THE UNIVERSITY *of* EDINBURGH

Genome-Scale Transcriptomic and Epigenomic Analysis of Stem Cells

Florian Halbritter

Thesis submitted for the degree of Doctor of Philosophy
The University of Edinburgh
2012

Abstract

Embryonic stem cells (ESCs) are a special type of cell marked by two key properties: The capacity to create an unlimited number of identical copies of themselves (self-renewal) and the ability to give rise to differentiated progeny that can contribute to all tissues of the adult body (pluripotency). Decades of past research have identified many of the genetic determinants of the state of these cells, such as the transcription factors *Pou5f1*, *Sox2* and *Nanog*. Many other transcription factors and, more recently, epigenetic determinants like histone modifications, have been implicated in the establishment, maintenance and loss of pluripotent stem cell identity.

The study of these regulators has been boosted by technological advances in the field of high-throughput sequencing (HTS) that have made it possible to investigate the binding and modification of many proteins on a genome-wide level, resulting in an explosion of the amount of genomic data available to researchers. The challenge is now to effectively use these data and to integrate the manifold measurements into coherent and intelligible models that will actually help to better understand the way in which gene expression in stem cells is regulated to maintain their precarious identity.

In this thesis, I first explore the potential of HTS by describing two pilot studies using the technology to investigate global differences in the transcriptional profiles of different cell populations. In both cases, I was able to identify a number of promising candidates that mark and, possibly, explain the phenotypic and functional differences between the cells studied.

The pilot studies highlighted a strong requirement for specialised software to deal with the analysis of HTS data. I have developed *GeneProf*, a powerful computational framework for the integrated analysis of functional genomics experiments. This software platform solves many recurring data analysis challenges and streamlines, simplifies and standardises data analysis workflows promoting transparent and reproducible methodologies. The software offers a graphical, user-friendly interface and integrates expert knowledge to guide researchers through the analysis process. All primary analysis results are supplemented with a range of informative plots and summaries that ease the interpretation of the results. Behind the scenes, computationally demanding tasks are handled remotely on a distributed network of high-performance

computers, removing rate-limiting requirements on local hardware set-up. A flexible and modular software design lays the foundations for a scalable and extensible framework that will be expanded to address an even wider range of data analysis tasks in future.

Using *GeneProf*, billions of data points from over a hundred published studies have been re-analysed. The results of these analyses are stored in an web-accessible database as part of the *GeneProf* system, building up an accessible resource for all life scientists. All results, together with details about the analysis procedures used, can be browsed and examined in detail and all final and intermediate results are available and can instantly be reused and compared with new findings.

In an attempt to elucidate the regulatory mechanisms of ESCs, I use this knowledge base to identify high-confidence candidate genes relevant to stem cell characteristics by comparing the transcriptional profiles of ESCs with those of other cell types. Doing so, I describe 229 genes with highly ESC-specific transcription. I then integrate the expression data for these ES-specific genes with genome-wide transcription factor binding and histone modification data. After investigating the global characteristics of these "regulatory inputs", I employ machine learning methods to first cluster subgroups of genes with ESC-specific expression patterns and then to define a "regulatory code" that marks one of the subgroups based on their regulatory signatures.

The tightly co-regulated core cluster of genes identified in this analysis contains many known members of the transcriptional circuitry of ESCs and a number of novel candidates that I deem worthy of further investigations thanks to their similarity to their better known counterparts. Integrating these candidates and the regulatory code that drives them into our models of the workings of ESCs might eventually help to refine the ways in which we derive, culture and manipulate these cells – with all its prospective benefits to research and medicine.

Declaration

I have read and understood The University of Edinburgh guidelines on plagiarism and declare that the work presented is my own, except where otherwise indicated, and has not been submitted for any other degree or professional qualification.

Florian Halbritter

Edinburgh, October 22, 2012

Acknowledgements

This thesis would not have been possible unless with the constant support of others.

First and foremost I offer my sincerest gratitude to my supervisor, Dr. Simon R. Tomlinson, who has supported me throughout my work with his knowledge, advice and patience. I am deeply indebted to Dr. Tomlinson for creating an environment conducive to the free development of ideas, productive criticism and unencumbered research. My thanks extend to the present and former members of the Tomlinson group for stimulating discussions and assistance: Dr. Ed Curry, Dr. Laura Skylaki, Sofia Morfopoulou, Aidan McGlinchey, Will Bowring, Harsh Vaidya, Hina Dalal and Duncan Godwin.

I wish to thank also my other supervisors, Prof. Ian Chambers and Dr. Keisuke Kaji. I owe my gratitude to the many people I had the pleasure to work and collaborate with: Violetta Karwacki-Neisius, Nicola Festuccia, Rodrigo Osorno, Pablo Navarro, Alessia Gagliardi, Nick Mullen, Gary Loake, Thomas Waibel, Byung-Wook Yun, Raymond Poot, Debbie van den Berg, Johan Brandsma, Claus Nerlov, Alexander Medvinsky, Clare Blackburn, Stephanie Tetelin and many more.

Of course, none of my work would have been possible without the generosity of my host institution, the Institute for Stem Cell Research / Centre for Regenerative Medicine, and of the funding bodies that have supported me (and paid for all those computers!): The Medical Research Council and the European Union via its seventh framework programme "EuroSyStem".

The remainder of my sanity I owe to the support and love of my family and friends, Monika, Kurt, Daniel and Michael, and to my wife, Mei Sze Lam, who not only cheered me up in difficult times, but also helped by proof-reading this thesis and drew the GeneProf logo. I wouldn't have made it without you!

Dedicated to snowdrops.

* * *

Habe nun, ach! Philosophie,
Juristerei und Medizin,
Und leider auch Theologie
Durchaus studiert, mit heißem Bemühn.
Da steh ich nun, ich armer Tor!
Und bin so klug als wie zuvor;
Heiße Magister, heiße Doktor gar
Und ziehe schon an die zehen Jahr
Herauf, herab und quer und krumm
Meine Schüler an der Nase herum –
Und sehe, dass wir nichts wissen können!

(I've studied now, alas! Philosophy, jurisprudence, and medicine, and unfortunately even theology, all through and through with ardour keen! Here now I stand, poor fool, and see I'm just as wise as formerly. Am called a Master, even Doctor, too, and now I've nearly ten years through pulled my students by their noses to and fro and up and down, across, about, and see there's nothing we can know!)

J. W. v. Göthe, *Faust*

Contents

Chapter 1

Introduction and Background

This thesis explores the regulatory mechanisms underlying the cellular identity of embryonic stem cells (ESCs). The work I describe has been of a largely *in silico* nature, drawing heavily on the computational meta-analysis of large amounts of genomic data, both in-house and public, generated using high-throughput sequencing (HTS) technologies. After reviewing some of the background information pivotal for the understanding of the subsequent chapters (**Chapter 1**), I will proceed in chronological order and first discuss some of the early data analysis work I did in an attempt to gauge the utility of HTS technologies for the study of stem cell biology (**Chapter 2**). Specifically, I will talk about two pilot studies conducted in collaboration with other research groups at the University of Edinburgh: The first one on transcriptional targets dependent on the expression of a well-known stem cell regulator gene, *Nanog*, in mouse ESCs and the second pioneering transcriptional assessment of proliferating cell populations in the Japanese yew. My experience in these studies highlighted a distinct lack of streamlined data analysis methods to match the high-throughput data generation. **Chapter 3** introduces the *GeneProf* software, a novel data analysis framework that has been developed to address these issues. To lay further groundwork for following investigations, this tool has been applied for the large-scale reanalysis of a numerous published experiments, building up a valuable resource for life scientists interested in gene expression, transcriptional regulation and epigenetics (**Chapter 4**). In the penultimate chapter (**Chapter 5**), an extensive meta-analysis of these data is presented, integrating information about gene expression with the regulatory inputs of ESCs in order to track down a unique signature of gene regulation that distinguishes genes central to ES identity from the rest of the transcriptome. Finally, I conclude this thesis with a review of the primary research achievements and an outlook on future work (**Chapter 6**).

A summary of abbreviations and terms used throughout this thesis is given in **Appendix A**.

The remainder of this first chapter is structured as follows: First, a brief overview of some of the core concepts of stem cell biology relevant to the work in this thesis will be given in **Section 1.1**. I will start with a summary of early developmental processes and continue to details about ESCs. In particular, I will focus on the genetic and regulatory factors that define them. The second part of this chapter focuses on HTS technology (**Section 1.2**). After describing the technology itself and explaining the primary methodological approaches to its utilisation, I will conclude this chapter bringing the focus back to stem cells by highlighting some groundbreaking research made possible with the use of HTS.

1.1 Embryonic Stem Cell Biology

Stem cell research has undergone remarkable growth over the recent decades. The field has attracted great scientific, commercial and public interest, not least thanks to its promise for regenerative medicine and drug development. I shall now briefly review some of the fundamentals of stem cell biology. I will first give an overview of early development in the mouse, followed by details about embryonic stem cells discussing how exactly they are defined, how they were discovered and how they can be derived from an embryo. Lastly, I shall discuss the key regulators and mechanisms that are the driving forces behind embryonic stem cell state.

1.1.1 Early Mammalian Development

Stem cell biology essentially comes down to the understanding, modelling and (targeted) recapitulation of early developmental embryology. Questions such as what defines stem cells, how do they maintain their state and how to they give rise to their differentiated progeny might perhaps be best addressed by having a closer look at how equivalent processes happen naturally *in vivo*. We will look here at the embryonic development of the mouse (*M. musculus*) that for many years has served as a model system closely mimicking human development. Nevertheless, it must be acknowledged that there are notable differences in the developmental process and conclusions derived from one organism should be translated to another only with caution – after all, men and mice end up quite differently indeed.

That being said, let us now look at what is known about early mouse development starting from the fertilised egg (unfertilised: oocyte; fertilised: zygote; reviewed in[30,86,146,154,418]). During the first three to four days after fertilisation, the zygote travels to the uterus. In the meantime, a series of cell divisions (cleavages) occur (**Figure 1.1**). These early cells in the embryo are called "blastomeres". Since much of the cytoplasm is derived from the maternal oocyte, many of the early developmental decisions are believed to be controlled by maternal

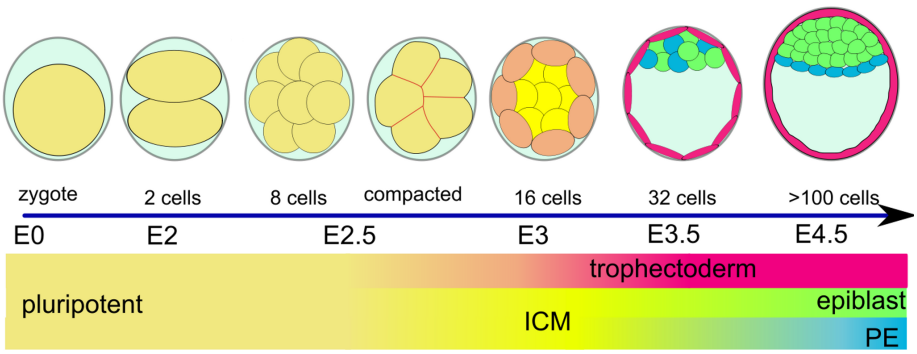

Figure 1.1: Early mouse development. Progression of the zygote through repeated cell divisions into a blastocyst with increasingly narrowed down fate. Adapted with permission from reference[281].

gene products. Usually after the 8-cell stage, the formerly loosely bound cells compact and are held together by gap junctions formed by connexins. The gap junctions allow for molecule exchange between cells, which may contribute to the establishment of polarisation and thus the spatial patterning of the embryo[238]; in fact, it has been suggested that the anterior-posterior axis might already be established at this point in time[30,145,146].

