THE UNIVERSITY
of EDINBURGH

Large-Scale Analysis of Microarray Data to Identify Molecular Signatures of Mouse Pluripotent Stem Cells

Aidan McGlinchey BSc (Hons) MRes

A Thesis submitted for the degree of Doctor of Philosophy
The University of Edinburgh
2017

Abstract

Publicly-available microarray data constitutes a huge resource for researchers in biological science. A wealth of microarray data is available for the model organism – the mouse. Pluripotent embryonic stem (ES) cells are able to give rise to all of the adult tissues of the organism and, as such, are much-studied for their myriad applications in regenerative medicine. Fully differentiated, somatic cells can also be reprogrammed to pluripotency to give induced pluripotent stem cells (iPSCs). ES cells progress through a range of cellular states between ground state pluripotent stem cells, through the primed state ready for differentiation, to actual differentiation.

Microarray data available in public, online repositories is annotated with several important fields, although this accompanying annotation often contains issues which can impact its usefulness to human and / or programmatic interpretation for downstream analysis. This thesis assembles and makes available to the research community the largest-to-date pluripotent mouse ES cell (mESC) microarray dataset and details the manual annotation of those samples for several key fields to allow further investigation of the pluripotent state in mESCs.

Microarray samples from a given laboratory or experiment are known to be similar to each other due to batch effects. The same has been postulated about samples which use the same cell line. This work therefore precedes the investigation of transcriptional events in mESCs with an investigation into whether a sample's cell line or source laboratory is a greater contributor to the similarity between samples in this collected pluripotent mESC dataset using a method employing Random Submatrix Total Variability, and so named RaSToVa. Further, an extension of the same permutation and analysis method is developed to enable Discovery of Annotation-Linked Gene Expression Signatures (DALGES), and this is applied to the gathered data to provide the first large-scale analysis of transcriptional profiles and biological pathway activity of three commonly-used mESC cell lines and a selection of iPSC samples, seeking insight into potential biological differences that may result from these.

This work then goes on to re-order the pluripotent mESC data by markers of known pluripotency states, from ground state pluripotency through primed pluripotency to earliest differentiation and analyses changes in gene expression and biological pathway activity across this spectrum, using differential expression and a window-scanning approach, seeking to recapitulate transcriptional patterns known to occur in mESCs, revealing the existence of putative “early” and “late” naïve pluripotent states and thereby identifying several lines of enquiry for in-laboratory investigation.

Lay Summary

Humans are living organisms made of much smaller sub-units called cells. These cells come in different types, depending on their function, for example cells of the liver are involved in energy storage, detoxification and even immunity. Among the many different subtypes of cells, stem cells have important properties that make them of great interest to biomedical research. Stem cells can provide a supply of other cell types, almost indefinitely. For example, skin cells are constantly shed from the body but are replaced from a small number of stem cells residing in the deeper skin.

Embryonic stem cells (ESCs) are the stem cells present in the early embryo that create all of the cells of the adult, so are called “pluripotent”. A great deal of research has been carried out on mouse ESCs (mESCs) to pick apart their properties of self-renewal and creating child cells to build and maintain an entire mouse. Looking at the genes that are “on or off” is a common way to do this, and microarrays give simultaneous measurements of the average levels of gene activity of millions of cells at a time, this forming one “sample”. Much microarray data is available in public, online locations, but the way in which information about individual mESC samples is sometimes confusing or obscure to both humans and computers trying to analyse it. This thesis brought together the largest-to-date pluripotent mESC microarray dataset and the samples within it were all manually annotated to make analysis much easier, and also created a simple way to record experimental details that humans and computers can work with and understand easily.

Samples from the same lab or cell line are known to be similar to each other, which can get in the way of looking for patterns in the data that explain the biology we are researching. This work created two new tools, called RaSToVa (Random Submatrix Total Variability) and DALGES (Discovery of Annotation-Linked Gene Expression Signatures). RaSToVa measures how much one annotation (e.g. cell line) makes samples more similar to each other. The tool does this in a way that allows researchers to compare this increase in similarity of samples due to different annotations. DALGES takes this further by showing researchers which individual genes appear to be more “on” or “off” in samples that share the same annotation (e.g. being from the same lab.)

Finally, this work devises a way to sort all of the samples in this large mESC dataset between their earliest known stage of the pluripotent state, to the state right before the development of the mouse truly begins, finding new steps that the cells go through on the way, and showing patterns of which genes turn on and off, and by how much, as the cells go across these states. This approach holds a lot of promise for using the large amounts of data already out there and publicly available to do similar things to gain new insights about other biological processes, including repeating this in human ESCs, which would be of great use to medical science indeed.

Acknowledgements

The work in this thesis was partly funded by Biotechnology and Biological Sciences Research Council (BBSRC). Grateful and sincere thanks are given here to so many over the years. To my supervisor, Dr Simon Tomlinson, for his guidance and patience and for also standing by me when more time was needed for this work. To Professor Ian Chambers, my second supervisor and committee member for his constructive criticism, honesty and belief in me.

To Dr Edward Curry, friend and ex-colleague of the Tomlinson lab and now research fellow at Imperial College, London's Bioinformatics Hub, for his support and discussion of the writing of this thesis during his own hectic work. To the other members of the Tomlinson lab who contributed discussion, criticism and witticism: Duncan Godwin, Florian Halbritter, Stavroula Skylaki and Anastasia Kousa.

To Professor Anura Rambukkana for his support and encouragement and flexibility whilst I was working and publishing part-time with his lab.

To Dr Stephen Bush, old friend from Plymouth University and current research fellow at the Roslin Institute, for his discussion of this work and insistently-upbeat nature.

From my important, formative days at Plymouth University, sincerest thanks to my first research supervisor, Professor Tamara Galloway whose encouragement and belief are still with me today. To Dr. Andrew Foey, immunology lecturer who, in the absence of any parents to attend my first graduation, shouted an unforgettable and gloriously inappropriate “Well done, Aidan!” from the lecturer's area. To Dr. Simon Fox, bone pathology and molecular cell biology lecturer, for his enthusiasm and excitement in his teaching. To Dr. Stephen Thompson, my previous neurobiology of pain and molecular biology lecturer for his undying inspiration in managing to teach at a speed to match the timetable he was given as well as weathering my daily question barrage. To Dr John Moody, my biochemistry and metabolism lecturer for his encouragement and his understanding in my lateness to lectures due my having part-time jobs instead of financial support. To Professor Roddy Williamson, neurobiology lecturer and Peter Brooks, my personal supervisor at Plymouth, for their kindness and encouragement when my circumstances were dire.

To my oldest friend from Plymouth University, David Boon, for many laughs, over the years and time taken to be there for me when things appeared hopeless. To Sophie Gallagher, though she be an ex-partner, our exchange of support, kindness, care and love during our time at Plymouth University is never forgotten. To Bernice Fawell for being a friend to talk to during early and difficult, solitary times after my move to Edinburgh University.

To Mick Ludwig, finance director at Plymouth University, for naming me “the Harry Potter of Plymouth University” for my orphaned background and refusal to give up. Many a fee payment to the University of Plymouth would have been impossible without his flexibility. The same for Sally St. John, still working to this day in student finance support at Plymouth University, that the less fortunate can hope for better in life. Without such support, I would never have achieved even my BSc.

Although purely emotional support, thanks are very much due to Kate Fletcher, my counsellor for the last year, who offered me a very important kind ear and heart.

Final thanks now to those whose support was both material and strongly emotional. To my partner, Layla, for your support, understanding, kindness and love during the most testing and terrifying times of my PhD. You not only gave me four walls and a roof, but filled that home with love, compassion, fun and rest; things which I often gave myself little of and that I had not known for considerable time. You endured my company during the low periods during this PhD, but were there with me riding the highs. You are now there at the end of this time, also having been waiting for me for these last months of my near-total absence due to writing this thesis.

Lastly, thanks are due to someone whom I shall have to wait a much longer time to reunite with. Mum, though you died when I was 17, at my school, on October the 2nd 2001, you have never been away from my heart or mind. You could not be there for my first or second graduation, and I will again be paying for a seat to be kept empty at my graduation for this PhD. You dressed me, loved me, fed me, shouted at me, protected me, held me, encouraged me, believed in me and, as a teacher by profession and nature, always expounded the vital importance of education, urging me to go as high as I could. I have been on that mission since you left and it has been unbearable at times without you, but now at the end of my education I hope that you will be there watching from that empty seat to see it happen.

I kept my promise.

“The true sign of intelligence is not knowledge but imagination”

- Albert Einstein

“If a path to the better there be, it begins with a full look at the worst.”

- Thomas Hardy,

“An expert is someone who has made all the mistakes which can be made, in a narrow field”

- attributed to Niels Bohr by Edward Teller

“Start by doing what's necessary; then do what's possible; and suddenly you are doing the impossible”

- Francis of Assisi

Declaration

I have read and understood the University of Edinburgh's guidelines on plagiarism and declare that the work presented in this thesis is my own, unless otherwise stated, and that it has not been submitted towards any other degree or professional qualification.

Aidan James McGlinchey BSc (Hons) MRes

Table of Contents

Illustration Index

Chapter 1 –

Introduction

1.1 Embryonic Stem Cells

Before discussing the motivation driving the study of embryonic stem (ES) cells, it is first necessary to define the embryonic stem cell. Again, prior to this, the definition of the "stem cell" is required. A stem cell is a cell which is capable of both self-renewal and differentiation. The terminology applied to stem cells has, over recent years, required ever more careful use. For example, it is not equivalent to say that a stem cell is any cell capable of self-renewal and *proliferation*, as these properties combined would not facilitate the function of the stem cell; to renewably create new or replace old tissue under strict control. The existence of the "cancer stem cell", for example, is reason enough to stress the judicious use of such terminology, in that a stem cell capable only of creating copies of itself and a mass of differentiation-incompetent progenitors which fail to respond to cues is a very different cell indeed.

The term "embryonic stem cell" was coined when they were first cultured through the use of media conditioned by mouse teratocarcinoma cells (Martin, 1981). The first experiments in culturing ES cells took place this same year (Evans and Kaufman, 1981). Mouse ES cells (mESCs), specifically, are those cells which, in the *in vivo* case, populate the inner cell mass (ICM) region of a developing mouse embryo and eventually give rise to the tissues of the three canonical germ layers of the adult organism: the endoderm, mesoderm and ectoderm. The etymology of the prefix "embryonic" is therefore obvious. These cells occur at around day E3.5 (see 1.2 for more detail on the developmental context.)

However, given that techniques now exist to generate and manipulate these cells *in vitro*, there is sometimes usage of the term "embryonic stem cell" when, in actuality, the cells in question are not taken directly from an embryo at all (e.g. they may be cells long-passaged from an original isolation), or may be generated by nuclear transfer methods or, particularly in the case of much of the data compiled in this thesis, maintained in an *ES-like* state *in vitro* for the purposes of manipulation and study.

It is of fundamental interest to present stem cell science in general as well as, consequently, this thesis, what the true identity of an ES cell is, if such a thing can be said to exist. This question, like so many in biological science, has had to be qualified and redefined time and time again. It is now accepted, although perhaps not always stated or stressed sufficiently for accuracy, that the ES cell is

only a transient cell state in the development of certain organisms (including mouse and human), and that our attempts, as scientists, to maintain this ES cell state likely actually removes that cell considerably from any naturally-occurring *in vivo* state. This is, however, necessary for the study of and subsequent harnessing of the abilities of the ES cell. Therefore, throughout the course of this thesis, for convenience, the term mESC may be used when referring to samples when, in the great majority of cases, these cells are removed from their original *in vivo* state of being true mESCs, however convention in the literature still has them referred to as mESCs rather than mESC-like or mESC-derived.

The mouse ES cell has historically been, and currently still is, the most widely-available and best understood ES cell and forms a great part of the basis for translating our understanding of stem cell maintenance, differentiation and pluripotency to the human, for ultimate use (see 1.7). It is for this reason that mouse data was chosen for use in this thesis; in order to amass as large-scale an analysis as possible, given the amounts of data publicly available for different organisms.

1.2 Developmental Origins

The mESC is a cell of the ICM of the developing mouse embryo which is to be found around the day E3.5 time point and is defined as a cell which can self-renew (that is, give rise to identical progeny) and differentiate, that is, generate daughter cells which are fated to give rise to the main three germ layers of endoderm, mesoderm and ectoderm. This means that the mESC is, ultimately, what gives rise to the entire adult animal. A detailed review of the developmental events related to mESCs can be found in (Nichols and Smith, 2012).

Starting as the fertilised egg, the diploid zygote, division of this cell occurs repeatedly into several cells known as blastomeres, but no real increase in volume occurs. Being binary in nature, these cell divisions proceed to generate first 2, then 4, then 8 cells and so on. The next stage of development of the mammalian embryo is compaction, involving the maturation of cell adhesion complexes to bind the cells together. At this stage, the 16-cell stage, the whole entity is referred to as the morula, with the innermost cells fated to become the ICM.

Division proceeds in doubling cell numbers as the outer cells (distinct now from the ICM) begin to form into the trophectoderm (TE); the support tissue for the embryo proper. Fluid is actively

produced by the trophectoderm and pumped into the centre of what is rapidly becoming the blastocyst stage, forming a cavity on one side, giving rise to the easily-recognised shape distinguishing between the areas of the ICM and the TE. This, in the mouse, occurs at the E3.5 time point and it is here that we find our *bona fide* mESCs in the ICM.

It is crucial to appreciate at this point that the state of the mESC, derived from the ICM at this time, is, being a developmental *stage,* therefore, *by definition*, a *transient* state *in vivo.*

Following this stage of development, the ICM proceeds to both proliferate and differentiate into the three aforementioned germ layers, with ectoderm to the nervous system and the skin, its "opposite", the endoderm, to the gastrointestinal tract and other viscera, and the intermediary layer, the mesoderm, to the bone, muscle, cartilage and the haematopoietic system. At any stage following on from the undifferentiated ICM stage, our mESCs essentially cease to exist in the pluripotent state as their progeny begin to restrict their possible fates and progress towards final cellular identities. This work concentrates on mESCs, their maintenance and early exit from pluripotency, and thus any development beyond this ICM stage is outside the scope of this thesis.

From this explanation of the earliest developmental events is made apparent the definition of "pluripotent" for our purposes, being the ability to give rise to the 3 major germ layers defined above. It is therefore not necessary for a "pluripotent" cell to be able to give rise to trophectoderm, although this is a matter of convention in stem cell biology, not in keeping with the meaning of the prefix itself, where "pluri" simply means "many", as does "multi" from "multipotent". Multipotency, by convention, however, refers to the ability of a stem cell to give rise to multiple types of more differentiated progeny, but further towards adult stem cells than the pluripotency of mESCs. Cells capable of giving rise to all 3 germ layers and also trophectodermal tissue are referred to as "totipotent", where the prefix "toti-" refers to "all."

1.3 Embryonic Stem Cell Transcriptional Machinery

The ESC transcriptome is required to be able to perform three major functions. The first is the universal requirement for the so-called "housekeeping" processes of cellular life, such as metabolism, repair and maintenance of cell structures and cellular homeostasis. The second is the driving of the undifferentiated self-renewal program, providing replicatively-vigorous identical

progeny with their genetic material intact. The third is to detect and respond correctly to developmental cues which instruct the cell to exit from the self-renewal program and start down a differentiation pathway. This makes the mESC a highly-specialised cell in its own right. Given the complexity of internal and external signalling events constantly processed and responded to by mESCs, it is inaccurate and misleading to describe mESCs as “unspecialised” cells.

The study of the state of, and mechanisms underlying, ES cell transcriptional machinery is central to the area of basic stem cell research. Transcriptional “networks” in ES cells have therefore been the subject of intense research in recent years. This brief overview of the transcriptional workings of mESCs is given to provide context and background understanding of the major factors currently known to be involved in the maintenance of pluripotency and exit from it.

1.3.1 Oct4, Sox2 and Nanog Core Transcriptional Circuitry

Agreement has been present in the literature for some years now concerning three transcription factors, octamer-binding transcription factor 4 (Oct4) (also known as Oct4), sex-determining-region Y box 2 (Sox2) and Nanog (pronounced “nanOg”), named after the mythical Irish land of the forever young, as forming the core transcriptional circuitry underpinning pluripotency in mESCs (Young RA, 2011), (Nichols and Smith, 2012), (Yeo and Ng, 2013) (Hackett and Surani 2014).

In the context of mESCs, Oct4 expression is closely associated with the identity of true ICM cells, and is downregulated in the TE, concomitant with the expression of Cdx2 and eomesodermin (Eomes), driven by expression of TEA domain family member 4 (Tead4) in the TE. This is due to Hippo signalling being activated in cells that are surrounded by other cells at this point in development. Those cells on the surface of the developing blastocyst, not being surrounded, do not have active Hippo signalling, helping to partition the TE from the ICM (Wada et al. 2011). Furthermore, knockdown of Oct4 causes a failure to generate an ICM, favouring TE generation instead (Nichols et al. 1998). In this same work it was also demonstrated that a lack of Oct4 caused a near-total failure of transcription of FGF4, a transcription factor required for the generation of the hypoblast through paracrine signalling and also known to be required for later exit from pluripotency by way of its activation of MAPK/ERK signalling (Kunath et al. 2007). Whilst it may be expected that underexpression of Oct4 would cause a loss of pluripotency, it has also been demonstrated firstly that overexpression by 50% of Oct4 does not reinforce pluripotency, but in fact

promotes differentiation towards mesodermal and endodermal lineages (Niwa et al. 2000). Secondly, the level of expression of Oct4 in stably-pluripotent mESCs was further explored in work by (Karwacki-Neisius et al. 2013), and this work identified a level of Oct4 expression which coincided with both robust pluripotency and greatest enhancer occupancy by Oct4 and Nanog.

Oct4 has also been conclusively shown to act as a heterodimer in conjunction with the next of the three canonical pluripotency factors: Sox2. The initial work that found co-operative binding between Oct4 and Sox2 was by (Yuan et al. 1995) in the case of co-operative binding to the FGF4 enhancer. Indeed, there is such overlap in the targets of Oct4 and Sox2 (Chew et al. 2005) that expression of Oct4 from a transgene is capable of rescuing pluripotency in Sox2-null embryos. The use of a transgene in this work by (Masui et al. 2007) to express Oct4 was necessary in that Sox2 and Oct4 promote each other's expression, and therefore it would be expected that Oct4 expression would decrease considerably in the absence of Sox2. Sox2-null embryos are, however, capable of forming normal blastocysts, but die shortly after implantation due to a failure to generate a true epiblast, with development only progressing as far as implantation by the grace of maternally-derived Sox2 from the initial zygote (Avilion et al. 2003). This co-operative binding of Oct4 and Sox2 has been shown to take place at a great many pluripotency-related genes (Chambers and Tomlinson 2009).

The third of these three canonical pluripotency factors, Nanog, initially known as Ecat4, was initially identified in a cDNA screen for factors which could confer self-renewal to leukaemia inhibitory factor (LIF) receptor (LIFR) knockout ($Lifr^{-/-}$) mESCs, dubbed LRK1 cells (Chambers et al. 2003). In this work, both of the pools of transfected cells which gained self-renewal capability contained cDNA for Nanog. Parallel work in which the name Nanog was first coined also demonstrated Nanog's ability to sustain mESCs in the absence of LIF (Mitsui et al. 2003). Loss of Nanog results in embryos which do not have a pluripotent ICM and is embryonic lethal (Chambers et al. 2007). However, this work also demonstrated that it is possible to delete Nanog from self-renewing mESCs *in vitro*, and that whilst there is a greatly increased propensity towards differentiation, loss of Nanog does not prevent continued passage of self-renewing mESCs. Nanog is said, therefore, to be required for the development of pluripotency, but is not strictly required for its maintenance (Chambers et al. 2007). This same work also demonstrated that Nanog fluctuates in its expression in individual cells, from a $Nanog_{high}$ state to a $Nanog_{low}$ state, cyclically. This correlates with a cycle of robust self-renewal and increased propensity to differentiate, respectively. Furthermore, work by (Navarro et al. 2012) demonstrated that the fluctuation of Nanog is the result

of autorepression and that Nanog does not self-promote, being dependent instead on Oct4/Sox2. Nanog has also been shown to be both required for and capable of reprogramming EpiSCs to naïve pluripotency (Silva et al. 2009) (see section 1.5 for details on EpiSCs.)

A further indication of the pluripotent ICM-specific nature of Nanog is in the initial specification of epiblast versus hypoblast in the developing embryo. From an initial stochastic mosaic of cells positive for either Nanog or Gata6 in the ICM, the centre of the ICM remains Nanog positive while Gata6 positive cells that find themselves on the periphery of the forming hypoblast continue to express Gata6 and progress toward hypoblast differentiation. This differentiation towards hypoblast specification appears, again, to be dependent on FGF signalling through the MAPK/ERK pathway, as prevention of MAPK/ERK signalling (through deletion of the downstream growth factor receptor bound protein 2 (Grb2) gene) results in a failure to form hypoblast with all ICM cells becoming stably positive for Nanog and a concomitant elimination of Gata6 positive cells destined for hypoblast formation (Chazaud et al. 2006). Conversely, the addition of sufficient exogenous FGF4 directs all would-be ICM cells to form hypoblast (Yamanaka et al. 2010). This again indicates that paracrine signalling through the fibroblast growth factor receptor and MAPK/ERK pathways are key signals in the developing mouse embryo at this stage.

1.3.2 The role of c-Myc

Whilst the transcription factors Oct4, Sox2 and Nanog (OSN) form what is agreed upon as the core pluripotency network, transcription of OSN target genes is affected by the c-Myc. Whilst transcription factors are responsible for facilitating the binding of RNA polymerase II (RNApol II) to the OSN target gene loci, transcription has been shown to only proceed a short distance from the transcription start site (TSS) (around 35 base pairs (bp)) (Rahl et al. 2010). c-Myc functions by permitting this short transcription event to proceed beyond terminating at around 35bp, to completion of a full transcript. Work by groups such as (Nie et al. 2012) confirmed c-Myc's general transcription-promoting activity. This general transcription-promoting activity of c-Myc explains why it is associated both with ES cells and with cancer stem cells ((Lin et al. 2012), (Kim et al. 2010), (Rothenberg et al. 2010)) and also explains its ability to reinforce pluripotency and/or the efficiency of reprogramming of somatic cells to pluripotency (Takahashi and Yamanaka 2006) (see section 1.6 on induced pluripotent stem cells.))

In a review of transcriptional workings of mESCs, (Young RA, 2011) demonstrated that of a total of 9355 genes found to be active in mESCs, 3497 (37%) were targets of both c-Myc and the OSN factors. 2504 (27%) of active genes were targets of the OSN factors but not c-Myc and 1847 (20%) of active genes were targets of c-Myc but not all of the OSN factors (figure 3A of (Young RA, 2011).)

1.3.3 External factors and signalling pathways involved in mESC biology

Whilst an in-depth review of all molecular processes involved in individual signalling pathways is not necessary for the interpretation of the work presented in this thesis, mESC biology cannot be discussed in any meaningful manner without at least understanding the relevance of those signalling pathways known to be involved in the maintenance of pluripotency and/or exit from it. More information on the mechanisms involved in some of the pathways mentioned here can be found in the review by (Dreesen and Brivanlou 2007).

Initial efforts to culture mESCs used mitotically-inactivated feeder cells and culture media containing serum (Martin, 1981), (Evans and Kaufman, 1981) . At this time it was not known what factor(s) in the added serum were responsible for maintenance of the cells. Work by (Williams et al. 1988) and by (Smith et al. 1988) identified LIF as one of the factors involved. Typical signalling in this pathway involves the binding of LIF to it's heterodimeric receptor, consisting of the LIF receptor (LIFR) and glycoprotein 130 (gp130). This results in the downstream phosphorylation of Janus kinase (JAK) and the subsequent activation of the Janus kinase / Signal transducer and acitvator of transcription (JAK/STAT) pathway. From this it is already clear to see that active JAK/STAT signalling is implicated in the maintenance of pluripotency. STAT3, in particular, has been shown to be highly important for the maintenance of self-renewal in mESCs (Niwa et al. 1998), which demonstrated, amongst other things, that inhibition of STAT3 presents a block to self-renewal and promotes differentiation. Conversely, work by (Matsuda et al. 1999) demonstrated that mESCs were maintained in the pluripotent state when a doxycycline (DOX)-inducible STAT3 was expressed, by using a STAT3-estrogen receptor fusion protein. One mechanism behind this requirement for STAT3 was shown in work by (Cartwright et al. 2005) in that STAT3 was demonstrated to promote expression of Myc (see section 1.3.2 for the importance of Myc), and, in the absence of LIF (read: JAK/STAT signalling), Myc is targeted for degradation by glycogen synthase kinase beta (GSK3β), one of the best-known antagonists of Wnt signalling (discussed later

in this section.) Work was also being carried out at this time by (Sekkai et al. 2005) in identifying downstream targets of LIF/STAT3 signalling, by either removing LIF or by expression of inactive Stat3 to observe which genes changed their expression as cells progressed towards differentiation. This revealed a connection to the transforming growth factor beta (TGFβ) signalling pathway as well as the Id family of proteins (discussed later in this section as being downstream also of BMP signalling.)

It was work by (Ying et al. 2003) which identified BMP4 as the other major factor in serum which, when coupled with LIF, could substitute for serum entirely in the culture of mESCs. It therefore follows that signalling through both the LIF and BMP pathways are involved in mESC biology. BMP4 was subsequently found to inhibit MAPK/ERK signalling in mESCs by (Qi et al. 2004), with small molecule inhibitors of MAPK/ERK allowing the maintenance of mESCs even in the absence of bone morphogenetic protein receptor 1 alpha (BmprIa), confirming that the relevant function of BMP4 in maintenance of mESC self-renewal is to inhibit MAPK/ERK signalling. In mESCs, the BMP signalling pathway acts through SMADs (named as a combination of “sma” - small body type gene in *C. elegans*, and “mothers against decapentaplegic” (MAD)). In the case of mESCs, BMP signalling acts through SMADs and is responsible for the expression of inhibitor of differentiation (Id) genes (Ying et al. 2003). In fact, this work also neatly demonstrated that culture of mESCs with LIF, but not BMP4, results in cells proceeding towards neural differentiation. It is the action of BMP/SMAD signalling and particularly the expression of these Id genes that prevents this when culturing with both LIF and BMP, as induced expression of Id genes could compensate for a lack of added BMP when cells were cultured with only LIF.

As regards TGFβ signalling, including BMP and Nodal signalling, which are part of the TGF family, an excellent review of the effects of these inter-related pathways can be found in (Watabe and Miyazono 2009) and, more recently at great length in (Sakaki-Yumoto et al. 2013). To summarise a couple of relevant facts as regards the work in this thesis, however, Activin/Nodal (which are extracellular signalling molecules) signal primarily through phosphorylation of Smad2, and Smad2 was shown by (Lee et al. 2011) to be phosphorylated to different levels in mESCs by differing levels of Activin/Nodal signalling. This, in turn, changed the sets of genes targeted by the phosphorylated Smad2, including Oct4. These differing levels of Nodal signalling ranged in their effects from the maintenance of pluripotency (middling level of Activin/Nodal) to promotion of mesoderm / endoderm differentiation (high Activin/Nodal) (Lee et al. 2011). As overexpression of Oct4 is also associated with mesoendodermal differentiation, this may be therefore how high

Activin/Nodal signalling causes mesoendodermal differentiation. This work also showed that total blockade of Nodal signalling surprisingly resulted in differentiation to the trophectoderm fate, again reinforcing the importance of close titration of signalling to maintain mESC self-renewal. Work by (Vallier et al. 2009) showed that Nodal signalling, particularly in the case of primed pluripotent (EpiSC) mESCs, is linked to the expression of Nanog, which, in turn, prevents differentiation. Taken together, this explains, although far from fully, how a basal level of Nodal signalling is required in naïvely-pluripotent mESCs for self-renewal, but this changes to a more definite requirement for strong Nodal signalling in primed mESCs for their maintenance in culture.

As regards MAPK/ERK signalling, with the aforementioned discovery that MAPK/ERK inhibition promotes self renewal, it follows that activation of MAPK/ERK must be involved in the exit from pluripotency. This is indeed the case, as shown in pivotal work by (Kunath et al. 2007). This work also links mESC expression of autocrine FGF4 to the cells' ability to differentiate, with inhibition of FGF signalling causing a resistance to differentiation, which can be restored by the administration of exogenous FGF4. In turn, this begins to give an overview of how mESCs can self-renew in the appropriate environment (read: activation of LIF and BMP signalling), but remain poised to differentiate when MAPK/ERK signalling is uninhibited when BMP signalling is deactivated. Work by (Lanner et al. 2010) demonstrated the autocrine nature of FGF signalling and its relevance to mESCs biology by demonstrating that mESC differentiation was prevented by the removal of mESCs ability to sulphate proteoglycans on their cell surface which are critical for FGF signalling. This could be mimicked by treatment with $NaClO_3$, which blocks the same sulphation from taking place and pushes mESCs towards naïve pluripotency. A review of MAPK/ERK signalling in the context of mESCs produced by the same group is available in (Lanner and Rossant 2010). Recent work by (Hamilton et al. 2013), however, demonstrated that deletion of ERK2 from mESCs was not sufficient as a block to differentiation, as functioning ERK1 granted the MAPK/ERK signalling pathway redundancy. However, this work did also show that deletion of ERK2 (the primary ERK in mESCs) did reinforce markers of pluripotency and reduced heterogeneity in the cells, in agreement with discoveries such as the ability of the “2i” culture condition to promote naïve pluripotency (discussed later in this section.) Even more recently, (Yeo et al. 2014) demonstrated that inhibition of Klf2 is downstream of MAPK/ERK and explains another way in which inhibition of MAPK/ERK (e.g. in 2i culture) facilitates the maintenance of the ground state of pluripotency in mESCs.

Another pathway intricately involved in mESC biology is the Wnt signalling pathway. Named as a portmanteau of “Wingless” and “integration” genes investigated in *D. melanogaster* and mouse respectively, Wnt signalling occurs through, with present knowledge, three ways. A review of these can be found in (van Es et al. 2003). A recent review with an excellent section on Wnt signalling in ES cells can be found in (Munoz-Descalzo et al. 2015), also.

Briefly, “canonical” Wnt signalling involves the release of β-catenin from its usual state of being bound to a complex involving Axin and adenomatous polyposis coli (APC), which target β-catenin for ubiquitin-mediated proteasomal destruction. Binding of Wnt ligand to the Frizzled receptor acts through intermediary factors such as Dishevelled and Frat to cause inhibition of GSK3β. In the absence of active GSK3β, β-catenin is not targeted for destruction and can accumulate in the cytoplasm and be translocated to the nucleus, where it interacts with TCF/LEF to alter transcription of target genes (van Es et al. 2003).

Following on from work comparing hESCs and mESCs (Sato et al. 2003), which identified Wnt genes as being important in hESC biology, further work by (Sato et al. 2004), used a specific inhibitor of GSK3β, a derivative of the naturally-occurring compound indirubin, 6-bromoindirubin-3′-oxime (BIO), to maintain self-renewing mESCs, keeping levels of Oct4, Nanog and the naïve pluripotency (not described until 2008 in (Ying et al. 2008)) marker Rex1 (Zfp42) in a high state of expression. Wnt signalling was also found by (Ogawa et al. 2006) to complement, but be unable to replace LIF as concerns the maintenance of mESCs in their self-renewing state, whether by the addition of exogenous Wnt3a, or the use of constitutively-active β-catenin, Wnt signalling could only reinforce, but not replace, LIF. The use of the specific Wnt ligand in this work, Wnt3a, is significant in that not all Wnts are capable of contributing to self renewal in mESCs, for example as demonstrated by (Singla et al. 2006). Wnt5a and Wnt6 were also found to be able to contribute to self-renewal of mESCs by (Hao et al. 2006), and were identified as factors in serum generated by feeder cells in their work. (Hao et al. 2006) went on to recapitulate the actions of Wnt5a and Wnt6 through the activation of β-catenin and demonstrate a link between Wnt signalling and the expression of STAT3, again linking LIF and Wnt signalling. This is not the only link that Wnt has into the heart of pluripotency, as the core transcription factor Oct4 is linked to Wnt signalling also, as demonstrated by (Takao et al. 2007), who showed that not only does LIF signalling itself increase the nuclear localisation of β-catenin, but that constitutively-active β-catenin could maintain mESCs in the absence of LIF. In this work, constitutively-active β-catenin upregulated Nanog, although this upregulation was dependent on Oct4. The interaction of β-catenin with Oct4 was shown in this

work also. This work is therefore at odds with the findings of (Ogawa et al. 2006), who were unable to maintain mESCs through their activation of β-catenin in the absence of LIF, although the reason for the difference in findings may be the choice of cell line, as (Takao et al. 2007) were not able to repeat their maintenance of mESCs with mutant β-catenin alone in another cell line. This may in turn suggest that the ability to maintain mESCs in the absence of LIF, using only mutant β-catenin could be an artefact of the use of a cell line which produces its own LIF. (Takao et al. 2007) attempted to address this through the use of an antibody for LIF, although even their experiments which successfully gave self-renewal to mESCs in the absence of LIF showed that their self-renewal and maintenance of an undifferentiated state was not complete, lending support again to the idea that Wnt signalling cannot, at least fully, replace LIF.

The aforementioned work identifying the ground state of pluripotency (Ying et al. 2008) was also the first time that the culture condition, 2-inhibitor, "2i", was invented. In this work it was found that dual inhibition of MAPK/ERK signalling and inhibition of GSK3β was capable of maintaining mESCs in a highly-homogeneous, self-renewing state. For example, whilst Nanog is known to fluctuate between high and low expression in mESCs (Chambers et al. 2007), the low-Nanog state being highly permissive to differentiation, "ground state" (naïve pluripotent) mESCs cultured in 2i are maintained in a homogeneous $Nanog_{high}$ state.

With these results in mind, it is somewhat unsurprising that Wnt signalling has also been shown to be involved in the naïve to primed pluripotency transition (ten Berge et al. 2011). This work also showed that activation of Wnt signalling through the use of exogenous Wnts could allow for the derivation of mESCs from a strain normally considered to be non-permissive, the inbred FVB/N strain (normally used in transgenics due to their large litter numbers). (ten Berge et al. 2011) further demonstrate that some ESC cell lines, such as CGR8 and E14 produce their own Wnts, this suggests that production of Wnt ligands (and, consequently, self-activation of Wnt signalling) may be behind some observed differences in the permissivity of different strains to ESC derivation.

Whilst the activation of Wnt signalling through the inhibition of GSK3β may provide a way to maintain mESCs in their naïve pluripotent state, the role of Wnt signalling in mESC biology through naïve pluripotency through to differentiation is highly complex, not fully understood and current opinion is that Wnt signalling is a highly context-dependent factor in mESC transcriptional regulation (Sokol 2011). Amongst other findings, work by (Wray et al. 2011) interestingly found knockout of the downstream target of canonical Wnt signalling, Tcf3, made mESCs able to self-

renew robustly in the presence of either LIF or a MAPK/ERK inhibitor without the requirement for both. This suggests that a crucial role of Wnt signalling is to inhibit the actions of Tcf3, which, in turn plays its part in promoting differentiation through the repression of the pluripotency network, although it is likely that there is a lot more to the story.

The most current opinions surrounding Wnt pathway signalling in mESCs centre on the activity of β-catenin in the context of ground state mESCs and EpiSCs, wherein it has been demonstrated that activity of β-catenin in mESCs is associated with the reinforcement of self-renewal and pluripotency, this is reversed in the case of EpiSCs, where activation of β-catenin drives these cells towards differentiation, as shown by (Kurek et al. 2015). This work demonstrated not only that Wnt signalling in EpiSCs drives them to differentiation, but also that the inhibition of Wnt stabilises the EpiSC state and allows for these EpiSCs to contribute to mouse chimera, when EpiSCs were previously found to not be able to do this (see section 1.5.) Finally, as regards EpiSCs and Wnt signalling, it has very recently been demonstrated that prevention of nuclear localisation of β-catenin is capable of reprogramming EpiSCs to the ground state (Murayama et al. 2015).

In summary, the actions of TCF/LEF and Wnt signalling are far from fully elucidated in the context of mESCs, and further research is required.

The Notch pathway is also active in mESCs, although with a role much less defined and investigated compared to other signalling pathways, it is known to participate in stem-cell related phenomena such as proliferation (Androutsellis-Theotokis et al. 2006). Notch signalling occurs through one of four subtypes (Notch1 to Notch4) which are transmembrane receptors, but also using primarily transmembrane ligands for those receptors, meaning that Notch signalling is an overwhelmingly cell-cell-contact-dependent process. Binding of a Notch ligand causes the proteolytic cleavage of the intracellular part of the Notch receptor, which then interacts with recombining binding protein suppressor of hairless (RBPJ) and mastermind-like protein 1 (MAML1) to affect gene expression. Work by (Lowell et al. 2006) notes that mESCs with constitutively active Notch signalling were not noticeably different from wild-type until withdrawal of differentiation inhibitors, whereupon there was mass conversion to a neural phenotype. Their work hypothesises, therefore, that a cell-cell interaction-dependent mechanism is responsible for the observed conversion of mESCs to neural fate on removal of external factors designed to maintain self-renewal.

Phosphoinositide 3-kinase (PI3K) signalling has been demonstrated in mESCs also. Initially, PI3K signalling was found to be involved in the survival and proliferation of mESCs (Burdon et al. 2002). Later, (Paling et al. 2004) found that whilst the inhibition of PI3K did not interfere with STAT3 signalling, it actually enhanced LIF-induced activation of the MAPK/ERK pathway. This may suggest that a role for PI3K exists in inhibiting MAPK/ERK, as inhibition of MAPK/ERK reversed the effects of inhibiting the PI3K pathway. Later work by (Watanabe et al. 2006) demonstrated, in fact, that constitutively active Akt could maintain mESCs in a self-renewing, undifferentiated state in the absence of LIF signalling, again suggesting that PI3K/Akt signalling acts downstream of LIF. Counter-intuitively, this work also found that constitutively active Akt seemed to activate ERK signalling without causing differentiation, supporting the notion that other mechanisms are active downstream of Akt signalling that prevent ERK signalling from driving differentiation. Nanog has also been directly found (by (Storm et al. 2007)) to be downstream of PI3K signalling, through a GSK3β-dependent mechanism. This was shown by observing the reduction of Nanog expression when PI3K was inhibited, which was mimicked by GSK3β inhibition. Blockade of GSK3β mimicked PI3K signalling when PI3K was inhibited, and Nanog expression was restored, strongly suggesting that Nanog expression was reduced by a lack of Wnt signalling, and that this Wnt signalling was itself promoted by PI3K signalling. Lastly, this work demonstrated that while PI3K inhibition was sufficient for the inhibition of Nanog expression and a loss of self-renewal, forced expression of Nanog, even while PI3K was inhibited rescued this phenotype, cementing the link between PI3K to Nanog promotion, very likely via GSK3β inhibition (Storm et al. 2007).

The interconnectedness of the JAK/STAT3, MAPK/ERK and PI3K pathways is investigated elegantly in work by (Niwa et al. 2009), who demonstrated that all three pathways are activated downstream of LIF/gp130 dimerisation and signalling. Here it was shown that JAK/STAT signalling stimulates the expression of Klf4, while PI3K signalling stimulates Tbx3 expression. MAPK/ERK signalling, however, plays its role by inhibiting the expression of Tbx3, downstream of which is Nanog. Klf4's role was found to be mainly to promote the expression of Sox2, but also to promote Nanog. Tbx3's main role was found to be the promotion of Nanog, with a lesser role in promoting Sox2. Sox2 and Nanog can then promote the expression of Oct4. However, Tbx3 and Klf4 are unlikely to be the only members of their respective families capable of maintaining mESCs in an undifferentiated state, as Oct4 expression can be maintained without Tbx3 and Klf4, suggesting built-in redundancy from other factors such as Klf2 and/or Tbx4; this has not yet been investigated. A more recent review of PI3K signalling in ES cells can be found in (Welham et al. 2011).

As a last note on signalling pathways in the context of the work presented in this thesis, unfortunately, whilst microarray analyses such as those used in this thesis can (and indeed, do) repeatedly find enrichment of signalling pathways in changing genes across the spectrum between naïve pluripotency and differentiation, precious little can be inferred about signalling pathways' activities as they most often involve actions such as phosphorylation, nuclear translocation, protein-protein interaction, protein-protein-DNA binding *et cetera* that do not necessarily change the level of mRNA expression of their constituent factors and thusly do not allow microarray analysis to find these interactions / activities. The exact functions and cross-interactions of signalling pathways are therefore unfortunately beyond the scope of purely microarray-based methods such as the ones in this work and remain as future laboratory-based work.

1.4 Epigenetic regulation in mESCs

Whilst epigenetic regulation is known to play a significant role in mESC biology, the only technology used in this thesis is the Affymetrix Mouse 430 v2 microarray, which only detects mRNA transcripts. Therefore this work does not seek to investigate epigenetic markers or regulation as it is beyond the scope of the technology used. The most that epigenetics enters into this work is in the form of chromatin modification / organisation biological pathway enrichments which result from gene lists investigating transcriptional events between naïve pluripotency and early differentiation. For a recent review of epigenetics in mESC biology, see (Morey et al. 2015).

1.5 Primed pluripotency / Epi-Stem Cells (EpiSCs)

Aside from specific mention in relevant genes in section 1.3.3, most of the this introduction has concentrated on naïvely pluripotent mESCs, said to be in the “ground state” (Ying et al. 2008). The “ground state” was actually named after the discovery of mouse epi-stem cells (EpiSCs), as, in seminal work by (Tesar et al. 2007), cells were taken from post-implantation (E5.5) mouse epiblasts and it was found that they could be cultured in conditions similar to those used for human ES cells (hESCs), requiring activation of Activin/Nodal signalling and FGF signalling, and this was also demonstrated by (Brons et al. 2007). The usual culture condition for mESCs (mitotically-inactivated feeder cells plus LIF) did not sustain these EpiSC cells in culture. Furthermore, whilst teratoma formation could be successful on injection with these EpiSCs, all early attempts to generate

chimeric mice from EpiSCs failed, and were noted to even impair development of the embryos into which they were injected (Tesar et al. 2007).

Other features of these cells found at the time included that they expressed similar levels of key pluripotency factors such as Nanog, Oct4, Sox2, but a low level of Rex1, compared to mESCs, was found, along with high levels of Otx2, Brachyury and Fgf5 were found, among others, in addition to the aforementioned differences in signalling pathways required to maintain them (Tesar et al. 2007), (Brons et al. 2007). Rex1 is a marker of naïve pluripotency that is often used to mark mESC samples that have not yet even progressed towards a post-implantation epiblast, EpiSC-like state (Pelton et al. 2002) Work by (Hayashi et al. 2008) noted that in mESC cultures, it was possible to observe difference cellular states that ranged between the ICM-like state towards and epiblast like state (read: naïve to primed pluripotency), with individual cells fluctuating between the two extremes, and highlighted the epigenetic differences between EpiSC-like and more naïve-mESC-like cells in culture. In addition to differences in the level of expression of pluripotency factors, the regulation of transcription was found to be different, including a change in the linkage of enhancers to Oct4 expression, with a distal enhancer of Oct4 associated with Oct4 transcription in mESCs, but there was a pronounced change towards the driving of Oct4 expression from a proximal enhancer in EpiSCs (Tesar et al. 2007).

Crucially, EpiSCs, being reminiscent of a more advanced state of development in the mouse, are also different in that they have undergone (in the case of cells derived from females) X-inactivation, being XaXi, instead of XaXa, as mESCs are (Nichols and Smith 2009).

Later work on EpiSCs by (Acampora et al. 2013) demonstrated the role Otx2 in EpiSCs. Otx2 expression is present in mESCs and opposes self-renewal as it is not only one of the first genes that responds to LIF withdrawal, but also knockout of Otx2 generated a Nanog-homogeneous state in mESCs, as seen in mESCs maintained in the ground state by 2i. Otx2 knockout was even shown to be able to maintain this homogeneous state in the absence of LIF and in the presence of STAT3 inhibition. Conversely, constitutive activity of Otx2 generated a strikingly EpiSC-like state, with high expression of canonical EpiSC markers Fgf5 and Brachyury.

While conversion from mESC to EpiSC can be spontaneous or driven (e.g. by culture in FGF2 and Activin), or by overexpression of Otx2, reversion to the mESC (ground, naïve) state from the EpiSC (primed) state is more problematic. Work by (Guo et al. 2009) used culture in 2i and the

introduction of Klf4 as a transgene to attempt this, although conversion rates were low (around 1%). Klf4 was also used by (Hanna et al. 2009), but this work also found success in reverting EpiSCs to mESC-like state in nonobese diabetic (NOD) mice, known to be refractory to mESC derivation. Klf2 can also revert EpiSCs to mESCs (Hall et al. 2009) and a later screen by (Guo and Smith 2010) identified Nr5a1 and 2 as able to revert EpiSCs to mESCs with greater potency than the other single factors here. Work by (Silva et al. 2009) demonstrated that ectopic expression of Nanog could revert EpiSCs to ground state mESCs in the presence of 2i + LIF (known to kill off EpiSCs), but also in LIF/BMP4 without 2i. The use of more than one factor can reprogram EpiSCs to mESCs with higher efficiency still, as shown by the synergistic action of Klf2 and Prdm14 by (Gillich et al. 2012), wherein Prdm14, although not capable of reversion on its own, greatly enhanced the reprogramming capability of Klf2.

(Bao et al. 2009) also had success in reversion of EpiSCs to an mESC-like state, simply through prolonged culture in the presence of LIF to activate STAT3. Despite taking multiple weeks of culture, mESC-like cells were recoverable which demonstrated all the hallmarks of reprogramming to mESCs, with reactivated X (XaXa), the use of the distal enhancer to drive expression of Oct4 and, crucially, these cells regained the ability to contribute extensively to chimeras.

Very recently, this change to the use of the distal enhancer in the ground state has also been investigated in mESCs cultured in serum + LIF, with the change to homogeneous ground state mESCs being accompanied by the switch to the use of the distal enhancer for Oct4 by 2i conditions (Galonska et al. 2015).

Finally, recently, it has been demonstrated that it is indeed possible to generate chimeric mice using EpiSCs that have been stabilised through the inhibition of Wnt signalling, suggesting that earlier failures to generate chimeric mice may have been down to the production of Wnts by the EpiSCs used in those experiments (Kurek et al. 2015).

1.6 Induced Pluripotent Stem Cells (iPSCs)

One of the fundamental goals of stem cell research is to gain control over cell state to the point where it becomes possible to generate any and all cell types in the laboratory, with a view to their use in a myriad of basic (research) and applied (e.g. clinical) roles. Cells that have been

reprogrammed from a later developmental state, be it fully differentiated somatic cells, or even reprogramming of EpiSCs to mESCs, results in cells dubbed induced pluripotent stem (iPS) cells. Whilst these extremely useful cells are one of the most prized promises of stem cell research, with great advances in both understanding and generating them being made, they are not at all a focus of this work. They are mentioned here, therefore, only briefly, for both completeness, and because there are some iPS samples in the data in this work, and these samples were analysed separately, if only briefly, in chapter 3.

The seminal paper associated with the generation of iPS cells from somatic cells is (Takahashi and Yamanaka 2006), who generated iPS cells from mouse embryonic fibroblasts through the ectopic expression of the now-canonical "Yamanaka factors" Oct4, Sox2, Klf4 and c-Myc. However, extremely important work was done leading up to this, wherein work by (Mitsui et al. 2003) identified 24 transcripts which appeared to be found abundantly specifically in mESCs, by comparing available online data from mESCs to data from more adult tissues. It was from this set of 24 that the groundbreaking work by (Takahashi and Yamanaka 2006) drew its inspiration, by a process of elimination determining that combination of four of these factors (Oct4, Sox2, Klf4 and c-Myc) generated (albeit seldomly) colonies which survived selection. Selection here was obtained by linking the expression of an mESC-specific gene (Fbx15) to a geomycin resistance gene.

Later, however, it was found that such selection does not necessarily select for cells that are truly mESCs by the definition that they can contribute to chimeras, as was demonstrated by (Okita et al. 2007). This work changed the selection criterion from the previous Fbx15 to Nanog, which resulted in the recovery of cells that had improved doubling time and where at least some clones contributed to chimeras. Furthermore, DNA methylation was closer to that of mESCs than the Fbx15-selected cells, with promoter regions of Fbx15, Nanog and Oct4 lacking methylation.

The full history of iPS generation technology is not relevant to this work, and yet, no comment on induced pluripotent stem cells would be complete without the mention of non-transgenic methods for creating iPSCs, as a critical barrier to clinical application is that the use of viral vectors and, indeed, any method of iPSC generation that modifies the genome will face extremely stiff resistance to being used in any downstream application. Therefore brief mention is made here that there have been successful generations of iPSCs without the use of viral vectors or modifying the genome in any way. The use of non-integrating viruses to generate iPSCs was shown by (Stadtfeld et al. 2008) and, on the very next page of the same journal, a non-viral approach using an inserted plasmid by

(Okita et al. 2008). Totally transfection-free reprogramming has been achieved even in human cells by appending cell penetrating peptides (arginine repeats) to reprogramming factors by (Kim et al. 2009). A somewhat recent review of advances in iPS generation and application is available in (Robinton and Daley 2012).

1.7 The Promises of Stem Cell Research

Stem cells have the capacity to self-renew and give rise to more differentiated progeny. When considering that embryonic stem cells have the capacity to generate all of the tissues of the adult organism (Evans and Kaufman, 1981), it becomes clear that the applications of mastery over embryonic stem cells are varied and exciting, with great promise for several areas of biomedical research (Robinton and Daley 2012). Some of these areas, with selected examples from the literature, are mentioned here to give frame of reference to why improving understanding of transcriptional events in mESCs, such as is the focus of the larger, biology-oriented part of this work, contributes to the wider advancement of stem cell research.

Regeneration of tissue

Regeneration is an ongoing process to counter general accumulation of damage / wear; even part of aging itself is the simple slowdown in cellular turnover and accumulation of senescent cells (Jeyapalan et al. 2007). As tissue becomes damaged / old or otherwise in a state of disrepair / reduced homeostatic resilience or otherwise reduced functionality, pathology is the natural result. Understanding of the mechanisms of stem cell biology can grant the ability to induce regeneration in muscle where normally stagnation or deterioration would occur (Conboy and Rando 2005). With improved understanding of events in regenerative processes also comes the ability to induce that regeneration. In this vein, even adult tissue also demonstrates a remarkable ability to undergo a form of reprogramming and re-differentiation, but only when given appropriate signals, such as was demonstrated with wound-induced induction of Wnt signalling to generate skin appendages (hair) *de novo* (Ito et al. 2007), (Celso et al. 2004).

Generation of replacement tissue

Possibly one of the most exciting promises of embryonic stem cell research is the ability to generate tissues and / or organs for use in clinical settings (Yoshida and Yamanaka 2010), (Taylor 2009). With the demand for organ donations outstripping available supply, the ability to generate both organs and even just functional tissues brings with it the promise of alleviating and eventually eliminating the need for donor organs. Organ transplantation also carries with it the risk of rejection, often requiring organ recipients to adhere to regimens of immunosuppressant drugs, which carry their own inherent risks of increased susceptibility to infection. Through the reprogramming of somatic cells, the generation of iPS cells offers hope for the replacement of tissues, and eventually whole organs as technology and technique improves (Nishikawa et al. 2008). All of these medical marvels begin with basic research and are sped by developments in understanding the mechanisms that drive pluripotency, exit from pluripotency and differentiation to desired lineages.

Applications to drug testing / discovery

The ability to generate tissues / organs in the laboratory does not only have direct implications for the implantation of those tissues for the purposes of replacing lost or damaged tissue. Generation of specific cell types from iPS cells could provide an effectively limitless supply of differentiated model organism (e.g. mouse, rat) or human cells pertaining to the tissue / organ in question, speeding testing of drug testing and / or discovery. For a recent, comprehensive review, see (Inoue and Yamanaka 2011).

Applications in cancer

Embryonic stem cells have been noted to share similarity with cancer cells, particularly with their replicative capacity and proliferative vigour, to the point where the field of cancer research has taken on board, as well as found the existence of, “cancer stem cells”, which, with their defining trait of immortality, appear to maintain cancers (Clevers 2011). Bringing the fields of cancer research and basic stem cell research together is already opening promising avenues in advancing understanding of, and thus expediting improved treatments for, cancer (Scatena et al. 2011).

Disease modelling

The promise of iPS cells to grant researchers the ability to generate differentiated cells mimicking somatic tissues of the adult body presents opportunity to generate cells / tissues specifically to model disease processes, such as through genetic modification of the resulting cells / chemical insult *et cetera* (Tiscornia et al. 2011), (Bellin et al. 2012). iPS cells are already being harnessed to provide models and enhance understanding of diseases such as Miyoshi myopathy (Tanaka et al. 2013) and vitelliform macular dystrophy (Best disease) (Singh et al. 2012).

1.8 Brief Overview of Transcriptomics

The analysis of the mRNA content of a cell or population of cells is transcriptomics. Following on from the sequencing of an organism's genome, transcriptomics can give us large amounts of biologically-relevant knowledge by determining the amount of mRNA for a particular product, be that a protein or be it so that the mRNA is the product itself, in the case of untranslated “non-coding” mRNAs. This, in turn, is used to study the state of the cell or cell population by taking a snapshot of the overall transcriptome (literally, the quantities of which mRNA transcripts are present.) Experiments can therefore be designed to investigate the effect of a plethora of combinatorial factors (e.g. cell type, cell state, timepoint, treatment, disease state) on gene expression, which is usually predicted by, but not identical to, mRNA expression (Gygi et al. 1999).

1.8.1 The Microarray

There are many technologies used for the study of transcriptomics, however the only technology relevant to this thesis is the microarray. Microarrays have the ability to provide information about the transcriptional profile of a sample of cells, a function of great utility in a wide range of biological research roles (Ekins and Chu 1999), (Brown and Botstein 1999).

For being able to carry out, in greatly multiplexed fashion, simultaneous analysis of thousands of genes’ expression, microarrays are considered to be lab-on-a-chip technologies. Physically, a microarray is a chip synthesised by methods such as photolithography onto a quartz glass base (in the case of Affymetrix arrays), onto which many thousands of oligonucleotide probes are attached. These probes are designed to be complimentary to parts of mRNA sequences expected to be found in the cells of the organism for which the chip is designed to be used.

Following reverse transcription of mRNA from the biological sample and subsequent biotin labeling to give biotinylated cDNA. These biotinylated cDNAs then undergo in-vitro transcription (IVT) in order to convert these into biotinylated cRNAs. It is these biotinylated cRNAs (which are anti-sense compared to those original mRNAs from which they were generated) that therefore are able to bind by complimentary base-pairing to their matching oligonucleotide probes on the chip. Antibody staining is then used to complete the microarray preparation steps before they can be scanned by confocal laser to quantify binding. Quantification of the fluorescence resulting from laser excitation is analysed by proprietary computer imaging software to estimate the amount of matching mRNA present at each particular probe.

For a given region of the genome (read: expected mRNA transcript), several (25mer) oligonucleotide probes are designed which map to regions in close proximity of one another. These probes are grouped together in a single “probe set” which is then, at a higher level of abstraction, often connected to a particular gene or microRNA. For example, there may be 11 oligonucleotide sequences which map to locations all in the first exon of a gene. This probeset would be associated with that gene, but there may be other probes which are also associated with that same gene. Therefore it is not only possible to test for the presence of mRNAs coding for a certain gene or untranslated genomic region, but we may have several probesets, of many oligonucleotide probes each. This can give rise to improved detection, for example in the case of a probeset which may later become known to not give adequate detection or may be prone to binding cRNAs from another gene.

In the case of the microarray type used in this work, the Affymetrix Mouse 430v2 array (see section 1.8.2), there are both perfect match (PM) and a mismatch (MM) probes for each oligonucleotide sequence. This is not a feature of all microarrays. The MM probe has a single nucleotide in the middle of the oligonucleotide sequence replaced with its complimentary base. This MM probe is included in order to test for the aforementioned cross-hybridisation (non-specific binding) which may occur as a technical issue and thus provides us with a measurement of “background” against which actual mRNA binding signals may be contrasted.

The choice of microarray for this work (the Affymetrix Mouse 430 v2 chip) was guided by the numbers of mouse microarray samples available in public, online repositories for download (see section 1.8.2). Consulting the GEO website and sorting the available data in mouse by platform (read: chip type), showed that the Affymetrix Mouse430v2 array (platform code GPL1261) was the

dominant platform, being the most widely-used array for the mouse, with 34,037 uploaded samples at time of writing, this being more than 3 times the number of samples in the next most-used mouse chip (MoGene 1.0 ST Array, platform code GPL6246), with only 10,275 uploaded samples. This made platform GPL1261 a clear choice for this work.

1.8.2 Online Repositories of Microarray Data

Two prominent repositories of microarray data are the European Bioinformatics Institute (EBI)'s ArrayExpress (https://www.ebi.ac.uk/arrayexpress) (Brazma et al 2003), (Parkinson et al. 2007) and the US National Centre for Biotechnology Information (NCBI)'s Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo) (Edgar et al. 2002).

These repositories came into existence in response to the burgeoning amounts of microarray data being generated by research groups worldwide in an effort to both store, catalogue and make them available for other groups' use (Edgar et al. 2002), (Brazma et al 2003).

These repositories offer both graphical, manual browsing and downloading via their websites, but also provide programmatic access for search and download (e.g. by file transfer protocol (FTP)) directly. These repositories hold many thousands of microarray samples from a plethora of species and different microarray technologies. Users can browse by organism, platform (read: individual microarray chip design) or enter search terms to look for samples of interest to them. Programmatic access and scripted FTP download was used in this work.

In the case of GEO, a system of platforms, samples, and series is used. The term "platform" describes the specific subtype of microarray chip used (GEO uses "GPL1261" as the platform code for the Affymetrix Mouse 430v2 chip used in this work), and "sample" intuitively refers to one specific instance of that platform, with an actual biological sample's cRNAs hybridised to it and subsequently quantified. The term "series" needs a little more explanation; in GEO, a "series" is a set of samples which are grouped together as forming a particular experiment and so concerning a particular question (e.g. the response of mESCs to a particular agent in culture at varying concentrations.) Individual samples begin with the code "GSM", while series begin with the code "GSE". It is worthy of note that different samples may fall under the umbrella of more than one series, if that sample pertains to more than one experiment or question. Programmatically

downloading large amounts of data, such as in this work, must take this into account, for example by removing duplicate samples that may otherwise result from retrieval of experiments of interest, given that different experiments ("series") may contain overlaps of samples.

ArrayExpress is similar in that individual arrays are grouped under experiment numbers (albeit in a different naming format) and platforms are also categorised (again, with different codes such as "A-AFFY-45" for the Affymetrix Mouse 430 v2 chip.) ArrayExpress codes begin with "A" for array designs and "E" for experiments. A hyphen and four character descriptor follows these codes, detailing the source of the data (other database and certain institutes) or its annotation standard. There is such acknowledged overlap in the samples between GEO and ArrayExpress, that ArrayExpress has a code "E-GEOD" for data that is outrightly imported from GEO itself. One notable, difference, however, is that with ArrayExpress, protocols are given accession numbers also. Full details of these codes and their meanings are given and kept up-to-date at:

http://www.ebi.ac.uk/arrayexpress/help/accession_codes.html

It is worth noting here that some of the codes (e.g. "E-MEXP-" codes) are used to denote the way in which annotations were given to ArrayExpress, and as annotations are crucial to the work carried out in this thesis, special mention is given to the MIAME / MAGE-ML standards in section 1.8.3.

Both the GEO and ArrayExpress samples are annotated in different ways, to individual sample level, and several different formats and stages of analysis are available for download. Annotations from GEO are given in files in SOFT format, detailed on the GEO website. These files provide both summarised expression values, but also contain fields to detail experimental summaries and associated publications. These fields, as can be seen from sections 2.1.2 and 2.1.3, are often incomplete, missing, ambiguous or otherwise do not facilitate retrieval of full experimental details. Crucially, certain experimental details such as genetic modification of cells as well as culture condition are not explicitly present in current annotation systems, including MIAME (see section 1.8.3), necessitating the manual search carried out in chapter 2.

ArrayExpress annotations are available for each experiment as text (.txt) files. These include an "investigation description", "sample and data relationship" and "array design." The array design is a brief description of the array technology itself and a list of Ensembl transcript identifiers that this platform probes for.

The sample and data relationship file provides basic information on experimental design, such as which samples form a treatment or control group, protocol codes and details which samples occur in which data archives.

The investigation description has space for a written summary of the experimental design and objectives, associated literature, upload and update dates, and protocol codes that apply to the experiment. ArrayExpress' protocol codes are a noticeable departure from the formatting of GEO, in that every sample in an experiment can have several protocol codes attached to it. These codes are prefixed with a "P-", e.g. "P-MTAB-53395". This example was chosen as it is from an experiment which dealt with human pluripotent stem cells (E-MTAB-5367) to demonstrate that even experiments uploaded in 2015 (and further updated in 2017), many years after the establishment of standards such as MIAME (see section 1.8.3), microarray experiment annotations often still do not contain usable details of experimental details such as culture condition. This protocol code is, in fact, the chosen experiment's "growth protocol", and simply informs the reader that the "Cellartis® Definitive Endoderm Differentiation Kit with DEF-CS™" system was used. This does not contain information on actual culture conditions and still, therefore, requires researchers to perform significant manual work to obtain culture condition information.

1.8.3 Microarray Annotation Standard MIAME and MAGE-ML

With the volume of microarray data being produced by laboratories around the world, the problem of ensuring a meaningful, standardised way of annotating microarray data arose. A group, the Microarray and Gene Expression Data society (MGED) came together in 1999 in response to this and, in 2001, a proposal was put forward to ensure that publicly-available microarray data would contain certain minimum information, and have that information provided in such a way as to facilitate its cataloguing in repositories (Brazma et al. 2001). This standard was named after its intended purpose; to ensure the provision of the minimum information about a microarray experiment (MIAME.) The initial standards put forward in this initial publication were very much up for discussion and, in fact, the publication itself threw the question out to the research community as a whole, but suggested six core fields.

These six fields were 1) experimental design, 2) array design, 3) sample information, 4) hybridization information, 5) the measurements themselves and 6) information on control methods used. Briefly, experimental design is a description of the experiment itself, with its aims and overall structure. Array design details all arrays used in an experiment and, for each array, gives the layout of the chip and locations of features upon it. Whilst this may seem to warrant the inclusion of a huge amount of information, MIAME allows, and indeed suggests, that this be done by way of naming commercial manufacturers, so long as there is provision for retrieving the full details of the array from the manufacturer. Custom arrays would still need to be detailed fully. Thirdly, the sample information itself should contain the details of each biological sample, including its source / cell line with any treatments / modifications applied to it, nucleic acid extraction and labeling. Fourth is a section detailing the method / conditions used to perform the hybridisation. Fifth come the actual measurements themselves, and MIAME divided these originally into raw image scans of the arrays, detail on image quantification and then the gene expression matrix itself. Finally, MIAME specifies its sixth section as details of normalisation controls. This includes information about the use of inter-array normalisation methods such as normalising to the expression of housekeeping genes or the use of spike-in data.

Whilst MIAME, as a proposal, suggested these data fields as being required and pushed for their adoption and requirement for acceptance of microarray data submission as well as for publication / funding, MIAME did not intend to specify in what format those details should be provided. This was instead left up to future discussion and refinement based on observing the trends in submission formats as they came in. A term used by the authors refers to “controlled vocabularies”, being defined sets of accepted, standard words or tags used to describe data annotation features. As the area of microarray annotation was young at the time, it was likely wise to postulate the utility of controlled vocabularies without trying to define any with the limited data submissions that were available at the time.

An update to the use of MIAME occurred just a year later, in 2002, (Spellman et al. 2002) with the advent of microarray gene expression markup language (MAGE-ML) which took the concepts of MIAME and sought to create an extensible markup language (XML)-based document type definition (DTD), setting out both classes and child structures within those classes to hold information on microarray experiments to satisfy MIAME standards. MAGE-ML itself is based on the MAGE object model (MAGE-OM), which is the underlying structure that was designed to store relevant experimental details to MIAME standards. These data fields would then be populated by

end-users performing microarray experiments, ensuring a standardised set of informative annotations, sufficient to communicate, re-analyse and repeat investigations.

The syntax of the MAGE-ML adheres to a limited set of rules. The main packages defined my the MAGE-OM contain sets of related classes. An exhaustive list of these is given as table 1 in (Spellman et al. 2002). A class can be either a physical entity (such as a microarray) or a process (such as hybridization). Every class has a list of attributes attached to it. MAGE-OM classes also contain relationships to other classes. (Spellman et al. 2002) provide an example of a class of type “Person”. This given example makes MAGE-OM's complex abstraction clearer:

<Person identifier=”Person1” name=”John Doe”>

<Affiliation_assnref>

<Organization_ref identifier=”ABC Inc.” />

</Affiliation_assnref>

</Person>

This defines a person, an instance of the class “Person”, and assigns the property of “name”, setting it to “John Doe”. This person has the attribute of their “Affiliation”. MAGE-OM rigidly enforces that tags sugs as a suffixing “assn” for something that is an association of that class (in this case it can be read that a person is literally associated with a given organization.) The further, concatenated suffix “ref” denotes that the association itself is by reference. From the above example it can be seen that what is “referred” to is in fact given here as “ABC Inc.” . Each tag, as is the format of UML-based languages such as XML, enclosed in inequality symbols “<” and “>”, with the opening of a section given just a name (e.g. “<Affiliation_assnref”) and the end of this section denoted with the same name, preceded by a terminating slash (e.g. “</Affiliation_assnref>”) common to other markup-style languages such as the ubiquitous HTML. The complexity of MAGE-ML is therefore considerable, given that there are 132 classes organised into 17 packages.

A noteworthy mention should be given, however, to the package “BioAssayData”, which holds the results of the microarray experiement(s). In this package, the class “BioDataCube”, said to be a

cube as this class has three dimensions: 1) an axis for what is being measured (e.g. individual probes) called (DesignElements), 2) an axis ("QuantitationTypes") for the types of value measured / calculated for that probe such as intensities and error values and 3) an axis made up of all of the microarrays (each being a 2D slice of the cube).

Whilst MIAME only stated required information and gained acceptance in the microarray community for setting out these provisions, MAGE-ML, with its 132 classes in 17 packages was acknowledged to be prohibitively complex for laboratories without dedicated informatics support able to decipher their meanings, intended contents and the methods to populate the data fields for submission (Rayner et al. 2006). Thus was created MAGE-TAB, a simple, spreadsheet-based method of compiling and exchanging microarray information to MIAME standards (Rayner et al. 2006). This work even candidly puts forward in it's abstract that "the complexity of MAGE-ML format has made its use impractical for laboratories lacking dedicated bioinformatics support".

One of the many acknowledged weaknesses of the MAGE-OM / MAGE-ML concept was that it was entirely possible to encode the same information several different ways using MAGE-ML , and required the creators of MAGE-ML to begin releasing advisory documents on how to avoid such issues (Rayner et al. 2006). This work also acknowledged the greatest weakness of the MAGE-ML system was the "complexity of the MAGE-ML files, making it difficult to interpret or produce MAGE-ML files in the absence of a dedicated software development effort." The same update that brought MAGE-TAB also admits that MAGE-ML would still require as-yet-undeveloped text mining approaches in order to facilitate any kind of automatic analysis of microarray annotations. Having seen the simpler, tab-delimited SOFT format (used by GEO at the time), MAGE-TAB was clearly easier to use and adoption of this was far greater. MAGE-TAB has four types of files, 1) the investigation description format (IDF), 2) array design format (ADF), 3) sample and data relationship format (SDRF) and 4) raw and processed data files. These tab-delimited descriptor files are easily interpretable by eye, even without reading (Rayner et al. 2006) and are easy to read into spreadsheet software packages (e.g. Microsoft Excel, LibreOffice Calc) used by non-specialists, as well as not requiring and special packages / proprietary code in order to be easily read into more complex software such as R. With the increasing focus on "big data" and large meta-analyses coming to the biosciences, this removes a vast barrier to the effective use of microarray annotation information.

Finally, relating this to the present work, despite repeated mention in the literature from the time of MIAME to MAGE-TAB, of the crucial importance of recording detailed experimental conditions, there is (including currently) no specification in MIAME that stresses the need for a “controlled vocabulary” for data tags such as genetic modification, presence of reporter constructs, cell lines' constitutive expression / knockout / knockdown of particular genes. Neither is there requirement for such a controlled vocabulary or data-mineable syntax to chronologically record experimental, detailed in-vitro culture conditions and exposure times. This puts crippling constraints that this puts upon automated meta-analysis (and, often, manual attempts at meta-analysis) of microarray data. Comment on this and details of this work's creation of a simple, spreadsheet-style annotation syntax to include and make mineable all of this information is given at length, therefore, in chapter 2.

1.8.4 Microarray Data Processing Overview

The microarray chip itself presents fluorescence as its quantification of mRNA detection. Fluorescence Is first measured through proprietary scanning machines (in the case of this work, mostly an Affymetrix device), and this is output as a raw image as a computer file given a “.DAT” extension (suffix). This is then processed by proprietary Affymetrix software into a “.CEL” extension file. The processing here is primarily to quantify the detection of fluorescences into numerical format with some further information relevant to downstream processing (for example the number of pixels of image data used to calculate that individual probe's intensity.)

This is the point at which data was retrieved for this work. A CEL file contains all raw data necessary for the analysis of microarray data and, given that it is a standard format for each type of Affymetrix microarray, it was possible to simply compile relevant samples from large online repositories of microarray data in order to make comparisons, observations and useful inferences.

Further downstream, however, processing is still required to combine individual oligonucleotides into their probe sets and summarise them and further processing again required in order to render comparable different individual samples (each comprising, in the case of the Affymetrix Mouse 430v2 array, the most commonly used array in this work, 45,101 individual probe sets).

The processing for this task is neatly encapsulated in the tools provided by the Bioconductor project (Huber et al. 2015), which, in turn, is based on the highly-flexible and widely-used statistical programming language, R (R Team, 2015).

Using the R language and the Bioconductor packages, data used in this work was processed into a final, usable form by combining the desired microarrays (as their CEL files) and using the robust multi-chip average (RMA) (Irizarry et al. 2003) method of generating a matrix of numerical outputs for each probeset in each sample. RMA is a widely-used technique which has become something of a gold-standard to compare to in the area of microarray analysis. RMA makes microarrays comparable in a meaningful manner by first subtracting out background from each sample (without actual resort to the mismatch probe data), then ensuring that the intensity ranges of the measurements across all samples are brought into line by quantile normalisation and finally, summarisation and log2 transformation (Irizarry et al. 2003). This deals with technical issues that could otherwise cloud biological signals. Dealing with such technical issues is imperative as, for example, in the case of a specific experiment yielding slightly more or less mRNA for use on the chip, a generalised change in all probe intensities would be observed, which would, in turn, be entirely misleading if treated as directly-comparable intensities (this is fixed by quantile normalisation.)

1.8.5 Meta Analysis / Multidimensional Analysis of Microarray Data

With the explosion of available microarray data that has occurred in biological research, the possibility of large-scale meta-analyses became a reality. Meta-analysis refers to the combination of multiple datasets in order to attempt to gain extra insight that could not be seen (or was not robustly-observed enough) from the smaller datasets that make it up. Meta-analyses also, by their definition, can be used to observe differences between different datasets. Meta-analysis is therefore a large, umbrella term and requires defining in the scope of this work. In the case of this work, the focus was on the potential for finding transcriptional signatures that may or may not exist between cell lines represented in the data amassed.

When attempting to discover patterns in complex, high-dimensional gene expression data, analysis approaches are often built by starting from simple distance metrics or variability scores and then applying these in dimensionality reduction methods (discussed later in this section.)

One such simple distance metric used in this analysis is euclidean distance. Euclidean distance measures the distance between any two datapoints in euclidean space. In the case of microarray

data, each array can be thought of as a vector of datapoints. To compare one array's intensity values to another's by way of euclidean distance is simply to compare one vector to another. As each probe is present in each microarray (in this work, at least, which uses only the one array design), euclidean distance can be calculated between two arrays X and Y by:

$$d(X,Y)=\sqrt{(X_1-Y_1)^2+(X_2-Y_2)^2+\ldots(X_n-Y_n)^2}$$

Here, X_1 and Y_1 refer to the first feature (read: probe) of the arrays X and Y respectively. A simple subtraction of the values between corresponding datapoints is performed for all probes of X versus all probes of Y. These values are squared and summed and the root of this sum taken.

This is a distance measure commonly used in simple pattern-visualisation / data grouping analyses such as the ubiquitous hierarchical clustering used in bioscience today. Euclidean distance is used in this work as a way to quantify dissimilarity / similarity between groups of microarrays taken together, where euclidean distance was used pairwise; an all-against-all comparison of every microarray in a group versus all others in that group, then normalised to the size of the group of microarrays by dividing the summed pairwise euclidean distance by the number of microarrays, to give a result which expressed a general dissimilarity measure of the microarrays in the group.

In addition to measures of distance, a measure of variability is often used in meta-analyses, quantifying the amount by which a given variable varies. This measure, and the related measure of covariance, are not used *per se* in this work, but are often used as the basis for techniques mentioned later in this section for dimensionality reduction. Briefly, variance is calculated by first centring all values in a vector to their mean by dividing every value of the vector by the mean. The difference (literally a subtraction) between the new values and the mean is squared, and the mean of these squared numbers is the variance. Covariance expands this to examine how two variables vary together using the same method, but combining corresponding datapoints of each vector to quantify how one variable varies with another. Variance is mentioned here as it is a simple metric that is used as part of principal component analysis (PCA) mentioned later in this section.

Meta-analyses become ever more complex when the datasets collected have a large number of variables (alternately called “features”), taking such analyses into the realm of large-scale, high-dimensional analyses.

A microarray is, itself, a high-dimensional entity, with a large number (45,101 in the case of this work) of probes providing simultaneous measurements (see section 1.8.1). As each microarray contains a snapshot of a transcriptional profile of a sample, the simultaneous intercomparison of a great many transcriptional profiles has been an approach that has generated valuable insights, particularly in the case of highly-complex conditions such as cancer (Rhodes et al. 2004), (De Cecco et al. 2014), (Clarke et al. 2008).

Current, prominent methods that can find and attempt to meaningfully display patterns and groupings in such data do so by literally reducing this extreme dimensionality of the data and so are duly called “dimensionality reduction” methods. Briefly put, dimensionality reduction methods reduce the number of variables in data by generating a few new, summarising variables which capture the variation in the data whilst keeping any information loss to a minimum. One such dimensionality reduction method, used widely in meta-analysis of gene expression data, is principle component analysis (PCA) (Massy et al. 1965).

PCA functions by finding the eponymous “principle components” of the data. Briefly, PCA considers all variables to be features. In the case of microarray data, these features will be the probes themselves representing cRNA detection. Principal components (PC) are essentially orthogonal mathematical vectors with specific geometric direction and magnitude which explain the variance within data, considering all features simultaneously. Typically, the first PC explains the most variation in data and last PC explains the least. A PC can be traced back to original influencing features by representing their contribution (called “loadings”). PCA analysis has proven to be extremely useful in biological data for summarising and visualising large numbers of samples and PCA plots are a common sight in publications dealing with, for example, visualising groupings of samples (including microarrays) together without prejudicing the algorithm towards and kind of “desired” correct answer.

However, PCA has its drawbacks with respect to particular kinds of analysis. Despite its excellent utility in grouping datasets such as microarrays, the complex nature of the algorithm means that the principle components it identifies cannot be directly translated back to meaningful variables such as gene expression values. That is to say that whilst PCA may well confirm to the user the groups that appear to exist in their data, PCA is not designed to, particularly in the case of gene expression data, give the user a meaningful list of genes and relative expression values which define and separate the

groups in their data. This is the most fundamental drawback of the PCA method which prevents its use in this work. This work specifically desires to extract and quantify changes in gene expression between groups as part of its objectives (see section 1.9). PCA can provide only “loadings” of different variables used to define its principle components, and therefore is unsuitable. Other dimensionality reduction algorithms similar to PCA also are unsuitable for this work for the same reason and are thus excluded from discussion here. In fact, this relative inability to derive biological meaning from the results of dimensionality reduction, particularly PCA has been noted in the literature (Hibbs et al. 2007). This is not to say that dimensionality reduction is not without merit for biological meaning at all; far from it, in fact PCA analysis has been the driving method behind several efforts in the stem cell field and include work which concerned cell lineage along differentiation paths (Aiba et al. 2009). However, again, PCA is not designed to answer the transcriptomic signature questions posed in this work.

Though other methods for finding patterns in high-dimensional data exist, ranging from older methods such as multidimensional scaling (Torgerson 1952), which is a linear transformation method like PCA, to newer, more complex, machine-learning based approaches such as t-distributed stochastic neighbour embedding (t-SNE) (van der Maaten & Hinton), an exhaustive list outlining their mathematical methods would not serve this work, as the same drawback applies. This is not a failing of these methods at all; they simply are not designed to answer the question put forth in this work.

