration against its rank of best fit (herein referred to as a 'likelihood-ranks' plot). In these plots the best case scenario is either a flat line for when there is only a single global minimum or more commonly, a monotonically increasing step-like function where each step marks a different minimum (Raue *et al.*, 2013). Horizontal lines in the likelihood-ranks plot indicate that many iterations of the same optimization problem have located the same minimum, which increases our confidence that the problem is well-posed. In contrast a smooth curve indicates that estimations have not converged to a minimum.

If the likelihood-ranks plot shows a smooth curve, it is a good idea to either rerun the parameter estimation using a different algorithm or different algorithm settings. Alternatively, while others (Raue *et al.*, 2013) employ a multi-start Latin-hypercube strategy with a local optimizer to ensure strategic and uniform sampling of the parameter space, given the choice of algorithms in COPASI it is easy to first run a global and then switch to a local algorithm. This strategy, here referred to as a 'chaser estimation', can be performed on all or a subset parameter sets to drive them closer to their respective minima.

In addition to profile likelihoods and time course ensembles, viewing distributions of parameter estimation data and correlations between parameters can provide information about an optimization problem. Box plots provide immediate information about the range of parameter estimates and how they compare to other parameters. Often a box plot can provide clues to a parameter's identifiability status. Histograms on the other hand provide a more detailed view of parameter distributions and can identify behaviour (e.g. bimodal parameters) that would not be identified with box plots. Moreover, a combination of Pearson's correlation heat maps and scatter graphs can be used to locate linear or log-linear relationships between parameters.

An important aspect of visualizing parameter estimation data is that not all parameter sets fit the model equally well. Parameter sets with higher objective function values can distort the distribution of better performing parameter sets or the shape of a relationship. For this reason PyCoTools implements flexible means of subsetting parameter estimation data before plotting.

3.2 A demonstration: extending the Zi and Klipp (2007) model

To demonstrate PyCoTools, we define a model selection problem to extend a published model of canonical TGF-β signalling (Zi and Klipp, 2007) (Fig. 1). As an alternative demonstration, we also provide an another model selection problem in the supplementary content, as described in the methods.

TGF-β binds to the autophosphorylated homodimeric type 2 TGF-β receptors which phosphorylate and heterodimerize with homodimers of type 1 TGF-β receptors (De Crescenzo *et al.*, 2001). This event leads to internalization of the ligand–receptor complex into one of two types of membrane bound intracellular compartment: early endosomes or caveolae. Evidence in Di Guglielmo *et al.* (2003) suggests that ligand–receptor complexes in the early endosome, rather than the caveolae, are responsible for conveying the TGF-β signal, via phosphorylation, to the Smad second messenger system. Phosphorylated Smad2/3 binds to Smad4, translocates to the nucleus and induces transcription of TGF-β responsive genes (Schmierer *et al.*, 2008). Smad7 is a well characterized negative regulator of the Smad system and is transiently produced in response to TGF-β (Hayashi *et al.*, 1997; Nakao *et al.*, 1997). Multiple mechanisms of negative regulation by Smad7 have been reported, including the recruitment of E3 ubiquitin ligases to either Smad2/3 in competition with Smad4 (Yan *et al.*, 2016) or to activated TGF-β receptors in caveolae (Di Guglielmo *et al.*, 2003; Kavsak *et al.*, 2000). Many biological entities have been proposed as regulators of this process, including PPM1A (Lin *et al.*, 2006), NEDD4L (Gao *et al.*, 2009), SNoN (Stroschein *et al.*, 1999) and Ski. Ski acts as co-repressor at Smad regulated genes by recruiting histone deacetylases which leads to epigenetic constriction of Smad-responsive genes (Akiyoshi *et al.*, 1999).

The Zi and Klipp (2007) model (Fig. 1a) combines work by Vilar *et al.* (2006) describing TGF-β receptor internalization and recycling dynamics with a Smad nuclear-cytoplasmic translocation module. In this model, an explicit representation of the Smad7 negative feedback was not included, but was instead incorporated into the rate law for the reaction describing the degradation of the activated ligand–receptor complexes from within caveolar compartments ('LRC_Cave' in Fig. 1a). The purpose of the model selection problem presented here is to investigate the feasibility of three alternative mechanisms of negative regulation (Fig. 1) in explaining the experimental data (Fig. 2).

After calibration, the 'viz.ModelSelection' class was used to calculate and visualize the Akaike information criteria (AIC) corrected for small sample sizes (AICc) (Fig. 3a) and the Bayesian information criteria (BIC) (Supplementary Fig. S1). With these statistics, a lower value indicates a better agreement with the data and thus a better