At the 16-cell stage (from now on also called "morula"), the embryo begins showing distinct pattern formation with the outer cells forming a ring of cells, the trophectoderm (TE), which will eventually constitute the trophectoderm and extraembryonic ectoderm[30,144,418]. The cells on the inside are called the inner cell mass (ICM), destined to develop into the fetus and extraembryonic mesoderm and endoderm[30,144,418]. Around day 3 post fertilisation, a cavity (blastocoel) begins to form, which together with the physically and structurally separated TE and ICM makes up the "blastocyst" at day 3.5[30].

After approximately four days, the blastocyst arrives in the uterus, but does not yet implant, because it is still enclosed by a protective layer, the "zona pellucida". This layer is then shed off and the blastocyst implants into the uterine wall at day 4.5. The ICM now becomes separated into the hypoblast and the epiblast. The hypoblast will later develop into the primitive endoderm (PE) and the epiblast harbours cells that will develop into all parts of the actual embryonic body[86]. At day 6, the embryo is made up of what is now called the trophoblast, the epiblast (or primitive ectoderm) and the PE. The primitive ectoderm contains cells that will differentiate into the three primary germ layers, endo-, meso- and ectoderm. This stage of development is called gastrulation,

After gastrulation, increasingly specialised structures begin to form. The ectoderm will eventually give rise to the skin and nervous system, the mesoderm will differentiate into bone and cartilage as well as muscle tissues and blood, and the endoderm is the basis for the development of internal organs.

1.1.2 Embryonic Stem Cells, Pluripotency and Differentiation

The work in this thesis is concerned with the study of stem cell biology. But what are stem cells and where do they come from? In fact, what is a stem cell and what is not is a matter of some discussion, but for the purposes of this work I shall describe a stem cell in terms of the following two key properties[69,509]:

Definition 1. ***Potency:*** *The ability of a cell to differentiate into heterogeneous subtypes. The derived cell types ("progeny") may be limited in their potency and exhibit phenotypic and functional differences. A cell shall be called* ***totipotent*** *if it can give rise to all embryonic and extraembryonic tissues ever observed at any point of an organism's natural development and* ***pluripotent****, if it can constitute any tissue in the actual embryonic and adult body, including the germline.*

Definition 2. ***Self-renewal:*** *The ability of a cell to divide indefinitely giving rise to identical daughter cells that also have the potential to self-renew.*

Putting these properties together, stem cells can be defined most generally as[69,509]:

Definition 3. ***Stem cell:*** *An undifferentiated progenitor cell that has an unlimited potential for self-renewal and is pluripotent, according to the definitions given before.*

Now, where in the process of embryonic development do stem cells occur? As we have seen in the previous section, mouse embryonic cells commit early on their future fate. It has been shown that cells taken from a later stage in development can no longer reconstitute all tissues of the body, they are said to be restricted in their potency. Only the zygote itself can with certainty be said to be totipotent. That is, only this mother-of-all-cells can indeed give rise to all different embryonic and extraembryonic lineages observed during development. Cells following the early cleavages may or may not be totipotent still, but certainly the last cells in the embryo that can positively give rise to any cell of the embryo proper, occur for a short period of time only in the early, pre-implantation blastocyst around E3.5[418]. These cells are called "pluripotent". Cells from later stages of development as well as a number of adult cells can still give rise to differentiated progeny of various types, yet they are greatly reduced in their potency to only specific lineages (they are "multipotent").

This insight has led to the hypothesis of the existence of undifferentiated, pluripotent cells. Indeed it has later been proven possible to derive such cells from the ICM of the pre-implantation blastocyst of mice[123,344] and, years later, from the outgrowth of in vitro fertilised human eggs[545]. Thanks to the origin of these cells and their potential as the stem population for all the tissues of a mature organism, they were subsequently called embryonic stem cells (ESCs). Because of their unique key properties – self-renewal and pluripotency – ESCs can be maintained in cell cultures (given appropriate culture conditions) and they

can divide both symmetrically into undifferentiated daughter cells as well as asymmetrically into undifferentiated and differentiated progeny. Additionally, ESCs can contribute to (viable) chimeras if injected back into a blastocyst (reviews:[42,69,70,505]).

For the purposes of this thesis, the definition of a stem cell as a cell that is pluripotent and capable of self-renewal shall be sufficient. To more rigorously characterise ESCs, the cells have to satisfy a number of additional criteria (adapted from[509]):

- ESCs must be derived without transformation or immortalisation from the ICM of the blastocyst,
- they ought to be karyotypically stable and diploid,
- clonogenic and capable of unlimited self-renewal, with a high amplification capacity.
- ESCs can demonstrate pluripotency *in vitro* and in teratomas,
- have two active X-chromosomes in female cells (no X-inactivation),
- have no G1 cell cycle checkpoint,
- are be able to contribute to all parts of chimera and can colonise and transmit to the germ line,
- and they remain undifferentiated in the presence of suitable external stimuli (see **Section 1.1.3**).

ESCs have been derived from numerous mouse strains or individual human embryos and primates, however, this was achieved only much later with the use of improved culture conditions[380,619] and some controversy exists as to whether non-mouse pluripotent cell lines are indeed equivalent to mESCs[42,147]. While all "ESCs" share the same basic defining properties (self-renewal and pluripotency), there are considerable differences in their transcriptional and epigenetic characteristics, their cell culture viability, proliferation rate and other phenotypic attributes. Moreover, they depend on different external signals and culture conditions for their maintenance[42,541,560]. Importantly, it has been noted that human ESCs differ substantially from mouse ESCs and it has been suggested that they do actually more closely resemble cells derived from the post-implantation epiblast (EpiSCs) of the mouse. In fact, when mouse EpiSCs were first derived, researchers used hESC culture conditions, which exhibit different maintenance requirements than mESCs[42,55,541].

Differences between mESCs and mEpiSCs may well be due to the different developmental stage they were derived from. After implantation, cells in the ICM undergo rapid and vast changes, for instance, (female) cells randomly inactivate one copy of the X-chromosome and they are transcriptionally and epigenetically poised to differentiation[173,381,387].

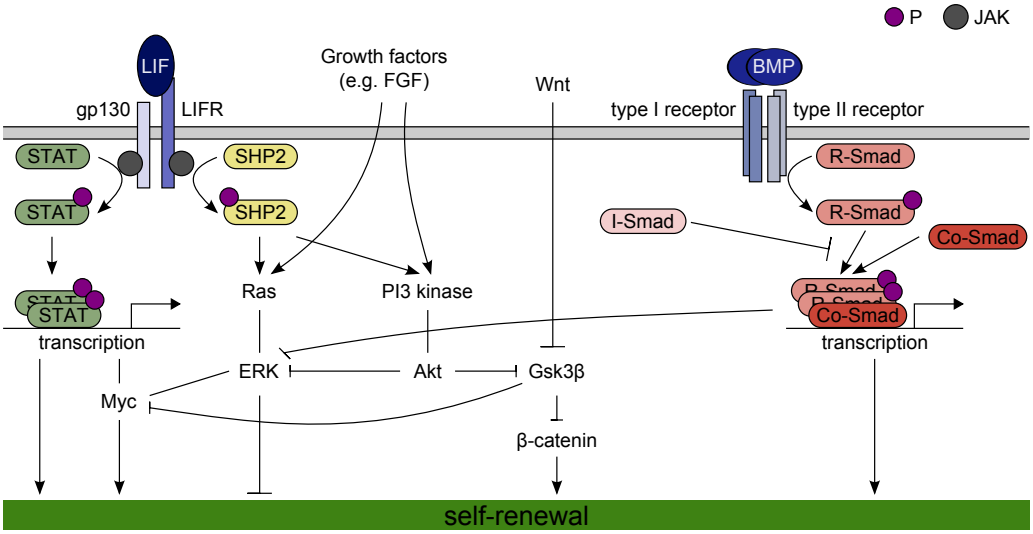

Figure 1.2: ESC signalling pathways. Intracellular signalling pathways relevant to ESC self-renewal and pluripotency with their downstream effectors in mouse. Abbreviations: JAK = Janus kinase; P = Phosphorylation; STAT = STAT-family proteins, primarily *Stat3*; R-/Co-/I-SMAD = Receptor-regulated, cooperating and inhibitory SMAD-proteins; Based on reference[402].

1.1.3 Self-Renewal and Differentiation of Stem Cells

The first stem cell lines were derived in serum on a layer of feeder cells (inactivated fibroblasts), initially without knowing much about the benefits that these conditions offered to the cells[42,56,123,147,344]. Only later, the cytokines *leukaemia inhibiting factor (LIF)* and *bone morphogenic protein 4 (Bmp4)* were identified as the main contribution of feeder cells[510,593] and as a substitute for serum[618], respectively, allowing to culture ESCs without recourse to serum and feeders.

But how do *LIF/Bmp4* confer the self-renewal properties of ESCs? LIF binds to a heteromeric receptor complex made up of *LIF receptor (Lifr)* and *gp130*. Both units have attached tyrosine kinases *Janus Kinase (JAK)* which upon binding phosphorylate STAT-family protein *Stat1* and *Stat3* (reviewed in[402]; see **Figure 1.2**). Phosphorylation induces *Stat3-Stat3* dimerisation and migration to the nucleus, where *Stat3* binds to DNA and supports the transcription of genes, e.g. *Myc* (also known as *c-Myc*) with a demonstrated positive effect on self-renewal[64,348,389,402]. Another factor activated by this pathway appears to be *Klf4*, which in turn drives expression of *Sox2*[391]. Contrary to expectations, the mitogen-activated protein kinase (MAPK)/extracellular signal-regulated kinase (ERK) signalling pathway, another downstream effector of *LIF* and of several growth factors (e.g. *FGF4*[285]), was found to encourage differentiation[58]. On the other hand, recent evidence suggests that phosphatidylinositol-3-OH kinaseAkt (PI3/Akt) and MAPK pathways support expression of *Tbx3*, which in turn encourages *Nanog*[391].

BMP family proteins bind to two types of tyrosine kinase receptors inducing the phospho-

rylation of the receptor-regulated SMAD-proteins, *Smad1*, *Smad5* and *Smad8* (**Figure 1.2**). After associating with the cooperating SMAD-protein *Smad4*, the phosphorylated proteins bind to DNA and drive the expression of, for instance, *inhibitor of differentiation (Id)* proteins[69, 402]. Via this action, BMP signalling is believed to suppress neural differentiation and encourage self-renewal[618]. Interestingly, over-expression of *Id* abolishes dependence on *Bmp4* to suppress differentiation, arguing that *Id* might indeed be a main effector of this signalling pathway in the context of self-renewal[473].