In summary, whilst it may be possible to efficiently observe gross patterning of high-dimensional data using methods such as PCA and others that are based on similar principles these methods are not able to fulfill the objectives of this work, specifically regarding the assessment of an annotation’s “contribution to sample similarity”. Also, these methods do not perform or facilitate measurement of amplitudes of changes in expression of specific genes which can be, with quantified statistical significance at the gene level, attributed to a specific annotation, such as a particular cell line. As one of the core objectives of this work is to attempt to discern transcriptional signatures of mESC cell lines, methods which share core approaches were developed to answer both the “annotation contribution to samples similarity” and “annotation-linked gene expression signature” questions.

1.9 Research Objectives Overview

The work presented in this thesis is motivated by several questions concerning mouse embryonic stem cell (mESC) biology, in conjunction with the availability of large amounts of public microarray data.

Briefly, the work in this thesis addresses first the feasibility of conducting large-scale analysis of publicly available mESC microarray data. As pluripotency is an extremely promising property of mESCs (with a view to eventual translation to human work) (Nishikawa et al. 2008), (Robinton and Daley 2012) (and see section 1.7), this work assembles the largest to-date number of microarray samples selected for annotation as mESCs that also show the highest levels of expression of three canonical markers of pluripotency, Oct4, Sox2 and Nanog.

Whilst such large and heterogeneous datasets are known to be useful in and of themselves in biological research (Quakenbush 2001), (Ramasamy et al. 2008), this work improves the utility of this dataset by developing an informative, easy-to-use annotation system for recording vital information about every microarray that makes up this dataset. Full, manual annotation of every sample is then carried out in order to maximise utility of this new dataset of high-pluripotency-marker (HPM) mESC microarray samples. The existence of this dataset immediately makes possible the two avenues of investigation which are carried out in this work in chapters 3 and 4.

The first avenue of investigation concerns whether or not it becomes possible, using the manual annotations of the generated HPM matrix to ask questions of the contribution to sample similarity of two key annotations – those of cell line and source laboratory. Laboratory-specific transcriptional profiles have already been found to exist in mESC data (Newman and Cooper 2010). Other work has shown differences in human iPS lines appearing to represent what may be “memory” of the donor material (Marchetto et al. 2009). This dataset, and therefore the work presented herein, can therefore offer the first attempt on this scale at analysing relative contributions to sample similarity of cell line annotations and source laboratory annotations. This work does not seek to remove the effects on transcriptional profile of a samples being from one laboratory / cell line or another, but offers a first comparison of the relative contribution of these annotations to sample similarity using a method that has three key features. The method devised for this is named RaSToVa, as it uses Random Submatrix Total Variability. The first of this method's key features is the lack of requirement of advanced statistical understanding on the part of any researcher using or interpreting

the results of this method. Secondly, the method devised uses publicly-available data in a format that is available to biological researchers without the need for any preprocessing other than annotation. Thirdly, the method devised uses only freely-available software and the use of scripting, meaning that this method can be carried out by any laboratory with sufficient computing power without needing to pay for any proprietary software, and using a method that is understandable to non-statisticians / bioinformaticians.

An additional possibility that arises further to the completion of full manual annotation of the HPM matrix is the ability to mine deeper than “sample similarity” in this data to attempt to identify transcriptional profiles that appear to be associated with specific cell lines. No microarray analysis of public data has yet investigated the differences between commonly-used mESC cell lines on this scale, while the work presented here attempts this using a novel method dubbed DALGES – Discovery of Annotation Linked Gene Expression Signatures.

Another major question that is asked in this work is whether or not it is possible to use this manually-annotated HPM matrix to investigate potential changes, at a transcriptional level, in high-pluripotency-marker mESCs between naïve pluripotency, primed pluripotency and exit from primed pluripotency. No analysis on this scale of these transcriptional profiles across these cellular states has yet been carried out on mESC microarray data.

As this work progressed, it became clear that such an analysis was a very real possibility, given that the data matrix was found to contain samples which, by both annotation and by transcriptional marker profile, appeared to cover this broad spectrum of cellular state from naïve pluripotency to the exit toward differentiation from primed pluripotency. In order to investigate transcriptional events, as well as higher-order biological pathway activities across these cellular states, it is necessary to, in some way, meaningfully arrange the data in such a way as to provide a progression between the cellular states of interest. This was achieved in this work through the careful selection of a candidate gene for sorting the data. After confirming the successful broad sorting of the data between naïve pluripotency and exit from primed pluripotency toward differentiation, both by marker profile and by ranked annotation referencing, it became possible to interrogate the data as to both the transcriptional and biological pathway changes between these cellular states so crucial to the utility of mESCs.

This thesis therefore asked what major changes can be seen between these cellular states using a purely informatic approach. Two approaches were used, the first being the common “differential expression” method between different sections of the data found to be, by marker profile, representative of the aforementioned cellular states, but also a scanning window approach was devised, calibrated to the data and used to attempt to improve upon the results of standard differential expression analysis. This greatly improved the detection of transcriptional changes and biological pathway enrichments, most strikingly across the region of the data whose samples bore markers of naïve pluripotency, identifying signalling pathways and changes in gene expression which may indeed be able to separate “early” and “late” naïve pluripotency. Whilst the use of a purely informatic approach prevents the findings from this work being taken as direct proof of biological reality, a set of novel markers of this putative “late naïve pluripotent state” which is proposed to exist immediately prior to the switch from naïve to primed pluripotency, are reported. The scanning of the data for transcriptional changes and biological pathway enrichments culminated in a full-matrix scan using the scanning window approach which successfully identified nearly every other biological pathway enrichment of interest to mESCs found by the smaller analyses in this work, warranting the use of this method in other datasets that can be broadly sorted across a biological phenomenon of choice.

Finally, proof-of-concept is demonstrated in this work wherein the scanning window method's results can be used to visually mark changes in expression of significantly-enriched biological pathway genes across the sorted data matrix, with intent for use in guiding future work on the chronology of transcriptional events in mESC (and, later translation to hESC) pluripotency and differentiation.

Chapter 2 –

Assembly and Annotation of a Matrix of High-Pluripotency-Marker mESC Microarrays

2.1 Research Questions

2.1.1 Do sufficient numbers of publicly-available mESC-annotated microarray samples have high Oct4, Sox2 and Nanog for the generation of an HPM matrix?

Within this first research question here to be answered is whether or not it is possible to amass a large number of mESC / mESC-like microarray samples from public data sources prior to even filtering for high levels of expression of pluripotency markers.

The ability to source data is not in question, as the use of public data is a commonplace affair. However, whether or not a sufficiently-large number of mESC samples exist to enable truly large-scale analysis is not certain. The second part of this question to be addressed in this chapter is whether or not, of those samples which may contain online annotation for the terms “embryonic stem cell”, do a sufficient number of those samples will carry a pluripotent signature, as marked by highest expression of the pluripotency markers Oct4, Sox2 and Nanog (see section 1.3.1)? The reason behind the filtering for these pluripotency markers was two-fold.

First and most obviously is the desire to study the pluripotent state which, in the absence of the ability to test for functional pluripotency, pluripotency markers are the closest way that a bioinformatic approach has (in conjunction with checking sample annotations) of ensuring that the samples gathered do in fact pertain to the pluripotent state.

Secondly, but no less importantly, it is expected that an automated search of the online repositories for microarray data will undoubtedly retrieve some samples which contain the appropriate search term (being “embryonic stem cell”) but may not in fact be ES cells at all. It may be, for example, that the description of the uploaded samples mentions that phrase in a secondary manner, such as “these cells are known to have a slow turnover and were used to investigate the influence of gene X on proliferation rate, as this is known to be involved in the proliferative vigour of other cell types such as the embryonic stem cell.”. It may also be that the annotations are simply incorrect, or copied wholesale across all samples of an experiment, some samples of which are embryonic stem cells and others are cells to which the ESCs are being compared. Filtering for highest Oct4, Sox2 and Nanog should go a long way to eliminating these samples from the analysis in this work as the transcriptional profile of the sample, not its annotation, is the ultimate indicator of its state.

After observing the number of samples downloaded prior to filtering for pluripotency markers, distributions of the expression of each of the chosen pluripotency markers (Oct4, Sox2 and Nanog) can be ascertained. From these can be derived a reasonable cutoff / threshold for filtering for only those samples which are highest in the expression of all three. This filtering can then answer the second question: whether or not there are sufficient samples within the first dataset which are high for the selected markers of pluripotency (Oct4, Sox2 and Nanog.)

As one of the main advantages of a large-scale analysis is to piece together sufficient samples so as to observe subtle phenomena which would not otherwise be necessarily found, it is imperative that enough samples meet this high Oct/Sox/Nanog requirement. There was no preconceived idea as to what would constitute “enough” samples, however, a brief look in the literature of mESC microarray analyses found that considerable efforts into large scale analyses had indeed already been undertaken, with larger projects such as the FunGenES project (Schulz et al. 2009) which amassed a total of 258 microarrays from a consortium of 20 research groups. Therefore, it is hoped that at least around this number of samples would be found to contain the search term “embryonic stem cell” as well as be found to have high expression of Oct4, Sox2 and Nanog. In an ideal world, the number of “high pluripotency marker samples” would exceed the number of samples found in studies such as this.

2.1.2 Generation of a set of manually-curated annotations for all HPM mESC microarray samples currently publicly available for the Affymetrix Mouse 430v2 Array

By far the greater amount of work in this chapter, however, was dedicated to the second research outcome; providing manual annotation of the data matrix generated in the first part of this chapter. In order to make meaningful connections between annotations and sample similarity, it was decided that this annotation effort would trace sample annotations to the online repository and then, further, to any available literature, taking the accompanying literature as the highest (most accurate) source of information about each sample.

It was decided to include information pertaining to the sample's source laboratory, cell line, any genetic modification made to the cell, any interference with gene expression (knockdown / siRNA *et cetera*), any reporter constructs and expression / inhibition of genes under control of an external

factor. Furthermore, rather than try to find a simple, representative culture condition which detailed what culture condition the cells were exposed to for the longest, and in preference to trying to deduce which culture condition the cells were immediately prior to submission for microarray, all information found in the online annotations / source literature pertaining to culture conditions would be included for completeness and, where available, time spent in each culture condition would be recorded. Further details of the annotation system are given in section 2.2.2.

Quite possibly the greatest challenge when annotating this data was (apart from its sheer size, which turned out unexpectedly to contain 1,101 high-pluripotency marker samples) having to invent annotation syntax whilst “going in blind”, as the format, quality and conventions of the available online annotations were entirely unknown and proved to be highly inconsistent. The annotation system therefore had to be developed as each sample was added, sometimes requiring restructuring of previous data fields or formats.

However, this task was seen as a necessary step, particularly as the planned downstream analyses for this thesis included investigation of transcriptional events from naïve pluripotency to early differentiation. A crucial reason for performing these manual annotations, therefore, was to be able to trace back to individual samples any later result. This allows for quick reference to what any given sample is, its culture condition, genetic modification etc., in case an effect of one of these annotations may be responsible for a given result, rather than being an interesting discovery about mESC biology *per se*.

Finally, despite its inherent utility to the work carried out in the preparation of this thesis, one of the most tangible outcomes of this chapter is, in fact, the simple existence of both the high-pluripotency-marker matrix of mESC microarrays and full, manual annotation of this data. These combined may make for a highly-useful resource for other researchers to go data mining with. It is essentially impossible for the work of one PhD student, let alone one discrete thesis, to mine all possible useful information from a resource such as this, making the generation of the data and its full annotation extremely useful beyond the scope of this work, hopefully for others to use to make further discoveries in the field of mESC biology.

2.1.3 Collation of a list of examples of discovered errors / omissions in available annotations for high-pluripotency-marker mESC microarrays

The final research outcome from this chapter is a detailed description of selected examples of the sorts of errors, omissions and / or confusing annotations that were found when attempting to perform the manual annotation. It is to the highly likely detriment of wet-lab biological science investigators, bioinformaticians and their automated systems that there is currently no immediate, tangible incentive for the uploading of standardised, accurate, intuitive, human and machine-friendly annotations for microarray data. Therefore, time spent on accurate, intuitive reporting of sample annotations may come to be seen as a time burden, rather than a priority for the betterment of science in general.

The presence of errors / omissions / confusing issues in both online annotations and sometimes even the literature referenced by online annotations is somewhat more understandable when this lack of incentive is borne in mind. Scientists, with often very limited time and manpower, can be expected to only naturally prioritise those tasks which further their work and allow them to meet their hard, enforced targets. This section of this work therefore is intended to provide some insight into commonly-made but quite easily-avoided mistakes that were picked up while manually annotating every sample in the high-pluripotency-marker matrix. This can be used to inform researchers who quite likely have the best of intentions while completing online annotations for their data, or referencing where to find their methodologies, as to which parts of their annotations require that extra care for them to be usable by others.

Stress is also given in the results section of this part of the work as to how automated searches / analyses may well be affected severely by the common mistakes encountered during this work. It is intended, therefore, that drawing examples of these errors / omissions / confusing issues may allow data uploaders to provide better annotations for their data through increased understanding of which errors cause the most human / machine frustrations and why.

2.2 Methods

2.2.1 Automated assembly of mESC microarray sample dataset

In order to maximise the chance of successfully delineating the effects on sample similarity of source laboratory and cell line, it was necessary to assemble the largest-possible matrix of mESC microarray data. By using as many samples as are available, the number of laboratories and cell lines is maximised, which, by definition, includes as much information about the transcriptional state associated with samples from different annotations as possible. However, manually assembling such a matrix from publicly-available repositories would be prohibitively-time consuming. Thence it was decided to use programmatic, scripted access to both GEO and ArrayExpress.

A scripted search was performed on all GPL1261 platform samples, requiring the term “embryonic stem cell” to be found. It was noteworthy, also, that performing this search using the standard web-browser user interface returned a fluctuating number of results, depending on the syntax used, but always below 1000 in number, while the scripted access consistently found the same number over 2500. Manual reading of a selection of the annotations of samples found using the scripted approach which were not found through the website's “advanced search” page confirmed that the extra samples found using scripted access did indeed contain the required term.

Further to this issue, GEO itself contains an admission that, due to the issues involved in placing large volumes of files in the same directory (a limit is imposed by some filesystems), a certain syntax of directory structure is used in order to provide for programmatic access to files, based on their experimental accession number. This syntax had to be dynamically mimicked in order to allow for automatic download of the data and, even so, came across pathing issues due to inconsistent truncation and padding of the directory names. ArrayExpress did not have these issues, although a smaller number of samples were downloaded from here as only those which were labelled as being unique to ArrayExpress were included in the final results passed to the download scripts.

One last issue was found during test runs of automatic download of data from the GEO repository, in that initial attempts at automatic download returned multiple failure messages. This was traced to an issue regarding the extensions of the files, wherein there was no accepted standard in the GEO repository as to whether or not file extensions should be capitalised or not. Without a way to predict which samples would have their file extensions in either lower or upper case (“.cel” or “.CEL”), it

was decided to simply use two simultaneous download scripts which both attempted to pull all desired .CEL files, with one script with all uppercase extensions, the other all lowercase, and to combine the results. No desired files failed to download after both of these scripts were run.

2.2.2 Manual annotation of matrix N1101

In order to allow for the interrogation of matrix N1101 for the purposes of delineating the contributions to sample similarity of source laboratory and cell line, annotation of matrix N1101 would, technically, have only required each sample to be annotated as to the status of those factors under investigation, those of source laboratory and cell line. However, the scope of work in this thesis was always intended to go beyond that, fulfilling other objectives such as the production of the first manually-annotated high-pluripotency-marker mESC microarray dataset and investigation of the transcriptional effects of other annotations, such as presence or absence of serum or other culture conditions. Therefore, a far more detailed annotation process was undertaken, including full, chronological culture conditions, control/treatment status, cell sorting status of the sample and genetic modification flags. This comprehensive annotation is provided in full on the included DVD in the “Chapter 2/N1101 Annotations” folder. The annotation system developed records sample details as follows:

Laboratory Group

This field was completed by using only the *last author* on any published paper associated with the sample. If no paper was linked to or cited by the annotations available online, usually one could be found by searching for terms included in the brief descriptions included in the online annotations.

Failing this, further searching for the experimental details was done until a paper was found. Confirmation of the paper's link to the samples in question was ascertained by linking treatment conditions, culture conditions, cell type and experimental design before an author was assigned to the annotation.

This resulted in nearly all samples having a clearly-associated “group”. It was decided to use the last author, rather than the first, to keep samples true to the overall lab group, rather than trying to piece together later which lab group a particular first author worked in.

Finally, those few samples which did not specify a paper, could not be found by lengthy, intuitive searching online and/or whose annotations were totally obscure or contradictory were assigned to the name of the individual chosen for correspondence on that sample's summary page online.

Date

Date of original upload of the data is detailed here in YYYY.MM.DD format, for example, 2006.04.26 for April 26th, 2006.

Cell Line

Information regarding cell lines was identified in the same per-sample manner as was annotated by the contributing authors. This labels those samples which are using common cell lines (for example the standard CGR8 cells, E14 cells, ES-D3 cells) but also maintains the names given to other lines by individual authors for comparison. Often the name given to a line is an abbreviation of the genetic background (where direct sampling is done from a live donor). This information is included in order to observe not only patterns which may arise from the use of less common cell types, but also to capture any relationships that may be due to cell origin. Of course, the opposite is also true in that any annotated property may prove to be unimportant, or linked to another. This is the driving reason behind maintaining such detail when annotating cell line.

Furthermore, cell line names are sometimes standard lines and sometimes not. In the case of a standard cell line from a catalogue, its details are provided *without* inverted commas (“”), for example the commonly-used ES-D3 cell line, which is available, standardised from the American Type Culture Collection (ATCC), catalogue reference number CRL-1934. In this case, cell name is provided as: ES-D3, ATCC: CRL-1934. Other cell lines from catalogues follow the same convention.

However, in many cases, cell lines are simply provided with names given by authors, particularly in the case of freshly-sampled cells from donor mice. Here, often authors name their cells after the genetic background, such as “C57BL/6”. Where such names are given, they are quoted directly and no attempt to standardise them is made in order to preserve uniqueness; clustering and similarity grants insight into which cell types may well be highly similar or even identical, but individual

names must be preserved for fairness. Some names are altogether uninformative as to genetic background, however, and it remains to be seen as to what they resemble most. An example of this kind of naming is found in the annotations for sample with filename “GSM258655.CEL”, which bears the cell line annotation of “embryonic_stem_line_Bruce4_p13”.

In addition, a condition most pertinent to the area of ES cell culture is the use of feeder cells. Whilst methods of ES cell culture may have largely moved on from the unnecessary complication of feeder cells, many samples in the data used in this work were cultured with the use of feeders. Unfortunately, the annotation of the use of feeders is not always clear. Therefore, like other data fields, it is possible to infer from the manual annotation whether or not feeders cells were *definitely* used. It is not, possible, however, to *always* infer the reverse; that samples that do not include the annotation or feeders are truly without their presence. This is only likely a very small minority of cases in this data, however, both due to the progression towards feeder-free culture and because authors, unlike with culture condition where components are often omitted from annotation due to their prevalence, the use of feeders is unlikely to be so commonly assumed.

Combined with the cell line annotation, an example of the syntax of the manual annotation when feeders were used is as follows, for the sample with filename “GSM747184.CEL”, cell line:

“KH2”, FEEDERS

Following the described syntax, it is therefore clear that KH2 is a name given by the authors only, in that inverted commas surround the KH2 name, and the experiment involved the use of feeder cells.

Annotation here of “N/C”, as in other fields, indicates that the cell line used is not clear from the uploaded annotations or accompanying literature. Contradictory, obscure or unreliable annotations in any way result in the use of this annotation, where “N/G” is where cell line is simply not specified at all (from “Not Given”). This latter annotation is somewhat rare and usually is attached to samples wherein no annotations are uploaded and no paper can reasonably be attributed to the sample at all, for example with unique filename “affs447.4.201005.CEL”.

Finally, other annotations are sometimes provided here in the event that samples should confuse any downstream analysis, such as “haploid” for cells said to be as such, or “parthenogenetic”. These are included only for completeness of annotation and are extremely few in number.

Genetic Modification

An extremely important field, genetic modification details, in a general sense, what changes have been made to the cell's genetic makeup, but also some information as to the technique and implications. This field also contains any information about markers, the engineered driving of certain genes by culture condition, constitutive expression and so on.

For samples with a relatively simple over-expression of a gene or other product, the name of this product is included without *any further syntax.* For example, in sample with file name "IPK.1_4_Pax4_ESC.CEL" are modified for constitutive expression of paired box gene 4 (Pax4). This is therefore annotated only with "Pax4". This type of annotation is particularly important in the case of iPS cells included in the data, where the factors used to generate them are recorded. If more than one name for the same product is used, or where one is used commonly in the ES cell field, both are provided, separated with *only* a slash "/" and no spaces (e.g. "Rex1/Zfp42" .)

Syntax here, as with all of these annotations, is crucial. In the case of simple genetic modification resulting in either a heterozygous or homozygous state for a particular gene, the usual convention is followed, with "Oct4 +/-" for heterozygous Oct4 and "Oct4 -/-" for Oct4 homozygous knockout. Convention is followed as regards the placement of the positive "+" sign before negative "-" always, so heterozygosity will never be marked as "-/+".

This annotation also includes important details regarding linkages between gene expression and culture condition. Conditional expression cell lines, for example, tie the expression of a certain product to components added to culture for the purposes of experimental manipulation. In this case, such as the example of sample with filename "GSM739488.CEL", annotation of this is given in the following manner:

DOX|Msgn1

This details that the mesogenin-1 (Msgn1) product is *driven* by the addition of doxycycline (DOX) to the culture condition. Annotation in this manner provides extra resolution of information when querying the data in that it is possible to separate this experiment's controls from treatment conditions, as both the culture condition components and modification are detailed, it can be seen

whether or not Msgn1 would be induced in these samples. This annotation is interchangeable with a similar case where distinction is made between activation and repression of a given product. This is annotated by simply adding "_ON" or "_OFF" to the driven gene. If neither is provided, the product is turned *on*. Lists never follow pipe characters, in that if a sample has multiple factors induced by the same factor, these will still be listed as separate modifications, separated by a comma and a space, as is convention here.

The pipe character "|" is used in a consistent manner in this way, always implying that what follows the pipe is driven or activated by what precedes it. This is therefore the way in which another common cell modification is recorded; the use of reporters. Reporters are intended, for example, to confirm the expression of a gene and are tagged onto the end of that gene; a common technique in biological research. Whilst reporters may not be apparent in microarray data, in that the green fluorescent protein (GFP), for example, a commonly-used reporter, will not be assayed for by the microarray, it is still necessary to record two facts, however; firstly, that the cell has been genetically-modified and, secondly, what this modification is.

Another important piece of information recorded here is whether or not the cells in the sample have been sorted in any way. Cell sorting is a common way of purifying a sample, selecting for a particular trait. For example, cells may be sorted for markers in order to purify cells for further use, or, indeed, to test the validity of a proposed marker for a given cell type / cell capability. Coupled with the above information on reporters driven by certain factors, this allows combinatorial querying of the data as it is known whether or not a certain reporter is present, what drives this reporter *as well as* whether or not the conditions for the expression of this reporter are met by the culture conditions. This provides much more information than the sum of its parts when it comes to interpreting the microarray data.

Cell sorting information is included with simple syntax, for example in the case of sample with filename "UKOE.5_sc5_CGR8.es_06.CEL", sorted for the presence of CD31, written as "CD31+". This is to differentiate cell sorting from any heterozygosity annotation (which would be marked "CD31 +/-".) Cell sorting is done while selecting both for and against a particular marker and, following convention, cells sorted for the *absence* of CD31 would be annotated "CD31-".

Information is also to be found here concerning other techniques such as RNA interference (RNAi), the usage of small-hairpin RNAs (shRNAs), endoribonuclease-prepared siRNAs (esiRNAs) and, of particular relevance to stem cell science, microRNA (miRNA) expression.

In the case of RNA interference, the pipe syntax is again followed, for example, sample with filename “GSM210974.CEL” has a small, interfering RNA (siRNA) for Renilla luciferase. Counter-intuitive as this may sound, this is, in effect, a control sample compared to others in this experiment (accession number GSE8503) where the siRNA was designed to interfere with microRNA-290 cluster (miR-290), known to be involved in the control of pluripotency in mESCs. Renilla luciferase is not present in the cells used at all, making the “siRNA|Renilla luciferase”-annotated samples the siRNA controls, to control for the effect of using any RNAi technique. Again, inclusion of these allows more accurate interrogation of the data than simply “genetically-modified or not” and allowing for the teasing apart of individual effects (in this case, of miR-290) and of any observable effects of RNAi technique in general.

The same syntax is followed for shRNAs, esiRNAs, and miRNA (which is modification for the expression of a given miRNA) and so this needs no further elaboration. Controls for any of these techniques are detailed in the same syntax with “CONTROL” after the pipe if no control product is detailed.

Finally, the use of the Cre/LoxP system is detailed here as it was commonly found while annotating the data. In this case, flanking a given region with locus of X-over P1 (LoxP) sites allows for the later use of the Cre recombinase to delete the region contained within the LoxP sites. Linking the expression of Cre recombinase to a particular component given in culture (commonly agents such as 4OHT or DOX), allows for conditional knockouts to be made. In the data used in this work, the flanking of a given gene or product with LoxP sites is denoted as follows, using the example of sample with file name “GSM648807.CEL”:

Rbp2 fl/fl

It is crucial to note that, in this case, the retinol-binding protein 2 (Rbp2) gene gas been “floxed”, but is *still active*. The annotation of such samples is to provide the knowledge that some genetic modification *has* taken place (in the insertion of the LoxP sites), but that Cre recombination *has yet* to take place. When recombination *has* taken place, the standard annotation for knockout will apply:

“Rbp2 -/- ” in this case. It should also be taken into consideration that, where possible, the factor responsible for the expression of Cre (when not a culture component) are listed, for example, in the case of sample with filename “GSM338367.CEL”, Cre recombinase is under the control of Mox2. This, in the case of this experiment, was achieved by replacing one allele of Mox2 with Cre recombinase. Three products are mentioned in the experiment's accompanying literature as being floxed for deletion in this experiment, resulting in the annotation:

Mox2|Cre, Rb fl/fl, p107 fl/fl, p130 fl/fl

In this case, it is more problematic, by definition, to tell from the manual annotations whether or not Cre recombination has taken place in these samples, as Cre recombinase is expressed when Mox2 is expressed, not when a specific action was carried out by the experimenters. This is, in fact, the only example of endogenously-controlled Cre recombinase expression in the data that were gathered, however, and would therefore require confirmation in the microarray data to predict the likelihood of Cre recombination having taken place.

Culture Condition

Culture condition annotation is required in order to detail the milieu of factors to which cells in a sample are exposed. This, by definition, can have profound effects on the cells themselves, as evidenced by the very existence of the science of maintaining ES cells (not only of the mouse) by small molecule inhibitors or other factors. This is in addition to the fundamental biological fact that cells can, do and indeed *must* respond to external cues.

Culture condition thereby provides both solution and problem. Culture of cells is a fundamental technique throughout biological science and is thus of immeasurable utility, but the effect of cell culture condition is often overlooked when considering results and inter-experimental comparisons. As is true of the design of any experiment, maintaining all factors constant except for the factor(s) being tested is implicit with good scientific method. This therefore applies to culture condition as much as any other factor. However, this is *profoundly* different from implying that the cell culture condition is having *no effect at all*, merely that that effect is applied across all samples, effectively masking any effect it may have.

Given the wide range of culture conditions used when culturing mESCs, there exists, therefore, a lot of potential for effects, masked within individual experiments, becoming apparent when comparing across large amounts of data. This, coupled with the nature of the area of stem cell research, where culture conditions for stem cell manipulation often take centre stage, makes it clear that such detailed cataloguing of culture condition is highly necessary as well as desirable.

To this end, as much detail as possible for each sample's culture condition was included in the manual annotation process, often consuming the majority of the time taken to annotate each. Indeed, with the vast majority of samples used, in this work, culture condition manipulation was the only change made between control and treatment. This is further complicated by the quality of the annotations available online. For example, it is most common, as observed from the compilation of the matrix used in this work, for uploaded annotations to only contain the culture condition that was used *at the time of harvesting of mRNA*. This does not provide the full story or "history" of the cell as previous culture conditions can reasonably be expected to have had considerable effects on cell state.

To this end, annotation syntax was developed which allows for the accurate, exhaustive and chronological recording of all components stated to be present in each sample's culture condition(s). However, there is still issue with this approach which must be elaborated upon. The key word with this is that every effort has been made to include every *stated* component. It must therefore be emphasised that the manual annotations, whilst exhaustive, must be considered to be somewhat incomplete in this regard.

Early in the process of annotating all samples, it was noticed that some cell culture components were omitted from sample annotations online which were clearly stated in the associated paper. For example, the inclusion of sodium pyruvate (a carbon source) is not always mentioned either in the literature or annotations. It is likely that this component is simply so ubiquitous that many authors do not feel the need to report its inclusion. However, this is the case for many culture components, including, but not limited to, non-essential amino acids (NEAA), β-mercaptoethanol and antibiotics, particularly in the cases of penicillin and streptomycin. These components, by their very presence, will have an effect on the transcriptome of the cells in the sample, to a greater or lesser degree.

Annotation of the culture conditions is formatted in chronological order by the use of square brackets "[ ]". All components listed within a set of such brackets are present at the same time. All

components in the culture conditions during that time are listed in no particular order (although usually listed in the same order as listed in the given annotations online or in attached literature, following that convention to a large degree.)

Each component is formed of both the concentration and the component name. These are separated by an underscore "_", in an attempt to avoid confusion between termination of concentration information and beginning of component name. Components are separated with a comma and a space ", ". Wherever a duration of time spent in a particular culture condition is given, this is denoted *after* the square brackets containing the details of that condition, and a period "." . There is *no* space between the square brackets of culture conditions for a given sample.

In the case of the use of Greek characters such as those for "micro" and for "beta", the regular Roman characters "u" and "b" are used, as is convention in the absence of advanced character typesets, as the annotations aim to be query-able by simple plaintext searching. Capitalisation is unimportant and provided for user-readability only, except in the case of units, for example "molar" requiring the use of the capital, and so on.

Crucially, culture conditions are annotated *in full*. That is, components are carried over from one culture condition to its successor for that sample when this is the description in the accompanying paper or online description. This is included for completeness so that every component present is annotated as being such, without the need for any guesswork or assumption as to what is carried over from one time point to the next. Furthermore, this is one of the many reasons why exhaustive reading was required of all accompanying literature.

The first culture condition that a sample is in does not carry with it a duration and, indeed, many subsequent culture conditions are not given a timepoint if this is ambiguous in any way from reading the associated literature.

As an example of all of this syntax, the annotations provided for sample file "E.MTAB.300.100_.Mouse430_2..CEL" (from Array Express experiment E-MTAB-300), are as follows:

[NC_DMEM, 20%_FBS, 1%_NEAA, 1%_Penicillin/Streptomycin, 2mM_L-Glutamine, 0.1mM_b-mercaptoethanol, 0.2%_DMSO],[NC_DMEM, 20%_FBS, 1%_NEAA,

1%_Penicillin/Streptomycin, 2mM_L-Glutamine, 0.1mM_b-mercaptoethanol, 0.2%_DMSO, 1.1mM_warfarin].24H

This represents all the information that was available for this sample compiled from both the uploaded annotations to ArrayExpress (observed as being more often the less-complete annotations) and from reading the methods sections of the attached paper(s). Quite often, attached literature simply quotes the use of method from another paper. Where accessible, these papers were also consulted until a culture condition was found.

In the event of discrepancy between the two (or indeed any whole culture condition that is unclear), the entire culture condition field is marked with “N/C” (not clear). In some cases, it is still possible to draw some information from unclear annotations in order to have made all reasonable effort. For example, in the case of sample with filename “HOXB4.ERT2.tmxf.pls.A4.cel”, it is clear that a treatment condition involving 4-hydroxytamoxifen (4OHT) was used after a control condition, so this annotation simply becomes:

[N/C],[NC_OHT]

It is still necessary to annotate in this manner as it can at least be inferred from this that the sample is *not* a “control” sample, but most definitely a “treatment” sample from an experiment. In addition, this may leave intact the ability to interrogate the data for any associated effects of 4OHT, given enough samples.

In the case of components such as Dulbecco's Modified Eagle's Medium (DMEM) or non-essential amino acids (NEAA) abbreviation is used and, particularly in the case of media, this is always accompanied by “NC” as the “concentration”. This is the “not clear” annotation to prevent potential confusion regarding the “concentration” of base media, but not annotating as “NG” for “not given”, as this is not the case. “NC” therefore simply provides any interrogation of the manual annotations with the idea of “concentration of this cannot be discerned for analysis.” In the case of the “NG” annotation, the entire field is changed to “N/G” if no annotation is given whatsoever.

Furthermore, manual annotations can be found in this work which include “N/C” for “not clear”, where excessively-confusing, contradictory, obscure or otherwise poor annotations are found in either a sample's accompanying paper or the annotations uploaded with the sample. It is also the

case that some samples have no clear culture condition associated with them in the uploaded annotations, and several culture conditions are mentioned in the attached paper, with insufficient link between them for confident reporting of cell culture condition for that sample (or, indeed, entire experiment.) In this case, the choice to use “N/C” was preferred to risking damaging real relationships found in our data for the sake of completeness. Nevertheless, even where “N/C” is sometimes used, a change of culture condition to definitely *add* a certain factor has still been annotated. For example, in the case where culture condition is unclear, but the experiment was clearly to assess the effects of adding LIF at a particular timepoint, this has still been recorded. This is necessary for downstream flagging of samples as having had components added to or removed from culture before harvesting.

A clearer annotation is found when “N/G” or “not given” is used, where it has proven impossible, to locate any given annotation. This often arose as a result of having no attached paper as well as no solid ground on which to base any manual attempt to link a paper to the experiment in question.

2.3 Results

2.3.1 A large number of publicly-available Affymetrix Mouse 430v2 microarray samples contain the search term “embryonic stem cell”

The initial download scripts for both repositories (GEO and ArrayExpress) obtained 3,321 individual .CEL files, after decompression of all archives. It was assumed at the time that these .CEL files contained unique samples, as the options used while searching ArrayExpress included the option to omit all samples which were hosted by GEO. This was mostly true, although during the manual annotation of the data, one duplicated experiment (ArrayExpress accession number E-GEOD-3653) (n = 16 samples) was identified as having identical annotation to a set of GEO samples and these were subsequently removed. This left 3,305 samples downloaded from GEO and ArrayExpress. To these were added 7 microarrays from the Prof. Ian Chambers' lab's own investigation of Nanog-related genes (for a total of n = 3312 samples), including 3 Nanog knockout samples, 2 Nanog-overexpressing samples and 2 Nanog-wild-type samples. These samples were added as there was initial discussion of observing where within the rest of the data these samples would cluster. This investigation was not taken forward, and the samples remain in matrix N3312, although as soon as it was decided that the aforementioned investigation would not proceed, these samples were restricted from being included in the high-pluripotency marker (HPM) matrix. These samples therefore contribute nothing to downstream analyses and do not affect the work in this thesis which almost exclusively uses the high-pluripotency marker matrix. All of the conclusions from this thesis are therefore based only on the automatically-downloaded publicly-available data. A list of all filenames that made up matrix N3312 is available as “Chapter 2/N3312 Filenames List/N3312.Filenames.List.csv” on the accompanying DVD, while all filenames which made it into the HPM matrix can be found in the manual annotation file, provided as both the commonly-used spreadsheet format (.xlsx) and as a plaintext tab-delimited file (commonly called (though a misnomer in this case) a “comma-separated value” (.CSV) file) under “Chapter 2/N1101 Annotations” on the accompanying DVD. The 7 samples in matrix N3312 which were from Prof. Ian Chambers' lab are clearly visible in here as beginning with the string “Chambers”. The order of the filenames in the file list for matrix N3312 corresponds to their column indexes in the compressed R object version of matrix N3312 provided on the DVD as “Chapter 2/Matrix N3312/N3312.RObject”.

2.3.2 Distribution of levels of gene expression of matrix N3312

Taking mRNA detection to be representative of gene expression, figure 2.1 shows the distribution of all probes' mRNA detection across all samples in matrix N3312. This clearly shows that, as would be expected, the highest frequency peaks occur toward the lower end of the mRNA detection spectrum. These values represent genes that are likely to be off or, at the very most, expressed at levels that are indistinguishable from background noise. As the RMA values increase to about around the 5 mark, genes are now in the “lowly expressed” range and, at around the 8 mark, the histogram tails off as gene expression becomes very strong. This is in keeping with the phenomenon that most genes in any cell are either off of lowly expressed, and that only a fraction of the total potential compliment of genes that can be expressed are actually very highly expressed. It should be further noted that, given that the RMA values used in such plots are log2 values, an increase of 1 along the x-axis represents a doubling of mRNA detection. This makes clearer the interpretation that at the very highest levels of expression, around the 14 mark, representing a $(2^{(14-5)}) = 512$-fold upregulation compared to those around the 5 mark, one expects very few genes to be in this region (figure 2.1).

Concerning three of the canonical pluripotency markers, Oct4, Sox2 and Nanog, the distributions of these genes is shown in figure 2.2. Vertical lines here represent the thresholds chosen to represent the very highest levels of expression of these three pluripotency markers in the later filtering of matrix N3312 in an effort to leave only pluripotent ES samples for further analysis. In the case of these 3 pluripotency factors, a clear bimodal-like distribution is seen across all the data. In the case of Oct4, however, there is another peak of expression lower than the highest peak. When thresholding for the highest pluripotency markers, it was decided to include at least a little of the shoulder of this peak for interest's sake, to observe how these samples might be related and/or different to those highest in expression for Oct4. The fourth canonical Yamanaka factor, Klf4's expression distribution is also shown here. Klf4 was not used in the filtering for pluripotent ES samples as it does not have a clearly-visible bimodal-like distribution amenable to such filtering. Klf4 instead shows a distribution which appears to be three peaks, the central of which is much larger than the two at either side of it.

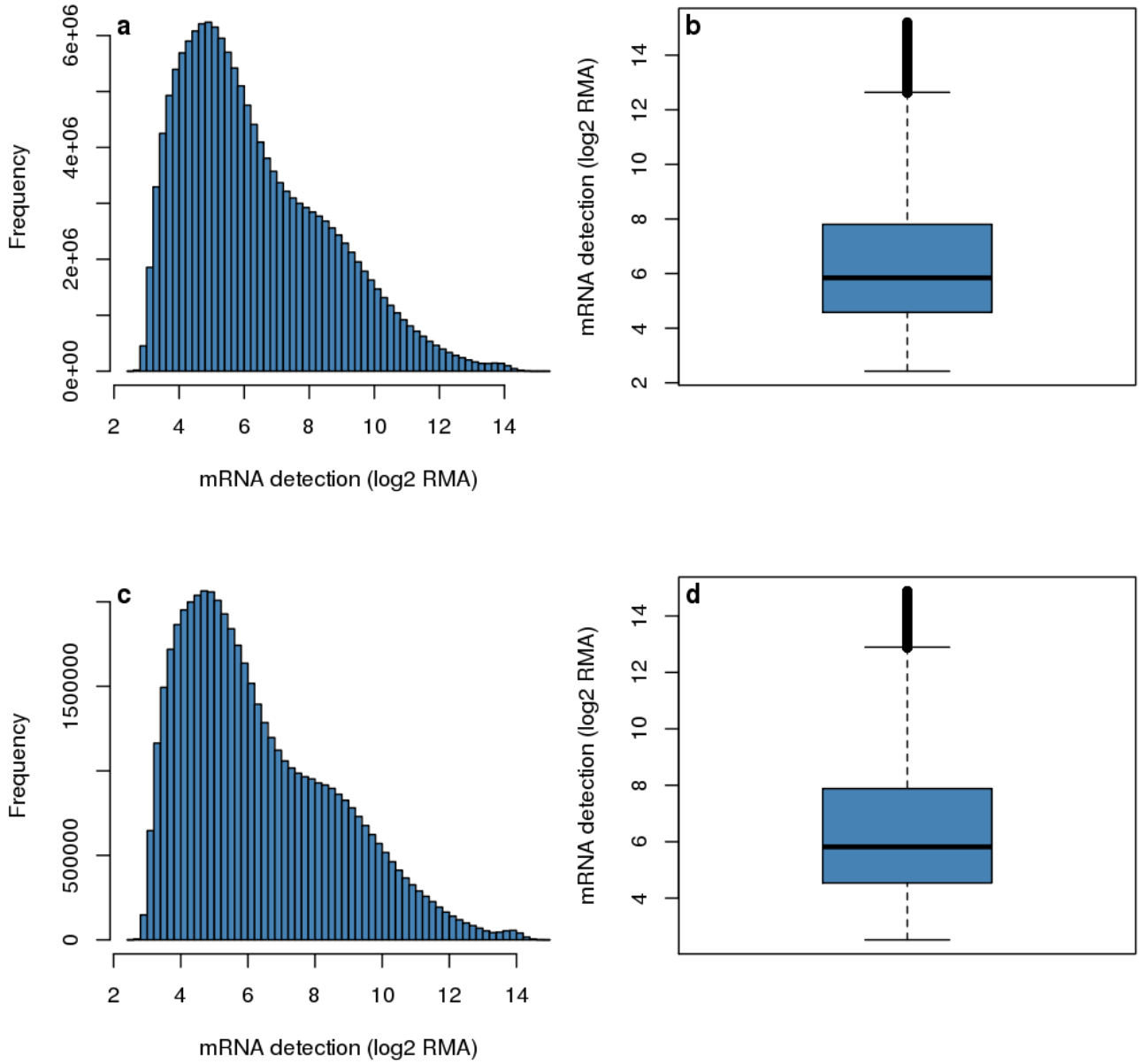

Figure 2.1: Distributions of all robust multichip average (RMA) values as both histograms (a) and (c) and as boxplots (b) and (d) for matrices N3312 (top panels) and the HPM matrix (bottom panels.)

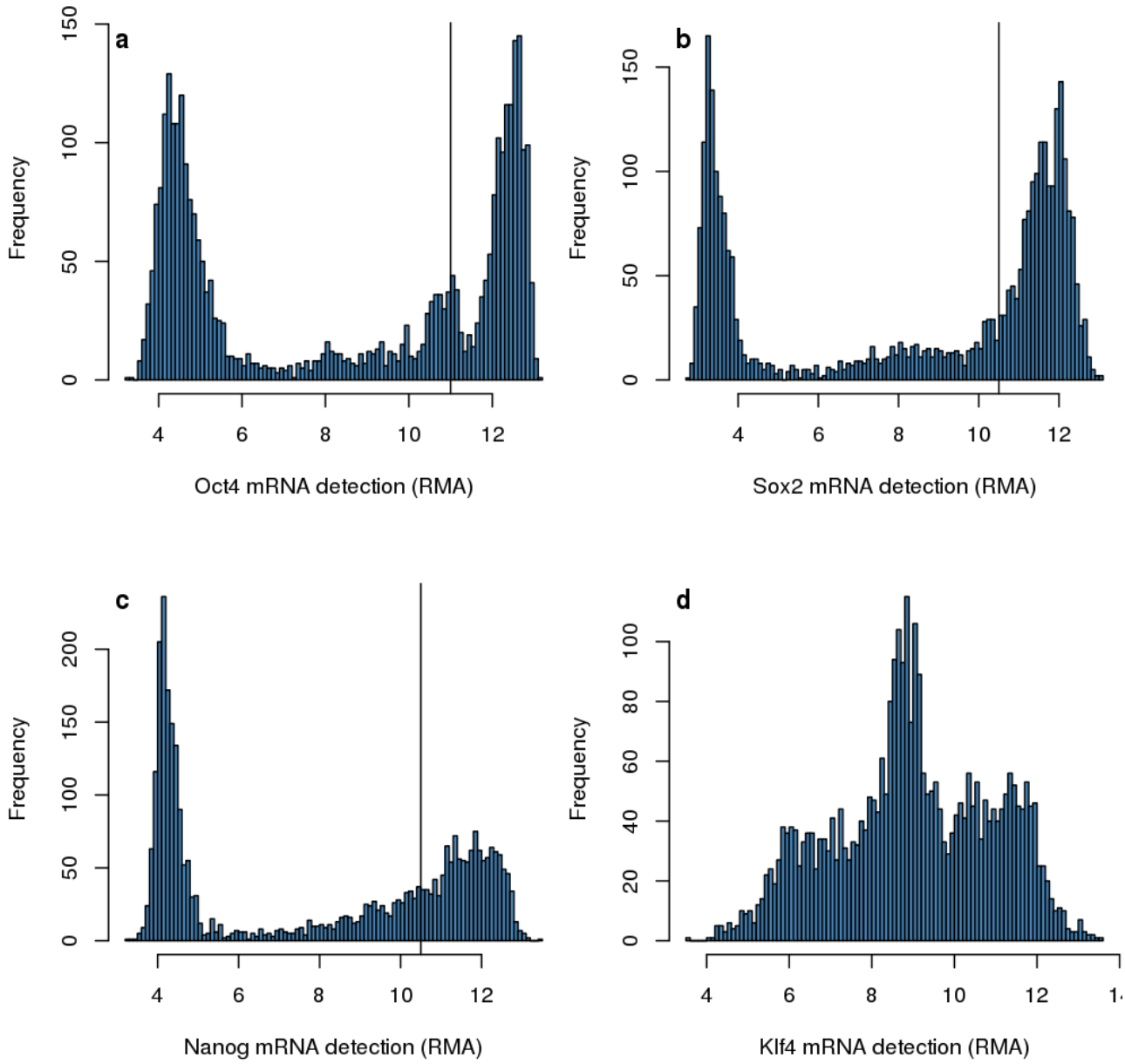

Figure 2.2: Distribution of expression of canonical Yamanaka factors Oct4, Sox2 and Nanog in matrix N3312 (a, b and c respectively), with vertical black lines denoting selected cutoffs for making matrix N1101 (except in the case of Kl4 (figure d), which, in the absence of clear "low/high" states, was not used for filtering.)

2.3.3 Filtering of matrix N3312 for high-pluripotency-marker-only samples

The pluripotency markers Oct4, Sox2 and Nanog are well-accepted markers of the pluripotent state. The automated assembly of matrix N3312, whilst the requirement for the search term “embryonic stem cell” being satisfied was indeed enforced, this automated search cannot be relied upon to guarantee that only *bona fide* embryonic stem cell samples were retrieved. Furthermore, as pluripotency and the point of immediate exit from pluripotency are the foci of this work, it therefore stands to reason to retain only those samples which are *bone fide* ES cells, as evidenced by high Oct4, Sox2 and Nanog expression. Thresholds for log2 RMA values were set for Oct4 at 11, Sox2 and Nanog at 10.5, again, marked by vertical lines on figure 2.2 and only samples which satisfied all *three* of these thresholds were kept to form the high-pluripotency-marker matrix, leaving behind 1101 samples. This matrix was therefore named N1101. A copy of matrix N1101 is available as an R object on the DVD as “Chapter 2/Matrix N1101/N1101.RObject”.

2.3.4 Distribution of levels of gene expression of matrix N1101

This is mentioned mostly as it would be hoped that filtering for high Oct4, Sox2 and Nanog should not majorly alter the distribution of mRNA detection from that which was observed across matrix N3312. This is indeed the case, as can be seen in figure 2.1, bottom 2 panels, which show a histogram and boxplot showing the distribution of detection of all probes across matrix N1101 (the HPM matrix). The same comments apply to this distribution as applied to the upper panel plots, as mentioned in section 2.3.2.

2.3.5 Examples of each type of issue encountered in online microarray data annotations / accompanying literature

The assessment of the state of available annotations for microarray data was carried out during the manual annotation of matrix N1101. A variety of issues were identified and are detailed below. Not every incidence of a certain kind of issue is elaborated upon and only issues which can interfere with downstream analysis or otherwise prove problematic were considered to be issues. Furthermore, the list of issues here should not be taken to be exhaustive, in that those samples in matrix N1101 not linked with any issues in this section should not be assumed to be free of any issues. This is because the primary objective of generating the manual annotation was to analyse the

effect of source laboratory and cell line on sample similarity; recording every occurrence of every issue along with the details of the efforts made to retrieve accurate annotations from the source literature would have proven prohibitively time-consuming as well as generating an unacceptably large amount of text / tables for inclusion in a thesis. Formatting, quality of English and other issues which may be bothersome, but are unlikely to have any significant effect on the ability to use / analyse the data (including by a machine) are considered to be irrelevant for the purposes of this work and so are also omitted. The discovery of these issues and detailing them on as large a scale (n=1101) as has been done here is, to the author's knowledge, the first attempt to provide both an assessment of the quality of mouse ES cell microarray data annotations, and also provide example reasons why any observed failings of these annotations affects downstream usage of the data in order to underscore need for more thorough, thoughtful, accurate and standardised annotation of microarray samples, although this section can also inform future practice of annotating other cell culture samples not destined for use only in microarrays. It is also the objective of highlighting specific types of error relevant to the use of mESC microarray annotations to allow authors to concentrate on ensuring accuracy where it is most relevant and saves the most frustration for other researchers.

Spelling

Whilst spelling may not seem initially to be a major issue with microarray annotation and can, indeed, often be overlooked in the case of non-mission-critical words, spelling was often found to be incorrect in important terms during the manual annotation of N1101. In E-TABM-562, for example, an ArrayExpress experiment, a crucial term “STAT3” was mis-spelled as “STATA3”. Whilst a researcher in the embryonic stem cell field may well recognise this as a mistake, as STAT3 is a critical pathway in ES cell biology, any attempt at automated use of online sample annotations would not be able to make such a judgement and would miss this sample. Spelling mistakes in such crucial terms may well have an even more deleterious effect, were STATA3 to be some other, valid signalling pathway. This spelling mistake occurs in the experiment description wherein a cell line is labeled as being “E14-STATA3-CTL”, and so would interfere with looking for cell line names as well as the “STAT3” string being broken with the erroneous “A.”

Sample naming such as this is not the worst area to be affected as regards annotation spelling, as there are errors present in actual compound names when details of culture conditions are recorded. This is evidenced, for example, in GSE20575 which incorrectly spells “l-glutamine” as “l-

glutamin", which will, again, disrupt searches performed on the annotations as they are. In the case of this work, for example, where a method was envisaged for the investigation of the relationship between culture conditions, source laboratories etc. with a transcriptional profile, there is scope to look deeper into the data (although beyond the scope of a PhD thesis), at a very high resolution, down to individual culture components, something which would require errors such as these to be corrected in the available annotation. It was the observation of the issues in the annotations that fuelled the decision not to attempt to analyse culture condition effects on sample similarity in chapter 3, as the annotations could not be completed to a satisfactory level.

Issues concerning the general spelling / checking over of annotations also came to notice with experiments such as from ArrayExpress, accession number E-TABM-674, wherein the description contains markup language, such as chevrons and markup codes for formatting. It would appear, therefore that these have been hastily copy-pasted from another source. Worse, the failure of superscript formatting in this copy-paste effort lead to the inclusion of claims that the cells were "then stimulated with 103U/ml mLIF". This is extremely likely to have been intended to read "10^3". Without checking on the annotations, the uploader has left this highly incorrect concentration on their online annotation, which will need to be manually detected by anyone re-using this data. Such problems continue in the next subsection of examples with units:

Units

Critical to the annotation of cell culture samples are the culture medium details. Mistakes in units can be, at best, confusing to manual readers and, at worst, mis-informative to attempts at automated processing of annotations. For example, ArrayExpress experiment P-TABM-4268 contains an annotation in the cell culture details wherein an abbreviated unit "nanomolar", correctly written as "nM", is written as "nm". Far from being an issue solely for the pedant, this unit, the nano*metre* is not only inapplicable to the subject, but is mis-informative. Case-sensitivity is not optional when recording units.

An example from GEO sample GSM185513 further demonstrates the point concerning unit annotations as here a culture condition is labeled as having 15M mercaptoethanol present. This is clearly a concentration which is highly unlikely and likely the annotation meant to read a smaller unit, probably "15mM", common in other annotations. Whilst this may seem inconsequential to the manual reader, again, automated attempts to marry annotation to effect will incorrectly generate a

sample assumed to have this “15M” concentration of mercaptoethanol in the culture condition. This may, at best, waste a great deal of time at a later date, combing through an analysis to find the mistakes in the annotation that have generated anomalous or impossible results, or, at worst, the error goes unnoticed and affects all other results, leading to false conclusions, at even greater cost as regards time and trustworthiness of results. Similar issues regarding units were found wherein 5 micro- and 5 milli- molar concentrations of LY294002 (PI3K signalling inhibitor) were reported in ArrayExpress experiment with accession number E-TABM-673 in the accompanying paper and the online annotations respectively, leaving uncertainty as to which should be treated as true.

Some online annotations also differ from a standard format of expressing a concentration of a given factor as a simple “number per unit”, instead adding a number to the unit as well, such as the case with GSE27341, which mentions “5µl/500ml β-mercaptoethanol”. Whilst this is at least informative, the use of a standard “number per unit” annotation here would remove the need for either a manual researcher or automatic search to perform calculations in order to render concentrations comparable.

Annotation access issues

Whilst referencing previous or other work for in order to detail a method is not an inaccuracy *per se*, it costs a great deal of time to chase down individual culture conditions or methods when a reference to another paper is given. This is another jump which would be highly problematic, if not impossible, to automate. This applies, for example, to an experiment found from GEO: GSE8128. The online annotation for the culture protocol simply states “ES cells grown as per protocol as described at the UCSF Bay Genomics website”. Aside from, again, the extra work involved in tracking this down and, again, preventing any automated retrieval of samples and their annotations, there is no link whatsoever provided with the online annotations in order to ease manual searching. Going to the source literature for this sample and looking up the reference manually gives “http://www.baygenomics.ucsf.edu”. Aside from simply being unhelpful in that this link does not go directly to any protocol (as can be discerned simply by looking at the link; it's a homepage, not a specific page), the link simply does not work and searching manually for a “Bay genomics ES cell protocol” or other permutations of this looking for the protocol were entirely unfruitful. This has remained the case for at least a year at time of writing. The use of internet links for methodology is therefore a practice which not only renders futile any attempt to automate annotation retrieval or

meta-analysis, but is also highly vulnerable to dynamic changes in webpage addresses. Whilst it may be understandable to have methodology written in the accompanying literature and therefore a simple reference to this put on the online annotations, two major issues arise from this. First and foremost is that there are specific spaces in the online annotation forms for both GEO and ArrayExpress which request cell culture condition details. It can only be assumed, therefore, that authors choosing to reference only their literature, rather than provide the details in the appropriate forms simply do not wish to take the time to fill these forms out, again, rendering large-scale automated analysis impossible or severely hampered were this practice to be replicated across all samples. Secondly, the use of data from publicly-available repositories should *not* presuppose the availability of the source literature as this is not always the case, even in academic institutions with seemingly-comprehensive journal subscriptions. Ideally, the online annotations should be completed to the point wherein they are sufficient to inform the reader to the extent that they are able to replicate the experiment and attempt to reproduce the results without recourse to other (often several) sources.

Another, more common issue concerning access to annotations is where seemingly the “trail runs cold” wherein a sample's annotation is first missing in the online annotations, and on further investigation of the sample by tracing to the source literature, there is no direct mention of that sample, or even the cell type which that sample purports to be in the online literature. An example of this can be seen with sample GSM258655, part of experiment with accession number GSE10246, with accompanying literature apparently as (Lattin et al. 2008). Even tracing through the accompanying paper, there was no mention of the sample by accession number or even a reference to the cell line used, being annotated online only as “embryonic_stem_line_Bruce4_p13”. A search therefore had to be done of all accompanying supplementary files from the paper to find one solitary reference to this cell line, in the third available supplementary file, which only detailed that this was “sample number 37, of unknown gender, a technical replicate, using 2000ng of RNA, and single amplification.” None of this effort, in the end, provided any information that could prove enlightening in the manual annotations. This sort of access issue, wherein annotations are missing or sometimes simply paraphrased (e.g. with GSE30561, where the culture condition is described as “standard conditions maintaining the undifferentiated state, i.e. KO-EM supplemented with LIF”, allows only the most rudimentary reconstruction of what the cells were exposed to), occurred many times in other uploaded experiments not listed here in the interests of space. Furthermore this section is to provide examples, rather than be an exhaustive list.

Contradiction

One of the most confusing and potentially-disruptive errors to make in annotation is that of contradiction. A stark example can be found in sample GSM266063 from experiment GSE10553. Whilst to a human reader, it is possible to discern the meaning of the contradictory sample name “BAF250 knock-out ES cells, wild type”, this presents severe problems to any attempt to retrieve annotations for certain kinds of cells/experiments automatically. The suggestions from the name is that these are BAF250 knockout cells, but with “wild type” appended to it, which is an impossible cell type; a wild-type knockout. The rest of the sample description for this sample later states “knock-out ES cell lines are derived in vitro from wild type E14g ES cell line”. Therefore the annotation should simply detail the knockout status of the cell line, rather than confusingly adding the “wild-type” marker to it as well. These are clearly *not* wild-type cells. Contradictions such as this pose possibly one of the greatest hurdles to larger-scale analysis, whether automated or not, as contradictory annotations not only mean that both mutually-exclusive states must be considered to have *no* annotation, but this calls into question the accuracy of the rest of that sample / experiment's annotation. Direct contradiction such as this was a relatively rare occurrence, however, compared to issues of spelling, access and units, but compared to the next issue mentioned, simply obscurity:

Obscurity

An example of this is prominent in experiments such as GEO experiment GSE10210. GEO accepts (and, indeed, requires) sample names to be uploaded for each microarray. In order to make some sense of the samples present in the data, either the experimental description or the sample name best make clear which sample pertains to what condition. In the case of GSE10210, sample names such as “HB1” and “HP2” are used. Details of individual samples contain a little more information, and these appear to change along with the abbreviations, but not in any meaningful manner. Here, HB = VEGFR2+, day 2.5, HP = CD41+, day 3.5. The sample names are only referred to in the full text of the source literature ((Nikolova-Krstevski et al. 2008)), whereas using the description of the samples by their cell line, sorting and timepoint is *far* more common in other uploaded experiments and is far more helpful to the reader.

Another form of obscurity no less frustrating to both manual and automated annotation retrieval is exemplified perfectly by an experiment with accession number GSE3653, where, for each sample

uploaded (n = 16), the uploader has simply “copy-pasted” the same paragraph under the “description” field for every sample. This paragraph describes all 16 uploaded samples and contains a description of the overview of the entire experiment. With all samples described in every “description” field, any automated search is going to be frustrated at this point and require manual intervention. In fact, the experiment's summary page has this same paragraph under the “Overall design” field, where it is helpful, and this should not have been simply appended to every sample.

Missing literature

Both GEO and ArrayExpress clearly have sections demarcated for uploading experimenters / authors to detail the published literature associated with their samples. However, associated literature is often missing or incorrect papers are cited, even when correct literature both exists and is searchable by other means. This is noted in order to distinguish between any experiments which may not have generated any published literature, and those which have. In the case of those with associated published literature, it is imperative that the literature be cited here, in order to facilitate any further interrogation of the data and also to simply inform the reader. In the case of samples from GSE10574 (such as GSM266837), for example, it is even recorded (at time of writing) that the citation is missing. For this sample, only an author list exists. Manual searching for this author list has an exact match with a published paper by (Endoh et al. 2008), which, in fact, references only a similarly-numbered experimental accession number, GSE10573, which doesn't actually contain sample accession GSM266837 at all.

ArrayExpress also had examples of missing literature, such as with E-MEXP-2238, which has no associated paper (although under the “citation” field, there is a sentence which reads like the title of a possible article, there is no such article, even when corrected for the presence of a question-mark in this line, likely the result of a failure to process a non-standard character, as also happened with the contact e-mail address.) By downloading the experimental details, using the author name and sub-parts of this possible article title along with the experimental accession number, a paper matching this data could eventually be found, however, in (Caillier et al. 2010).

A special mention on the presence / absence of feeder cells

As the first forays into ES cell culture required the use of mitomycin-C-inactivated fibroblasts as a feeder layer (Evans & Kaufman 1981), the presence or absence of feeder cells is a culture condition annotation worthy of special mention. As knowledge of ES cell culture improved and alternatives to feeder cell culture became available (Williams et al. 1988), it is understandable that a tendency to avoid feeder cell culture would arise in that this removes one more “undefined” presence in the culture medium, much in the same way that a move from fetal calf serum (an undefined medium) to only defined media (*id est* with known constituents at known concentrations) would reduce variations in cell culture. However, the presence or absence of feeders is not always mentioned in the samples annotated in N1101. As mentioned in the section dealing with “cell line” annotations, absence or presence of feeders was annotated only where such information was available. It is not strictly a criticism of the available annotation that the presence or absence of feeders cells is not always given. Some experimenters clearly are aware of this issue and make special mention of the fact that no feeder cells were used (e.g. experiment GSE22637 on GEO makes this very clear.) This is why, despite the recording of any available information of presence/absence of feeders in the manual annotations, it was decided not to try to investigate any effect of the absence or presence of feeder cells from this data and why special mention of this case is made here; it simply is not possible to, with enough confidence, place the samples from matrix N1101 into clear “feeders”/ “no feeders” categories.

2.3.6 Summary of Research Outcomes

2.3.7 Assembly of a large dataset of mouse embryonic stem cell microarrays

This chapter details the assembly of a large (n = 3,312 samples) dataset of mouse microarrays which were annotated online in the public repositories GEO and ArrayExpress as containing the string “embryonic stem cell” (and 7 microarrays from the Ian Chambers' lab.) Whilst the vast majority of samples downloaded were, in fact, unique, there was one instance of duplication and other minor issues concerning complex paths for file download from GEO, and capitalisation of file extensions. These minor issues were overcome and the duplicate samples removed. As previously mentioned, detailed annotation of the large, unfiltered N3312 matrix was not carried out, as it

cannot be ruled out, and is indeed likely that, several samples within this data matrix were downloaded due to the phrase “embryonic stem cell” being detected in online annotations, rather than every downloaded microarray being a *bona fide* mESC sample. This was part of the reason behind downstream filtering for pluripotency factors Oct4, Sox2 and Nanog.

2.3.8 Generation of a high-pluripotency-marker (HPM) mESC microarray matrix

Following a brief check that the distribution of probe detection values made intuitive sense (see figure 2.1) the large matrix of all 3,312 microarrays was filtered in such a way as to leave behind only samples which had the highest expression of all of three core pluripotency factors Oct4, Sox2 and Nanog (see figure 2.2). This left 1,101 high-pluripotency-marker (HPM) samples, forming matrix N1101, which retained a similar distribution of expression values as did matrix N3312 from which it was made (figure figure 2.1.)

2.3.9 Full manual annotation of the HPM matrix

The vast majority of work in this chapter was in the generation of manual annotations for the HPM matrix, providing details of source laboratory, cell line, upload date, genetic modification, cell sorting of sample, experimental accession number, sample file name and details of culture conditions to which the samples were exposed, attempting to list as many factors in the culture as could be gleaned from the online annotation or, failing that, accompanying literature. The combination of the HPM matrix and the annotation of all samples within it makes for a powerful tool for investigating transcriptional events in mESCs, as chapter 3 and particularly chapter 4 go on to demonstrate. An example screenshot of these annotations loaded into the freely-available spreadsheet editor LibreOffice Calc (available at https://www.libreoffice.org)is given in figure 2.3.

These manual annotations also confirm that the combination of using the search string “embryonic stem cell” and downstream filtering for highest expression of Oc4, Sox2 and Nanog left behind only mESC and mESC-like samples. That is, all samples in this matrix are annotated as being mESCs or iPSCs at various degrees of pluripotency / early differentiation. The fact that all of these samples still exhibit the highest levels of expression of Oct4, Sox2 and Nanog allows this matrix to be said to contain only pluripotent samples, with a qualifying remark that these are markers, rather than functional confirmations, of pluripotency.

The availability of this data matrix and accompanying manual annotation is, to the author's knowledge, the largest manually-annotated dataset of high-pluripotency-marker mESC microarrays to date, enabling detailed analyses of transcriptional profiles and events in mESCs, with some first steps on both of these paths being taken in the later chapters of this work.

	A	B	C	D	E	F	G	H
430	**Filename**	**Matrix Order**	**Experimental Accession Number**	**Lab**	**Date online**	**Cell line**	**Genetic Modification / Sorting**	**Culture Condition Details**
431	GSM272753.CEL	450	GSE10806	Schoeler H	2008.06.28	"OG2/ROSA26", FEEDERS	Oct4|GFP	[NC_DMEM, 15%_FBS, NC_NEAA, L-glutamine, NC_penicillin/streptomycin, NC_b
432	GSM272836.CEL	451	GSE10806	Schoeler H	2008.06.28	"OG2/ROSA26", FEEDERS	Oct4|GFP	[NC_DMEM, 15%_FBS, NC_NEAA, L-glutamine, NC_penicillin/streptomycin, NC_b
433	GSM272837.CEL	452	GSE10806	Schoeler H	2008.06.28	"OG2/ROSA26", FEEDERS	Oct4|GFP	[NC_DMEM, 15%_FBS, NC_NEAA, L-glutamine, NC_penicillin/streptomycin, NC_b
434	GSM272839.CEL	453	GSE10806	Schoeler H	2008.06.28	IPS	Oct4|GFP, Oct4, Klf4	[NC_DMEM, 15%_FBS, NC_NEAA, L-glutamine, NC_penicillin/streptomycin, NC_b
435	GSM272846.CEL	454	GSE10806	Schoeler H	2008.06.28	IPS	Oct4|GFP, Oct4, Klf4	[NC_DMEM, 15%_FBS, NC_NEAA, L-glutamine, NC_penicillin/streptomycin, NC_b
436	GSM272890.CEL	455	GSE10806	Schoeler H	2008.06.28	IPS	Oct4|GFP, Oct4, Klf4	[NC_DMEM, 15%_FBS, NC_NEAA, L-glutamine, NC_penicillin/streptomycin, NC_b
437	GSM279200.CEL	459	GSE10806	Schoeler H	2008.06.28	IPS	Oct4|GFP, Oct4, Klf4, Sox2, c-myc	[NC_DMEM, 15%_FBS, NC_NEAA, L-glutamine, NC_penicillin/streptomycin, NC_b
438	GSM279201.CEL	460	GSE10806	Schoeler H	2008.06.28	IPS	Oct4|GFP, Oct4, Klf4, Sox2, c-myc	[NC_DMEM, 15%_FBS, NC_NEAA, L-glutamine, NC_penicillin/streptomycin, NC_b
439	GSM279202.CEL	461	GSE10806	Schoeler H	2008.06.28	IPS	Oct4|GFP, Oct4, Klf4, Sox2, c-myc	[NC_DMEM, 15%_FBS, NC_NEAA, L-glutamine, NC_penicillin/streptomycin, NC_b
440	GSM277757.CEL	456	GSE10970	Gearhart JD	2009.03.05	"B6C3F1"	Nkx2-5|GFP	N/G
441	GSM277758.CEL	457	GSE10970	Gearhart JD	2009.03.05	"B6C3F1"	Nkx2-5|GFP	N/G
442	GSM277759.CEL	458	GSE10970	Gearhart JD	2009.03.05	"B6C3F1"	Nkx2-5|GFP	N/G
443	GSM371661.CEL	535	GSE11274	Schoeler H	2009.07.15	N/G	NONE	[NC_DMEM, 15%_FBS, NC_b-mercaptoethanol, 1000U/ml_LIF]
444	GSM371662.CEL	536	GSE11274	Schoeler H	2009.07.15	N/G	NONE	[NC_DMEM, 15%_FBS, NC_b-mercaptoethanol, 1000U/ml_LIF]
445	GSM371663.CEL	537	GSE11274	Schoeler H	2009.07.15	N/G	NONE	[NC_DMEM, 15%_FBS, NC_b-mercaptoethanol, 1000U/ml_LIF]
446	GSM294970.CEL	462	GSE11628	Ruiz-Ontanon P	2009.12.31	ES-D3, ATCC: CRL-1934	NONE	N/G,[NC_LIF]
447	GSM294971.CEL	463	GSE11628	Ruiz-Ontanon P	2009.12.31	ES-D3, ATCC: CRL-1934	NONE	N/G,[NC_LIF]
448	GSM294972.CEL	464	GSE11628	Ruiz-Ontanon P	2009.12.31	ES-D3, ATCC: CRL-1934	NONE	N/G,[NC_LIF]
449	GSM294973.CEL	465	GSE11628	Ruiz-Ontanon P	2009.12.31	ES-D3, ATCC: CRL-1934	NONE	N/G,[NC_LIF,50nM_lithium chloride]
450	GSM294974.CEL	466	GSE11628	Ruiz-Ontanon P	2009.12.31	ES-D3, ATCC: CRL-1934	NONE	N/G,[NC_LIF,50nM_lithium chloride]
451	GSM294975.CEL	467	GSE11628	Ruiz-Ontanon P	2009.12.31	ES-D3, ATCC: CRL-1934	NONE	N/G,[NC_LIF,50nM_lithium chloride]
452	GSM294978.CEL	468	GSE11628	Ruiz-Ontanon P	2009.12.31	ES-D3, ATCC: CRL-1934	NONE	N/G
453	GSM294981.CEL	469	GSE11628	Ruiz-Ontanon P	2009.12.31	ES-D3, ATCC: CRL-1934	NONE	N/G,[50nM_lithium chloride]
454	GSM314038.CEL	470	GSE12499	Schoeler H	2009.02.01	"OG2/ROSA26", FEEDERS	Oct4,Oct4+, FEEDERS	[NC_DMEM, 15%_FBS, NC_NEAA, L-glutamine, NC_penicillin/streptomycin, NC_b
455	GSM314039.CEL	471	GSE12499	Schoeler H	2009.02.01	"OG2/ROSA26", FEEDERS	Oct4,Oct4+, FEEDERS	[NC_DMEM, 15%_FBS, NC_NEAA, L-glutamine, NC_penicillin/streptomycin, NC_b
456	GSM314040.CEL	472	GSE12499	Schoeler H	2009.02.01	"OG2/ROSA26", FEEDERS	Oct4,Oct4+, FEEDERS	[NC_DMEM, 15%_FBS, NC_NEAA, L-glutamine, NC_penicillin/streptomycin, NC_b
457	GSM325390.CEL	473	GSE12982	Orkin SH	2008.12.12	, ATCC: SCRC-1010, FEEDER	NONE	[NC_DMEM, 15%_FBS, 0.1mM_b-mercaptoethanol, 2mM_l-glutamine, 0.1mM_NEA
458	GSM325391.CEL	474	GSE12982	Orkin SH	2008.12.12	, ATCC: SCRC-1010, FEEDER	NONE	[NC_DMEM, 15%_FBS, 0.1mM_b-mercaptoethanol, 2mM_l-glutamine, 0.1mM_NEA
459	GSM325392.CEL	475	GSE12982	Orkin SH	2008.12.12	"CJ7", FEEDERS	NONE	[NC_DMEM, 15%_FBS, 0.1mM_b-mercaptoethanol, 2mM_l-glutamine, 0.1mM_NEA
460	GSM325393.CEL	476	GSE12982	Orkin SH	2008.12.12	"CJ7", FEEDERS	NONE	[NC_DMEM, 15%_FBS, 0.1mM_b-mercaptoethanol, 2mM_l-glutamine, 0.1mM_NEA
461	GSM325394.CEL	477	GSE12982	Orkin SH	2008.12.12	E14, ATCC: CRL-1821	NONE	[NC_DMEM, 15%_FBS, 0.1mM_b-mercaptoethanol, 2mM_l-glutamine, 0.1mM_NEA
462	GSM325395.CEL	478	GSE12982	Orkin SH	2008.12.12	E14, ATCC: CRL-1821	NONE	[NC_DMEM, 15%_FBS, 0.1mM_b-mercaptoethanol, 2mM_l-glutamine, 0.1mM_NEA
463	GSM325396.CEL	479	GSE12982	Orkin SH	2008.12.12	E14, ATCC: CRL-1821	NONE	[NC_DMEM, 15%_FBS, 0.1mM_b-mercaptoethanol, 2mM_l-glutamine, 0.1mM_NEA
464	GSM325397.CEL	480	GSE12982	Orkin SH	2008.12.12	"CJ7", FEEDERS	Ezh2 -/-	[NC_DMEM, 15%_FBS, 0.1mM_b-mercaptoethanol, 2mM_l-glutamine, 0.1mM_NEA
465	GSM325398.CEL	481	GSE12982	Orkin SH	2008.12.12	"CJ7", FEEDERS	Ezh2 -/-	[NC_DMEM, 15%_FBS, 0.1mM_b-mercaptoethanol, 2mM_l-glutamine, 0.1mM_NEA
466	GSM325399.CEL	482	GSE12982	Orkin SH	2008.12.12	"CJ7", FEEDERS	Ezh2 -/-	[NC_DMEM, 15%_FBS, 0.1mM_b-mercaptoethanol, 2mM_l-glutamine, 0.1mM_NEA

Figure 2.3: Example screenshot of manual annotations of matrix N1101.

Chapter 3 –

Investigation of Links Between Annotations, Sample Similarity and Transcriptional Profiles in the HPM Matrix Using RaSToVa and DALGES

3.1 Research questions

The “batch effect” is already known to have a potentially large impact on the results of, and, consequently, inferences drawn from microarray data (Leek et al. 2010). The batch effect, as a phenomenon, primarily affects samples that come from the same laboratory and experiment and are processed, therefore, in “batches” sent for processing either at the same time or within a short space of time of each other (Leek et al. 2010). Potential sources of this batch effect have been both demonstrated and others hypothesised in the literature (for a recent review, see (Lazar et al. 2012), particularly figure 1). As batch effects are primarily a within-experiment phenomenon, the performing of large-scale analyses may be one way to mitigate its effects, as the combination of many samples may “smooth out” effects which occur within one experiment.

With the assembly of the HPM matrix in chapter 2, the first question of this chapter is whether or not, in this data, similarity between samples appears to be more strongly linked to one annotation or another. The term “annotation” here is similar to the use of this term in chapter 2, where “annotation” pertains to the available annotations of a given microarray sample, for example as to the cell line used. For example, whilst samples from the same laboratory may well be more similar to each other due to said batch effects, these samples may share a given amount of similarity as they all used the same cell line in their analyses, or simply were from the same experiment. Calculation of a metric which both quantifies and allows for the direct intercomparison between the amounts of sample similarity that any annotation is responsible for in a given data matrix would be a very useful tool, particularly in the field of ES cell biology as considerable interest exists in the field of mESC biology regarding differences between mESC cell lines (Schulz et al. 2009). It is critical to note that this work is not directed at all towards correcting for batch effects, as this is already the subject of much competing work by teams of bioinformaticians (Lazar et al. 2012). This exploratory work rather attempts to ask which of two annotations, those of “cell line” or “source laboratory” appears to be responsible for more sample similarity in the HPM matrix generated in chapter 2, as lab-specific signatures of microarray samples in mESC biology are already a known phenomenon (Newman and Cooper 2010).

This chapter then goes on to ask whether or not this methodology (described in section 3.2.2) can be extended in order to link transcriptional profiles to samples annotated as being from three mESC cell lines (ESD3, E14, CGR8), that are highly represented in the HPM matrix, as well as for iPS cell

lines all generated by forced expression of the canonical Yamanaka factors Oct4, Sox2, Klf4 and c-Myc (Takahashi and Yamanaka 2006). The relevance to mESC biology behind including this iPS cell line, as exploratory work, is apparent when considering that in an analysis of 2 human iPS lines, work by (Marchetto et al. 2009) suggested that “memory” may exist in iPS cell lines due to donor cell type and, crucially, reprogramming method. It is quite likely that such differences exist between mouse iPS cell lines also, although insufficient numbers of different iPS lines were in the HPM matrix to compare multiple mouse iPS lines, and this therefore remains as future work, following proof-of-concept in this work. Finally, using the resulting lists of genes which appear to be linked, in their expression, to these four selected cell lines, this chapter analyses these lists for enrichment for biological pathways which may therefore be suggested to be also linked to these chosen cell lines in this data. Should this prove possible in this dataset, this methodology could therefore be applied in larger datasets with a view to investigating whether these genes and / or enrichments for biological pathways translate into functional differences between mESC cell lines, as differences between cell lines are already known to have significant effects on their biology, such as the endogenous production of Wnts by some mESC lines rendering them more permissive to derivation (ten Berge et al. 2011).

3.2 Methods

3.2.1 Development of RaSToVa: a method to quantify contribution to sample similarity of annotations in microarray data

In order to ascertain the extent to which source laboratory and cell line contribute to sample similarity, a method was required which directly quantified these effects in a manner which allowed intercomparison of contribution to sample similarity of different annotations. That is to say that whilst it may prove trivial to demonstrate that the annotations of source laboratory or chosen cell line somewhat unsurprisingly have an effect on the transcriptional profile of any given sample, it is considerably more difficult, and therefore the objective of developing this method, to ascribe a relative “strength” of those effects in any meaningful manner.

The method first requires the provision of full, manual annotation of the data, as was carried out in 2.2.2. Whilst the approach taken in annotating the HPM matrix was done with best effort using available online literature, this need not be done to ask questions of large-scale data such as those in this chapter. For example, the manual annotations developed in chapter 2 included all information available about each culture condition(s) that could be found to which the cells had been exposed. This took up the vast majority of the time in performing the manual annotation, but would not need to be done to ask simpler questions of the data regarding source laboratory or choice of cell line, as in this chapter. Therefore the method need not be as costly in terms of the prior manual groundwork necessary as the manual annotations in chapter 2 would suggest.