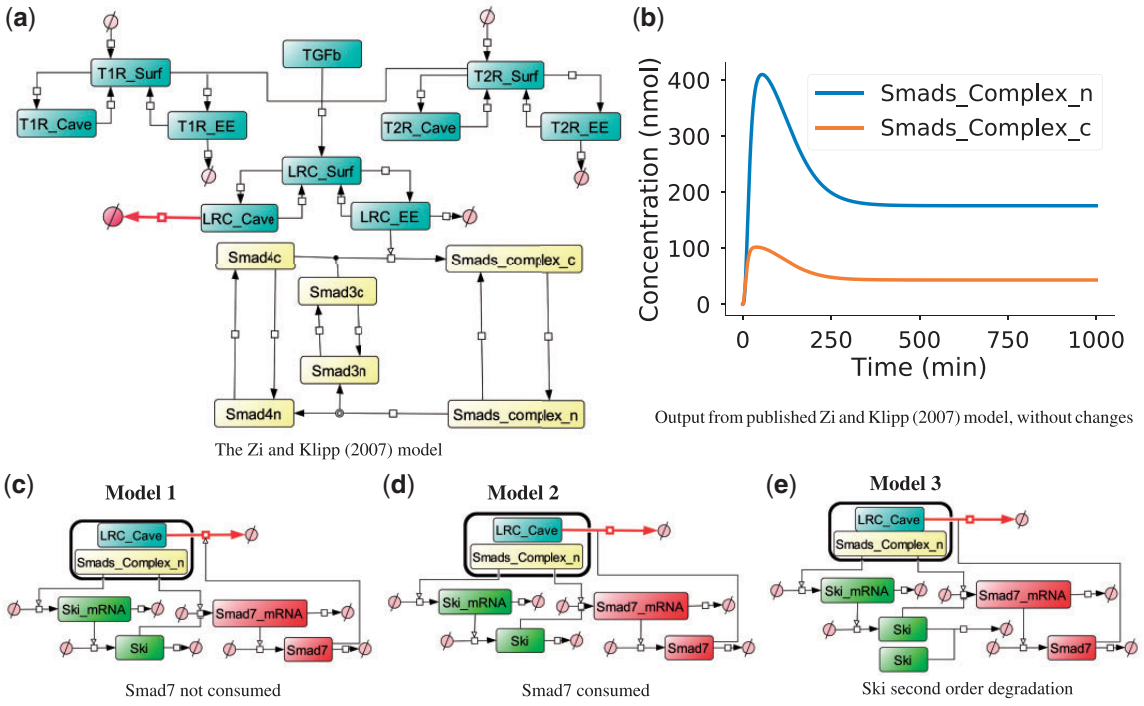


Fig. 1. Network representation of ODE networks used in model selection problem. (**a**) The Zi and Klipp (2007) model is a common component of each model variant. (**b**) Simulation output from the Zi and Klipp (2007). (**c–e**) The model variable 'Smads_Complex_n' is responsible for transcription reactions in model variants while 'LRC_Cave' is degraded by Smad7 protein, thus completing the explicit representation of the Smad7 negative feedback loop. In (c) Model 1, Smad7 participates in but is not consumed by the reaction with LRC_Cave while in (d) Model 2, Smad7 is consumed by this process. In (e) Model 3, the same topology as Model 2 is assumed but it also incorporates second order mass action degradation kinetics for Ski protein

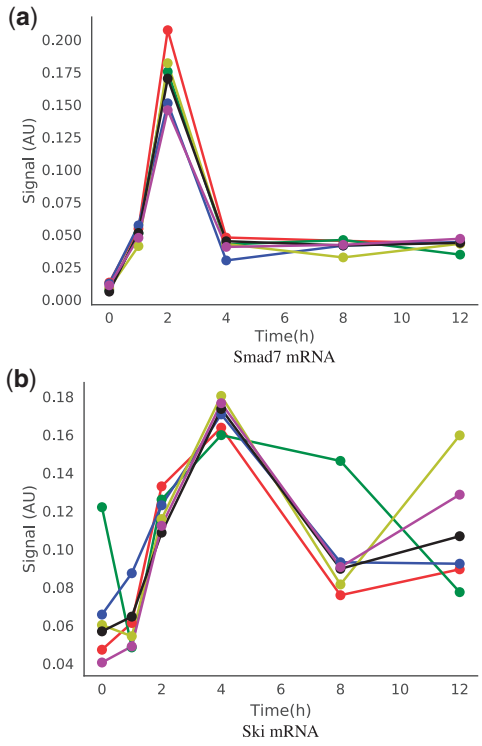
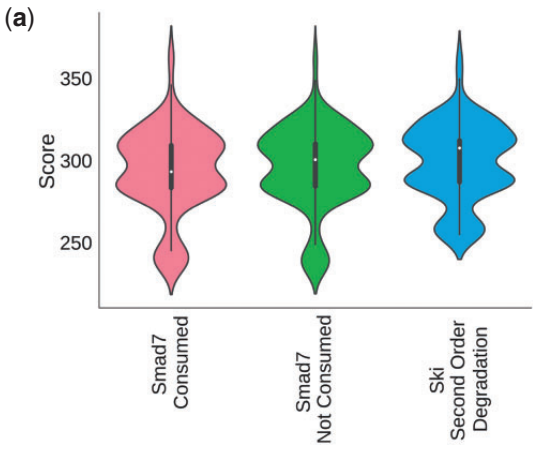


Fig. 2. Experimental data used for model calibration. Neonatal human dermal fibroblasts were treated with 5 ng ml^{-1} TGF-β for 0, 1, 2, 4, 8 and 12 h. Shown are profiles of 6 biological replicates for (**a**) Smad7 and (**b**) Ski messenger RNA, measured by high throughput quantitative PCR as described in the methods



(**b**)

Model Selection Criteria	Model 1	Model 2	Model 3
RSS	614.037	615.003	698.592
AICc	234.751	234.996	254.877
BIC	264.25	264.495	284.376

Model selection criteria for best ranked parameter sets for each model.

Fig. 3. Model selection criteria. (**a**) Distribution of Akaike information criteria (AICc) per model displayed as violin plot. The central white dot represents the median; the thin centre line is the 95% confidence interval; the thick central bar is the interquartile range and the width represents the frequency with which a score was observed. These graphs were produced with 'viz.ModelSelection'. (**b**) A comparison of model selection criteria for the best ranking parameter sets in each model

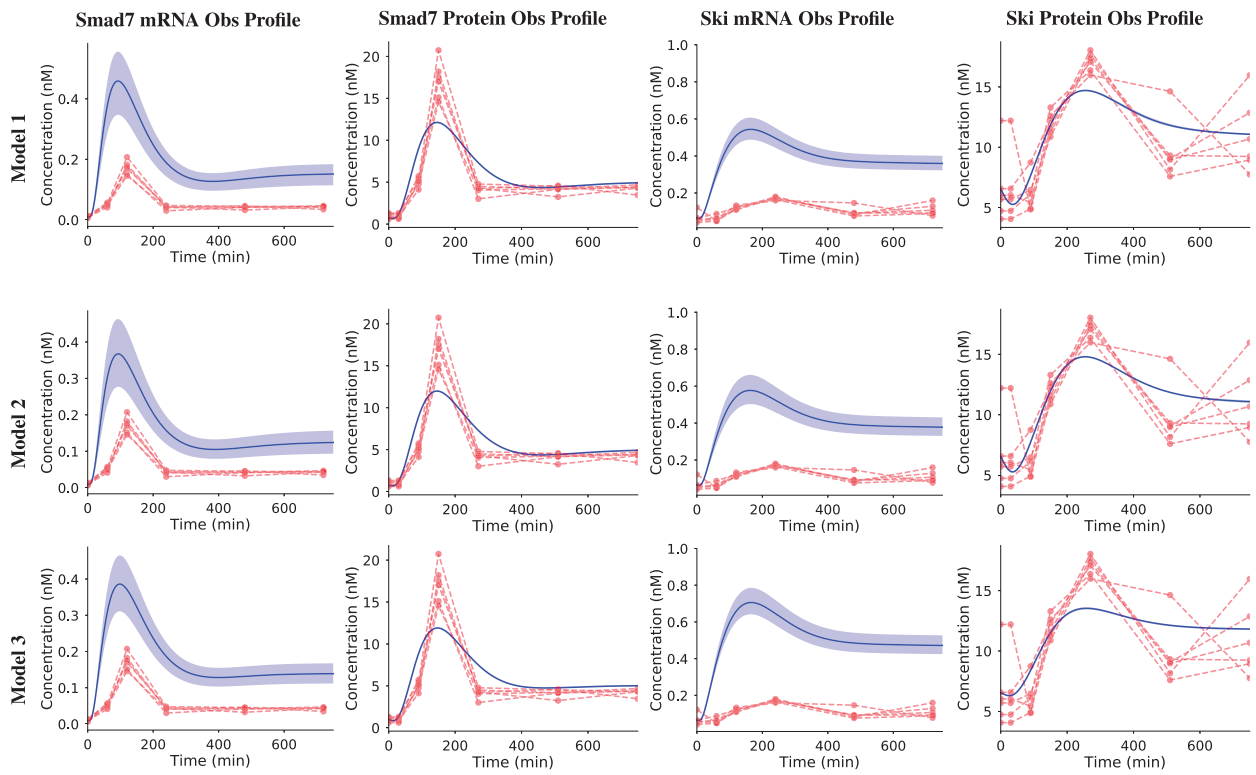


Fig. 4. Ensemble time courses produced with 'viz.PlotTimeCourseEnsemble'. The top 10 best parameter sets for each model were sequentially inserted into their respective models. Time courses were simulated with each parameter set and averaged. Red profiles indicate experimental data while solid blue lines are simulated profiles. Shaded areas represent 95% confidence intervals