Other signalling pathways shown to be involved in the maintenance of self-renewal and suppression of differentiation are downstream of growth factors and Wnt-protein activity[69, 402]. Wnt-signalling prevents the phosphorylation of β-catenin by various enzymes, e.g. *glycogen synthase kinase 3 β (Gsk3β)*. In consequence, unphosphorylated β-catenin will no longer be degraded and can therefore influence transcription via the transcription factors *lymphoid enhancer factor (LEF)* and *T-cell factor (TCF)* . Via this pathway and a number of alternative routes ("non-canonical pathways"), Wnt has been shown to, on the one hand, promote self-renewal and proliferation[13, 479], but also be involved in various differentiation processes[511].

Recently, an alternative to the *LIF/Bmp4* media, called *2i*, has been developed[619], which utilises small molecule inhibitors of *Fgf4*-mediated ERK-signalling (otherwise resulting in differentiation[285]) and *Gsk3β* (interfering with aforementioned Wnt-singalling cascades).

ESCs represent an *in vitro* phenomenon and, if they ever exist *in vivo*, do so for only a very short period of time. To maintain this precarious state in culture, as so often, a complex interplay between the signalling networks outlined above (and others) and important endogenous factors (see next section) is required. Further research is yet required to disseminate the exact roles of individual proteins and to identify missing links and downstream targets.

1.1.4 Core Embryonic Stem Cell Transcriptional Regulators

Over the past twenty years, in-depth investigations into the molecular biology of stem cells have revealed great insights into the core transcriptional circuitry responsible for the establishment and functioning of self-renewal and pluripotency. Although many additional elements have been determined, it appears that the wider transcriptional network concerned, revolves around the expression of three core regulators, the transcription factors *Pou5f1*, *Sox2* and *Nanog* (P-S-N).

Interestingly, the three factors bind to each other's and their own promoter and enhancer elements, suggesting that they might be regulating each other to a certain degree, probably to strike the right balance of dosage necessary to maintain ESC identity. Furthermore, the three factors share many binding targets across the genome indicating that they might either control target genes cooperatively or redundantly[187, 621]. Transcription factors (TFs) can encourage

transcription via at least three routes, either by recruiting elements of the transcriptional machinery to the promoter of genes, by inducing the restructuring of chromatin (euchromatin instead of heterochromatin) or its associated elements (histone modifications, etc.) in such a way that is permissive to transcription or by releasing transcriptionally paused polymerase to allow productive elongation[91,187,245,437,621]. Alternatively, they may counteract transcription by blocking any of these routes.

Many of the genes involved in groundstate pluripotency encode TFs, but there are also co-factors and further genes that exert their function in ways other than by binding to DNA. I will now try to review some of the most important known genes implicated in ESC state.

1.1.4.1 Pou5f1

One of the most well-known key regulators of ESCs is *POU domain, class 5, transcription factor 1* (*Pou5f1*; also known as *octamer-binding transcription factor 4* or *Oct4*; reference[487,488]; reviewed in[66,69]). *In vivo*, *Pou5f1* is expressed during the earliest stages of development starting from the unfertilised egg and observed still in the ICM and even after implantation in the epiblast, but not TE or later outer embryonic layers[410,423]. Later on its expression is restricted to primordial germ cells (PGCs).

Loss of *Pou5f1* does not disrupt blastocyst formation *per se*, but disrupts the developmental potency of the cells contained and no PE or germ cells are generated: As confirmed *in vitro*, the loss of *Pou5f1* leads to differentiation into trophectoderm only[382,390]. Interestingly, overexpression (more than 1.5× the normal level) was found to lead to differentiation towards endoderm and mesoderm. Thus, fine control of *Pou5f1* expression levels is essential to maintain ESCs in a pluripotent, self-renewing state and variations in expression lead to spontaneous differentiation in a dose-dependent manner.

LIF withdrawal in ES cell cultures, leading to differentiation, correlates with a rapid drop in *Pou5f1* gene expression. However, experiments in cells in which *Pou5f1* expression has been engineered to be under the control of tetracycline, that is, in which expression can be maintained even without LIF, have shown that expression of *Pou5f1* alone is not sufficient to prevent ESC differentiation[388]. *Pou5f1* is therefore a requirement of ESC maintenance, but in itself is not sufficient for their survival.

The protein product of *Pou5f1* contains two DNA-binding domains, a low-affinity "Pit", "Oct" and "Unc" domain (POU) and a higher-affinity homeodomain. Together the two domains "encircle" the DNA, binding to a `ATGCAAAT` consensus motif[70,272,425], although an alternative `TATGCGCATA` motif might also exist[16,347,544].

1.1.4.2 Nanog

In 2003, the homeobox transcription factor *Nanog* was identified by both, computational analysis of expression data and functional cDNA expression cloning as a novel regulator of pluripotency[67,363]. It is specifically expressed in cells of the ICM in the early blastocyst, with declining expression still observed post-implantation, especially in the proximal posterior region of the epiblast[189]. During days 9-13, *Nanog* expression is further observed in migratory PGCs and in genital ridges, the expression however ceases later on and no expression is detected in adult gametes[67,608].

It has been shown that *Nanog* is capable of conferring LIF-independent self-renewal if over-expressed beyond levels usually observed in ESCs[67], yet it was later discovered that the deletion of the gene did not abolish self-renewal and that *Nanog*$^{-/-}$ ESCs could still be maintained in culture[68]. However, *Nanog* does occur naturally at variable expression levels ("mosaic expression") and cells expressing low levels of *Nanog* are more prone to differentiate[68]. Interestingly, it was also demonstrated that *Nanog* expression is not activated by *Stat3* and neither does *Nanog* drive *Stat3* expression (cp. **Section 1.1.3**), arguing for a different mode of action than might have been expected[69].

In maintaining ESCs, *Nanog* is dependent on *Pou5f1* expression and even its over-expression does not prevent differentiation into TE, if *Pou5f1* is deleted[67], however, and although both proteins show evidence of binding in each other's promoter or enhancers regions[75], the expression of neither is essential for the other[69].

Like *Pou5f1*, *Nanog* contains a DNA-binding homeodomain, but no other DNA-binding elements[232]. Consequently, the DNA sequence motif bound to by the TF is likely to be shorter and it is not yet clear which site is actually recognised *in vivo* or whether there might be alternative binding sequences[70]. Proposed motifs include the core homeodomain sequence `TAAT`[363], an extended version `TAATGG`[232] or completely different motifs `CAAT`[327] / `ccAT(C/T)A`[16,193,544]. Which of these motifs is correct, or whether, in fact, all might be valid remains an open issue.

1.1.4.3 Sox2

The third protein commonly attributed a core role as a pluripotency TF is *SRY (sex determining region Y)-box 2 (Sox2)*. Unlike *Pou5f1* and *Nanog*, expression of *Sox2* is not limited to early pluripotent or largely uncommitted cells, but it has, in fact, also a rather crucial role in the development on the neural lineage and is strongly expressed in neural progenitor cells[164]. *Sox2* expression does appear to be dispensable for the establishment of ESC identity, but this might be due to the presence of maternal *Sox2* proteins at early developmental stages[70].

Sox2 binds to DNA via a high mobility group (HMG) domain and numerous lines of

evidence suggest that it (often) binds DNA cooperatively with *Pou5f1*[5,6,75,386]. Structural studies performed with the highly similar protein *Pou2f1* and *Sox2*[448,591] and the fact that the binding sites for both factors are frequently found together and with the same orientation[75], suggest that this binding occurs cooperatively at a protein level[70].

1.1.4.4 Other Genes Relevant to Stem Cells

Numerous other genes have been implicated with pluripotency and self-renewal, some in a role as downstream effectors of P-S-N or as their interaction partners and others without any apparent, direct connection to the three at all. Of these, *Stat3*, *Klf4/2*, *Myc* and *Esrrb* might be of particular interest, since they have been able to confer LIF-independent self-renewal[183,630]. Rather than going into a lengthy discussion, I shall give here only a concise summary table of important ESC- and differentiation-linked genes and others relevant to this study (**Table 1.1**). In addition to these genes, differentiation and knock-down experiments and computational meta-analysis of genome-wide expression data have identified many additional candidates whose roles in stem cells are still poorly understood[12,157,227,276,363]. Amongst these candidates rank *Manba*, *Hck*, *Gbx2*, *Spp1*, *Otx2*, *Cldn7*, *Rrp12* and many more. It is an exciting prospect that future research into these factors might help us to extend our understanding of the core transcriptional circuitry that controls stem cell identity.

1.1.5 Epigenetic Control of Stem Cell State

It is becoming increasingly evident that TFs are not the only control mechanism driving gene expression. Rather, it is a complex network of the interactions of TFs and the epigenetic markup of a cell that allow active transcription to happen. One aspect that has received much attention over the last years is the role of epigenetic influences in regulating the balance that marks the switch from pluripotency to differentiation. Note that there is a considerable difference in the way the term "epigenetics" is used by different researchers[462], but we shall not get hung up about the definition and refer to "epigenetics" as the stable activity of genes across many generations (cell divisions) and, importantly, to the mechanisms that are controlling this stability.

In this section I will briefly review the most important (known) epigenetic factors and point to their role in stem cells and their progeny. More specific studies will be discussed later on in the context of the applications of sequencing technologies (**Section 1.2.3.4**).

1.1.5.1 DNA Methylation

The earliest discovered epigenetic regulatory mechanism is the methylation of cytosine residues in DNA, a reaction catalysed by DNA methyltransferases (DNMTs; reviews:[36,38,230,353,462,621]).