In the case of this work, it came to light as the annotations were being completed, that some source laboratories / cell lines are more represented than others, sometimes greatly so. This is an integral problem when dealing with the problem of investigating the effect of different annotations to transcriptional profiles; the data that is publicly available is not generated for this purpose and thus does not come in neat, controlled, equally-sized groups which would facilitate easy intercomparison. It was therefore decided to use a resampling and random permutation-based method in order to ask questions of the link between annotation and sample similarity.

The method which was developed was named RaSToVa, as the method involves the use of submatrices of the data being assessed for their total variability, giving the name Random Submatrix Total Variability (RaSToVa.)

3.2.2 RaSToVa methodology overview

The methodology of RaSToVa begins with the selection of a single annotation field (e.g. source laboratory) which is chosen for the first analysis. In this case, a list of source laboratories that contributed which samples to the data is generated from the full annotations.

For each laboratory, a submatrix consisting of all the samples from this laboratory is copied from the full matrix. Next, a metric is calculated to quantify the amount of dissimilarity/variability in this submatrix. It was decided to use two different metrics here in order to ascertain the behaviour of RaSToVa when different metrics were used. This was done with a view to demonstrating that the method does not return different conclusions when different metrics are applied. The two metrics chosen were firstly one of total dissimilarity (summed Euclidean distances), and another, more complex-to-calculate metric of information content / variability (as quantified by a normalised form of Shannon entropy (Shannon 1948)). Euclidean distance was chosen as the first metric as Euclidean distance is already a common metric used in grouping similar microarray samples together for analysis (Quakenbush 2001). Shannon entropy was chosen as the other metric to test as it is a direct measure of variability which lends itself to being expressed in a meaningful manner, *id est*, scaled between 0 (no information content, no unpredictability) and 1 (maximal information content, maximum unpredictability.) For how this normalisation is carried out, see section 3.2.4.

This calculated metric (whether Euclidean distance or normalised Shannon entropy) serves as an indicator of how much variability there is in a submatrix made up entirely of samples from this one laboratory. The process is then repeated by sourcing samples from the whole matrix at random, allowing for the choosing of samples more than once, *id est* with replacement, until another submatrix is created containing the same number of samples as the first submatrix which contained only the samples from the current laboratory of interest. These randomly-permuted submatrices have the same calculations for normalised Shannon entropy and summed Euclidean distance performed on them. All of the results from these randomly-permuted submatrices come together to form an array of results referred to as the “expecteds”. By “expected” is meant the resulting total

Euclidean distances / normalised Shannon entropies of the randomly-permuted submatrices. The word “expected” is used as the distance / variability metrics calculated from these randomly-permuted submatrices represent the expected amount of total distance / variability likely to be found if the annotations (as to source laboratory) were to be randomised. This is equivalent to asking the question “If there were no effect on transcriptional profile due to the source laboratory, what variability would we expect to see in a matrix with the same distribution as our entire dataset, given x number of samples, where x is equal to the number of samples in that one laboratory?”

This results in a set of values, each representative of the amount of variability / distance present within submatrices of the data. As the numbers of samples contributed by any one source laboratory are likely to be different, these numbers may vary in their magnitude considerably. To address this, each “expected” value is expressed as a ratio to the “observed”, where the observed value is the total Euclidean distance or Shannon entropy of the submatrix wherein the samples are all from the same laboratory (the first submatrix that was made.) This brings all of the resulting values into line for direct comparability.

It would be expected that if, in this case dealing with source laboratory as the annotation of interest, the source laboratory has an effect on the transcriptional profile wherein samples from the same laboratory are more similar than those from different laboratories, that we would observe that submatrices generated from randomly-selected samples would have a larger variability than observed in the submatrix made of only samples from that laboratory. Conversely, we would reject this idea of laboratory increasing sample similarity if the randomly-permuted submatrices are of similar (or, indeed, less) total Euclidean distance / normalised Shannon entropy to the “observed” values for their associated laboratories.

This analysis can be performed for any annotation, with identical methodology. In the case of this work, comparisons of contributions to transcriptional profile were carried out for the annotations of source laboratory and sample cell line. A graphical representation of this methodology is provided in figure 3.1.

3.2.3 Discretisation of the HPM matrix for use with RaSToVa

As RaSToVa uses Shannon entropy as one of its metrics for quantifying variation in a submatrix of a dataset, the HPM matrix required discretisation. Calculation of Shannon entropy as a measure of

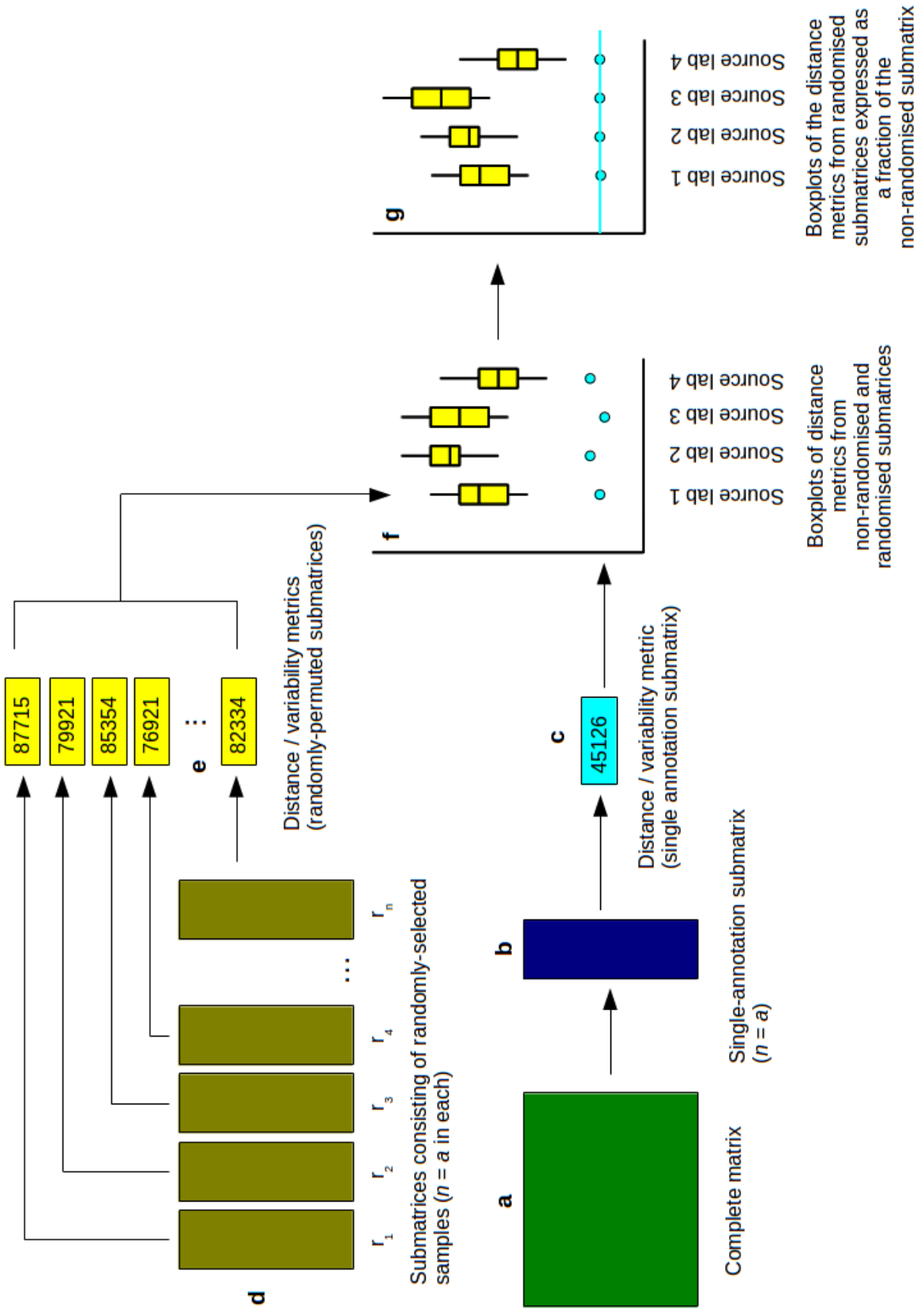

Figure 3.1: RaSToVa methodology overview. From a complete dataset of many microarray samples (a), a single annotation (e.g. all samples from one cell line) submatrix is copied (b) and a variability metric calculated (c). Randomly-permuted submatrices of equal size to (b) are also drawn from (a) and variability calculated for each of these (e). These are combined into a results table (f) for each annotation (e.g. cell lines). All results are then expressed as ratios of the total variability of the random submatrices chosen in (d), to the intact annotation matrix chosen in (b).

information content or variability requires the discretisation of continuous data into bins. It is usual to have these bins as being of equal size between the minimum value and maximum value of the data. The selection of an appropriate number of “bins” into which the data is divided essentially is a balancing act between two extremes. At one extreme, where all data is in one bin, there can be no variability to report; all data falls into this one bin and thus provides no informative result. As the number of bins increases, the resolution of the data which is captured increases, providing an ever more accurate picture of the variability of the data. Low numbers of bins, however, may group datapoints crudely, failing to reflect subtler patterns. At the opposite extreme, Shannon entropy may return meaningless results as the bins are extremely small in size, effectively causing every datapoint to reside in its own bin. This would result in great amounts of entropy being reported, tempered only by the large number of bins which would remain empty. In this case, however, all entropies would appear to be the same, uniform values, where there would be a number (where this number is equal to the number of datapoints) of bins with a single datapoint each and all other bins would be empty. Between these extremes must, therefore, reside a point which we can estimate to be an acceptable number of bins wherein the data's variation is captured sufficiently, but where increasing the number of bins beyond this point begins to suffer from diminishing returns as regards added information / variability detection in the data. Therefore, looking for this point avoids arbitrarily choosing a number of bins which may suffer from a milder form of one of the above issues.

Practical issues also bring to light the need to optimise the number of bins used in calculating Shannon entropy in this work. As more bins are added, the computational power and computational memory required increases greatly, as the discretisation of datapoints is done on a per-probe basis, meaning that an increase of only one extra bin translates to a further 45,101 bins (given that the Affymetrix Mouse 430 v2 microarray contains 45,101 probes) potentially filled by some data points. This can make discretisation of a whole matrix with a large number of bins a very memory-intensive task. This computational memory issue was greatly lessened by not performing discretisation of the entire matrix, but one probe at a time. This is important if this method is to be used outside of specialist bioinformatic computing facilities. The processing time remains an issue which cannot be bypassed (although some mitigation of it is possible as was necessary in section 3.5.3) and greater numbers of bins will increase processing time. Therefore it is of great importance, particularly if these methods are to be applied to ever larger datasets beyond the scope of this work, and even as more annotations are analysed from the same matrix, to optimise the number of necessary calculations. Given that this work also uses approaches involving random permutations,

the problems of performing unnecessary calculations is compounded multiplicatively for every random permutation carried out.

3.2.4 Normalisation of Shannon entropy

Normalisation of the results of calculating Shannon entropy was also carried out. This normalisation is necessary in order to render results for entropy calculation comparable when using differing numbers of bins. This was therefore necessary in order to make sense of the values generated when looking for the optimum number of bins for discretisation of the HPM matrix (see 3.2.5.) In this normalisation, a maximum value for entropy ($H(X)_{max}$) is calculated wherein it is assumed that every bin contains a single datapoint. This reflects the maximum possible entropy (for this number of bins) as the distribution is perfectly spread across all bins with no bins empty; effectively maximum variability where, from an information theoretic point of view, there is no greater or lesser likelihood of a datapoint from the given distribution occurring in one bin or another. By having this as a set maximum, it is trivial to then express any result of calculating Shannon entropy as a fraction of this maximum possible entropy:

$$H(X)_{observed} / H(X)_{max}$$

This gives an easy-to-compare and considerably more human-friendly metric which ranges from 0 (for no variability at all, perfect predictability) to 1 (maximum variability, no predictability) and means that entropies can now be directly compared even when bins numbers are different. This is crucial for the following steps wherein the results of performing entropy calculations, searching for the “ideal” number of bins, must be compared to one another.

3.2.5 Justification of number of bins into which the HPM matrix should be divided for entropy measurement

In order to estimate the optimal number of bins into which the HPM matrix's mRNA detection data should be divided, many runthroughs of normalised Shannon entropy calculation for the HPM matrix were required. For each runthrough, each probe is discretised into the test number of bins and its entropy calculated. When a list of the entropies of all probes is complete, they are totalled and divided by the number of probes on the microarray (in this case, 45,101.) This represents,

therefore, the mean amount of variability found in probes of the HPM matrix when this number of bins is used. The mean entropy of probes in the HPM matrix when different bin numbers were used is plotted against these test numbers of bins in figure 3.2. This figure shows that it is around the 110-bin mark that increasing the number of bins begins to result in far less noticeable increases in mean probe entropy. This was therefore chosen as an acceptable balance between the two extremes mentioned earlier in section 3.2.3.

3.2.6 Justification of the number of permutations required for RaSToVa analysing the HPM matrix

With any resampling or random permutation-based method, a question arises concerning the number of permutations which may be deemed sufficient in order for the method to make reasonable, defensible inferences (some of which can have levels of statistical significance attributed to them) about the patterning of the data. With too few permutations, the data's distribution will not be adequately captured or assessed, leaving its conclusions weak and of little use. However, it is not feasible, or indeed desirable, to simply choose an arbitrary, and very large, number of permutations for the reasons of time constraints, computing power and, importantly, extension to future, larger data matrices. By justifying the choice of an appropriate (*id est* sufficient), but not excessive, number of permutations, computational resources are used efficiently. When planning the use of this method on other / larger matrices in future work (see end-of-chapter discussion), it will likely prove very useful to have this clearly-defined way of ascertaining the number of permutations which is sufficient, rather than the aforementioned, misguided “as many as possible” approach. As the size and distribution of the data changes in other analyses, the “acceptable” number of permutations will likely change also.

In the case of RaSToVa's analysis of the HPM matrix, an acceptable number of permutations was chosen by mimicking the core functionality of RaSToVa by randomly selecting submatrices (of size n = 5 samples and then size n = 20 samples) from the HPM matrix and calculating the total Euclidean distance of each one. For each set of total Euclidean distances for one number of permutations, the standard deviation of these values was calculated. By plotting the standard deviations of each runthrough (being for a different number of permutations), a point could be chosen at which there was stability in the standard deviations calculated. This point would suggest

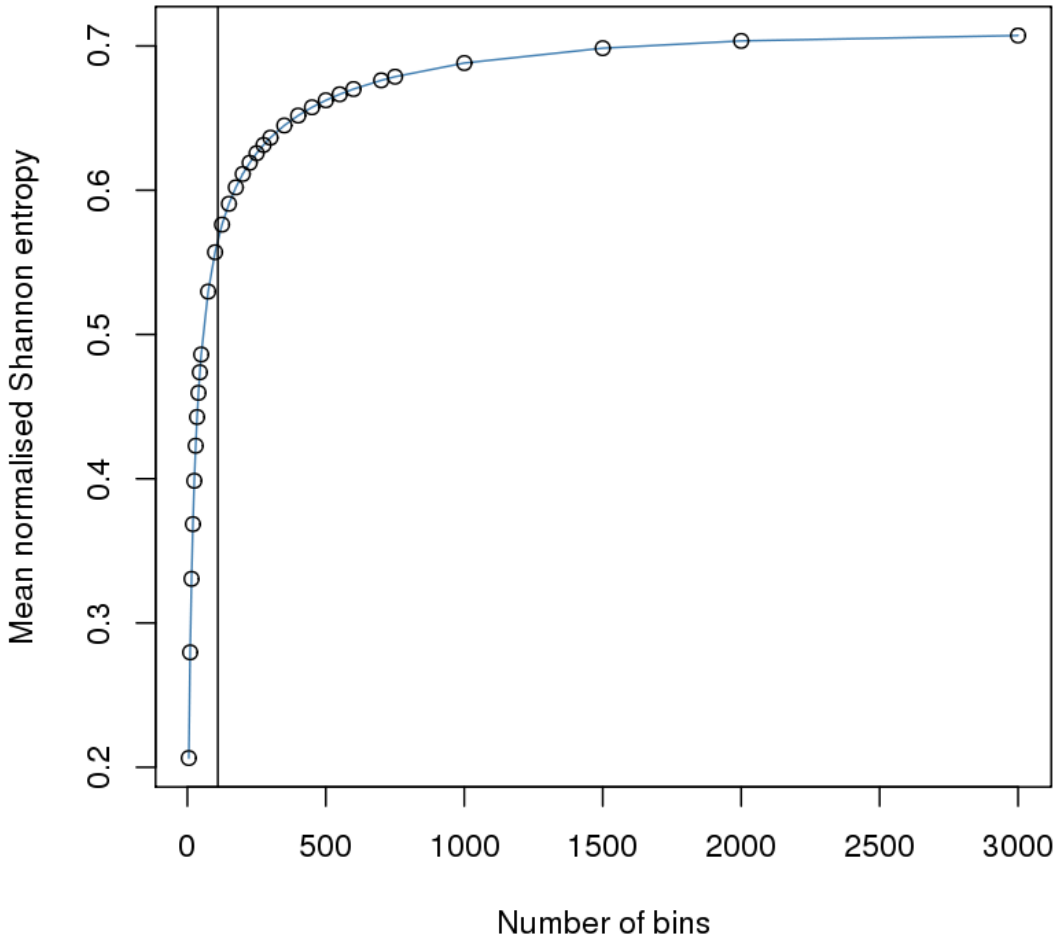

Figure 3.2: Relationship between the number of bins into which the HPM matrix was divided, and the mean normalised Shannon entropy of each probe. The vertical line denotes the number of bins that was chosen wherein the mean normalised Shannon entropy of each probe began to show progressively less increase when increasing the number of bins used for discretisation, being 110 bins.

that the number of permutations was sufficient to capture the overall pattern of the data. The results of this test are shown in figure 3.3.

The chosen number of permutations which results from this test can then be used in running RaSToVa either with another distance metric, or when using RaSToVa to analyse the same matrix for the contribution to sample similarity of other annotations. If the same annotations ("source laboratory") is to be analysed in this matrix in any other way, the same number of permutations can be used in order to capture the patterning of the HPM matrix, such as was done in sections 3.5 and after.

3.3 RaSToVa results investigating cell line versus source laboratory annotations in the HPM matrix

To quantify the contribution to sample similarity of the "source laboratory" annotation in the HPM matrix, RaSToVa was run using the accompanying annotations for all source laboratories which contributed to the HPM matrix, but did not attempt to quantify the contribution to sample similarity for any source laboratories with fewer than 5 samples contributed by that laboratory to the data. This was due to the fact that RaSToVa compares the "intact annotation matrix", *id est* the set of samples which make up that source laboratory, and compares it repeatedly to randomly-permuted submatrices of the same size. As the number of microarray samples included in these submatrices, both intact and randomised, decreases, the distance metrics calculated, and, consequently, the ratios of total distance between them, become less and less able to capture the distribution of the data. The nature of the data, in this case the HPM matrix, will dictate the minimum number of samples required in order for RaSToVa to deliver meaningful results. With a minimum of 5 samples contributed to the HPM matrix in order for a source laboratory to qualify for inclusion, it would still be possible, at a later time, to exclude those laboratories which contributed the lowest numbers of samples to the HPM matrix if this was necessary. The number 5 was therefore simply chosen to save RaSToVa from spending a great deal of time generating results which were extremely likely to be excluded.

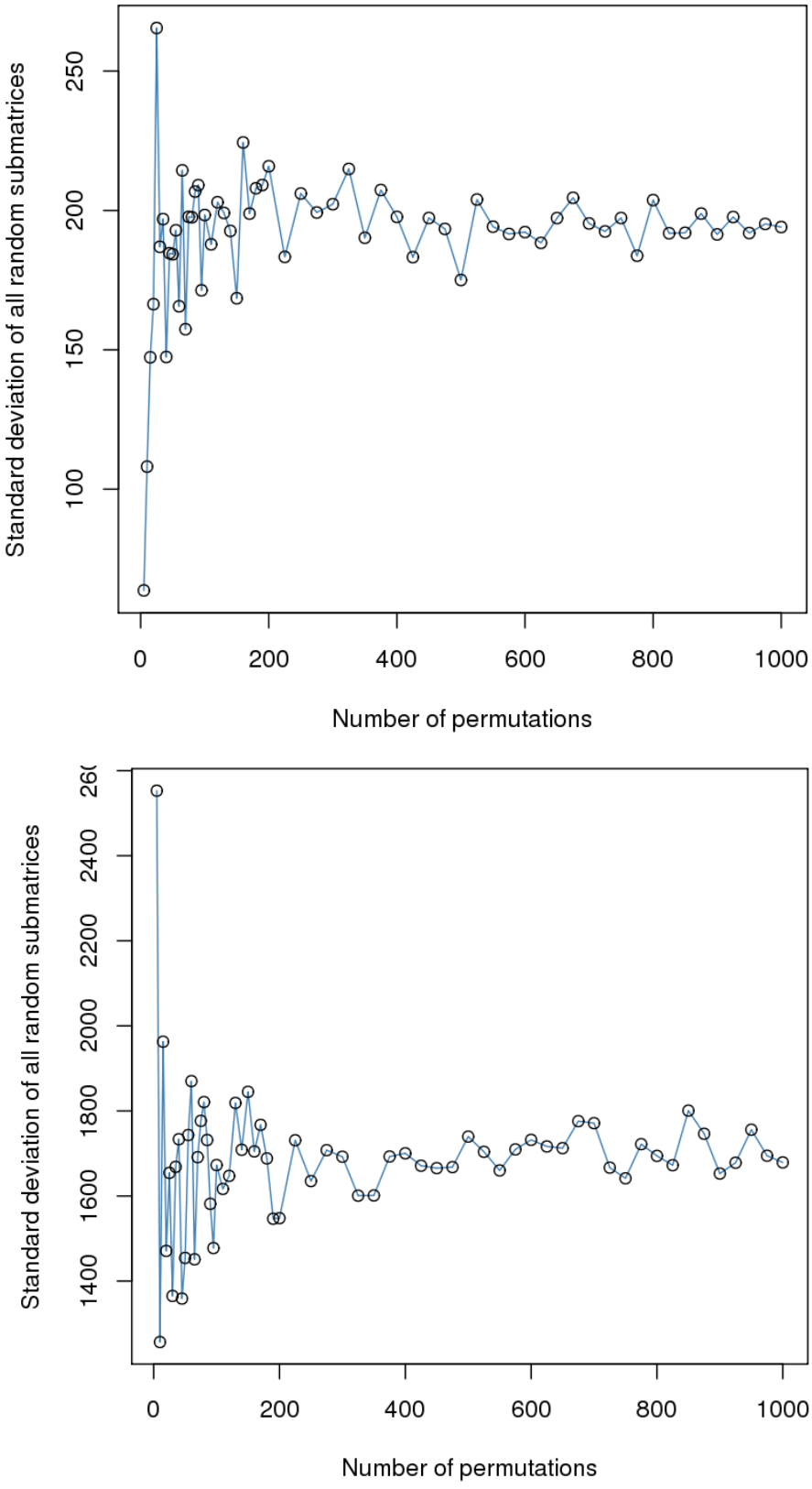

Figure 3.3: Standard deviations of total euclidean distances of randomly-chosen submatrices drawn from the HPM matrix plotted against the number of submatrices generated, showing that after around the 200-permutation mark, the standard deviation of the lists of total euclidean distances stabilises. This was true both when using 5 samples in each random submatrix (upper plot) and when using 20 samples in each random submatrix (lower plot.)

3.3.1 More than 200 permutations captures patterning in the HPM matrix for cell line and source laboratory analyses

The standard deviations which result from the testing of RaSToVa as detailed in 3.2.6 are shown in figure 3.3. Here it is clear to see that at the lower numbers of permutations, the standard deviation of all total Euclidean distances of the submatrices selected fluctuates greatly. This large variation in the standard deviations continues up until around the 200-permutation mark. These lower numbers of permutations are therefore not likely to capture the pattern of the data sufficiently for RaSToVa's purposes.

As the number of permutations increases past the 200 mark in figure 3.3, this fluctuation greatly reduces and the large fluctuations do not recur. This suggests that the number of permutations is now adequately capturing the patterning of the HPM matrix such that further increases in permutation numbers do not noticeably affect the ability of RaSToVa to capture the “lay of the land” of the data, at least with respect to the distribution of source laboratories. With this observation in mind, 500 permutations was deemed more than sufficient for RaSToVa to use when analysing the HPM matrix for the contribution to sample similarity of the two annotations analysed.

The number of permutations carried out in this section was actually slightly over 500 at 504. This was due to the way in which RaSToVa was run on a large, multi-core bioinformatics server. This machine was capable of running 7 concurrent instances of RaSToVa, allowing for faster completion. As the development of RaSToVa took place over a long period of time, it was, at time of coming to the final runs, most efficient to keep this practice of dividing the total number of permutations desired (in this case, 500) divided over multiple CPU cores in separate instances, and then later combining all of the calculated intact / random ratios into one results table for final plotting and analysis. In order to achieve 500 permutations, each run was performed 72 times, making for the total of (7 instances x 72 permutations = 504 permutations total).

The exact same methodology was applied for both the running of RaSToVa with Euclidean distance as the variation metric and with normalised Shannon entropy (see methodology section 3.2.2) as the variation metric. For the running of RaSToVa to investigate the same phenomenon using “cell line” as the annotation of interest, the same number of permutations (504) was run and the minimum number of samples contributed to the HPM matrix for any given cell line was kept at 5.

3.3.2 RaSToVa results overview

The results of running RaSToVa (504 permutations, minimum of 5 samples contributed from an annotation for inclusion) for the “source laboratory” annotation, using Euclidean distance as the variability metric, can be seen in figures 3.4 and 3.5. This was repeated using normalised Shannon entropy, using the recommended 110 bins from section 3.2.5, and the results of this are shown in figures 3.6 and 3.7. For the investigation of the “cell line” annotation's contribution to sample similarity as detectable in the HPM matrix by RaSToVa, the results are plotted for Euclidean distance as the variability metric in figure 3.8 and with normalised Shannon entropy in figure 3.9.

These boxplots represent the ratio of measured total variability (whether by normalised Shannon entropy or euclidean distance), of individual laboratories / cell lines whose samples contributed to the HPM matrix. For “source laboratory” and “cell line”, and both Euclidean distance and normalised Shannon entropy, the boxplots in all RaSToVa results figures are ordered for each of these experiments by increasing order of the mean of the boxplot, for clarity, trend observation and intercomparison.

In all of RaSToVa's boxplots, the y-axis depicts the ratio of the intact (that is, all from one laboratory) submatrix divided by the randomly-permuted submatrices of the same size as the intact matrix. Each boxplot is generated from 504 points, corresponding to the number of permutations that RaSToVa was run with in this work. There is a line provided across the x-axis of RaSToVa plots where the ratio (intact / random) would be 1. Any point in any boxplot falling below this ratio line indicates that the intact matrix for the denoted source laboratory / cell line contains less variability than would be expected by randomly resampling samples from the HPM matrix. These randomly permuted matrices were made of samples outwith those which make up the laboratory / cell line in question. However, selecting the same sample more than once (but, again, not from the source laboratory / cell line in question) is allowed, *id est*, sampling was done with replacement. Therefore, given the distribution of the HPM matrix outwith the source laboratory / cell line in question, a data point falling below the ratio line supports the hypothesis that samples which are from the same group of the annotation in question are more similar to each other than would be expected if the sample labels were randomised, with the magnitude of that increased similarity related to the distance below the ratio line. This, in turn, can be taken to mean that any data point falling below the ratio line supports the notion that the annotation (“source laboratory” or “cell line”) is definitely

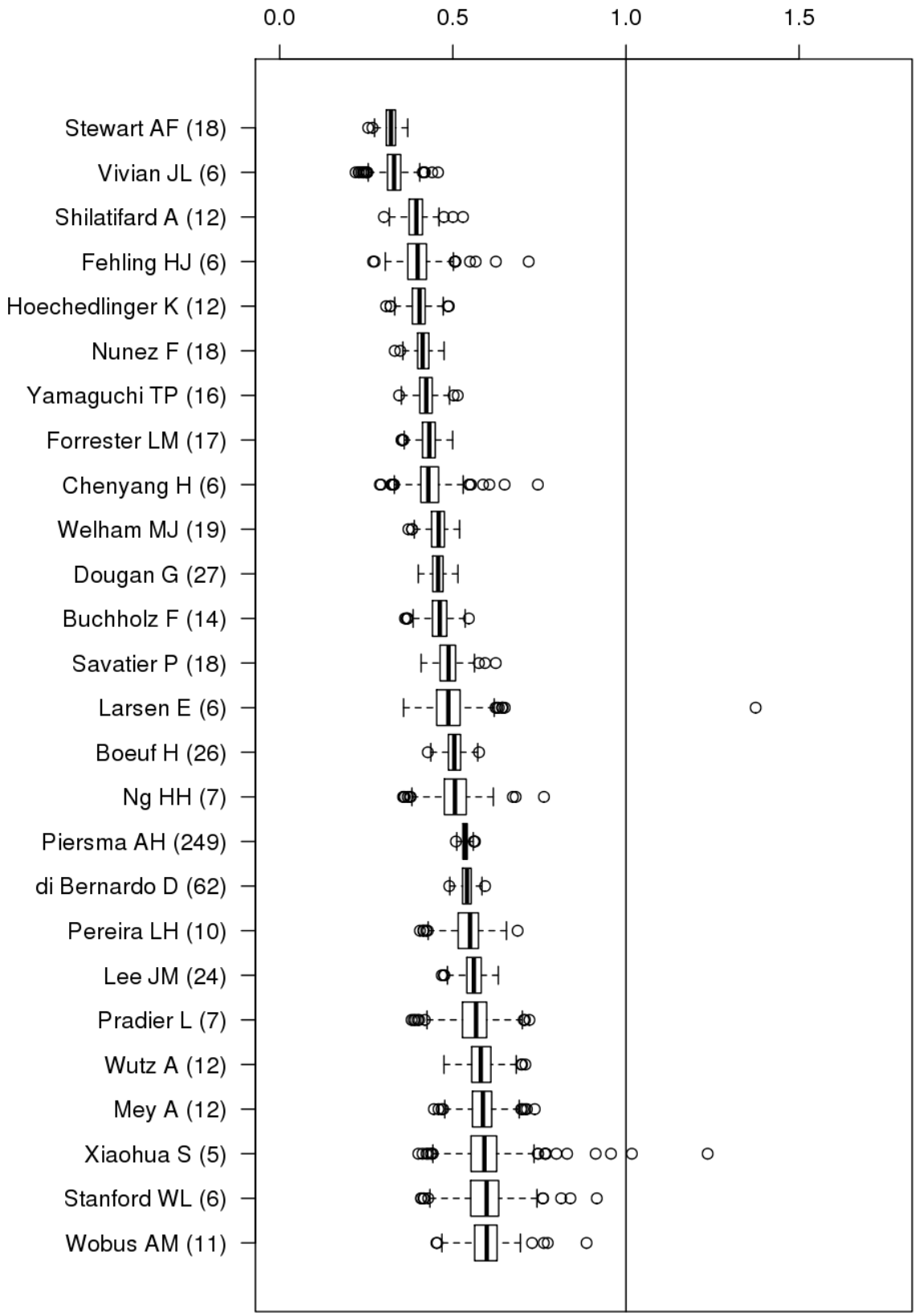

Figure 3.4: Results of running RaSToVa on the HPM matrix (504 total permutations), analysing contribution to sample similarity of the "source laboratory" annotation. Boxplots are made up of ratios between randomly-permuted submatrices to each intact, single-laboratory submatrix (laboratory name provided along with the number of samples this laboratory contributed to the HPM matrix.) Similarity metric: euclidean distance. Plot 1 of 2.

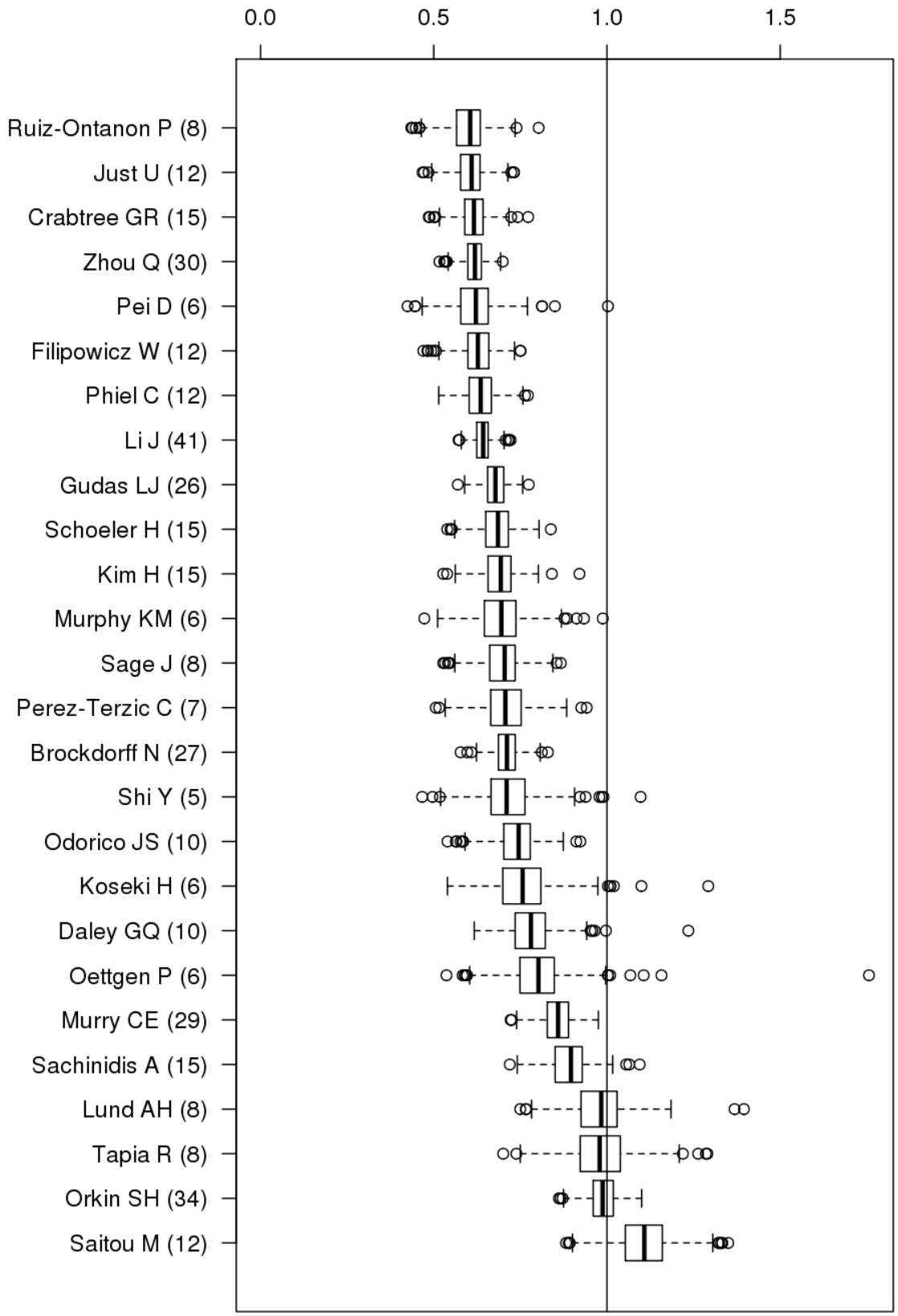

Figure 3.5: Results of running RaSToVa on the HPM matrix (504 total permutations), analysing contribution to sample similarity of the "source laboratory" annotation. Boxplots are made up of ratios between randomly-permuted submatrices to each intact, single-laboratory submatrix (laboratory name provided along with the number of samples this laboratory contributed to the HPM matrix.) Similarity metric: euclidean distance. Plot 2 of 2.

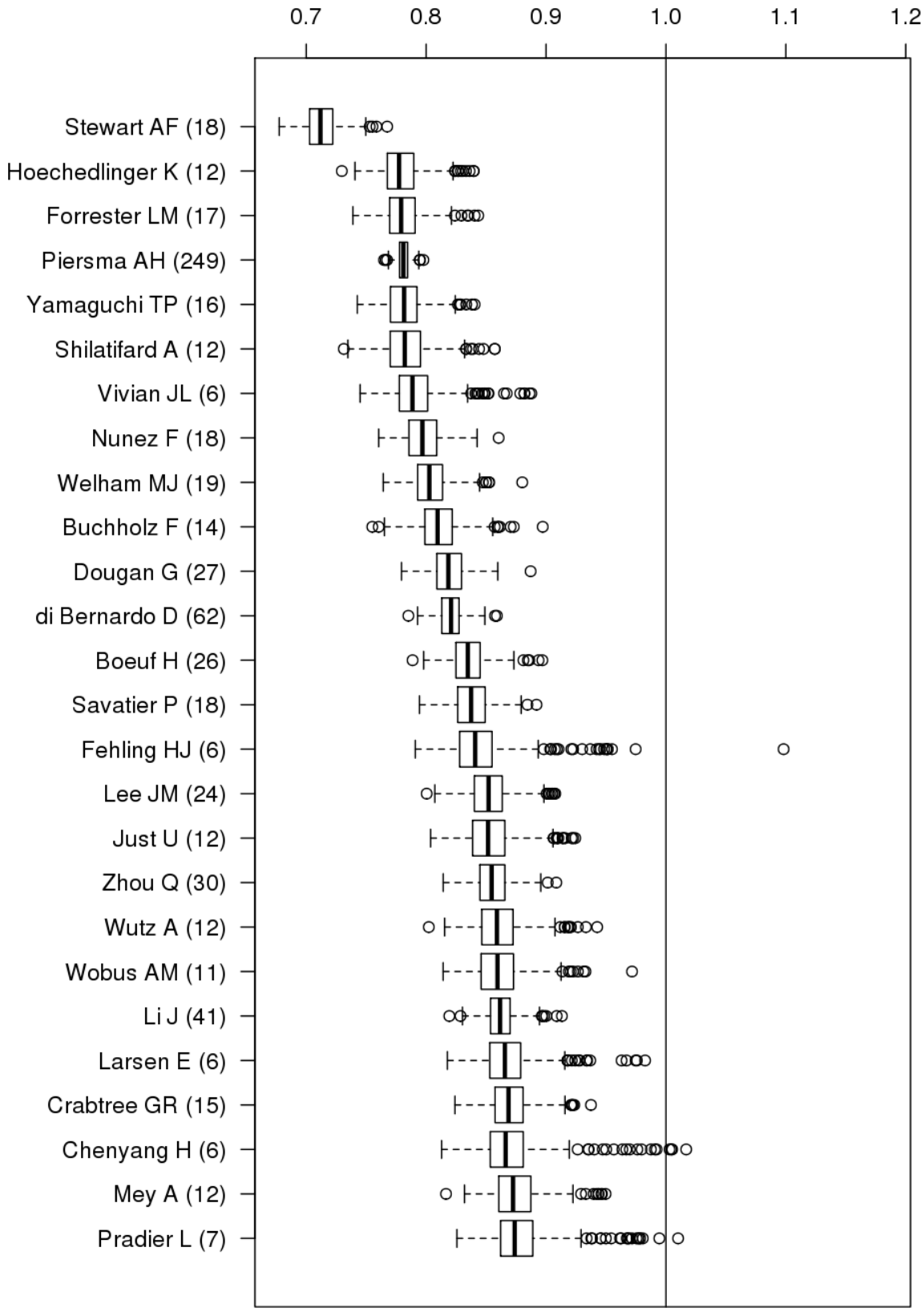

Figure 3.6: Results of running RaSToVa on the HPM matrix (504 total permutations), analysing contribution to sample similarity of the "source laboratory" annotation. Boxplots are made up of ratios between randomly-permuted submatrices to each intact, single-laboratory submatrix (laboratory name provided along with the number of samples this laboratory contributed to the HPM matrix.) Similarity metric: normalised Shannon entropy. Plot 1 of 2.

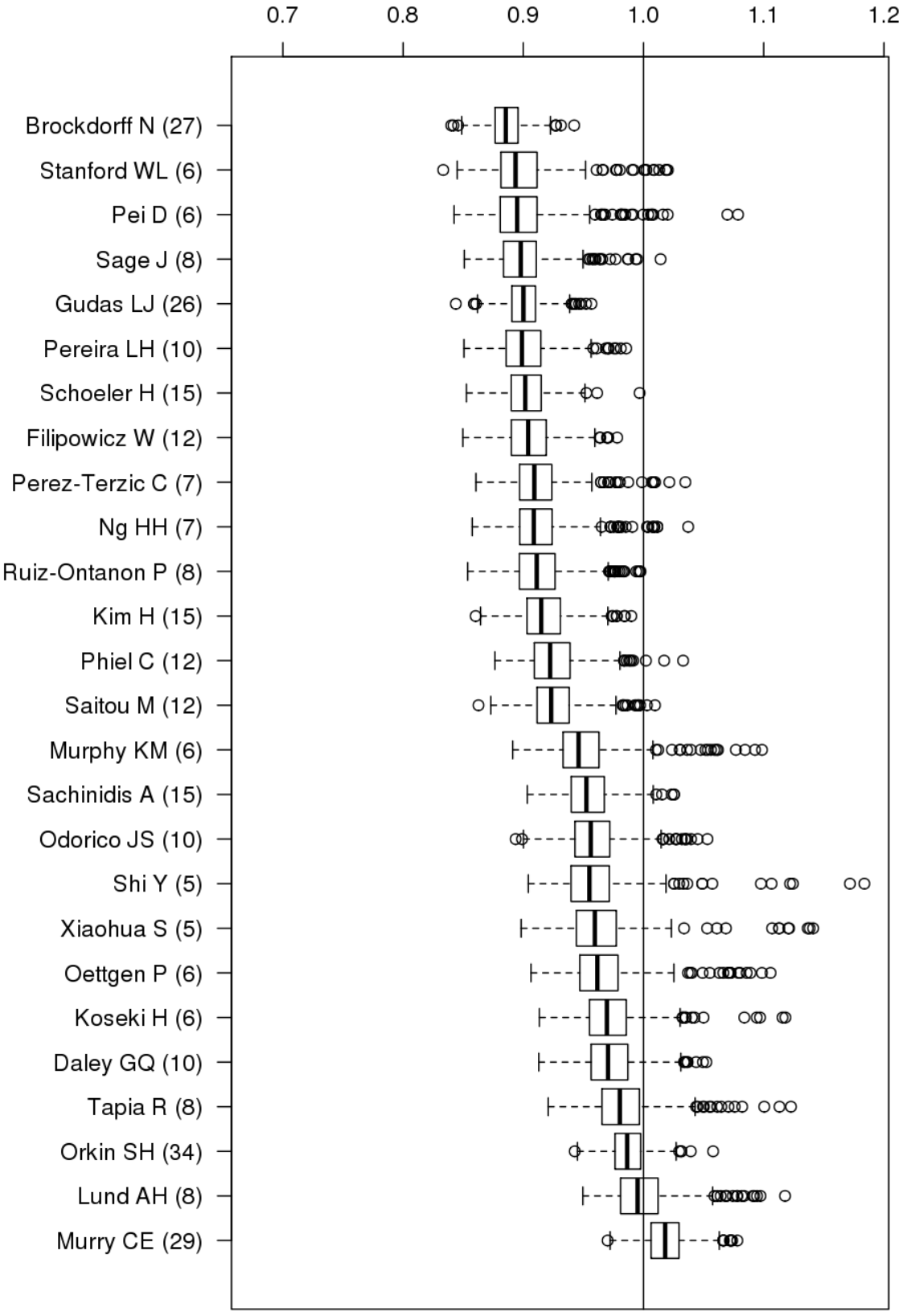

Figure 3.7: Results of running RaSToVa on the HPM matrix (504 total permutations), analysing contribution to sample similarity of the "source laboratory" annotation. Boxplots are made up of ratios between randomly-permuted submatrices to each intact, single-laboratory submatrix (laboratory name provided along with the number of samples this laboratory contributed to the HPM matrix.) Similarity metric: normalised Shannon entropy. Plot 2 of 2.

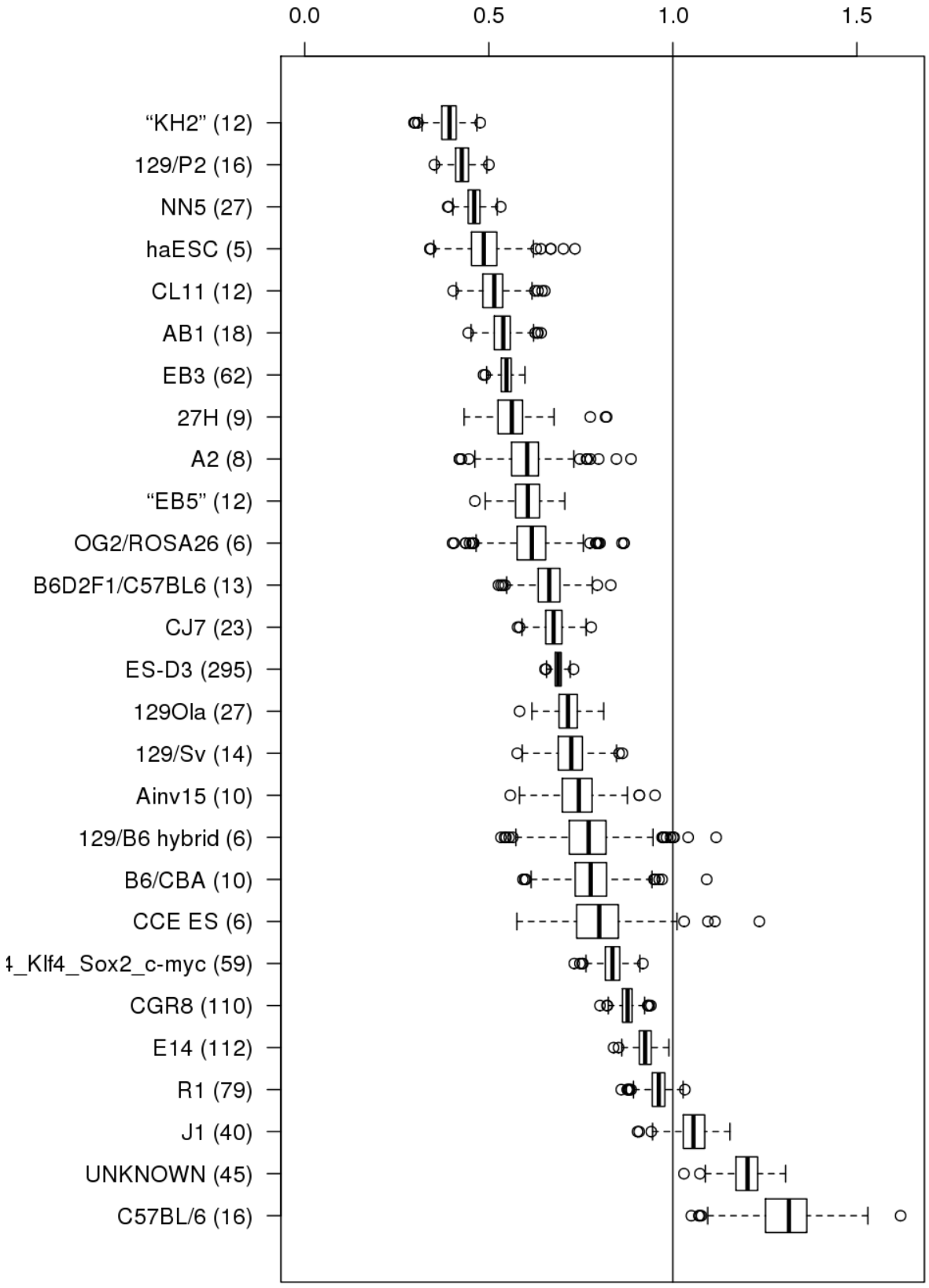

Figure 3.8: Results of running RaSToVa on the HPM matrix (504 total permutations), analysing contribution to sample similarity of the “cell line” annotation. Boxplots are made up of ratios between randomly-permuted submatrices to each intact, single-laboratory submatrix (cell line name provided along with the number of samples this cell line contributed to the HPM matrix.) Similarity metric: euclidean distance.

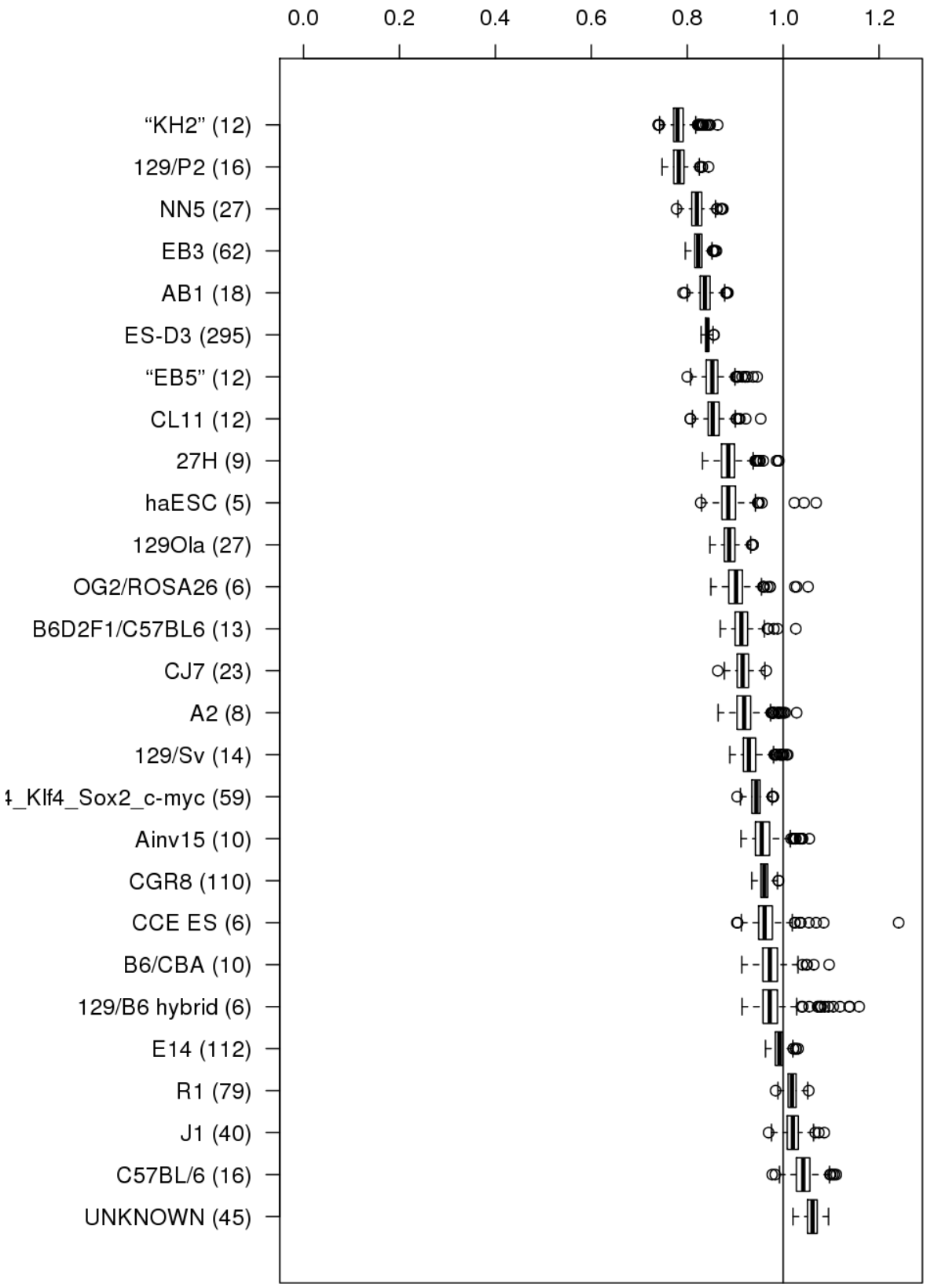

Figure 3.9: Results of running RaSToVa on the HPM matrix (504 total permutations), analysing contribution to sample similarity of the “cell line” annotation. Boxplots are made up of ratios between randomly-permuted submatrices to each intact, single-laboratory submatrix (cell line name provided along with the number of samples this cell line contributed to the HPM matrix.) Similarity metric: normalised Shannon entropy.

associated with samples in the intact matrix being more similar to each other than would be expected if there was no effect on similarity arising from sharing the queried annotation.

The inverse is also true, wherein any data point which rises above the ratio line suggests that being from the same source laboratory / cell line is actually cause for more variability than would be expected if all samples in the HPM matrix had their “source laboratory” or “cell line” labels reassigned randomly.

It was expected that, due to the nature of random permutation, RaSToVa boxplots would occasionally have outlying datapoints generated as a consequence of RaSToVa having compared an annotation-intact submatrix to a randomly-permuted submatrix which included one or more samples with a high similarity or high dissimilarity to the rest of the randomly-permuted submatrix. This would manifest itself as an outlying datapoint on the RaSToVa boxplot results, but would likely, for reasons discussed in section 3.3.3, preferentially affect laboratories or cell lines with lower levels of representation in the HPM matrix. A perfect example of these random occurrences can be seen as wildly-deviant points in the results for the “Larsen E” laboratory in figure 3.4 and “Fehling HJ” laboratory in figure 3.6. Were these random occurrences not chance results due to the assembly of different randomly-permuted submatrices, and due to the distribution of the intact matrix, it would be expected that such outliers would occur regardless of the variability metric chosen. This, as can be seen from the lack of outlying points for the these laboratories when the variability metric is switched, is not the case.

Two important properties of each boxplot should be taken into account when viewing RaSToVa outputs. Firstly and most importantly is the distance between the mean of the boxplot to the ratio line. The larger this distance (if the boxplot mean lies below the ratio line), the stronger the evidence for an “effect” RaSToVa found for the annotation of “source laboratory” or “cell line” increasing the similarity between samples. The opposite is also true, where a mean line noticeably above the ratio line would be a potential cause for concern, with regards to the hypothesis in this work, implying that a contributing laboratory's samples were, in fact, so widely different to everything else in the HPM matrix, that randomly-permuted submatrices of equal size to this hypothetical laboratory actually would be less variable than the submatrix which had that annotation (one laboratory or one cell line) intact. As can be seen from figures 3.4 to 3.9, this, reassuringly, is not in any way the trend amongst laboratories or cell lines.

The second important property of the boxplots that RaSToVa generates is their spread. The more spread a given boxplot has, the more variation there was when comparing randomly-permuted submatrices to the intact matrix of only the samples from that one laboratory or cell line. The behaviour of these two properties, and RaSToVa in general is discussed in section 3.3.3.

3.3.3 Assessment of the behaviour of RaSToVa

The behaviour of RaSToVa was investigated by determining any relationships present between several input and output values.

The first behaviour that was investigated, on viewing the layout of results in figures 3.4 to 3.9, was whether or not a relationship existed between the mean and the standard deviations of the boxplots which RaSToVa generated. This essentially investigated whether or not RaSToVa finding more evidence for an effect of an annotation was related to the spread of the boxplot. There is something of a relationship between the means and standard deviations of the boxplots in both the Euclidean distance and normalised Shannon entropy runs of RaSToVa, with Pearson correlations of ($r = 0.645$) and ($r = 0.576$) respectively for these two variability metrics. This relationship between mean and standard deviation of boxplots did not hold for the investigation of “cell line”'s contribution to sample similarity. The scatterplots from which these two correlations derive are shown in figure 3.10 for the investigation of “source laboratory” and in figure 3.11 for the investigation of “cell line”. It is not known whether or not a relationship would emerge between means and standard deviations of the boxplots were the “cell line” annotation to have been divided into a similar number of subgroups as “source laboratory” was (27 cell lines vs 72 laboratories respectively.)

The second behaviour of RaSToVa that was investigated was possibly the most crucial; whether or not RaSToVa found different “strengths” of the effect on sample similarity dependent on the number of samples that a given laboratory or cell line contributed to the HPM matrix. It is reassuring to see that, regardless of variability metric employed, and across the investigations of both “source laboratory” and “cell line”, there is no discernible relationship between the mean of a boxplot and the number of samples in the HPM matrix bearing that annotation. This is visualised in the scatterplots of boxplot mean versus number of samples, middle two plots of figure 3.10 for the “source laboratory” investigation, and middle two plots of figure 3.11 for the “cell line” investigation. This was reassuring as any relationship between the number of samples contributed to

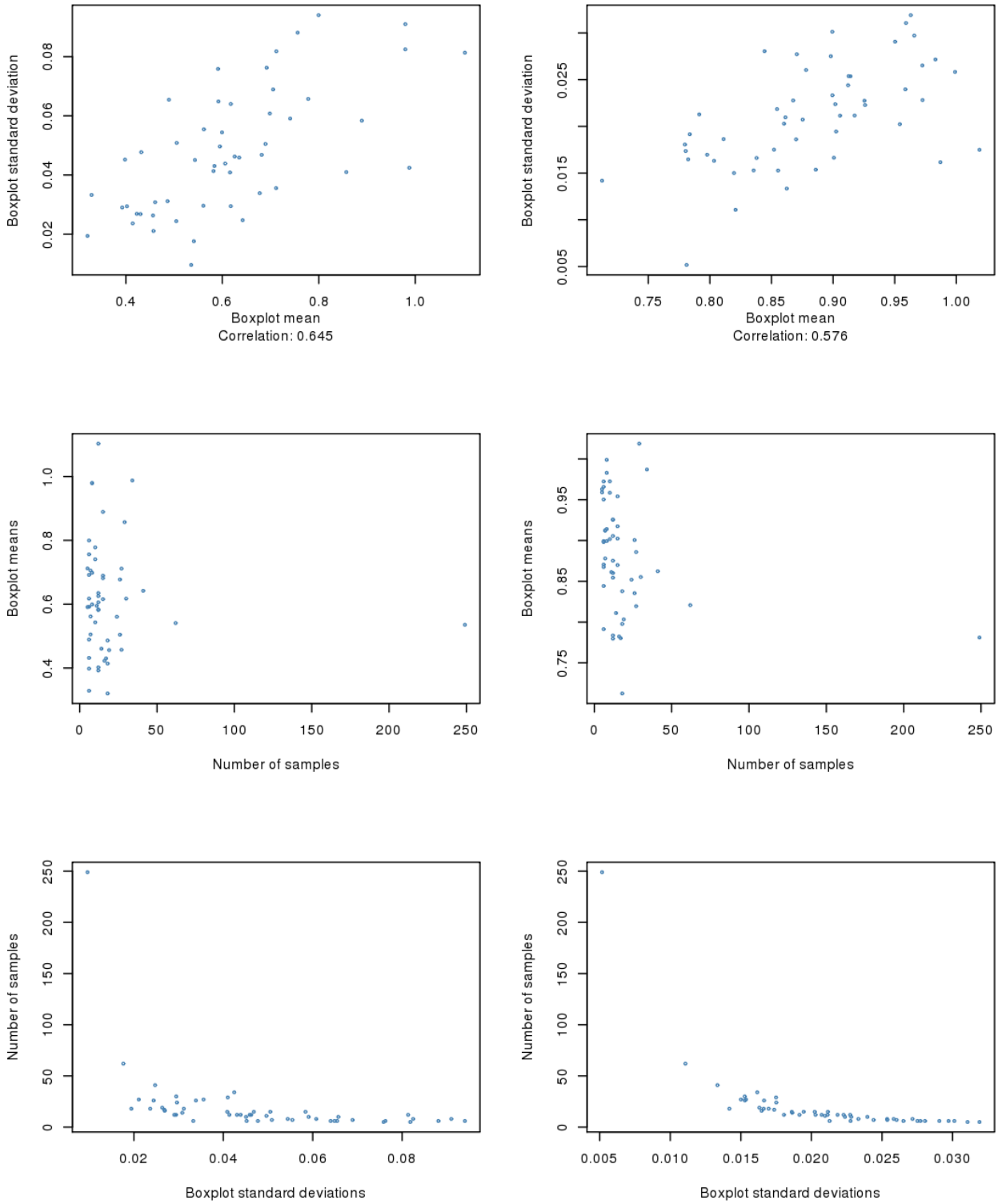

Figure 3.10: Behaviour of the RaSToVa method when using Euclidean distance as the variability metric (left plots) and Shannon entropy metric (right plots) investigating the contribution to sample similarity of the source laboratory annotation. The top two scatterplots show the relationship between boxplot means and standard deviations. The middle two scatterplots show the relationship between the number of samples that a given laboratory contributes to the HPM matrix, and the mean of the boxplot generated for that laboratory. Finally, the lowest two scatterplots show the inverse relationship between the number of samples contributed by laboratory and the standard deviation (spread) of the boxplot which RaSToVa generates for that laboratory.

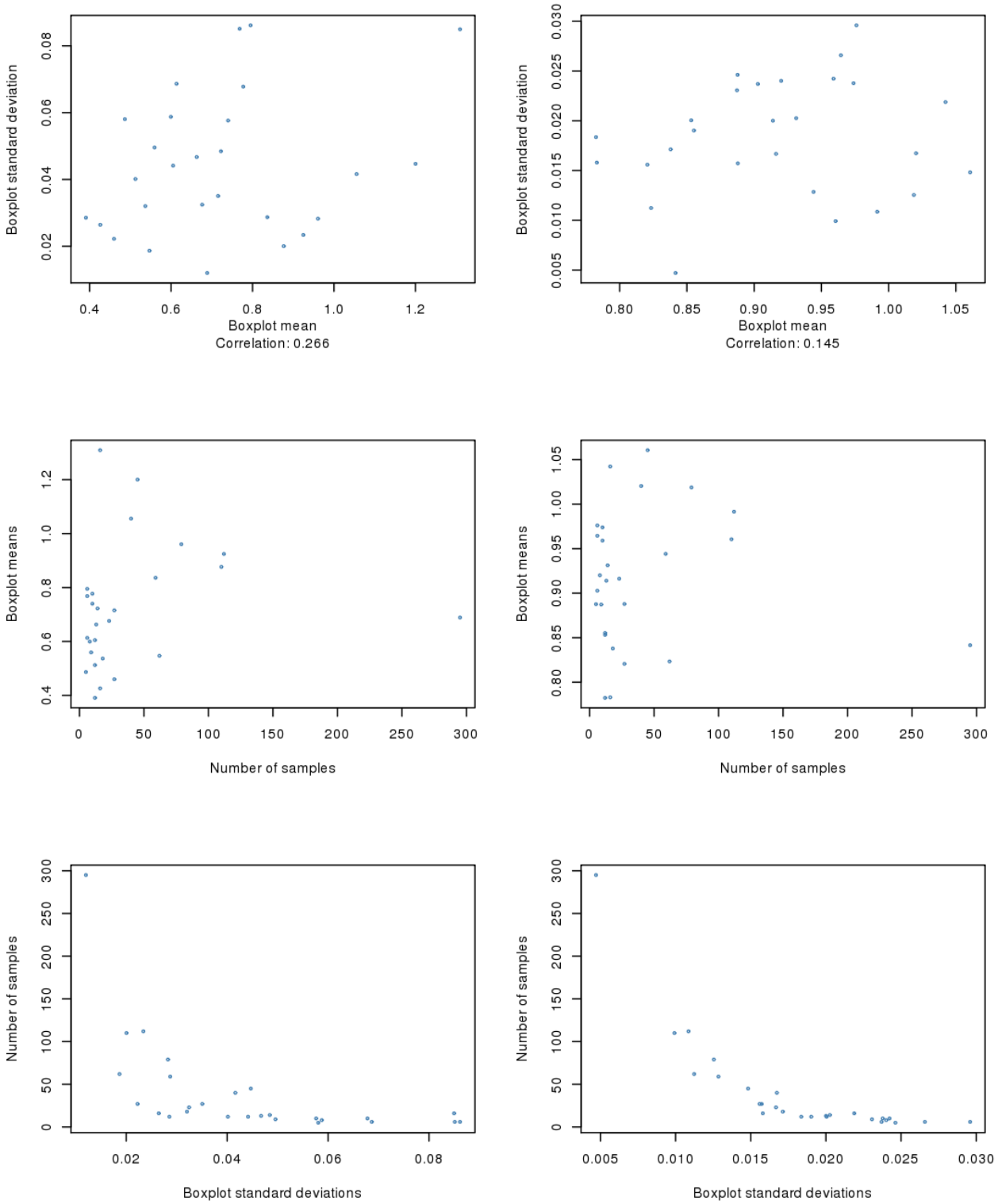

Figure 3.11: Behaviour of the RaSToVa method when using Euclidean distance as the variability metric (left plots) and Shannon entropy metric (right plots) investigating the contribution to sample similarity of the cell line annotation. The top two scatterplots show the relationship between boxplot means and standard deviations. The middle two scatterplots show the relationship between the number of samples that a given cell line contributes to the HPM matrix, and the mean of the boxplot generated for that cell line. Finally, the lowest two scatterplots show the inverse relationship between the number of samples contributed by cell line and the standard deviation (spread) of the boxplot which RaSToVa generates for that cell line.

the HPM matrix and the resulting mean of the boxplot (read: evidence for effect on sample similarity) would have made real an initial concern that perhaps only larger numbers of samples with the same annotation would generate evidence for an effect on sample similarity, with less-represented annotations being attributed either very little evidence for any effect on similarity, or, worse, an inverse relationship between sample number and similarity may have rendered RaSToVa all but meaningless, at least with regard to the distribution of the HPM matrix.

For example, the laboratory which contributed the greatest number of samples to the HPM matrix was the “Piersma AH” laboratory (n = 249 samples). However, this source laboratory does not by any means have the lowest mean. The mean of the boxplot for the Piersma laboratory is 0.535 (3 dp.). This is actually the 17th lowest mean, rather than the lowest. It is the boxplot for the “Stewart AF” laboratory which holds the lowest mean of 0.320 (3 dp.) here, and this laboratory contributed only 18 samples to the HPM matrix. This also holds true when normalised Shannon entropy is used as the variability metric, again with the “Piersma AH” laboratory having a mean of 0.781 (3 dp.), 4th lowest, and the “Stewart AF” laboratory having the mean of 0.712 (3 dp.).

This lack of a relationship between number of samples and boxplot mean is maintained in the “cell line” investigation. Here, in the case of Euclidean distance used as the variability metric, the lowest boxplot mean (0.391 (3 dp.)) is had by the “KH2” cell line, with only 12 samples contributed to the HPM matrix, but a mean of 0.689 (3 dp.) (14th lowest) found for the ES-D3 cell line, with the largest (n = 295) number of samples contributed to the HPM matrix.

When normalised Shannon entropy is used, the ES-D3 cell line has only the 6th lowest mean (0.842 (3 dp.)). The lowest mean, again, is found for the “KH2” cell line, at 0.782 (3 dp.). In addition, it is reassuring to see that the subgroup of cell lines labeled as “UNKNOWN” (therefore likely of mixed actual cell line), has the highest mean here of 1.061 (3 dp.). Being around the 1 mark, this ratio indicates that the samples in the HPM matrix annotated as “UNKNOWN” for the cell line field are roughly as similar to each other as other submatrices randomly pulled from the HPM matrix of an equal size (where size here is n = 45.) This is exactly in accordance with what would be expected. The next highest mean was found for the “C57BL/6” cell line (n = 16 samples), also around the 1 mark (1.200 (3 dp.)), with a similar inference as to their similarity as was inferred for the “UNKNOWN” cell line.

These ranks for “UNKNOWN” and “C57BL/6” are swapped when Euclidean distance is used as the variability metric, however. It is reassuring to see that there is agreement between the two variability metrics tested, as to which cell lines appear at this end of the list.

The third behaviour of RaSToVa which was investigated was any link between the standard deviation of the resulting boxplots and the number of samples which a particular cell line or laboratory contributed to the HPM matrix. Here it was hoped that a relationship would emerge, with higher numbers of samples contributed resulting in less of a spread of that cell line / laboratory's boxplot. Indeed this is the case, as can be seen from the bottom two scatterplots in figure 3.10 for the “source laboratory” investigation and in the bottom two scatterplots of figure 3.11 for the “cell line” investigation. As the number of samples increases, it follows that RaSToVa will be comparing an intact matrix to ever-larger randomly-permuted submatrices. As these randomly-permuted submatrices grow in size, it will be less and less likely that RaSToVa may choose a greater number of similar samples by chance, or a greater number of dissimilar samples by chance. As such, even when an outlying (read: very similar or dissimilar to a sample already chosen) sample is added to a larger randomly-permuted submatrix, this will have less of an effect on the variability metric calculated for that random submatrix.

An extreme of this is clearly visible when observing the boxplots for the “Piersma AH” laboratory and the “ES-D3” cell line in figures 3.4 to 3.9, with both having extremely tight boxplots with hardly any, if any, outlying datapoints. The “Oettgen P” laboratory has the widest spread, with a standard deviation of 0.0940 (4 dp.) when Euclidean distance is used, and has only 6 samples to its name. The same laboratory is 4th from top of the largest standard deviations when Shannon entropy is used, with a standard deviation of 0.0297 (4 dp.). The source laboratory with the largest standard deviation when Shannon entropy is used, however, is the “Xiaohua S” laboratory, with only 5 samples to its name and a standard deviation of 0.0319 (4 dp.). Interestingly, there is a clearer relationship between standard deviation of boxplots and the number of contributed samples when using Shannon entropy than when using Euclidean distance, arguing that perhaps Shannon entropy is better capturing the variability present in these submatrices. This can be seen from the smoothness and lack of noise enjoyed by the Shannon entropy method compared to Euclidean distance in the bottom two plots of figures 3.10 for “source laboratory” and 3.11 for “cell line”.

The final assessment of the behaviour of RaSToVa looks at the agreement between means and standard deviations of all boxplots in all four runs of RaSToVa (2x annotations, 2x variability

metrics). Scatterplots depicting these agreements are shown in figure 3.12. This indicates that, despite any differences in the ranks of means or standard deviations of boxplots generated by RaSTOvA when investigating either “source laboratory” or “cell line”, there is overall very good agreement whether RaSToVa uses Euclidean distance or normalised Shannon entropy as the variability metric.

Full lists of boxplot means and standard deviations for all 4 runs of RaSToVa (2x annotations, 2x variability metrics) are available electronically in “Chapter 3/Tables/RaSToVa/Boxplot Summaries”.

3.3.4 “Source laboratory” annotation effect on HPM matrix sample similarity

The results of RaSToVa in figures 3.4 to 3.7 unequivocally confirm that being from the same source laboratory is most definitely having an effect on sample similarity. When using Euclidean distance as the distance metric for the “source laboratory” analysis (figures 3.4 and 3.5), this is easily seen as the vast majority (all but one, 98.1% (1 dp.)) of the boxplots having their mean lines below the ratio line. Only the “Saitou M” laboratory had its mean above the ratio line. Some other contributing laboratories approached this ratio line, however, such as the Lund AH, Tapia R and Orkin SH laboratories. Potential reasons for this are discussed in section 3.3.3. When using normalised Shannon entropy as the variability metric, the results of RaSToVa are as presented in figures 3.6 and 3.7.

With normalised Shannon entropy as the chosen variability metric, the overall result is still overwhelmingly in favour of source laboratory greatly increasing sample similarity (again, with the caveats put forward in section 3.4 and 3.10.5.) However, there are some notable differences. Firstly, when using Euclidean distance as the distance metric, the “Saitou M” laboratory analysis resulted in a boxplot whose mean lay above the ratio line, whereas when normalised Shannon entropy is used as the variability metric, the mean of the boxplot for this source laboratory is comfortably below the ratio line. This peculiarity, however, does not change the extremely strong evidence for “source laboratory”'s effect being overwhelmingly to increase sample similarity. The “Lund AH” laboratory, however, when normalised Shannon entropy is used, is now the one source laboratory found to be over the ratio line. It is highly unlikely that this is simply a result of differences in the random resampling of the HPM matrix.

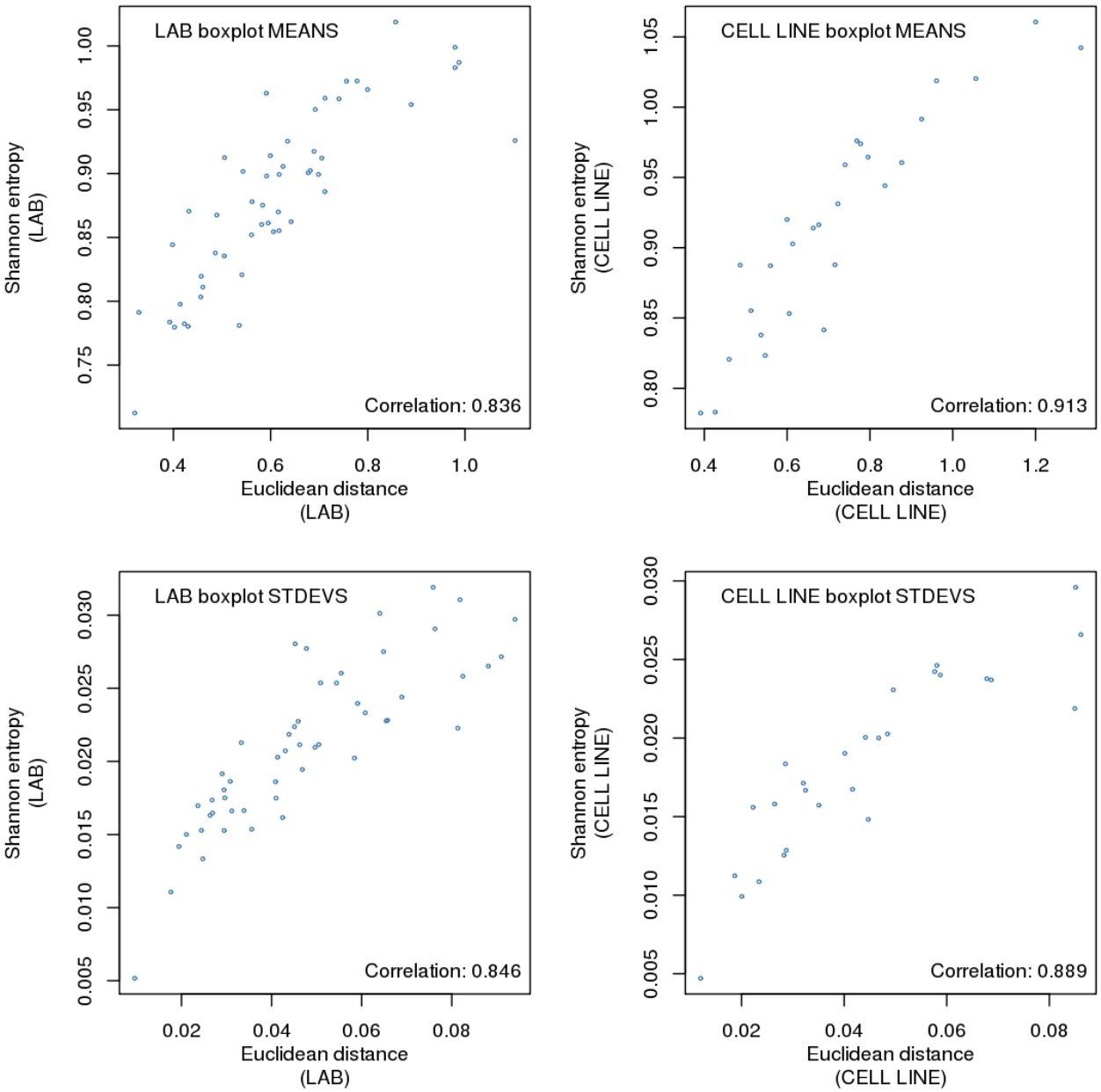

Figure 3.12: Agreement between means and standard deviations of boxplots generated as results of RaSToVa analysis. Good agreement is found with the means and boxplots, whether normalised Shannon entropy or Euclidean distance is used, as shown by the strong positive Pearson correlations found between these metrics. This good agreement was found in both the "source laboratory" (2 leftmost plots) and "cell line" investigations (2 rightmost plots.)

3.3.5 “Cell line” annotation effect on HPM matrix sample similarity

The same overwhelming trend of increasing sample similarity as was found with “source laboratory” is found for cell line, as was expected. From figures 3.8 and 3.9, it can be seen that, again, regardless of variability metric employed, the vast majority of samples grouped by “cell line” show that sharing this cell line annotation increases sample similarity. In the case of using Euclidean distance as a distance metric (figure 3.8), out of 27 cell lines which met the minimum number of 5 contributed samples to the HPM matrix, only 4 of these resulted in a RaSToVa boxplot whose mean fell above the ratio line and only 3 fell above the ratio line when using normalised Shannon entropy (figure 3.8), for majorities of 85.2% (1 dp.) and 88.9% (1 dp.) of boxplots in support of a tendency for the “cell line” annotation to increase sample similarity.

The contributions to sample similarity are further summarised by taking all of the means of the boxplots for the “source laboratory” analysis, and plotting a further boxplot of these means. The result of this is shown in the comparative plots in figure 3.13. Here it can, again, be seen that the evidence is clearly in favour of source laboratory having a greater effect on sample similarity as the boxplots (both when using Euclidean distance or normalised Shannon entropy), are overwhelmingly beneath the ratio line. The same is true of the analysis investigating the “cell line” annotation (see figures 3.8 and 3.9 for Euclidean distance and Shannon entropy analyses respectively).