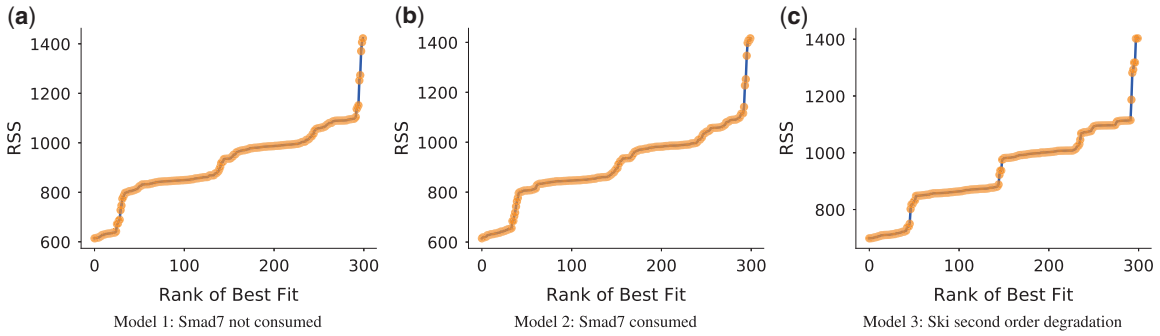


Fig. 5. A 'likelihood-ranks' plot. The residual sum of squares objective function value is plotted against the rank of best fit for each parameter estimation iteration for each model (**a–c**). Graphs were produced with 'viz.LikelihoodRanks'

model. In the current problem, a closer inspection of the best model selection values (Fig. 3b) indicates that from a purely statistical perspective, the topologies of Models 1 and 2 are indistinguishable in terms of the experimental data (Fig. 2) while Model 3 is worse.

The simulated profiles for each model (Fig. 4) supports the model selection results. While the Smad7 mRNA and Ski mRNA profiles are slightly greater in Model 1 and Model 3 respectively, all profiles are virtually indistinguishable between all the models. It is likely that the difference in the Ski mRNA profile in Model 3 accounts for the difference observed in the best model selection criteria (Fig. 3b). Regardless of this slight difference, the same qualitative interpretation holds for each model: the speed and magnitude of both Smad7 and Ski mRNA induction profiles are overestimated while the protein level data fits each model to a high degree of confidence.

When looking at model predictions it is important to consider whether the parameter sets used to produce them are actually the best parameter sets. This is important because it is quite common for parameter estimation algorithms to find sub-optimal parameters. Here, while improvements can still be made, the algorithm and settings were reasonably well-chosen because the likelihood-ranks plot produced a step-like shape for each model (Fig. 5), heuristically mapping out where the local and global minima are.

Profile likelihoods are only meaningful when calculated from a minimum with respect to the data. For this reason the best three parameter sets from the stochastic genetic algorithm in Model 2 were 'chased' with a Hooke & Jeeves algorithm (tolerance $= 1e^{-10}$ and iteration limit $= 1000$) using the 'PyCoTools.tasks.ChaserParameterEstimations' class. Profile likelihoods were then computed

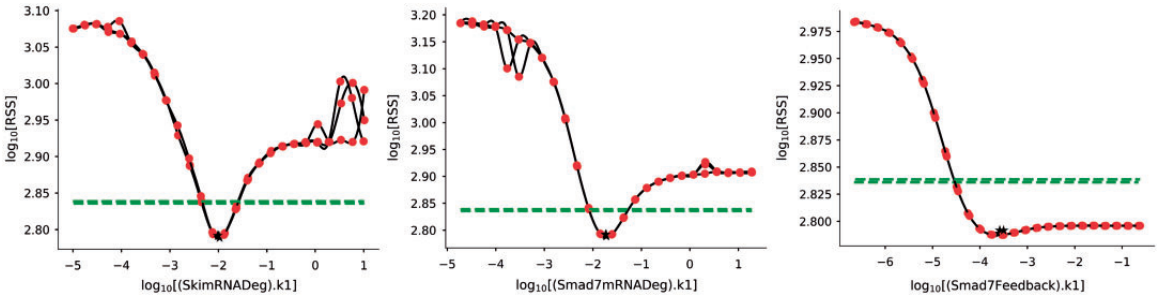


Fig. 6. Profile likelihoods were calculated using the 'tasks.ProfileLikelihood' class for the top three parameter sets of Model 2 and visualized using 'viz.PlotProfileLikelihood'. The black stars indicate the best estimated parameter. The dotted green line indicates the 95% confidence level and the red spots are the minimum RSS value achieved after re-optimization of all parameters except the parameter of interest (x-axis). Lines between red spots have been interpolated using a cubic spline

around these three parameter sets, again using the Hooke & Jeeves algorithm (tolerance = $1e^{-6}$ and iteration limit = 50). Sampling was conducted on a log10 scale over 6 orders of magnitude, $1e^3$ times above and below the best estimated parameter values. For brevity, profile likelihoods for Models 1 and 3 are not discussed. The identifiability analysis shows that seven of the ten parameters are identifiable and the remaining three are practically non-identifiable (Fig. 6 and Supplementary Fig. S2).

To investigate the source of these non-identifiabilities, two strategies were employed: Pearson's correlation analysis and the 'profile likelihood model reduction' approach as described in Maiwald *et al.* (2016). The Pearson's correlation approach identified several parameter pairs as putative linear correlations (Supplementary Fig. S3). Of these, only the most correlated pair, the k_m and I_{50} parameters of Smad7 transcription, was verified to be log-linearly related in both scatter graphs (Fig. 7a) and profile likelihood traces (Fig. 7b). To resolve this issue, one could replace one of the free parameters in the relationship with the algebraic equation resulting from the fit of a linear model to the profile likelihood trace (Fig. 7b). The other putative relationships suggested by the Pearson's correlation analysis (Supplementary Fig. S3) were also investigated but the relationships were more difficult to interpret. As an example, Supplementary Figure S4 shows the relationship between '(SkiDeg).k1' and '(SkimRNADeg).k1' parameters. While the scatter graph shows a reasonable linear correlation (Supplementary Fig. S4a), it is defined on a very small interval and the profile likelihood is clearly non-linear, albeit linear on a sub-domain of the parameter space (Supplementary Fig. S4b).