Gene	Roles, Functions and Pathways
Atrx	SWI/SNF chromatin remod.; X inact.; trophoblast dev.[28, 149, 296, 599]
Cbx7	PRC1; DNA methylation; gene silencing, inhibits differentiation[364, 404]
Cdx2	Induction of TE; mutually inhibitive with *Pou5f1*[392, 522, 539]
Chd7	Chromatin remod., ES gene activation; TrX; Sox2 cofactor[17, 121, 485]
Ctcf	Diverse functions; transcriptional activator, repressor and insulator[186, 424]
Ctr9	PAF1-subunit; transcription elongation, mRNA processing[437]
Dnmt3a/b/l	DNA methylation; transcriptional silencing[353, 363, 451]
Dppa4/5a	Suppresses differentiation; early dev.; euchromatin formation[346, 363, 531, 587]
E2f1	DNA-repair, cell cycle, tumor suppressor; coop. binds with other DBPs[41, 82]
Ep300	TF cofactor; proliferation, diff.; HAT; chromatin remod.[70, 75, 117, 396, 637]
Esrrb	Self-renewal (LIF-independent); targets ES core factors[75, 227, 631]
Fbxo15	ES-specific marker, but dispensable for self-renewal and pluripotency[549]
Fgf2/3/4/5/8	FGF/ERK pathway; early dev., differentiation; progression to EpiSC[285, 293]
Gata4/6	Induction of PrE; mutually exclusive with high *Nanog*[50, 139, 507, 627]
Jarid2	HMs; PRC2 subunit; blocks differentiation[308, 417]
Jnk1/3 (Mapk8/10)	MAPK pathway; differentiation; H3S10ph[548]
Klf2/4/5	LIF target; self renewal; iPS factor; act redundantly[75, 183, 236, 315]
Lefty1/2	TGFβ family; early dev., patterning, antagonistic to *Nodal*[372]
Lin28	Proliferation, self-renewal; iPS; early dev.; miRNA control[420, 421, 622]
Luzp1	ATAC-mediator complex; neural development[279, 301]
Mcaf1 (Atf7ip)	Heterochromatin; gene silencing[215, 331]
Med1/12	Mediator complex; at enhancers and promoters of active genes[245]
Mtf2 (Pcl2)	PRC2; transcriptional silencing; differentiation[308, 574]
Myc	*Stat3* target; self renewal; proto-oncogene; Pol2 pause release; recruits HATs; DNA replication[64, 75, 111, 137, 269, 437, 479]
Mycn	Chromatin remodelling; H3K4 methylation and acetylation[94]
NelfA (Whsc2)	NELF-complex; transcriptional pausing[437]
Nfya	Open chromatin; recruits Pol2 and TFs to promoters[548]
Nipbl	Cohesin loading factor; cooccupies with Mediator/Cohesin[245]
Nodal	Key regulator in early dev., suppresses neural lineage[300, 560, 561]
Nr0b1 (Dax1)	Dev.; gender spec.; pluripotency; neg. regulator of *Nr5a2*[229, 258, 261, 350, 363]
Nr5a2	Blocks differentiation; self-renewal; iPS / reprogramming[47, 170, 172, 198, 538]
Phc1	PRC1; gene silencing, differentiation[51, 224]
Prdm14	Blocks endoderm differentiation; targets ES core factors[79, 332, 556, 609]
Rest	Self-renewal; blocks neural differentiation[18, 508]
Ring1b (Rnf2)	Chromatin compaction; PRC1; silencing; blocks differentiation[540, 566]
Sall4	Blocks TE differentiation; cooperates with *Nanog*[603, 621, 628]
Smad1/2/3	BMP, TGFβ/Activin/Nodal signalling; growth, dev., survival[29, 560, 618, 621]
Smarca4 (Brg1)	SWI/SNF; chromatin accessibility, activation; self-renewal[21, 200, 265, 349, 581]
Smc1/3	Cohesin complex; DNA loop formation[245]
Spt5	DSIF-complex; transcriptional pausing, but also elongation[437]
Suz12	Histone variants; PRC2; transcriptional silencing[61, 75, 342]
Tbx3	LIF signalling; self-renewal; blocks meso- and ectoderm[227, 329, 391, 621]
Tcf3	Wnt signalling; pluripotency, differentiation[87, 342, 621]
Tcfcp2l1	Little known; interacts with HDAC proteins[75, 564]
Tcl1	Self-renewal, growth; proto-oncogene; blocks neural diff.[227, 327]
Tdh	Threonine ctabolism; rapid cell growth; highly active in ESCs[578]
Tet1	DNA methylation; 5mC $\rightarrow$ 5hmC[528, 600]
Thap11 (Ronin)	Self-renewal (LIF-independent); chromatin remod. and HMs[106, 469]
Utf1	Differentiation, pluripotency; chromatin-associated[363, 403, 565]
Yy1	Docks *Xist* onto chromosomes; recruiter of TFs, PRC and TrX[226, 233, 354, 405]
Zic3	Pluripotency; positvely regulates *Nanog*[318, 319]
Zfp42 (Rex1)	Common ES marker; inhibits differentiation; X-inact.[31, 363, 376, 491, 501, 607]
Zfx	Self-renewal; also in adult SCs; targets *Tcl1* and *Tbx3*[142]

Table 1.1: ES- and differentiation genes. Genes with known implication in stem cells, differentiation or otherwise relevant to this study.

Different methyltransferases might serve different purposes, for instance, *Dnmt3a* and *Dnmt3b* are believed to confer *de novo* methylation[399]. The propagation of methylation states ("maintenance methylation"), that is, methylation of hemi-methylated CpG dinucleotides during DNA replication, on the other hand, is facilitated by *Dnmt1*[83,307,379]. More recent research, however, disputes this strict distinction between the functions of the individual enzymes and suggests that all of them might be involved in all mechanisms[190,266,450,462]. Demethylation may occur passively, i.e. without the maintenance of methylation, or might be directed by DNA glycolase activity or direct removal[365,438,462,559,572,642]. DNA methylation is thought to carry out its effects through the transcriptional regulator Kaiso (mouse gene *Zbtb33*) and proteins with a methyl-CpG-binding domain (MBD; e.g. *Mecp2*) and their interaction with other co-regulators[197,429,463].

DNA methylation is commonly associated with transcriptional silencing. Failure of proper methylation is linked to developmental defects and involved in cancer[129]. Recent evidence suggests that active demethylation by *activation-induced cytidine deaminase (AID)* might indeed be a requirement for the generation of induced pluripotent cells (**Section 1.1.6**), supporting the concept of an epigenetically "permissive" groundstate in ESCs. During natural development, global DNA demethylation occurs at two stages: After fertilization in genome of the zygote, which remains largely unmethylated until after implantation, and later on during the formation of primordial germ cells[191,353]. It may hence be reasoned, that demethylation is generally associated with the resetting of epigenetic signatures to a "tabula rasa" state.

Upon differentiation, ESCs are thought to silence pluripotency genes and those important for other lineages by methylating their promoters[127]. Much of the functionally relevant methylation appears to happen in the context of so-called CpG-islands (CGIs), preferentially promoter-associated regions of the genome with a high content of CpG pairs that are under permissive circumstances unmethylated[217]. The methylation of CGIs has been linked to X-inactivation, genomic imprinting and tissue-specific silencing[119,217,443]. In two interesting studies, researchers looked at the promoter methylation status of cells during the *in vitro* differentiation of ESCs into the three early germ layers[222] and in ESC, embryonic germ cells, sperm, trophoblast stem cells and embryonic fibroblasts[127]. They noted significant differences in *de novo* methylation of target genes consistent with lineage as well as a specific demethylation of pluripotency-related genes at the onset of development. Further supporting the importance of methylation for the silencing of pluripotency genes, it has been observed that ESCs can be derived in the absence of methyltransferases, but that the differentiation of these cells is impaired, probably due to the failure to silence pluripotency genes[131,621].

5-methylcytosine (5-mC) may be further modified to 5-hydroxymethylcytosine (5-hmC)[353]. This reaction is catalysed by *ten-eleven translocation* proteins, e.g. *Tet1*[528,600]. Both, the concentration of *Tet1* and the frequency of 5-hmC decrease upon differentiation of ESCs and

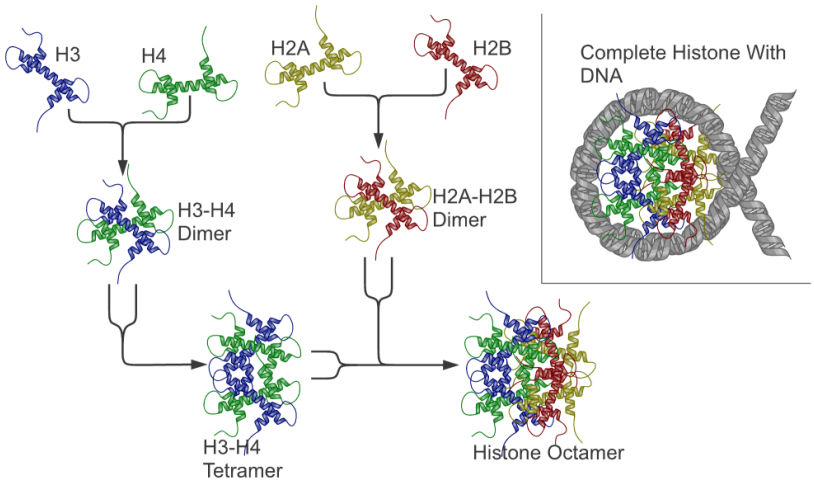

Figure 1.3: Nucleosome composition. In chromatin, DNA is wrapped around nucleosomes composed of eight histones, two of each type (H2A, H2B, H3, H4). The structure is stabilised by linker histones (H1). Adapted from Richard Wheeler (WikiMedia Commons, 2005, `http://en.wikipedia.org/wiki/File:Nucleosome_structure.png`).

the knock-down of *Tet1* reduces self-renewal efficiency[225, 528], implicating 5-hmC directly with the core functional network of ESCs[525, 600].

1.1.5.2 Nucleosomes, Histones and Chromatin

In eukaryotes, DNA together with various proteins is packaged into a higher-order structure called chromatin. The primary architectural scaffold of chromatin are nucleosomes, small protein complexes that DNA wraps around. Nucleosomes are then further compacted together using linker proteins and other structural elements. The structure of chromatin changes during cell cycle (major decompaction is necessary for mitosis), but also in response to regulatory mechanisms. For instance, chromatin may be loosely packed ("euchromatin"), allowing the active transcription of the DNA code by polymerases or more tightly packed ("heterochromatin"), preventing such activity[195, 462]. Chromatin may be remodelled by many factors, amongst others changes in DNA methylation (see previous section).

Each nucleosome is composed of eight proteins called "histones". There are five different types of histones (known) in mammals: H1, H2A, H2B, H3 and H4. The latter four make up the nucleosome octamer (**Figure 1.3**), while H1 acts as a structural linker protein[462]. Histones might occur in different structural variants, potentially with different functions[379], but they can also be enzymatically modified and the mechanisms and effects of these modifications are better understood and have been implicated with important roles in stem cells (reviews in[50, 110, 230, 353, 462, 621]). Interestingly, the structure of nucleosomes and histones is highly conserved across virtually all eukaryotic species[558].

Past research has revealed at least eight different types of modifications to histone pro-

teins: Methylation (*me*), acetylation (*ac*), phosphorylation (*ph*), ADP ribosylation (*ar*), ubiquination (*ub*), sumolyation (*su*), deimination/citrullination (*ci* or *cit*) and biotinylation (*bio*)[264,462]. The abbreviations given in the brackets follow the Brno nomenclature[558], according to which type of modification is given in the end following the designation of the histone (H2A, H2B, H3, H4) and amino acid (K: lysine, R: arginine, S: serine, T: threonine, Y: tyrosine) concerned. More than one methylation, ubiquination or ADP ribosylation can be applied to the same amino acid; to distinguish the variants an additional number is inserted after the modification code, e.g. *me3* for trimethylation or, more general, *ubn* for polyubiquination. Lastly, the dimethylation of arginine can be either symmetrical or asymmetrical, indicated by addition of another letter in the end, i.e. *me2s* or *me2a*, respectively. Histone modification codes are, in general, not italicised and I do so here only to distinguish the individual letter codes from the rest of the text. Not all modification work for all amino acids, **Table 1.2** gives an overview of known modifications in human (mostly equivalent for mouse).

There is a great number of known histone modifications with diverse functional roles (**Table 1.2**) – and it appears likely that further modifications might be found in future and alternative roles discovered for the modifications already known. Arguably, the best-studied histone modifications in ESCs are the methylation and acetylation of various lysines and arginines on histones 3 and 4 (see reviews[33,353,462,621]). Generally speaking, lysine acetylation and arginine methylation alike are implicated in functionally active genes and the consensus appears to be that the relationship is causal or at least permissive, rather than a consequence[33,462]. Histone deacetylases (HDACs) repress transcription by removing these activating histone marks and the inhibition of these enzymes has been demonstrated to block stem cell differentiation due to failure to silence pluripotency genes[299]. Recently, HDAC inhibitors have been used to increase the efficiency of the reprogramming of somatic cells to a pluripotent state[212] (**Section 1.1.6**).