In order to have accepted a null hypothesis here, for either annotation, in which case samples sharing an annotation would not have increased their similarity, the boxplots for both individual laboratories / cell lines and then when plotting all of these boxplots' means as one boxplot would have needed to centre about the ratio line, indicating that there would be no difference in sample similarity in submatrices chosen from within “source laboratory” or “cell line” groups compared to just randomly picking submatrices of the same size. Limitations and caveats of this approach are to be found in the discussion (5.1.4).

As was the intention from the outset, as all of the boxplots for both laboratory and cell line are expressed as variability ratios between intact annotations and randomly-permuted submatrices, it becomes possible to directly compare these two annotations' contribution to sample similarity, which enabled them to be put on figures opposite each other (figure 3.13) for easy intercomparison.

From this comparative figure, it is suggested that source laboratory has a marginally stronger effect on sample similarity than cell line does. When Euclidean distance is used as the variability metric, a simple Mann-Whitney U test between the values which make up the boxplots in 3.13, the laboratory-versus-cell-line comparison gives a p-value of 0.04895 (when using Euclidean distance to estimate similarity), suggesting that there is a significant different difference in these two annotations' effect on sample similarity. When using normalised Shannon entropy, however, the same Mann-Whitney U test gives a p-value of 0.06065. This second method falls only marginally short of a $p < 0.05$ significance test. Whilst both annotations clearly affect sample similarity, therefore, it is strongly suggested that source laboratory is the factor which has a stronger effect on sample similarity.

This is not as clear cut as it may first appear, however, as the division of the dataset into the annotations of “cell line” and “source laboratory” does not divide the matrix an equal number of times and so this must be taken into account when interpreting these results. For example, there are 72 source laboratories which were selected by RaSToVa for contributing more than 5 samples. On the other hand, only 27 cell lines were included in the analysis by RaSToVa. These 72 source laboratories represented 1009 total samples, and the 27 cell lines represented 1052 total samples. Despite, therefore, including a similar number of samples in their analyses, and including 91.6% (1 dp.) and 95.5% (1 dp.) of the HPM matrix respectively, an important concept here is the number of times the HPM matrix was divided during these analyses by the different annotations.

If the “cell line” annotation, when performing the RaSToVa analysis, is only dividing the matrix 27 times, compared to “source laboratory”'s ability to divide the matrix 72 times, it stands to reason that an annotation which divides the HPM matrix into smaller submatrices would appear better able to explain variation in that data. Taken to its logical extreme, an annotation which only has two different values, and thus only partitions the data into two sets will most likely have no chance of explaining the data as well as an annotation that is a unique identifier for every sample. Therefore, as the “cell line” annotation is only able to divide the matrix 0.375 the number of times that “source laboratory” is able to, “cell line” may require further investigation in other data where these “division numbers” are not so different. This, and other issues regarding data distribution and interpretation of RaSToVa's results here are discussed in section 3.4 and 5.1.4.

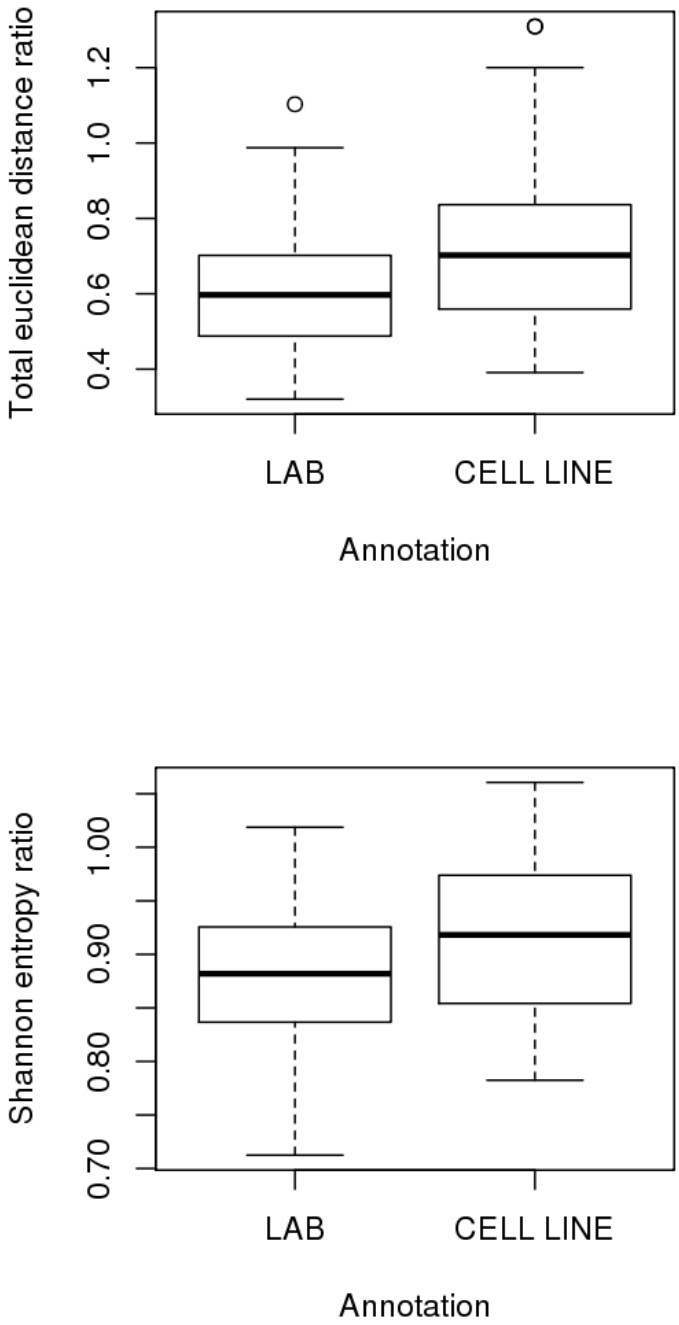

Figure 3.13 Summary of RaSToVa results using both total Euclidean distance (upper plot) and normalised Shannon entropy (lower plot.) Source laboratory, in both instances, therefore is found by RaSToVa to contribute marginally more to sample similarity in the HPM matrix than the cell line annotation, with a p-value of 0.049 for the differential expression analysis, and a tendency toward significance for the entropy change method with p = 0.061.

3.4 Discussion of RaSToVa results

RaSToVa was designed with the intention of ascribing inter-comparable “strengths” to the effects on sample similarity of annotations in a given data matrix. Whilst it has been shown to be able to do this, the results which it generates must be taken along with an understanding of what these results mean and, critically, what these results do not mean, particularly in the case of the HPM matrix used in this work.

Through the use of the RaSToVa method on the HPM matrix, the “source laboratory” annotation was shown to have a stronger effect on sample similarity than the “cell line” annotation did. This was somewhat expected, for two reasons. Firstly, the source laboratory annotation is likely very close to being a surrogate, in any data, for individual experiments, in which batch effects are likely to exist, increasing sample similarity. Experiments, and therefore, to a large extent, source laboratories, are likely to have certain interests and perform similar experiments, similar manipulations and therefore have highly similar samples. Secondly, the “source laboratory” annotation, in the data used here, breaks the matrix into a greater number of subgroups than the “cell line” annotation does, giving it the edge in explaining the patterning of the data. Despite this, and despite the “cell line” annotation dividing the matrix into roughly only a third of the number of groups that “source laboratory” does, similar contributions to sample similarity were found, as pointed out at the end of section 3.3.5.

RaSToVa cannot change the layout of the data that it is given, and the distribution of annotations in the data that it is given largely determines how the results of RaSToVa should be interpreted. In this case, with the disparity in the numbers of groups attributed to source laboratory and cell line, it may be tempting to declare the cell line annotation effectively more responsible for sample similarity than source laboratory, as it has only a third the number of divisions in the matrix to work with. As intuitive as this may sound, it is the conclusion of this half of the chapter that “source laboratory” is to be taken as the annotation which is responsible for more sample similarity. There are two major reasons for this.

First is the clear difference between the metrics calculated for the contribution to sample similarity of both of the investigated annotations, which, when Euclidean distance was used as the variability metric, satisfied a stringent $p <= 0.05$ threshold. With normalised Shannon entropy used as the variability metric, the result tended towards satisfying this threshold, with a p-value of 0.06065.

Secondly, the only potential cause for uncertainty as to which annotation contributes more to sample similarity is the issue of the disparity between the number of subgroups into which the two annotations investigated divide the HPM matrix. Whether or not the “cell line” annotation is at a perceived disadvantage when assessed for contribution to sample similarity, this does not affect the basic question as to “which annotation explains the patterning in the data best”. The answer to this question is clearly “source laboratory”. An annotation which divides data into more subgroups does not necessarily affect the transcriptional profile of the samples, however. Simply put, biological experiment annotations will likely include a range of fields, and nothing can be done about the way in which they divide the data.

Finally, it was discovered after the design of RaSToVa was nearing finalisation, that the HPM matrix's annotations for cell line and laboratory were highly confounded (see 3.9.) This may explain why similar levels of evidence were found for effects on sample similarity for both annotations. It is to RaSToVa's credit, therefore, that despite a near 3-fold difference in the number of times each annotation divided the matrix, the contribution to patterning of the data was still robustly detected in both cases. Final discussion on the implications of the confounded annotations of the HPM matrix are to be found in the final chapter conclusion in 3.10.5. The datapoints that make up every boxplot in figures 3.4 to 3.9 are provided on the accompanying DVD under “Chapter 3/RaSToVa/Permutation Results”.

3.5 Development of DALGES

A question which naturally follows on from ascribing an effect on sample similarity to any given experimental annotation is whether or not it is possible to go a step further and investigate the effects of an experimental annotation at the level of individual genes.

To investigate this possibility, a method was developed using some of the core methods which were used in RaSToVa. By using random permutation and resampling, the construction of a profile for an individual annotation can be built over the course of many permutations until a distinct profile emerges. When a sufficient number of permutations has been run for any one annotation, a transcriptional profile should emerge which, given the data from which it was generated, gives an annotation-linked gene expression signature; hence the name DALGES – Discovery of Annotation-

Linked Gene Expression Signatures. As with RaSToVa, the methodology that DALGES employs was chosen in such a way as to allow for the method to be extended to larger datasets and to other technologies such as the recent explosion in interest in RNAseq. The method was also conceived with a view to keeping it able to be carried out using only the data itself and the freely-available R statistical programming language and the Bioconductor suite of R-related tools.

3.5.1 Methodology overview

DALGES' methodology was originally intended to find annotation-linked gene expression signatures by way of differential expression. However, as this branched into two separate methodologies, first using the differential expression approach as detailed in section 3.5.2, and then using normalised Shannon entropy (see section 3.5.3) for the purposes of building up annotation-linked gene expression signatures for cell lines in the HPM matrix. The methodology remains essentially the same between them, and so is described only once in section 3.5.2. The use of Shannon entropy in building up a cell line's gene expression profile in this data requires only the changing of differential expression for Shannon entropy, but, critically, particular modifications are made to the normal pipeline for calculation of Shannon entropy in the interests of speed (see section 3.5.3).

3.5.2 Differential expression approach

The method starts, again, with a fully-annotated dataset. In the case of this section of chapter 3, the only required annotations from the manual annotation of the HPM matrix are the annotations for cell line. Firstly, DALGES copies a single-annotation submatrix from the whole matrix (e.g. all samples of the CGR8 cell line) and treats this as the “intact” submatrix (*id est*, where the annotation is left intact). The next step is to copy another, randomly-selected (with replacement) submatrix from the dataset of equal size to the aforementioned annotation intact submatrix. Rather than, as was the case with RaSToVa, simply calculating one metric for both the intact and random submatrices, differential expression analysis is performed between the intact submatrix and the randomly-permuted submatrix. This is done by calculating the means for all probes in both of these submatrices and then subtracting the randomly-permuted submatrix probe means from the intact submatrix's probe means. The resulting values, therefore, show the relative expression of each probe in the intact matrix as compared to randomly-permuted matrices. These numbers can have their

signs inverted (negative to positive and vice versa) to ask the question “what is the rest of the data like, compared to the samples sharing this annotation?”, although this chapter uses the former method, as it is concerned with finding transcriptional signatures of cell lines.

In addition to the calculation of the change in expression expected when a certain annotation is considered, DALGES seeks to ascribe a level of statistical significance to these measures. An advantage of using random permutation and resampling methods such as this is that p-values can be calculated which do not use any external model (such as the expectation of a normal distribution) and the distribution of the data itself is directly used when generating these p-values. Quantifying the change in gene expression between the intact annotation matrix and randomly-permuted submatrices is important as not all changes in gene expression need be large in order to be part of the signature of a cell line. If, for example, a cell line in the HPM matrix is only marginally lower in its expression of Gene X, if Gene X is consistently lower in this given cell line when compared to other cell lines, then this ought to be included. P-values allow for this. Therefore, in addition to calculating the means of each probe in both matrices, every time that a differential expression value is calculated (intact minus random), a running tally is kept, for each probe separately, as to whether or not the randomly-permuted mean expression level is higher, lower or identical to the mean expression of that probe in the intact matrix. This therefore allows DALGES to build up a profile, on a per-probe basis, of whether or not the intact mean expression level for that probe is consistently higher or lower than when considering the rest of the HPM matrix.

As it is the intention of this chapter to design analysis methods to be taken forward in future work, and on other genomics technologies, it was important to avoid unnecessary use of large amounts of compute resources. DALGES, therefore, does not store each individual differential expression value as calculated between intact and randomly-permuted submatrices, but rather simply keeps arithmetically adding the next differential expression value to a single list of running differential expression for all probes. These running totals of differential expression are then divided by the number of permutations performed. This, therefore, provides exactly the same result as would have been found by storing each differential expression result individually and taking a mean at the end of all permutations, but allows for a much smaller working memory footprint. This is particularly useful when considering that when DALGES is, as is the intention, used on much larger datasets, the number of permutations required for clear results may increase, multiplicatively increasing the amount of memory which would be required in these future cases.

A graphical summary of the DALGES method overview is provided in figure 3.14.

3.5.3 Normalised Shannon entropy approach

Whilst DALGES uses differential expression as the means to identify whether or not a given gene is associated with a particular annotation (e.g. a particular cell line or source laboratory), there is always the possibility that simple differential expression may miss the linkage between a gene's expression level and a particular annotation if the level of expression of that gene is to be found somewhat in the middle of other expression levels found for that gene. For example, if Gene X is consistently higher in the intact annotation matrix than in all other annotations, then Gene X will be given a p-value of zero and its mean differential expression when compared to randomly-permuted submatrices will be shown. This is clearly also the case for genes that are consistently lower in their expression when compared to all other annotations; DALGES will have no problem assigning it a mean differential expression value and zero p-value. As the distinction becomes less clear, and Gene X is not always above or not always below the level of expression found in other, randomly-permuted submatrices, the p-value will increase and the mean differential expression value will reduce in magnitude. However, in the hypothetical case of Gene Y, where Gene Y is usually expressed at a middle level when the cell line is “Cell Line A”, while all other cell lines have Gene Y expressed to a greater or lesser degree, there is still a distinct signature of Cell Line A, potentially, in that Gene Y is at this middle level. However, as DALGES performs comparisons between intact (Cell Line A – only) and randomly-permuted submatrices, there will tend to be a mix of greater and lesser means for Gene Y in the randomly-permuted submatrices. This will result in a very high p-value, causing the results outputted by DALGES to essentially advise the disregarding of Gene Y as potentially part of a signature of Cell Line A. This is particularly true if, for example, the randomly-permuted submatrices contain a somewhat equal mix of higher and lower values for Gene Y, causing their mean to average out somewhere quite close to, or possibly within, the spread of values for Gene Y found within Cell Line A itself! This would again result in DALGES essentially “writing off” Gene Y as uninteresting, or at least as not being linked to Cell Line A.

To ask whether or not these scenarios can be avoided, a second method of linking a gene's expression to a given annotation was added. By observing changes in the predictability of expression values, DALGES seeks to detect a more predictable level of expression for any given gene, once the annotation is known. That is, that if the aforementioned Cell Line A's Gene Y was

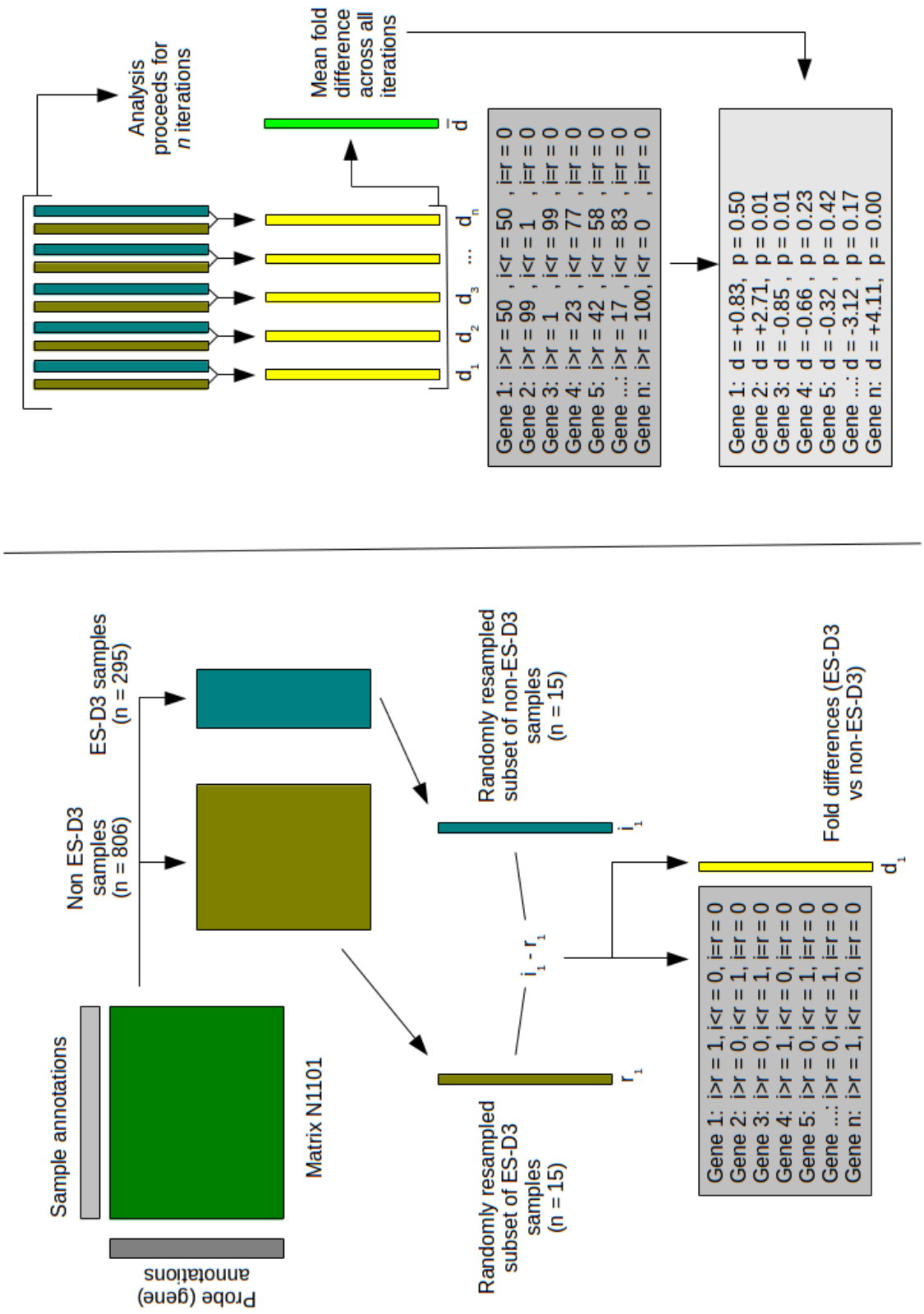

Figure 3.14: Methodology overview for DALGES. A submatrix comprised of all samples of one cell line is copied from the whole matrix. A randomly-permuted submatrix of the same size as the first submatrix is also copied. Differential expression analysis is carried out between the intact annotation submatrix and the random submatrix using means of each probe. When all permutations are complete, each differential expression result's positive and negative changes are tallied and used to generate p-values for each probe. Summed differential expression is then divided by the number of permutations for a final result for that probe.

consistently at any level, then the observance of that consistency versus the wider spread of expression values observed for Gene Y in randomly-permuted submatrices, DALGES would report this gene as being potentially linked to Cell Line A as its entropy is lower in the intact matrix of Cell Line A samples when compared to randomly-permuted submatrices. Normalised Shannon entropy was used to measure the amount by which either keeping the annotation intact or randomising it made any one probe more or less predictable. Again, the same method of calculating p-values can be applied where greater and lesser values for normalised Shannon entropy (between intact and randomly-permuted submatrices) are tallied as the permutations are completed. This was also intended to mitigate the issue that DALGES (using differential expression only) may encounter when assessing whether or not a gene is linked to an annotation if that gene is mostly higher or mostly lower than the expression found in any other annotation.

A combination of these two methods should, therefore, allow for both the assessment of whether or not a particular gene is, given the distribution of the matrix provided to DALGES, a particular gene appears to be related to any given annotation and, if so, pull out genes which are consistently higher or lower in expression than those of other annotations. In the case of those genes whose entropy greatly decreases when only samples of a certain cell line are used, but whose differential expression analysis is somewhat inconclusive from the differential expression method of running DALGES (due to being of intermediate expression compared to random submatrices), this may also imply that changes in variability (entropy) of genes may be of interest (subject to downstream gene ontology analyses.)

As the calculation of normalised Shannon entropy is quite complex, involving the discretisation of the HPM matrix, at a per-probe level, DALGES' analyses can be sped up with some modifications away from simply performing these calculations when they are needed. For example, for every randomly-permuted submatrix that DALGES uses, each probe is discretised into the chosen 110 bins and then entropy calculated and then normalised to the maximum entropy possible for this number of bins. This step will be performed 45,101 times for each new submatrix. With 1,001 permutations, this results in the sorting of a given number of values for each probe (dependent on the number of samples gathered into the random submatrix) into 110 bins, for 45,101 probes, 1001 times. An obvious solution when dealing with these issues of repetition in calculation would be to use pre-calculated entropies for all probes. This is of no help here as DALGES uses submatrices in its method, not the whole matrix at a time. Further, the samples which make up each submatrix change each time, forcing the recalculation of entropy each time. In addition, simply precalculating

the occurrences in each bin for all probes cannot help as there would be no way to include only those occurrences in the discretised bins which belong to the samples which are used in each submatrix. A solution, however, is possible, by precalculating bin numbers for each value in the HPM matrix. Where normally an expression value is found, a bin number is instead. By running through each probe of the HPM matrix and, for each sample, calculating which bin this expression value falls in (given the equally-spaced 110 bins that DALGES is using for this matrix), a precalculated matrix of bin numbers can be made. This allows for later steps of DALGES to completely avoid discretising up to 4,966,071,110 numbers into bins, and simply fill the bins using the bin numbers directly from this precalculated matrix, leaving only the entropy calculation still to do. As with other speed or memory improvements worked into the development of RaSToVa and DALGES, these were not necessarily required for the successful running of these methods in the scope of this work, but allow for a much easier and more timely application of these methods to larger matrices outwith this thesis, or to facilitate another aim of the project; making the method work on lower-end systems available to more researchers.

DALGES (Shannon entropy mode) was run on the HPM matrix with a total of 1001 permutations for both cell line (the main focus of the work) and source laboratory. These permutations were divided over 7 running instances of DALGES, each performing 143 permutations of each cell line or source laboratory. The results of these, being the summed entropies and the summed tallies for generating p-values were then pulled together into one instance of R and the p-values recalculated using all of these tallies. This confounding in the data only emerged late in the development of the methodologies of both RaSToVa and DALGES as the manual annotations in section 2.2.2 were still being completed. More discussion on the implications of the confounded nature of these two annotations is forthcoming in section 3.10.5.

3.6 Cell line specific gene expression signatures using DALGES – differential expression mode

An R object is provided on the accompanying DVD under “Chapter 3/DALGES/Differential Expression Method/Results R Object”, containing summed differential expressions as DALGES was run, along with p-values for all genes.

Further, a summary of the differential expression values for all 4 analysed cell lines can be found on the accompanying DVD in “Chapter 3/DALGES/Differential Expression Method/Differential Expressions Summary”

3.6.1 ES-D3 cell line

With the ES-D3 cell line, of which there are 295 samples in the HPM matrix, there were 468 probes which were found to have a greater than 1.5 log2 fold increase (2.83 (2 dp.) absolute fold increase) in expression compared to randomly-permuted submatrices of the same size. Enrichment for biological pathways was assessed using DAVID (Dennis et al. 2003), and the results of this are given in full on the accompanying DVD as “Chapter 3/DALGES/Differential Expression Method/ES-D3/Chapter.3.DALGES.ES.D3.POSITIVE.1.5.LOG2.FC.Genes.csv”. A similar naming convention is given for the results of all other DALGES analyses. The top of these was the generic “pattern specification process” GO term, (n = 33 probes, q-value 1.9×10^{-13}. The same 33 genes also contribute to the GO pathway “embryonic morphogenesis” with a q-value here of 1.1×10^{-10}. The list, as a whole, contains a great number of developmental processes. However, enrichment was also found for “regulation of Wnt receptor signalling pathway” (n = 9 probes, q-value 4.6×10^{-5}) and “Wnt receptor signalling pathway” in general, this time with (n = 13 probes, q-value 3.2×10^{-4}).

Genes which satisfied a threshold of a greater than 1.5 log2 fold lower expression in the ES-D3 cell line, compared to randomly-permuted submatrices, numbered only 180. There was also no enrichment for any non-generic pathways here, according to DAVID. The only pathways achieving a q-value of <= 0.05 were the three “stem cell” related pathways of “stem cell development”, “stem cell maintenance” and “stem cell differentiation” with the same 5 genes counted towards all of these, but respective q-values of 7.2×10^{-3}, 1.2×10^{-2} and 1.6×10^{-2}. The full list of pathway enrichments is available on the accompanying DVD.

Relaxing the threshold from the stringent +/- 1.5 log2 fold change to +/- 0.585 log2 (1.5 absolute) fold change generated lists of enriched pathways which showed a strikingly opposing signature to the CGR8 and E14 cell lines as analysed below. Using those genes found to be more than 0.585 log2 fold higher in ES-D3 cells, compared to random, a long list of developmental pathways is found to be enriched, available in full as on the accompanying DVD. 3,698 probes make up this list, implying that just under 8.2% of all probes on the microarray were found by DALGES to be more

highly expressed in ES-D3 samples compared to the distribution of the rest of the data. Despite this large number of probes, the pathway enrichments are very similar to those generated by using gene lists from the CGR8 and E14 lists of genes which tend to be more lowly expressed in those cell lines (see following sections.) The most significantly-enriched pathway whose genes are found to be more highly expressed in ES-D3 cells is the generic "embryonic morphogenesis" pathway (n = 109 probes, q-value $1.6x10^{-14}$). As generic as such terms may be, there are many more specific pathway GO terms are in this list such as "Wnt receptor signalling pathway" (n = 43 probes, q-value $1.9x10^{-6}$), "Notch signalling pathway" (n = 20 probes, q-value $1.4x10^{-3}$), "MAPKK cascade" (n = 29 probes, q-value $1.8x10^{-2}$), "death" (n = 92 probes, q-value $4.3x10^{-2}$). This may be indicative of the ES-D3 cell line having enhanced signalling in these pathways, compared to the other cell lines in the HPM matrix, but these may also be due to confounding of the data, as discussed in 5.1.4. It is interesting that there is an apparent increase in expression of 92 "death" related genes in the ES-D3 pathway. This would suggest that the ES-D3 cell line has modified apoptotic signalling / processes, particularly when it is considered that the other well-represented cell lines in the HPM matrix appear to have enrichments for pathways such as "death", "programmed cell death" in their lists of genes found to be more lowly expressed, compared to "randoms".

As these analyses of cell lines are comparative (that is, best thought of as showing signatures of cell lines relative to each other in a given dataset), it may also be that the patterning of the data is responsible for these enrichments (see 3.10.5). Substantiation through running of DALGES, in future work, on larger datasets, may answer this. In the meantime, however, there is clear suggestion from this data that the ES-D3 cell line appears to have altered apoptotic processes, compared to the CGR8 and E14 cell lines, with the usual caveats as given in 3.10.5 and 5.1.4.

There were 1,934 probes found to be more lowly expressed (threshold: -0.585 log2 fold change), compared to randomly-permuted submatrices, in the ES-D3 cell line. Enrichment for biological pathways in this list is provided on the accompanying DVD. This list, again, has highly-generic terms listed as the most significantly enriched, such as "regulation of transcription" (n = 232 probes, q-value $1.9x10^{-8}$). The more specific GO terms found to be significantly enriched in this list include "regulation of cell proliferation" (n = 76 probes, q-value $7.0x10^{-7}$, "cell cycle" (n = 74 probes, q-value $4.0x10^{-4}$), "negative regulation of cell differentiation" (n = 30 probes, q-value $2.0x10^{-3}$) and "cellular response to stress" (n = 50 probes, q-value $5.7x10^{-3}$). These pathways are of interest as they suggest at potential peculiarities of the ES-D3 cell line. The relatively-lower expression of cell cycle and proliferative genes may indicate that the ES-D3 cell line is inherently less vigorous than

other cell lines represented in the HPM matrix. Relatively-lower expression of genes linked to “negative regulation of differentiation” may be indicative of the ES-D3 line having relatively increased propensity for differentiation. It is also highly interesting that the “cellular response to stress” pathway genes found here are comparatively more lowly expressed in the ES-D3 line. Cellular stressors are being shown to promote cellular reprogramming in cancer, for example, in the case of nutrient stress (Ma et al. 2013) or inflammation (Song and Balmain 2015), it is interesting to find with this method that 50 genes involved in this pathway are comparatively lower in this cell line. With the usual caveats concerning distribution of data and random permutation methods (see sections 3.10.5 and 5.1.4), this suggests that there may be a relationship between an increased propensity to differentiate, and a dampened stress response. It would logically follow that “response to stress” genes may well be linked to a less-differentiated, naïve state in mESCs. This would be in agreement with literature which suggests that the induction of a stress response, such as hypoxia renders cells more amenable to reprogramming (Yoshida et al. 2009).

Finally, other pathways such as several chromatin reorganisation pathways were found in this list of relatively-lower expressed genes in ES-D3 samples, but these were of a generic nature and as such did not present opportunity for speculation as to their biological relevance.

3.6.2 CGR8 cell line

This cell line is represented 110 times in the HPM matrix. Interestingly, the results of DALGES (differential expression mode) on the CGR8 cell line did not result in many genes (n = 3) being more than 1.5 log2 fold more highly expressed in this line, when compared to randomly-permuted submatrices. Only Mid1 and two probes for the Ddx3y genes were found to be more highly expressed in this cell line by this threshold. At the other end of the scale, however, considering those genes found to be more than 1.5 log2 fold lower expressed in the CGR8 line, 47 probes achieve this level of fold change, but are mostly unnamed (40 of 47) and those which are annotated being Rian (3 probes), D7Ertd715e, Plalg1, Meg3 and Rbp1. As 1.5 log2 fold change corresponds to an absolute change of 2.83 (2 dp.), relaxing this threshold to 1.5 absolute fold change in either direction allows for some enrichment analysis using DAVID.

For those genes found to be 0.585 log2 fold higher expressed in CGR8 cells, of which there are 266 probes, an enrichment was found for quite high-level pathway names. These very broad terms

include “transcription” (n = 43 probes, q-value $2.4x10^{-3}$) and “cell cycle” (n = 21 probes, q-value 1.2x10-2). The full table is available on the accompanying DVD, but is very generic and rapidly loses statistical significance for enrichment, failing the q-value < 0.05 test after the third entry, being “regulation of transcription” (n = 46 probes, q-value $2.1x10^{-2}$). It was not expected that DALGES would generate lists with necessarily any striking enrichment for biological pathways, as the goal of DALGES was simply to ascertain whether or not it is possible to ascribe gene expression levels to annotations with a useful degree of certainty. This is an objective which it still definitely achieves, with certain important caveats (see section 3.10.5.)

At the other end of the mean fold change scale, being those genes found to be more than 1.5 absolute (0.585 log2) fold more lowly expressed in the CGR8 line, 894 probes meet this criterion. Enrichment for biological pathways here reveals some more specific pathways. Here, a striking enrichment is seen for the “epithelium development” pathway (n = 31 probes, q-value $4.5x10^{-5}$). The biological significance of this is unclear, however this result opens up the possibility that certain cell lines (in this case CGR8) may be found have less aptitude for developing into certain tissue types. These sorts of signatures in cell lines, if substantiated using larger datasets, can be used to drive experimental investigations concerning a cell line's efficiency in generating certain lineages. Other pathways found here are quite generic and are of no real surprise in our data, as our data concerns progression toward differentiation of embryonic stem cells. These pathways are given hierarchical names such as “tissue morphogenesis” (n = 28 probes, q-value $6.9x10^{-5}$). Several other development-related pathways are to be found here however, mixed in amongst these, however, is “negative regulation of cell proliferation” with (n = 25 probes, q-value $7.2x10^{-4}$). These genes being downregulated would imply that there is a signature here consistent with a pro-proliferative state. This may warrant substantiation of this pro-proliferation signature in larger datasets to test whether or not the CGR8 cell line tends to be more highly proliferative. This may confer the CGR8 an advantage in stem cell-related manipulations. Any putative proliferative tendency, if confirmed in larger data, would need to be taken into account by researchers working with the CGR8 line. In addition to this one pro-proliferative pathway, enrichments are also found for the downregulation of genes involved in “programmed cell death” (n = 37 probes, q-value $1.4x10^{-3}$) and “regulation of programmed cell death” (n = 40 probes, q-value $3.1x10^{-3}$). Interestingly, however, a pathway also found to be enriched here is “positive regulation of cell death” (n = 23 probes, q-value $5.9x10^{-3}$). “Wnt receptor signalling pathway” (n = 14 probes, q-value $3.1x10^{-2}$) is also to be found here, although the hierarchical nature of the database which DAVID uses to name pathways renders this somewhat inconclusive. In-depth analyses of all of the potential implications, at a phenotypic level,

of these differentially-expressed genes is outside the scope of this work, as this chapter only seeks to find whether or not these signatures can be found, and comment on the behaviour of the method used to find them.

3.6.3 E14 cell line

The E14 cell line is represented 112 times in the HPM matrix. Similar to the CGR8 cell line, very few genes achieve the 1.5 log2 fold change increase compared to randomly-permuted submatrices. In fact, only 8 genes make this threshold, including Mid1, a gene also found to be 1.5 log2 fold higher in CGR8 cells, compared to random. The use of the 1.5 absolute fold (0.585 log2 fold) threshold for being more highly expressed in the E14 cell line results in a list of 234 probes with absolutely no significant enrichment for any GO pathways. The most significant result in this list is "regulation of transcription", a term which is highly generic. However, this is still an interesting result when considering the results of investigating the higher-expressed genes in the CGR8 cell line, which were similar to this. A first suspicion here may be that DALGES is finding identical results as both E14 and CGR8 contribute a similar number of samples to the data, which, if the distribution of the data is obscuring DALGES from seeing gene-by-gene signatures of different cell lines, would be the case. However, DALGES is very unlikely to be being swayed, in this instance at least, by the distribution of the data in confidently ascribing these signatures to these two cell lines. This can be demonstrated by observing the lists of genes which DALGES finds to be likely higher in both the CGR8 and E14 cell lines. When sorting the data to bring those genes most likely more highly expressed in E14 cells, for example, the 13th rank gene is Fbxo15, which is found to be, when compared to random submatrices, 1.27 log2 fold (2.41 fold absolute) higher in E14 samples, with a p-value of zero (quite likely rounded when processed by R.) This same gene, in the case of CGR8 cells is, when compared to random submatrices, found to be -0.04 log2 (0.972 absolute) fold less expressed, and has a p-value of 0.38 (2 dp.). Other examples are present, even in this top list, such as Aass, which is found to be 1.17 (2 dp.) log2 (2.25 absolute) fold higher in E14 samples, with a p-value of another rounded-to-zero. In CGR8 samples, this is found to be, when compared to random submatrices, -0.02 log2 (0.986 absolute) fold lower, with a p-value of 0.428, indicating no real relationship between the Aass gene and the CGR8 annotation. If DALGES was simply displaying results that are, in fact, a surrogate of the distribution of the data, this could not happen. However, the distribution of the data is still highly important, as discussed in 3.10.5 and 5.1.4.

Similarity is also found between the E14 signature and the CGR8 signature when considering the list of genes found to be more lowly expressed in E14 cells. Passing the -0.585 log2 fold change (-1.5 absolute fold change) mark were 819 probes. Whilst the full list is provided on the accompanying DVD, it is largely comprised of developmental pathways, with the generic pathway "pattern specification process" (n = 41 probes, q-value $3.1x10^{-12}$) as the most significantly-enriched pathway. Genes are downregulated also for the "Wnt receptor pathway" with (n = 17 probes, q-value $3.9x10^{-4}$). The "cell death" pathway makes an appearance in this list, similar to the situation in CGR8 cells, this time with (n = 32 probes, q-value $1.6x10^{-2}$). Seeing this similarity between CGR8 and E14 cell lines is interesting as it may imply that both of these cell lines have an increased propensity for proliferation in that the transcriptional networks involved in apoptosis appear to be, from these results, somewhat muted when compared to the HPM matrix's distribution as a whole. This warrants further investigation in larger datasets for substantiation and suggests sets of genes that can be shortlisted for analysis in experimental settings to further confirm the differences between these cell lines. The implications of finding the downregulation of these apoptotic genes in CGR8 and E14 samples are discussed more in 3.10.4.

3.6.4 iPS (OSKM) cell line samples

These samples, induced pluripotent cells transfected with the 4 Yamanaka factors Oct4 (Oct4), Sox2, Klf4 and c-myc, are mentioned here from the DALGES results as they satisfy (when grouped) the inclusion criteria of having more than 50 samples contributed to the HPM matrix (with 59 samples). In addition, the inclusion of iPS samples in the DALGES analysis is of interest as the signatures of iPS lines, as compared to other mESC lines, is an area of great interest in the area of mESC biology, as this may impact on their utility, but also, critically, on their safety (Miura et al. 2009). It must be noted here that the iPS samples grouped together for this analysis are from more than one experiment, although all were generated through forced expression of the canonical Yamanaka factors (Oct4, Sox2, Klf4, c-Myc). These samples are labeled in the annotations file for the HPM matrix under the column "Cell Line Name (Simple)" as "IPS_Oct4_Klf4_Sox2_c-myc". This is therefore not an attempt to investigate the characteristics of an individual cell line contributed by one laboratory and that laboratory's individual methodology, but a first look at a whether a more general OctSoxKlfMyc (OSKM) iPS signature emerges at all in this data. If a distinctive signature is found for these iPS using this method, even in this somewhat limited dataset (in that it is limited to only the highest levels of Oct4, Sox2 and Nanog, rather than how said iPS

cells would behave when differentiating), then this would merit the use of the DALGES method in larger datasets to find signatures of individual iPS lines for intercomparison. In the interests of ease of reading, these grouped iPS samples are referred to as a "cell line" hereafter.

Using those genes found to be, compared to randomly-permuted submatrices, 0.585 log2 fold higher expressed in these iPS samples, a strikingly-different picture emerges from the other cell lines analysed in this chapter. 1,457 probes pass this threshold. First and foremost, there is a strong enrichment for the "chromosome organisation" pathway (n = 48 probes, q-value $1.1x10^{-3}$). This is unlike the enrichments for any other cell line analysed here. The second most significantly-enriched pathway is along a similar theme, being "meiosis 1" (n = 11 probes, q-value $1.3x10^{-2}$), and the third again, named "chromosome organisation involved in meiosis" (n = 8 probes, q-value $1.3x10^{-2}$.) This is of interest as this enrichment suggests that there is an upregulation of genes involved in meiosis in these iPS samples, which may be indicative of altered cell division machinery in these iPS samples. The other three cell lines (CGR8, E14 and ES-D3) analysed in detail in this chapter are commonly-used mESC lines. It was therefore expected that these cell lines would share a great deal of similarity, possibly hampering DALGES' ability to find any difference in the signatures found between them. By extension, there was concern that DALGES, in the absence of a suitably contrasting cell line occurring in the HPM matrix, may have been rather unfairly seen to be unable to find useful transcriptional, and, potentially, functional differences between cell lines. With the analysis of this iPSC group of samples, however, a clear and distinct signature separates this cell line, both at a transcriptional and possibly at a functional level, with these enrichments. Another enrichment which occurs only in this line, out of the four analysed, is "extracellular matrix organisation" (n = 18 probes, q-value $1.2x10^{-2}$). There also appears to be an enrichment here for pathways involved in gametogenesis, such as "spermatogenesis" (n = 30 probes, q-value $2.7x10^{-2}$). These same 30 genes are also counted toward the "male gamete generation" pathway. DALGES also may have captured the fact that these iPS samples have had their differentiation machinery suppressed as a likely possible of their having forced expression of the four Yamanaka factors. This is demonstrated by the enrichment for "negative regulation of cell differentiation" (n = 24 probes, q-value $2.6x10^{-2}$.) Enrichment is also found for the upregulation of genes involved in the "cell cycle" pathway (n = 58 probes, q-value $1.2x10^{-2}$.) This may be indicative of a propensity for increased cellular turnover in iPSCs generated by OSKM factors, although this requires further investigation.

Those genes which DALGES finds to be at least 0.585 log2 fold less expressed (n = 944) in these iPS samples were expected, given the previous strikingly-different signature for those found to be

more expressed, very different again to the signatures found for the other cell lines. Interestingly, this was not the case, with a very similar list of generic "development" pathways emerging and the familiar "pattern specification process" (n = 42 probes, q-value $1.1x10^{-10}$) making its way to the top of the list with the most significant enrichment. Even the familiar "Wnt receptor signalling pathway" (n = 17 probes, q-value $1.0x10^{-3}$) is found here, similar to the signature found for the other analysed cell types. The full table is available on the accompanying DVD. The fact that the list is quite so similar to other cell lines for the "downregulated" genes, yet so different for the "upregulated" genes, is encouraging as this shows that DALGES is indeed capable of finding unique signatures for the annotations that are queried. Had the enrichment list for those genes found to be "downregulated" in the iPS samples been very different to those of other cell lines, there may have remained a possibility that there was some effect of either the number of samples (where the iPS line has only 59, compared to the greater numbers of samples found in the other cell lines analysed) or some other systematic problem with DALGES. Therefore, it is ideal that DALGES has given a similar list of enrichments for the "downregulated" genes, while demonstrating its ability to find a unique list of enrichments for those "upregulated" in this cell line. Further discussion of the results of DALGES can be found in 5.1.4

3.7 Cell line specific gene expression signatures using DALGES – normalised Shannon entropy mode

An R object is provided on the accompanying DVD under "Chapter 3/DALGES/Entropy Change Method/Results R Object", containing summed differential expressions as DALGES was run, along with p-values for all genes.

Further, a summary of the summed entropy change values for all 4 analysed cell lines can be found on the accompanying DVD in "Chapter 3/DALGES/Differential Expression Method/Entropy Changes Summary"

Distributions of entropy changes found by DALGES for all probes are sorted into ascending order for each cell line and displayed in figure 3.15 with accompanying red horizontal lines depicting thresholds chosen for entropy reduction and entropy increase analysis.

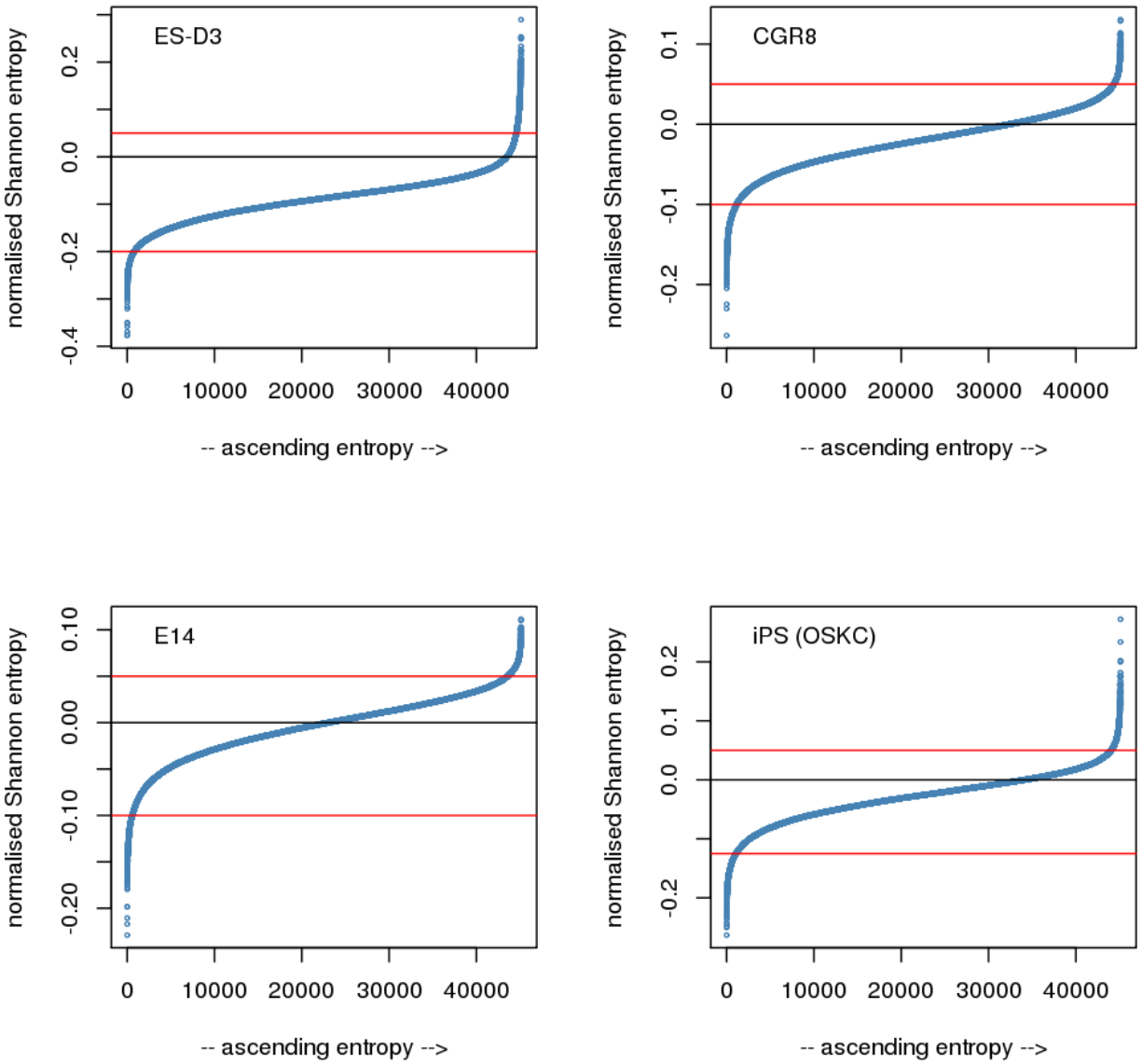

Figure 3.15: S-plots of all changes in probe entropy as calculated by DALGES for all 4 analysed cell lines (ESD3, CGR8, E14 and OSKM/OSKC_iPS) (OSKM and OSKC are equivalent, the last letter standing for c-Myc), red horizontal lines show cutoffs used for considering genes to show either strong positive or negative entropy change. Datapoints signify the normalised Shannon entropies of probes in each cell type.

3.7.1 ES-D3 cell line

There were 811 probes which pass the threshold of an entropy change of at least less than -0.2. Contrary to the findings when DALGES used the differential expression method, DALGES finds a different signature here, being strongly enriched for oxidative and metabolic pathways. Here, the generic term “generation of precursor metabolites and energy” pathway (n = 30 probes, q-value $3.1x10^{-5}$) is the most significantly enriched. Following on from this are more specific pathways such as “electron transport chain” (n = 19 probes, q-value $2.2x10^{-5}$), “cellular respiration” (n = 13 probes, q-value $3.2x10^{-4}$) and “oxidation reduction” (n = 46 probes, q-value $3.9x01^{-3}$). The full table is available on the accompanying DVD. This respiration / oxidation enrichment is strikingly different to the enrichment found when using the differential expression method, suggesting that entropy change captures a different set of phenomena than differential expression.

It was expected also that those genes which increase in their entropy within one annotation would be uninteresting. After all, genes which become less predictable within a given annotation should intuitively be considered to be unrelated to that annotation. For completeness, however, those probes with a larger than 0.05 positive entropy change in the ES-D3 samples, compared to random submatrices, were assessed for any pathway enrichments. The aforementioned assumption that there would be no interesting enrichments was dispelled by the extremely significant enrichments found in the list of 532 probes which passed the entropy threshold of >= 0.05. These enrichments are detailed in full on the accompanying DVD. These enrichments include a long list of developmental pathways, with an extremely low q-value given to the generic “pattern specification process” (n = 47 probes, q-value $6x10^{-25}$). Aside from the long list of developmental pathways which make up the vast majority of the significantly-enriched pathways, the “Wnt receptor signalling pathway” (n = 11 probes, q-value $7x10^{-3}$) is to be found, along with “cell adhesion” (n = 25 probes, q-value $1x10^{-2}$) and “regulation of BMP signalling pathway” (n = 5 probes, q-value $1.3x10^{-2}$). An increase in the entropy of the genes which contribute to these pathway enrichments suggests that those samples in the HPM matrix which use the ES-D3 cell line may be from experiments in which these signalling pathways are changing activation states and the changing in state of developmental pathways in the ES-D3 samples would be parallel to this. Indeed, the increased entropy enrichments are logical when it is considered that the majority (n = 249 or 295) of ES-D3 samples are from the “Piersma AH” laboratory, and the experiments contributed from this laboratory to the HPM matrix were from investigations of embryoid body cultures during differentiation (van Dartel et al. 2011). This, in fact, was one of the first and strongest suggestions in this work that the use of normalised Shannon

entropy in large-scale microarray data such as the HPM matrix may well be a method which can, in and of itself, indicate the activity of signalling pathways and / or transcriptional networks across the data.

3.7.2 CGR8 cell line

Probes which satisfied a threshold of a change in entropy of -0.1 numbered n = 1149. The enrichments found here are given in full in on the accompanying DVD. The most significantly-enriched pathway here was found to be “blood vessel development” (n = 38 probes, q-value 7.8×10^{-7}) Other developmental pathways occur in this set of genes found to be less entropic in the CGR8 cell line, including “skeletal system development” (n = 32 genes, 4.4×10^{-3}), “neural crest cell development” (n = 9 probes, q-value 1.2×10^{-2}), “heart development” (n = 23 probes, q-value 4.0×10^{-2}) and “cartilage development” (n = 12 probes, q-value 4.1×10^{-2}). The implications of this are not clear simply from a decrease in entropy, but warrant future analysis of the differences in developmental pathway genes between this and other mESC lines.

The CGR8 cell line's less entropic genes also contain enrichment for cell-cycle related pathways, being “cell division” (n = 33 probes, q-value 2.8×10^{-3}), “cell cycle process” (n = 38 probes, q-value 9.2×10^{-3}), “cell cycle phase” (n = 34 probes, q-value 9.6×10^{-3}) and others (see accompanying DVD.) These enrichments suggest that the CGR8 cell line may have altered propensity for proliferation when compared to other cell lines in the HPM matrix. This is in accordance with DALGES' findings for this cell line when using the differential expression method, which found cell cycle-related genes to be upregulated, compared to randomly-permuted submatrices. There is also a signal detected here for the MAPKK pathway (n = 13 probes, q-value 4.8×10^{-2}), a pathway known to be involved in mESC differentiation. There is also an enrichment for “cell response to stress” (n = 41 probes, q-value 4.7×10^{-3}). Just as the ES-D3 cell line was found to have a potentially dampened response to stress, it is equally possible that, as this analysis compares cell lines to each other, the CGR8 cell line may have an increased tendency toward the activation of stress-related pathways. This may simply be present in samples in this matrix due to the conditions to which the CGR8 samples herein were exposed.

Those probes which increased their normalised Shannon entropy by a mean of +0.05 (n= 752), when analysed for pathway enrichments, contained no enrichments at all which passed a threshold of significance of q <= 0.05. The full table is available on the accompanying DVD.

3.7.3 E14 cell line

There were 597 probes which appear to reduce in entropy in the E14 cell line, compared to the distribution of the HPM matrix as a whole, by at least -0.1. Pathway enrichments for this gene list included a great many developmental pathways. The "Wnt receptor signalling pathway" (n = 15 probes, q-value $8.2x10^{-4}$) was enriched along with "negative regulation of cell proliferation" (n = 17 probes, q-value $7.9x10^{-3}$), "negative regulation of cell differentiation" (n = 18 probes, q-value $6.4x10^{-4}$) and "cell fate commitment" (n = 15 probes, q-value $1.7x10^{-3}$). Together these pathways suggest that the E14 cell line (or the samples using it which are included in the HPM matrix at least) should be looked at more closely regarding their propensity for proliferation and tendency to differentiate. The fact that the Wnt pathway is the only obvious, named signal transduction pathway here to pass the q-value threshold of q <= 0.05 may link these two phenomena, in that the E14 cell line may have this altered proliferation / differentiation behaviour due to differences in Wnt signalling. This possibility may apply to the other cell lines here found to have signatures involving the Wnt signalling pathway, apoptosis, proliferation and cell cycle phenomena. After all, differences in endogenous Wnts between different cell lines are a known phenomenon in mESCs (ten Berge et al. 2011).

Probes which satisfied a threshold of an apparent change in entropy of at least +0.05 (n = 1,552), when analysed for pathway enrichments, similar to the case for CGR8, contain no enrichments at all which satisfy the q-value threshold of q <= 0.05. The full table is available on the accompanying DVD.

3.7.4 iPS (OSKM) line

1,078 probes satisfied the threshold of a decrease in normalised Shannon entropy of -0.125. In stark contrast to the enrichments found when using the differential expression analysis method (see 3.6.4), a drop in entropy is observed in genes which contribute to pathways much more in line with those pathways commonly linked to the other cell lines analysed here. Pathway enrichments for "stem cell maintenance" (n = 11 probes, q-value $8.5x10^{-6}$), "stem cell differentiation" (n = 12

genes , q-value $2.3x10^{-5}$), “stem cell development” (n = 12 probes, q-value $1.1x10^{-6}$) are included here. As these iPS cells have forced expression of Oct4 (Oct4), Sox2, Klf4 and c-myc, it is encouraging to see DALGES finding enrichments for these three pathways with greater levels of significance than in any of the other 3 cell lines analysed (see full enrichment tables for exact figures.) There is also an extremely strong enrichment for “regulation of cell proliferation” (n = 62 probes, q-value $3.7x10^{-7}$), which dwarfs the enrichments found for this pathway in the other cell lines analysed, in terms of significance. This may suggest a proliferative / stem cell identity “overdrive” induced by forced expression of the Yamanaka factors. Enrichment also exists for the BMP pathway (n = 7 probes, q-value $2.2x10^{-2}$), “Wnt receptor signalling pathway” (n = 17 probes, q-value $1.6x10^{-2}$), “cell fate commitment” (n = 19 probes, q-value $9.0x10^{-3}$) and “cell adhesion” (n = 49 probes, q-value $3.9x10^{-3}$). Much like the results for other cell lines, differences in Wnt signalling appear to be part of the signature of this iPS line's samples.

1,062 probes satisfied the threshold of an increase in normalised Shannon entropy of +0.05. Despite such a large number of probes increasing in entropy with this cell line, there were, as with the CGR8 and E14 cell lines, no biological pathways enriched at all to the required threshold of q <= 0.05.

A summary of the selected pathways mentioned in the search for pathway enrichment signatures for these four cell lines is presented in figure 3.16.

3.8 Behaviour of the DALGES method

3.8.1 Behaviour of DALGES - differential expression mode

In the previous analyses using RaSToVa, certain metrics were found to be affected by the number of samples that the annotation in question contributed to the HPM matrix (see section 3.3.3). It was expected that, much like with RaSToVa, this would not affect the overall message to be derived from the results, but confirmation was nonetheless required that DALGES' results would fare similarly well. A relationship was expected to be found between the p-values found for individual genes and the number of samples that the annotation in question contributed to the HPM matrix, in that statistical significance of the differential expression of any given gene should become clearer the more samples that DALGES has to work with for that annotation. Conversely, it would be highly detrimental to DALGES' ability to meaningfully link gene expression levels (or, later,

	ES D3	CGR8	E14	iPS (OSKC)
Expression increase	Wnt receptor signalling pathway MAPKK cascade Notch signalling pathway Death Many development pathways	Cell cycle	NONE	Cell cycle Chromosome organisation Meiosis I Extracellular matrix organisation Gametogenesis / spermatogenesis
Expression decrease	Negative regulation of cell proliferation Cell response to stress	Epithelium development Negative regulation of cell proliferation Programmed cell death Positive regulation of cell death Wnt receptor signalling pathway	Wnt receptor signalling pathway Cell death	Wnt receptor signalling pathway
Entropy decrease	Generation of metabolites Electron transport chain Cellular respiration Oxidation / reduction	Cell division Cell migration Regulation of cell proliferation Cell response to stress MAPKK cascade	Many developmental pathways Wnt receptor signalling pathway Negative regulation of cell differentiation Negative regulation of cell proliferation Cell fate commitment	Cell proliferation BMP signalling pathway Wnt receptor signalling pathway Cell fate commitment Cell adhesion Stem cell differentiation Stem cell development Stem cell maintenance
Entropy increase	Many development pathways Wnt receptor signalling pathway VEGF signalling pathway BMP signalling pathway	NONE	NONE	NONE

Figure 3.16 : Summary of significantly-enriched (q <= 0.05) biological pathways identified by both differential expression and entropy change modes of the DALGES methodology in the four analysed cell line annotations: ESD3, CGR8, E14 and iPS samples.

entropy signatures) of genes to annotations if both of those metrics (differential expression and entropy change) were also related to the number of samples a given cell line contributed to the HPM matrix. Observation of the relationship between these parameters confirmed the presence of the desired relationship between “number of samples contributed” and the resulting p-values (see figure 3.17, topmost plots).From this initial analysis, it is clear to see that those annotations with larger numbers of samples generate results with much lower p-values (see figure 3.17.) This is an encouraging sign as one would expect those annotations which have a larger number of samples available to be most amenable to having their gene expression signature determined.

A relationship is also observed between the p-values for differential expression values and the differential expression values themselves. This again is indicative of DALGES behaving as intended as it would be expected that those genes which consistently are found to be much more highly expressed in the intact matrix when compared to randomly-permuted matrices would have an accompanying zero p-value, as no permutation performed ever resulted in a randomly-permuted mean for this gene which was greater than that mean found in the intact annotation submatrix. The same is true for genes which are always more lowly expressed in the intact submatrix when compared to randomly-permuted ones. This can be seen by the shapes of the plots in figures 3.18 and 3.19 when using the differential expression mode and figure 3.20 when using the change in normalised Shannon entropy.

Support is also given to DALGES performing as intended by these same figures by the fact that around the 0 mark for differential expression and entropy change, there is a gap where the p-value is also 0. This shows that DALGES does not ascribe any significance (read: no way in which that gene is linked to the cell line in question) for those genes which are not differentially expressed (or differently entropic) in the intact matrix compared to randomly-permuted ones. It is interesting to note the difference in the relative sizes of these around the 0 mark on the x-axis, where in figures 3.18 and 3.19 this space is considerably smaller than the space in figure 3.20, indicating that larger changes in normalised Shannon entropy between intact and random submatrices are required before significance is attributed to any gene, potentially making entropy change more conservative a method than differential expression, although this requires further investigation.

These basic tests of DALGES' functionality aside, it was surprising to find that out of 1,217,727 total p-values calculated (27 cell lines x 45101 probes), nearly 40% (n = 479,391) of these were equal to or less than the stringent threshold of $p <= 0.01$. Relaxing this threshold to $p <= 0.05$

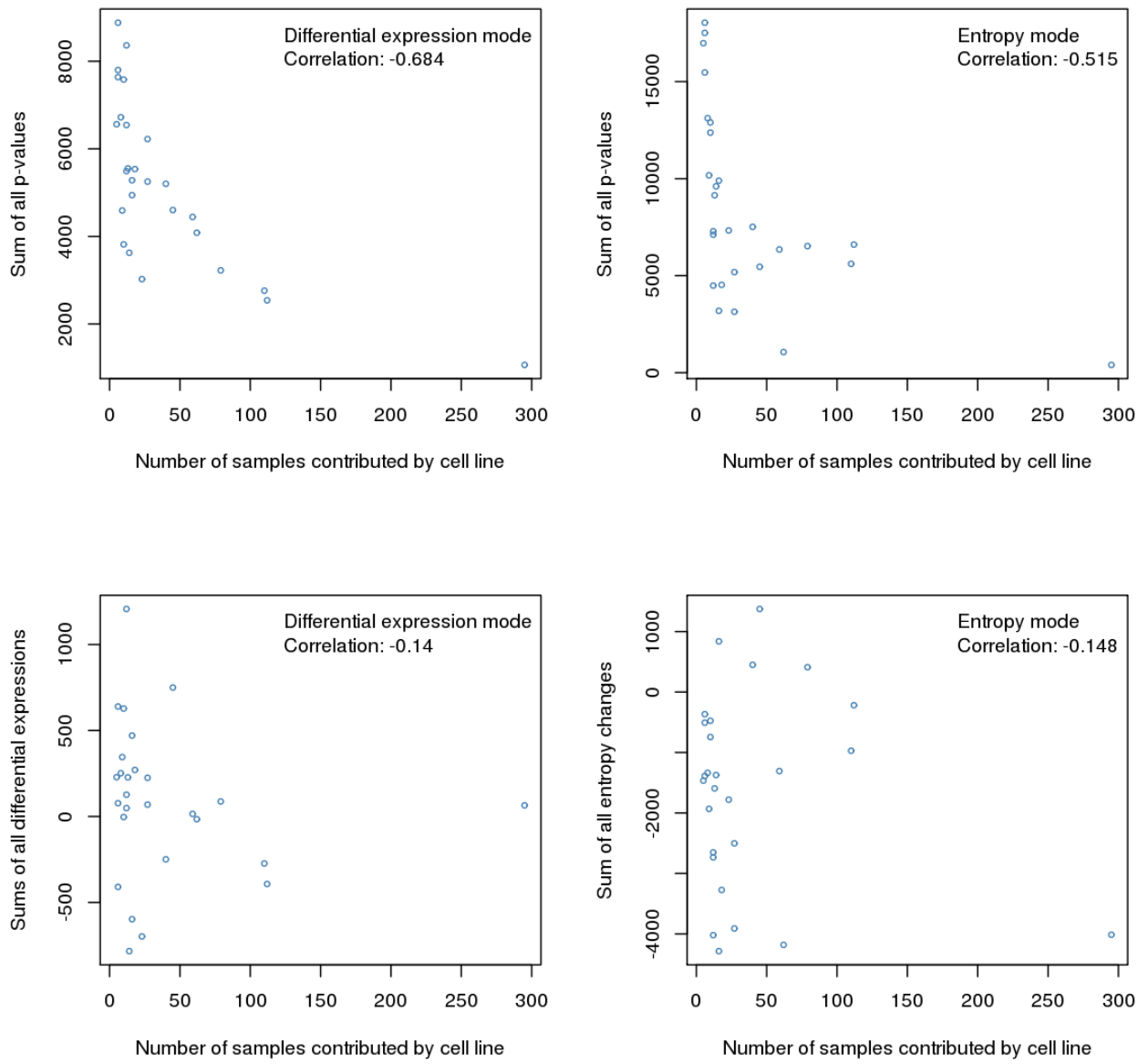

Figure 3.17 Behaviour of the DALGES method for different numbers of samples contributed by the different cell lines present in the HPM matrix when using both differential expression (leftmost plots) and normalised Shannon entropy (rightmost plots). Pearson correlations shown in the top right of each plot demonstrate the existence of the expected effect of a greater number of contributed samples for a given cell line having an overall effect of increasing the statistical power with which genes are ascribed differential expression or entropy change values. Reassuringly, however, no such relationship, exists between number of samples contributed for a given cell line and the magnitudes of the changes in either expression or entropy themselves.

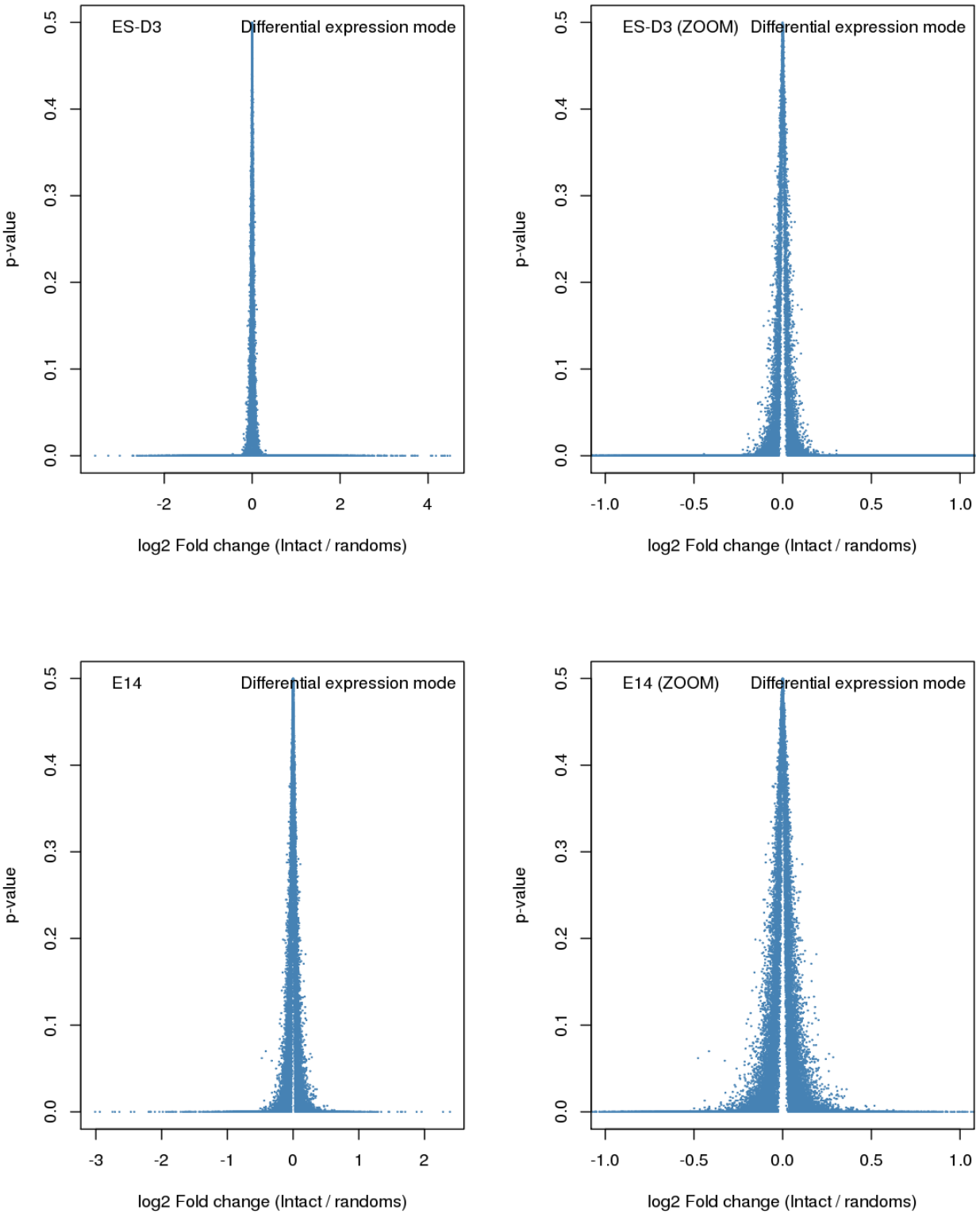

Figure 3.18: Behaviour of DALGES when analysing ES-D3 and E14 cell lines in the HPM matrix. Relationship between differential expression (x-axis) and p-value (y-axis) is shown. Leftmost plots include the full range of values while, for clarity of the centre of the plots, the rightmost plots show the data only between limits of -1 and 1 on the x-axis, for detail.

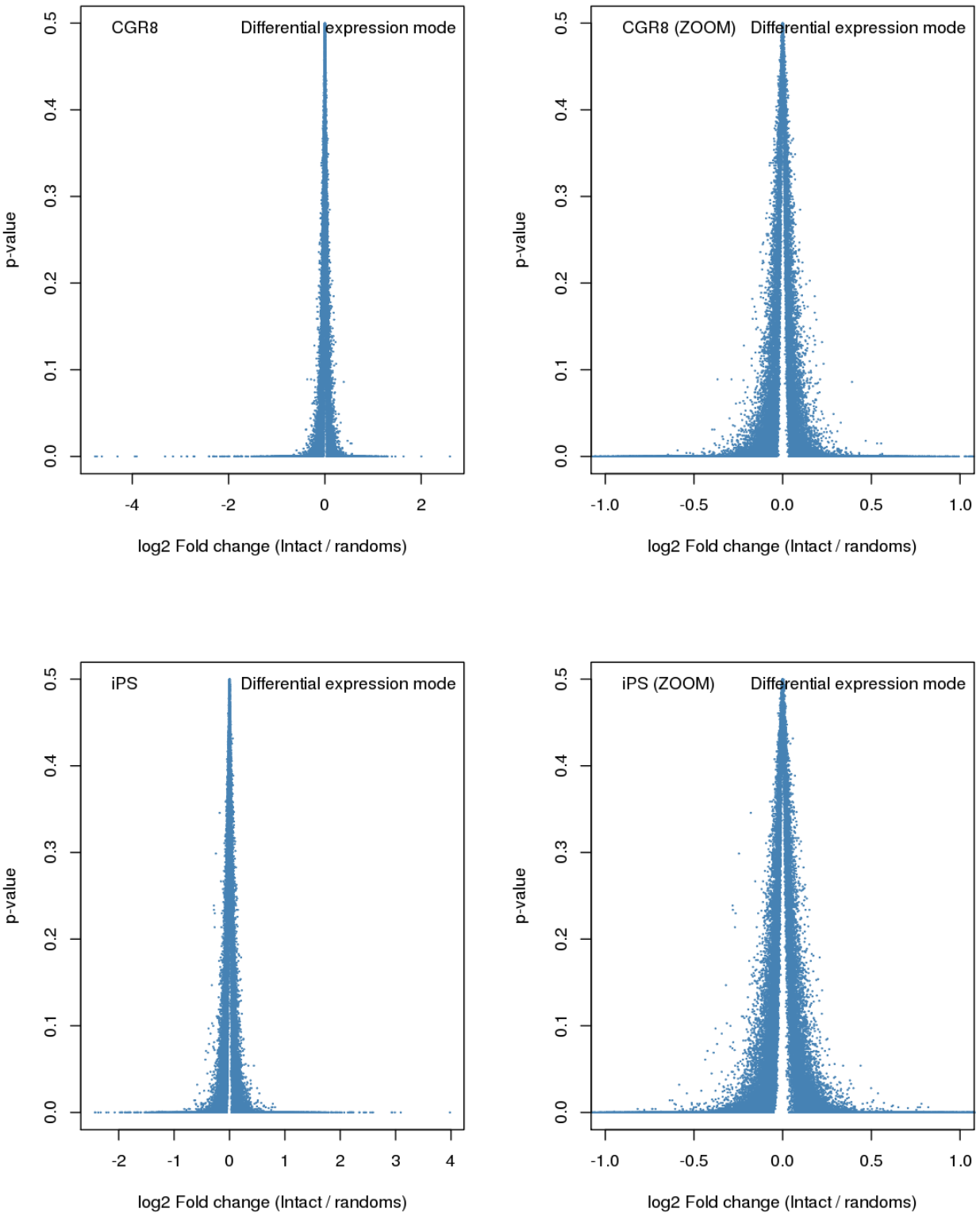

Figure 3.19: Behaviour of DALGES when analysing CGR8 and iPS (OSKM) cell lines in the HPM matrix. Relationship between differential expression (x-axis) and p-value (y-axis) is shown. Leftmost plots include the full range of values while, for clarity of the centre of the plots, the rightmost plots show the data only between limits of -1 and 1 on the x-axis, for detail.

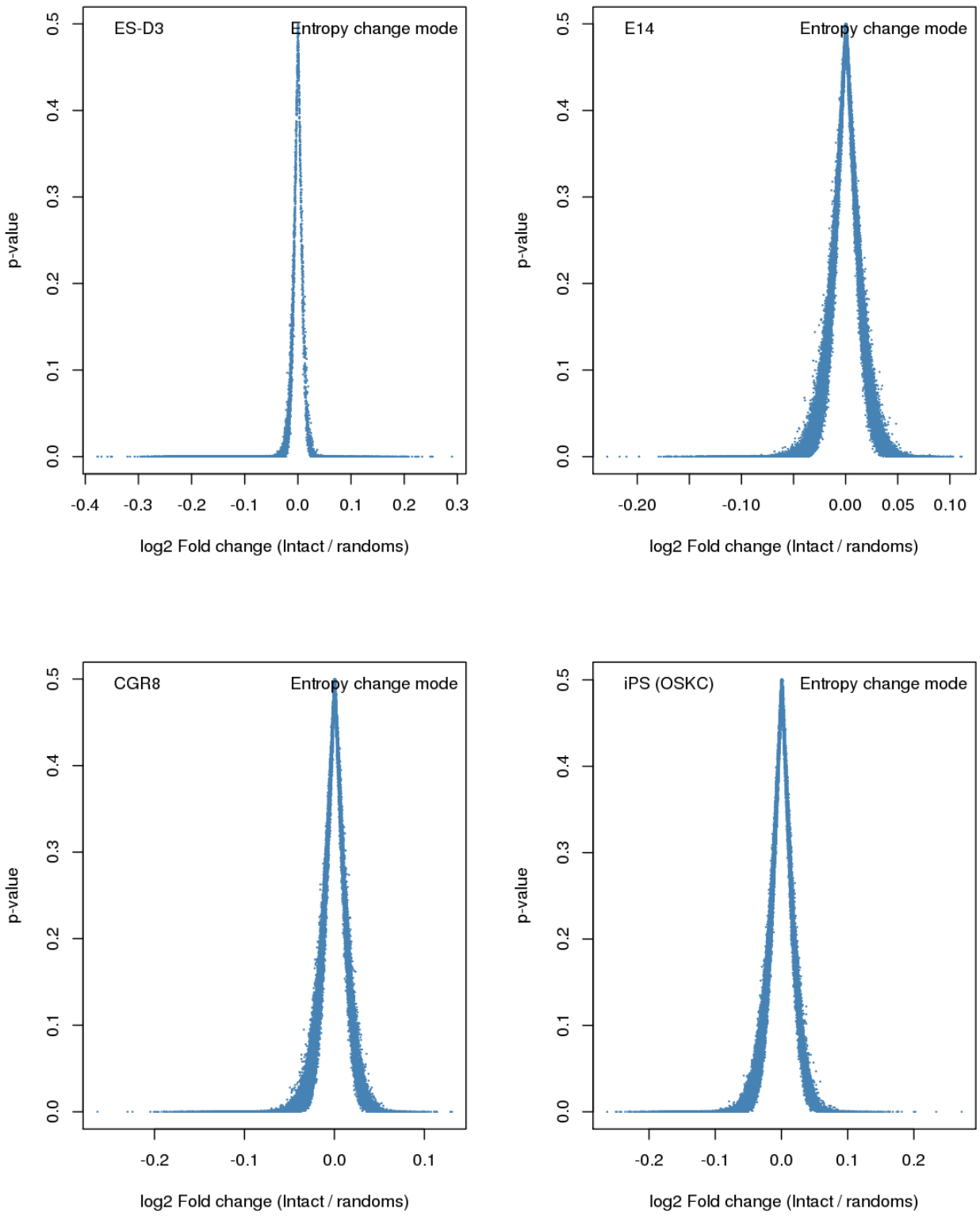

Figure 3.20: Behaviour of DALGES when analysing all four cell lines which contributed at least (n >= 50) samples to the HPM matrix. Relationship between observed entropy change change (x-axis) and p-value (y-axis) is shown.

increases this further to 650,532 p-values satisfying this threshold, representing over 53% of all p-values calculated. Initially this was cause for suspicion as to DALGES' performance. However, the reason for this is that a gene need not have its level of expression conclusively linked to only one cell line. Indeed, it would be highly unlikely for there to be a gene which is only ever elevated or depressed in one cell line, with respect to the entirety of the rest of the data. If cell line-specific signatures can be detected with these sorts of methods, it is far more likely that different levels of expression of the same gene may be associated with different cell lines. The p-values retain their inherent utility, however, in giving any researchers using this methodology the ability to discard mean fold changes which may arise due to chance.