Lastly, distributions of parameter estimates were visualized using box plots (Supplementary Fig. S5) and histograms (Supplementary Fig. S6). Despite being presented last, these are computationally inexpensive to generate and are good to view prior to more involved analyses such as profile likelihoods. To demonstrate the effect of sub-optimal parameter sets, a comparison is made between box plots generated for Model 2 using all parameter estimation data (Supplementary Fig. S5a) to those using only the top 10% ranking parameter sets (Supplementary Fig. S5b). Supplementary Figure S5 demonstrates that suboptimal parameter sets can distort the insight that can be gained from visually exploring parameter estimation data. Without truncating the parameter estimation data, the observation that the distributions of parameters from the best parameter sets reflect the identifiability status of the model, would be missed.

4 Discussion

PyCoTools is an open source Python package designed to assist COPASI users in the task of modelling biological systems.

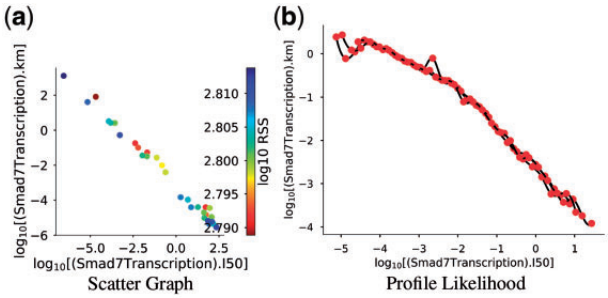


Fig. 7. Identification of a log-linear relationship between '(Smad7Transcription).km' and '(Smad7Transcription).I50' (k_m and I_{50}, respectively). (**a**) Scatter graph showing that as k_m increases, I_{50} decreases ($r^2 = 0.995$, *P*-value = $1e^{-39}$). (**b**) The path traced by k_m is plotted as a function of I_{50} during the profile likelihood calculation. Graphs were produced using 'viz.Scatters' and 'viz.PlotProfileLikelihood' respectively

PyCoTools offers an alternative high level interface to COPASI tasks including time courses, parameter scans and parameter estimations. While COPASI implements the heavy computation, PyCoTools automates task configuration and execution, thereby promoting efficiency, organization and reproducibility.

PyCoTools bridges COPASI with the Python environment allowing users to take advantage of Python's numerical computation, visualization, file management and code development facilities. One tool in particular, the Jupyter notebook, allows annotation of code blocks with rich text elements and is a powerful environment from which to develop and share annotated workflows. The combination of Jupyter notebooks, COPASI and PyCoTools therefore enables the production of reproducible and shareable models that are annotated with justifications.

PyCoTools supports model editing using both an object-oriented approach and with Antimony, a model specification language for building SBML models (Smith *et al.*, 2009). The Antimony and COPASI user interface are complementary and can be used together to enhance the modelling process. For example, models in Antimony format can be used as a 'hard copy' while a parallel COPASI model can be used for exploratory changes that are 'committed' to the hard copy when satisfactory.

PyCoTools supports the configuration of 'composite' tasks which are those comprised of a combination of other tasks. These tasks can be configured using the COPASI user interface but generally take time and are vulnerable to human error. For example, users can automatically configure repeat parameter estimations, chaser parameter estimations and model selection problems, thereby circumventing the requirement for manual configuration.

Another composite task supported by PyCoTools is the profile likelihood method of identifiability analysis (Raue *et al.*, 2009). Models with non-identifiable parameters are common in systems biology and it is useful to have a means of assessing which parameters are reliably defined by an estimation problem. PyCoTools automates the procedure outlined by Schaber (2012) for conducting profile likelihoods in COPASI, thereby enabling COPASI users to perform an identifiability analysis more efficiently and in a way less amenable to errors than manual configuration. PyCoTools also enables users to calculate profile likelihoods from multiple parameter sets thereby enabling users to address one of the shortcomings of the profile likelihood approach: that it is a local method of identifiability analysis.

One alternative to COPASI and PyCoTools is Data2Dynamics (Raue *et al.*, 2015). While Data2Dynamics provides an excellent range of model analysis tools, the transfer of files between COPASI and Data2Dynamics is imperfect, often necessitating that a COPASI user redefine their model within the Data2Dynamics environment. PyCoTools allows COPASI users to stay within the COPASI environment, thereby making profile likelihood analysis more accessible to COPASI users.

In this work we have demonstrated PyCoTools by posing a model selection problem to discriminate between three model topologies (Fig. 1) with respect to some experimental data in response to TGF-β (Fig. 2). Rather than using synthetic data, our aim was to demonstrate in a 'real world' scenario how PyCoTools can be used together with COPASI to calibrate a set of models and discriminate between them.

As this is primarily a software demonstration and not a biological investigation, the model selection problem proposed was designed to be as simple as possible whilst still being non-trivial. Mechanistically the three models (Fig. 1) are alternative hypotheses which attempt to address the dynamics of the Smad7 (Fig. 2) negative feedback. Model alternatives were based on a published dynamic model of TGF-β signalling (Zi and Klipp, 2007) that was adapted to incorporate Smad7. Since the decay of Smad7 is transient and fast (Fig. 2a), the simplest mechanism involving only Smad7 with first order mass action degradation kinetics would not be able to account for the observed decline in Smad7. Therefore Smad7 degradation was assumed to be an active process. Since Ski is a known Smad corepressor (Akiyoshi *et al.*, 1999) and Smad7 is a Smad responsive gene (Hayashi *et al.*, 1997), Ski was proposed to be transcribed in response to TGF-β (Fig. 2b) and inhibit Smad7 transcription. The model alternatives are slightly different representations of this hypothesis (Fig. 1).

In this model selection problem it is clear that the model topologies chosen are too similar to be discriminated with the experimental data and therefore the models are virtually indistinguishable (Fig. 4). Generally, with model selection, the strongest statement that can be made about a model is a rejection, since accepting the hypothesis does not necessarily guarantee that it is correct. By comparing the performance of multiple models in model calibration it is possible reject one or more topologies in favour of another. Here, however, because the models are so similar, it was not possible to provide support for any model being worse than any other, despite the minor differences in model selection criteria for Model 3 (Fig. 3a). In a more comprehensive investigation many more topologies would be similarly compared to iteratively reject topologies until the model is capable of making useful, validatable predictions.