Perhaps one of the most interesting observations regarding histone modifications is the presence of both activating H3K4me3 and repressive H3K37me3 ("bivalent domains") in the promoters of many developmentally related genes in ESCs[14,35,50]. Bivalently marked genes are transcriptionally repressed, but the presence of the activating marks indicates that they are ready to be transcribed once the repressive mark disappears. Thus, bivalent genes are captured in a special state "poised" for transcription. Bivalent domains are exceptionally highly conserved between species[35], advocating an important biological role. Some controversy exists as to whether these HMs actually ever occur simultaneously in the same cells or whether they are indeed present in different cells, possibly from different subpopulations, although studies using sequential chromatin immunoprecipitation[71,141,355,554] (that is, pulling out DNA that is enriched for both marks at the same time) have shown that the two marks do indeed occur together in at least some promoters[35,103,412].

There are two groups of histone modifying enzymes that are particularly well understood:

Histone H1		
K25	ac	heterochromatin
	me1	heterochromatin
S17	ph	cell cycle interphase
S26	ph	euchromatin
S171	ph	mitosis
S172	ph	mitosis
S186	ph	rDNA activation
S188	ph	cell cycle interphase
T10	ph	mitosis
T17	ph	mitosis
T30	ph	mitosis
T137	ph	mitosis
T145	ph	mitosis
T153	ph	mitosis
T154	ph	mitosis
Histone H2A		
K5	ac	DNA repair
K9	ac	activation
	bio	heterochromatin?
K13	ar	DNA repair, histone-DNA
	bio	heterochromatin?
K119	ub	silencing
K121	ub	silencing, X inact.
K125	bio	heterochromatin?
K127	bio	heterochromatin?
K129	bio	heterochromatin?
R3	ci	silencing?
	me2	activation?
S137	ph	mitosis
S139	ph	apoptosis, DNA repair
T120	ph	metaphasic centromeres
Y142	ph	DNA repair
Histone H2B		
K5	ac	activation
	me1	cor. w. active genes
K12	ac	cor. w. DNA methylation
K15	ac	activation?
K16	ac	cor. w. DNA methylation
K20	ac	at TSS, activation?
K30	ar	DNA repair, histone-DNA
K46	ac	cor. w. DNA methylation
K120	ac	at TSS, activation?
	ub	elongation, H3K4/79me
S14	ph	apoptosis
Histone H4		
K5	ac	activation
K8	ac	at TSS + in gene, activation?
	bio	heterochromatin
K12	ac	at TSS + in gene, activation?
	bio	heterochromatin
K16	ac	DNA repair, H3K79me
	ar	DNA repair, replication
Histone H3		
K4	ac	at TSS + in gene, activation?
	me1/2	activation
	me3	activation, elongation
K9	ac	activation, elongation
	bio	heterochromatin?
	me1/2	silencing
	me3	silencing?
K14	ac	chromatin remodelling
K18	ac	activation
	bio	heterochromatin?
K20	me1	cor. w. inactive genes, mitosis
	me2	mitosis
	me3	heterochromatin
K23	ac	activation
K27	ac	activation
	ar	DNA repair, histone-DNA
	me1	cor. w. active genes
	me2/3	silencing
K36	ac	activation
	me1	cor. w. active genes
	me2	DNA repair, restricts H3K27me
	me3	restricts H3K27me
K37	ar	DNA repair, histone-DNA
K56	ac	DNA repair
K79	me1/2	activation?
	me3	? (different from me1/me2)
R2	ci	silencing
	me1	activation
	me2	H3K4me3 antagonist
R8	ci	silencing
	me2	rRNA regulation
R17	ci	silencing
	me1/2	activation
R26	ci	silencing
	me1	activation
S6	ph	cell cycle
S10	ph	mitosis, genomic stability
S28	ph	mitosis, H3K27me-¿ac
S31	ph	metaphasic centromeres
T3	ph	mitosis
T6	ph	keeps H3K4me
T11	ph	mitosis, activation
T45	ph	nucleos. structure, apoptosis
Y41	ph	euchromatin
	ub	DNA damage protection
Histone H4 (cont.)		
K91	ac	activation
R3	ci	silencing?
	me1/2	activation
S1	ph	DNA repair

Table 1.2: Histone modifications. An overview of (human) histone modifications with their associated, putative biological function. Labelling according to Brno nomenclature[558]. From the HIstome database[264].

- Polycomb group (PcG) proteins, which can be further divided into polycomb repressive complex 1 and 2 (PRC1/2), and are, as the name suggest, believed to have a repressive function. PRC1 members facilitate mono-ubiquitination of H2AK119[577] and PRC2 (e.g. *Ezh2, Eed, Suz12*) the trimethylation of H3K27[61]. PcG-related silencing, especially via PRC2, has been demonstrated repeatedly to be essential for many stages of normal development as well as the establishment, maintenance and differentiation of ESCs[50, 308, 353, 395, 416, 417, 498, 574]. More details about some relevant recent studies will be mentioned later on (**Section 1.2.3.2**).

- Trithorax group (trxG) proteins, on the other hand, might be responsible for gene activation by conferring H3K4me3[50, 343, 353, 452].

Both protein complexes have been shown to be associated with *Nanog* and *Pou5f1* binding[110, 302] and so have the chromatin remodelling complexes SWI-SNF[580] (switch-sucrose non-fermentable) and NuRD[247, 316] (nucleosome remodelling and deacetylase), that influence chromatin structure in a way that is conductive or repressive with respect to gene expression, respectively[110, 621].

1.1.5.3 Non-coding RNA

Transcripts that are not being translated into proteins had traditionally been considered non-functional and mere effects of transcriptional noise. This concept has been repeatedly challenged over the past decade or so and important roles for various species of non-coding transcripts (ncRNAs) have been discovered. Perhaps one of the best-known examples of an ncRNA with proven importance in development is *Xist/Tsix*. *Xist*, an ncRNA itself, is essential for X inactivation in female cells. Its function is blocked by an anti-sense ncRNA transcribed from the opposite strand, *Tsix*[110, 374, 376]. Importantly, reactivation of the inactive X chromosome is a hallmark of ESCs and *Xist* seems to be repressed also by *Pou5f1, Sox2* and *Nanog*[110]. Similar repressive anti-sense transcription has been reported for other imprinted genes[65, 444, 445].

Micro-RNAs (miRNAs), in particular, have attracted much attention in the stem cell field[110, 617]. miRNAs are pieces of single-stranded RNA of only 18-25 nucleotides in length. They have been reported to interact with messenger-RNA (mRNA) resulting in degradation (via RNA-induced silencing complex, RISC[617]), deadenylation or the repression of translation[113]. Alternatively, they may interact with DNA or histones and might create heterochromatin[462]. miRNA expression is often specific to tissues or cell types[273].

Disruption of the orderly processing of miRNAs by enzymes such as *Dicer*, has been shown to cause severe defect in proliferation and differentiation[250, 370]. Moreover, several miRNAs have recently been reported to induce the transformation of somatic cells into stem cell-like,

RNA	Expression / Regulation / Function	Reference
	Stem Cell-Specific	
miR-290 to miR-295	down-regulated upon differentiation; balances ESC maintenance/differentiation by regulating DNA methylation via *Rbl2*; regulated by PSNT	72, 110, 206, 342
miR-291-3p, miR-295, miR-294	can generate iPS cells	243
miR302/367	can generate iPS cells	8, 23, 286
miR-205	supports mammary gland adult SC self-renewal by suppression of PTEN	617
	Differentiation- / Tissue-Specific	
miR-134, miR-470	up-regulated upon differentiation; targets CDS of PSN	110, 342, 537
miR-296	expressed specifically during differentiation; targets CDS of N	110, 342, 537, 621
miR-155	expressed specifically in immune system	342, 621
miR-375	expressed specifically in pancreatic islets	342, 621
miR-124 and miR-9	expressed specifically in neural cells	342, 617, 621
miR-145	represses PKS; silencing of self-renewal	617
let-7	represses *Lin28* and *Myc*; silencing of self-renewal	617
miR-1, miR-133	upon differentiation, represses *Dll-1* (non-muscle fate) and therefore promotes cardiomyocyte differentiation	617
miR-203, miR-124, miR-1/miR-206	promotes adult (epidermal, neuronal, muscle) SC differentiation by repressing of *p63*, *Sox9* and *Pax7*, respectively	617
miR-125b	promotes hair follicle adult SC differentiation into various lineages by targeting *Blimp1, VDR* and others	617

Table 1.3: miRNAs implicated in stem cell functions. Non-exhaustive list of miRNAs and miRNA clusters with their associated putative function. Extracted from reviews and papers[110, 243, 286, 617, 621]. P = *Pou5f1*, S = *Sox2*, N = *Nanog*, T = *Tcf3*, K = *Klf4*.

reprogrammed cells[8, 243, 286] (**Section 1.1.6**). Both lines of evidence stress the key functional role of miRNAs in many natural processes and also for stem cell identity. A summary of several known miRNAs is given in **Table 1.3**.

1.1.6 Restoration of Pluripotency

The dedifferentiaton and "reprogramming" of somatic cells to pluri- or even totipotency has been a topic of active research for many years[44, 95, 174, 486, 527, 594] and two major methodologies (with variations) have been established for this purpose (**Figure 1.4**; reviews:[175, 176, 187, 513, 611]):

- Nuclear transfer of somatic cell contents into oocytes[594] (somatic cell nuclear transfer, SCNT), even of different species. Upon transfer, pluripotency markers are rapidly induced[175, 176].
- Cell fusion of somatic cells with ESCs[95, 527] leads to the "dominant" ESC imposing its expression on the somatic cell. Fused cells may be multinucleaic heterokaryons, which will not survive long, or hybrid cells with fused tetraploid nuclei. These hybrids can proliferate and form euploid (same species) or aneuploid (different species) offspring[611].

More recently, in 2006, groundbreaking research led by Shinya Yamanaka achieved the reprogramming of a somatic cell (a fibroblast) to a self-renewing state mimicking that of ESCs using retroviral transduction of only four defined factors, *Pou5f1, Sox2, Klf4* and *Myc*[529]. The cells, termed "induced pluripotent stem cells" (iPS cells), could at this point not contribute

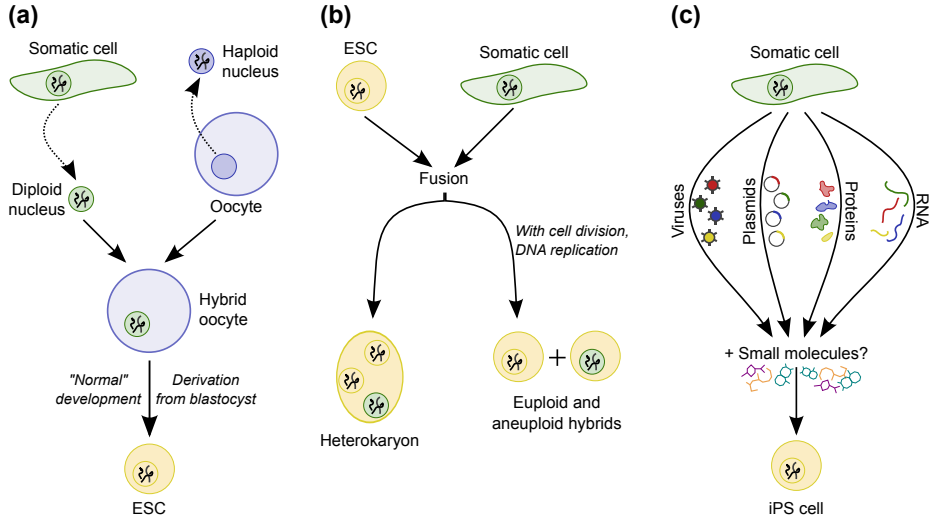

Figure 1.4: Somatic cell dedifferentiation strategies. Schematic representation of somatic cell reprogramming strategies. **(a)** Somatic cell nulear transfer, **(b)** cell fusion and **(c)** various types of induction by defined factors have all been used to reinstantiate pluripotency and self-renewal in somatic cells. Inspired by reference[611].

to chimeras. However, when the cells were subsequently selected for those that successfully reactivated *Nanog*, they were overall transcriptionally and epigenetically more similar to "real" ESCs and did contribute to viable chimeras[400]. This demonstrates that, although *Nanog* was dispensable for the induction of iPS cells, the reactivation of its expression might serve as (or at least mark) an important stepping stone to groundstate pluripotency, a concept later supported by strong additional evidence[504]. Selection for other markers, such as SSEA-1 and *Fbxo15* was not sufficient to demarcate fully reprogrammed cells[54,400]. It appears that iPSCs need to be epigenetically "reset", that is, histone and DNA methylation and acetylation (and possibly other modifications) need to be reorganised in a way permissive to pluripotency and removing marks specific to the differentiated cell of origin[37,202,360] (see previous section, **Section 1.1.5**).