3.8.2 Behaviour of DALGES – Normalised Shannon entropy mode

The running of DALGES using Shannon entropy was carried out in exactly the same manner as was done for the differential expression runthrough, using a total of 1,001 permutations over 7 separate instances, each carrying out 143 permutations using the precalculated bin matrix. The results of this are given in full, including all entropy changes and p-values on the accompanying DVD, as an R object, under “Chapter 3/DALGES/Entropy Change Method/Results R Object”. In the case of differential expression, commonly-used fold changes may be used for the selection of interesting genes. In order to select potentially interesting entropy changes in this data, however, the number of genes chosen for each cell line's analysis was chosen by plotting all entropy changes in ascending order and observing the two “tails” that form in each case, as can be seen in 3.15. These tails were used as a guide for selecting an appropriate cutoff to select genes with strong (negative or positive) entropy changes.

The cutoffs chosen for the four cell lines (ES-D3, CGR8, E14 and iPS (OSKM)) were -0.2, -0.1, -0.1 and -0.125 respectively for decreases in probe entropy, and +0.05 for increases in probe entropy all cell lines analysed. These cutoffs are shown as horizontal red lines on the plots in 3.15.

It was expected that there would be a relationship between the values obtained for differential expression and entropy, in a way not unlike that found for RaSToVa. However, an altogether different relationship is apparent between the values calculated for normalised Shannon entropy and differential expression when DALGES was run on these four cell lines. The relationship between these two metrics is depicted for all four analysed cell lines across figures 3.21 and 3.22. From

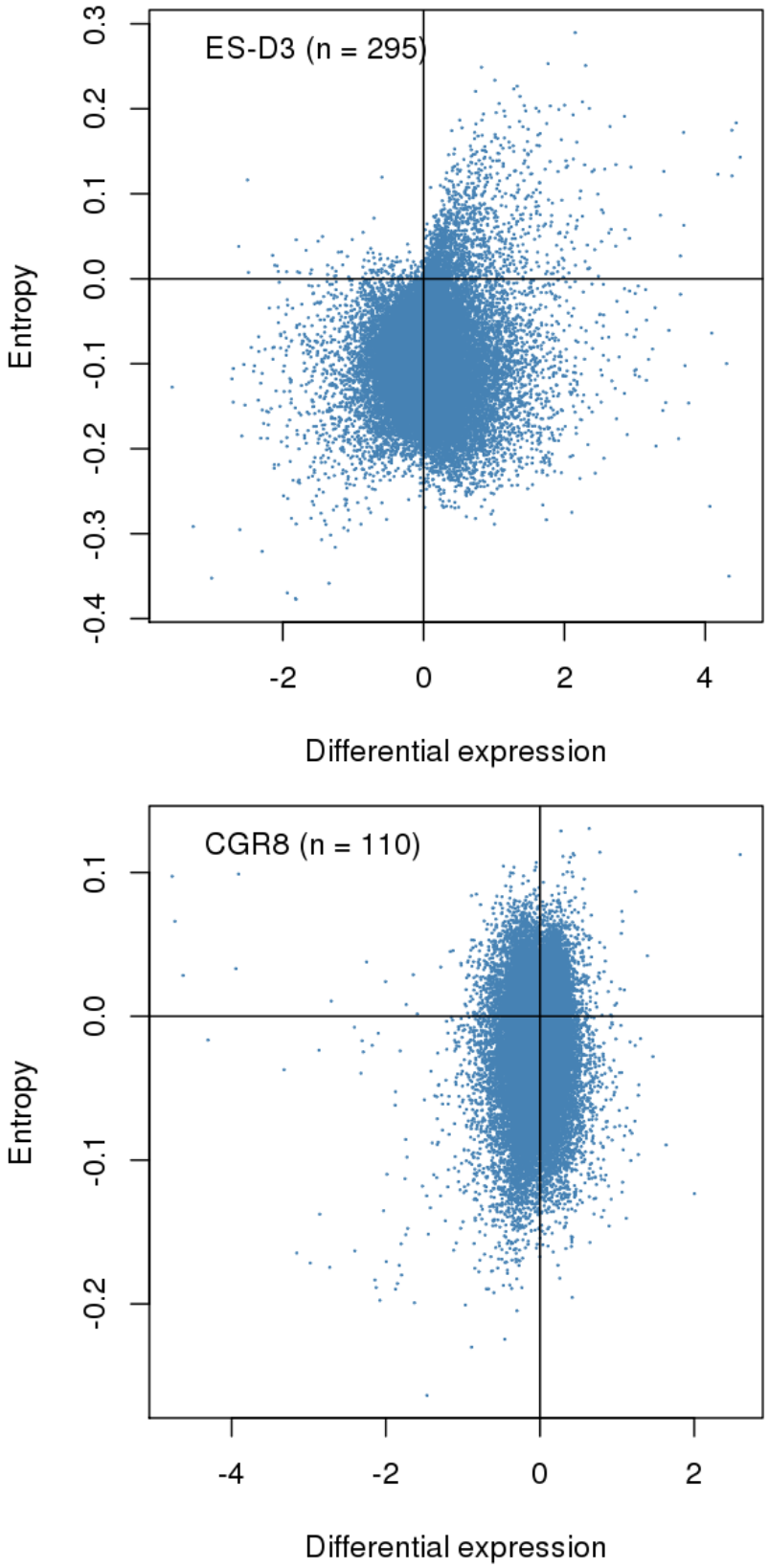

Figure 3.21: Relationship between differential expression and entropy change analyses for all probes in analyses by DALGES for ES-D3 and CGR8 cell lines.

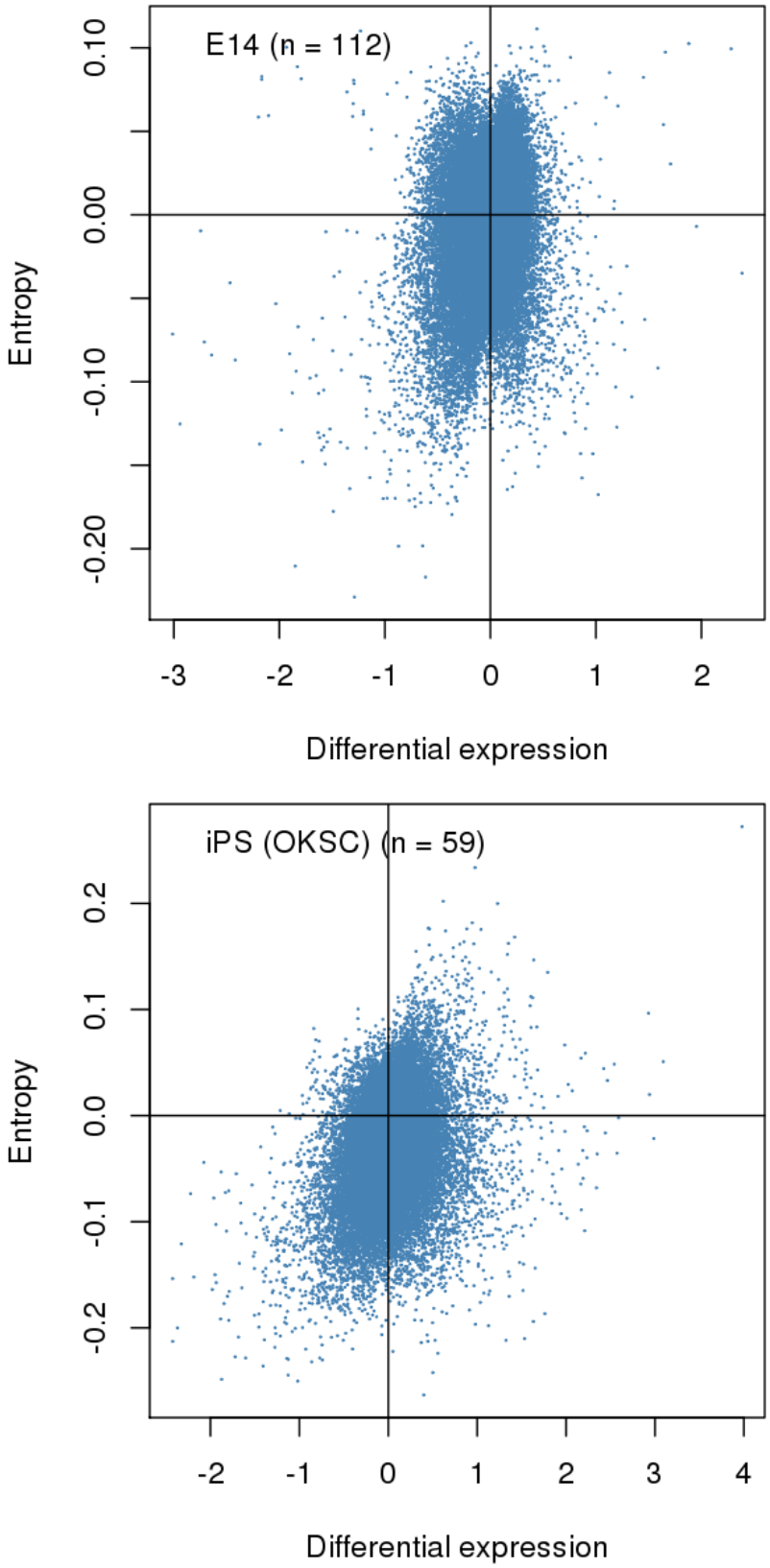

Figure 3.22: Relationship between differential expression and entropy change analyses for all probes in analyses by DALGES for E14 and iPS samples

these, it is clear to see that differential expression and Shannon entropy appear to agree about a majority of the genes being unrelated to any one annotation, as can be seen from the dense centres of the plots for E14, CGR8 and OSKM_IPS cell lines. A crucial observation, therefore, to make from the results of the ES-D3 plot in figure 3.21, is that the cloud of datapoints is not centred near the 0,0 mark on the x and y axes, whereas the results for other cell lines are generally a lot closer to this. This is likely to be due to the fact that the ES-D3 cell line is used mostly by the one laboratory, the “Piersma AH” laboratory (see the data confounding assessments in figures 3.23 and 3.24.) This explains why the vast majority of probes were found to be less entropic in the ES-D3 cell line in figure 3.21; the ES-D3 samples are likely to be more homogeneous than the other cell lines analysed here, as they are mostly from the same source laboratory. Comment on this is made in 5.1.4.

However, what is also clear from these plots is that there are a large number of genes which occupy areas of the plots indicative of having a large negative entropy change (that is, a gene becoming much more predictable within a cell line), yet differential expression analysis suggests that the mean expression of this gene is close to that which would be expected in this matrix, given randomly-permuted submatrices. These genes occur at the zero mark of the x-axis, but stretch below the y-axis' zero mark. Conversely, there are those genes on which the two methods disagree wherein there is a large apparent change in mean expression, according to the differential expression method, but no real change in entropy, implying that the gene becomes no more predictable (or, at least, stays quite variable within the cell line), but the mean expression of that gene changes with respect to what would be expected, given the distribution of the HPM matrix. These genes occur toward the zero mark on the y-axis, but occur away from the zero mark on the x-axis.

Shannon entropy and differential expression methodologies, as used here, may be capturing different aspects of the data, as Shannon entropy also finds in the data many gene lists which are significantly-enriched for biological pathways, for different cell lines, as seen below. Regardless, the ability to demonstrate upregulated or downregulated genes as associated with a cell line by using the differential expression mode of DALGES is believed by the author to be the more useful methodology here, as it more directly suggests differences in cellular function for investigation in the laboratory. The entropy measure may prove useful, however, in detecting confounded analyses, perhaps, as is discussed in 5.1.4.

3.9 The HPM matrix's annotations of source laboratory and cell line are highly confounded

As the development of RaSToVa and DALGES were done in parallel with the annotation of the HPM matrix in chapter 2, it could not be foreseen that there was a high degree of confounding in the HPM matrix which may bring the different effects on transcriptional profile of both source laboratory and cell line so close together that they may not be able be strikingly different. Despite this, RaSToVa still found a conclusive, significant result in that the source laboratory annotation affects sample similarity more than the cell line annotation does, in this matrix. The caveat that the results in this chapter pertain only to and are affected by the distribution of the annotations and samples in this matrix is something which has been mentioned throughout the chapter for emphasis. However, it is difficult to visualise the extent to which the source laboratory and cell line annotations are, in fact, confounded.

To show this in a meaningful manner, a heatmap was generated in which all cell lines and laboratories were shown (figure 3.23). This heatmap represents every sample in the HPM matrix and the annotations for cell line and source laboratory that are attached to them. This is an important point as both RaSToVa and DALGES eliminated some laboratories and cell lines from their analyses for not being represented sufficiently in the HPM matrix, and this is the reason for the occurrence of hitherto unmentioned cell lines and contributing laboratories. For each laboratory, cell lines which its samples used were counted. Then, for each laboratory again, these numbers of samples, as they spread across cell line annotations, were converted into percentages for easy intercomparison (and to enable meaningful colour coding on a heatmap). This heatmap is therefore a “source-laboratory”-centric view of the distribution of annotations. This can be seen by looking at the vertical axis of, for example, the ES-D3 cell line. Some laboratories contributions to the HPM matrix are made up entirely of samples bearing the “ES-D3” cell line annotation. Therefore these are shown in red. Were the heatmap to be “cell line”-centric, then the vertical row of red boxes shown for the ES-D3 line would not be red, as these “cell lines” would be divided amongst “laboratories”, and thus not red. The heatmap was clustered by the default settings of the R function which was used to generate it (the “pheatmap” package) (Euclidean distance, complete linkage), but the dendrograms were removed as they do not represent anything relevant here; clustering was performed simply to improve readability of the heatmap.

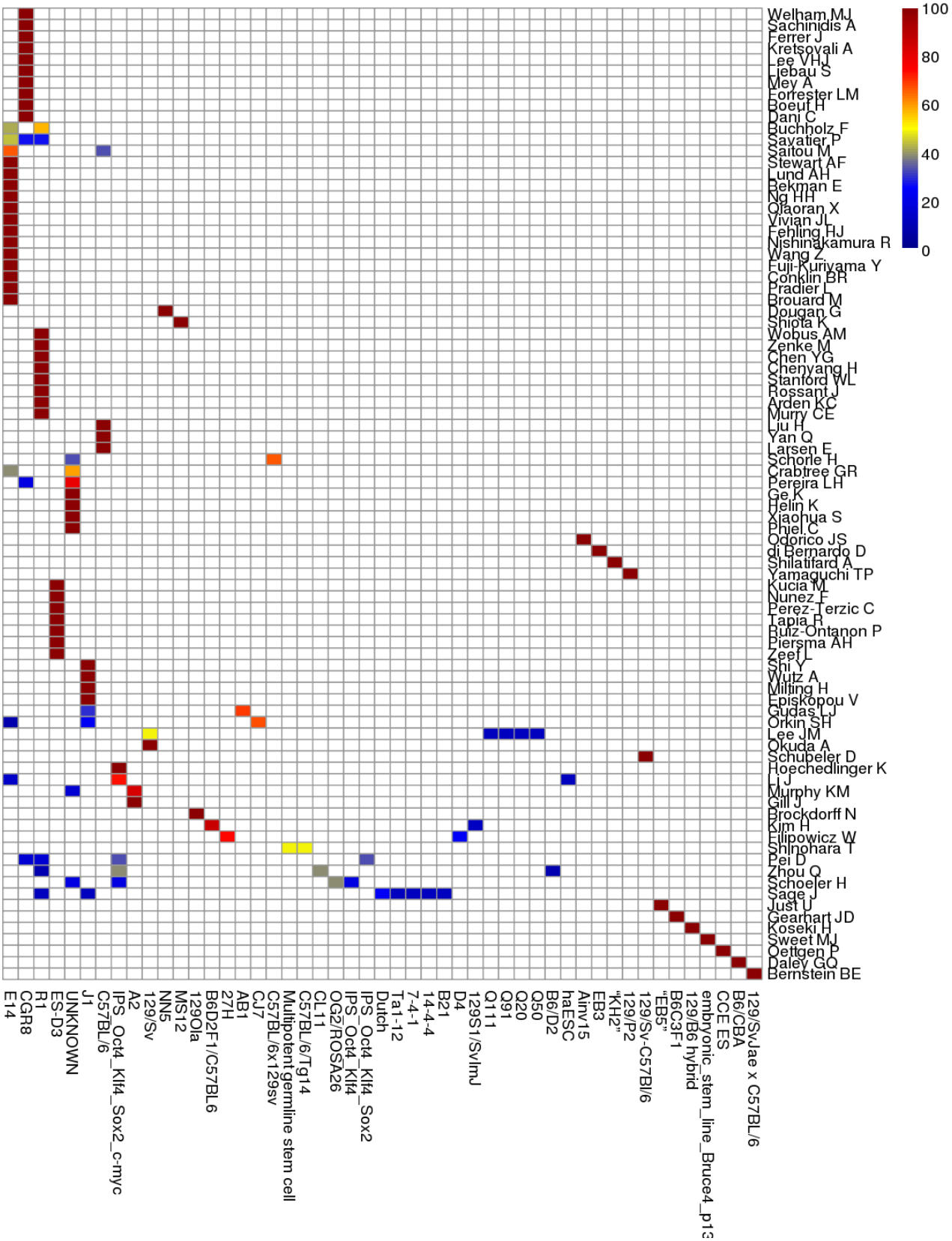

Figure 3.23: Heatmap assessing confounding of the annotations of source laboratory (vertical axis) and cell line (horizontal axis). Values are representative of the percentage of a given source laboratory's samples that are annotated as being from a given cell line.

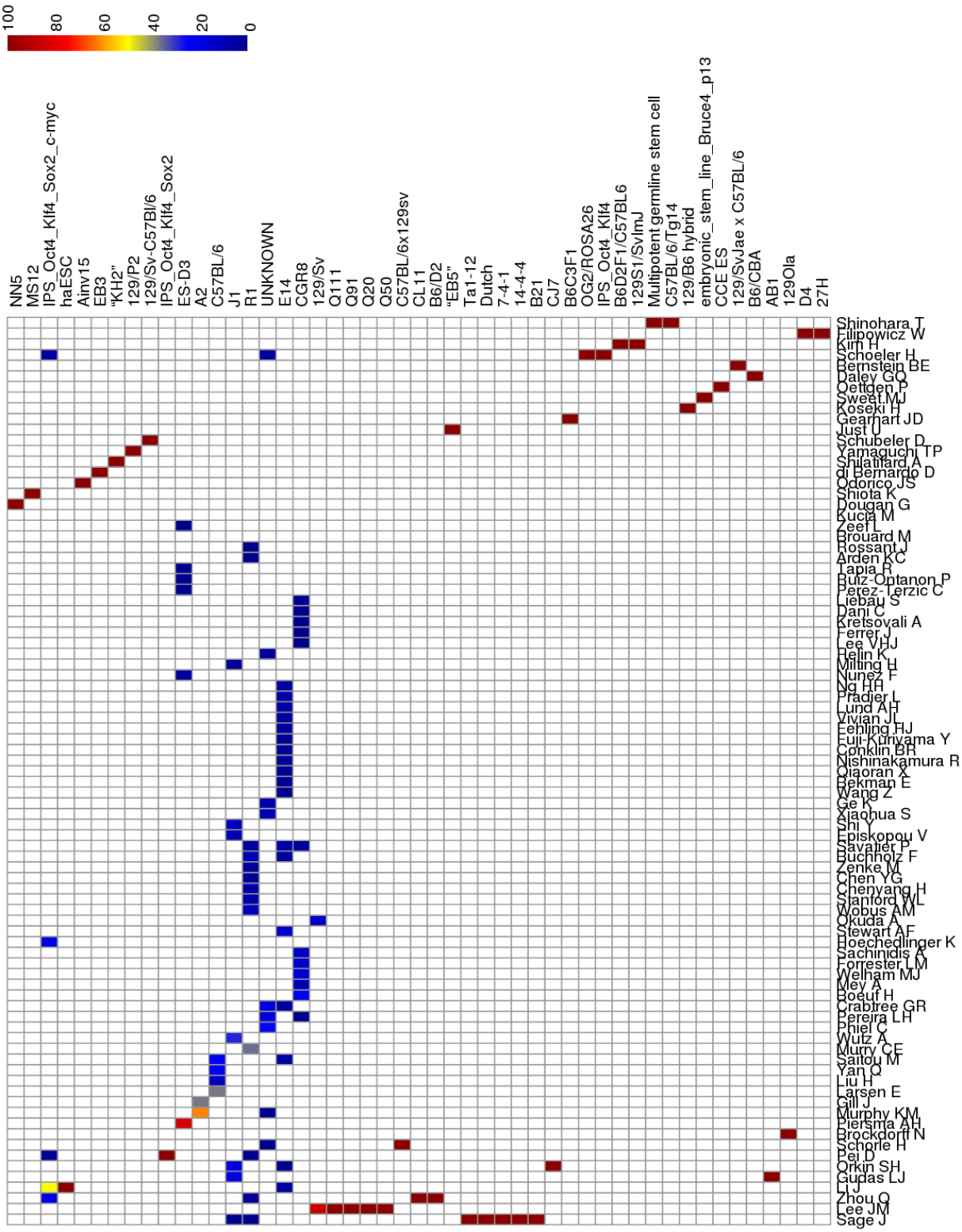

Figure 3.24: Heatmap assessing confounding of the annotations of source laboratory (vertical axis) and cell line (horizontal axis). Values are representative of the percentage of a given cell line's samples that are annotated as being from a given source laboratory. For example the E14 cell line is spread among many laboratories, while the ESD3 cell line is mostly used by the Piersma AH laboratory. At the extreme, cell lines only used by one laboratory are in dark red (100% confounding.)

This process was repeated, although this time with a “cell-line”-centric view of the confounded nature of the annotations. Here it is shown how cell lines are distributed between laboratories, as a percentage 3.24.

It can also be observed from this heatmap that there are those contributing laboratories which not only use only one cell line, but where this cell line is never used by any other laboratory. These can be seen as red squares (where 100% of the samples from the laboratory occur in this one cell line), but where there are no other non-zero (read: non-deep-blue) squares in any of the four cardinal directions from the single square of red.

A final note on the confounded nature of the HPM matrix, as regards source laboratory and cell line annotations, is mentioned in 3.10.5.

3.10 Summary of Research Outcomes

3.10.1 Summary of RaSToVa and DALGES methodologies

The two methods, RaSToVa and DALGES were developed with differing objectives in mind, although DALGES can be seen as an outgrowth of the RaSToVa method. Here an intercomparison is given as to their objectives, strengths and weaknesses. Pseudocode scripts are given for both RaSToVa and DALGES in appendix A.

Objectives

RaSToVa was developed first, and is a novel method for quantifying how much a specific annotation in a given dataset appears to affect the similarity of samples within that dataset.

In the context of this work, RaSToVa was used to investigate whether the annotation fields of “source laboratory” or “cell line” appeared to cause samples to be more similar to each other. This it achieves by a combination of random resampling of subsets of the data, and comparing these to the samples which bear the annotation (e.g. a specific cell line) in question, observing how the samples from the annotation in question are either more or less similar when compared to these randomly-resampled subsets. The final result of RaSToVa is therefore indicative of which annotation appears to make samples most similar. Whilst RaSToVa runs across all of one kind of annotation (e.g. all cell lines) and then another (e.g. all source laboratories) in order to compare cell line and source laboratory, DALGES’ objectives are much more specific.

DALGES, developed second, extends the methodology of RaSToVa by taking the same random-resampling idea and calculates differential expression between the samples that bear one annotation of interest, and all of the other samples, at each iteration. This asks a totally different question to RaSToVa and is their major defining difference. DALGES builds up, over many permutations, a transcriptomic (read: gene expression) signature that appears to be linked with the specific annotation (e.g. the “ES-D3” cell line) in question. In the case of this work, the chosen annotations were specific cell lines chosen for being sufficiently represented in the assembled data. This work therefore investigated whether or not DALGES could find transcriptomic (read: gene expression) differences between the aforementioned cell lines. This it achieved and these gene expression signatures were then successfully mined for GO pathway enrichment.

Strengths

One of the primary, greatest strengths that these two methods share is their fitting to the specific data given to them. As the number of permutations (resamplings and comparisons) increases, both methods gain increasingly-accurate pictures of the patterns within the data. Therefore, the exact distribution of the data is what is taken into account when results are generated. There are no assumptions of distributions, no fitting models applied and no information is lost due to dimensionality reduction.

With both methods, the resampling part uses randomly-resampled subsets of the data which are equal in size to the subset which contain the annotation of interest. This is an approach which therefore is able to contextually assess the statistical significance of the results, which is of particular relevance with DALGES, by repeatedly comparing the subset of interest (e.g. “ES-D3” cells) only ever against a random subset of the data of the same size. This is reflected in the statistical power reported, where comparing a very small number of samples to that same very small number of samples is less informative compared to a larger number of samples. Methods which compare a small subset of the data to a bulk statistic calculated over the rest of the entire dataset have issues concerning the statistical inferences which can be drawn.

A second strength of both methods is the ease of understanding of their results. One does not need an advanced understanding of machine learning, model-fitting or the complex mathematics of dimensionality reduction. In the case of RaSToVa, the measures of inter-sample variability are commonly-used and easy to understand (Euclidean distance and Shannon entropy). In the case of DALGES, which returns gene expression signatures and significances, there is no need for any

deconvolution or reprocessing of the results to attempt to tease out an estimate of a gene expression signature; a signature is provided and statistical significance provided on a probe-by-probe basis. This makes both methods easy to use without the need for a dedicated bioinformatician or statistician and the results are not affected by the method of choice for deconvolution.

A further strength of both RaSToVa and DALGES is that they can be applied to other types of data, not just microarrays, for example NGS technologies such as RNAseq. The flexibility of these methods is another strength, in that the annotations which are fed into the method are in plain text format. This makes it easy for researchers to modify these annotations and even create their own that may not have been in the original dataset, be these derived variables or simply new information. For example, researchers can add new knowledge or test new hypotheses as to what may be driving the patterning in the data. This is as simple as adding a new line of annotations to the annotation matrix.

For example, in the case of stem cell research, the DALGES methodology could be used to investigate the changes in gene expression which appear to accompany certain methods of cellular reprogramming. This is where the use of large datasets becomes integral to the utility of DALGES, where a researcher need not perform a large number of their own experiments to investigate such a phenomenon; they can perform their own treatment / control-paired experiment for their reprogramming method of interest, and then use DALGES to place the gene expression signature of their chosen reprogramming method in the context of however much public data concerning other reprogramming methods that they wish (or can handle.)

Weaknesses

Both RaSToVA and DALGES deliver results based on the exact distribution of the datasets provided to them. Whilst this is a great strength of the methods for the aforementioned reasons, this also brings with it certain limitations.

First and foremost is that RaSToVa will only provide the user with the “relative contribution to sample similarity” for the specific dataset used. It may be that in the data used in this work, there was no clear difference in whether “source laboratory” or “cell line” annotations seemed to make samples more similar to each other. Another, hypothetical dataset may show a clear advantage to one or the other annotation. This does not invalidate RaSToVa’s utility or results, but it does require that any researcher using it be aware of this limitation. Very small datasets would be highly likely to give results that do not hold up to more general analysis. Therefore, RaSToVa results must not be

taken as absolutes across all data. In this hypothetical situation, RaSToVa must be re-run on an expanded dataset which brings together sufficient numbers of samples from those annotations which the researcher wants to interrogate. RaSToVa is therefore limited in that the conclusions that it comes to are only true for the data upon which it is run.

One simple way to potentially mitigate this weakness, or at least make the results applicable to more general datasets, is to include relatively transcriptionally-diverse samples when running RaSToVa, something which should be easy to do with so much public data available. This is still limited, however, by the fact that the available annotations for public data suffer from a myriad of inadequacies and inaccuracies, potentially requiring the user to perform laborious curation, as was necessary in this work. With simpler annotations such as “cell line” or “source laboratory”, however, this should not be a prohibitive time cost (compared to annotations such as culture conditions as were gathered for this work.)

DALGES carries a similar weakness, in that the signatures that it builds are, by definition, relative to the distribution of the data that it is given. The same mitigation strategy could be applied to DALGES as to RaSToVa, by including transcriptionally-diverse samples in a larger dataset. This is the intended method of application of both DALGES and RaSToVa.

Similarly, a weakness, not of the methods *per se*, but of their interpretation, concerns confounded data. If the annotations investigated by the user are highly confounded with one another, this presents a problem for RaSToVa particularly.

When one annotation (e.g. cell line) is largely overlapped by at least one other, it is not possible for RaSToVa to discern which of these two annotations mechanistically is driving sample similarity. Whilst it is well known in the biological research community that “correlation is not causation”, an extra level of caution is required here, where multiple annotations may overlap and appear to be driving sample similarity. This is why RaSToVa should be run on multiple annotations in a given dataset and the contribution to sample similarity compared between these annotations. In the case of this work, where there wasn’t a clear advantage to either “cell line” or “source laboratory”, further investigation was required to reveal the levels of confounding present in this data. More heterogeneous data, again, would be the way to address this issue.

RaSToVa has another weakness, which is that there is no clear, acceptable way to address the fact that the annotations may partition the data into different numbers of subsets. For example, if there are 100 laboratories represented in the data, but only 30 cell lines, then it stands to reason that “laboratory” will appear to be responsible for greater sample similarity than “cell line” would, as

dividing the data into 100 subsets is likely to create more homogeneous subsets than splitting the data only 30 ways (as would be done to split it into subsets for “cell line.”) So is it really true, in the case of this work, that “cell line” and “source laboratory” didn’t contribute strikingly-different amounts to sample similarity, given that “source laboratory” had the advantage of dividing the data many more times?

This limitation of RaSToVa must also inform user’s interpretations of the method as it stands now. Ways to try to control for this kind of effect are considered as future work and will require advanced statistical approaches and substantial testing.

3.10.2 RaSToVa can quantify and make directly comparable the contribution to sample similarity of annotations in microarray data

An important advance of RaSToVa is the ability to directly and meaningfully compare the contribution of different annotations to sample similarity. Obviously, this would not be possible without metrics that are directly comparable, which is exactly what this method is designed to deliver. Furthermore, this method's metrics are calculated in a manner which is intended to be understandable by audiences without expert knowledge of statistics. The method can also be readily used without the need for proprietary software or great knowledge of computer programming. Only basic file manipulation, knowledge of statistics and information theory and some advanced scripting are required in order to use Bioconductor and the statistical programming language R and repeat this analysis on other data.

Limitations of this approach which must be declared boil down, unavoidably, to complications arising from the data and the nature of the samples themselves. For example, it stands to reason that samples from a certain laboratory may well be more similar to each other than samples from without that laboratory simply because the samples from one laboratory are likely to include samples in similar conditions being manipulated in similar manners. If all of the samples from one laboratory are a single experiment investigating the effects of a certain treatment, then it would be expected that these samples would bear great similarity to one another, certainly when compared to other samples not from this experiment. Conversely, if one source laboratory has contributed several different experiments to the gathered data, then more variability would be introduced simply because multiple experiments are present. However, it also stands to reason that the effect of such

sampling issues is diminished the more data is gathered, in the case of this work, 1,101 individual samples.

Another potential issue arises from the overlap between different annotations. That is to say that certain laboratories will use many cell lines, whereas other laboratories may well only use a couple, or even just one. At first appearance, this may seem to be greatly disruptive to our efforts to delineate the effects of the two. However, this is not the case as the limits seemingly imposed on our ability to ask our questions of the data are actually controlled for within the method. If it is the case that a large laboratory only uses one cell line (a worst-case scenario for trying to delineate the effects of laboratory *vs* cell line), this will actually be "subtracted out" in the final conclusion as, when the analysis is run for laboratory and then for cell line, highly-similar (they are extremely unlikely to reach being identical due to the random sampling steps) values would be produced for these samples, meaning that, exactly as intended, the contribution to sample similarity of both laboratory and cell line for these samples would result in a "draw", which may actually reflect perfectly the issues surrounding our fundamental question when it comes to the field of microarray data analysis. That is to say that precisely because of the inherent complexities of the annotations of microarray experiments, it is extremely difficult to ascribe a greater or lesser effect to annotations such as laboratories or cell lines without specifically-designed experiments for that purpose, as the distribution of data available simply does not allow for the unambiguous and definite delineation of these effects. This is what urges caution against drawing conclusions regarding the important sources of sample similarity before large-scale analysis such as this has been carried out, such as has been done in the case of suggesting source laboratory as the primary factor in (Newman & Cooper 2010). Our analysis' conclusions are therefore robust in the face of the complex nature of the data as the statistics calculated by it show, without bias, whether or not annotations can be observed to be associated with more or less of the similarity in that data. To summarise – if the nature of the data masks the relative effects of source laboratory / cell line etc., then this will be reflected in the results and no false positive will be found, whereas if the data is capable of showing it, delineation of the effects will be successful.

3.10.3 Source laboratory contributes marginally more sample similarity to samples in the HPM matrix than the cell line annotation

It was found during this chapter that, given the distribution of the data in the HPM matrix and particularly, with respect to the distribution of the annotations that accompany this data, the contribution to sample similarity of the "source laboratory" and "cell line" annotations were similar, as can be seen from the final comparison figure 3.13, although source laboratory was found to contribute marginally more sample similarity. In euclidean distance mode, a significant difference was found between the contributions to sample similarity of the "cell line" and "source laboratory" annotation with $p = 0.049$, whereas the entropy change method tended towards significance with a p-value of $p = 0.061$.)

This suggests that, from this data, being from the same laboratory makes samples more similar to each other than does simply being of the same cell line. The confounded nature of the data used was not apparent during the development of the methods in this chapter as the annotations were still being completed. Special notes on the confounded nature of the data are given in 3.10.5. After all, RaSToVa can only ask questions about annotations and their contribution to sample similarity given the distribution of the data that it is given.

3.10.4 The DALGES methodology finds transcriptional profiles which may be linked to cell lines

This chapter also details the use of the DALGES methodology (see section 3.5) to investigate the possibility of linking cell lines with transcriptional profiles. Again, despite the development and assessment of the behaviour of a method which is suited to this function, it must be stressed that all methods in this chapter were developed and made ready while the annotations of the data were still being completed and, therefore, the level of confounding between source laboratory and cell line was not known at the time. This chapter therefore contains a first attempt at the assignment of transcriptional profiles to cell line annotations in this data, but the points set out regarding the confounded nature of the data in section 3.10.5 must be stressed.

Regardless, some interesting results are obtained through the use of random permutation and differential expression / entropy change analyses here, as are summarised in figure 3.16. This chapter does not attempt to infer too much about the biology which may be behind the finding of

the biological pathway enrichments that resulted from the use of DALGES in investigating these four different cell lines, for two reasons. Firstly is the aforementioned issue of the confounded nature of the data (see section 3.10.5), but also because the goal of this chapter was more to test the possibility of using a method such as random permutation followed by differential expression / entropy change to such a question. The method itself showed its utility in this chapter, but it was decided not to concentrate on making likely premature statements about the biological differences between these cell types in this data due to the confounded data issue.

Even so, it is interesting to see that the Wnt receptor signalling pathway is so prevalent among these enrichments, possibly suggesting that there are indeed differences in endogeneous Wnt activity between these cell lines. This would be in agreement with (ten Berge et al. 2011) who found that certain cell lines, including CGR8 and E14, produce their own Wnts.

Other potential differences were found in this analysis and so brief comment is offered here on what enrichment for different pathways may mean, to underscore the utility of investigating using this same methodology in more heterogeneous data: Cell lines found with upregulation of developmental pathway genes may suggest that these cell lines are more prone to differentiation than others. Likewise, those found with increased expression of proliferative genes may have advantages in cell proliferation and turn over more quickly than other cell lines. Other named signalling pathways occur in this summary figure also, such as BMP, VEGF and MAPK/ERK signalling. Inherent differences in these signalling pathways would be very interesting to investigate and could prove most informative in future work when choosing different cell lines to investigate phenomena that may be affected by such signalling pathways. Interestingly, there was also enrichment found in this summary figure for “stress response”. Enrichment for the stress response biological pathway here may suggest that some cell lines have decreased DNA repair, antioxidant or other defences. As a reduction in DNA repair may be concomitant with a general increase in replicative vigour, the finding of an enrichment for stress response warrants future investigations of the differing capacities of commonly-used cell lines in mESC research for their relative levels of DNA repair enzymes, antioxidant capacity *et cetera*. The presence of an enrichment for apoptosis-related pathways, for example, may indicate that a given cell type is more prone to cell death, which may be of particular interest when, for example, comparing data contributed by groups generating iPS cells using different methods, which neatly brings about the last point in this summary:

One final and extremely interesting use of this methodology may be in comparing the efforts of groups involved in generating iPS cells. Intercomparison of the transcriptional profiles of iPS cells from different groups, along with ground state mESCs may offer new insight into the effects of different iPS generation technologies, and how they affect different pathways. For example it may be found that iPS generation methods that are harsher on the cells in question may select for and subsequently recover and multiply cells with enhanced DNA repair capabilities and resistances to other stresses, but these may be compromised in their ability to, for example, proliferate. A methodology such as DALGES may be able to use transcriptomic data (whether microarray or, as is rapidly becoming the fashion, RNAseq) to make testable predictions about these cells which could, in turn, provide early warning to laboratory groups in the form of seemingly upregulated and downregulated pathways in their iPS cells, as compared to mESCs and other groups iPS generation efforts. With increasing amounts of public data available, methodologies such as DALGES, designed to be easily understandable by non-bioinformaticians and, indeed, which do not require any non-free software, would ideally allow groups in future to rapidly compare their own efforts to this rapidly-expanding amount of public data, increasing robustness and providing insight into which methodologies / cell lines *et cetera* appear to be associated with which biological phenomena. Again, it was decided to treat only as a suggestion, at this point, the results summarised in figure 3.16, due to the confounding of the data, which can be visualised in figure 3.23.

3.10.5 Special addendum on the confounded nature of the HPM matrix

The most important caveat to take into consideration when interpreting all of the results from this chapter, as has been repeatedly pointed out in the previous text, is the confounded nature of the annotations in the HPM matrix. In brief, the manual annotations took a great deal of time to reliably complete and the methodologies in this chapter were being tested out informally, during that time, on small subsections of the data. Randomised data with spiked-in “similarities” or “annotation associated levels of gene expression” were also sometimes used to test whether or not RaSToVa / DALGES were successful in finding these spiked-in phenomena. Therefore, the methodology of RaSToVa, in its entirety, and most of the functionality of DALGES (excepting only the use of pre-calculated matrices) were in place at about the same time as the annotations were coming to an end. After final checking of the annotations, it became clear the level of lab / cell line confounding that was present in the annotations (see figure 3.23).

Therefore it must be said of both RaSToVa and DALGES, crucially, that these methods both have performed their tasks to the letter for the data and annotations that were available. To the author's knowledge, this is still the largest to-date analysis of high-pluripotency-marker mESC microarray data. There was initial concern that this confounding would result in absolutely no difference being found between the contribution to sample similarity of lab or cell line, and worse, that there would be wholly uninteresting pathway enrichments found when applying DALGES to the analysis of transcriptional profiles associated with cell lines. Luckily this was not the case and both methods have generated very interesting initial results; RaSToVa found that “lab” contributes slightly, but significantly ($p = 0.049$, by euclidean distance method), more sample similarity than “cell line” does, and also there were pathway enrichments assigned to the different cell lines (E14, CGR8, ESD3 and an iPS cell line for interest's sake). These pathways appear to be relevant to mESC biology, being concerned with signalling pathways, proliferation, stem cell-related pathways, apoptosis, stress response *et cetera*. This caveat is placed here simply as a precaution against taking these results to be reproducible in larger data as it is the authors wish to carry out both the RaSToVa and DALGES methods on much larger data with far less confounding between crucial annotations.

Chapter 4 –

Investigation of Transcriptional Events Associated with Progression from Pluripotency to Early Differentiation

4.1 Research Questions

4.1.1 Overview

Mouse embryonic stem cells are not one particular cell type in one particular state. *In vivo*, mESCs progress from a naïve, ground state (Ying et al. 2008) through to a primed state of pluripotency (Brons et al. 2007), (Tesar et al. 2007). The objective of this chapter as a whole is to leverage the large (n = 1,101) high-pluripotency-marker matrix of microarrays in an attempt to investigate the transition between these two states. An overview of and details of the experiments behind the characterisation of the primed pluripotent state are already given in this work (see section 1.5). For completeness, however, a section in this introduction is given over to summarising the key differences which define the naïve and pluripotent mESC states (see section 4.1.7).

Whilst the pluripotent state and the transcriptional networks that underpin it have been the subject of much focussed research already (see (Nichols and Smith, 2012), (Yeo and Ng, 2013), (Chambers and Tomlinson 2009), (Young RA, 2011) for excellent reviews and the relevant parts of chapter 1 for a brief overview), there remains comparatively less understanding of the exit from pluripotency (Young RA, 2011), particularly regarding the very earliest transcriptional events which may occur as naïve pluripotency starts to progress towards primed pluripotency. Most experimental observations of the initiation of differentiation and the beginning of the collapse of pluripotency are made alongside a drop in known pluripotency markers such as Oct4, Nanog or especially in naïve pluripotency markers such as falling Rex1 and rising FGF5 (Sene et al. 2007).

The work undertaken in this chapter can be summarised into the essential steps given below.

4.1.2 The presence of sufficient and relevant information in the HPM matrix for useful interrogation

As the HPM matrix has been filtered for only the highest levels of Oct4, Sox2 and Nanog, there was a risk that such filtering rendered the dataset too homogeneous for meaningful interrogation, perhaps losing vital information concerning progression between naïve and primed pluripotency. Two questions required answering to address this potential pitfall. Firstly, a general assessment of the loss of information as a result of the aforementioned filtering was undertaken. It would be expected that the matrix would become significantly (but not drastically) more homogeneous than if samples were only removed at random, as the intention of the filtering is to leave only *bona fide*

pluripotent samples. This was also tested (and found to be true), see section 4.2.1 for methods. Secondly, even if the information content lost was not deemed to be excessive, meaningful relationships between genes relevant to mESC biology should still exist. This was assessed by investigating whether canonical pluripotency factors maintained strong, expected relationships with other genes in the data relevant to both maintenance of mESC pluripotency and mESC priming / differentiation.

4.1.3 Determination of a suitable gene for ordering the HPM matrix between pluripotency and early differentiation

To interrogate the HPM matrix for information regarding progression from naïve to primed pluripotency, the data required sorting broadly between samples bearing the transcriptional profile of naïve pluripotency, to those of primed pluripotency and beyond. As experimental annotations could not be easily used or relied upon to achieve this, the approach taken to order the HPM matrix was to identify a single gene whose expression broadly sorted the HPM matrix between naïve and primed pluripotency. Details on the criteria for the selection of this gene are given in the appropriate methods section, section 4.2.2.

4.1.4 Assessment of the ordering of the HPM matrix between naïve and primed pluripotency

The success of the broad ordering of the HPM matrix between naïve and primed pluripotency was gauged through a combination of methods, given in more detail in section 4.2.3. These methods used both known naïve / primed pluripotency markers and experimental annotation cross-referencing to confirm that the desired broad sorting of the HPM matrix was successful. This assessment was entered into without the unreasonable expectation that a single sorting gene (or even a set of genes) would or could create a perfect spectrum between pluripotency and early differentiation (see method section 4.2.3 and results section 4.3.4).

4.1.5 Use of the ordered HPM matrix to observe transcriptomic changes between naïve and primed pluripotency by differential expression analysis

With matrix N1101 broadly sorted between naïve to primed pluripotency and beyond, interrogation of this matrix for transcriptional events occurring between these states was undertaken. Differential expression analysis is ubiquitous in gene expression research throughout the biological sciences, and involves calculating differences in the level of expression of genes in one group of samples versus another, where a particular biological phenomena of interest separates the two groups.

associating the resulting changes in gene expression with the biological event. In this case, four areas of interest exist: “early” naïve pluripotency, “late” naïve pluripotency, primed pluripotency and early exit from primed pluripotency.

As the matrix’s samples were broadly sorted into a progression of these cellular states, marker profiles for those cellular states were used to identify groups of samples in each cellular state for the differential expression analysis (e.g. low levels of Rex1 and high levels of brachyury (T), Otx2 and FGF5 are known to define the primed pluripotent state (see sections 1.5 and 4.1.7)). Using these different regions, several differential expression analyses were performed and finally, to investigate the relevance of the resulting lists of differentially-expressed genes to mESC biology, pathway enrichment analysis was carried out. For further details of the methodology, see 4.2.4.

4.1.6 Potential of a scanning window approach for investigation of transcriptional events, states and pathway enrichments across the Klf4-ordered HPM matrix

Following the differential expression analysis approach, the final work in this chapter asks the most crucial question: can the HPM matrix, ordered broadly by cellular state, reveal novel genes which change their expression significantly at any given point between these cellular states, specifically the “early” to “late” naïve pluripotency transition? Whilst differential expression will likely capture transcriptional profiles pertaining to the beginning and end points of any cell state transition as the expression of the ordering gene decreases, there is great interest in mapping which changes take place at which times / in what order, between early and late naïve pluripotency. There may well be genes whose expression changes only transiently from early naïve to late naïve / primed pluripotency, and therefore may be missed by a more simplistic “start-versus-end-point” analysis. To scan across the change from naïve to primed pluripotency as a process, rather than a switch, is precisely the utility of such an ordered matrix.

Therefore the final approach taken in this chapter is a scanning-window approach to ask: which genes significantly change their expression at any time as these cellular states change, particularly between early naïve and late naïve pluripotency? Does a scanning window approach across the whole matrix find all of the pathways enrichments that other analyses found? Which of these genes were not already known in the literature? These genes can be further investigated experimentally as potentially sensitive markers of different stages of the transition between naïve to primed puripotency in mESCs. The specifics of the methodology are given in section 4.1.6.

4.1.7 Summary of defining attributes of naïve versus primed mESCs

Several good markers are already known in the literature which signal the exit from naïve pluripotency and are used to effectively separate naïve and primed mESCs. Most canonically of these are the moderate drop in pluripotency markers Oct4 and Nanog, but large drop in Rex1, along with a concomitant large rise in expression of Otx2, FGF5 and brachyury (T) (Sene et al. 2007). These markers are the ones used in this work to definitively draw a line between the two states as these are amenable to easy quantification in microarray data. The objective of this chapter is to view the transition between these two states as a gradual change, and look for any as-yet unknown factors which may accompany this transition. This is one of the advantages of using large numbers of samples which span a biological phenomenon of interest (Ramasamy et al. 2008).

Though a detailed literature review is beyond the scope of this thesis, here is given a brief summary of several crucial, defining differences between the naïve and primed mESC states. naïve mESCs are not dependent on MEK-ERK signalling, FGF2 or TGFβ/Activin A signalling , while primed mESCs are. There is also a pronounced switch between the use of the distal to the proximal enhancer of Oct4 as primed pluripotency arises. Another defining point of ground-state mESCs is their globally hypomethylated state, which gives way to increasing methylation and therefore transcriptional restriction as the primed state comes about. naïve mESCs have not undergone X-inactivation, while primed mESCs have (X_aX_a vs X_aX_i respectively). Nanog, Oct4 and Sox2 drop moderately from naïve to primed pluripotency, as do the Klfs and Esrrb. The Prdm14/Nanog-based maintenance of pluripotency is also lost from naïve to primed pluripotency. Further, there is a switch from the expression of E-cadherin molecules to N-cadherin molecules. The cells also change their metabolic activities as a whole, moving from a mixture of oxidative phosphorylation to a glycolytic state. Finally, another property that is striking between these two states is that whilst naïve mESCs are readily capable of being coaxed towards primordial germ cell-like cells (PGCLCs), whilst, in contrast, primed mESCs are notoriously refractory to this.

4.2 Methods

4.2.1 Confirmation that the HPM matrix retains transcriptional information relevant to pluripotency

For the general information loss assessment, a measure of the overall information content in matrices N3312 and N1101 (the HPM matrix) was taken. This was done by simply calculating the normalised Shannon entropy of all probes in both matrices.

The distribution of entropies within both matrices was then compared with a view to demonstrating that there was not a drastic reduction in the amount of information in the HPM matrix after filtering all samples for the highest levels of OSN. Means of probe entropies were calculated for both the N3312 matrix and the filtered N1101 matrix.

Furthermore, as touched upon in section 4.1.2, the purpose of the filtering was to leave behind only *bona fide* pluripotent samples, and so it would be expected that, now representing only a few cellular states and the transition between them, the HPM matrix would be more homogeneous than if samples were simply removed at random, but not drastically so. This was confirmed using a random permutation approach to ask what the expected remaining information content of the dataset would be if samples were removed at random, rather than in a targeted manner by filtering for the highest expression of pluripotency factors Oct4, Sox2 and Nanog.

Datasets were therefore generated by randomly sampling 1,101 samples from the N3312 matrix (without replacement) and the same calculation of all probe entropies was performed on these. A mean probe entropy was calculated for each randomly resampled matrix. This was repeated for 50 random resamplings. The mean probe entropies of the 50 randomly-resampled matrices (n = 1,101) were then shown in comparison to the observed mean probe entropy of the Oct4, Sox2, Nanog-filtered matrix.

To confirm that the filtered dataset still contained information pertaining to mESC pluripotency, Pearson correlation of all three pluripotency factors was calculated for each OSN factor, to all 45,100 other probes. This resulted in 3 lists of correlations that were then, for enrichment analysis, separated into positive and negative correlations to each of the OSN factors. Lists of genes found to have absolute Pearson correlations stronger than +/-0.5 for each pluripotency factor were considered to have strong relationships with the OSN factors. Pathway enrichment analysis was

then carried out on these lists separately to ascertain whether or not the OSN factors still maintained detectable transcriptional relationships with genes enriched for pathways related pluripotency and exit from it. Pathway enrichment analysis was carried out using the DAVID bioinformatics tool and pathways found to have a q-value of less than 0.05 were considered to be significantly enriched.

4.2.2 Selection of a suitable guide gene for ordering the HPM matrix

The method for selection of a suitable ordering gene involved four criteria:

Firstly, the ordering gene should have a wide range of expression in the data; this larger range in a guide gene providing for gradation between one cellular state and another. This was calculated, per probe, as that probe's maximum expression in the HPM matrix, minus the minimum expression.

Secondly, the ordering gene should occupy a maximal number of states between its minimum and maximum value and thus provide for a spectrum of values from high to low, again providing for as smooth a high-to-low transition as possible across the data.

Thirdly, the ordering gene must change its expression in direct correlation to pluripotency as a biological phenomenon. This was scored by Pearson correlation to one of the canonical pluripotency factors Oct4, Sox2 or Nanog, whichever retained the greatest information content (read: entropy) in the HPM matrix (see section 4.3.3 for details.)

Fourth, the ordering gene should ideally, when chosen by the above scoring methods, pass a final criterion of having a known, publsihed role in mESC pluripotency / exit from it.

To score and rank genes for both their correlation to one of the canonical OSN pluripotency factors as well as their information content in the data, a multiplicative score was calculated as the product of the absolute Pearson correlation to the pluripotency factor and the normalised Shannon entropy of that gene (absolute correlation multiplied by entropy.) A candidate list of the top scoring genes was generated and then the best candidate gene with a role in mESC pluripotency was selected.

4.2.3 Confirmation of the broad sorting of samples between naïve and primed pluripotency by the selected ordering gene

To verify that ordering the HPM matrix using the selected ordering gene (Klf4) does indeed generate a matrix broadly sorted between naïve and primed pluripotent samples, observation of known marker patterns was undertaken. Known markers which differentiate between naïve and primed pluripotency (FGF5, Rex1 and brachyury (T)) were observed across the ordered matrix, to confirm progression from $FGF5_{low}$ $Brachury_{low}$ $Rex1_{high}$ (naïve) samples to $FGF5_{high}$, $Brachyury_{high}$ $Rex1_{low}$ samples (primed). The chosen markers were therefore plotted as smoothed lines alongside the decreasing values of gene expression of Klf4.

To confirm that this desired progression of change in known naïve / primed marker genes is, in fact, due to the choice of Klf4 as the ordering gene, this same method of plotting the expression of these marker genes was performed when ordered instead by a housekeeping gene, GAPDH. Conversely, to demonstrate that a gene correlated to, but with higher information content than, the pluripotency marker with the highest information content (Nanog), achieves a smoother progression of change in naïve / primed pluripotency marker profiles, the same plot was repeated, but this time ordering the matrix by Nanog.

To demonstrate the validity of using the multiplicative (entropy x correlation) scoring method, and rule out that Klf4 was only coincidentally useful for matrix ordering purposes, another gene from the candidate list of ordering genes was chosen (Jam2) and the same progression of markers plot generated.

Experimental annotations were checked across several regions of the Klf4-ordered matrix as further verification of the broad sorting of samples between naïve and primed pluripotency. Regions where Klf4 was notably changing expression were chosen, and annotations cross-referenced for a randomly-chosen sample that falls in this area. In addition to the individual sample, if other samples from the same experiment are to be found in the HPM matrix, these were cross-referenced as well and annotations checked.

Results of this (detailed in 4.3.4) were generated both as a table and graph for visualisation. The table consists of a Klf4-rank-ordered list of experimental accession numbers, sample accession numbers, and summary of sample annotations, a Klf4-rank (where 1 is the highest expression of

Klf4) and a colour code on a scale of green to dark red between naïve pluripotency and differentiation respectively (see section 4.3.4 for details).

For the graphical form and to represent the progression of naïve pluripotency to early differentiation as being concomitant with decreasing Klf4, the marker method and the annotation method were combined into a single figure. Here, the same colour coding from the aforementioned table was superimposed (as vertical lines at their corresponding Klf4-rank positions) over smoothed lines showing the Klf4-ordered changes in FGF5, Brachyury (T) and Rex1 (Zfp42).

4.2.4 Gross differential expression analysis of guide-gene ordered HPM matrix

For the differential expression analysis between regions of the Klf4-ordered HPM matrix representing different cellular states, four such regions of interest were defined using both Klf4 decrease and the aforementioned published markers that differentiate between naïve and primed pluripotency, these being FGF5, brachyury (T), and Rex1. See results section (4.3.5) to see these regions along the Klf4 spectrum.

Between each region, differential expression was calculated by subtraction of the mean of the log2-transformed RMA values of one block of samples from another. A 50-sample wide region of the Klf4-ordered HPM matrix was used to represent each region / cellular state (early naïve pluripotency, late naïve pluripotency, primed pluripotency and early exit from primed pluripotency.) Differential expression was calculated in this manner and coupled to a test of statistical significance using a Mann-Whitney U test, on a probe-by-probe basis. A threshold of significance was set at $p < 0.05$. The relationship between the observed fold change in gene expression and the p-value was also plotted for the first analysis (highest-Klf4 region vs lowest-Klf4 region) to ensure expected behaviour of such an analysis.

Specific attention was focused on the region between early and late naïve pluripotency, as this was the primary research question of interest in this chapter. Manual interrogation of the genes which changed their expression significantly between these two defined regions of the Klf4-ordered HPM matrix was carried out. Full lists of those genes which significantly changed expression across this region of the matrix were retained for this thesis, and both a heatmap of selected changes and a

barplot summarising good candidate genes was plotted. Details of these are found in the corresponding results section, section 4.3.7.

As with previous analyses, pathway enrichment analysis was carried out using the DAVID bioinformatics tool for further interrogation of the meaning of the groups of genes found to be significantly changing their expression ($p < 0.05$). For these pathway analyses, a q-value of < 0.05 was considered to be significant.

4.2.5 Development and calibration of a scanning window approach for the detection of transcriptional events across the Klf4-ordered HPM matrix

Whilst this method is referred to as a "window-scanning" approach, the method actually involves two separate windows; a "start" window and a "moving" window. These two windows can be thought of as being analogous to the "from" and "to" comparisons of a traditional differential expression analysis.

The "from" window encompasses the first group of adjacent samples (at the high Klf4 side of the spectrum) and the moving, "to" window starts also at the high Klf4 side of the spectrum, but incremented by one sample towards the lower Klf4 end. Effectively, if 50 was chosen as the number of samples in each window, then the "from" window would be comprised of samples 1 to 50 along the Klf4 spectrum (where 1 is highest Klf4) and the moving, "to" window would begin by containing samples 2 to 51.

A mean expression value for the current probe in each of the two currently selected groups of samples is calculated, and differential expression analysis performed between them. If mean expression of that gene (probe) has passed a set threshold (e.g. 1 log2 fold change), then the same statistical test as in the previous analyses (section 4.2.4) is performed, the Mann-Whitney U test, to assess the statistical significance of this change.

The "from" window then is moved to the current position where the change was found (essentially, the "from" window now references exactly the same group of samples as the moving, "to" window does, and then the moving, "to" window is moved along again by one sample, to start looking for the next significant change in the expression of the current probe (if any.)

The results of the scanning window analysis are stored in two matrices. In both matrices, each row represents a probe on the microarray (in this case, 45,101 probes.) The second dimension has a size of:

$$s - (w-1)$$

Here *s* is the number of samples in the data, being 1,101 microarrays, and *w* is the width of the scanning window. This corresponds to the number of windows that will actually be scanned by the method before the moving window iterates its way to hitting the end of the matrix (in this case, the lower end of Klf4 expression.) A graphical representation of this method is given in figure 4.1.

To arrive at satisfactory values for both the scanning window widths and the threshold of expression change which should be used, it was decided to run the scanner over the whole matrix with different values for each, as a form of calibration. In order to render this section of the work directly comparable to the previous, differential expression analysis section (see), the 1 log2 fold change was one of the thresholds used, although tests of the method were also run with the threshold set at 1.5, 2 and 2.5 log2 fold change. As concerns the scanning window width, tests were carried out using 3, 5, 10, 15, 20, 25, 30, 35, 40, 45 and 50-sample-wide windows.

The number of genes significantly changing their expression at any given scanning window was also recorded, as a way to observe which regions along the Klf4 expression spectrum the most interesting changes may be taking place. As a mean is calculated in both the “from” and “to” windows, these means may be affected by outliers as the scanning window width is made smaller. The scanning window size tests therefore sought to find the smallest window that could be used before the plots of “number of genes changing expression in this scanning window” appeared to become noisy, see results section 4.3.9 for detail on these plots.

With an acceptable scanning window width and an expression change threshold set, the scanning window analysis was performed on the ordered HPM matrix between the same regions detailed previously: the transition of greatest interest (early to late naïve pluripotency), but also late naïve to primed pluripotency, primed pluripotency to early differentiation, and a full run across all of the data. For each of these, the total number of genes that changed expression by more than 1 log2 fold and that also passed the significance threshold ($p < 0.05$) were recorded. Pathway enrichment

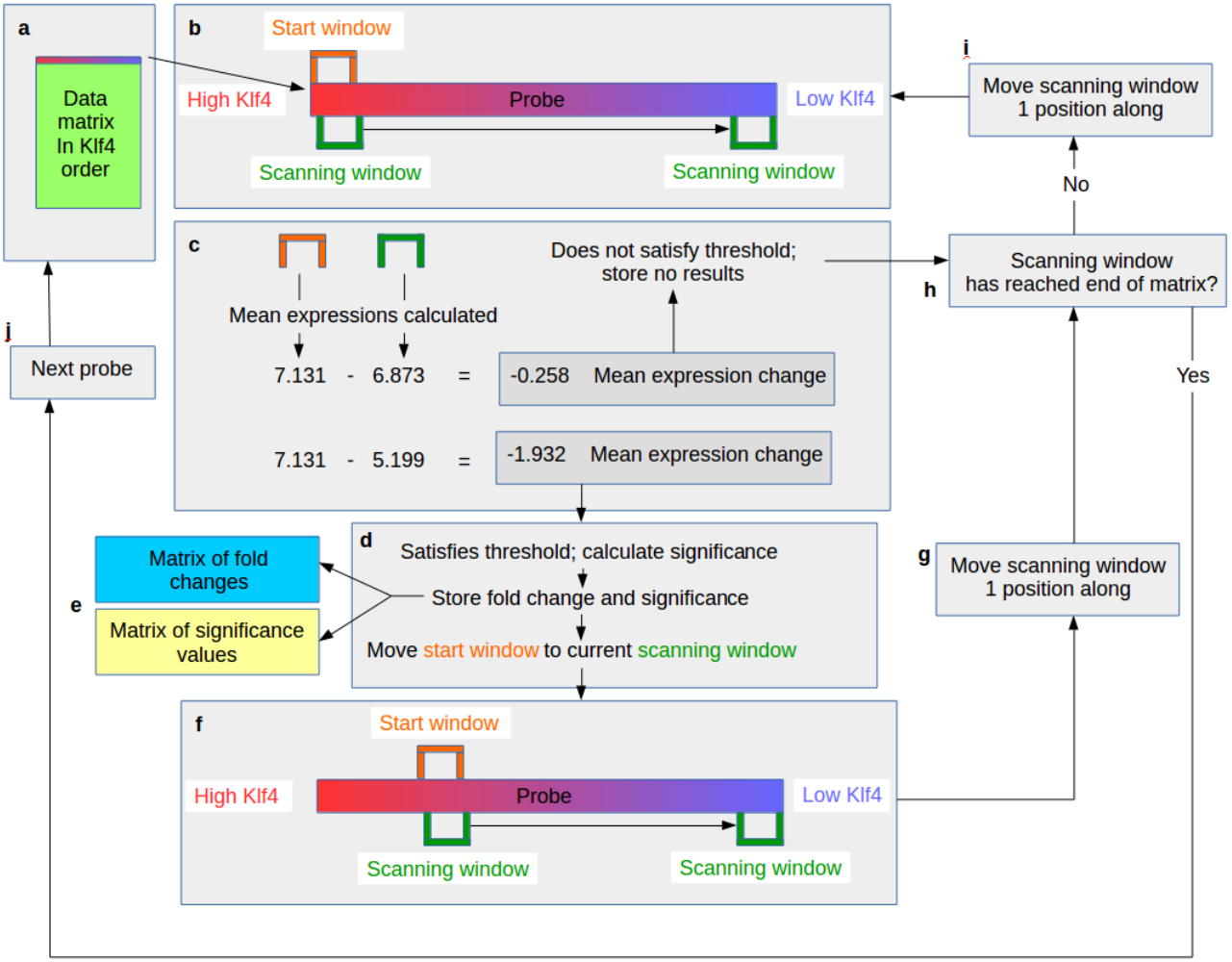

Figure 4.1: Graphical overview of the window scanning approach to analysis of transcriptional events in the Klf4-ordered HPM matrix. For this description, a window width of 25 samples and a threshold of 1 log2 fold change is used. From the whole data matrix **(a)**, for each probe in turn, a starting window is drawn across the 25 highest-Klf4 values for this probe. A scanning window is also drawn, but offset by one value towards the lower Klf4 end of the matrix from the starting window **(b)**. This scanning window is moved towards the low-Klf4 values for this probe and, at each increment, the mean expression (mean of values in the window) of the scanning window is compared to the start window **(c)**. If the scanning window has hit the end of this probe **(h)**, then that ends the analysis of this probe, and the next probe is analysed, with scanning and starting windows returned to their initial positions from **(b)**. If the observed fold change (mean of scanning window minus mean of starting window) does not satisfy the threshold of 1 log2 fold change in either direction, the scanning window is moved towards the low-Klf4 end of this probe **(h),** then **(i),** then **(b)**. However, if the fold change is found to be greater (in either positive or negative direction) from the starting window **(c)**, then a p-value is calculated for this change by Mann-Whitney U test **(d)**. These results are then recorded for this scanning window for this probe **(e)** and the starting window is moved to the current position of the scanning window **(f)**. The scanning window is then moved along **(g)** and the process repeated from **(c)**, unless this was the last scanning window **(h)**, in which case the process starts from the next probe **(j),** with scanning and starting windows returned to their initial positions from **(b)**.

analyses were carried out on those genes for each of the 4 analyses for comparison to the differential expression analysis described in section 4.2.4.

Finally, a summary comparison between the total numbers of genes found to significantly change their expression, as well as the total numbers of pathways found to be significantly enriched between the different regions of the Klf4-ordered HPM matrix, for both the differential expression analysis and scanning window analysis were computed and are discussed in results section 4.3.13.

4.3 Results

4.3.1 The HPM matrix retains substantial information content post-filtering

Normalised Shannon entropy was calculated on a per-probe basis for both the HPM matrix and for the matrix from which is was made, the N3312 matrix. A histogram showing the distribution of probe entropies for both of these matrices is given in the left side panels of figure 4.2. The distribution of entropies of probes in both matrices is also shown as “s-plots” provided in the right hand panels of figure 4.2 Here, the entropies of all probes for each matrix are sorted into ascending order. The mean probe entropy for both matrices was 0.629 (3 d.p.) and 0.565 (3 d.p.) for N3312 and the HPM matrix respectively. These means are marked on both the histograms and the “s-plots” by black lines (figure 4.2).

Comparing the mean probe entropy of the HPM matrix (0.565 (3.d.p.)) to the distribution of mean probe entropies from the randomly-permuted (n = 1101 sample) submatrices from N3312, it is clear that a larger reduction in mean probe entropy has occurred than that which would be expected by simply drawing 1101 samples at random from matrix N3312 (see figure 4.3). A direct comparison is shown between the distribution of probe entropies in these two matrices in the form of the final part of figure 4.4 in the form of boxplots.

This preliminary look at the variability of the unfiltered (N3312) and filtered high-pluripotency-marker (HPM) matrices shows that there is a small overall reduction in the amount of variability present. Mean probe entropy reduces by 11.317% (3 d.p.) (0.064 (3 d.p.)).

Between matrix N3312 and the HPM matrix, there is a reduction of 2211 samples, reducing the number of samples by two-thirds (66.757% (3 d.p.)), it might be expected that entropy would decrease considerably.

The mean of the mean probe entropies of the randomly permuted submatrices was 0.627 (3 d.p.), providing a benchmark of the amount of entropy expected to remain in a matrix of size n = 1,101 if the samples remaining were no more homogeneous than those removed. The reduction in mean probe entropy in the HPM matrix due to the specific nature of samples removed from matrix N3312

is therefore estimated as (0.0627 – 0.0565 = 0.062 (3 d.p.)). This confirms that the HPM matrix has become more homogeneous as a result of the specific samples remaining, rather than simply as an

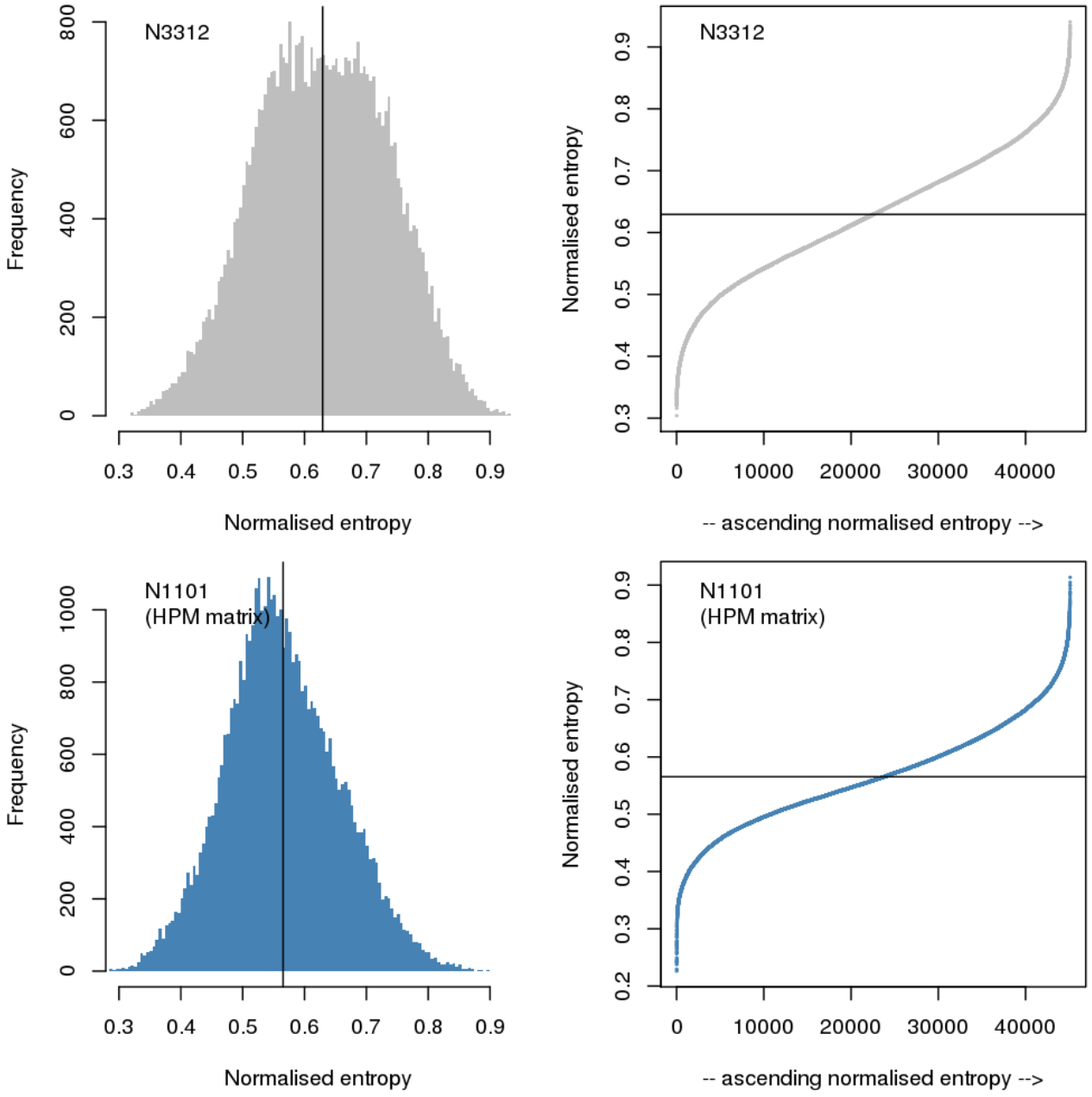

Figure 4.2: Distribution of individual probe entropies in matrices N3312 and the HPM matrix (N1101). Upper panels pertain to matrix N3312 (grey) and lower panels to N1101 (blue). Mean probe entropy for both matrices is denoted by vertical black lines in the case of histograms and horizontal black lines in the case of the accompanying "s-plots". S-plots are generated by ordering all probe entropies into ascending order and plotting. X axes in "S-plots" are therefore simple numeric indices indicating normalised entropy rank. Mean normalised probe entropy was 0.629 (3 d.p.) and 0.565 (3 d.p.) for N3312 and the HPM matrix respectively, a very small, but significant ($p < 0.01$) decrease of 0.064 (Mann-Whitney U test .)

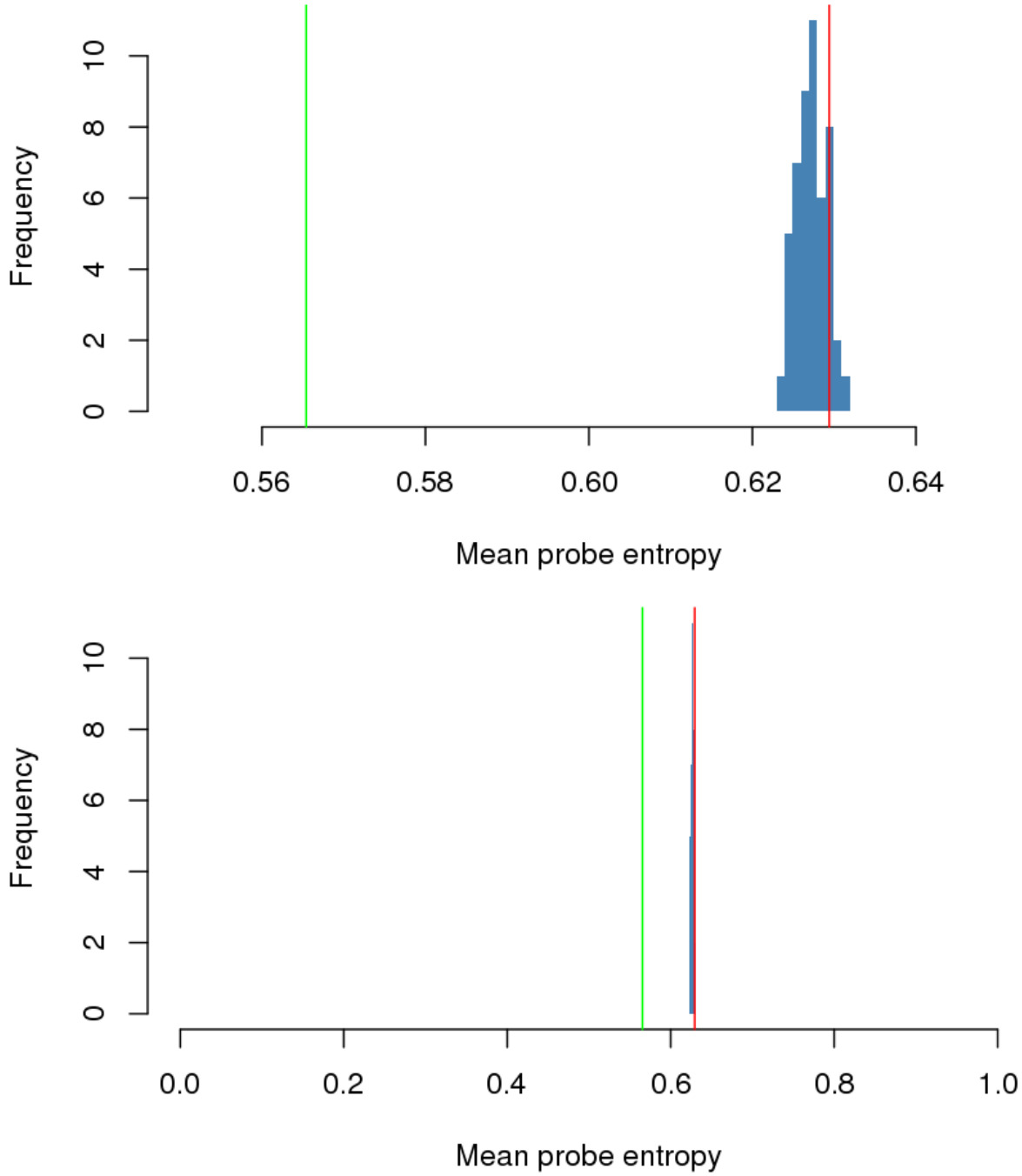

Figure 4.3 Upper panel: Histogram (blue) showing the distribution of mean probe entropies of 50 randomly-permuted submatrices (size: n=1101 samples) drawn, without replacement, from matrix N3312. Red vertical line shows mean probe entropy of intact matrix N3312. Green line shows mean probe entropy of intact HPM matrix, showing reduction of mean probe entropy in the HPM matrix below what would be expected were 1101 samples randomly selected from matrix N3312. Lower panel shows the same data as upper panel, but with x-axis scaled between 0 and 1, to put mean probe entropy difference between matrices N3312 and N1101 (HPM matrix) into perspective, relative to minimum and maximum possible values for mean probe entropy. N1101 mean probe entropy: 0.565 (3 d.p.). N3312 mean probe entropy: 0.629 (3 d.p.). Mean of mean probe entropies (randomly permuted submatrices): 0.627 (3 d.p.). Difference between mean randomly-permuted mean probe entropy and observed mean probe entropy of HPM matrix: 0.062 (3 d.p.)

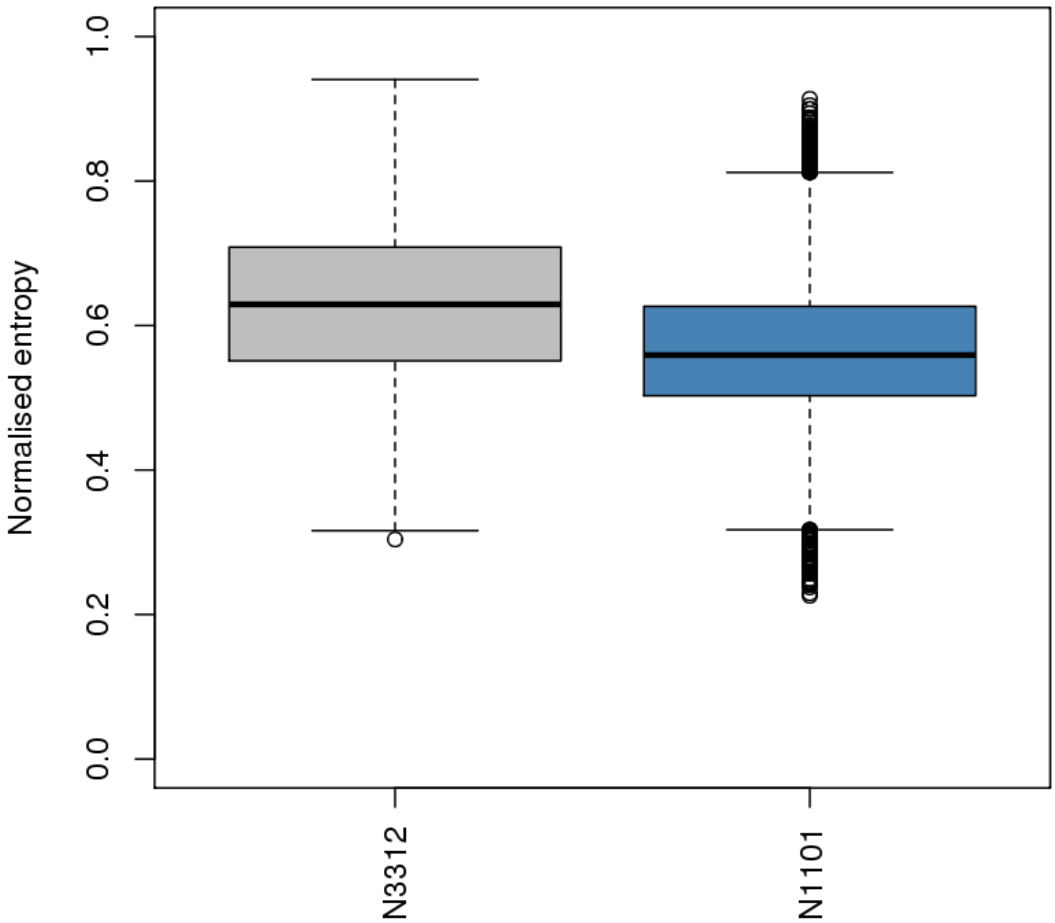

Figure 4.4: Boxplots showing distribution of all individual probe entropies in matrices N3312 (grey) and N1101 (blue.) Mean normalised probe entropy was 0.629 (3 d.p.) and 0.565 (3 d.p.) for N3312 and the HPM matrix respectively, a very small, but significant ($p < 0.01$) decrease of 0.064 (Mann-Whitney U test .)

effect of reducing the number of samples. This is as expected, given the fact that the filtering was carried out to leave only samples in the pluripotent state in the data.