Regardless of the biological interpretation, we have demonstrated the process of using PyCoTools and COPASI to discriminate between model alternatives and to critically assess the parameter estimation process. Model calibration is an essential part of a systems modelling investigations, but it is often limited by a vast, underdetermined parameter space and therefore, procedures that provide a measure of uncertainty are valuable. In PyCoTools, we have implemented a number of features aimed towards gauging confidence and uncertainty in the optimization process so that COPASI users can diagnose problems and make better informed decisions based on their parameter estimation output. These tools include: the likelihood-ranks plot (Fig. 5) which enables evaluation of an optimization algorithm and settings on a specific problem (Raue *et al.*, 2013); ensemble time courses (Fig. 4) which calculate confidence intervals from predictions made from multiple best parameter sets and propagates uncertainty from parameter estimates to model predictions; profile likelihoods for assessing identifiability (Fig. 6, Supplementary Fig. S2) and for model reduction (Fig. 7b) (Maiwald *et al.*, 2016); Pearson's correlation heat maps (Supplementary Fig. S3) and scatter graphs (Fig. 7a) for identifying relationships, and box plots (Supplementary Fig. S5) and histograms (Supplementary Fig. S6) for visualizing distributions of parameter estimates. Together these tools provide detailed information about an optimization problem that can be used to guide the modelling process.

5 Conclusion

PyCoTools is an open-source and extensible Python package designed to facilitate the use of COPASI, particularly for model calibration. PyCoTools supports a range of tools which are either wrappers around COPASI tasks, an ordered workflow of task configurations, or plotting facilities for exploratory data analysis on parameter estimation data. Use of PyCoTools can enhance the effectiveness with which one can calibrate models to experimental data and discriminate between alternate hypotheses.

Funding

This work was funded by Procter & Gamble. The contribution from AGM and CJP was supported by the Medical Research Council (https://www.mrc.ac.uk/) and Arthritis Research UK (http://www.arthritisresearchuk.org/) as part of the MRC-Arthritis Research UK Centre for Integrated research into Musculoskeletal Ageing (CIMA) (MR/K006312/1). The work builds on a BBSRC LINK grant awarded to SAB (BB/K019260/1).

Conflict of Interest: none declared.

References

Adamson,A. *et al.* (2016) Signal transduction controls heterogeneous nf-[kappa] b dynamics and target gene expression through cytokine-specific refractory states. *Nat. Commun.*, **7**, 12057.

Akiyoshi,S. *et al.* (1999) c-ski acts as a transcriptional co-repressor in transforming growth factor-β signaling through interaction with smads. *J. Biol. Chem.*, **274**, 35269–35277.

Ashall,L. *et al.* (2009) Pulsatile stimulation determines timing and specificity of nf-κb-dependent transcription. *Science*, **324**, 242–246.

Balsa-Canto,E. and Banga,J.R. (2011) Amigo, a toolbox for advanced model identification in systems biology using global optimization. *Bioinformatics*, **27**, 2311–2313.

Choi,K. *et al.* (2016) Tellurium: a python based modeling and reproducibility platform for systems biology. *bioRxiv*, 054601.

Dada,J.O. and Mendes,P. (2012) *ManyCell: A Multiscale Simulator for Cellular Systems*. Springer, Berlin, Heidelberg, pp. 366–369.

Dalle Pezze,P. and Le Novère,N. (2017) Sbpipe: a collection of pipelines for automating repetitive simulation and analysis tasks. *BMC Syst. Biol.*, **11**, 46.

Dalle Pezze,P. *et al*. (2012) A dynamic network model of mtor signaling reveals tsc-independent mtorc2 regulation. *Sci. Signal*, **5**, ra25.

Dalle Pezze,P. *et al*. (2016) A systems study reveals concurrent activation of ampk and mtor by amino acids. *Nat. Commun*., **7**, 13254.

De Crescenzo,G. *et al*. (2001) Real-time monitoring of the interactions of transforming growth factor-β (tgf-β) isoforms with latency-associated protein and the ectodomains of the tgf-β type ii and iii receptors reveals different kinetic models and stoichiometries of binding. *J. Biol. Chem*., **276**, 29632–29643.

Di Guglielmo,G.M. *et al*. (2003) Distinct endocytic pathways regulate tgf-β receptor signalling and turnover. *Nat. Cell Biol*., **5**, 410–421.

Flöttmann,M. *et al*. (2008) Modelmage: a tool for automatic model generation, selection and management. In: Jonathan,A. and See-Kiong,N. (eds) *Genome Informatics 2008: Genome Informatics Series Vol. 20*. World Scientific, Imperial College Press, London, pp. 52–63.

Gao,S. *et al*. (2009) Ubiquitin ligase nedd4l targets activated smad2/3 to limit tgf-β signaling. *Mol. Cell*, **36**, 457–468.

Hayashi,H. *et al*. (1997) The mad-related protein smad7 associates with the tgfβ receptor and functions as an antagonist of tgfβ signaling. *Cell*, **89**, 1165–1173.

Hoops,S. *et al*. (2006) Copasi—a complex pathway simulator. *Bioinformatics*, **22**, 3067–3074.

Hunter,J.D. (2007) Matplotlib: a 2d graphics environment. *Comput. Sci. Eng*., **9**, 90–95.

Kavsak,P. *et al*. (2000) Smad7 binds to smurf2 to form an e3 ubiquitin ligase that targets the tgfβ receptor for degradation. *Mol. Cell*, **6**, 1365–1375.

Kent,E. *et al*. (2012) Condor-copasi: high-throughput computing for biochemical networks. *BMC Syst. Biol*., **6**, 91.

Liepe,J. *et al*. (2010) Abc—sysbio—approximate bayesian computation in python with gpu support. *Bioinformatics*, **26**, 1797–1799.

Lin,X. *et al*. (2006) Ppm1a functions as a smad phosphatase to terminate tgfβ signaling. *Cell*, **125**, 915–928.

Livak,K.J. and Schmittgen,T.D. (2001) Analysis of relative gene expression data using real-time quantitative pcr and the 2- $\delta\delta$ct method. *Methods*, **25**, 402–408.

Maiwald,T. *et al*. (2016) Driving the model to its limit: profile likelihood based model reduction. *PLoS One*, **11**, e0162366.

Matsuoka,Y. *et al*. (2014) Modeling and simulation using celldesigner. *Methods Mol. Biol*., **1164**, 121–145.

Nakao,A. *et al*. (1997) Identification of smad7, a tgfβ-inducible antagonist of tgf-β signalling. *Nature*, **389**, 631–635.

Nelson,D. *et al*. (2004) Oscillations in nf-κb signaling control the dynamics of gene expression. *Sci. Signal*, **306**, 704.

Olivier,B.G. (2005) *Simulation and database software for computational systems biology: PySCeS and JWS Online*. PhD Thesis, University of Stellenbosch.

Palmisano,A. *et al*. (2015) Jigcell run manager (jc-rm): a tool for managing large sets of biochemical model parametrizations. *BMC Syst. Biol*., **9**, 95.

Pedregosa,F. *et al*. (2011) Scikit-learn: machine learning in python. *J. Mach. Learn. Res*., **12**, 2825–2830.

Purvis,J.E. *et al*. (2012) p53 dynamics control cell fate. *Science*, **336**, 1440–1444.

Raue,A. *et al*. (2009) Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, **25**, 1923–1929.