Initial excitement about iPS cells was slightly hindered by low reprogramming efficiency and the requirement of retroviral transduction and stable expression, in particular, of the proto-oncogene *Myc*, preventing direct application to regenerative medicine (see next section). A considerable amount of subsequent research has focused on identifying (a) less problematic "cocktails" of reprogramming factors and (b) reversible and genomically stable ways of administering these factors in a manner that is (c) effective for reprogramming purposes.

Shortly after the Yamanaka group, Yu *et al.* reported induction of pluripotent human cells using a combination of *POU5F1, SOX2, NANOG* and *LIN28*[622]. They used a lentivirus instead of a retrovirus and confirmed normal karyotype as well as telomerase activity and expression of markers consistent with hESCs.

In a first step towards clinically applicable iPSCs, Huangfu *et al.* showed that DNMT- and HDAC-inhibitor valproic acid (VPA) greatly increased the efficiency of reprogramming and eliminated the requirement for *Myc* in the process[212], consistent with the role of HDAC inhibitors in differentiation of ESCs which I have discussed earlier[299] (**Section 1.1.5.2**). Similarly, using a combination of other small molecules it was even possible to reprogram fibroblasts using only two factors, *Pou5f1* and *Klf4*[502]. Interestingly, also a small molecule inhibitor of *GSK3β*, the inhibition of which had previously been shown to support mESC self-renewal, was reported to increase reprogramming efficiency in human cells, while replacing any requirement for *SOX2*[314].

The first group to demonstrate integration-free reprogramming made use of a adenovirus to transiently express the original four defined factors[514]. While removing the need to stably integrate reprogramming factors into the target genome is desirable for clinical purposes, the efficiency of iPS induction suffered, though, making this approach hardly viable for large-scale application. Adenoviruses were later on also used to induce pluripotency in human fibroblasts[640]. In the same year, Yamanaka's own group suggested the use of plasmids to facilitate reprogramming without viral integration[401]. Two expression plasmids were used to transfect *Pou5f1*, *Sox2* and *Klf4* and *Myc*, respectively, yet efficiency was unfortunately again suboptimal.

To address the efficiency issue, while avoiding permanent integration of exogenous factors, Kaji *et al.* used non-viral transfection with a single *Pou5f1/Sox2/Klf4/Myc*-vector to reprogram human and mouse fibroblasts[248]. The combination of this vector with a *PiggyBac*-transposon[584,598] enabled robust induction of pluripotency markers. Importantly, exogenous factors could be completely removed after the reprogramming process.

An alternative to the induction of factor expression in the somatic cells is to simply introduce the relevant proteins directly into the cells. Zhou and colleagues used recombinant proteins in which a poly-arginine transduction domain had been fused to *Pou5f1, Sox2, Klf4* and *Myc* proteins (enabling penetration of the plasma membrane) to introduce the gene products into the target cells[639], presenting a simple, quick and safe method for generating iPSCs.

In 2009, research led by Robert Belloch[243] used the miRNAs *miR-291-3p*, *miR-294* and *miR-295* to improve reprogramming efficiency by *Pou5f1, Sox2* and *Klf4*. Interestingly, they found that this led to more homogeneous iPSC populations and that additon of *Myc* did not further increase efficiency. They argued that the miRNAs are likely downstream targets of *Myc* (which binds in their promoter), offering a mechanism by which *Myc* might otherwise have facilitated reprogramming. Similar findings were obtained by studies with miRNAs in human and mouse by another group[8]. In fact, they showed that lentiviral expression of the miRNA cluster *miR302/367* in combination with the suppression of *Hdac2* can directly reprogram cells without the transduction of any TFs.

The overview given here is merely meant to give an impression of the timeline of research into iPS cells over the past years. This has been an incredibly active field and the number of studies is by far too high to present here. For recent, excellent reviews please refer to references[513,610,611].

1.1.7 Uses of Stem Cells in Research and Medicine

Stem cells offer many prospective uses, including:

Developmental biology Embryonic stem cells represent an (artificially maintained) state reminiscent of cells in an early stage of development (**Section 1.1.1**). As such, they provide a useful tool to study developmental mechanisms *in vitro* and, in particular, to trace molecular mechanisms that would otherwise be difficult to disseminate *in vivo*[120,377,471].

Cancer research Stem cells share certain characteristics with cancer cells, to a degree that some researchers even refer to certain cancer cells, that exhibit the potential for indefinite self-renewal as "cancer stem cells"[481]. As such, stem cells may find use as models for cancer research, e.g. to study oncogenes, shared signalling pathways, abnormal cell division and differentiation.

Disease research Effective modelling of diseased cells in culture can provide a tool for studying the causes and cellular effects of genetic disorders. Stem cells, that can be differentiated into any cell of the body and that can be genetically engineered comparatively easily provide the ideal starting point for such research[116,120,219,377,547,604].

Tissue-regeneration and cell therapy Demand for organ transplants, sadly, exceeds supply. Regenerative medicine offers one potential avenue to address this issue in future, with stem cells potentially being useful to regenerate tissues and organs[377,604] (possibly using patient-derived somatic cells; **Section 1.1.6**). Even where the transplant of entire organs is not feasible, cell therapy may be beneficial to counteract the effects of disease and ageing, e.g. to fight neurodegenerative disorders like Alzheimer's or Parkinson's disease[330].

Drug development The pharma-industry has developed a great interest in stem cells for the purposes of drug development. Stem cell-based disease models can be used for large-scale screens with small molecule compounds to identify and test the efficiency, side-effects and potential toxic effects of new drugs[219,377,471,604]. Not only positive effects of medicinal drugs are an active area of research: For the development of new pesticides and food additives, trials using cultures of stem cells can give crucial insights into the implications on human health.

Personalised medicine A combination of the former points, the use of iPS cells derived from the patient's own body (**Section 1.1.6**), offers the potential to take medicine to a whole new level[116,377]. With each of us being different, often the effects of drugs on any given individual can vary drastically and are not always predictable. Similarly, the success rate of organ transplants declines with growing genomic dissimilarity. Using stem cells, it may be possible in the near future to test drugs patient-specifically, to customise or even custom-develop effective treatments and to grow tissue that is fully compatible with the recipient's body system.

1.2 High-Throughput Sequencing

The study of gene expression patterns has revealed great insights into the workings of cellular systems. In the past decade, most research has relied on the use of microarray technology to monitor expression levels indirectly by hybridising transcript libraries to oligonucleotide probes on an array[114,520]. Microarrays made the simultaneous measurement of thousands of genes possible and both, the technological hardware as well as the software and algorithms for their downstream analysis have undergone drastic development over time. More and more probes were placed on the slides and sophisticated tools were invented to account for technological short-comings, but nevertheless some issues remain unsolved, foremost an unavoidable bias towards those genes for which probes have been incorporated into the platform. Microarrays furthermore suffer from issues like cross-hybridisation and partly poor reproducibility.

An alternative to the hybridisation-based approach is the direct read-out of transcript sequences. Early methods include SAGE[567] and MPSS[447], but they were hindered by comparatively high costs and a difficult and time-consuming methodology limiting their use to large genome sequencing centres. More recently a new generation of high-throughput sequencing (HTS) platforms have revolutionised the field and they now offer the opportunity to overcome earlier barriers by greatly reducing expenses and making large-scale sequencing projects available to a wider scientific audience[499]. It is now feasible, even for smaller laboratories, to sequence large libraries of expressed sequence tags (ESTs) or even entire transcriptomes. Previous studies have revealed major improvements of the deep sequencing approach to conventional microarray analysis in terms of robustness and resolution[340,506,526,585].

In this section, I will first review the major high-throughput sequencing platforms available at present and subsequently go further into the applications they make possible – in themselves, largely independent of the specific platform employed. I will then also highlight some noteworthy previous applications of sequencing platforms for the study of stem cell biology.

1.2.1 Technologies

The recent years have seen the development and (commercial) launch of numerous new sequencing platforms (reviews:[9, 104, 336, 337, 499]). While the individual technologies differ greatly in the details of the mechanisms involved, they all share some common characteristics, foremost an unparalleled increase in throughput accompanied by a massive drop in costs as compared to conventional, "Sanger-style" capillary sequencing[155, 477, 478]. When in the past, it took years to sequence a single genome and the costs were in the millions, for instance, for the Human Genome Project[90], the same depth of sequencing can now be achieved within weeks and for a fracture of the costs. But the prospects of the new technology reach far beyond *de-novo* and re-sequencing of genomes. For the first time it is affordable to read out not only whole genomic sequences, but also short fragments thereof or transcripts. The applications henceforth include gene expression profiling, the analysis of short transcripts – not before measurable at a reliable level – and the unbiased analysis of chromatin immunoprecipitation and epigenetic data[506, 526, 585, 597].

1.2.1.1 Roche / 454

As the first next-generation sequencing technology to be launched commercially in 2005, 454 Life Sciences' (454; Branford, CT, USA; now Roche, Basel, Switzerland) *FLX* pyro-sequencer revolutionised the field[339, 467]. In comparison to capillary sequencing, a simplified sample preparation protocol utilising bead-based emulsion-PCR for the creation of adapter-flanked sequencing libraries facilitates a cost-effective, rapid experimental workflow. The beads are placed onto a micro-fabricated solid support of picoliter-scale wells. Even though impressively miniaturised, the size of the wells still limits the amount of distinct sequences read out in parallel.

The solid platform supports a constant flow of sequencing reagents ("flow-cell"), therefore enabling rapid sequencing reactions. The concept of the flow-cell has been adopted by all other manufacturers. The actual sequencing in the FLX platform is based on the detection of pyrophosphate release upon the incorporation of extra nucleotides into a sequence. The pyrophosphate release triggers an enzymatic cascade ending in luciferase and emitted light can be detected by the machine. The advantage of this approach over the alternative, the step-wise incorporation of labelled nucleotides, is that the sequencing reactions appear to be more stable resulting in the successful establishment of longer read sequences (average read length with a *Titanium*-generation instrument is about $400bp$).