It is reassuring to see, however, that the distribution of probe entropies in the HPM matrix still extends towards a similar maximum as that found in N3312, as can be seen in the histograms as the upper parts of the “s-plots” and the outlier datapoints in the boxplot for the HPM matrix.

4.3.2 Genes correlated to pluripotency factors Oct4, Sox2 and Nanog are enriched for pluripotency and developmental pathways in the HPM matrix

Nanog

A total of 374 probes were strongly (>= 0.5) correlated to Nanog. Analysis of these using DAVID revealed “stem cell differentiation”, “stem cell maintenance” and “stem cell development” at the top of the enrichment list, confirming that Nanog is still strongly related to pluripotency genes in the HPM matrix. See table 4.5 for the top 20 pathways positively correlated to Nanog and their q-values for significance testing.

Other pathways in this list are somewhat generic, however, the “negative regulation of cell differentiation” is good to see here as it reinforces the notion that Nanog remains positively correlated with the maintenance of pluripotency in the HPM matrix.

281 probes were found to be strongly negatively (<= -0.5) correlated to Nanog. Only 3 biological pathways were significantly enriched on analysis of this list, and these were again relevant to development with “pattern specification” and “embryonic morphogenesis” as the top 2 pathways, although these are also still generic in nature (see table 4.6).

These observations confirm that enough variability does indeed remain in the HPM matrix for Nanog to maintain strong positive transcriptional relationships with pluripotency-related genes and negative transcriptional relationships with genes involved in development. This also suggests that the HPM matrix contains samples which occupy various states of pluripotency.

GO Identifier	GO Term	Probe Count	Q-Value
GO:0048863	stem cell differentiation	8	0.0003666531
GO:0019827	stem cell maintenance	7	0.0004074534
GO:0048864	stem cell development	7	0.0003630761
GO:0045596	negative regulation of cell differentiation	14	0.0009710921
GO:0010605	negative regulation of macromolecule metabolic process	23	0.0013858654
GO:0010558	negative regulation of macromolecule biosynthetic proce	19	0.0100389704
GO:0031327	negative regulation of cellular biosynthetic process	19	0.0123903548
GO:0009890	negative regulation of biosynthetic process	19	0.0122127128
GO:0010629	negative regulation of gene expression	18	0.0175203962
GO:0045934	negative regulation of nucleobase, nucleoside, nucleotid	17	0.0341304401
GO:0051172	negative regulation of nitrogen compound metabolic proc	17	0.0349265405
GO:0016481	negative regulation of transcription	16	0.0448353961
GO:0045449	regulation of transcription	52	0.0534723809
GO:0006357	regulation of transcription from RNA polymerase II promo	21	0.0695225272
GO:0051252	regulation of RNA metabolic process	38	0.0752671583
GO:0006355	regulation of transcription, DNA-dependent	37	0.1024060666
GO:0045892	negative regulation of transcription, DNA-dependent	13	0.1527650631
GO:0051253	negative regulation of RNA metabolic process	13	0.1528982041
GO:0006020	inositol metabolic process	4	0.1535277825

Table 4.5: Top 20 pathway enrichments to be found in those genes with strong (>= 0.5) correlation to Nanog (1429388_at) in the HPM matrix. Green shaded items satisfy the q-value of <= 0.05, red shaded items fall short of this threshold.

GO Identifier	GO Term	Probe Count	Q-Value
GO:0007389	pattern specification process	15	0.0024415436
GO:0048598	embryonic morphogenesis	15	0.017354873
GO:0006928	cell motion	15	0.0147399277
GO:0001944	vasculature development	11	0.0950920398
GO:0003002	regionalization	10	0.1039757003
GO:0030326	embryonic limb morphogenesis	7	0.1053291318
GO:0035113	embryonic appendage morphogenesis	7	0.1053291318
GO:0007167	enzyme linked receptor protein signaling pathwa	11	0.1072771416
GO:0007169	transmembrane receptor protein tyrosine kinase	9	0.1387809296
GO:0016477	cell migration	10	0.1288570923
GO:0001568	blood vessel development	10	0.1300514974
GO:0035108	limb morphogenesis	7	0.1373425752
GO:0035107	appendage morphogenesis	7	0.1373425752
GO:0060173	limb development	7	0.1489690888
GO:0048736	appendage development	7	0.1489690888
GO:0001763	morphogenesis of a branching structure	7	0.173909216
GO:0060429	epithelium development	10	0.1843859941
GO:0016055	Wnt receptor signaling pathway	7	0.1825593299
GO:0051674	localization of cell	10	0.2161113282

Table 4.6: Top 20 pathway enrichments to be found in those genes with strong (<= -0.5) anticorrelation to Nanog (1429388_at) in the HPM matrix. Green shaded items satisfy the q-value of <= 0.05, red shaded items fall short of this threshold.

Oct4

In the case of probes positively correlated to Oct4, surprisingly, only 39 achieved the required threshold of a Pearson correlation of +0.5 or greater. No pathway enrichments resulted from this list, however, in that they did not pass the threshold of a q-value of <= 0.05. Significance aside, the strongest enrichment was for “ncRNA metabolic process”, not clearly related to mESC pluripotency whatsoever. This unexpected result is in stark contrast to the enrichments found when using a list of probes strongly correlated to Nanog.

As for those probes strongly negatively correlated to Oct4 in the HPM matrix, there are only 29. These probes, however, do contain significant (q <= 0.05) biological pathway enrichments, but only 7 (see table 4.7). These pathways include a clear developmental signal, with “pattern specification process”, “anterior / posterior pattern formation” and “regionalization”. As generic as they are, it still appears that Oct4 is negatively-correlated with progression towards organismal development in the HPM matrix. Given that the genes strongly anticorrelated to Nanog were found to be enriched for “pattern specification process”, it is reassuring to see that genes anticorrelated to another core pluripotency factor, Oct4, would, if anything, show similar kinds of enrichment.

Sox2

The probes found to be positively correlated to Sox2 are much more numerous than those found for Nanog and Oct4, with 763 probes passing the threshold of a Pearson correlation of >= +0.5. The topmost 27 of these pathways are shown in table 4.8, truncated after the emergence of those pathways which began to fail the q-value threshold of (q <= 0.05).

Reinforcing the notion that the HPM matrix contains information relating to pluripotency and exit from it, Sox2's role as a transcriptional regulator is clearly apparent, based on this set of biological pathways enrichments, despite Sox2 being filtered for expression only at its highest levels. The top two pathway enrichments “transcription” and “regulation of transcription” achieve two orders of magnitude of significance greater than the next most significant pathways, being the “stem cell maintenance”, “stem cell differentiation” and “stem cell development” pathways. Interestingly, the “cellular response to stress” pathway is enriched here, which would imply that the stress response may be involved in pluripotency maintenance or exit from it. This is supported by the literature which shows that such a link exists already (reviewed in (Tower 2012)). The presence here of an

GO Identifier	GO Term	Probe Count	Q-Value
GO:0007389	pattern specification process	6	0.0052354157
GO:0009952	anterior/posterior pattern formation	5	0.0047607001
GO:0003002	regionalization	5	0.0117232144
GO:0016477	cell migration	5	0.0136827644
GO:0048870	cell motility	5	0.0208080411
GO:0051674	localization of cell	5	0.0208080411
GO:0006928	cell motion	5	0.0452314654
GO:0048514	blood vessel morphogenesis	4	0.0768528769
GO:0001568	blood vessel development	4	0.1193766609
GO:0001944	vasculature development	4	0.1140253971
GO:0007423	sensory organ development	4	0.111152388
GO:0006355	regulation of transcription, DNA-depende	7	0.1517592812
GO:0051252	regulation of RNA metabolic process	7	0.1505307221
GO:0014855	striated muscle cell proliferation	2	0.2127647374
GO:0055017	cardiac muscle tissue growth	2	0.2127647374
GO:0060038	cardiac muscle cell proliferation	2	0.2127647374
GO:0060419	heart growth	2	0.2127647374
GO:0045449	regulation of transcription	8	0.2190172779
GO:0014706	striated muscle tissue development	3	0.2142271474

Table 4.7: Top 20 pathway enrichments to be found in those genes with strong (<= -0.5) anticorrelation to Oct4 (1429388_at) in the HPM matrix. Green shaded items satisfy the q-value of <= 0.05, red shaded items fall short of this threshold.

GO Identifier	GO Term	Probe Count	Q-value
GO:0006350	transcription	93	0.000033215
GO:0045449	regulation of transcription	107	9.86625029713029E-005
GO:0048863	stem cell differentiation	9	0.0010954051
GO:0048864	stem cell development	8	0.0011390779
GO:0019827	stem cell maintenance	8	0.0013544643
GO:0010558	negative regulation of macromolecule biosy	31	0.0016365935
GO:0010605	negative regulation of macromolecule metal	35	0.0019067664
GO:0031327	negative regulation of cellular biosynthetic	31	0.0024922886
GO:0009890	negative regulation of biosynthetic process	31	0.0027153538
GO:0045934	negative regulation of nucleobase, nucleosi	29	0.0032687432
GO:0051172	negative regulation of nitrogen compound m	29	0.0035589491
GO:0010629	negative regulation of gene expression	29	0.0048987656
GO:0016481	negative regulation of transcription	27	0.0058369992
GO:0042127	regulation of cell proliferation	34	0.0070531874
GO:0006357	regulation of transcription from RNA polyme	37	0.0082153588
GO:0045892	negative regulation of transcription, DNA-d	23	0.0136753938
GO:0051253	negative regulation of RNA metabolic proce	23	0.0140807971
GO:0045596	negative regulation of cell differentiation	16	0.0328496834
GO:0051252	regulation of RNA metabolic process	67	0.0355069218
GO:0033554	cellular response to stress	26	0.0373216472
GO:0006355	regulation of transcription, DNA-dependent	66	0.0379339181
GO:0007093	mitotic cell cycle checkpoint	6	0.0496997581
GO:0007049	cell cycle	33	0.0882390692
GO:0016570	histone modification	10	0.1150521012
GO:0016568	chromatin modification	17	0.1206253269
GO:0031328	positive regulation of cellular biosynthetic p	30	0.124343158
GO:0006974	response to DNA damage stimulus	19	0.1259572895

Table 4.8: Top 27 most significantly-enriched pathways to be found in those genes with strong (>= 0.5) correlation to Sox2 (1416967_at) in the HPM matrix. Green shaded items satisfy the q-value of <= 0.05, red shaded items fall short of this threshold.

enrichment for the stress response pathway informed later decisions to check for stress response enrichments in downstream analyses of the HPM matrix.

Probes strongly negatively correlated to Sox2 in the HPM matrix were more numerous again, at n = 1559. This extensive list contains a large number (n = 411) of pathways which satisfy the requirement for statistical significance of q <= 0.05. This therefore would be wastefully large to include as a figure, and so only selected pathways, including all of those discussed in the text, from the full list are shown in table 4.9. The majority of the full list reads like a catalogue of developmental processes, which is extremely encouraging as regards the nature of the HPM matrix.

However, it is not only enrichments for developmental pathways to be found in genes negatively correlated to Sox2. Excitingly, other enrichments found here pertain to specific signalling pathways, namely the Wnt, Notch, vascular endothelial growth factor (VEGF) and BMP pathways. Enrichment also exists for the MAPKKK cascade, which is downstream of FGF signalling required for exit from pluripotency in mESCs. The TGFβ pathway is listed, but did not achieve significance.

It is highly interesting that this initial analysis of the pluripotency factors Oct4, Sox2 and Nanog in the HPM matrix has already shown signs of pointing towards the activity of specific signalling pathways.

In summary, the HPM matrix appears to be fit for purpose and seems to have captured a range of pluripotency-related processes in mESCs which can be interrogated for meaningful biological information, at a minimum, providing potential transcriptional relationships and the activity of signalling pathways before a large drop in Oct4, Sox2 or Nanog has occurred, meaning that the samples, and therefore the phenomena found in this data, can therefore be reasonably assumed to still be occurring pre-differentiation.

4.3.3 Klf4 satisfies all criteria for use as an ordering gene of the HPM matrix

The choice of guide gene was ultimately made when considering those genes which ranked highest for their multiplicative score for normalised entropy and correlation to Nanog. The multiplicative score using Nanog was preferred over those of Oct4 and Sox2 as it was found that Nanog, of the pluripotency factors for which the HPM matrix was filtered (OSN), maintained the greatest

GO Identifier	GO Term	Probe Count	Q-Value
GO:0007389	pattern specification process	63	9.18790E-018
GO:0048598	embryonic morphogenesis	68	8.18795E-016
GO:0048729	tissue morphogenesis	53	5.02522E-015
GO:0016055	Wnt receptor signaling pathway	32	1.18817E-009
GO:0007167	enzyme linked receptor protein signaling pathway	45	3.39466E-008
GO:0030111	regulation of Wnt receptor signaling pathway	14	7.50321E-006
GO:0042127	regulation of cell proliferation	61	1.40663E-005
GO:0007169	transmembrane receptor protein tyrosine kinase	30	5.29017E-005
GO:0045944	positive regulation of transcription from RNA pol	43	0.0001908645
GO:0045165	cell fate commitment	24	0.0003038948
GO:0045596	negative regulation of cell differentiation	27	0.0004050975
GO:0007178	transmembrane receptor protein serine/threonine	16	0.0008064064
GO:0007219	Notch signaling pathway	13	0.001052487
GO:0048010	vascular endothelial growth factor receptor signa	7	0.0018177997
GO:0000165	MAPKKK cascade	18	0.004665708
GO:0030509	BMP signaling pathway	8	0.0059278705
GO:0006916	anti-apoptosis	14	0.0211733329
GO:0043406	positive regulation of MAP kinase activity	11	0.02603797
GO:0001945	lymph vessel development	4	0.0502856697
GO:0060284	regulation of cell development	19	0.0534004272
GO:0009100	glycoprotein metabolic process	18	0.0545471968
GO:0051347	positive regulation of transferase activity	17	0.0551778307
GO:0051338	regulation of transferase activity	22	0.0582013598

Table 4.9: Selected significantly-enriched pathways to be found in those genes with strong (< 0.5) anticorrelation to Sox2 (1416967_at) in the HPM matrix. Green shaded items satisfy the q-value of <= 0.05, red shaded items fall short of this threshold.

information content (as measured by normalised entropy), with a normalised entropy of 0.646 (3 d.p.), versus 0.550 (3 d.p.) and 0.618 (3 d.p.) for Oct4 and Sox2 respectively. For the purposes of deciding on the ordering gene, the absolute correlation to Nanog was used to generate the multiplicative scores. This was to allow for selection of genes with a strong relationship with Nanog, regardless of whether or not that relationship was positive or negative. A truncated list of the top candidates found by this method is given in table 4.10.

This list of multiplicative scores included, as its 8^{h} rank candidate (when excluding Nanog itself), Klf4 (probe ID 1417394_at). With a normalised entropy of 0.816 (3 dp.), this gene is considerably more information-rich than Nanog. This Klf4 probe also has a Pearson correlation to Nanog of 0.646 (3 dp.), for a multiplicative score of 5.527 (3 dp.). This score gives this Klf4 probe an extremely high rank, when considering that there are 45,101 probes. This gene fulfils all of the ideal requirements for a guide gene as set out in 4.1.3.

The distribution of correlations to Nanog and normalised entropies was investigated and is displayed in a scatterplot as the top panel of figure 4.11. Black crosshairs in this figure depict the location of the Klf4 probe chosen as the ordering gene. As can be seen from this figure, this Klf4 probe is near the maximum correlation to Nanog and also has one of the highest of all entropies in the HPM matrix. The distribution of multiplicative entropy and correlation scores themselves are presented as a histogram in figure 4.11, with the score of the chosen Klf4 probe evidenced by the vertical black line.

In addition to information content, Klf4 retains a large difference between its minimum and maximum expression, between a minimum log2 RMA value of 5.005 (3 dp.) to a maximum of 13.344 (3 dp.), a fold change across the data of 323.781 (3 dp.) absolute (non-log2) fold difference between the minimum and maximum values. This large range of expression, combined with high information content (entropy), implies that Klf4 has a desirable “progression” of expression values. This, therefore suggested this Klf4 probe to be a potentially highly-sensitive marker of cellular state between naïve pluripotency and early differentiation, subject marker gene / annotation confirmation (performed in section 4.3.4).

Finally, much like was done with the OSN factors, verification of the relationship to pluripotency / development pathways was carried out for this Klf4 probe, through the same pathway enrichment

Probe ID	Gene Name	Normalised Entropy	Absolute correlation to Nanog	Multiplicative score
1429388_at	Nanog	0.64568093	1	0.64568093
1431416_a_at	Jam2	0.8324310469	0.6851375067	0.570329732
1422458_at	Tcl1	0.8508280959	0.6565661202	0.5586249019
1449408_at	Jam2	0.8328247603	0.6655813748	0.554312649
1444390_at	Prdm14	0.8373711493	0.6539618911	0.5476088203
1429366_at	Lrrc34	0.8259522731	0.6575799162	0.5431296265
1460471_at	Ooep	0.7977445058	0.6633949245	0.5292196562
1418091_at	Tcfcp2l1	0.8040440933	0.6558461683	0.5273292377
1417395_at	Klf4	0.8155173858	0.6459001784	0.5267428249
1456329_at	Prtg	0.8421412726	0.6239222722	0.5254306963
1455300_at	Tet2	0.8038035484	0.6522767651	0.5243023783
1426858_at	ɔ /// LOC10004	0.8426430779	0.6214739275	0.5236807031
1456521_at	---	0.8126907923	0.6432830201	0.5227901873
1436926_at	Esrrb	0.8071064897	0.6442316482	0.5199635441
1424719_a_at	Mapt	0.789057297	0.652032107	0.5144906919
1438410_at	Prtg	0.8416658286	0.6105755654	0.5139005892
1449434_at	Car3	0.843895808	0.6086631895	0.5136483141
1436568_at	Jam2	0.8586770998	0.5964507084	0.5121585644
1435374_at	Cdyl2	0.7944764435	0.6427694777	0.5106652086
1431417_at	Jam2	0.7983050266	0.6382490939	0.5095174599
1436030_at	Cachd1	0.8040339697	0.6309788744	0.5073284492
1448949_at	Car4	0.8671082343	0.5842002473	0.5065648449
1456326_at	:3c1 /// LOC67(	0.8438414497	0.5969705016	0.5037484535
1418533_s_at	Fzd2	0.7832908466	0.6400561185	0.5013500989
1453628_s_at	Lrrc2	0.8003404435	0.6255568764	0.5006584679
1429377_at	2410004A20Rik	0.800821543	0.6237676516	0.4995265732
1417394_at	Klf4	0.8153679382	0.6119460916	0.498961223
1418094_s_at	Car4	0.8389928857	0.5928739524	0.4974170281
1434917_at	Cobl	0.791536874	0.6280735598	0.4971433822

Table 4.10: Normalised Shannon entropies and correlations-to-Nanog of top scoring genes in multiplicative score test. Probes for Klf4 are highlighted in green, and the higher scoring Klf4 probe was the one chosen for ordering of the HPM matrix.

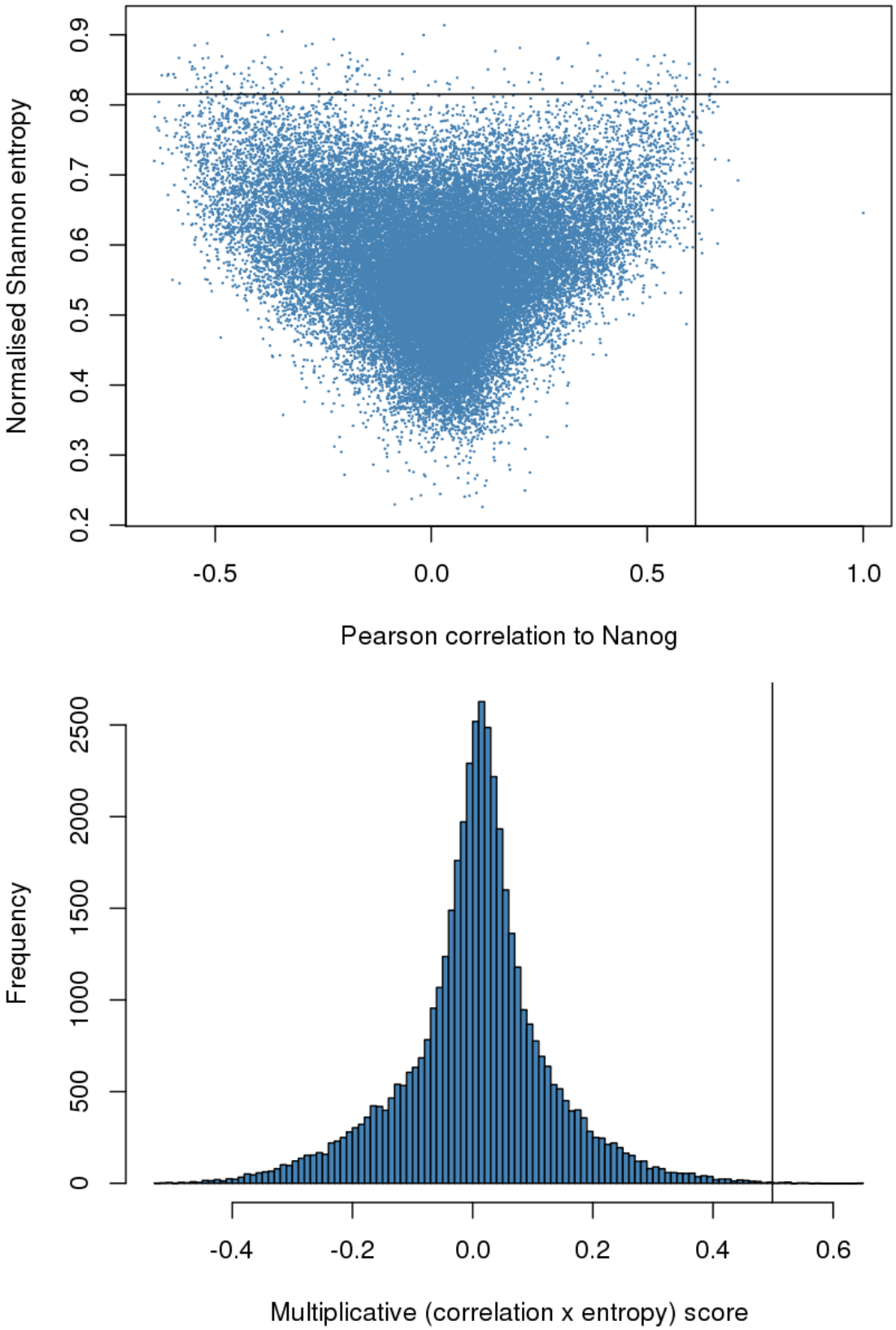

Figure 4.11: Relationship between correlations-to-Nanog and normalised Shannon entropies of all probes in the HPM matrix (upper plot). Distribution of all multiplicative scores (correlation-to-Nanog x normalised Shannon entropy), with vertical black line indicating the location of the score for the chosen Klf4 probe for the ordering of the HPM matrix (1417395_at), with a multiplicative score of 0.5627 (4 d.p.)

analysis using DAVID for those genes highly correlated (>= 0.5 Pearson correlation) and anticorrelated (<= -0.5 Pearson correlation) to this Klf4 probe.

Probes strongly positively correlated (>= 0.5) to Klf4 numbered n = 918. A truncated list of the top enrichments is given in table 4.12. The enrichments found here include the three “stem cell”-related pathways previously seen (stem cell differentiation, maintenance and development). Other pathways given in this figure show Klf4’s transcriptional regulatory properties. This is reassuring in that it can be expected that ordering by Klf4 gene would reveal much information about the transcriptional events involved in mESC pluripotency and exit from it.

Those genes strongly anticorrelated (<= -0.5 Pearson correlation) to Klf4 numbered n = 1045. For space reasons, only selected pathways discussed in the text, along with some generic developmental pathways and some examples of the first pathways which fail the significance threshold are given in table 4.13. The Wnt signalling pathway is at the top of this list, and other pathways enriched here, much like with Sox2 (see section 4.3.2), include a great many developmental pathways.

In summary, Klf4 is, in this data, primarily linked strongly to the Wnt signalling pathway by pathway enrichment analysis, as well as cellular proliferation, a large role in transcriptional regulation and also a linked to developmental pathways, as was hoped. The differences in pathway enrichments for the factors Nanog, Oct4, Sox2 and Klf4 supports the idea that the methods being employed in this work do not simply find the same enrichments for any gene correlated with pluripotency. This, in turn, reinforces the notion that this data matrix contains useful information on pluripotency and exit from it. See the following section (4.3.4) for the further marker profile and annotation cross-referencing confirmation of the utility of the HPM matrix.

GO Identifier	GO Term	Probe Count	Q-Value
GO:0010605	negative regulation of macromolecule metabolic pro	42	0.0003036769
GO:0006350	transcription	97	0.0003148161
GO:0045934	negative regulation of nucleobase, nucleoside, nucl	34	0.0003769537
GO:0042127	regulation of cell proliferation	42	0.0003853392
GO:0031327	negative regulation of cellular biosynthetic process	36	0.0003943944
GO:0010558	negative regulation of macromolecule biosynthetic p	36	0.0004035138
GO:0009890	negative regulation of biosynthetic process	36	0.0004107856
GO:0051172	negative regulation of nitrogen compound metabolic	34	0.0004158772
GO:0045449	regulation of transcription	116	0.0005410792
GO:0045892	negative regulation of transcription, DNA-dependent	28	0.0010445504
GO:0051253	negative regulation of RNA metabolic process	28	0.0010741086
GO:0016481	negative regulation of transcription	31	0.0012874656
GO:0010629	negative regulation of gene expression	33	0.0013647003
GO:0010604	positive regulation of macromolecule metabolic proc	43	0.0026551132
GO:0006357	regulation of transcription from RNA polymerase II p	41	0.0063920651
GO:0048863	stem cell differentiation	8	0.0078314242
GO:0019827	stem cell maintenance	7	0.0080669518
GO:0048864	stem cell development	7	0.0094849968
GO:0051173	positive regulation of nitrogen compound metabolic	36	0.0099834391
GO:0031328	positive regulation of cellular biosynthetic process	37	0.01065359
GO:0009891	positive regulation of biosynthetic process	37	0.0121559661
GO:0045596	negative regulation of cell differentiation	18	0.0125879407
GO:0051252	regulation of RNA metabolic process	76	0.0158188834
GO:0008284	positive regulation of cell proliferation	23	0.0185360787
GO:0010557	positive regulation of macromolecule biosynthetic p	35	0.0187811475
GO:0045935	positive regulation of nucleobase, nucleoside, nucle	34	0.0191335533
GO:0019953	sexual reproduction	28	0.0194699867
GO:0006355	regulation of transcription, DNA-dependent	74	0.0227974051
GO:0010628	positive regulation of gene expression	32	0.0344479432
GO:0045941	positive regulation of transcription	31	0.0450794752
GO:0048232	male gamete generation	20	0.0634587725
GO:0007283	spermatogenesis	20	0.0634587725
GO:0051276	chromosome organization	27	0.0718824783

Table 4.12: Selected biological pathway enrichments resulting from the gene list of strong positive (Pearson correlation > 0.5) Klf4 correlations in the HPM matrix.

GO Identifier	GO Term	Probe Count	Q-Value
GO:0016055	Wnt receptor signaling pathway	21	9.988553E-05
GO:0007389	pattern specification process	30	0.0003907111
GO:0048729	tissue morphogenesis	26	0.0007079068
GO:0007167	enzyme linked receptor protein signaling pathway	27	0.0014958646
GO:0048598	embryonic morphogenesis	32	0.0014013296
GO:0000904	cell morphogenesis involved in differentiation	23	0.0016245322
GO:0003002	regionalization	23	0.0016354609
GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway	20	0.0066669491
GO:0043434	response to peptide hormone stimulus	13	0.0120239916
GO:0042127	regulation of cell proliferation	37	0.0190228592
GO:0030879	mammary gland development	11	0.0579037154
GO:0021511	spinal cord patterning	5	0.065308208
GO:0009952	anterior/posterior pattern formation	15	0.0695084286

Table 4.13: Selected biological pathway enrichments resulting from the gene list of strong negative (Pearson correlation <= -0.5) Klf4 correlations.

4.3.4 Ordering the HPM matrix by Klf4 expression broadly sorts samples from naïve pluripotency toward early differentiation

Now ordered by decreasing Klf4, patterns of expression of 3 markers (FGF5, Rex1, Brachyury (T)) were observed across the Klf4-ordered matrix and compared to the patterning of these same crucial markers when ordered by GAPDH (see figure 4.14). Ordering the matrix by a the housekeeping gene GAPDH shows no patterning or progression from one state to another of any of the three chosen markers. FGF5 and Brachyury (T) increase as naïve pluripotency progresses toward primed pluripotency, while Rex1 decreases across this same phenomenon (Ying et al. 2008) (Nichols and Smith 2009) (Yamaji et al. 2013). This is precisely what is seen in the rightmost panels of figure 4.14.

Figure 4.15 puts the two other concerns identified in section 4.2.3 conclusively to rest. Firstly, Klf4 ordering is not simply a surrogate for Nanog ordering, even though Klf4 was partly chosen for its correlation to Nanog. This can be seen from the top three plots in 4.15, where Nanog ordering of the matrix actually results in quite woeful patterning of FGF5 and Brachyury and a much less clear line of progression of a drop in Rex1 (Zfp42), although a trend is still somewhat visible, it is nothing like as clear as when sorting by Klf4.

This second concern was that the choice of Klf4 by the fourth criterion in section 4.2.3, being to choose a gene with known pluripotency-related activity, might be the only criterion that mattered, and that were another gene with only a high multiplicative (entropy x correlation to Nanog) score chosen, no such impressive sorting of the data would have occurred. This would imply that the scoring method presented in this work was largely irrelevant.

To address this, the gene with the best multiplicative entropy and correlation-to-Nanog score was tested for its ability to generate the desired smooth progression of the FGF5, Brachyury, Rex1 markers. Jam2 (“1431416_a_at”) was this best scoring probe had a multiplicative score of 0.570 (3 d.p.), higher than Klf4's score of 0.527 (3 d.p.).

This probe’s ability to generate the desired marker progression across the data is demonstrated in the bottom-most plots of figure 4.15. If anything, Jam2 appears slightly superior to Klf4 (middle panels of figure 4.15) in its ability to sort these canonical markers between naïve pluripotency and

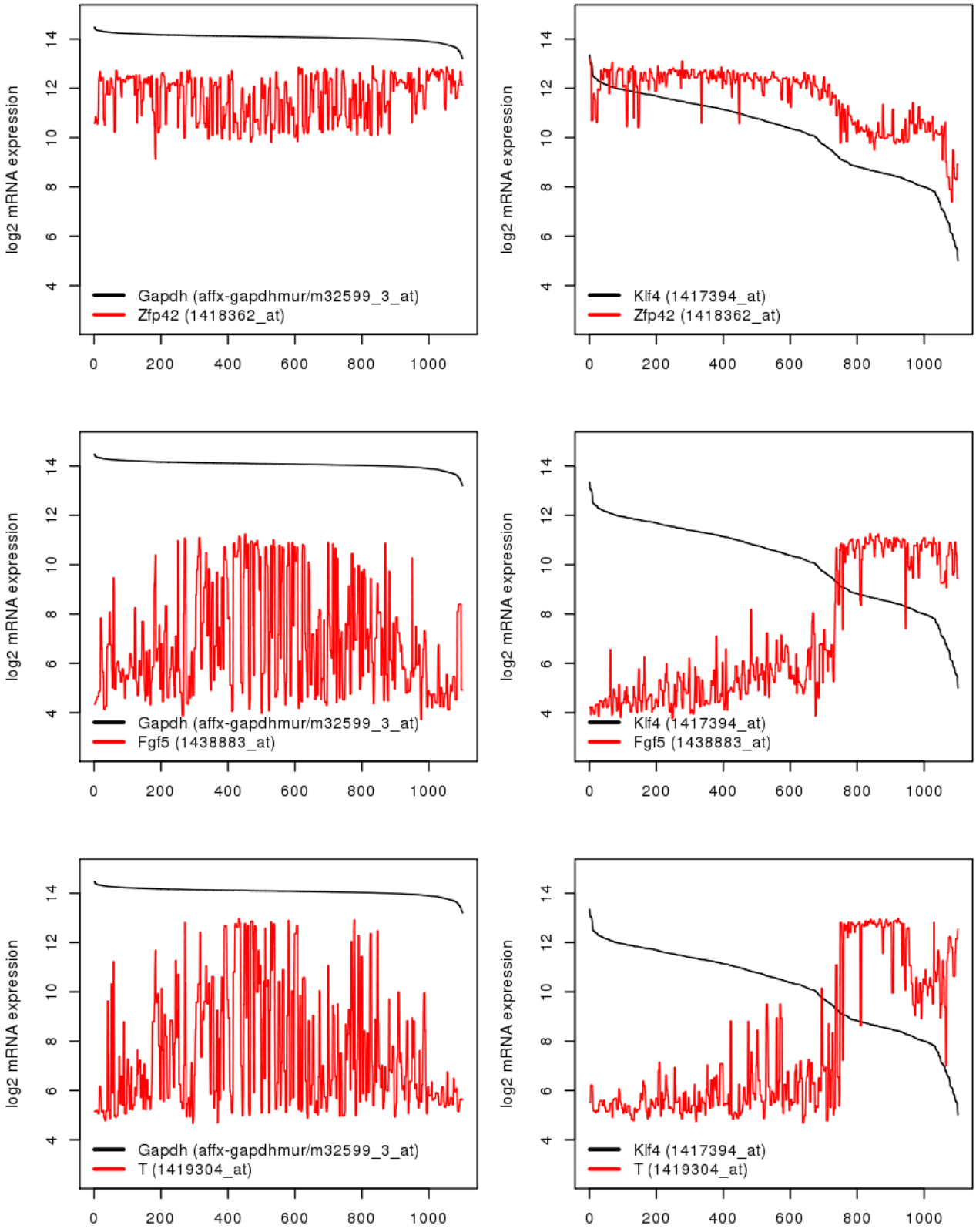

Figure 4.14: Matrix N1101 ordered by housekeeping gene GAPDH (leftmost panels) or the chosen ordering gene Klf4 (rightmost panels.) Canonical markers of the naïve versus primed pluripotent state, Rex1 (Zfp42) (top panels), FGF5 (middle panels) and Brachyury (T) (bottom panels) are also shown for both ordered matrices. Smoothed lines are used for clarity and clearly demonstrate Klf4s ability to sort the matrix between naïve to primed pluripotency, while the negative control ordering by GAPDH shows no such patterning.

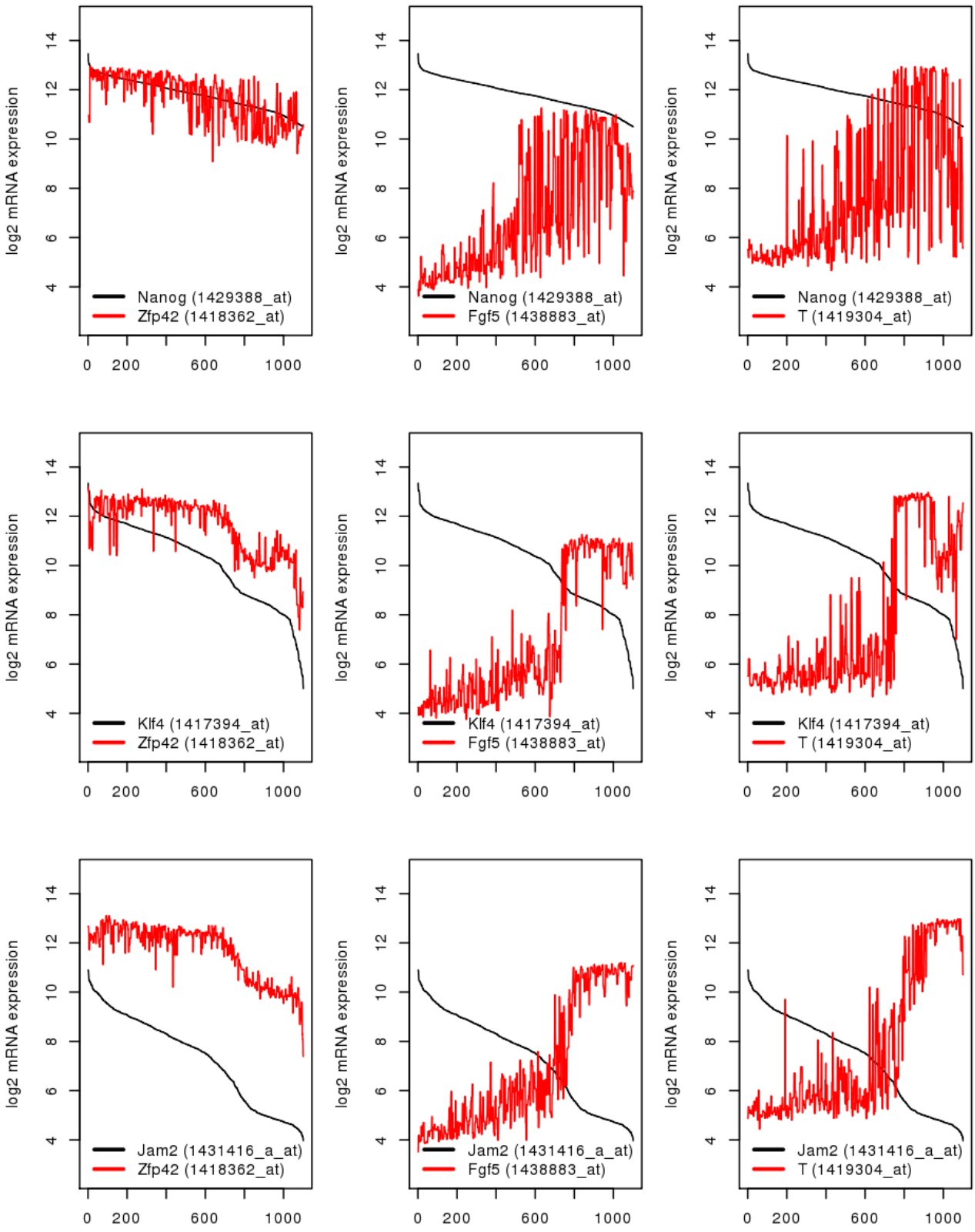

Figure 4.15: Patterning of the naïve pluripotency marker Rex1 and primed pluripotency markers Brachyury and FGF5 when all 1101 samples of N1101 are ordered by Nanog (top plots), Klf4 (middle plots) or Jam2 (bottom plots). Klf4 was plotted in the middle panels again here for easy by-eye comparison, but are identical plots to the rightmost panels of figure 4.14**.**

early differentiation. This manifests as Jam2-sorting resulting in a smoother line for FGF5 towards the rightmost side of the plot, and this is more pronounced when Brachyury (T) is ordered by Jam2.

It was decided to remain with the choice of Klf4 as the ordering gene as this is in keeping with the prescribed fourth criterion as laid down in section 4.2.3. That said, a novel finding from this work is that Jam2 is suggested by this methodology to be a highly-sensitive marker for where a particular high OSN sample may lie between naïve pluripotency and early differentiation.

In addition to verifying the utility of ordering by Klf4 using naïve vs primed pluripotency markers, the annotations associated with samples at different parts of the Klf4 order were investigated. Klf4-order ranks of all samples mentioned in this section and their accompanying annotations are all summarised in tables 4.16 and 4.17 while all annotations for all samples are given on the accompanying DVD in "Chapter 2/N1101 Annotations". Comment is now given on randomly-selected samples from a spread of regions across the decreasing-Klf4 gradient.

At the highest levels of Klf4, represented at the leftmost side of 4.18, samples such as GSM381301 occur (as 4th highest for Klf4), from experiment GSE15267, from work by (Chen et al. 2010), an investigation of culture conditions for the generation of mouse iPSCs. Sample GSM381301 itself is a control culture of CGR8 cells. Another sample at the leftmost side of 4.18 is GSM277757 (second highest rank for Klf4), which is part of experiment GSE10970. This experiment concerns mESC differentiation to cardiac lineages (Miller et al. 2008). Sample GSM277757 itself is a day 0 control. It should be noted that these samples are genetically modified with a GFP reporter for Nkx2-5. Other samples from GSE10970 are present in the HPM matrix, such as at ranks 6 and 9 with GSM277758 and GSM277759 respectively.

As is visible in figure 4.18 there is a noticeable spike on the far left where Klf4 is at its highest. The samples mentioned above are from this very high level of Klf4. In case this spike was a group of outlying samples, a random group of samples was chosen from near the 100th rank mark, where the Klf4 line first settles into a lower-gradient decline (see figure 4.18). Sample GSM241847 is at rank 83 for Klf4 expression and is annotated as being "ESC, Undifferentiated, Pool-1, Biological Rep-1" from experiment GSE9563 by (Sampath et al. 2008) whose work specifically concerned mESC differentiation and had 12 undifferentiated samples and 12 embryoid body samples. It would be expected that any of the undifferentiated samples samples from this experiment that occur in the HPM matrix should occur toward the left side of the plot in 4.18 and those that are EBs would

group together, but further to the right where markers Rex1, Fgf5, and Brachyury have changed to low, high and high expression respectively. This is indeed the case, as can be seen from tables 4.16 and 4.17 and figure 4.18.

It is also important to note that samples from the same experiment do not appear so close to each other in Klf4 ranking that they are essentially consecutive, which would have suggested that those samples are grouping simply due to being from the same lab / experiment. Samples from experiment GSE9563 do group by their progress toward differentiation and clear separation is seen, with naïve pluripotent samples between ranks 33 to 183 and primed pluripotent samples between ranks 550 to 813, interspersed with samples from other experiments (see tables 4.16 and 4.17.) This puts to rest the concern that perhaps samples grouped by experiment / lab across the Klf4-ordered matrix.

Annotations were also drawn from the HPM matrix to verify the middle-to-lower end of the Klf4 spectrum. Here, samples such as GSM597445, GSM597446 and GSM597448 occur, which are part of experiment GSE24289. This experiment, carried out by (Gill et al. 2011), was an investigation into the induction of differentiation in mESCs through the forced expression of miR-200c/miR-141, driven by the addition of tamoxifen. Sample GSM597446 involved 3 days of induction with tamoxifen and GSM597448 was induced for 4 days. Sample GSM597445 however, whilst annotated as an untreated control sample, was actually differentiated for 3 days. Furthermore, the untreated control sample had a lower level of Klf4 expression than the two tamoxifen-treated samples. This is not unexpected, however, as all of these samples are the results of culture of mESCs as embryoid bodies (Gill et al. 2011) and would therefore be expected to exhibit a transcriptional state close to early differentiation. This underscores the need for detailed annotations, without which an “untreated control” sample may have seemed out of place, rather than supportive of Klf4-ordering.

Another experiment, GSE12982, carried out by (Shen et al. 2008), investigated the effects of knocking out both Ezh2, the catalytic subunit of polycomb-repressive-complex 2 (Prc2), and Eed (embryonic ectoderm development). This experiment has 53 total microarrays uploaded and 28 of these appear in the HPM matrix. As can be seen from the spread of GSE12982 samples across the tables in tables 4.16 and 4.17 and figure 4.18, strong support is given that Klf4-ordering is not only broadly sorting the samples appropriately between naïve pluripotency and primed pluripotency /

Experiment accession	Sample accession	Description	Klf4 Rank
GSE10970	GSM277757	Day 0 control pluripotent cells (Nkx2-5-GFP)	2
GSE15267	GSM381301	Pluripotent CGR8 cells	4
GSE10970	GSM277758	Day 0 control pluripotent cells (Nkx2-5-GFP)	6
GSE10970	GSM277759	Day 0 control pluripotent cells (Nkx2-5-GFP)	9
GSE9563	GSM241853	ESC, Undifferentiated, Pool-3, Biological Rep-1	33
GSE9563	GSM241856	ESC, Undifferentiated, Pool-4, Biological Rep-1	36
GSE9563	GSM241848	ESC, Undifferentiated, Pool-1, Biological Rep-2	49
GSE9563	GSM241855	ESC, Undifferentiated, Pool-3, Biological Rep-3	63
GSE9563	GSM241858	ESC, Undifferentiated, Pool-4, Biological Rep-3	64
GSE9563	GSM241872	Total_RNA_ESC, Undifferentiated, Biological Rep-2	75
GSE9563	GSM241847	ESC, Undifferentiated, Pool-1, Biological Rep-1	83
GSE9563	GSM241857	ESC, Undifferentiated, Pool-4, Biological Rep-2	86
GSE9563	GSM241854	ESC, Undifferentiated, Pool-3, Biological Rep-2	89
GSE9563	GSM241871	Total_RNA_ESC, Undifferentiated, Biological Rep-1	97
GSE9563	GSM241873	Total_RNA_ESC, Undifferentiated, Biological Rep-3	108
GSE9563	GSM241849	ESC, Undifferentiated, Pool-1, Biological Rep-3	117
GSE9563	GSM241851	ESC, Undifferentiated, Pool-2, Biological Rep-2	120
GSE12982	GSM325396	E14tg1 wild-type ES cells at day 0 (undifferentiated), biological rep 7	141
GSE12982	GSM325407	Eed-null ES cells at day 0 (undifferentiated), biological rep 5	163
GSE9563	GSM241850	ESC, Undifferentiated, Pool-2, Biological Rep-1	165
GSE9563	GSM241852	ESC, Undifferentiated, Pool-2, Biological Rep-3	183
GSE12982	GSM325405	Eed-null ES cells at day 0 (undifferentiated), biological rep 3	196
GSE12982	GSM325406	Eed-null ES cells at day 0 (undifferentiated), biological rep 4	268
GSE12982	GSM325399	Ezh2-null ES cells at day 0 (undifferentiated), biological rep 3	281
GSE12982	GSM325398	Ezh2-null ES cells at day 0 (undifferentiated), biological rep 2	299
GSE12982	GSM325404	Eed-null ES cells at day 0 (undifferentiated), biological rep 2	328
GSE12982	GSM325403	Eed-null ES cells at day 0 (undifferentiated), biological rep 1	354
GSE12982	GSM325395	E14tg1 wild-type ES cells at day 0 (undifferentiated), biological rep 6	403
GSE12982	GSM325394	E14tg1 wild-type ES cells at day 0 (undifferentiated), biological rep 5	405
GSE12982	GSM325397	Ezh2-null ES cells at day 0 (undifferentiated), biological rep 1	409
GSE12982	GSM325392	CJ7 wild-type ES cells at day 0 (undifferentiated), biological rep 3	466
GSE12982	GSM325393	CJ7 wild-type ES cells at day 0 (undifferentiated), biological rep 4	485
GSE12982	GSM325402	Ezh2-null ES cells at day 0 (undifferentiated), biological rep 6	490

Table 4.16: Part 1 of 2 of table detailing samples mentioned in the main text (section 4.3.4). Pluripotent, undifferentiated samples are in green highlight, early differentiation samples in yellow highlight, progressing to orange for those differentiated for 2 days, red for 3 days and dark red for later than 3 days. All sorted by decreasing Klf4.

Experiment accession	Sample accession	Description	Klf4 Rank
GSE12982	GSM325401	Ezh2-null ES cells at day 0 (undifferentiated), biological rep 5	495
GSE12982	GSM325400	Ezh2-null ES cells at day 0 (undifferentiated), biological rep 4	513
GSE9563	GSM241875	Total_RNA_EB, Biological Rep-2	550
GSE12982	GSM325390	J1 wild-type ES cells at day 0 (undifferentiated), biological rep 1	563
GSE9563	GSM241874	Total_RNA_EB, Biological Rep-1	577
GSE12982	GSM325391	J1 wild-type ES cells at day 0 (undifferentiated), biological rep 2	587
GSE9563	GSM241876	Total_RNA_EB, Biological Rep-3	611
GSE9563	GSM241867	EB, Pool-3, Biological Rep-3	668
GSE9563	GSM241865	EB, Pool-3, Biological Rep-1	670
GSE9563	GSM241868	EB, Pool-4, Biological Rep-1	673
GSE9563	GSM241866	EB, Pool-3, Biological Rep-2	681
GSE12982	GSM325429	Ezh2-null ES cells at day 8 of LIF withdrawal, biological rep 2	686
GSE9563	GSM241862	EB, Pool-2, Biological Rep-1	692
GSE9563	GSM241870	EB, Pool-4, Biological Rep-3	701
GSE9563	GSM241863	EB, Pool-2, Biological Rep-2	709
GSE9563	GSM241859	EB, Pool-1, Biological Rep-1	734
GSE12982	GSM325428	Ezh2-null ES cells at day 8 of LIF withdrawal, biological rep 1	743
GSE9563	GSM241860	EB, Pool-1, Biological Rep-2	745
GSE9563	GSM241864	EB, Pool-2, Biological Rep-3	810
GSE9563	GSM241861	EB, Pool-1, Biological Rep-3	813
GSE12982	GSM325422	Ezh2-null ES cells at day 2 of LIF withdrawal, biological rep 4	959
GSE12982	GSM325420	Ezh2-null ES cells at day 2 of LIF withdrawal, biological rep 2	974
GSE12982	GSM325411	CJ7 wild-type ES cells at day 2 of LIF withdrawal, biological rep 4	1025
GSE12982	GSM325419	Ezh2-null ES cells at day 2 of LIF withdrawal, biological rep 1	1039
GSE12982	GSM325410	CJ7 wild-type ES cells at day 2 of LIF withdrawal, biological rep 3	1042
GSE12982	GSM325421	Ezh2-null ES cells at day 2 of LIF withdrawal, biological rep 3	1043
GSE12982	GSM325430	Eed-null ES cells at day 2 of LIF withdrawal, biological rep 1	1053
GSE24289	GSM597448	EB, 3 day differentiated	1054
GSE24289	GSM597445	EB, 3 day differentiated	1058
GSE24289	GSM597446	EB, 3 day differentiated	1059
GSE12982	GSM325431	Eed-null ES cells at day 2 of LIF withdrawal, biological rep 2	1062

Table 4.17: Part 2 of 2 of table detailing samples mentioned in the main text (section 4.3.4). Pluripotent, undifferentiated samples are in green highlight, early differentiation samples in yellow highlight, progressing to orange for those differentiated for 2 days, red for 3 days and dark red for later than 3 days. All sorted by decreasing Klf4.

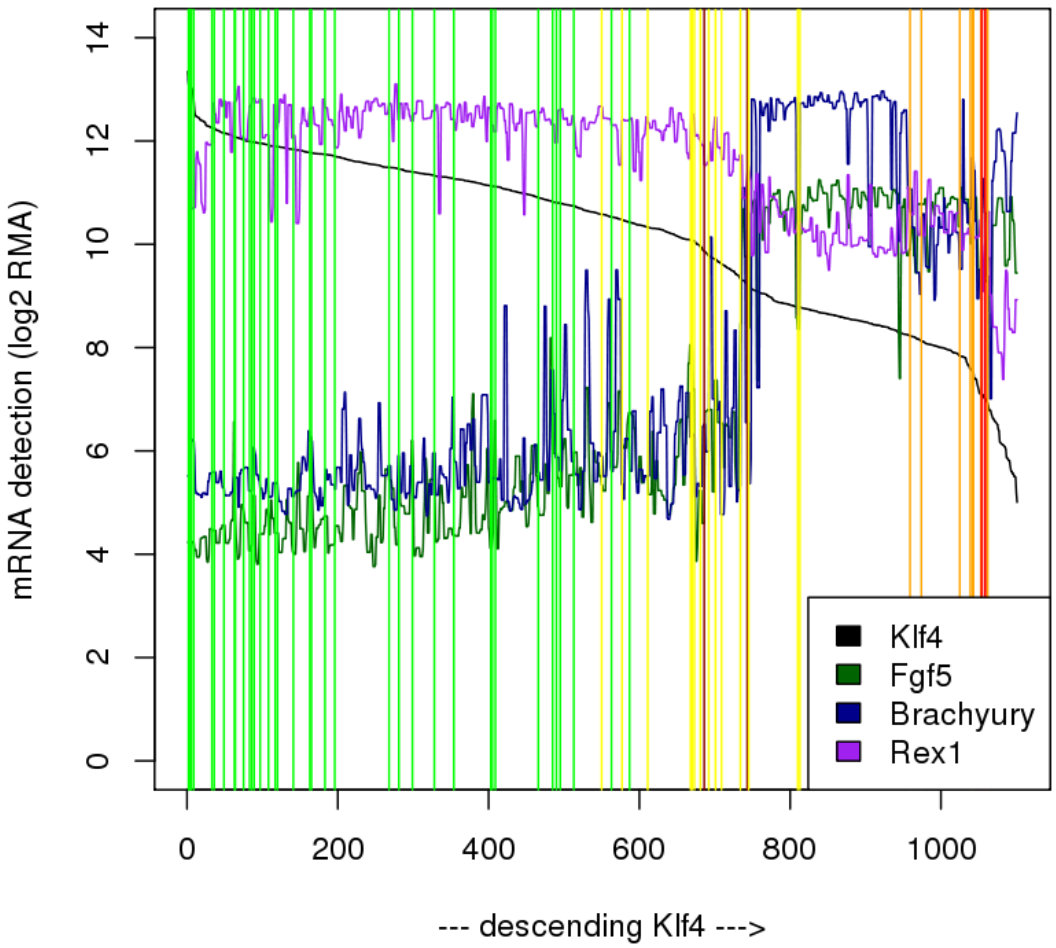

Figure 4.18: Combined plot showing and overview of the HPM matrix, as ordered by Klf4 (black line), with smoothed lines shown for naïve pluripotency marker Rex1 (light purple line), FGF5 (dark green line), Brachyury (dark blue line). This overview of progression from naïve pluripotency to exit from it is overlaid with the locations (by Klf4 rank) and cellular states (see figures 4.16 and 4.17), with undifferentiated annotations represented with green lines, early differentiation / embryoid body samples in yellow, and differentiating samples thereafter in orange (2 day differentiated), red (3 day differentiated) and aubergine colour (8 days of differentiation.) Note that whilst there is excellent agreement as a general trend, as seen by the intact progression of green → yellow → orange → red, there are two samples (GSM325429 and GSM325428, dark red vertical lines) which are found earlier in the Klf4 ranking than would be expected, showing that the ordering by Klf4 is, as predicted, not perfect.

beyond, but also strongly refutes the concern that perhaps samples may group by experiment in this sorted dataset.

The sorting was not expected to be perfect, and in table 4.17, two samples, GSM325429 and GSM325428 occur at a lower Klf4 rank than would have been ideal. A perfect ordering is not to be expected nor is it required, as the ability to broadly sort samples in this manner is still likely to allow for the analysis presented in this chapter to draw inferences from the trends visible across the data as regards the activities of known signalling pathways. Furthermore, as this chapter's later objectives focus on the finding of genes related to those signalling pathways found to be of interest across this ordered data, these can be found by observing any apparent relationship between genes identified in known pathways as being "of interest" and potential candidate genes. This would remain mostly unaffected by any samples not being in a "perfect order" of pluripotency/pre-differentiation status.

In summary, tables 4.16 and 4.17 and figure 4.18 show that ordering by Klf4 broadly sorts the samples between naïve pluripotency (lower Klf4 rank) through primed pluripotency to early differentiation (high Klf4 ranks). It must be stressed, however, even whilst mentioning "early differentiation" that all of these samples still have the highest levels of Oct4, Sox2 and Nanog, and so it is reasonable to assert that they all likely remain pluripotent.

4.3.5 Gross differential expression analysis of the highest-Klf4 state versus the lowest-Klf4 state shows enrichment for generalised differentiation

In this analysis, the selected highest 50-Klf4 and lowest 50 Klf4 samples were used (see the green and red vertical sections of figure 4.19 respectively for a representation of where on the Klf4 spectrum these samples were chosen from). Differential expression analysis was carried out as detailed in 4.2.4.

Brief verification of the behaviour of this approach was undertaken by observing the relationship between the p-values and fold changes generated. This relationship is shown in the top panel of figure 4.20. The bottom panel zooms in on the plot in the upper panel, close to where the observed fold change is close to zero. Here it can be seen that whilst there is a strong tendency for probes with a very low fold-change to be shown to be insignificant (only visible in the zoomed, lower

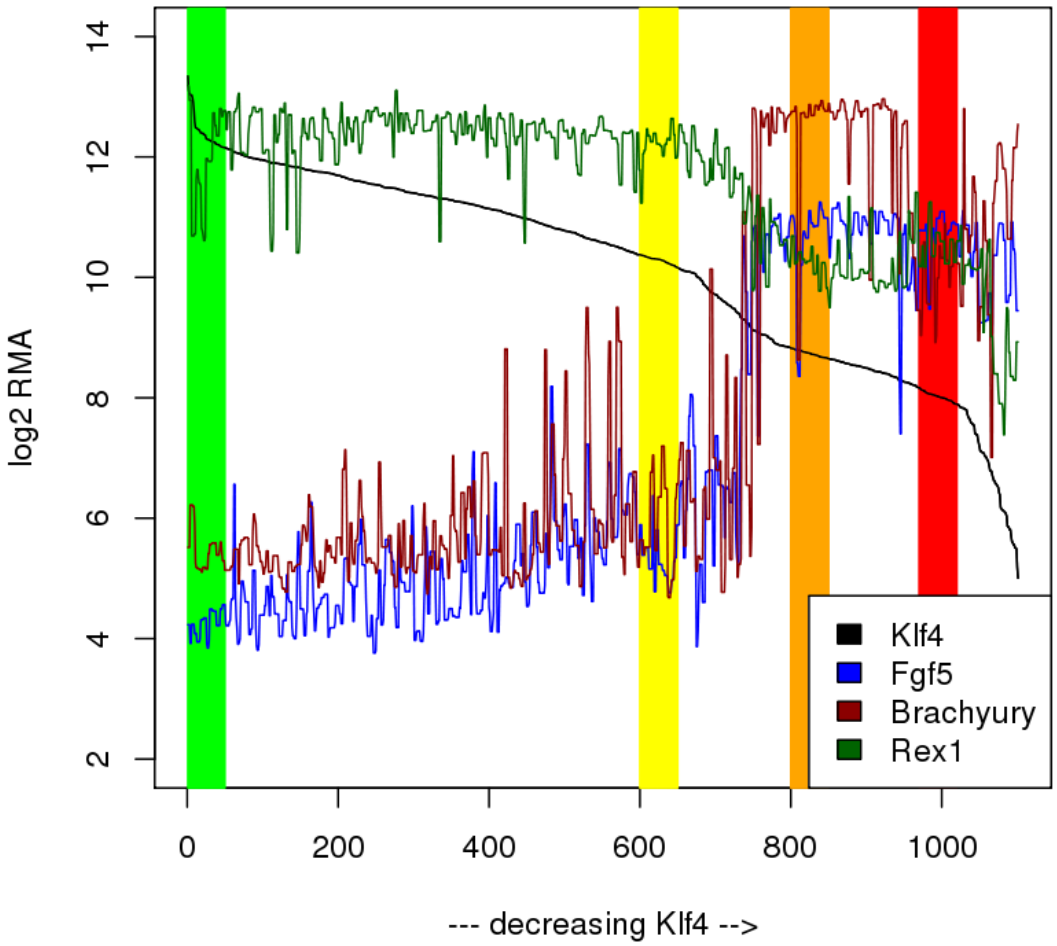

Figure 4.19: Markers of the change between naïve and primed pluripotency Fgf5, Brachyury and Rex1 plotted as smoothed lines across the Klf4-ordered HPM matrix. Vertical highlights demonstrate areas of the matrix which were selected to represent early naïve pluripotency (green), late naïve pluripotency (yellow), primed pluripotency (orange) and early differentiation (red).

panel of figure 4.20), there are yet some probes found to have a small fold change, but for that fold change to be found to be statistically significant. This confirms the sensitivity of the method, but that the general trend is as expected, that probes with very small fold-changes tend not to be called as significant. Conversely, those genes with large fold changes do not get called as insignificant. The behaviour of the approach confirmed, this type of plot will not be repeated for the other differential expression analyses, as mentioned in section 4.2.4.

Of all 45,101 probes, 1,141 probes were found to have significantly ($p < 0.05$) increased in their expression in the low-Klf4 group by at least 1 log2 fold (2 absolute fold), compared to the high-Klf4 group. 1,108 of these were found in DAVID's mouse database. Conversely, there were 870 probes found to have a fold change of -1 or larger for the investigation of genes whose expression drops between highest Klf4 and lowest Klf4. The pathways enrichments resulting from the list of upregulated genes contains many strong enrichments for developmental processes and also the crucial Wnt signalling pathway (see section 1.3.3). Selected, mESC pluripotency / signalling-relevant pathways resulting from significantly upregulated genes in this analysis are given in table 4.21, and similar pathways resulting from significantly downregulated genes given in table 4.22. With the confirmation of the utility of the Klf4-ordered matrix in section 4.2.3, results such as these become unsurprising and predictable, and so will not be discussed at length.

In the following analyses across the Klf4-ordered matrix, both using differential expression and the scanning window approach described in section 4.2.5, similar pathway enrichment tables were generated, but are not reproduced exhaustively in this thesis, nor is exhaustive description of the observed pathways given in the main text, to avoid expounding repetitive, predictable results. In place of exhaustive listing / discussion of pathway enrichments, the final figures of this chapter's results section (figures 4.34 to 4.36) neatly summarise significantly up/downregulated gene counts and the resulting categories of developmental, stem cell signalling, stress, apoptosis and gene expression control pathways significantly enriched in each analysis carried out. Any detailed discussion in the main text is, therefore, limited to the biological interpretation of selected pathways for each analysis.

Of the three "stem cell" pathways (stem cell differentiation, development and maintenance) that often appear in DAVID outputs of pluripotency factors (see section 4.3.2), none of them achieved significance here. This is mentioned as being of particular interest as it an objective of this work to

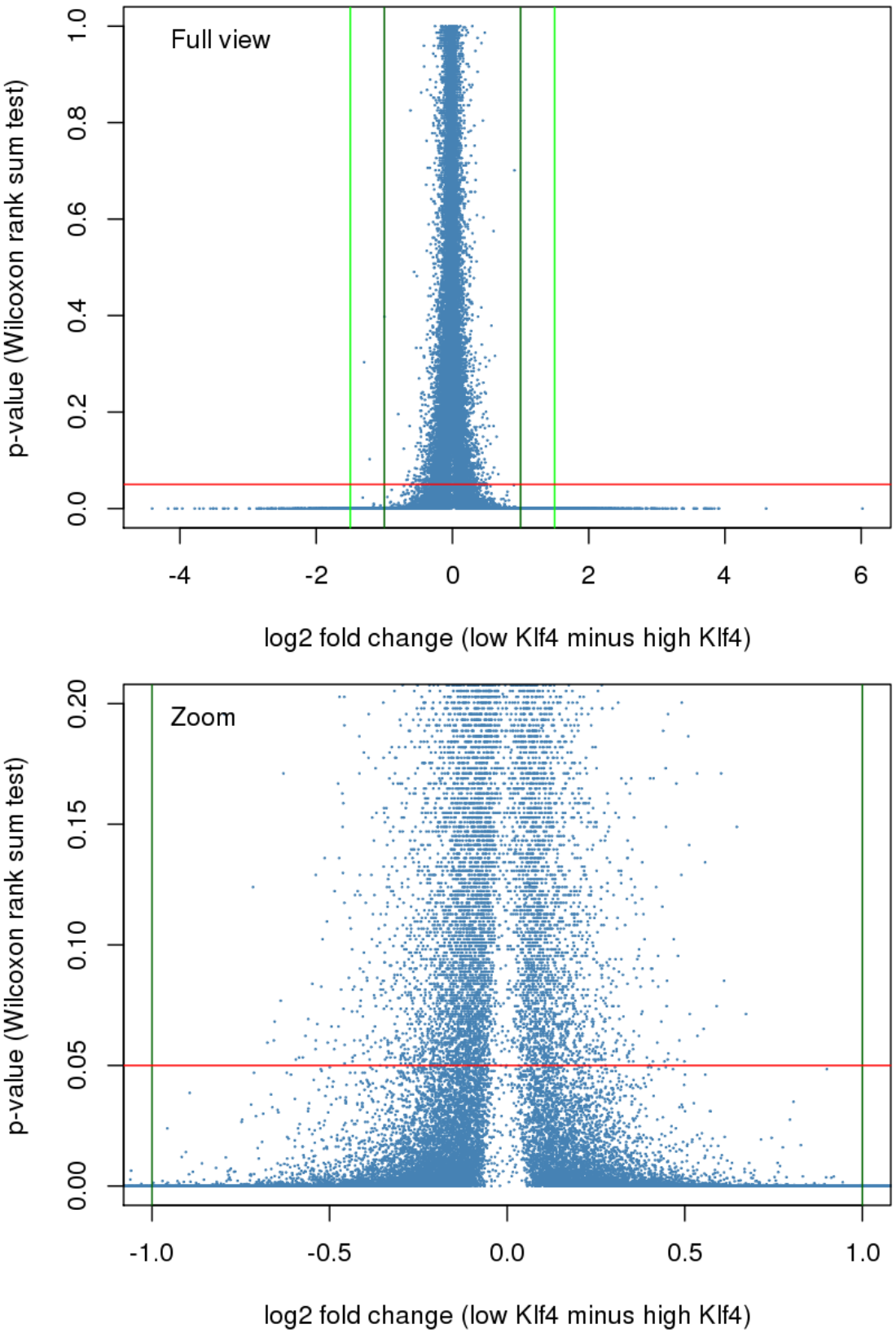

Figure 4.20: Scatterplots depicting the relationship between significance (Mann-Whitney U test) and observed fold change between the 50 highest-Klf4 samples and the 50 selected low-Klf4 samples in the Klf4-ordered HPM matrix. Red lines show the p = 0.05 mark, dark green lines show the log2 fold change = 1 mark, light green lines show the log2 fold change = 1.5 mark. Top panel is all datapoints, bottom panel is a zoomed version.

GO Identifier	GO Term	Probe Count	Q-value
GO:0007389	pattern specification process	36	0.00000682
GO:0048729	tissue morphogenesis	33	7.24994103484811E-006
GO:0000904	cell morphogenesis involved in differentiation	28	0.0001291903
GO:0016055	Wnt receptor signaling pathway	21	0.0001496236
GO:0031175	neuron projection development	26	0.0007089474
GO:0035239	tube morphogenesis	22	0.0012429839
GO:0060429	epithelium development	29	0.0013163323
GO:0030182	neuron differentiation	37	0.0013235519
GO:0007409	axonogenesis	21	0.0015842651
GO:0045449	regulation of transcription	131	0.0016816584
GO:0048667	cell morphogenesis involved in neuron differentiatio	22	0.0023617706
GO:0030111	regulation of Wnt receptor signaling pathway	10	0.0024675237
GO:0035295	tube development	27	0.0033344299
GO:0021915	neural tube development	14	0.0033427498
GO:0006350	transcription	107	0.0034410031
GO:0048812	neuron projection morphogenesis	21	0.0034720254
GO:0007411	axon guidance	15	0.0040551067
GO:0007507	heart development	24	0.0040689781
GO:0007167	enzyme linked receptor protein signaling pathway	27	0.0046893564
GO:0048666	neuron development	28	0.0054659545
GO:0051094	positive regulation of developmental process	22	0.0118156897
GO:0045941	positive regulation of transcription	37	0.0160137907
GO:0030178	negative regulation of Wnt receptor signaling pathw	7	0.0251371982
GO:0042127	regulation of cell proliferation	39	0.0356556346
GO:0045944	positive regulation of transcription from RNA polyme	29	0.0364865899
GO:0006357	regulation of transcription from RNA polymerase II	43	0.0371580925
GO:0043405	regulation of MAP kinase activity	11	0.0684411563
GO:0000165	MAPKKK cascade	12	0.1538194186
GO:0030509	BMP signaling pathway	5	0.2611128572
GO:0043408	regulation of MAPKKK cascade	9	0.4285691743
GO:0000187	activation of MAPK activity	6	0.5189439491
GO:0043409	negative regulation of MAPKKK cascade	3	0.6726345067

Table 4.21: Selected biological pathway enrichments resulting from a gene list made of those genes found to be upregulated (increase of 1 log2 fold expression) between 50 highest vs 50 lowest Klf4 samples.

GO Identifier	GO Term	Probe Count	Q-value
GO:0042127	regulation of cell proliferation	51	1.52228E-008
GO:0006357	regulation of transcription from RNA polymerase II prom	45	0.0002874579
GO:0010629	negative regulation of gene expression	34	0.0004960271
GO:0045596	negative regulation of cell differentiation	20	0.0009700733
GO:0010628	positive regulation of gene expression	36	0.0009870139
GO:0006355	regulation of transcription, DNA-dependent	77	0.0012405369
GO:0008285	negative regulation of cell proliferation	22	0.0012902969
GO:0045941	positive regulation of transcription	35	0.0013077032
GO:0001568	blood vessel development	23	0.0013089926
GO:0001944	vasculature development	23	0.0016453912
GO:0045892	negative regulation of transcription, DNA-dependent	26	0.0016926587
GO:0008284	positive regulation of cell proliferation	24	0.0033558614
GO:0016481	negative regulation of transcription	28	0.0047964846
GO:0045449	regulation of transcription	102	0.0059009418
GO:0045944	positive regulation of transcription from RNA polymeras	27	0.0060337084
GO:0001525	angiogenesis	15	0.0061080122
GO:0045893	positive regulation of transcription, DNA-dependent	29	0.0107934924
GO:0035295	tube development	21	0.018984997
GO:0048514	blood vessel morphogenesis	17	0.0332051926
GO:0001763	morphogenesis of a branching structure	13	0.0343960781
GO:0048863	stem cell differentiation	6	0.1172494743
GO:0019827	stem cell maintenance	5	0.1578286883
GO:0048864	stem cell development	5	0.1782723857
GO:0017015	regulation of transforming growth factor beta receptor s	5	0.318268833

Table 4.22: Selected biological pathway enrichments resulting from a gene list made of those genes found to be downregulated (decrease of 1 log2 fold expression) between 50 highest vs 50 lowest Klf4 samples.

demonstrate the increased utility of scanning over a large, ordered dataset rather than doing more traditional differential expression analyses between defined start and end points of a biological phenomenon of interest. This is borne out in later sections where the scanning window method is applied across the entire dataset in section 4.3.13.

In summary, it would appear that differential expression analysis between the highest and lowest Klf4 states finds a strong developmental signal, but does not reveal much of the internal workings (pathways or processes) known to operate in mESCs that other analyses in this work have much greater success with.

4.3.6 Gene expression changes between naïve pluripotency and primed pluripotency show a preference for gene activation of developmental and signalling pathways

This analysis was carried out as described in section 4.2.4, by calculating the fold changes of all probes between a first set of samples between Klf4 rank 1 and 50 (see figure 4.19, green vertical bar.), to a second set of samples between Klf4 ranks 800 and 849 (see figure 4.19, orange vertical bar.) Of all 45,101 probes, 2,254 were found to have significantly ($p < 0.05$) increased their expression by at least 1 log2 fold. This number of significantly changing genes is in striking contrast to the differential expression analysis between the highest and lowest Klf4 states (n =). This, again, supports the notion that more focussed methods are necessary to draw out the detailed transcriptional events occurring in the data.

From all 153 significantly-enriched pathways, this analysis returned significant enrichments for pathways concerning development, proliferation, transcription, and activity of the Wnt, VEGF, Notch and MAPK pathways. An appearance here of enrichment for “negative regulation cell death” was unexpected, but makes sense in that increased expression of anti-apoptosis markers implies a pro-survival signal.

Finally, there was enrichment here for “chromatin organization”. Chromatin remodeling is known to take place between naïve and primed pluripotency (see sections 1.4 and 4.1.7) and so it is good to see this enrichment occurring here, again lending support to the idea that events known in the literature are recapitulated here.

Only 874 probes were significantly ($p < 0.05$) downregulated by at least 1 log2 fold. Only four pathways achieved enrichment here, all concerning cell proliferation and transcription.

Therefore it appears that, compared to the "highest versus lowest Klf4" analysis, this more focussed analysis on naïve to primed pluripotency reinforced the importance of Wnt signalling, but added the aforementioned VEGF, Notch and MAPK pathways and found implied changes to chromatin organisation. This recapitulates what is known about the naïve to primed switch (see sections 1.3.3, 1.4, 1.5 and 4.1.7), and reinforces the utility of combining wider and more focussed analyses across ordered datasets.

4.3.7 Differential expression analysis of "early" to "late" naïve pluripotency suggests separate cellular states and markers defining these states

Fold changes for all probes were calculated between "early" naïve pluripotent (Klf4 rank 1 to 50 (figure 4.19, green vertical bar.)) and "late" naïve pluripotent (Klf4 ranks 600 to 649 (see figure 4.19, yellow vertical bar)).

This returned far fewer differentially-expressed genes than the previous analyses, with only 44 genes significantly ($p < 0.05$) upregulated, and 61 genes significantly ($p < 0.05$) downregulated by at least 1 log2 fold.

There were no pathways found to be enriched in the 44 upregulated genes list but 6 pathways significantly enriched in the list of downregulated genes. These 6 pathways concerned positive regulation of gene expression, cell proliferation, positive regulation of transcription. The downregulation of positive regulators of gene expression implies shutting down of gene expression, which is in line with the trend of restricting expression as mESCs move away from the naïve pluripotent state.

Despite this lacklustre showing of GO term enrichment in this data, one of this investigation's most exciting objectives was to test the ability to use this data and method to identify early responding genes that change their level of expression at the earliest stages of mESCs moving away from naïve pluripotency, ideally before the "usual suspects" such as FGF5, Brachyury (T) and Rex1 (Zfp42) do. This section will not exhaustively discuss each candidate gene, but rather some of the more

striking / novel ones as well as those with known parts to play in mESC biology. Tables are provided of the full lists of upregulated and downregulated genes across this region, in tables 4.23 and 4.24 respectively. Together, these form a list of candidate genes that this work puts forward for further investigation as useful markers to differentiate between, and monitor the change between, early and late naïve pluripotency. Probes for genes with no clear / mechanistic link to mESC pluripotency in the literature to date are highlighted in cyan.

Of particular interest is the identification of Tbx3 (fold change -2.23, p-value $2.8x10^{-13}$) as one of the earliest responders before the drop in Rex1 (Zfp42) or the rise of FGF5 / Brachyury (T), although FGF5 does also make the list with a fold change of 1.004, just over the threshold of 1 log2 fold change. Tbx3 is a known factor which both confers LIF independence and promotes the expression of Nanog (Niwa et al. 2009), (Ivanova et al. 2006). This data suggests that Tbx3 can also be used as a sensitive marker for the early naïve pluripotent state, and its expression used to gauge progress towards “late” naïve pluripotency. A plot showing a smoothed line for denoting the level of Tbx3 across the Klf4-ordered matrix is given in figure 4.25, along with the next gene of interest, carbonic anhydrase 4 (Car4.)

The first wholly-novel marker of early to late naïve pluripotency transition is Car4. Car4 is differentially expressed between early and late naïve pluripotency, being found to be significantly ($p = 2.3x10^{-5}$) increased in its expression by 1.45 log2 fold, 2.74 absolute fold change. This is depicted across the Klf4 scale in figure 4.25 upper panel. It's function in mESC biology is not known at present, so the significance of the rise of carbonic anhydrase 4 before the onset of primed pluripotency is unclear, although it may be that Car4's function of catalysing the production of bicarbonate and protons from carbon dioxide may be related to the hypoxic environment normally experienced by the ICM. Indeed, the hypoxic environment at the centre of some cancers has been cause for the linkage of the expression of other carbonic anhydrases to cancer, though carbonic anhydrases have no defined function in mESC biology as yet. Further, the fact that there is a crossover point between Car4 and the aforementioned Tbx3 in figure 4.25 suggests that there are indeed changes in trancriptional profiles occurring as naïve pluripotency moves from an earlier to later stage, but before the pronounced flip in currently-known markers occurs.