Raue,A. *et al*. (2013) Lessons learned from quantitative dynamical modeling in systems biology. *PLoS One*, **8**, e74335.

Raue,A. *et al*. (2015) Data2dynamics: a modeling environment tailored to parameter estimation in dynamical systems. *Bioinformatics*, **31**, 3558–3560.

Sauro,H.M. *et al*. (2003) Next generation simulation tools: the systems biology workbench and biospice integration. *Omics J. Integrative Biol*., **7**, 355–372.

Sauro,H.M. *et al*. (2013) libroadrunner: a high performance SBML compliant simulator. *Bioinformatics*, **31**, 3315–3321.

Schaber,J. (2012) Easy parameter identifiability analysis with copasi. *Biosystems*, **110**, 183–185.

Schmierer,B. *et al*. (2008) Mathematical modeling identifies smad nucleocytoplasmic shuttling as a dynamic signal-interpreting system. *Proc. Natl. Acad. Sci. USA*, **105**, 6608–6613.

Smith,L.P. *et al*. (2009) Antimony: a modular model definition language. *Bioinformatics*, **25**, 2452–2454.

Somogyi,E.T. *et al*. (2015) libroadrunner: a high performance sbml simulation and analysis library. *Bioinformatics*, **31**, 3315–3321.

Stroschein,S.L. *et al*. (1999) Negative feedback regulation of tgf-β signaling by the snon oncoprotein. *Science*, **286**, 771–774.

Sun,T. *et al*. (2011) Modeling the basal dynamics of p53 system. *PLoS One*, **6**, e27882.

Takahashi,K. *et al*. (2003) E-cell 2: multi-platform e-cell simulation system. *Bioinformatics*, **19**, 1727–1729.

Vilar,J.M.G. *et al*. (2006) Signal processing in the tgf-β superfamily ligand–receptor network. *PLoS Comput. Biol*., **2**, e3.

Wang,J. *et al*. (2014) The self-limiting dynamics of tgf-beta signaling in silico and in vitro, with negative feedback through ppm1a upregulation. *PLoS Comput. Biol*., **10**, e1003573.

Yan,X. *et al*. (2016) Smad7 protein interacts with receptor-regulated smads (r-smads) to inhibit transforming growth factor-β (tgf-β)/smad signaling. *J. Biol. Chem*., **291**, 382–392.

Zi,Z. and Klipp,E. (2007) Constraint-based modeling and kinetic analysis of the smad dependent tgf-β signaling pathway. *PLoS One*, **2**, e936.

Zi,Z. *et al*. (2014) Quantitative analysis of transient and sustained transforming growth factor-β signaling dynamics. *Mol. Syst. Biol*., **7**, 492.

Appendix B

Model Equations

B.1 Base Model Equations

The following set of equations was used as the base model in Chapter 6 and as such are a part of every model within the model selection problem.

$$
\begin{aligned}
V_{\text{Medium}} &= \frac{0.004}{\text{NumberOfCells}} \\
\frac{\mathrm{d}\left([\text{R2}] \cdot V_{\text{Cytoplasm}}\right)}{\mathrm{d}t} &= +V_{\text{Cytoplasm}} \cdot (\text{vR2}) \\
&\quad -V_{\text{Cytoplasm}} \cdot (\text{kcd} \cdot [\text{R2}]) \\
&\quad -(V_{\text{Cytoplasm}} \cdot \text{kaTGFb} \cdot [\text{R2}] \cdot [\text{TGFb}]) \\
&\quad +(V_{\text{Cytoplasm}} \cdot \text{kdTGFb} \cdot [\text{R2\_TGFb}]) \\
\frac{\mathrm{d}\left([\text{R2\_TGFb}] \cdot V_{\text{Cytoplasm}}\right)}{\mathrm{d}t} &= +(V_{\text{Cytoplasm}} \cdot \text{kaTGFb} \cdot [\text{R2}] \cdot [\text{TGFb}]) \\
&\quad -(V_{\text{Cytoplasm}} \cdot \text{kdTGFb} \cdot [\text{R2\_TGFb}]) \\
&\quad -V_{\text{Cytoplasm}} \cdot (\text{kaR2\_TGFbBindR1} \cdot [\text{R2\_TGFb}] \cdot [\text{R1}]) \\
&\quad +V_{\text{Cytoplasm}} \cdot (\text{\_kbR2\_TGFbUnBindR1} \cdot [\text{LRC}])
\end{aligned}
$$

$$\frac{\mathrm{d}\left([\mathrm{R1}]\cdot V_{\mathrm{Cytoplasm}}\right)}{\mathrm{d}t} = +V_{\mathrm{Cytoplasm}}\cdot(\mathrm{vR1})$$
$$-V_{\mathrm{Cytoplasm}}\cdot(\mathrm{kcd}\cdot[\mathrm{R1}])$$
$$-V_{\mathrm{Cytoplasm}}\cdot(\mathrm{kaR2\_TGFbBindR1}\cdot[\mathrm{R2\_TGFb}]\cdot[\mathrm{R1}])$$
$$+V_{\mathrm{Cytoplasm}}\cdot(\mathrm{\_kbR2\_TGFbUnBindR1}\cdot[\mathrm{LRC}])$$

$$\frac{\mathrm{d}\left([\mathrm{Smad3}]\cdot V_{\mathrm{Cytoplasm}}\right)}{\mathrm{d}t} = -V_{\mathrm{Cytoplasm}}\cdot(\mathrm{\_kSmad3Phos}\cdot[\mathrm{Smad3}]\cdot[\mathrm{LRC}])$$
$$-(V_{\mathrm{Cytoplasm}}\cdot\mathrm{kimp\_Smad3}\cdot[\mathrm{Smad3}])$$
$$+(V_{\mathrm{Nucleus}}\cdot\mathrm{kexp\_Smad3n}\cdot[\mathrm{Smad3n}])$$
$$+V_{\mathrm{Cytoplasm}}\cdot(\mathrm{kSmad3\_prod})$$
$$-V_{\mathrm{Cytoplasm}}\cdot(\mathrm{kSmad3\_deg}\cdot[\mathrm{Smad3}])$$

$$\frac{\mathrm{d}\left([\mathrm{LRC}]\cdot V_{\mathrm{Cytoplasm}}\right)}{\mathrm{d}t} = -V_{\mathrm{Cytoplasm}}\cdot(\mathrm{\_kSmad7\_degrade\_LRC}\cdot[\mathrm{Smad7mRNA}]\cdot[\mathrm{LRC}])$$
$$+V_{\mathrm{Cytoplasm}}\cdot(\mathrm{kaR2\_TGFbBindR1}\cdot[\mathrm{R2\_TGFb}]\cdot[\mathrm{R1}])$$
$$-V_{\mathrm{Cytoplasm}}\cdot(\mathrm{\_kbR2\_TGFbUnBindR1}\cdot[\mathrm{LRC}])$$