However, the continuity of the process poses a problem for the sequencing of homopolymeric sequences (consecutive stretches of identical bases), since there are no clear boundaries between cycles and multiple occurrences of the same base can hence only be inferred by signal

intensity[499].

Nevertheless, the 454/Roche instruments have been from the outset arguably the platform of choice for *de novo* genome sequencing thanks to comparatively long read sequences.

1.2.1.2 Illumina / Solexa

The *Illumina Genome Analyser* (San Diego, CA, USA; originally Solexa, Essex, UK) was the next platform to reach the market in 2006 and has since largely dominated the field[128, 557]. Here, adapter-ligated nucleotide sequences are amplified using the *Illumina ClusterStation* to form patches of identical sequences (called "colonies" or "polonies") on a flow-cell that is covered with a dense lawn of single-stranded oligonucleotides that correspond to the adapters ligated to the probe sequences during sample preparation.

On Illumina's flow cells, amplification and cluster-formation is achieved through repeated cycles of Bridge-PCR (as opposed to 454's emulsion PCR). The flow-cell is subsequently inserted into the *Genome Analyser* instrument (now called *HiSeq* in the latest generation), which performs the actual sequencing fully automatically, by incorporating one labelled, reversibly terminable nucleotide complementary to the probe sequences at a time. Each sequence extension step is followed by high-resolution imaging to read out the latest addition to the sequence of each cluster. The procedure is repeated to obtain a read sequence of the desired length. Effectively, the sequence is being read out while a second complementary sequence is being synthesised ("sequencing-by-synthesis").

While the sequencing may theoretically be continued for arbitrarily many cycles, experimental evaluation has shown that the quality of the base calls drops with read length and good results can currently only be obtained for about $100 - 150$ sequencing cycles, thus producing reads of $100 - 150bp$ length.

1.2.1.3 ABI SOLiD

As the last of the three major competitors to enter the field, Applied Biosystems (Foster City, CA, USA) introduced their *SOLiD* system in 2007[499, 500], now incorporated in Life Technologies (Grand Island, NY, USA). Like the 454 platform, SOLiD relies on bead-based emulsion-PCR to create clonal sequencing features which are subsequently immobilised to a solid substrate.

Sequencing is performed making use of a DNA ligase (not a polymerase) that ligates fluorescently labelled octamers to the complementary probe strands. After each ligation cycle, images from four colour channels are read out creating sequences in so-called *'colour-space'*. The octamers are thereafter cleaved and the procedure is repeated.

The colour-space model in combination with two-base encoding (an error correction scheme)

yields remarkably low error rates (according to the manufacturer). Like Illumina's Genome Analyser, SOLiD creates a high number of comparatively short reads making it particularly suitable for sequencing of transcript libraries (mRNAs, miRNAs, genomic fragments from ChIP, etc.).

1.2.1.4 Others

A number of further competitors have entered the market more recently, but have not yet, generally speaking, accrued any significant share of the market and shall hence be only mentioned for completeness' sake in this place.

A second subsidiary of Life Technologies (Grand Island, NY, USA), Ion Torrent, is approaching the sequencing problem from a slightly different angle than its competitors: Avoiding any need to detect light emission of any sort, Ion Torrent instruments exploit the fact that the incorporation of a nucleotide by polymerase releases a hydrogen ion (source: `http://www.iontorrent.com`). In combination with an array of DNA-templates that is sensitive to the release of these ions (measuring changes in the pH of the solution), this phenomenon can be used to read out rather long DNA sequences (about $200bp$) very quickly. Ion Torrent offers various semiconductor chips achieving increasing levels of sequencing depth.

Dover Systems (Salem, NH, USA) have recently started marketing the *Polonator G.007* system, developed in collaboration with the George Church laboratory (Harvard Medical School) as a low-cost, bench-top instrument advocating open standards and freely available, open-source software. Currently based on emulsion PCR-based amplification and ligation-based sequencing, the instrument offers a medium throughput at a very low read length ($2 \times 13bp$). A higher throughput is anticipated to be achieved with a switch to "rolony"-based amplification and longer reads are currently being worked on (source: `http://www.polonator.org`).

Promisingly, Helicos Biosciences (Cambridge, MA, USA) offer amplification-free sequencing of DNA and RNA using their HeliScope platform (source: `http://www.helicosbio.com`). Imaging billions of single molecules at a time, this sequencer might present an appealing solution for single-cell studies or other scenarios which are currently limited by the availability of sample material. Read lengths are currently still short, but are certainly going to be improved in future generations of the technology. Another real-time, single-molecule and amplification-free sequencing instrument has been developed by Pacific Biosciences (Menlo Park, CA, USA). Unlike with the HeliScope, Pacific Biosciences' focus is on longer reads with a lower throughput.

		Reads		
Name	**Technology**	**Length**	**Number**	**Time**
Illumina HiSeq 2000	bridge amplification, sequencing by synthesis, fluorescence	medium, paired	very high / 8 lanes	long
Roche GS FLX Titanium	emulsion PCR, sequencing by synthesis, luminescence	long	low	short
ABI SOLiD 3	emulsion PCR, sequencing by ligation, fluorescence	short, paired	very high / 8 lanes	long
Polonator G.007	emulsion PCR, sequencing by ligation, fluorescence	v. short, paired	medium / 8 lanes	medium
Helicos HeliScope	no amplification, sequencing by synthesis, fluorescence	short, paired	high / 25 lanes	medium
Pacific Biosciences	no amplification, sequencing by synthesis, fluorescence	very long	very low	N/A
Ion Torrent	emulsion PCR, sequencing by synthesis, change in pH	long	variable	short

Table 1.4: Overview of high-throughput sequencing platforms. Loosely based on[184, 358, 499]. Note that all values change so frequently that I decided to report qualitative rather than quantitative values. Time refers to the average time for a complete sequencing experiment, including sample preparation.

1.2.1.5 Comparison

The platforms of all mentioned manufacturers are under constant development and most of the systems are in their second or third release generation now. With every new version, reads become longer and more abundant and error rates drop further. Likewise, a gradual drop in maintenance costs thanks to optimised reagent usage has been announced by most (and delivered by some) manufacturers. **Table 1.4** compares the main platforms mentioned above.

The choice of sequencing platform should be guided by what sort of application (see next sections) a prospective user has in mind: For instance, DNA sequencing applications with the goal to assemble entire new genomes benefit from long reads which can be more easily connected into larger units. Roche's pyro-sequencers have therefore mostly been the platform of choice in this area of research. Assays of the active transcriptome for measuring gene expression changes, on the other hand, are a good example of an application in which sequencing depth is more important than read length: Even comparatively short reads are sufficient to identify transcript sequences, but a high coverage is required in order to detect even rarely transcribed genes and to pin down subtle changes in transcript counts between various conditions. The instruments provided by Illumina and Life Technologies offer the depth required for this goal. The same reasoning applies to surveys of specific small regions of the genome, such as TF binding sites or HMs. For the purposes of this thesis, I am interested in those latter types of applications, which is why I focus mostly on Illumina sequencing in the remainder of this chapter.

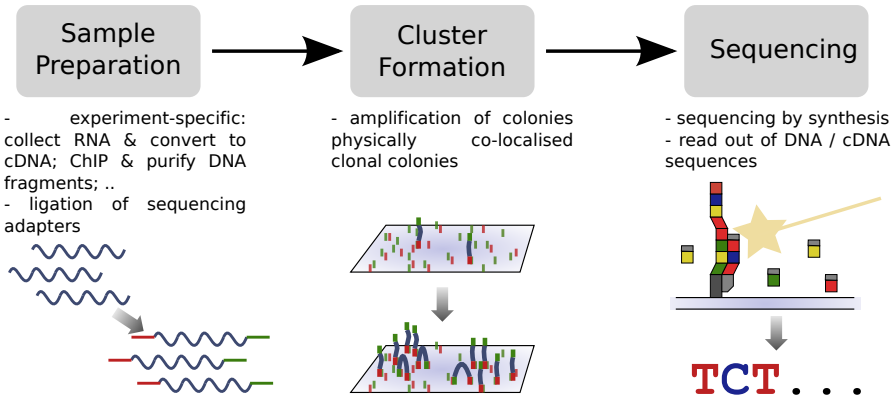

Figure 1.5: General HTS laboratory workflow. High-throughput sequencing, in the laboratory, is carried out in three steps: Sample preparation, clonal amplification of sequences and sequencing.

1.2.2 Protocols and Methodological Approaches

All HTS instruments described in the previous section can, in principle, be used to sequence DNA of any kind and from any source. I will now review the general workflow of HTS sequencing exemplified with the Illumina platform and then briefly describe common applications and methodological approaches which will be of major concern in the remainder of this dissertation. When my statements usually refer to Illumina's platforms, this is by no means intended to imply that the methods are limited to use with these instruments, but I do so merely for the sake of brevity. Equivalent processes exist equally for instruments provided by other manufacturers.

1.2.2.1 High-Throughput Sequencing by Synthesis Workflow

Before delving deeper into specific applications of HTS, let us first quickly review the general steps undergone in all HTS experiments.

The first step in every HTS workflow is the preparation of whatever biological material is to be studied in a way that makes it suitable for further processing in the sequencing instruments by the addition of sequencing adapters. Commonly, this is still the most labour- and often time-intensive step in the entire process and the only step that is truly application-specific. I will discuss different techniques in the following sections.

Sample preparation is followed by cluster (or "colony") formation, which, for the Illumina platforms, happens automatically inside an instrument called the "cluster station". Sequences that have been stuck to the solid surface of a flow cell, which is covered with a "lawn" of primers complementary to the adapters attached during sample preparation, are subjected to repeated cycles of bridge-PCR amplification. As a result, many copies of the same sequence with be physically co-located on the flow cell making it possible to more reliably read out the nucleotide sequences later on.

Once colonies have been formed, the samples are finally ready for the actual HTS. The sequencing process inside the Genome Analyser, MiSeq or HiSeq instruments works via "sequencing-by-synthesis", that is, the sequence of one (c)DNA-species is read out as a complementary strand is synthesised. Previously ligated adapters serve as sequencing primers and in repeated cycles one reversibly-terminated and fluorescently-labelled nucleotide is incorporated at a time. After each extension cycle, the latest addition to the sequence is read out with the help of a laser and high-resolution optics. Afterwards, unincorporated nucleotides and terminators are washed off and sequencing may continue into another cycle.

To summarise, the three principal steps of any HTS workflow are (1) sample preparation, (2) cluster / colony formation and (3) sequencing.

1.2.2.2 Expression: RNA-seq, DeepSAGE, miRNA-seq and GRO-seq

Large-scale assays of gene expression have for the past decade been the forte of microarrays – a position that is now increasingly being rivalled by HTS, which is offering more precise and unbiased quantification of gene expression levels and additional insights into the nature and structure of the transcriptome[340, 499, 506, 526, 585].

Transcriptomic assays using HTS may be broadly divided into four categories differing drastically in the object and aim of measurement and, as a consequence, in the protocols employed preparing the biological material for sequencing: RNA-seq, DeepSAGE, miRNA-seq and GRO-seq.

1.2.2.2.1 RNA-seq RNA-sequencing (RNA-seq) refers to the sequencing of mature RNA transcripts. In fact, it is usually reverse-transcribed cDNA that goes into the sequencing process (such is the case for the Illumina platform), although cases of direct sequencing of mRNA have been reported[407, 408].