Two further examples are given, with keratin 18 (Krt18) and inhibin beta B (Inhbb) changing expression across early to late naïve pluripotency, with Krt18 increasing, Inhbb decreasing. These

Probe.ID	Gene.Name	log2.Fold.change	Wilcox.p.val
1448169_at	Krt18	1.70619744555863	5.657095E-07
1448949_at	Car4	1.45481134948046	2.285171E-05
1417845_at	Cldn6	1.42338901304833	6.794322E-08
1423691_x_at	Krt8	1.40631203723095	5.877428E-05
1420647_a_at	Krt8	1.40606935857179	3.178046E-05
1421749_at	Lin28	1.38599996255796	1.904623E-05
1454681_at	Esrp1	1.37852863496619	7.307903E-09
1415938_at	Spink3	1.37116062786202	3.372524E-05
1435989_x_at	Krt8	1.34653357563372	3.08485E-05
1415801_at	Gja1	1.26345644596243	4.798612E-05
1418094_s_at	Car4	1.21773604943889	9.942155E-06
1456326_at	Fndc3c1 /// LOC676436	1.21625572927056	2.217082E-05
1443961_at	Gm4340	1.21200891025453	0.0060050036
1419086_at	Fgfbp1	1.20886099641398	1.904623E-05
1417061_at	Slc40a1	1.19054606181714	5.979713E-06
1417210_at	Eif2s3y	1.1790582391476	0.0019753194
1443256_at	---	1.15389869845337	2.150926E-05
1419018_at	Rhox6	1.11857431796383	2.217082E-05
1416034_at	Cd24a	1.11329572656806	1.096297E-06
1416832_at	Slc39a8	1.11294196492785	1.444936E-05
1423506_a_at	Nnat	1.09246430550637	0.0016810507
1448182_a_at	Cd24a	1.08697840112459	5.859907E-07
1450947_at	2610528J11Rik	1.07871647296722	4.770126E-06
1417156_at	Krt19	1.07679967619988	0.0016424719
1460330_at	Anxa3	1.07624810731618	2.820595E-05
1452320_at	Lrp2	1.07014038804717	6.069718E-07
1418240_at	Gbp2	1.06221805908015	0.0001948455
1417895_a_at	Tmem54	1.0554327943264	4.415401E-07
1418320_at	Prss8	1.05473264359882	1.027393E-07
1448566_at	Slc40a1	1.05174935329999	3.317481E-07
1418449_at	Lad1	1.0477123054772	2.173491E-11
1423049_a_at	Tpm1	1.04331593961162	4.326893E-06
1423523_at	Aass	1.03876121794039	0.0006421448
1435906_x_at	Gbp2	1.03599915145107	0.0002683301
1448612_at	Gm5279 /// Gm7850 /// Sfn	1.03572561500568	8.491997E-06
1437502_x_at	Cd24a	1.03332469484237	2.355247E-05
1450285_at	Gm2098 /// LOC100040390	1.02936416547357	0.0001702063
1420549_at	Gbp1	1.02566587187393	0.0023162129
1416242_at	Klhl13	1.022969318529	0.0014961643
1428804_at	Mfap3l	1.00867473617806	5.389606E-05
1427133_s_at	Lrp2	1.00634251007976	3.923276E-06
1438883_at	Fgf5	1.00439138529978	2.502053E-08
1418984_at	Inadl	1.00359614768412	2.728031E-10
1448690_at	Kcnk1	1.00176065252548	1.954765E-06

Table 4.23: Fold changes (upregulated >= 1 log2 fold) and p-values of probes resulting from differential expression analysis across early to late naïve pluripotency areas of the Klf4-ordered HPM matrix. Probes with cyan highlight are for genes with little current known link to mESC pluripotency states.

Probe.ID	Gene.Name	log2.Fold.change	Wilcox.p.val
1437479_x_at	Tbx3	-2.23209985334726	2.7689788261102E-13
1417395_at	Klf4	-2.19800162299615	6.927101010008E-18
1423100_at	Fos	-2.02078735099522	1.8567735828441E-07
1448736_a_at	Hprt1	-1.87663785136869	6.0639879169026E-08
1426858_at	Inhbb /// LOC10004680	-1.87273421200587	1.6901455185834E-08
1429524_at	Myo1f	-1.63373751222366	7.4009233345763E-10
1456735_x_at	Acpl2	-1.47815313384944	3.7969473410994E-06
1438781_at	Tet2	-1.44766984596176	1.0269695341346E-10
1422573_at	Ampd3	-1.44150705044951	3.111517548242E-10
1435204_at	Prmt8	-1.41713574280224	7.7548052082484E-07
1452142_at	Slc6a1	-1.33537267708939	1.7254533088917E-07
1416529_at	Emp1	-1.30671110697855	8.9114032776548E-07
1434917_at	Cobl	-1.29723938183136	3.0248910619924E-14
1422134_at	Fosb	-1.27369270480431	0.004769185865005
1427680_a_at	Nfib	-1.2642365121661	2.0575672042341E-08
1447831_s_at	Mtmr7	-1.26018913055295	9.0107426513298E-05
1449344_s_at	2210409E12Rik	-1.2544994434373	0.00447470462253
1418538_at	Kdelr3	-1.25351096493834	4.5751260451588E-07
1436742_a_at	Accsl	-1.23560531378227	4.1119336940054E-07
1452514_a_at	Kit	-1.23090581298702	4.5430634833731E-11
1434719_at	A2m /// LOC677369	-1.22979091489677	3.3725236443779E-05
1456182_x_at	Mela	-1.21770003464296	0.00090643077033
1424719_a_at	Mapt	-1.20209505400727	7.2321443408402E-07
1422937_at	Fzd5	-1.19083795328054	1.4047170982152E-10
1456887_at	Cmklr1	-1.1890268105006	9.8049555324702E-12
1420909_at	Vegfa	-1.17390042129202	2.7048121746121E-08
1456242_at	Gm7325	-1.17032238258192	5.0710890714306E-09
1444390_at	Prdm14	-1.15144164955731	6.0697182900748E-07
1454901_at	Ypel2	-1.14633127703364	1.4892190702711E-07
1423584_at	Igfbp7	-1.14630988366945	0.000109620683021
1431416_a_at	Jam2	-1.14022878332974	1.0592036229165E-06
1456609_at	Camk2n1	-1.14018141611979	6.5834852301104E-06
1416488_at	Ccng2	-1.13955235083266	0.000148520147157
1418645_at	Hal	-1.13800404134284	1.2339990476428E-09
1423176_at	Tob1	-1.12923136797559	1.058605497535E-05
1449146_at	Notch4	-1.12751803154829	1.9021315709154E-08
1436546_at	Lix1l	-1.12393836507941	3.4134309067322E-08
1418345_at	Tnfsf12 /// Tnfsf12-tnfsf	-1.1139943632653	1.9604968432006E-09
1456250_x_at	Tgfbi	-1.11023379929702	6.9837088366052E-07
1439881_at	C030013E06Rik	-1.1061325341709	3.7969473410994E-06
1417065_at	Egr1	-1.0966725915392	0.000184612660062
1448228_at	Lox	-1.09602322631905	0.202816995116694
1440692_at	Gm364	-1.09162835659487	0.005881775416865
1416121_at	Lox	-1.08666138172084	0.111203380460296
1436970_a_at	Pdgfrb	-1.08075528591853	0.000512206838224
1416405_at	Bgn	-1.08051178721812	0.02751593710251
1424037_at	Itpka	-1.07810148085244	4.4099926809132E-10
1418467_at	Smarcd3	-1.05816959845663	4.8095935680966E-10
1450297_at	Il6	-1.05796830416683	1.1627071063197E-05
1441045_at	Ddx43	-1.0284460269016	4.3268928067173E-06
1442489_at	D1Ertd564e	-1.02822404265232	0.001126223170162
1450857_a_at	Col1a2	-1.02232602907735	0.067169235475786
1442018_at	Btg1	-1.02055151880666	3.5558878054356E-06
1434624_x_at	Rps9	-1.00732187451079	0.060369337764021
1437188_at	Gabbr1	-1.00064475351948	1.9783678247906E-08

Table 4.24: Fold changes (downregulated <= -1 log2 fold) and p-values of probes resulting from differential expression analysis across early to late naïve pluripotency areas of the Klf4-ordered HPM matrix. Probes with cyan highlight are for genes with little current known link to mESC pluripotency states.

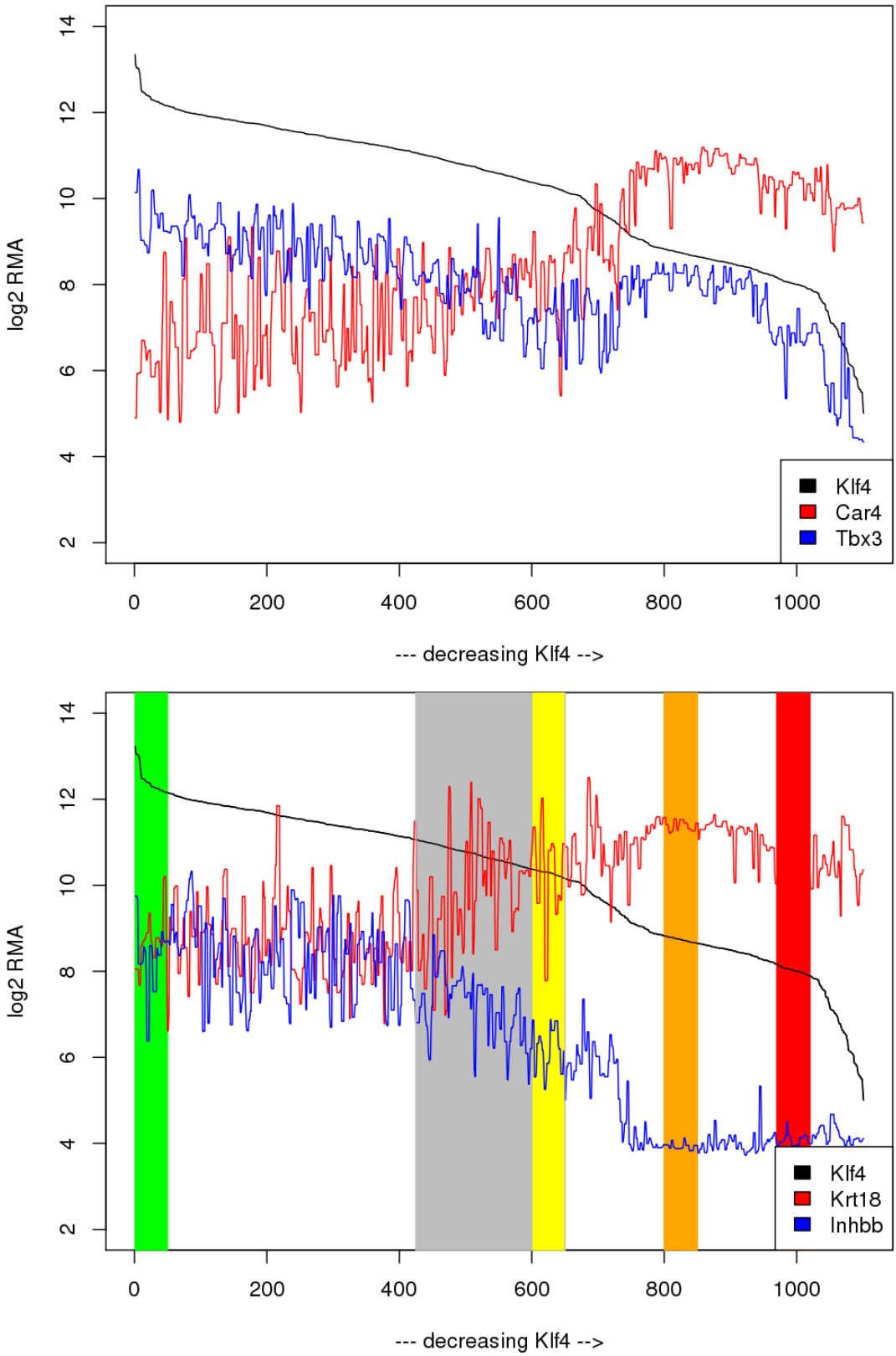

Figure 4.25: Expression of Car4, Tbx3 (upper plot), Inhbb and Krt18 (lower plot) as smoothed line plots across the Klf4-ordered HPM matrix demonstrating expression changes in these genes across early to late naïve pluripotency, see section 4.3.7. Vertical coloured highlight areas of lower plot match those from figure 4.19, grey area of lower plot draws attention to divergent changes in expression of Krt18 and Inhbb as examples of "late naïve pluripotency" changes.

changes are plotted across the Klf4 spectrum in figure 4.25's lower panel. The function of these genes in the context of naïvely-pluripotent mESCs is not discernible from this data, although Inhbb has been found to be downregulated upon LIF withdrawal and has therefore been hypothesised to be downstream of STAT3 signalling (Sekkai et al. 2005).

As can be seen from table 4.23, Krt18 was not the only keratin to follow this pattern; two other keratins, keratins 8 and 19, (fold changes of all three being 1.71, 1.41 and 1.08 log2 fold change respectively), which are markers of ectodermal / epithelial tissue, also follow this pattern. All three of these genes appear to remain somewhat steady in their expression at the earliest parts of naïve pluripotency, but then there is a marked rise in the expression of all three at around the 425 index mark (see figure 4.26, grey vertical line), long before the change point from naïve pluripotency to primed pluripotency, which occurs at around the 600 index mark (yellow vertical highlight bar.)

Lin28 also makes an appearance in this analysis, which is of interest given its role in suppressing the maturation of Let7 pri-miRNAs and its connection to cell cycle machinery in mESCs (Viswanathan et al. 2008), (Xu et al. 2009), (Hagan et al. 2009). This work suggests Its novel use as a marker of transition between early to late naïve pluripotency.

Claudin 6 (Cldn6) shows a fold change of 1.42 log2 (2.68 absolute) and a p-value of 6.8×10^{-8}. Claudin 6 is a tight junction protein which, in agreement with the findings presented here, has been linked to the expression of keratin 8 (Krt8) in mouse embryonic endothelium / embryoid bodies (Turksen et al. 2001). Claudin 6 is worthy of specific mention as (Tesar et al. 2007) shows high Claudin 6 as a marker of the EpiSC / primed state. Increasing expression of Krt8 and Cldn6 therefore recapitulates what is found in the literature. What is novel from this work, however, is now the knowledge that Cldn6 begins its increase in expression long before the flip in expression of the canonical naïve / primed pluripotency markers.

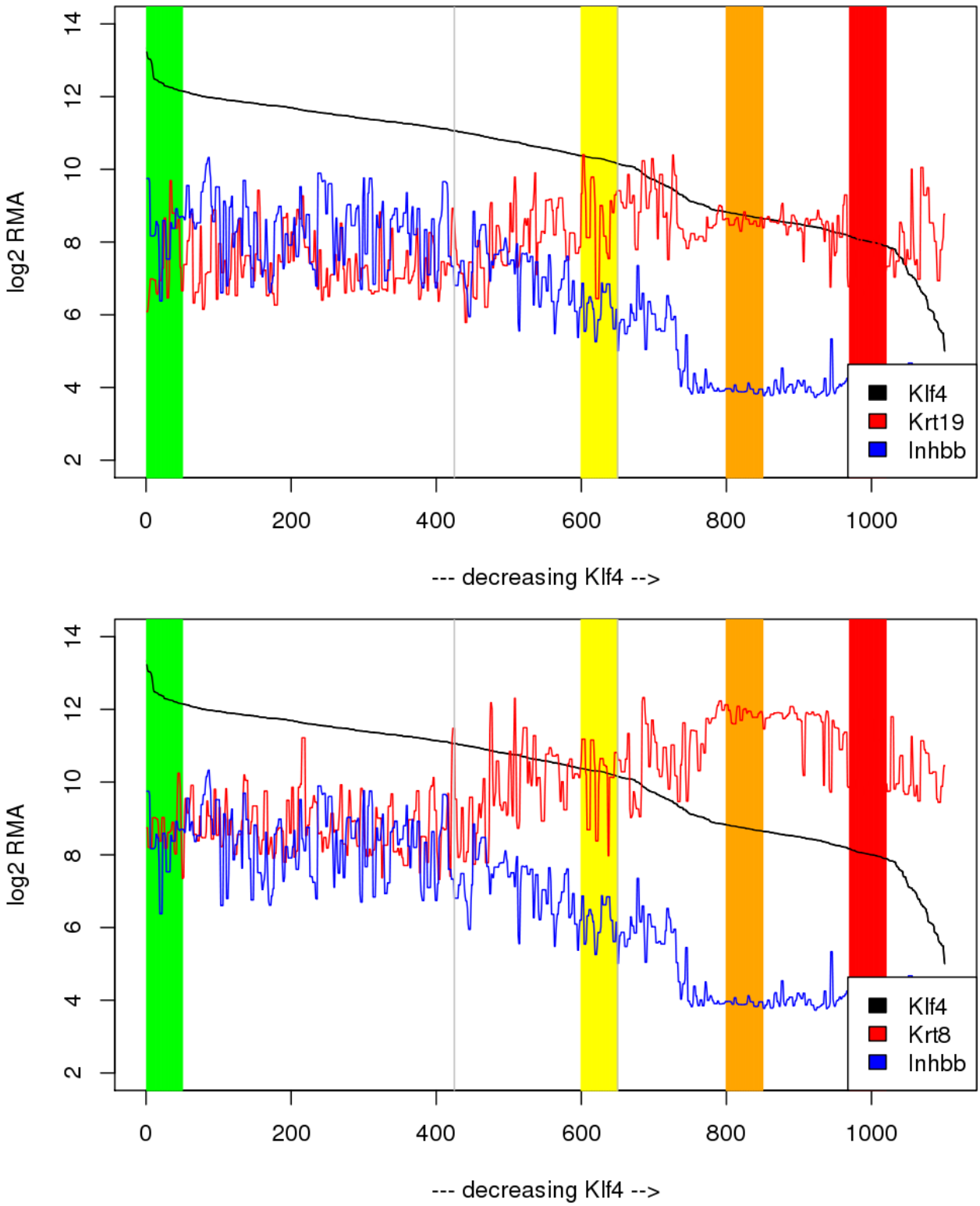

Figure 4.26: Expression of Krt8 and Krt19 as smoothed line plots across the Klf4-ordered HPM matrix demonstrating expression changes in these genes found in the putative "late naïve pluripotent" state identified in section 4.3.7. Vertical coloured highlight areas match those from figure 4.19. Grey vertical lines are drawn around this putative late naïve pluripotent state to highlight divergent changes in expression of the genes selected here. Full gene lists are given in figures 4.23 and 4.24.

Despite a lack of enrichment for any GO terms involving FGF signalling between early and late naïve pluripotency, epithelial splicing regulatory protein 1 (Esrp1) was found increase its expression 1.38 fold (p-value $7.3x10^{-9}$) across this period. Esrp1 is involved in the alternative splicing of fibroblast growth factor receptor 2 during the epithelial-to-mesenchymal transition (EMT) (Warzecha et al. 2009). A novel implication from this work, therefore is that a changing of the subtype of FGF receptor 2 may be occurring during the early to late naïve pluripotency state, although this would require experimental confirmation.
Furthermore, fibroblast growth factor binding protein 1 (Fgfbp1) occurs in table 4.23 with a 1.21 fold increase in its expression across the same period (p-value $1.9x10^{-5}$), which piqued interest in that analysis of the whole early to late naïve pluripotency region may find more to do with FGF signalling (see section 4.3.10.)

Tet methylcytosine dioxygenase 2 (Tet2) also made this list, and was shown to decrease in its expression -1.45 log2 fold ($p = 1.03x10^{-10}$). Tet2 is of interest to this work given it's role in maintenance of 5-methylcytosine epigenetic markers. Tet2 has been shown to interact directly with Nanog (Costa et al. 2013) and be regulated by Oct4, although is ultimately dispensable for pluripotency (Koh et al. 2011). It is interesting to see a marked drop in Tet2 as being associated with this putative novel cellular state, although microarray data can unfortunately provide no information on any epigenetic changes which may also identify this state.

Prdm14, a known pluripotency factor (Yamaji et al. 2013), is confirmed to drop towards late naïve pluripotency, with a fold change of -1.15 ($p = 6.07x10^{-7}$). This is worthy of mention here as Prdm14 may is known to promote the naïve pluripotent state through inhibition of both FGF signalling and epigenetic changes. Taken together with the concomitant increase seen in Fgfbp1 mentioned above, this suggests that, in accordance with the literature, FGF-related genes are involved in the progression from early to late naïve pluripotency, although this differential expression approach did not find an enrichment for the FGF signalling pathway. It will prove interesting, in future work, to experimentally manipulate genes identified in this section to observe effects on naïve pluripotency (see section 5.1.7). The expression pattern of the genes mentioned in this section are given in figure 4.27, except for those already in figure 4.26, to avoid repetition.

One of the most exciting results from this analysis came from simultaneous display of selected candidate genes as a heatmap generated from tables 4.23 and 4.24. This heatmap is provided as figure 4.28, with expression values from all 50 samples from each of early and late naïve regions of

the Klf4-ordered matrix (ranks 1 to 50 and 600 to 649 respectively). For clarity, this heatmap is row-scaled to demonstrate relative upregulation or downregulation of each gene. A final summary bar plot of the differential expression values of selected genes, including all of those mentioned in the text above is given in figure 4.29, representing the change in expression of those genes as early naïve pluripotency changes to late naïve pluripotency, but before the canonical flip in FGF5, Brachyury and Rex1 occur.

In summary of this section, it was interesting to find no significant enrichments across this part of the data when using differential expression. Regardless of there being no GO enrichments of any major significance in this section of the data, the Klf4-ordered matrix has, as was desired, proven useful in splitting the data into distinct “early” and “late” naïve pluripotent states (also see section 4.3.4). Furthermore it has allowed the successful identification of early responding genes and markers of these early and late naïve pluripotent states. A number of interesting future experiments are called for on the back of these findings (see section 5.1.7).

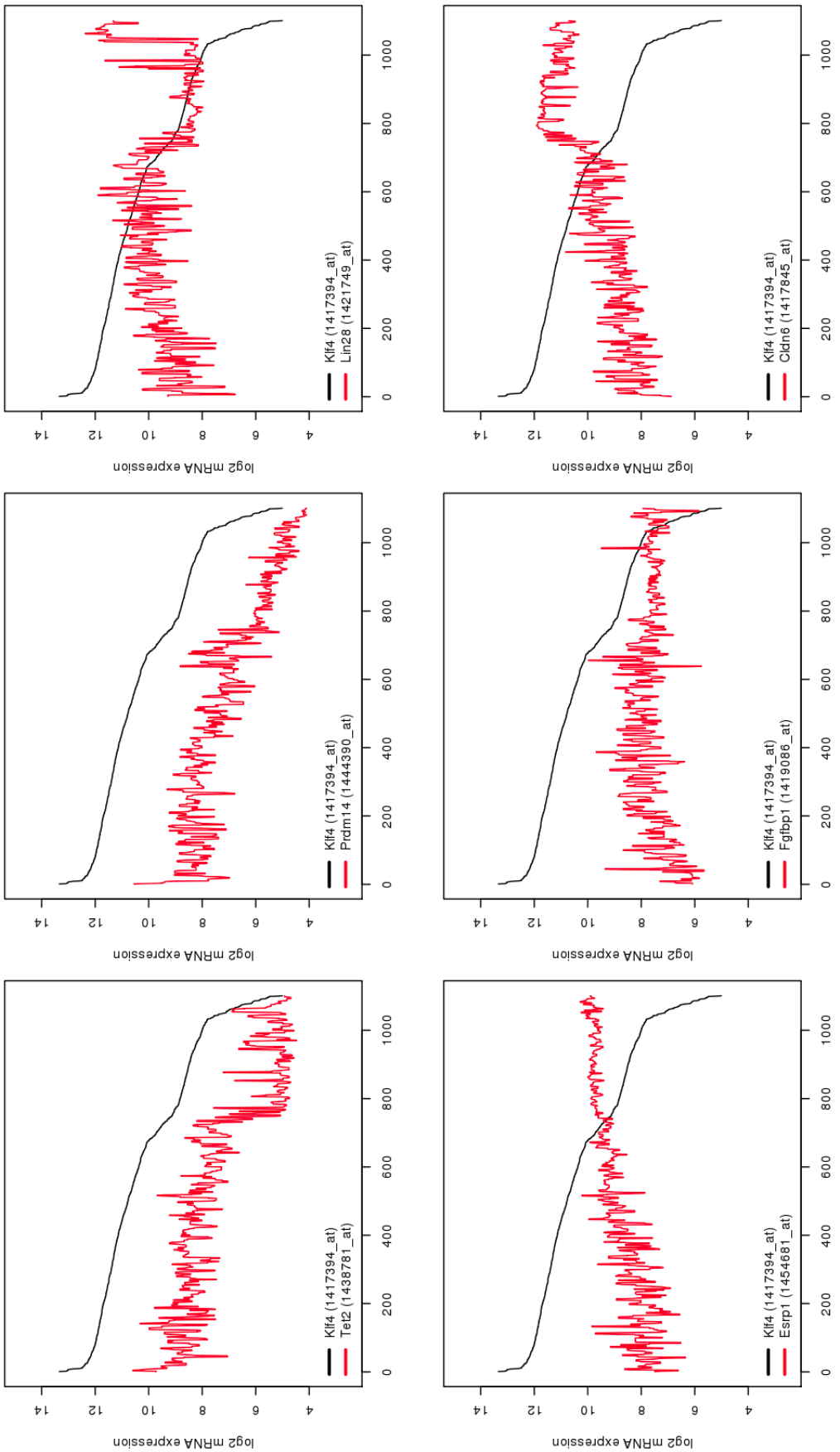

Figure 4.27: Smoothed line plots of selected genes identified in section 4.3.7 as potential markers of an identified late naïve pluripotent state prior to the change to primed pluripotency. Early naïve pluripotency is defined in this work as the area of highest Klf4-expression, which coincides with strongest expression of known naïve pluripotency markers (see figure 4.30) and late naïve pluripotency is defined here as the region on the Klf4 spectrum just prior to the reversal of the $FGF5_{low}$, Brachyury $(T)_{low}$, $Rex1_{high}$ state.

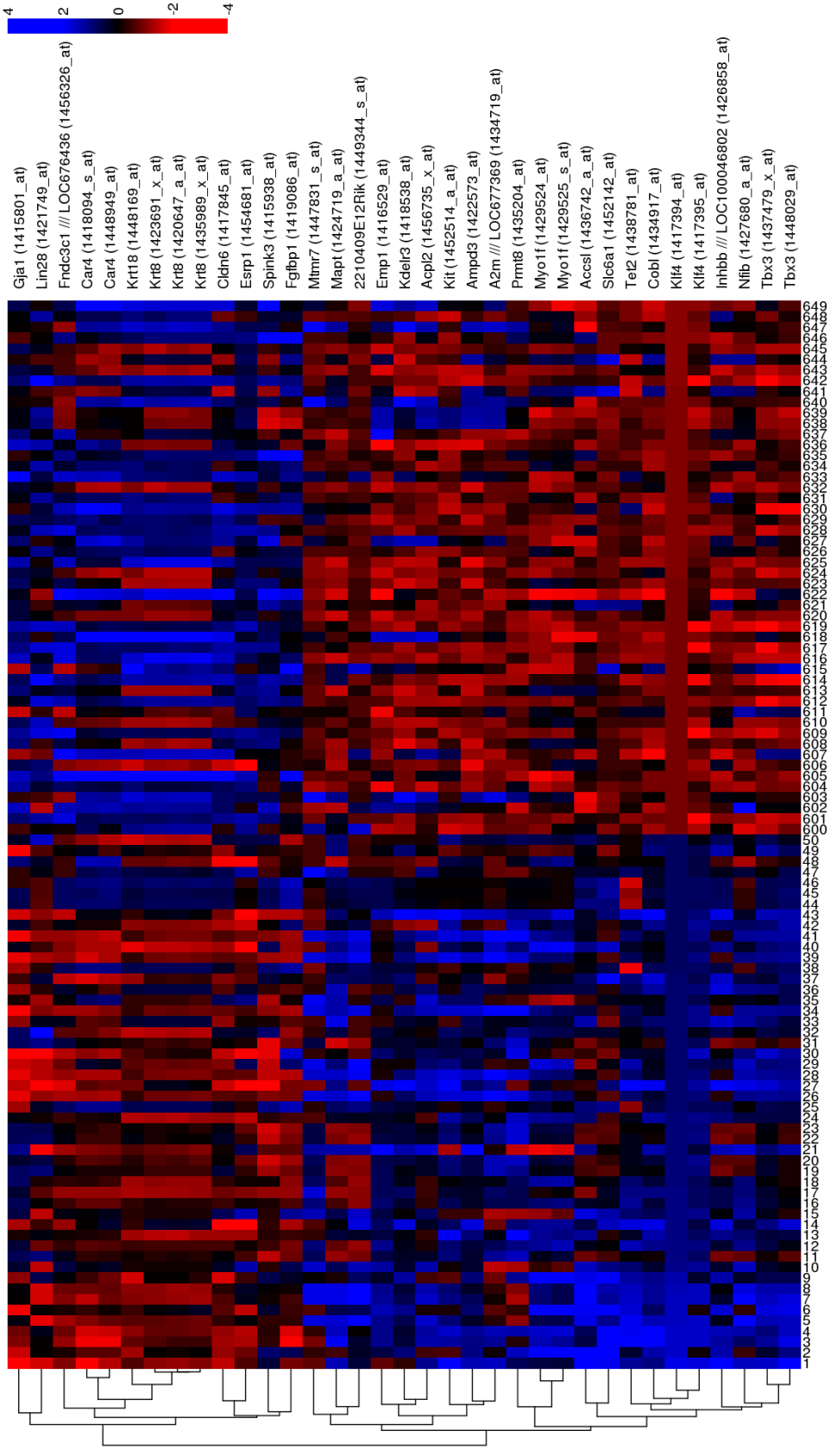

Figure 4.28: Heatmap showing change of expression of selected identified putative markers of the early and late naïve pluripotent states from the HPM matrix. Numeric indexes on bottom of heatmap are Klf4 ranks. Samples numbers are the same as those chosen for differential expression analysis of early vs. late naïve pluripotency. Row-scaling is applied for clarity and show relative upregulation or downregulation of all selected genes between these two regions of the Klf4 spectrum.

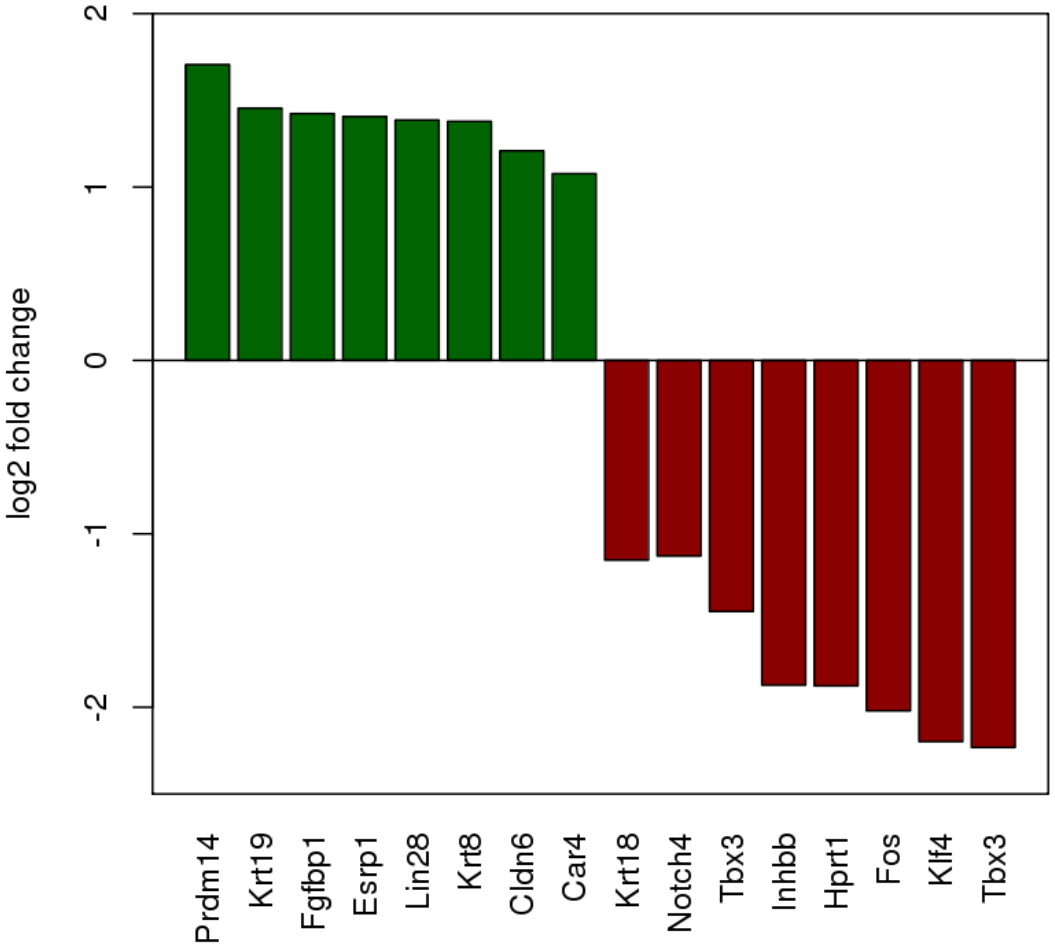

Figure 4.29: Barplot displaying change in expression of selected genes shortlisted as candidates for markers identifying a late naïve pluripotency state as being distinct from the early naïve pluripotent state. For details, see section 4.3.7. Numbers used in this plot can be found, along with changes in expression of other genes, in figures 4.23 and 4.24. Numbers were generated by the observation of significant ($p < 0.05$) changes in mean expression of these genes between groups of 50 consecutive samples from the early naïve pluripotent region and the late naïve pluripotent region of the Klf4 ordered data, subtracting the mean expression of the lower-Klf4 group from the higher-Klf4 group. Samples from the early and late naïve pluripotent region of the Klf4-ordered matrix can be seen as the green and yellow vertical bars in figure 4.19 respectively.

4.3.8 Differential expression analysis between primed pluripotency and early differentiation shows shutdown of transcription of developmental genes and activation of survival genes

This analysis was carried out by calculating the fold changes of all probes between a first set of samples between Klf4 rank 800 and 849 (see figure 4.19, orange vertical bar.), and a second set of samples between Klf4 ranks 970 and 1,019 (see figure 4.19, red vertical bar.) Only 106 probes were found to have significantly ($p < 0.05$) increased their expression by 1 log2 fold (2 absolute fold change.) Only 5 pathway enrichments were found in this list, pertaining to negative regulation of apoptosis (3 pathways), filament bundle assembly and "cellular component morphogenesis").

Those probes found to have significantly decreased their expression by 1 log2 fold across this region numbered 385, and gave 199 GO biological pathway enrichments. The signalling pathways found to be enriched in this list were Wnt, BMP, and VEGF pathways. Proliferation and regulation of gene expression pathways were also enriched and entered into the summary figures 4.35 and 4.36. Development of vasculature, skeletal system, urogenital system, heart, kidney, nerve, lung, gland, limb, pancreas, eye, gut and lymph system were found here, which is interesting to see, given that this region of the matrix was hypothesised to represent exit from primed pluripotency towards differentiation / devlopment. The earlier analysis of genes changing between naïve to primed pluripotency uncovered a large list of developmental pathway genes being upregulated, while here a great number of downregulated genes are developmental in nature.

One possible explanation for a large increase in the expression of developmental genes between naïve and primed pluripotency was offered in work by (Turner 2008) and particularly work by (Efroni et al. 2008), who put forward that the ES state has open chromatin and exhibits transcriptional hyperactivity. If this effect was strengthened during naïve to primed transition, it would explain why so many developmental pathways are upregulated between naïve pluripotency and primed pluripotency (section 4.3.6) and thus, later, as Klf4 drops and differentiation starts to proceed, a relative large downregulation of these developmental genes would become apparent here.

However, this explanation is unsatisfactory for two reasons. Firstly, it has not been shown in the literature that there is any increase in transcriptional hyperactivity across the naïve to primed pluripotency transition. In fact, chromatin is, if anything, already becoming less permissive to transcription at the primed pluripotency point, arguing against transcriptional hyperactivity, and

secondly, the very idea of transcriptional hyperactivity in ES cells has also failed to gather evidence (Marks et al. 2012). This caused concern when taking the results of this analysis together with those from across the naïve to primed pluripotent state, as the number of significantly enriched developmental pathways is so large across both of these analyses. However, an explanation for this phenomenon becomes apparent when it is considered that transcriptional pausing is an important transcriptional regulation method in mESCs (see the functions of Myc in mESCs in section 1.3.2). Given that autocrine FGF signalling is pronounced during naïve pluripotency, and that MAPK/Erk signalling is the downstream effect of this, it is far more likely that, during the naïve pluripotent state, transcriptional pausing is, if anything, increased compared to the primed state, when it is considered that ERK1/2 has been found in recent work to promote promoter proximal pausing in mESCs (Hackett and Surani 2014) (Tee et al. 2014).

This neatly ties together the idea of the pluripotent state as being “poised” for the transcription of many genes, particularly developmental pathways, in that RNApolII is able to access many of these sites, but FGF/MAPK/ERK signalling prevents the transcription of full-length mRNA transcripts from these genes. With these recent advances in understanding in mind, the data appears to elegantly recapitulate these phenomena, with a distinct increase in the number of upregulated genes associated with developmental pathways occurring at exactly the time of the switch-over to primed pluripotency which also coincides with the data where there is the only observable major drop in the detection of mRNA for FGF4, as shown in figure 4.30. The later observation of a large drop in the expression of developmentally-related genes can now be understood to be a natural progression towards a more differentiated cell state wherein there will be a progressive switching off of different developmental lineages as differentiation progresses. This last phenomena (the switching off of many developmental lineages) is, it must be said, may also be affected by the makeup of samples which form the lower Klf4 part of the data.

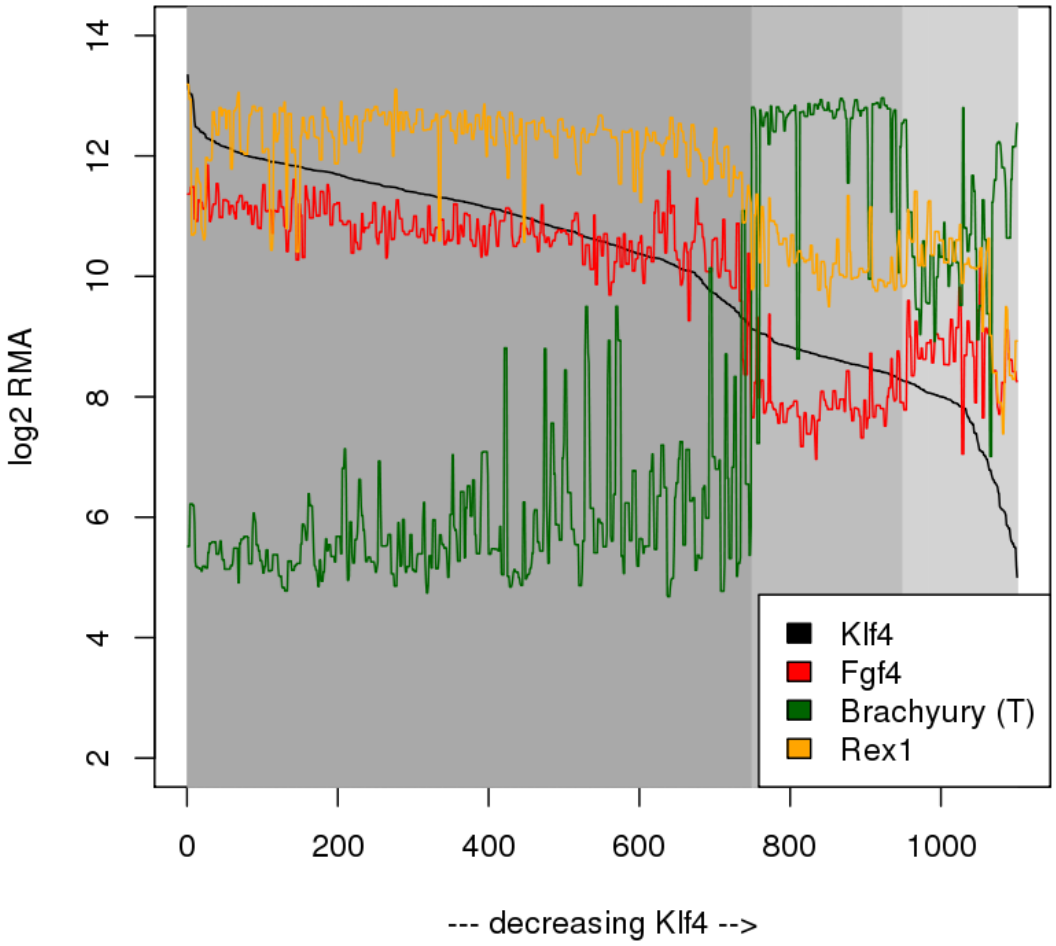

Figure 4.30: Smoothed line plots of naïve pluripotency marker Rex1, primed pluripotency marker Brachyury (T) and FGF4, demonstrating the drop in FGF5 occurring concurrently with the reversal of expression of the aforementioned Brachyury / Rex1 markers, suggesting, therefore, relief of proximal promoter pausing when FGF4 drops (see section 4.3.8.)

4.3.9 A scanning window method for detecting significant changes in gene expression captures the patterning of the Klf4-ordered matrix with a 25 sample window and threshold of 1 log2 fold change

The scanning window method detailed in section 4.2.5 and described graphically in figure 4.1 was run a total of 15 times. The first 12 runs were to investigate the influence of the width of the scanning window (read: the number of adjacent samples used to calculate a mean level of gene expression in each comparison.) Each time the method was run, the number of significant expression changes over 1 log2 fold (either up or down) in each scanning window was recorded. 11 of 12 runs investigating a suitable window width for the later analysis are shown in figures 4.31 and 4.32.

When the number of samples used in the scanning window was set at 3, there were no fold changes of more than 1 log2 fold in magnitude that were significant in any window. As such there is no plot of the expression changes across the Klf4-ordered matrix with a window width of 3 samples.

As can be seen from the plots in figure 4.31 when window widths of 5 and 10 were used, there was little discernible pattern in the data, particularly compared to when the window width was increased to 15 samples.

Now with the window width at 15 samples, an overall pattern becomes visible in figure 4.31. As the width of the scanning window is increased beyond 15, that pattern that emerged continues to be seen, albeit with a reduction in noise. The peak at around the 750-th scanning window remains while significant expression changes are still seen across other regions.

This represents, therefore, the first pattern to become apparent in the data when using the scanning window method; that a large number of changes in gene expression occur around the 750^{th} window. This was expected and shows that this approach is capturing patterning of the data, with a majority of changes in gene expression around the point which coincides with the change from naïve to primed pluripotency in the previous analyses. The change from a scanning window of 10 samples to a scanning window of 15 samples represents the single largest drop in the total number of observed significant gene expression changes, as can be seen from the last plot of figure 4.32, which depicts the relationship between this total number of expression changes and scanning window width.

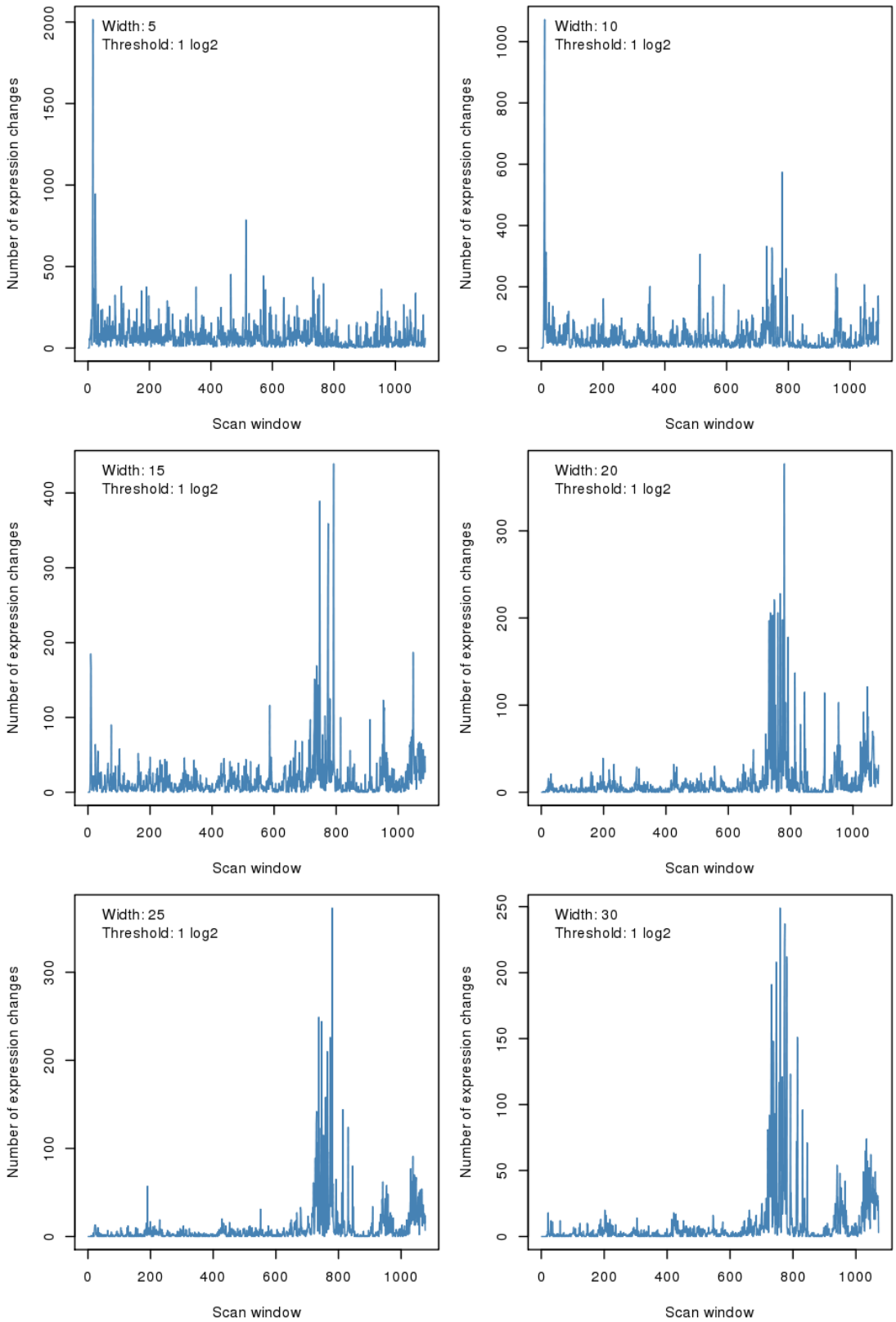

Figure 4.31: Assessment of the suitability of different scanning window widths using plots depicting the total number of genes found to change their expression by a minimum of 1 log2 fold (either upward or downward) in all scanning windows across the Klf4-ordered HPM matrix. Note the differences in scales on the y-axes. Plot 1 of 2.

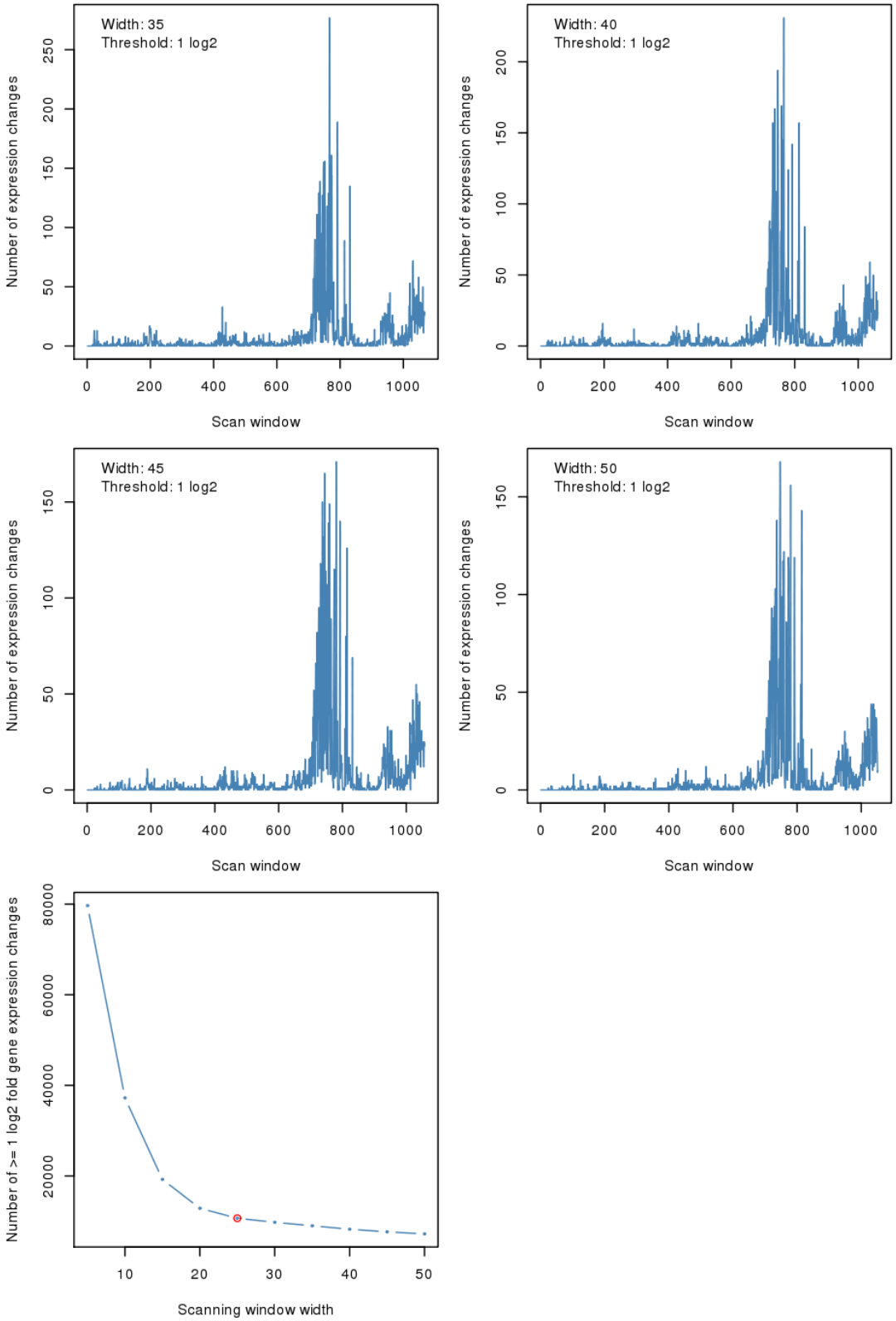

Figure 4.32: Assessment of the suitability of different scanning window widths using plots depicting the total number of genes found to change their expression by a minimum of 1 log2 fold (either upward or downward) in all scanning windows across the Klf4-ordered HPM matrix. Note the differences in scales on the y-axes. Lowermost plot shows the grand total (across all scanning windows) of how many genes significantly ($p <= 0.05$) change their expression, given each scanning window width, justifying the use of the 25-sample window width as having likely removed "noise". Panel 2 of 2.

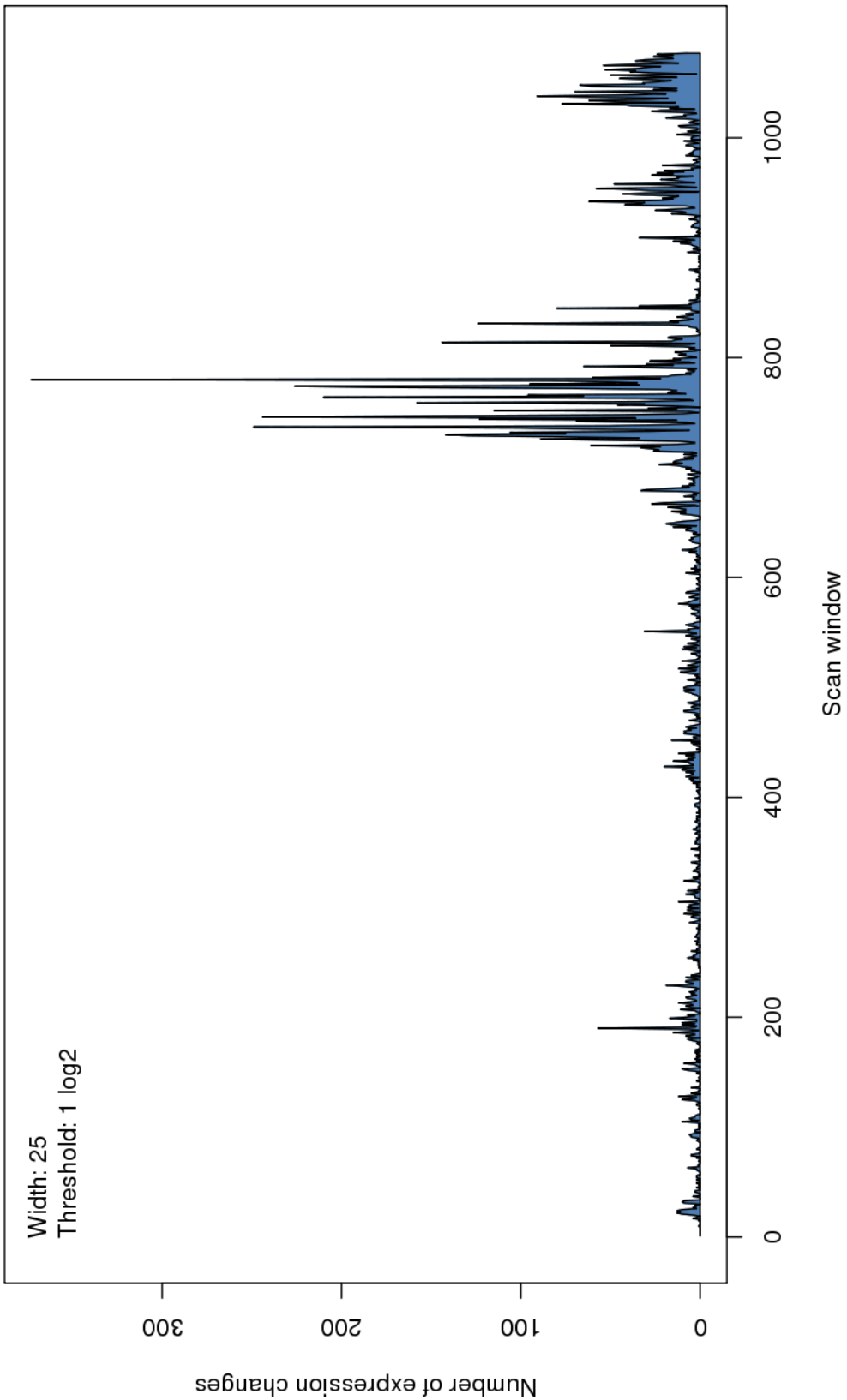

Figure 4.33: Filled curve of all significant (p <= 0.05) changes in mean expression of genes across the Klf4-ordered matrix when using window scanner with window width 25, threshold 1 log2 fold.

Increase of the scanning window width to 20 samples brings with it the first visually-striking reduction of “noise” in this scan, compared to the scanning window width of 15.

As can be seen from the last panel of figure 4.32, it is when the scanning window width is set to 25 or above, that the curve of the line flattens off. Taken together with the previous observations in this figure, a window width of 20 has removed most noise, while 25 or above begins to slowly lose the number of detected significant changes. It was therefore decided that this, the 25-wide scanning window mark, was an appropriate balance of reduction of noise, but retention of as many significantly-changing genes as possible.

Concerning an appropriate threshold to use when scanning the data, the previous differential expression analyses' biological pathway enrichments contained pathways that would be expected of a dataset capturing pluripotency and exit from it (see sections 4.3.5 to 4.3.8). This suggests that the use of a similar threshold (1 log2 fold) should be acceptable here. Indeed, fair comparison between the differential expression approach and the scanning window approach is only possible if this threshold is kept the same.

4.3.10 Scanning window analysis of early to late naïve pluripotency reveals a plethora of significantly-enriched biological pathways

As one of the most striking results in this thesis, the scanning window approach to investigating transcriptional events between early and late naïve pluripotency must be held in direct comparison to the results of using differential expression analysis between these two regions of the Klf4 spectrum (see section 4.3.7).

The differential expression analysis (see section 4.3.7) found a potential novel cellular state before the change to primed pluripotency and identified markers of this possible state change for downstream experimental investigation. However, there were only a total of 105 genes significantly changing their expression by at least 1 log2 fold between early and late naïve pluripotency and only 6, highly-generic pathway enrichments. The author's suspicion that a great many other changes may be missed by traditional differential expression analyses investigating such subtle changes in cellular state would appear to be justified, as between scanning windows 1 and 625 (which corresponds to all changes that occur between Klf4 ranks 1 and 650, given a scanning window

width of 25) 1,212 probes were found to have have a significant fold change of at least 1 log2 fold at some point.

Concerning pathway enrichments, these 1,212 probes resulted in a list of 187 biological pathway enrichments which satisfied a q-value of $q < 0.05$. This is a vast increase on the 6 found by differential expression. Prior to examining these enrichments, there were initially two concerns. First, that the scanning window not been wide enough to sufficiently reduce "noise". Secondly, that the central hypothesis of there being meaningful transcriptional events occurring between the putative "early" and "late" naïve pluripotent states was false. Had either of those concerns been true, then this would have generated a list of enrichments approaching that which would occur by choosing genes at random (*id est* "noise"), returning little to no statistical significance and / or pathways with no relevance to mESC biology.

These significantly-enriched pathways included cell proliferation, transcriptional regulation, developmental pathways (vasculature, muscle, epithelium, eye, skeletal, kidney and others).

An excellent result was to see that the FGF signalling pathway achieved a q-value of 4.7×10^{-2} here. From the prior disappointment that this pathway, known to be so critical in naïve and primed pluripotency (see section 1.3.3), was not found by differential expression analysis across this region, this greatly supported the utility of the scanning window approach. As for the other signalling pathways previously found in this chapter, none of VEGF, Notch, BMP, Wnt or MAPK passed the significance threshold of $q < 0.05$.

Naturally, this result begged the use of the window scanning method across the other areas of the Klf4 spectrum for comparison to the differential expression method. (see figure 4.19). These other window scanning runs were also to demonstrate agreement between differential expression and window scanning in areas of the Klf4-ordered data where changes in expression would likely be more pronounced, simply to validate that the window scanning method finds *bona fide* signals in the data. A large discrepancy in what the window scanning and differential expression approaches find in these areas of the Klf4-ordered data might suggest that there is issue with the methodology of the window scanner.

The significantly-enriched pathways here were added to summary figures 4.35 and 4.36, which also provide the direct comparison between the differential expression and window scanning approach.

4.3.11 Scanning window analysis of late naïve to primed pluripotency reveals strong enrichment for transcriptional, developmental, PDGF signalling, post-transcriptional gene expression regulation, stress response and stem cell pathways

Window scanning between windows 600 and 825 gives the window scanner chance to look for genes which change their expression significantly between late naïve and primed pluripotency, seen as the yellow and orange vertical bars respectively in figure 4.19.) This was carried out with a window width of 25 samples and thresholds of 1 log2 fold change, $p < 0.05$ for genes to be included, as defined in section 4.3.9.

This scan found 4,796 genes to change their mean expression significantly at some point between these areas of the Klf4 spectrum by at least +/- 1 log2 fold.

The list of pathway enrichments resulting from this extensive list included regulation of transcription and proliferation, differentiation, chromatin organisation and a long list of developmental pathways. In addition, the Wnt, VEGF, MAPK signalling pathways are significantly enriched in this list, but it is the appearance of the PDGF signalling pathway here that was unexpected. Some enlightenment from the literature may yet be found in that PDGF signalling is known to be relevant to the maintenance of human ES cells (Pebay et al. 2005), although PDGF signalling does not appear to be functionally characterised in mESCs as yet.

One other pathway to make a first appearance in this work here is “cellular response to stress”. The finding of an enrichment for stress response genes in the change between late naïve and primed pluripotency is interesting as it is already known that mESCs have considerably enhanced defences against stressors such as reactive oxygen species (ROS), DNA damage and heat shock (Saretzki et al. 2004).

Interestingly, this is also the first time in which the biological pathway “posttranscriptional regulation of gene expression” pathway has been found to be significantly enriched, although, this work is not well placed to observe post-transcriptional changes, being microarray-based.

Finally, there was enruchment found for stem maintenance, differentiation and development in this analysis, when none of these pathways achieved statistical significance when using the differential expression method across the same region.

4.3.12 Scanning window analysis of primed pluripotency toward early differentiation shows complete agreement with differential expression analysis of the same regions of the Klf4 spectrum

This window scan ran from Klf4 ranks 800 to 995 and, with a window width of 25 samples, this corresponds to observing changes between Klf4 ranks 800 and 1020.

The enrichments found here agree with those found by differential expression across the same part of the Klf4-ordered data (see figure 4.19, orange and red vertical bars, see section 4.3.8) Enrichments are found for the regulation of transcription and proliferation, apoptosis, differentiation and a large number of developmental pathways including vasculature, lung, gland, neuron, skeletal, urogenital, heart, eye, kidney and germ cell

As the directionality of these changes was not taken from the results of the window scanning analysis, it is not proven here that these developmental genes were found to change in broadly the same direction (downregulated) as was found in the differential expression analysis as conducted in section, although future work may involve bringing directionality to the window scanning results (see chapter 5.)

Exact agreement is also to be found concerning the signalling pathways found to be significantly enriched across this part of the Klf4 spectrum, with Wnt, BMP, and VEGF being significantly enriched.

All of the above-mentioned pathways are summarised and compared with those that occur in other analyses in this chapter in the summary figures 4.35 and 4.36.

4.3.13 Scanning window analysis across all of the Klf4-ordered data shows enrichment for nearly all pathways identified in previous analyses and adds others

As an interesting test of the functionality and capability of the window scanning approach, scanning across all of the data, from highest Klf4 to lowest Klf4 was carried out to check for its ability to recapitulate the results of all of the analyses across sub-regions of the data, with a view to recommending this approach for the interrogation of large, ordered datasets detailing biological phenomena of interest.

Across the entire matrix, 1,077 scanning windows were observed for significant changes in mean expression. Full results of this scan are provided in the accompanying R Object file, with filename "HPM.Matrix.In.Klf4.Order.Significant.WindowScan,Results.W25.T1.RObject", on the accompanying DVD in the "Chapter 4/WindowScanResults" folder. A total of 6,460 probes were found to significantly change their mean expression by at least 1 log2 fold in at least one scanning window across the Klf4 spectrum.

Enrichments across the entire matrix overwhelmingly met the expectation that the scanning window approach would be capable of finding the results of the other analyses performed across smaller regions of the matrix. Here, significant enrichments were found for regulation of transcription, proliferation, apoptosis, differentiation, post-transcriptional regulation, chromatin modification, cellular response to stress and the three stem cell pathways of differentiation, maintenance and development. The Wnt, BMP, MAPK and VEGF pathways were successfully identified as significantly enriched, but not the Notch or FGF pathways, which came close to but did not achieve statistical significance.

Two pathways which did not achieve significance in the other analyses in this chapter now appear, being the epithelial to mesenchymal transition (EMT) pathway and a previously-unseen signalling pathway; the TGFβ receptor signalling pathway. It is interesting to see that these enrichments appear for the first time in the same analysis, as the TGFβ pathway is known to play a role in the epithelial to mesenchymal transition (EMT) (for an excellent review on the pathways that can stimulate EMT, and the relevance of EMT to the cancer / stem cell phenotype, see (Polyak and Weinberg 2009)). For more information on the role of TGFβ signalling in mESCs, see section 1.3.3.

The ability of the scanning window approach to find so many of the significantly-enriched pathways from other analyses, using only one run, is a striking result. From other analyses' combined lists of pathway enrichments, only the Notch pathway and the FGF pathway did not make the final list of enriched pathways for the full matrix window scan (see summary figures 4.35 and 4.36).

Even then, the Notch pathway only barely failed the significance threshold ($q = 0.06$) and the FGF pathway was still detected, albeit with a clearly unacceptable q-value ($q = 0.22$). Whether or not this

argues for the relaxing of significance thresholds in the case of such a large window scan analysis, the introduction of some pruning / filtering method for the final list of genes which significantly change their expression, or suggests that the sheer number of probes / genes making it in to the final lists of the window scan results does not mesh best with enrichment tools such as DAVID remains to be determined in future work. These points are discussed in the summary of research outcomes in section 4.3.14.

A summary of the numbers of genes found to be significantly upregulated or downregulated (in the case of differential expression analyses) or those found to significantly change their mean expression during a window scan (for the window scanning approach) can be found, along with the numbers of pathway enrichments that these genes generated in figure 4.34. Further, a final summary result of these analyses is available displaying the most interesting enrichments found by all 4 differential expression analyses and all 4 window scanning analyses is provided in figures 4.35 and 4.36. This can be used to much more easily visualise the gist of the results discussed in the text.

	Diff expression across whole matrix	Diff Expression Early to Late naïve	Diff expression naïve to primed	Diff expression primed to early diff
Number of Genes UP	1141	44	2254	106
Number of Genes DOWN	870	61	874	385
Number of genes changing	2011	105	3128	491
UP enrichments	62	0	153	5
DOWN enrichments	40	6	4	199
Total enrichments	102	6	157	204

	Window scan across whole matrix	Window scan Early to Late naïve	Window scan naïve to primed	Window scan primed to early diff
Number of genes changing	6460	1212	4796	1665
Total enrichments	335	187	231	196

Figure 4.34: Summary table of all 8 analyses (4 differential expression, 4 scanning-window) of the Klf4-ordered HPM matrix, showing totals for genes found upregulated, downregulated (in the case of differential expression), changed expression (in the case of window-scanning), and the numbers of pathway enrichments these analyses generated.

	Diff expression high vs low Klf4	Diff expression early to late naïve	Diff expression naïve to primed	Diff expression primed to early diff		Window scan across whole matrix	Window scan early to late naïve	Window scan naïve to primed	Window scan primed to early diff
Control of gene expression pathway enrichments	Regulation of Transcription		Regulation of Transcription	Regulation of Transcription		Regulation of transcription	Regulation of transcription	Regulation of transcription	Regulation of Transcription
	Regulation of proliferation	Regulation of proliferation	Regulation of proliferation	Regulation of proliferation		Regulation of proliferation	Regulation of proliferation	Regulation of proliferation	Regulation of proliferation
	Neg reg. Differentiation	Pos reg. gene expression				Post-transcrip. reg gene expression		Post-transcrip. reg gene expression	
		Pos. reg transcription	Chromatin modification / organisation			Chromatin modification / organisation		Chromatin modification / organisation	
Stem cell / differentiation pathway enrichments			Regulation of differentiation	Regulation of differentiation		Regulation of differentiation	Regulation of differentiation	Regulation of differentiation	Regulation of differentiation
						Stress response		Stress response	
						Stem cell differentiation		Stem cell differentiation	
						Stem cell development		Stem cell development	
						Stem cell maintenance		Stem cell maintenance	

Figure 4.35: Table summarising biological pathway enrichment findings of all 8 analyses (4 differential expression, 4 scanning-window) on the Klf4-ordered HPM matrix. Table 1 of 2.

	Diff expression high vs low Klf4	Diff expression early to late naïve	Diff expression naïve to primed	Diff expression primed to early diff	Window scan across whole matrix	Window scan early to late naïve	Window scan naïve to primed	Window scan primed to early diff
Developmental pathway enrichments	Some general development		Many developmental processes	Many developmental processes	Many developmental processes	Many development al processes	Many development al processes	Many development al processes
	Vasculature development				Epithelial to mesenchymal transition			
	Nerve development							
Signalling pathway enrichments	Wnt		Wnt	Wnt	Wnt		Wnt	Wnt
			VEGF	VEGF	VEGF		VEGF	VEGF
				BMP	BMP			BMP
			MAPK		MAPK	FGF	MAPK	
			Notch		PDGF		PDGF	
					TGFβ			
Apoptosis-related pathway enrichments			Neg reg cell death	Neg reg cell death	Cell Death	Cell Death	Cell Death	Cell Death

Figure 4.36: Table summarising biological pathway enrichment findings of all 8 analyses (4 differential expression, 4 window-scan) on the Klf4-ordered HPM matrix. Table 2 of 2.

4.3.14 Summary of Research Outcomes

This chapter demonstrates that the amount of information lost by selecting samples for only the highest levels of Nanog, Oct4 and Sox2 was significantly higher than if the same number of samples had been filtered at random, indicating that Oct4, Sox2, Nanog filtering has not removed heterogeneity, but has likely reduced the types of samples that remain to fewer than originally captured in the larger matrix, N3312. Further, the remaining information was proven to include information relevant to pluripotency, as assessed by the existence of relevant relationships between pluripotency factors Oct4, Sox2 and Nanog to other genes, resulting in pathway enrichments strongly associated with pluripotency / differentiation.

A novel approach was taken to ordering these high-pluripotency-marker (HPM) samples from naïve pluripotency to primed pluripotency and beyond, identifying Klf4 as an ordering gene with known pluripotency function and scoring 8th overall among all genes with a novel scoring system of (absolute correlation to Nanog x normalised Shannon entropy.) Ordering by Klf4 sorted the samples appropriately, observed as appropriate, progressive changes in canonical naïve and primed pluripotency markers FGF5, Brachyury (T) and Rex1 (Zfp42). Further confirmation was demonstrated by cross-referencing samples from across the newly-ordered matrix with their experimental annotation, showing the annotated state of pluripotency / priming / differentiation was overwhelmingly in agreement with the ordering by Klf4. Thus an overall transcriptional progression between naïve pluripotency to early differentiation was achieved and, furthermore, samples from the same experiment and/or laboratory were found split across the Klf4 ordering by their pluripotency status, which allayed concerns that perhaps laboratory and cell line confounding, found to be unfortunately present in the data in the course of chapter 3

The choice of correlation-to-Nanog as part of the multiplicative scoring method used in this part of the work does not mean that the choice of Klf4 is simply a surrogate for Nanog, as it is demonstrated in this chapter that ordering the data by Nanog would not have so smoothly sorted Fgf5, Brachyury (T) and Rex1 either, while another gene given a high multiplicative score (Jam2) was shown to do so, demonstrating the utility of this multiplicative scoring method and that therefore the good ordering performance of Klf4 was not simply by chance.

With the utility of the data confirmed, analysis was then carried out using differential expression of groups (n = 50) of consecutive samples selected from representative areas of the Klf4-ordered

matrix. Four areas were chosen, being the highest-Klf4 samples (referred to as “early naïve pluripotency”), 50 samples taken from just before the switch from $Fgf5_{low}$, $Brachyury_{low}$, $Rex1_{high}$, representing samples that were deemed to be in a “late naïve pluripotenct” state, an area of 50 samples taken from the highest Fgf5, Brachyury (T) part of the Klf4-spectrum, representing “primed” (EpiSC) pluripotency, and then a final 50 samples chosen from the last Klf4 plateau to represent the “earliest differentiation” available in this data while OSN factors are still high.

These differential expression analyses identified genes that change their expression between these different points along the Klf4-spectrum and biological pathways that these genes enrich for. These pathways included significant enrichments for developmental, transcriptional, proliferative, chromatin-related, stress response, stem-cell related, differentiation-related, posttranscriptional-regulative and 8 defined signalling pathways, all of which were entered into the summary figures of the chapter for clarity.

This work then went on to calibrate a scanning window approach to discovery of enriched pathways across the Klf4-ordered matrix. Having devised the method, thresholds for the scanning window were investigated in order to remove noise but retain sensitivity. The window width decided upon which performed in this manner was a window of 25 consecutive samples.

This window scanning method was then used to carry out comparisons between the same cellular states as were done by differential expression. An astounding improvement in the number of genes found to change their expression significantly between early and naïve late pluripotency was found, with a concomitant massive increase in the number of biological pathways found to be enriched across this area of the data, including an enrichment for the FGF signalling pathway, amongst others. The window scanning method is in near-total agreement with the differential expression analyses in all other analysed parts of the matrix, and this method is therefore recommended by this work as an excellent approach to analysing transcriptional datasets which have, as here, been ordered as a biological phenomenon of interest takes place.

Bringing the results of this work into the context of current knowledge, the most important result of this chapter is the addressing of the current lack of detailed knowledge of transcriptional events which drive naïve pluripotent mESCs toward primed pluripotency. To that end, both the ordered dataset seeks to fill this gap in knowledge, especially in that the analysis of it returned a putative set of markers was identified between “early” and “late” naïve pluripotency. These may now be used to

define / investigate the changes between the early and late naïve pluripotent states immediately prior to the change to primed pluripotency. A comprehensive list of these putative marker genes is provided and genes unknown for any clear function in mESC pluripotency at present are highlighted. Those non-highlighted genes are not to be ignored, however, as the presence of so many mESC-pluripotency-related genes is very reassuring to see here and lends credence to the notion that the analysis across this point of the data is, in fact, showing *bona fide* transcriptional events that take place in mESCs. Selected genes clearly showing the divergence of upregulated and downregulated genes between early and late naïve pluripotency given as a heatmap and simplified bar plot.

This work therefore makes several contributions to the area of mESC biology. Firstly, this work has assembled the largest to-date set of pluripotent (HPM) mESC microarrays and provided detailed annotation of them. Another contribution is the method of verifiably, meaningfully ordering this matrix to investigate progression from naïve through primed pluripotency.

Next, there is demonstration that the *in silico* methods detailed herein can be successfully used to probe for transcriptional events and novel cellular states existing between those documented in the literature, even when none of the experiments making up the data matrix are specifically designed to identify or investigate these states, and originate from a plethora of different studies.

Further, there is no known mESC-related function of many of the candidate late / early naïve pluripotency markers found in this work. This work therefore contributes to the literature these candidate markers for further interrogation, ideally experimentally. These markers themselves may be linked to biological processes which can provide information on novel events which drive the changes between naïve and primed pluripotency (e.g. Car4 suggests that oxygen / bicarbonate levels may be involved). The ordered dataset itself is a result which brings the potential for investigation of ever more detailed events that take place during this critical transition process in mESCs; a task which will require future work and hopefully contribute many hypotheses for experimental validation.

As for those genes which were already known to the literature to be related directly to mESC pluripotency, this work now updates that knowledge by showing which of those genes can be used as sensitive markers of where along the transition between “early naïve”, “late naïve” and primed

pluripotency any given sample is, improving upon the use of only the current known markers, canonically being FGF5, Brachyury (T) and Rex1 (Zfp42).

In order to maximise the likelihood that what was predicted in this work *in silico* reflected biological reality, great effort was undertaken throughout this chapter to ensure that the expected events (read: expected naïve / primed marker changes, enrichment of relevant biological pathways) were detected across this ordered dataset, in strong agreement with what is known in the literature.

Further discussion of these results and future work to be carried out on the back of these findings to improve our knowledge of the transcriptional workings of mESCs, as well as to improve the methods put forward in this chapter, are given in future work sections 5.1.6 and 5.1.7.

Chapter 5 –

Discussion and Future Work

5.1.1 Overview

This final part of the thesis provides discussion of the most salient points arising from each chapter, while keeping the repetition of specific results to a minimum, as each previous chapter has appended to it its own summary of research outcomes. This section is given, therefore, to provide a very brief recap of the work in this thesis and for the author to criticise this work and identify areas for future work.

5.1.2 Chapter 2 Discussion

The assembly and annotation of the high pluripotency marker (HPM) matrix carried out in this work enabled the downstream work in this thesis in chapters 3 and 4. In addition, the generation of a matrix of mESC microarray data with full manual annotations of all samples is a valuable resource and outcome of this work in that many more analyses (some alluded to in the future work sections of this chapter) can be carried out on this data that were outwith the scope of this thesis.

Online annotations and often even the annotations in the accompanying literature were found to be wanting or frustrating for a variety of reasons, examples of which were discussed along with their relevance to researchers attempting to re-use and re-analyse public data. Particularly in the case of any attempts at automated retrieval and interpretation of available annotations, several key areas were identified as requiring attention. The identification of these key areas wherein occur the most potentially-deleterious effects on automated retrieval highlights the data fields and content to which time and effort on the part of uploaders could be most productively directed.

An annotation system which uses only plaintext characters was also developed and presented. This syntax can provide users of public data with quick reference to key information about sample cell line, origin, genetic modification, sorting and a chronology of exposure to detailed culture conditions.

There are some issues concerning this work that deserve comment here. The first of these is that the annotations were carried out by just one researcher; the author. It is therefore possible that some degree of human error has found its way into even this careful manual annotation of the HPM matrix, despite the repeat checking of these annotations by the author.

Secondly, the thresholds chosen for the filtering of the HPM matrix by high Oct/Sox/Nanog, were chosen by eye. Whilst this filtering did result in a useful dataset, as evidenced by the findings of chapters 3 and particularly of chapter 4, there remain questions as to which samples would have made it into the HPM matrix were these thresholds even marginally different. In retrospect, writing this discussion immediately prior to the submission of this thesis, other approaches than selection by eye could have been attempted, such as clustering of the data to identify high-pluripotency-marker samples may have gone some way to justifying and specifying thresholds which would be less open to questioning.

5.1.3 Chapter 2 Future work

There are several areas of work which the author would like to see undertaken that arise from the work presented in chapter 2, aside from any attempts to address the potential criticisms identified in the preceding discussion.