$$\frac{\mathrm{d}\left([\mathrm{pSmad3}]\cdot V_{\mathrm{Cytoplasm}}\right)}{\mathrm{d}t} = +V_{\mathrm{Cytoplasm}}\cdot(\mathrm{\_kSmad3Phos}\cdot[\mathrm{Smad3}]\cdot[\mathrm{LRC}])$$
$$-2\cdot V_{\mathrm{Cytoplasm}}\cdot(\mathrm{\_kpSmad3\_bind\_Smad4}\cdot[\mathrm{pSmad3}]\cdot[\mathrm{pSmad3}]\cdot[\mathrm{Smad4}])$$

$$\frac{\mathrm{d}\left([\mathrm{Smad4}]\cdot V_{\mathrm{Cytoplasm}}\right)}{\mathrm{d}t} = -V_{\mathrm{Cytoplasm}}\cdot(\mathrm{\_kpSmad3\_bind\_Smad4}\cdot[\mathrm{pSmad3}]\cdot[\mathrm{pSmad3}]\cdot[\mathrm{Smad4}])$$
$$-(V_{\mathrm{Cytoplasm}}\cdot\mathrm{kimp\_Smad4}\cdot[\mathrm{Smad4}])$$
$$+(V_{\mathrm{Nucleus}}\cdot\mathrm{kexp\_Smad4n}\cdot[\mathrm{Smad4n}])$$

$$\frac{\mathrm{d}([\mathrm{Smad4n}]\cdot V_{\mathrm{Nucleus}})}{\mathrm{d}t} = +V_{\mathrm{Nucleus}}\cdot(\mathrm{\_kpSmad3\_Smad4n\_unbind}\cdot[\mathrm{pSmad3\_Smad4n}])$$
$$+(V_{\mathrm{Cytoplasm}}\cdot\mathrm{kimp\_Smad4}\cdot[\mathrm{Smad4}])$$
$$-(V_{\mathrm{Nucleus}}\cdot\mathrm{kexp\_Smad4n}\cdot[\mathrm{Smad4n}])$$

$$\frac{\mathrm{d}([\mathrm{pSmad3n}]\cdot V_{\mathrm{Nucleus}})}{\mathrm{d}t} = +2\cdot V_{\mathrm{Nucleus}}\cdot(\_\mathrm{kpSmad3\_Smad4n\_unbind}\cdot[\mathrm{pSmad3\_Smad4n}])$$
$$-V_{\mathrm{Nucleus}}\cdot(\_\mathrm{kpSmad3\_dephos}\cdot[\mathrm{pSmad3n}])$$

$$\frac{\mathrm{d}([\mathrm{Smad3n}]\cdot V_{\mathrm{Nucleus}})}{\mathrm{d}t} = +V_{\mathrm{Nucleus}}\cdot(\_\mathrm{kpSmad3\_dephos}\cdot[\mathrm{pSmad3n}])$$
$$+(V_{\mathrm{Cytoplasm}}\cdot\mathrm{kimp\_Smad3}\cdot[\mathrm{Smad3}])$$
$$-(V_{\mathrm{Nucleus}}\cdot\mathrm{kexp\_Smad3n}\cdot[\mathrm{Smad3n}])$$

$$\frac{\mathrm{d}([\mathrm{pSmad3\_Smad4n}]\cdot V_{\mathrm{Nucleus}})}{\mathrm{d}t} = -V_{\mathrm{Nucleus}}\cdot(\_\mathrm{kpSmad3\_Smad4n\_unbind}\cdot[\mathrm{pSmad3\_Smad4n}])$$
$$+(V_{\mathrm{Cytoplasm}}\cdot\mathrm{kimp\_pSmad3\_Smad4}\cdot[\mathrm{pSmad3\_Smad4}])$$

$$\frac{\mathrm{d}\left([\mathrm{Smad7mRNA}]\cdot V_{\mathrm{Cytoplasm}}\right)}{\mathrm{d}t} = -V_{\mathrm{Cytoplasm}}\cdot(\_\mathrm{kSmad7mRNA\_deg}\cdot[\mathrm{Smad7mRNA}])$$
$$+V_{\mathrm{Nucleus}}\cdot\_\mathrm{kBasalSmad7mRNA}+V_{\mathrm{Nucleus}}\cdot\_\mathrm{kSmad7mRNA\_prod}\cdot[\mathrm{pSmad3\_Smad4n}]$$
$$-V_{\mathrm{Cytoplasm}}\cdot\_\mathrm{kSmad7\_degrade\_LRC}\cdot[\mathrm{Smad7mRNA}]\cdot[\mathrm{LRC}]$$

$$\frac{\mathrm{d}\left([\mathrm{CTGFmRNA}]\cdot V_{\mathrm{Cytoplasm}}\right)}{\mathrm{d}t} = -V_{\mathrm{Cytoplasm}}\cdot(\_\mathrm{kCTGFmRNA\_deg}\cdot[\mathrm{CTGFmRNA}])$$
$$+V\mathrm{Nucleus}\cdot\_\mathrm{kBasalCTGFmRNA}+V_{\mathrm{Nucleus}}\cdot\_\mathrm{kCTGFmRNA\_prod}\cdot[\mathrm{pSmad3\_Smad4n}]$$

$$\frac{\mathrm{d}\left([\mathrm{COL1A2mRNA}]\cdot V_{\mathrm{Cytoplasm}}\right)}{\mathrm{d}t} = -V_{\mathrm{Cytoplasm}}\cdot(\mathrm{kCOL1A2mRNA\_deg}\cdot[\mathrm{COL1A2mRNA}])$$
$$+V_{\mathrm{Nucleus}}\cdot\_\mathrm{kBasalCOL1A2mRNA}+V_{\mathrm{Nucleus}}\cdot\_\mathrm{kCOL1A2mRNA\_prod\_by\_Smads}\cdot[\mathrm{pSmad3\_Smad4n}$$
$$+V_{\mathrm{Cytoplasm}}\cdot\left(\frac{V_{\mathrm{Nucleus}}\cdot\_\mathrm{kCOL1A2mRNA\_prod\_by\_CTGF}\cdot[\mathrm{CTGFmRNA}]}{V_{\mathrm{Cytoplasm}}}\right)$$

$$\frac{\mathrm{d}\left([\mathrm{COL1A1mRNA}]\cdot V_{\mathrm{Cytoplasm}}\right)}{\mathrm{d}t} = -V_{\mathrm{Cytoplasm}}\cdot(\_\mathrm{kCOL1A1mRNA\_deg}\cdot[\mathrm{COL1A1mRNA}])$$
$$+V_{\mathrm{Nucleus}}\cdot\_\mathrm{kBasalCOL1A1mRNA}+V_{\mathrm{Nucleus}}\cdot\_\mathrm{kCOL1A1mRNA\_prod\_by\_Smads}\cdot[\mathrm{pSmad3\_Smad4n}$$
$$+V_{\mathrm{Cytoplasm}}\cdot\left(\frac{V_{\mathrm{Nucleus}}\cdot\_\mathrm{kCOL1A1mRNA\_prod\_by\_CTGF}\cdot[\mathrm{CTGFmRNA}]}{V_{\mathrm{Cytoplasm}}}\right)$$

$$\frac{\mathrm{d}[\mathrm{pSmad3\_Smad4}]\cdot V_{\mathrm{Cytoplasm}}}{\mathrm{d}t} = +V\mathrm{Cytoplasm}\cdot\_\mathrm{kpSmad3\_bind\_Smad4}\cdot[\mathrm{pSmad3}]\cdot[\mathrm{pSmad3}]\cdot[\mathrm{Smad4}]$$
$$-(V_{\mathrm{Cytoplasm}}\cdot\mathrm{kimp\_pSmad3\_Smad4}\cdot[\mathrm{pSmad3\_Smad4}])$$

$$
\begin{aligned}
\text{kcd} &= \tfrac{1}{36} \\
\text{kCOL1A2mRNA\_deg} &= \text{\_kCOL1A1mRNA\_deg} \\
\text{TGFbMoleculesPerLiter} &= \frac{\frac{\text{TGFb\_dose\_in\_ng\_per\_liter}}{1000000000}}{25000} \cdot 6.0223e+23 \\
\text{TGFbMoleculesPerCell} &= \text{TGFbMoleculesPerLiter} \cdot V_{\text{Medium}} \\
\text{Smad3Tot} &= [\text{Smad3}] + [\text{pSmad3}] + [\text{Smad3n}] + [\text{pSmad3n}] \\
&\quad +[\text{pSmad3\_Smad4}] + [\text{pSmad3\_Smad4n}] \\
\text{pSmad3Tot} &= [\text{pSmad3}] + [\text{pSmad3\_Smad4}] + [\text{pSmad3n}] + [\text{pSmad3\_Smad4n}]
\end{aligned}
$$

B.2 Extension Hypotheses

Four extension hypothesis were proposed in Chapter 6. Each hypothesis is a simple replacement of a single equation from the base model and represents the addition of a single biochemical reaction. Those single equations described in the following.

B.2.1 Hypothesis 1

$$
\begin{aligned}
\frac{\mathrm{d}\left([\text{R2}] \cdot V_{\text{Cytoplasm}}\right)}{\mathrm{d}t} &= + V_{\text{Cytoplasm}} \cdot (\text{vR2}) \\
&\quad - V_{\text{Cytoplasm}} \cdot (\text{kcd} \cdot [\text{R2}]) \\
&\quad - (V_{\text{Cytoplasm}} \cdot \text{kaTGFb} \cdot [\text{R2}] \cdot [\text{TGFb}]) \\
&\quad + (V_{\text{Cytoplasm}} \cdot \text{kdTGFb} \cdot [\text{R2\_TGFb}]) \\
&\quad - V_{\text{Cytoplasm}} \cdot (\text{Adult} \cdot \text{kReducedR2InAge})
\end{aligned}
$$

B.2.2 Hypothesis 2

$$\begin{aligned}\frac{\mathrm{d}\left([\mathrm{Smad3}]\cdot V_{\mathrm{Cytoplasm}}\right)}{\mathrm{d}t} &= +\left(V_{\mathrm{Nucleus}}\cdot \mathrm{kexp\_Smad3n}\cdot[\mathrm{Smad3n}]\right)\\ &\quad -\left(V_{\mathrm{Cytoplasm}}\cdot \mathrm{kimp\_Smad3}\cdot[\mathrm{Smad3}]\right)\\ &\quad -V_{\mathrm{Cytoplasm}}\cdot\left(\mathrm{\_kSmad3Phos}\cdot[\mathrm{Smad3}]\cdot[\mathrm{LRC}]\right)\\ &\quad -V_{\mathrm{Cytoplasm}}\cdot\left(\mathrm{kSmad3\_deg}\cdot[\mathrm{Smad3}]\right)\\ &\quad +V_{\mathrm{Cytoplasm}}\cdot\left(\mathrm{kSmad3\_prod}\right)\\ &\quad -V_{\mathrm{Cytoplasm}}\cdot\left(\mathrm{Adult}\cdot\mathrm{kReducedSmad3InAge}\cdot[\mathrm{Smad3}]\right)\end{aligned}$$

B.2.3 Hypothesis 3

$$\begin{aligned}\frac{\mathrm{d}\left([\mathrm{Smad3}]\cdot V_{\mathrm{Cytoplasm}}\right)}{\mathrm{d}t} &= +V_{\mathrm{Cytoplasm}}\cdot\left([\mathrm{CTGFmRNA}]\cdot\mathrm{kCTGF\_increase\_Smad3}\right)\\ &\quad +\left(V_{\mathrm{Nucleus}}\cdot\mathrm{kexp\_Smad3n}\cdot[\mathrm{Smad3n}]\right)\\ &\quad -\left(V_{\mathrm{Cytoplasm}}\cdot\mathrm{kimp\_Smad3}\cdot[\mathrm{Smad3}]\right)\\ &\quad -V_{\mathrm{Cytoplasm}}\cdot\left(\mathrm{\_kSmad3Phos}\cdot[\mathrm{Smad3}]\cdot[\mathrm{LRC}]\right)\\ &\quad -V_{\mathrm{Cytoplasm}}\cdot\left(\mathrm{kSmad3\_deg}\cdot[\mathrm{Smad3}]\right)\\ &\quad +V_{\mathrm{Cytoplasm}}\cdot\left(\mathrm{kSmad3\_prod}\right)\end{aligned}$$

B.2.4 Hypothesis 4

$$\begin{aligned}\frac{\mathrm{d}\left([\mathrm{Smad7mRNA}]\cdot V_{\mathrm{Cytoplasm}}\right)}{\mathrm{d}t} &= -V_{\mathrm{Cytoplasm}}\cdot\left(\mathrm{\_kSmad7mRNA\_deg}\cdot[\mathrm{Smad7mRNA}]\right)\\ &\quad +V\mathrm{Nucleus}\cdot\mathrm{\_kBasalSmad7mRNA}\\ &\qquad +V\mathrm{Nucleus}\cdot\mathrm{\_kSmad7mRNA\_prod}\cdot[\mathrm{pSmad3\_Smad4n}]\\ &\quad -V_{\mathrm{Cytoplasm}}\cdot\left(\mathrm{\_kSmad7\_degrade\_LRC}\cdot[\mathrm{Smad7mRNA}]\cdot[\mathrm{LRC}]\right)\\ &\quad -V\mathrm{Cytoplasm}\cdot[\mathrm{Smad7mRNA}]\cdot[\mathrm{CTGFmRNA}]\\ &\qquad \cdot\mathrm{kSmad7DegByCTGF}\end{aligned}$$