First among these would be the full annotation of the larger matrix, N3312, using the annotations of the HPM matrix as a starting point. This would identify non-mESC samples in that matrix and facilitate their removal. Downstream of this, analyses could begin to be carried out into identifying molecular signatures and transcriptional events that are associated with mESC-like samples that are not only highest in their expression for the pluripotency factors Oct4, Sox2 and Nanog, as was the selection criteria for inclusion in the HPM matrix. This would possibly involve the insertion of an additional field into the annotations for quick reference of cell type, as well as cell line. This would require a great deal of time, however, to make an educated judgement call on the state of individual samples.

Secondly, the HPM matrix and its accompanying annotations can be made available to the wider research community, enabling other groups with considerably more man-hours to contribute than the author had available to interrogate this data for other phenomena that lay outside the scope of one PhD thesis. Some investigations are immediately suggested by the work in this thesis, however, and are mentioned in the following future work sections of this chapter.

Third, a similar approach to the generation and annotation of mESC data as was detailed here can be carried out with human data. An advanced search on the GEO website at time of writing

(September 2015) returns 1,118 samples on the GPL570 platform (Affymetrix Human Genome U133 Plus 2.0 Array) which contain the string “embryonic stem cell”. The investigations carried out in later chapters 3 and 4 could then be carried out on this data, with particular emphasis and the author's own personal excitement concerning repeating the analyses from chapter 4 on human data to elucidate molecular signatures and investigation of biological pathway activities of hESCs.

5.1.4 Chapter 3 Discussion

The methods conceived in chapter 3 concern the comparison of contribution to sample similarity of two annotations in the HPM matrix assembled in chapter 2. Here, the annotations of “cell line” and “source laboratory” are evaluated for their apparent contribution to sample similarity in the HPM matrix and these quantified contributions are then compared, with the result that samples sharing a “source laboratory” are made more similar to each other than those which share a “cell line” annotation ($p < 0.05$, in the case of analysis using Euclidean distance as the similarity metric).

Previous work in mESCs has shown that source laboratories do indeed have their own, similar signature (Newman and Cooper 2010), while other work has demonstrated similarities arising from the choice of cell line in human iPS cells, proposed as being “memory” of the donor cell line (Marchetto et al. 2009). This exploratory analysis in this chapter of mESCs using the HPM matrix constitutes a first attempt at direct comparison between the effects of source laboratory and cell line on sample similarity in pluripotent mESC microarray data.

This chapter then goes on to propose a method for investigating the linkage between an annotation of interest (in the case of this work, “cell line”) and transcriptional signatures. The differences between mESC lines, and a group of miPSCs all generated by the same method; forced expression of the Yamanaka reprogramming factors Oct4, Sox2, Klf4 and c-Myc (Takahashi and Yamanaka 2006), are investigated using this proposed method. The resulting lists of genes found to be associated with these 4 different groups (ESD3, E14, CGR8 and the OSKM-iPSC group) were further analysed for biological pathway enrichment using the DAVID bioinformatics tool (Dennis et al. 2003).

The resulting enrichments reveal potential differences in those genes found to be comparatively upregulated or downregulated when compared to all other cell lines present in the HPM matrix, summarised in figure 3.16. Significant enrichment was found for signalling pathways known to be

involved in mESC biology, such as Wnt, Notch, MAPK/ERK and VEGF (figure 3.16 and see section 1.3.3 on signalling in mESCs). Differences were also found in expression of genes related to stress response, apoptosis-related and proliferation-related pathways. These initial results suggest that there may be phenotypic differences between these cell lines as regards (among others listed in figure 3.16 and in greater detail in sections 3.6 and 3.7) their endogeneous signalling activities, proliferative vigour and resilience to stressors. Interestingly, the OSKM-iPS group of samples was found to have comparative upregulation of cell cycle related genes, possibly suggesting that they had, during the course of their generation, been autoselected for a proliferative advantage.

Differences in entropy of genes associated with biological pathways are also reported here, although the implications of changing entropy of groups of genes between cell lines requires further investigation, it remains of interest that similar pathways emerged, being mostly concerned with signalling pathways, redox homeostasis and proliferation. As may have been expected, it was more likely to find, using DALGES, pathways whose constituent genes decreased in entropy, as opposed to increased, although there were some enrichments for the ESD3 cell line, which neatly brings about mention of the main criticisms the author has about chapter 3's work.

The first criticism of this chapter is the confounded nature of the annotations. The large amount of confounding between laboratory and cell line will likely be behind the similar results for contribution to sample similarity of the cell line and source laboratory annotations. The methodologies of both RaSToVa and DALGES were developed and tested concurrently with the annotation of the HPM matrix and took considerable time. It could therefore not be predicted that the (then future) analysis by RaSToVa would likely be so largely affected by this confounding. RaSToVa can only work with the data that it is given, and there was still a significant ($p < 0.05$, using Euclidean distance as the similarity metric) difference between the two annotations' contribution to sample similarity, with “source laboratory” being the stronger of the two. However, repeating this analysis with normalised Shannon entropy as the similarity metric did not meet this significance threshold, despite a tendency. Euclidean distance may simply be amplifying differences between samples, as Euclidean distance uses the square of differences between each probe. It is the author's suspicion that this is the case, as whilst there is general agreement between the boxplots for both Euclidean distance and Shannon entropy metrics being used by RaSToVa, the boxplots are considerably more spread out when using Euclidean distance. This leads the author to suspect that when comparing more pronouncedly-dissimilar samples, Euclidean distance assigns a

disproportionately higher number to that difference than Shannon entropy would when comparing those same samples.

DALGES was used in this chapter as a method proposed for the investigation of transcriptional profiles and their potential linkage to a cell line. It is tempting to declare that DALGES is wholly unaffected by the confounded nature of the annotations of cell line and laboratory, as DALGES is only comparing between cell lines. This freedom from confounding is not quite so, however. The author's major criticism of DALGES' investigation of transcriptional profiles of different cell lines is that if cell line can be considered at all to be simply a surrogate for a source laboratory, then there exists the likelihood that what DALGES is really finding and reporting to be a transcriptional profile of a cell line, is influenced more by the experiments / source laboratory that used that cell line, rather than a transcriptional signature of the cell line itself. Whilst this is not a fault at all of the methodology's conception or implementation, the results of the investigation into transcriptional profiles of cell lines must take this into consideration.

Fortunately, as can be seen in figure 3.24, however, the cell lines that were chosen for the analysis of transcriptional profile (CGR8, E14, ESD3 and OSKM_iPS group), have different levels of confounding with the “source laboratory” annotation. The E14 cell line is dispersed across different laboratories, as can be seen by the different shades of deep blue across multiple named laboratories. The CGR8 cell line also is mostly in the blue, with samples distributed across multiple (albeit less than with E14) laboratories. The ESD3 cell line, however, is considerably confounded with the “Piersma AH” laboratory. The OSKM_iPS group has some spread across laboratories, with the “Zhou Q” laboratory being more represented. This makes for an interesting take on the results of this part of chapter 3, in that the results of searching for cell-line specific transcriptional signatures may have found some success in the case of E14, CGR8 and even OSKM_iPS lines, in that these cell lines were used by multiple source laboratories. It is the results from the ESD3 analysis here that must be treated as highly likely confounded with the source laboratory, and therefore the transcriptional signature found by DALGES is likely to be heavily influenced by the experiments that make up the data. It is therefore interesting to see that, as can be seen in figure 3.16, only the ESD3 cell line analysis found genes related to biological pathways to increase their entropy in this cell line. This means that DALGES, when comparing ESD3 samples to non-ESD3 samples, found increased variability with genes associated with developmental pathways, and signalling pathways (VEGF, BMP and Wnt). It would normally be expected that if a cell line had a particular signature, pathway enrichments would result from those genes which were more predictable (decreased

entropy), rather than those that were less predictable (more entropy) within a given cell line. The question then becomes whether or not the increased entropy of genes in the ESD3 cell line is the result of the “Piersma AH” samples being highly varied, or whether this increased entropy signature would occur in other highly-confounded annotations. This is an open question which remains for future work.

The signatures from the CGR8 and E14 cell lines are far less confounded with source laboratory (see figure 3.24) and offer a first insight into potential differences between these two mESC cell lines. A suggestion is also given from this analysis that iPS cells generated using all four OSKM factors may be predisposed towards upregulation of cell cycle genes and a downregulation of Wnt-related genes. This is not, however, the same as decreased Wnt activity, as this would require pulling from the data the exact Wnt-related genes found to be downregulated. Even then, only an educated guess could be made as to whether this truly meant reduced Wnt signalling; the future testing of Wnt activity in these iPS samples would be far better served in the wet-lab, but would be of interest to the field of mESC biology.

As a final note on the ES-D3 cell line, it is interesting to note that it is the most confounded with “source laboratory” and also shows a movement of all probes, in figure 3.21 towards being less entropic, while the other plots for the less-confounded cell lines E14, CGR8 and the OSKM_iPS group are far closer to being centred on the 0,0 mark of the axes (see figures 3.21 and 3.22). It may be, therefore, that such a displacement may be a useful way of detecting issues of confounding when running DALGES in future analyses, although this requires more investigation to confirm.

An issue which should be raised regarding the RaSToVa and DALGES methods as described in this work is that they do not currently apply multiple hypothesis correction to their output p-values. This was a choice made by the author as not all analyses for which these methods are used will use groups of genes together. Whilst adjustment of the p-values is merited when using groups of these genes together to generate hypotheses, this was only done in this work when looking for biological pathway enrichment, which, given that DAVID was used, already applies the Benjamini-Hochberg q-value (Benjamini and Hochberg 1995) downstream. However, it can be argued that the grouped genes passed to the DAVID tool should have their p-values revised using such correction also, even with downstream multiple hypothesis correction being applied. Given the exploratory nature of these analyses, however, it was decided to report observations in this work without performing

multiple rounds of multiple hypothesis correction. Addition of this functionality is mentioned at the end of the following section.

5.1.5 Chapter 3 Future work

Some future work for chapter 3 is directly aimed at correcting for the aforementioned criticisms. This would most certainly include the re-running of RaSToVa on less-confounded data, annotated purely for cell line and source laboratory in a larger matrix, such as the full N3312 matrix. This is absolutely necessary before RaSToVa could contribute results to an acceptable standard of robustness.

Concerning the method of RaSToVa itself, rather than the unfortunate case of the confounded annotations, there is merit to be had in assessing the behaviour of the method even further than was done in this work. Specifically, even though there was only a little difference in the p-values between the outcomes of RaSToVa when using Euclidean distance and normalised Shannon entropy, these values frustratingly sat either side of the threshold of $p = 0.05$. It is therefore of interest to go back to the method and systematically observe calculation of both metrics (Euclidean distance and normalised Shannon entropy) and observe how much actual changes in individual data points between two samples affects the resulting Euclidean distance and Shannon entropy ratios. This would likely provide definitive explanation of the different spreads of the boxplots from chapter 3. Neither metric can be considered to be “right” or “wrong” as a result of this, however, but the most intricate understanding of their behaviour may influence the choice of one over the other in future work. Whilst development did involve testing both RaSToVa and DALGES on small, synthetic datasets, this was entirely for debugging purposes while the author was learning to script in R to develop these systems. With the attained experience of using R throughout the course of this work, future efforts could be undertaken quite rapidly to assess the behaviour of both of these metrics using much larger synthetic datasets with spiked-in differences.

Future work remains to be carried out using the DALGES methodology also, despite its early successes in investigating transcriptional signatures of E14, CGR8 cell lines and the OSKM_iPS group of samples. Extension of the method to larger data is a primary desire for future work, but also confirmation of the findings of DALGES in the wet-lab would provide vital confirmation of the utility of the method in linking *in silico* observations to *in vitro* truth. Also, re-running of DALGES on other confounded groups (as occurred with the ES-D3 group of samples), to observe

any possible repeat of the skewing of all probes towards a reduction in entropy, as was seen in figure 3.21 for the ESD3 plot, remains as future work to characterise this behaviour of DALGES. If this skewing occurs when data is too homogeneous for comparison to the rest of a dataset, then such a shift towards most probes showing entropy reduction would be a useful quality control method for DALGES results.

The development of an R package for the DALGES method is planned and in early stages as it could be useful for the wider research community, enabling other groups to investigate annotation-linked gene expression signatures by any annotation of choice, not just cell line. Concerning the continued use of DALGES in investigating mESC cell lines, however, future work would include more mouse ES cell lines, human ES cell lines, and the application of the same methodology to non-microarray data such as the explosion in the amount of RNAseq data being generated. In particular, future work comparing the signatures of iPSC lines generated by different methods and from different donor material could be an extremely useful contribution to stem cell research that DALGES may be able to fulfil. As was mentioned in the preceding discussion section, options for multiple hypothesis correction will be added to the DALGES method, so that it reports both p and q values for researchers to use, depending on which is appropriate.

5.1.6 Chapter 4 Discussion

The work in this chapter first seeks to take the HPM matrix and order it in a manner which allows for the broad sorting of all samples between naïve pluripotency and early differentiation. One of the potential pitfalls when attempting to do this comes from the facts which were the reasons behind much of the work in chapter 3, being that samples from the same experiment may well be so similar that ordering the matrix just orders experiments, rather than cellular states. It was therefore decided to not use any collection of genes to order the matrix, but to look for the possibility of using only one. It was the author's belief that if the expression level of many genes is slightly affected by samples being from the same experiment, then using only one gene should be far less prone to this. Indeed, genes which make samples from the same experiment more similar to each other need not necessarily even be genes of any real biological relevance to this work; those experimentally-related genes may be collections of housekeeping genes, unused probesets *et cetera*, although confirmation of this may form part of possible future work for chapter 3.

Choosing one gene to order the matrix was therefore undertaken through the use of four major selection criteria. The first was a large range of expression, being simply the difference between the maximum any minimum expression value found in the matrix. The second was a good “spread” of values between this minimum and maximum, which was assessed using the normalised Shannon entropy metric detailed in 3.2.4. The third was correlation to Nanog (Nanog was chosen over Oct4 and Sox2 as Nanog had higher information content (read: entropy) in the HPM matrix.) Fourth and finally, it would be best if this selected gene were to be a known pluripotency factor. To avoid lengthy repetition, these criteria are detailed, along with reasons for their choosing, in section 4.2.2.

Surprisingly, an excellent candidate for all 4 of these criteria was found in Klf4 (see section 4.3.3). The utility of ordering by this gene was confirmed in section 4.3.4, as Klf4 is able to broadly sort the samples of the HPM matrix between naïve pluripotency and early differentiation, with primed pluripotency markers occurring in the middle. Further, the markers chosen to observe progression from naïve to primed pluripotency (being Rex1, Fgf5 and Brachyury (T)) had only the one major change-point across the matrix, implying successful sorting. The annotations provided in chapter 2 were also cross-referenced with Klf4 ranks as the major way to confirm that, in addition to observation of marker profiles, experimental annotation supported the notion that Klf4 had broadly sorted the samples by their cellular state from naïve pluripotency, to primed pluripotency, to exit from primed pluripotency (see figures 4.16 to 4.18). It was also this cross-referencing with annotations that confirmed that using only a single gene for the sorting of the data did indeed prevent samples from different states of pluripotency grouping together simply by virtue of being from the same experiment (see figures 4.16 to 4.18 and section 4.3.4).

Coupled with the annotations generated in chapter 2, this Klf4-ordered HPM matrix makes for a highly-useful resource, which this chapter goes on to use in a first effort at identifying transcriptional changes between the states captured by the data, along with analysing these transcriptional changes for biological pathway enrichments. Two methods were used, the first being a “differential expression” approach between each identified area of interest (see figure 4.19). The second method was devised as a scanning window approach (detailed in section 4.2.5 and figure 4.1). The scanning window approach was then calibrated to this data to reduce the possible effects of noise (section 4.3.9).

The biological pathway enrichments found by the differential expression analyses and the scanning window analyses were in overwhelming agreement, as can be seen in the text and the many figures

depicting biological pathway enrichments, but neatly in the summary figures 4.35 and 4.36 . However, there was one striking difference between these two analysis methods. Whilst the differential expression analysis between “early naïve pluripotency” and “late naïve pluripotency” (both of these terms referring to areas of the Klf4-ordered HPM matrix, rather than a biological property of the samples referred to, without future investigation) identified hardly any (a total of 6) pathway enrichments, the window scanning method across the early to late naïve parts of the data returned a plethora of enrichments (n = 187), including a host of developmental pathways and, crucially, identifying changes in genes related to the FGF signalling pathway, known to be crucial to the progression to the primed state (Kunath et al. 2007) and also upstream of the MAPK/ERK signalling pathway, whose inhibition contributes to the maintenance of the very definition of the naïve, ground state (Ying et al. 2008). This is an extremely interesting outcome of this work in that the use of the window scanner across this part of the Klf4-ordered data was able to find so many pathway enrichments and genes significantly changing their expression, but that simply looking at a “start” and “end” point in such data may be missing out on a vast amount of information. This is also the first attempt at analysing, on a large scale, changes in gene expression / transcriptional profile which may be going on as the naïve / ground state progresses towards exit from that state towards primed pluripotency, as canonical markers of naïve and primed pluripotency (Rex1, Fgf5, Brachyury (T)) remained level throughout this part of the Klf4 spectrum. Other naïve / primed markers, such as Nr0b1, did change but were not plotted in the course of this work so as to avoid cluttering the plots, although their patterning can clearly be seen with a simple smoothed line plot across the HPM matrix, which is available on the accompanying DVD.

The differential expression analysis was not without its use, however, as the major advantage it has over the window-scanning method (in its current form, at least), is that it provides information on the directionality of the changes in gene expression which it finds. This allowed this work to observe the existence of changes in transcriptional profile as naïve pluripotency moves towards exit to primed pluripotency (see section 4.3.7), identifying markers of this state that appears to occur immediately prior to the exit from naïve pluripotency (see figures 4.28 4.29 for summary). This was cause for considerable excitement, as a great deal of future work can be spawned from this (see following section.)

The pathway enrichments found in this part of the work also included enrichments for genes involved in the stress response. Cellular reprogramming towards a cancer state has already been linked with the stressing of cells, such as by inflammation (Song and Balmain 2015) or nutrient

stress (Ma et al. 2013), and other stress, particularly hypoxia, has already been identified as being beneficial for the reprogramming of cells to iPS cells (Yoshida et al. 2009). This demonstrates that the approach taken in this work, even though a preliminary effort, not only finds signalling pathways which recapitulate knowledge about mESCs in the literature (e.g. Wnt, BMP, TGF, Notch, FGF), but also shows enrichment for pathways given less attention in the literature, such as PDGF and VEGF (see results sections of chapter 4 and also section 1.3.3 for details on signalling in mESCs).

Whilst this discussion has greatly truncated the findings from the analyses in chapter 4, the recapitulation of known phenomena in mESC biology from this data is highly exciting and opens the door to more analysis of this useful dataset and its annotations. See the following section on future work for more details on the directions of investigation warranted and enabled by the work in this chapter.

There are some areas of this work which the author wishes to comment on critically, however. Firstly, proof was offered in this work that the sorting by Klf4 was not simply a lucky choice, as other genes which scored highly on the multiplicative score method devised in 4.3.3, also ordered marker profiles in a similar (or sometimes smoother) manner to Klf4 (see 4.14 and 4.15 for the demonstration of ordering the matrix using Jam2, which scored more highly than Klf4.) Whilst this method proved useful in identifying a gene to order the HPM matrix by, evidence is not given in this chapter that ordering by other high-scoring genes would have performed just as well as (or better than) Klf4 when it comes to the downstream analyses of chapter 4. Repeat of this analysis using other genes from the high-multiplicative scoring genes would have been interesting, as this would confirm that other genes identified here would have performed the same (or better.) As it was not known at the time whether or not the choice of Klf4 would provide good results, no such alternative analyses were performed. Particularly, the use of a gene with good correlation to Nanog, but a much lower entropy, would reinforce the case for the utility of the multiplicative score system.

Secondly, there are potential issues to note about the window scanning method. Whilst differential expression through the use of mean expression is typical, the window scanning method was run on data that had been sorted in such a way as to broadly sort samples across markers of cellular states. It could therefore improve this work to attempt to identify and remove outliers as the window scanner progresses. Whilst it is unlikely that outliers in the window scanning method could, by chance, have given rise to a sufficient number of false significant changes in gene expression so as

to, again, by chance, provide enrichments for biological pathways that made a great deal of intuitive sense in the context of mESC biology, the question of outliers could be addressed through either a specific modification to the methodology of the window scanner, or perhaps more simply through the use of an alternative metric such as a median.

In addition to the window scanner's own inner workings, one potential issue was identified when observing the pathway enrichments that result from the window scanner's results. The summary figure 4.35 and 4.36 demonstrate that the full matrix window essentially found every biological pathway enrichment found by all other analyses, with the exception of FGF signalling and Notch signalling, as although these were found in the list of biological pathway enrichments, they did not achieve the required q-value of ($q <= 0.05$). However, without in-depth understanding of the exact methodology that DAVID employs, it remains possible that this lack of enrichment for these pathways could be simply due to the sheer number of probes that DAVID was given in the case of the full matrix scan (over 6000.) This is suspected by the author when considering that in the analysis of differentially-expressed genes between naïve pluripotency to primed pluripotency, the Notch signalling pathway has 14 probes associated with it, giving a q-value of 0.01. However, the full matrix scan found 22 Notch-related probes, yet DAVID returned a q-value of 0.06 for this pathway. It would make sense that DAVID takes into account the number of probes given to it, as, at the extreme, providing DAVID with every known gene would show enrichment for every single biological pathway, even though such a list would not be “enriched” for one pathway over another. The same issue may be responsible for the window scanner's failure to find enrichment for the FGF pathway across the whole matrix, when a window scan of early to late naïve pluripotency found 7 probes associated with FGF signalling to be deserving of a q-value of 0.04, while the full matrix scan's improvement on 7 probes by finding 13 probes associated with FGF signalling was only worthy of a q-value of 0.22. Regardless, the findings of the analyses in chapter 4 are overwhelmingly in agreement with both the literature and with each other, demonstrating the utility of both methods, but these observations imply that more work need be done before either method should be generalised to larger datasets, or perhaps the limiting of the window scanner to use only across a certain number of genes that change expression at a time.

Finally, and with the same reasoning behind its mention in the discussion and future work sections of chapter 3, multiple hypothesis correction was not applied to groups of genes passed to the DAVID bioinformatics tool for pathway enrichment as DAVID applies the Benjaminj-Hochberg multiple hypothesis correction (Benjamini and Hochberg 1995). The application of another round of

multiple hypothesis correction prior to this correction may be overly conservative in the case of the initial hypotheses generated in this work, although an option for this is to be added to the window scanner's functionality.

5.1.7 Chapter 4 Future work

First and foremost, for future work, are the exciting directions that this work points to for in-laboratory investigation. These suggestions for in-laboratory work were a major reason behind undertaking the work in this chapter (and thesis as a whole) in the first place. With a marker profile now available of cellular states which may pertain to “early” and “late” naïve pluripotency, a host of experiments to confirm the existence of these cellular states become warranted. Confirmation of a progression from early to late naïve pluripotency is warranted, most simply by observing the predicted changes in marker profiles.

Investigation can also be undertaken through the use of techniques such as conditional knockouts of markers for early / late naïve pluripotency to observe any effects on the progression from “early” to “late” naïve pluripotency. siRNA experiments could be carried out against sets of these markers to test the possibility of reversion from late to early naïve pluripotency. Finely-tuned inhibition or stabilisation signalling pathways named in figures 4.35 and 4.36 may be able to do this. Alternatively, it would be of great interest if it were found that something altogether different was driving this change from “early” to “late” naïve pluripotency. As the overwhelming majority of the annotations for the HPM matrix detail the use of some form of serum (read: non-2i conditions), it would be also interesting to directly compare these transcriptional profiles of “early” and “late” naïve pluripotency with cells that were cultured in 2i. Perhaps 2i cultured cells exist at one end or in the middle of these two putative states identified in this work?

The stress response genes identified in this work are also deserving of in-laboratory investigation. Experiments can be devised wherein these genes are experimentally manipulated and the effect on cellular proliferation / survival can be ascertained, but also it would be of interest to observe any effect on cells' tendency or, in fact, ability to differentiate / remain undifferentiated. The author suspects that there would indeed be effects on cells' fate decisions when manipulating these stress response genes, as cellular stress is already known to be relevant to cellular reprogramming (Ma et al. 2013), (Song and Balmain 2015), as well as improving reprogramming in iPS cells (Yoshida et

al. 2009). This may lead to both improved understanding of mESC biology, but also to improved methods for iPS generation / mESC maintenance and manipulation.

The creation of the Klf4-ordered HPM matrix also begs its use in more focussed bioinformatic analyses, such as data mining for correlated genes. Indeed, every list of pathway enrichments, every list of differentially expressed genes in this work has the potential for further investigation. So many analyses are beyond the scope of one person, particularly as these lists are likely to pique the interest of stem cell biologists working in other subsections of the field who may recognise patterns or single genes of interest to themselves. Therefore, the Klf4-ordered HPM matrix and its annotations are in the process of being prepared for publication so that other researchers might be able to mine this data generally, but also so that researchers with more specialised questions, but unfortunately a lack of pluripotent mESC data, might ask those questions of this matrix. Another benefit for other researchers of the use of this fully-annotated, ordered matrix is the ability to observe where their own would sort within it, or simply to compare their own mESC / iPSC data to it. By downloading the same files which make up this matrix (possible through using the accession numbers in the annotations file provided), the same processing can be carried out (running RMA on all files, including those contributed by the aforementioned researcher), followed by Klf4-ordering. With full annotations available, researchers can compare their own data directly to samples from a wealth of other samples from varied laboratories, cell lines and culture conditions.

The window scanner itself also has potential for improvement and expansion, in addition to adding options new metrics and / or parameter-based outlier elimination. The window scanner as used in this work did not provide indication as to a gene's direction of change at any given changepoint. Towards the end of this work, beginning efforts were underway to add to the window scanner the ability to use directionality information so that pathway enrichment analyses could again be given directionality similar to those generated by differential expression results. The publication of the window scanner as an R package for use by other biological researchers, complete with a vignette explaining its behaviour, is planned as a future outcome of this work also.

Another possibility that arises from this work is the potential to analyse the changing expression of genes from the point of view of ChIP targets. Analysis is already underway using the window scanner results, observing the chronology of significant changes in gene expression and looking for enrichment of known transcription factors which may be behind these grouped changes in expression, with some initial hypotheses already generated.

Possibly the largest bioinformatic effort that the author intends to undertake following on from the findings of this work is to take the improved methods from this work and apply them to a large dataset of curated human ES cell data, whether that data be generated by microarray, or, possibly more excitingly, using the explosion in the amounts of RNAseq data being generated. After all, it is the eventual translation of knowledge from mESCs, miPSCs to human ES cells and, finally, that knowledge's application to human disease, lifespan and healthspan that all embryonic stem cell research ultimately promises.

Appendix A
RaSToVA and DALGES pseudocode scripts

RaSToVa pseudocode script summary

Load data matrix
Load annotations file
Get unique annotations of interest (e.g. all different cell lines in the matrix)
Loop for each unique annotation (e.g. each cell line):

{

Retrieve samples from matrix matching annotation
Calculate variability metric of annotation-intact matrix
Loop for the total number of permutations:

{

Resample matrix using any samples that do not match current annotation of interest
Calculate variability of randomly-resampled matrix
Express random matrix's variability as index of intact matrix's variability
Store this index alongside that unique annotation (e.g. that specific cell line)

}

}

Return a list of all unique annotations (e.g. cell lines) and their calculated (random matrix variability / intact matrix variability) scores

Plot these as separate boxplots, named by annotation (e.g. one for each cell line)

DALGES pseudocode script summary

Load data matrix
Load annotations file
Get unique annotations of interest (e.g. all different cell lines in the matrix)
Loop for each unique annotation (e.g. each cell line):

{

Retrieve samples from matrix matching annotation (e.g. ES-D3 line)
Calculate mean expression value for all probes
Loop for the total number of permutations:

{

Resample matrix using any samples that do not match current annotation of interest
Calculate mean expression values for all probes in randomly-resampled matrix
Calculate differential expression between intact and random matrix for all probes
Store all differential expression changes for this permutation
Store higher than intact / lower than intact flags for all probes

}

Use higher than / lower than intact flags to calculate significance per probe
Store results for all probes' differential-expression-to-random, with significance

}

Return a list of all unique annotations (e.g. cell lines), each one having a list of all probes with their differential-expression-compared-to-random fold changes and significances.

Write out data matrix of these values for all annotations (e.g. cell lines) and append gene names to probe IDs

List of Abbreviations

Acronym	Meaning
ChIP	Chromatin immunoprecipitation
DALGES	Discovery of Annotation-Linked Gene Expression Signatures
DAVID	Database for Annotation Visualisation and Integrated Discovery
DMEM	Dulbecco's Modified Eagle's Medium
DMSO	Dimethylsulphoxide
DOX	Doxycycline
EB	Embryoid Body
EMT	Epithelial-Mesenchymal Transition
EpiSC	Epi-Stem Cell
ES	Embryonic Stem
ESC	Embryonic Stem Cell
FBS	Fetal Bovine Serum
GEO	Gene Expression Omnibus
GO	Gene Ontology
hESC	human Embryonic Stem Cell
hiPSC	human Induced Pluripotent Stem Cell
HPM	High Pluripotency Marker
ICM	Inner Cell Mass
iPSC	induced Pluripotent Stem Cell
KOSR / KSR	Knockout Serum Replacement
mESC	mouse Embryonic Stem Cell
NEAA	Non-Essential Amino Acids
OHT	4-Hydrotamoxifen
OSKM	Oct4/Sox2/Klf4/cMyc
OSN	Oct4/Sox2/Nanog
RaSToVa	Random Submatrix Total Variability
RMA	Robust Multichip Average
TE	Trophectoderm

Bibliography

Acampora D, di Giovannantonio LG, Simeone A (2013), "Otx2 is an intrinsic determinant of the embryonic stem cell state and is required for transition to a stable epiblast stem cell condition", Development, 140, pp43-55

Amano T, Hirata T, Falco G, Monti M, Sharova LV, Amano M, Sheer S, Hoang H, Piao Y, Stagg CA, Yamamizu K (2013), "Zscan4 restores the developmental potency of embryonic stem cells", Nature Communications, 4

Androutsellis-Theotokis A, Leker RR, Soldner F, Hoeppner DJ, Ravin R, Poser SW, Rueger MA, Bae SK, Kittappa R, McKay RDG (2006), "Notch signalling regulates stem cell numbers in vitro and in vivo", Nature, 442, pp832-826

Anokye-Danso F, Trivedi CM, Juhr D, Gupta M, Cui Z, Tian Y, Zhang Y, Yang W, Gruber PJ, Epstein JA, Morrisey EE (2011), "Highly efficient miRNA-mediated reprogramming of mouse and human somatic cells to pluripotency", Cell Stem Cell, 8, pp376-388

Avilion AA, Nicolis SK, Pevny LH, Perez L, Vivian N, Lovell-Badge R (), "Multipotent cell lineages in early mouse development depend on Sox2 function", Genes & Development, 126;17, pp126-140

Bagga S, Bracht J, Hunter S, Massirer K, Holtz J, Eachus R, Pasquinelli AE (2005), "Regulation by let-7 and lin-4 miRNAs results in mRNA degradation", Cell, 122, pp553-563

Bain G, Kitchens D, Yao M, Huettner JE, Gottlieb DI (1995), "Embryonic stem cells express neuronal properties in vitro", Developmental Biology, 168, pp342-357

Bao S, Tang F, Li X, Hayashi K, Gillich A, Lao K, Surani MA (2009), "Epigenetic reversion of post-implantaiton epiblast to plruipotent embryonic stem cells", Nature, 461, pp1292-1295

Bellin M, Marchetto MC, Gage FH, Mummery CL (2012), "Induced pluripotent stem cells: the new patient?", Nature Reviews Molecular Cell Biology, 13, pp713-726

Ben-Shushan E, Thompson JR, Gudas LJ, Bergman Y (1998), "Rex-1, a gene encoding a transcription factor expressed in the early embryo, is regulated via Oct3/4 and Oct-6 binding to an octamer site and a novel protein, Rox-1, binding to an adjacent site", Molecular and Cellular Biology, 18, pp1866-1878

Benjamini Y and Hochberg Y (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing", Journal of the Royal Statistical Society (Series B), 57;1, pp289-300

Brons IGM, Smithers LE, Trotter MWB, Rugg-Gunn P, Sun B, Lopes SMCS, Howlett SK, Clarkson A, Ahrlund-Richter L, Pedersen RA, Vallier L (2007), "Derivation of pluripotent epiblast stem cells from mammalian embryos", Nature, 448, pp191-195

Brown PO, Botstein D (1999), "Exploring the new world of the genome with DNA microarrays", Nature Genetics Supplement, 21, pp33-37

Burdon T, Smith A, Savatier P (2002), "Signalling, cell cycle and pluripotency in embryonic stem cells", TRENDS in Cell Biology, 12;9, pp432-438

Caillier M, Thenot S, Tribollet V, Birot AM, Samarut J, Mey A (2010), "Role of the epigenetic regulator HP1-gamma in the control of embryonic stem cell properties", Public Library of Science ONE, 5;11, pp

Cartwright P, McLean C, Sheppard A, Rivett D, Jones K, Dalton S (2004), "LIF/STAT3 controls ES cell self-renewal and pluripotency by a Myc-dependent mechanism", Development, 132;5, pp885-896

Celso CL, Prowse DM, Watt FM (2004), "Transient activation of B-catenin signalling in adult mouse epidermis is sufficient to induce new hair follicles but continuous activation is required to maintain hair follicle tumours", Development, 131;8, pp1781-1799

Chambers I, Colby D, Robertson M, Nichols J, Lee S, Tweedie S, Smith A (2003), "Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells", Cell, 113, pp643-655

Chambers I, Silve J, Colby D, Nichols J, Nijmeijer B, Robertson M, Vrana J, Jones K, Grotewold L, Smith A (2007), "Nanog safeguards pluripotency and mediates germline development", Nature, 450, pp1230-1235

Chambers I, Tomlinson SR (2009), "The transcriptional foundation of pluripotency", Development, 136, pp2311-2322

Chazaud C, Yamanaka Y, Pawson T, Rossant J (2006), "Early Lineage Segregation between Epiblast and Primitive Endoderm in Mouse Blastocysts through the Grb2-MAPK Pathway", Developmental Cell, 10, pp615-624

Chen C, Ridzon D, Lee CT, Blake J, Sun Y, Strauss WM (2007), "Defining embryonic stem cell identity using differentiation-related microRNAs and their potential targets", Mammalian Genome, 18, pp316-327

Chen J, Liu J, Han Q, Qin D, Xu J, Chen Y, Yang J, Song H, Yang D, Peng M, He W, Li R, Wang H, Gan Y, Ding K, Zeng L, Lai L, Esteban MA, Pei D (2010), "Towards an Optimized Culture Medium for the Generation of Mouse Induced Pluripotenct Stem Cells", The Journal of Biological Chemistry, 285;40, pp31066-31072

Chen X, Johns DC, Geiman DE, Marban E, Dang DT, Hamlin G, Sun R, Yang VW (2001), "Kruppel-like factor 4 (Gut-enriched Kruppel-like factor) inhibits cell proliferation by blocking G1/S progression of the cell cycle", The Journal of Biological Chemistry, 276;32, pp30423-30428

Chen X, Whitney EM, Gao SY, Yang VW (2003), "Transcriptional profiling of Kruppel-like Factor 4 reveals a function in cell cycle regulation and epithelial differentiation", The Journal of Molecular Biology, 326, pp665-677

Chew JL, Loh YH, Zhang W, Chen X, Tam WL, Yeap LS, Li P, Ang YS, Lim B, Robson P, Ng HH (2005), "Reciprocal transcriptional regulation of Oct4 and Sox2 via the Oct4-Sox2 complex in embryonic stem cells", Molecular and Cellular Biology, 25;14, pp6031-3046

Clevers H (2011), "The cancer stem cell: premises, promises and challenges", Nature Medicine, 17;3, pp313-316

Cole MF, Johnstone SE, Newman JJ, Kagey MH, Young RA (2008), "Tcf3 is an integral component of the core regulatory circuitry of embryonic stem cells", Genes and Development, 22, pp746-755

Conboy IM, Rando TA (2005), "Aging, Stem Cells and Tissue Regeneration", Cell Cycle, 4;3, pp407-410

Costa Y, Ding J, Theunissen TW, Faiola F, Hore TA, Shliaha PV, Fidalgo M, Saunders A, Lawrence M, Dietmann S, Das S, Levasseur DN, Li Z, Xu M, Reik W, Silva JCR, Wang J (2013), "NANOG-dependent function of Tet1 and Tet2 in establishment of pluripotency", Nature, 495, pp370-347

Dani C, Smith AG, Dessolin S, Leroy P, Staccini L, Villageois P, Darimont C, Ailhaud G (1997), "Differentiation of embryonic stem cells into adipocytes in vitro", Journal of Cell Science, 110, pp1279-1285

Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA (2003), "DAVID: Database for Annotation, Visualization, and Integrated Discovery", Genome Biology, 4;9, pp

Doetschman TC, Eistetter H, Katz M, Schmidt W, Kemler R (1985), "The in vitro development of blastocyst-derived embryonic stem cell lines: formation of visceral yolk sac, blood islands and myocardium", Journal of Embryology and Experimental Morphology, 87, pp27-45

Dreesen O & Brivanlou AH (2007), "signalling pathways in cancer and embryonic stem cells", Stem Cell Reviews, 3, pp7-17

Edgar R, Domrachev M, Lash AE (2002), "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository", Nucleic Acids Research, 30;1, pp207-210

Efroni S, Dttagupta R, Cheng J, Dehghani H, Hoeppner DJ, Dash C, Bazett-Jones DP, Le Grice S, McKay RDG, Buetow KH, Gingeras TR, Misteli T, Meshorer E (2008), "Global transcription in pluripotent embryonic stem cells", Cell Stem Cell, 2, pp437-447

Ekins R & Chu FW (1999), "Microarrays: their origins and applications", TIBTECH, 17, pp217-218

Endoh M, Endo TA, Endoh T, Fujimura YI, Ohara O, Toyoda T, Otte AP, Okano M, Brockdorff N, Vidal M, Koseki H (2008), "Polycomb group proteins Ring1A/B are functionally linked to the core transcriptional regulatory circuitry to maintain ES cell identity", Development, 135;8, pp1513-1529

Evans MJ, Kaufman MH (1981), "Establishment in culture of pluripotential cells from mouse embryos", Nature, 292, pp154-156

Feng B, Jiang J, Kraus P, Ng JH, Heng JCD, Chan YS, Yae LP, Zhang W, Loh YH, Han J, Vega VB, Cacheux-Rataboul V, Lim B, Lufkin T, Ng HH (2009), "Reprogramming of fibroblasts into induced pluripotent stem cells with orphan nuclear receptor Esrrb", Nature Cell Biology, 11;2, pp197-203

Festuccia N, Osomo R, Halbritter F, Karwacki-Neisius V, Navarro P, Colby D, Wong F, Yates A, Tomlinson SR, Chambers I (2012), "Esrrb is a direct Nanog target gene that can substitute for Nanog function in Pluripotent Cells", Cell Stem Cell, 11, pp477-490

Galan-Caridad JM, Harel S, Arenzana TL, Hou ZE, Doetsch FK, Mirny LA, Reizis B (2007), "Zfx controls the self-renewal of embryonic and hematopoietic stem cells", Cell, 129, pp345-357

Galonska C, Ziller MJ, Kamik R, Meissner A (2015), "Ground state conditions induce rapid reorganization of core pluripotency factor binding before global epigenetic reprogramming", Cell Stem Cell, 17, pp1-9

Geijsen N (2012), "Epigenetic reprogramming: Prdm14 hits the accelerator", The European Molecular Biology Organisation Journal, 31, pp2247-2248

Gill JG, Langer EM, Lindsley C, Cai M, Murphy TL, Kyba M, Murphy KM (2011), "Snail and the microRNA-200 family act in opposition to regulate epithelial-to-mesenchymal transition and germ layer fate restriction in differentiation ESC", Stem Cells, 29, pp764-776

Gillich A, Bao S, Grabole N, Hayashi K, Trotter MWB, Pasque V, Magnusdottir E, Surani MA (2012), "Epiblast stem cell-based system reveals reprogramming synergy of germline factors", Cell Stem Cell, 10, pp425-439

Grabole N, Tischler J, Hackett JA, Kim S, Tang F, Leitch HG, Magnusdottir E, Surani MA (2013), "Prdm14 promotes germline fate and naïve pluripotency by repressing FGF signalling and DNA methylation", European Molecular Biology Organisation Reports, 14;7, pp629-637

Gu P, Goodwin B, Chung ACK, Xu X, Wheeler DA, Price RR, Galardi C, Peng L, Latour AM, Koller BH, Gossen J, Kliewer SA, Cooney AJ (2005), "Orphan Nuclear Receptor LRH-1 Is Required To Maintain Oct4 Expression at the Epiblast Stage of Embryonic Development", Molecular and Cellular Biology, 25;9, pp3492-3505

Guo G, Smith A (2010), "A genome-wide screen in EpiSCs identifies Nr5a nuclear receptors as potent inducers of ground state pluripotency", Development, 137;19, pp3185-3192

Guo G, Yang J, Nichols J, Hall JS, Eyres I, Mansfield W, Smith A (2009), "Klf4 reverts developmentally programmed restriction of ground state pluripotency", Development, 136, pp1063-1069

Gygi SP, Rochon Y, Franza BR, Aebersold R (1999), "Correlation between protein and mRNA abundance in yeast", Molecular and Cellular Biology, 19;3, pp1720-1730

Hackett JA, Surani MA (2014), "Regulatory principles of pluripotency - from the ground state up", Cell Stem Cell, 15, pp416-430

Hagan JP, Piskounova E, Gregory RI (2009), "Lin28 recruits the TUTase Zcchc11 to inhibit let-7 maturation in mouse embroynic stem cells", Nature Structural & Molecular Biology, 16;10, pp1021-1026

Hall J, Guo G, Wray J, Eyres I, Nichols J, Grotewold L, Morfopoulou S, Humphreys P, Mansfield W, Walker R, Tomlinson S, Smith A (2009), "Oct4 and LIF/Stat3 Additively Induce Kruppel Factors to Sustain Embryonic Stem Cell Self-Renewal", Cell, 5;6, pp597-609

Hamilton WB, Kaji K, Kunath T (2013), "ERK2 suppresses self-renewal capacity of embryonic stem cells, but is not required for multi-lineage commitment", Public Library of Science One, 8;4

Hanna J, Markoulaki S, Mitalipova M, Cheng AW, Cassady JP, Staerk J, Carey BW, Lengner CJ, Foreman R, Love J, Gao Q, Kim J, Jaenisch R (2009), "Metastable pluripotent states in NOD-mouse-derived mESCs", Cell Stem Cell, 4, pp513-524

Hanna LA, Foreman RK, Tarasenko IA, Kessler DS, Labosky PA (2002), "Requirement for Foxd3 in maintaining pluripotent cells of the early mouse embryo", Genes & Development, 16, pp2650-2661

Hao J, Li TG, Qi X, Zhao DF, Zhao GQ (2006), "WNT/B-catenin pathway up-regulates Stat3 and converges on LIF to prevent differentiation of mouse embryonic stem cells", Developmental Biology, 290, pp81-91

Hayashi K, de Sousa Lopes SMC, Tang F, Surani MA (2008), "Dynamic equilibrium and heterogeneity of mouse pluripotent stem cells with distinct functional and epigenetic states", Cell Stem Cell, 3, pp391-401
>
Hosler BA, La Rosa GJ, Grippo JF, Gudas LJ (1989), "Expression of Rex-1, a gene containing zinc finger motifs, is rapidly reduced by retinoic acid in f9 teratocarcinoma cells", Molecular and Cellular Biology, 9;12, pp5623-5629

Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, Gottardo R, Hahne F, Hansen KD, Irizarry RA, Lawrence M, Love MI, MacDonald J, Obenchain V, Oles AK, Pages H, Reyes A, Shannon P, Smyth GK, Tenenbaum D, Waldron L, Morgan M (2015), "Orchestrating high-throughput genommic analysis with Bioconductor", Nature Methods, 12;2, pp115-121

Inoue H and Yamanaka S (2011), "The use of induced pluripotent stem cells in drug development", Clinical Pharmacology & Therapeutics, 89;5, pp655-661

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003), "Exploration, normalization, and summaries of high density oligonucleotide array probe level data", Biostatistics, 4;2, pp249-264

Ito M, Yang , Andl T, Cui C, Kim N, Millar S, Costarelis G (2007), "", Nature, 447

Ivanova N, Dobrin R, Lu Rong, Kotenko I, Levorse J, DeCoste C, Schafer X, Lun Y, Lemischka IH (2006), "Dissecting self-renewal in stem cells with RNA interference", Nature, 442, pp533-538

Jeyapalan JC, Ferreira M, Sedivy JM, Herbig U (2007), "Accumulation of senescent cells in mitotic tissue of aging primates", Mechanisms of Ageing and Development, 128, pp36-44

Jian J, Chan YS, Loh YH, Cai J, Tong GQ, Lim CA, Robson P, Zhong S, Ng HH (2008), "A core Klf4 circuitry regulates self-renewal of embryonic stem cells", Nature Cell Biology, 10;3, pp353-360

Kanellopoulou C, Muljo SA, Kung AL, Ganesan S, Drapkin R, Jenuwein T, Livingston DM, Rajewsky K (2005), "Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing", Genes and Development, 19, pp489-501

Karwacki-Neisius V, Goeke J, Osomo R, Halbritter F, Ng JH, Weisse AY, Wong FCK, Gagliardi A, Mullin NP, Festuccia N, Colby D, Tomlinson SR, Ng HH, Chambers I (2013), "Reduced Oct4 Expression Directs a Robust Pluripotent State with Distinct signalling Activity and Increased Enhancer Occupancy by Oct4 and Nanog", Cell Stem Cell, 12, pp531-545

Kim D, Kim CH, Moon JI, Chung YG, Chang MY, Han BS, Ko S, Yang E, Cha KY, Lanza R, Kim KS (2009), "Generation of human induced pluripotent stem cells by direct delivery of reprogramming proteins", Cell Stem Cell, 4;6, pp472-476

Kim J, Woo AJ, Chu J, Snow JW, Fujiwara Y, Kim CG, Cantor AB, Orkin SH (2010), "A Myc Network Accounts for Similarities between Embryonic Stem and Cancer Cell Transcription Programs", Cell, 143, pp313-324

Koh KP, Yabuuchi A, Rao S, Huang Y, Cunniff K, Nardone J, Laiho A, Tahiliani M, Sommer CA, Mostoslavsky G, Lahesmaa R, Orkin SH, Rodig SJ, Daley GQ, Rao A (2011), "Tet1 and Tet2 regulate 5-hydrozymethylcytosine production and cell lineage specification in mouse embryonic stem cells", Cell Stem Cell, 8, pp200-213

Kosaka N, Sakamoto H, Terada M, Ochiya T (2009), "Pleiotropic function of FGF-4: Its role in development and stem cells", Developmental Dynamics, 238, pp265-276

Kunath T, Saba-El-Leil MK, Almousailleakh M, Wray J, Meloche S, Smith A (2007), "FGF stimulation of the Erk1/2 signalling cascade triggers transition of pluripotent embryonic stem cells from self-renewal to lineage commitment", Development, 134, pp2895-2902

Kurek D, Neagu A, Tastemel M, Tuysuz N, Lehmann J, van de Werken HJG, Philipsen S, van der Linden R, Maas A, van Ijcken WFJ, Drukker M, ten Berge D (2015), "Endogenous WNT signals mediate BMP-induced and spontaneous differentiation of epiblast stem cells and human embryonic stem cells", Stem Cell Reports, 4, pp114-128

Lanner F and Rossant J (2010), "The role of FGF/ERK signalling in pluripotent cells", Development, 137, pp3351-3360

Lanner F, Lee KL, Sohl M, Holmborn K, Yang H, Wilbertz J, Poellinger L, Rossant J, Farnebo F (2010), "Heparan sulfation-dependent fibroblast growth factor signalling maintains embryonic stem cells primed for differentiation in a heterogeneous state", Stem Cells, 28, pp191-200

Lattin JE, Schroder K, Su AI, Walker JR, Zhang J, Wiltshire T, Saijo K, Glass CK, Hume DA, Kellie S, Sweet MJ (2008), "Expression analysis of G protein-coupled receptors in mouse macrophages", Immunome Research, 4;5

Lazar C, Meganck S, Taminau J, Steenhoff D, Coletta A, Molter C, Weiss-Solis DY, Duque R, Bersini H, Nowe A (2012), "Batch effect removal methods for microarray gene expression data integration: a survey", Briefings in Bioinformatics, 14, pp469-490

Lee KL, Lim SK, Orlov YL, Yit LY, Yang H, Ang LT, Poellinger L, Lim B (2011), "Graded Nodal/Activin signalling Titrates Conversion of Quantitative Phospho-Smad2 Levels into Qualitative Embryonic Stem Cell Fate Decisions", Public Library of Science Genetics, 7;6

Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA (2010), "Tackling the widespread and critical impact of batch effects in high-throughput data", Nature Reviews Genetics, 11, pp733-739

Li Y, McClintock J, Zhong L, Edenberg HJ, Yoder MC, Chan RJ (2005), "Murine embryonic stem cell differentiation is promoted by SOCS-3 and inhibited by the zinc finger transcription factor Klf4", Blood, 105;2, pp635-637

Li Z, Fei T, Zhang J, Zhu G, Wang L, Lu D, Chi X, Teng Y, Hou N, Yang X, Zhang H, Han JDJ, Chen YG (2012), "BMP4 signalling Acts via Dual-Specificity Phosphatase 9 to Control ERK Activity in Mouse Embryonic Stem Cells", Cell Stem Cell, 10;3, pp171-182

Lin CY, Loven J, Rahl PB, Paranal RM, Burge CB, Bradner JE, Lee TI, Young RA (2012), "Transcriptional Ambplification in Tumor Cells with Elevated c-Myc", Cell, 141, pp56-67

Lowell S, Benchoua A, Heavey B, Smith AG (2006), "Notch promotes neural lineage entry by pluripotent embryonic stem cells", Public Library of Science (PLoS) Biology, 4;5

Ma L, Tao Y, Duran A, Llado V, Galvez A, Barger JF, Castilla EA, Chen J, Yajima T, Porollo A, Medvedovic M, Brill LM, Plas DR, Riedl SJ, Leitges M, Diaz-Meco MT, Richardsom AD, Moscat J (2013), "Control of nutrient stress-induced metabolic reprogramming by PKC in tumorigenesis", Cell, 152, pp599-611

Ma Z, Swigut T, Valouev A, Rada-Iglesias A, Wysocka J (2011), "Sequence-specific regulator Prdm14 safeguards mouse ESCs from entering extraembryonic endoderm fates", Nature Structural & Molecular Biology, 18;2, pp120-127

Marchetto MCN, Yeo GW, Kainohana O, Marsala M, Gage FH, Muotri AR (2009), "Transcriptional signature and memory retention of human-induced pluripotent stem cells", Public Library of Science ONE, 4;9

Marks H, Kalkan T, Menefra R, Denissov S, Jones K, Hofemeister H, Nichols J, Kranz A, Stewart AF, Smith A, Stunnenberg HG (2012), "The transcriptional and epigenomic foundations of ground state pluripotency", Cell, 149, pp590-604

Marson A, Levine SS, Cole MF, Frampton GM, Brambink T, Johnstone S, Guenther MG, Johnston WK, Wernig M, Newman J, Calabrese JM, Dennis LM, Volkert TL, Gupta S, Love J, Hannett N,

Sharp PA, Bartel DP, Jaenisch R, Young RA (2008), "Connecting microRNA Genes to the Core Transcriptional Regulatory Circuitry of Embryonic Stem Cells", Cell, 134, pp521-533

Martello G, Sugimoto T, Diamanti E, Joshi A, Hannah R, Ohtsuka S, Goettgens B, Niwa H, Smith A (2012), "Esrrb is a pivotal target of the Gsk3/Tcf3 axis regulating embryonic stem cell self-renewal", Cell Stem Cell, 11, pp491-504

Martin GR (1981), "Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells", Proceedings of the National Academy of Sciences - Developmental Biology, 78;12, pp7634-7638

Masaki H, Nishida T, Kitajima S, Asahina K, Teraoka H (2007), "Developmental pluripotency-associated 4 (DPPA4) localized in active chromatin inhibits mouse embryonic stem cell differentiation into a primitive ectoderm lineage", The Journal of Biological Chemistry, 282;45, pp33034-33042
>
Masui S, Nakatake Y, Toyooka Y, Shimosato D, Yagi R, Takahashi K, Okochi H, Okuda A, Matoba R, Sharov AA, Ko MSH, Niwa H (2007), "Pluripotency governed by Sox2 via regulation of Oct3-4 expression in mouse embryonic stem cells", Nature Cell Biology, 9;6, pp625-635

Matsuda T, Nakamura T, Nakao K, Arai T, Katsuki M, Heike T, Yokota T (1999), "STAT3 activation is sufficient to maintain an undiffrentiated state of mouse embryonic stem cells", The European Molecular Biology Journal, 18;15, pp4261-4269

Melton C, Judson RL, Blelloch R (2010), "Opposing microRNA families regulate self-renewal in mouse embryonic stem cells", Nature, 463, pp621-626

Miller RA, Christoforou N, Pevsner, McCallion AS, Gearhart JD (2008), "Efficient Array-Based Identification of Novel Cardiac Genes through Differentiation of Mouse ESCs - PLoS One", Public Library of Science (PLOS) One, 3;5

Mitsui K, Tokuzawa Y, Itoh H, Segawa K, Murakami M, Takahashi K, Maruyama M, Maeda M, Yamanaka S (2003), "The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells", Cell, 113, pp631-642

Miura K, Okada Y, Aoi T, Okada A, Takahashi K, Okita K, Nakagawa M, Koyanagi M, Tanabe K, Ohnuki M, Ogawa D, Ikeda E, Okano H, Yamanaka S (2009), "Variation in the safety of induced pluripotent stem cell lines", Nature Biotechnology, 27;8, pp743-745

Morey L, Santanach A, di Croce L (2015), "Pluripotency and epigenetic factors in mouse embryonic stem cell fate regulation", Molecular and Cellular Biology, 35;16, pp2716-2728

Moss EG, Lee RC, Ambros V (1997), "The cold shock domain protein LIN-28 controls developmental timing in C. elegans and is reulated by the lin-4 RNA", Cell, 88, pp637-646

Munoz-Descalzo S, Hadjantonakis AK, Arias AM (2015), "Wnt/B-catenin signalling and the dynamics of fate decisions in early mouse embryos and embryonic stem (ES) cells", Seminars in Cell & Developmental Biology - article in press

Murayama H, Masaki H, Sato H, Hayama T, Yamaguchi T, Nakauchi H (2015), "Successful reprogramming of Epiblast stem cells by blocking nuclear localization of B-catenin", Stem Cell Reports, 4, pp103-113

Nakaki F, Saitou M (2014), "PRDM14: a unique regulator for pluripotency and epigenetic reprogramming", Trends in Biochemical Sciences, 39;6, pp289-296

Navarro P, Festuccia N, Colby D, Gagliardi A, Mullin NP, hang W, Karwacki-Neisius V, Osorno R, Kelly D, Robertson M, Chambers I (2012), "OCT4/SOX2-independent Nanog autorepression modulates heterogeneous Nanog gene expression in mouse ES cells", The European Molecular Biology Organisation Journal, 31, pp4547-4562

Navarro P, Oldfield A, Legoupi J, Festuccia N, Dubois A, Attia M, Schoorlemmer J, Rougelle C, Chambers I, Avner P (2010), "Molecular coupling of Tsix regulation and pluripotency", Nature, 468, pp457-460

Newman AM & Cooper JB (2010), "Lab-Specific Gene Expression Signatures in Pluripotent Stem Cells", Cell Stem Cell, 7, pp258-262

Newman AM, Cooper JB (2010), "Lab-specific gene expression signatures in pluripotent stem cells", Cell Stem Cell, 7, pp258-262

Nichols J and Smith A (2009), "naïve and primed pluripotent states", Cell Stem Cell, 4, pp487-492

Nichols J and Smith A (2012), "Pluripotency in the Embryo and in Culture", Cold Spring Harbour Perspectives in Biology, 4

Nichols J, Zevnik B, Konstantinos A, Niwa H, Klewe-Nebenius D, Chambers I, Schoeler H, Smith A (1998), "Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4", Cell, 95, pp379-391

Nie Z, Hu G, Wei G, Cui K, Yamane A, Resch W, Wang R, Green DF, Tessarollo L, Casellas R, Zhao K, Levens D (2012), "c-Myc Is a Universal Amplifier of Expressed Genes in Lymphocytes and Embryonic Stem Cells", Cell, 151, pp68-79

Nikolova-Krstevski V, Bhasin M, Otu HH, Libermann T, Oettgen P (2008), "Gene expression analysis of embryonic stem cells expressing VE-cadherin (CD144) during endothelial differentiation", BMC Genomics, 9

Nishikawa S, Goldstein RA, Nierras CR (2008), "The promise of human induced pluripotent stem cells for research and therapy", Nature Reviews Molecular Cell Biology, 9, pp725-729

Niwa H, Burdon T, Chambers I, Smith A (1998), "Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3", Genes & Development, 12, pp2048-2060

Niwa H, Miyazaki J, Smith AG (), "Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells", Nature Genetics, 24, pp372-376

Niwa H, Ogawa K, Shimosato D, Adachi K (2009), "A parallel circuit of LIF signalling pathways maintains pluripotency of mouse ES cells", Nature Letters, 460, pp118-122

Niwa H, Toyooka Y, Shimosato D, Strumpf D, Takahashi K, Yagi R, Rossant J (2005), "Interaction between Oct3/4 and Cdx2 Determines Trophectoderm Differentiation", Cell, 123, pp917-929

Ogawa K, Nishinakamura R, Iwamatsu Y, Shimosate D, Niwa H (2006), "Synergistic action of Wnt and LIF in maintaining pluripotency of mouse ES cells", Biochemical and Biophysical Research Communications, 343, pp159-166

Okita K, Ichisaka T, Yamanaka S (2007), "Generation of germline-competent induced pluripotent stem cells", Nature, 448, pp313-318

Okita K, Nakagawa M, Hyenjong H, Ichisaka T, Yamanaka S (2008), "Generation of mouse induced pluripotent stem cells without viral vectors", Science, 322, pp949-953

Paling NRD, Wheadon H, Bone HK, Welham MJ (2004), "Regulation of embryonic stem cell self-renewal by phosphoinositide 3-kinase-dependent signalling", The Journal of Biological Chemistry, 279;46, pp48063-48070

Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M, Mani R, Rayner T, Sharma A, William E, Srkans U, Brazma A (2007), "ArrayExpress - a public database of microarray experiments and gene expression profiles", Nucleic Acids Research, 35, pp747-750

Payer B, Rosenberg M, Yamaji M, Yabuta Y, Koyanagi-Aoi M, Hayashi K, Yamanaka S, Saitou M, Lee JT (2013), "Tsix RNA and the germline factor, PRDM14, lin X-reactivation and stem cell reprogramming", Molecular Cell, 52;6, pp805-818

Pebay A, Wong RCB, Pitson SM, Wolvetang EJ, Peh GSL, Filipczyk A, Koh KLL, Tellis I, Nguyen LTV, Pera MF (2005), "Essential roles of sphingosine-1-phosphate and platelet-derived growth factor in the maintenance of human embryonic stem cells", Stem Cells, 23, pp1541-1548

Pelton TA, Sharma S, Schulz TC, Rathjen J, Rathjen PD (2002), "Transient pluripotent cell populations during primitive ectoderm formation - correlation of in vivo and in vitro pluripotent cell development", Journal of Cell Science, 115;2, pp329-339

Pereira L, Yi F, Merrill BJ (2006), "Repression of Nanog Gene Transcription by Tcf3 Limits Embryonic Stem Cell Self-Renewal", Molecular and Cellular Biology, 26;20, pp7479-7491

Polyak K and Weinberg RA (2009), "Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits", Nature Reviews Cancer, 9, pp265-272

Qi X, Li TG, Hao J, Hu J, Wang J, Simmons H, Miura S, Mishina Y, Zhao GQ (2004), "BMP4 supports self-renewal of embryonic stem cells by inhibiting mitogen-activated protein kinase pathways", Proceedings of the National Academy of Sciences, 101;16, pp6027-6032

Quakenbush J (2001), "Computational Analysis of Microarray Data", Nature Reviews Genetics, 2, pp418-427

Rahl PB, Lin C, Seila AC, Flynn RA, McCuine S, Burge CB, Sharp PA, Young RA (2010), "c-Myc Regulates Transcriptional Pause Release", Cell, 141, pp432-445

Ramasamy A, Mondry A, Holmes CC, Altman DG (2008), "Key issues in conducting a meta-analysis of gene expression microarray datasets", Public Library of Science Medicine, 5;9

Rana TM (2007), "Illuminating the silence: understanding the structure and funciton of small RNAs", Nature Molecular Cell Biology, 8, pp23-36

Reya T & Clevers H (2005), "Wnt signalling pathways in stem cells and cancer", Nature, 434, pp843-850

Robinton DA, Daley GQ (2012), "The promise of induced pluripotent stem cells in research and therapy", Nature, 481, pp295-305

Rogers MB, Hosler BA, Gudas LJ (1991), "Specific expression of a retinoic acid-regulated, zinf-finger gene, Rex-1, in preimplantation embryos, trophoblast and spermatocytes", Development, 113, pp815-824

Rothenberg ME, Clarke MF, Diehn M (2010), "The Myc Connection: ES Cells and Cancer", Cell, 143, pp184-186

Rowland BD, Bernards R, Peeper DS (2005), "The KLF4 tumour suppressor is a transcriptional repressor of p53 that acts as a context-dependent oncogene", Nature Cell Biology, 7;11, pp1074-1082

Sakaki-Yumoto M, Katsuno Y, Derynck R (2013), "TGFβ family signalling in stem cells", Biochimica et Biophysica Acta, 1830, pp2280-2296

Sampath P, Pritchard DK, Reinecke H, Schwartz SM, Morris DR, Murry CE (2008), "A hierarchical network controls protein translation during murine embryonic stem cell self-renewal and differentiation", Cell Stem Cell, 2, pp448-460

Saretzki G, Armstrong L, Leake A, Lako M, von Zglinicki T (2004), "Stress defense in murine embryonic stem cells is superior to that of various differentiated murine cells", Stem Cells, 22, pp962-971

Sato N, Meijer L, Skaltsounis L, Greengard P, Brivanlou AH (2004), "Maintenance of pluripotency in human and mouse embryonic stem cells through activation of Wnt signalling by a pharmacological GSK-3-specific inhibitor", Nature Medicine, 10;1, pp55-63

Sato N, Sanjuan IM, Heke M, Uchida M, Naef F, Brivanlou AH (2003), "Molecular signature of human embryonic stem cells and its comparison with the mouse", Developmental Biology, 260, pp404-413

Scatena R, Bottoni P, Pontoglio A, Giardina B (2011), "Cancer stem cells: the development of new cancer therapeutics", Expert Opinion on Biological Therapy, 11;7, pp875-892

Schneider-Gaedicke A, Beer-Romero P, Brown LG, Mardon G, Luoh SW, Page DC (1989), "Putative transcription factor with alternative isoforms encoded by human ZFX gene", Nature, 342, pp708-711

Schulz H, Kolde R, Adler P, Aksoy I, Anastassiadis K, Bader M, Billon N, Boeuf H, Bourillot PY, Bucholz F, Dani C, Doss MX, Forrester L, Gitton M, Henrique D, Hescheler J, Himmelbauer H, Hubner N, Karantzali E, Krestovali A, Lubitz S, Pradier L, Rai M, Reimand J, Rolletschek A, Sachinidis A, Savatier P, Stewart F, Storm MP, Trouillas M, Vilo J, Welham MJ, Winkler J, Wobus AM, Hatzopoulos AK (2009), "The FunGenES Database: A Genomics Resource for Mouse Embryonic Stem Cell Differentiation", Public Library of Science One, 4;9

Scotland KB, Chen S, Sylvester R, Gudas LJ (2009), "Analysis of Rex1 (Zfp42) unction in embryonic stem cell differentiation", Developmental Dynamics, 238, pp1863-1877

Sekkai D, Gruel G, Herry M, Moucadel V, Constantinescu SN, Albagla O, Roux DTL, Vainchenker W, Bennaceur-Griscelli A (2005), "Microarray analysis of LIF/Stat3 transcriptional targets in embryonic stem cells", Stem Cells, 23, pp1634-1642

Sene KH, Porter CJ, Palidwor G, Perez-Iratxeta C, Muro EM, Campbell PA, Rudniki MA, Andrade-Navarro MA (2007), "Gene function in early mouse embryonic stem cell differentiation", BioMed Central Genomics, 8;85

Shannon CE (1948), "A mathematical theory of communication", Mobile Computing and Communications Review, 5;1, pp3-55

Shen X, Liu Y, Hsu YJ, Fujiwara Kim J, Mao X, Yuan GC, Orkin SH (2008), "EZH1 mediates methylation on histone H3 lysine 27 and complements EZH2 in maintaining stem cell identity and executing pluripotency", Molecular Cell, 32;4

Shi W, Wang H, Pan G, Geng Y, Guo Y, Pei D (2006), "Regulation of the pluripotency marker Rex-1 by Nanog and Six2", The Journal of Biological Chemistry, 281;33, pp23319-23325

Shimizu N, Yamamoto K, Obi S, Kumagaya S, Masumura T, Shimano Y, Naruse K, Yamashita JK, Igarashi T, Ando J (2008), "Cyclic strain induces mouse embryonic stem cell differentiation into vascular smooth muscle cells by activating PDGF receptor B", Journal of Applied Physiology, 104, pp766-772

Silva J, Nichols J, Theunissen TW, Guo G, van Oosten AL, Barrandon O, Wray J, Yamanaka S, Chambers I, Smith A (2009), "Nanog is the gateway to the pluripotent ground state", Cell, 138, pp722-737

Singh R, Shen W, Kuai D, Martin JM, Guo X, Smith MA, Perez ET, Phillips MJ, Simonett JM, Wallace KA, Verhoeven AD, Capowski EE, Zhang X, Yin Y, Halbach PJ, Fishman GA, Wright LS, Pattnaik BR, Gamm DM (2012), "iPS cell modeling of Best disease: insights into the pathophysiology of an inherited macular degeneration", Human Molecular Genetics

Singla DK, Schneider DJ, LeWinter MM, Sobel BE (2006), "Wnt3a but not Wnt11 supports self-renewal of embryonic stem cells", Biochemical and Biophysical Research Communications, 345, pp789-795

Smith AG, Heath JK, Donaldson DD, Wong GG, Moreau J, Stahl M, Rogers D (1988), "Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides", 336, pp688-690

Sokol SY (2011), "Maintaining embryonic stem cell pluripotency with Wnt signalling", Development, 138, pp4341-4350

Song IY and Balmain A (2015), "Cellular reprogramming in skin cancer", Seminars in Cancer Biology, 32, pp32-39

Stadtfeld M, Nagaya M, Utikal J, Weir G, Hochedlinger K (2008), "Induced pluripotent stem cells generated without viral integration", Science, 322, pp945-949

Storm MP, Bone HK, Beck CG, Bourillot PY, Schreiber V, Damiano T, Nelson A, Savatier P, Welham MJ (2007), "Regulation of Nanog Expression by Phosphotinositide 3-Kinase-dependent signalling in Murin Embryonic Stem Cells", Journal of Biological Chemistry, 282;9, pp6265-6273

Strumpf D, Mao CA, Yamanaka Y, Ralston A, Chawengsaksophak K, Beck F, Rossant J (2005), "Cdx2 is required for correct cell fate specification and differentiation of trophectoderm in the mouse blastocyst", Development, 132;9, pp2093-2102

Sutton J, Costa R, Klug M, Field L, Xu D, Largaespada DA, Fletcher CF, Jenkins NA, Copeland NG, Klemsz M, Hromas R (1996), "Genesis, a winged helix transcriptional repressor with expression restricted to embryonic stem cells", The Journal Of Biological Chemistry, 271;38, pp23126-23133

Takahashi K and Yamanaka S (2006), "Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors", Cell, 126, pp663-676

Takao Y, Yokota T, Koide H (2007), "B-catenin up-regulates Nanog expression through interaction with Oct-3/4 in embryonic stem cells", Biochemical and Biophysical Research Communications, 353, pp699-705

Tanaka A, Woltjen K, Miyake K, Hotta A, Ikeya M, Yamamoto T, Nishino T, Shoji E, Sehara-Fujisawa A, Manabe Y, Fujii N, Hanaoka K, Era T, Yamashita S, Isobe KI, Kimura E, Sakurai H (2013), "Efficient and reproducible myogenic differentiation from human iPS cells: prospects for modeling Miyoshi myopathy in vitro", Public Library of Science ONE, 8;4

Taylor DA (2009), "From stem cells and cadaveric matrix to engineered organs", Current Opinion in Biotechnology, 20, pp598-605

Team, R. C. (2015). "R: A language and environment for statistical computing." Vienna, Austria; 2014. URL http://www. R-project. org.

Tee WW, Shen SS, Oksuz O, Narendra V, Reinberg D (2014), "Erk1/2 activity promotes chromatin features and RNAPII phosphorylation at development promoters in mouse ESCs", Cell, 156, pp678-690

ten Berge D, Kurek D, Blauwkamp T, Koole W, Maas A, Eroglu E, Siu RK, Nusse R (2011), "Embryonic stem cells require Wnt proteins to prevent differentiation to epiblast stem cells", Nature Cell Biology, 13;9, pp1070-1077

Tesar PJ, Chenoweth JG, Brook FA, Davies TJ, Evans EP, Mack DL, Gardner RL, McKay RDG (2007), "New cell lines from mouse epiblast share defining features with human embryonic stem cells", Nature Letters, 448, pp196-202

Tiscornia G, Vivas EL, Belmonte JCI (2011), "Diseases in a dish: modeling human genetic disorders using induced pluripotent cells", Nature Medicine, 17, pp1570-1576

Tower J (2012), "Stress and stem cells", Wiley Interdisciplinary Reviews: Developmental Biology, 1;6, pp789-802

Turksen K and Troy TC (2001), "Claudin-6: A Novel Tight Junction Molecule is Developmentall Regulated in Mouse Embryonic Epithelium", Developmental Dynamics, 222, pp292-300

Turner BM (2008), "Open chromatin and hypertranscription in embryonic stem cells", Cell Stem Cell408-409

Vallier L, Mendjan S, Brown S, Chng Z, Teo A, Smithers LE, Trotter MWB, Cho CHH, Martinez A, Rugg-Gunn P, Brons G, Pedersen RA (2009), "Activin/Nodal signalling maintains pluripotency by controlling Nanog expression", Development, 136, pp1339-1349

van Dartel DAM, Pennings JLA, de la Fonteyne LJJ, Brauers KJJ, Claessen S, van Delft JH, Kleinjans JCS, Piersma AH (2011), "Evaluation of developmental toxicant identification using gene expression profiling in embryonic stem cell differentiation cultures", Toxicological Sciences, 119;1, pp126-134

van den Berg DLC, Snoek T, Mullin NP, Yates A, Bezstarosti K, Demmers J, Chambers I, Poot RA (2010), "An Oct4-Centred Protein Interaction Network in Embryonic Stem Cells", Cell Stem Cell, 6, pp369-381

van Es JH, Barker N, Clevers H (2003), "You Wnt some, you lose some: oncogenes in the Wnt signalling pathway", Current Opinion in Genetics & Development, 13, pp28-33

Viswanathan SR, Daley GQ, Gregory RI (2008), "Selective Blockade of MicroRNA Processing by Lin28", Science, 320, pp97-100

Wada KI, Itoga K, Okano T, Yonemura S, Sasaki H (2011), "Hippo pathway regulation be cell morphology and stress fibres", Development, 138, pp3907-3914

Wang Y, Medvid R, Melton C, Jaenisch R, Blelloch R (2007), "DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal", Nature Genetics, 39;3, pp380-385

Warzecha CC, Sato TK, Nabet B, Hogenesch JB, Cartsens RP (2009), "ESRP1 and ESRP2 are epithelial cell-type-specific regulators of FGFR2 splicing", Molecular Cell, 33, pp591-601

Watabe T & Miyazono K (2009), "Roles of TGFβ family signalling in stem cell renewal and differentiation", Cell Research, 19, pp103-115

Watanabe S, Umehara H, Murayama K, Okabe M, Kimura T, Nakano T (2006), "Activation of Akt signalling is sufficient to maintain pluripotency in mouse and primate embryonic stem cells", Oncogene, 25, pp2697-2707

Welham MJ, Kingham E, Sanchez-Ripoll Y, Kumpfmueller B, Storm M, Bone H (2011), "Controlling embryonic stem cell proliferation and pluripotency - the role of PI3K and GSK-3 dependent signalling", Biochemical Society Transactions, 39;2, pp674-678
>
Williams RL, Hilton DJ, Pease S, Willson TA, Stewart CL, Gearing DP, Wagner EF, Metcalf D, Nicola NA, Gough NM (1988), "Myeloid leukaemia inhibitory factor maintains the developmental potential of embryonic stem cells", Nature, 336, pp684-687

Wray J, Kalkan T, Gomez-Lopez S, Eckardt D, Cook A, Kemler R, Smith A (2011), "Inhibition of glycogen snythase kinase-3 alleviates Tcf3 repression of the pluripotency network and increases embryonic stem cell resistance to differentiation", Nature Cell Biology, 13;7, pp838-845

Xu B, Zhang K, Huang Y (2009), "Lin28 modulates cell growth and associates with a subset of cell cycle regulator mRNAs in mouse embryonic stem cells", RNA, 15, pp357-361

Yamaji M, Ueda J, Hayashi K, Ohta H, Yabuta Y, Kurimoto K, Nakato R, Yamada Y, Shirahige K, Saitou M (2013), "PRDM14 ensures naïve pluripotency through dual regulation of signalling and epigenetic pathways in mouse embryonic stem cells", Cell Stem Cell, 12, pp368-382

Yamanaka Y, Lanner F, Rossant J (2010), "FGF signal-dependent segregation of primitive endodermand epiblast in the mouse blastocyst", Development, 137, pp715-724

Yang J, Chai L, Fowles TC, Alipio Z, Xu Dan, Fink LM, Ward DC, Yupo M (2008), "Genome-wide analysis reveals Sall4 to be a major regulator of pluripotency in murine-embryonic stem cells", Proceedings of the National Academy of Sciences, 105;50, pp19756–19761

Ye S, Li P, Tong C, Ying QL (2013), "Embryonic stem cell self-renewal pathways converge on the transcription factor Tfcp2l1", The European Molecular BIology Organisation Journal, 32, pp2548-2560

Yeo JC and Ng HH (2013), "The transcriptional regulation of pluripotency", Cell Research, 23;1, pp20-32

Yeo JC, J J, Tan ZY, Yim GR, Ng JH, Goeke J, Kraus P, Liang H, Gonzales KAU, Chong HC, Tan CP, Lim YS, Tan NS, Lufkin T, Ng HH (2014), "Klf4 is an essential factor that sustains ground state pluripotency", Cell Stem Cell, 14, pp864-872

Yi R, Fuchs E (2011), "MicroRNAs and their roles in mammalian stem cells", Journal of Cell Science, 124;11, pp1775-1783

Ying QL, Nichols J, Chambers I, Smith A (2003), "BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3", Cell, 115, pp281-292

Ying QL, Smith AG (2003), "Defined conditions for neural commitment and differentiation", Methods in Enzymology, 365, pp327-341

Ying QL, Wray J, Nichols J, Batlle-Morera L, Doble B, Woodgett J, Cohen P, Smith A (2008), "The ground state of embryonic stem cell self-renewal", Nature, 453, pp519-523

Yoshida Y and Yamanaka S (2010), "Recent stem cell advances: induced pluripotent stem cells for disease modeling and stem cell-based regeneration", Circulation, 122, pp80-87

Yoshida Y, Takahashi K, Okita K, Ichisaka T, Yamanaka S (2009), "Hypoxia enhances the generation of induced pluripotent stem cells", Cell Stem Cell, 5;3, pp237-241

Young RA (2011), "Control of the embryonic stem cell state", Cell, 144, pp940-954

Yuan H, Corbi N, Basilico C, Dailey L (1995), "Developmental-specific activity of the FGF-4 enhancer requires the synergistic action of Sox2 and Oct-3", Genes & Development, 9, pp2635-2645

Zalzman M, Falco G, Sharova LV, Nishiyama A, Thomas M, Lee SL, Stagg CA, Hoang HG, Yang HT, Indig FE, Wersto RP, Ko MSH (2010), "Zscan4 regulates telomere elongation and genomic stability in ES cells", Nature, 464, pp858-863

Zhang J, Tam WL, Tong GQ, Wu Q, Chan HY, Soh BS, Lou Y, Yang J, Ma Y, Chai L, Ng HH, Lufkin T, Robson P, Lim B (2006), "Sall4 modulates embryonic stem cell pluripotency and early embryonic development by the transcriptional regulation of Oct4", Nature Cell Biology, 8;10, pp1114-1123