Although alternative, optimised protocols have been developed[335, 433, 434, 636], the principal steps of RNA-seq sample preparation most commonly involve*: (1) Isolation of mRNA. (2) Fragmentation of mRNA into random pieces using divalent cations. (3) Synthesizing double-stranded cDNA. (4) End-repair, adenylation and adapter ligation. (5) Purification and amplification of cDNA with correctly ligated sequencing adapters.

Thanks to the random fragmentation of transcript sequences, RNA-seq reads (given enough sequencing depth) can span the entirety of the active transcriptome allowing, in addition to the measurement of expression levels, the option to reconstruct characteristics of the transcriptome, e.g. in order to refine gene models (alternative start / termination sites, novel exons, non-protein coding transcription), to examine the interplay between expression and

*Source: `http://grcf.jhmi.edu/hts/protocols/mRNA-Seq_SamplePrep_1004898_D.pdf`

DNA-associating factors, to assess isoform expression or even to assemble *de novo* entire transcriptomes of organisms for which no genomic annotation exists (reviewed in references[340,585]).

1.2.2.2.2 DeepSAGE The DeepSAGE strategy is an expansion of a pre-HTS expression assay called serial analysis of gene expression (SAGE), hence the name. SAGE libraries are short, fixed-length cDNA sequence tags extracted from a reverse-transcribed RNA sample by digesting the cDNA with a combination of restriction enzymes (*MmeI* and either *NlaIII* or *DpnII*). Essentially the same approach has been carried forward with advancing sequencing platforms and optimised for HTS and, although slight variations might apply, is now known by many terms which are largely used synonymously, e.g. DeepSAGE (my name of choice owing to its similarity to SAGE), massively parallel signature sequencing (MPSS), Tag-seq (for "sequencing of tags") or digital tag profiling (as per the title of Illumina's official protocols).

In short, sample preparation for DeepSAGE involves four fundamental steps[†]: (1) Isolation of poly-A mRNA and generation of double-stranded cDNA attached to a magnetic bead. (2) Addition of the restriction enzyme *NlaIII* or *DpnII* cleaves the cDNA at every recognition site (`CATG` and `ATGC`, respectively) leaving only the 3'-most fragment. (3) An adapter containing a *MneI* recognition site is attached and this enzyme then cuts specifically 17bp downstream of the adapter-cDNA link (16bp for *DpnII*) creating well-defined sequence "tags" of a fixed length. As a result of the last restriction step, the tags are now not attached to the bead any longer. (4) Finally, a second adapter is ligated at the other end of the tag and the sequences will be amplified and purified before loading them into the cluster station for colony formation and, eventually, sequencing.

Sequencing of well-defined tags as compared to random fragments of transcripts (RNA-seq, see above) brings advantages and disadvantages: On the positive side, the "search space" to be covered when sequencing tags is only a minor fraction of the entire transcriptome. It is for this reason that DeepSAGE has attracted most attention in the early days of HTS, when the instruments had not yet been advanced enough to routinely produce the depth and coverage required for unrestricted assays. But even today, if RNA is not available in abundance, e.g. in single cell experiments, the approach may still well be worthwhile to pursue. However, tags come at the cost of losing additional information about their genomic context making them largely useless for transcriptome assembly, the refinement of known gene models, genomic comparison of the interplay between expression and DNA-associating factors and the assessment of isoform expression. Moreover, transcript without poly-A tails or without restriction sites for the enzymes used (*NlaIII* / *DpnII*) cannot be detected using this approach.

[†]Source: `http://grcf.jhmi.edu/hts/protocols/1004240_GEX_NlaIII_Sample_Prep.pdf`

1.2.2.2.3 shortRNA-seq / miRNA-seq Non-protein coding transcription of short RNAs is attracting more and more attention. The study of these new species of RNAs on the genome-wide scale has been made possible only by the refinement of protocols specialised for the detection of short RNAs (including, but not limited to, miRNA). Most mature miRNAs are cleaved by *Dicer* and other enzymes that leave the RNA with a phosphate and a hydroxyl group at the 5' and 3' end, respectively. Illumina's protocols[‡] exploit this structure by using specific adapter sequences that are ligated to these ends.

Illumina (and others) also encourage the use of multiplexing for shortRNA-sequencing. "Multiplexing" refers to the addition of sample-specific "bar-code" (a short nucleotide tag) that mark all sequences from the same sample, making it possible to load multiple biological samples onto the same lane of a flow cell without losing the ability to tell where they came from. The sequencing depth of modern instruments by far exceeds what is required for the measurement of the rather limited repertoire of short RNAs and read lengths are longer than most RNAs in question (miRNAs are typically no longer than 19-25bp[366]), thus multiplexing allows for a more economical use of the technology.

1.2.2.2.4 GRO-seq Another methodology is focusing on a different aspect of gene expression: Global run-on sequencing (GRO-seq; sometimes also "genome-wide run on sequencing") aims to measure nascent transcriptional events as they happen, that is, active transcription before splicing and further processing[91]. The technique is based on the sequencing of nuclear run-on assays (NRO) which have been optimised by Core and colleagues for use in genome-wide studies[91]. NRO extends RNA that is associated with active polymerase and prohibits its elongation by removing endogenous nucleotides from isolated nuclei and adding back radionucleotides that enable actively engaged polymerase to resume elongation, while no new transcription is initiated during short run-on times[148, 439, 468]. Additionally, new initiation events are suppressed by addition of the anionic detergent sarkosyl[91, 468]. For GRO-seq, NRO-RNA is marked with a BrU-tag, which is then used to immunopurify the sample[91]. Subsequently, ends are repaired in essentially the same fashion as for shortRNA sequencing, adapters are ligated to both ends and sequencing is carried out as usual.

In summary, GRO-seq presents a promising and exciting approach to examine active transcription, pausing of elongation and promoter architecture.

1.2.2.3 Regulation and Epigenetics: ChIP-seq

Chromatin immunoprecipitation coupled with HTS (ChIP-seq) has over the recent years established itself as the primary method of choice for the genome-wide study of gene regulation and

[‡]Source: `http://genome.med.harvard.edu/documents/illumina/TruSeq_SmallRNA_SamplePrep_Guide_15004197_A.pdf`

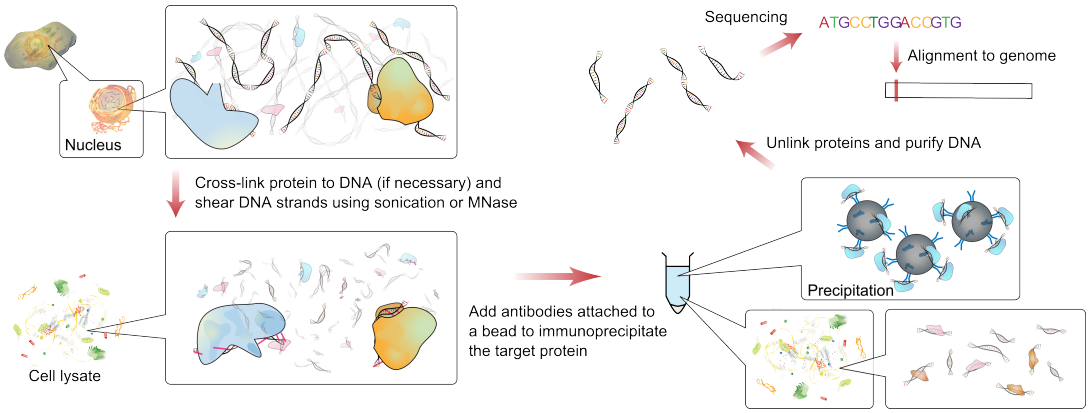

Figure 1.6: ChIP-seq laboratory workflow. Proteins are cross-linked to chromatin, which is then sheared into fragments. Fragments associated with a protein of interest are precipitated, proteins removed and the DNA is sequenced. Bioinformatics analysis identifies binding region in the genome. Adapted with permission from `http://en.wikipedia.org/wiki/File:ChIP-sequencing.svg`.

many epigenetic factors[88, 237, 263, 415, 633]. Chromatin immunoprecipitation (ChIP) is a technique whereby the binding sites of DNA-associated proteins, such as TFs, epigenetic regulators and histones, can be identified. To do so, proteins and associated chromatin are temporarily bonded, the DNA is then sheared to create small to medium-sized fragments (typically a few hundred base-pairs in length) and those fragments bound by a protein of interest are selectively immunoprecipitated with an antibody targeted at this protein and then purified and pulled out (**Figure 1.6**). One way of preparing chromatin is to reversibly cross-link sonication-sheared chromatin with formaldehyde or ultraviolet light. After immunoprecipitation, the DNA-protein cross-link can be reversed and proteins removed to leave only the DNA for subsequent processing. This technique is mainly applied for DNA-binding protein such as TFs. Alternatively, proteins that naturally link to chromatin, such as histones that wrap DNA in nucleosomes, can be investigated using native chromatin sheared by micrococcal nuclease (MNase) digestion.

Selected sequences have previously been hybridised to microarrays containing probes corresponding to regions of interest (ChIP-on-chip), but nowadays most researchers choose to utilise HTS instead in order to read out the sequences of all enriched DNA fragments. This approach offers major advantages in terms of resolution and does not require prior knowledge of putative target regions for DNA-protein association affording an unbiased screen of all genome-wide binding events. After ChIP, the HTS workflow is fundamentally very similar to RNA-sequencing approaches described before[§]: DNA ends are repaired using a combination of polymerases and the 3' end is adenylated to prepare the DNA for ligation. Adapter sequences are then added to both ends of the template. After selecting suitably sized fragments and removing excess adapters, adapter-coupled sequences are enriched by PCR and finally put

[§]Source: `http://grcf.jhmi.edu/hts/protocols/11257047_ChIP_Sample_Prep.pdf`

forward for cluster formation and sequencing as described before.

After sequencing, subsequent bioinformatics analysis can detect regions of DNA-protein association by mapping the sequence reads back to the genome and identifying enriched binding events. Unfortunately, ChIP-seq data is obscured by variations in fragment size, the inconsistent location of binding sites within these fragments (making it more difficult to pinpoint the exact location of binding) and imperfect precipitation leading to contamination of the signal with DNA or proteins incorrectly pulled out by antibodies. Downstream analysis therefore depends heavily on statistical methods to distinguish real binding from background, but nevertheless one must generally expect a high level of false positives (incorrectly identified binding events). One recent development promises to significantly reduce impurities and increase resolution: Research led by Rhee and Pugh at the Pennsylvania State University applied a lambda exonuclease to immunoprecipitated chromatin[449]. The lambda exonuclease digests unbound DNA starting from the 5'-to-3' direction, which gets rid of contaminating DNA and ensures that each sequenced read ends at the position of actual DNA-protein binding. I would expect that future research will increasingly make use of this technique to improve the quality of TF binding assays and the like.

1.2.2.4 Others

RNA-seq and ChIP-seq are the two methodologies of most relevance to the work described in this thesis, but many other application areas for HTS exist and have attracted an equal amount of attention from the community. Traditionally, sequencing has been applied to determine the sequence of genomic DNA (Human Genome Project: `http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml`). HTS has taken this endeavour to the next level, making it possible for smaller institutions or even individual research groups to compile entire new genome assemblies of up to mammalian scale[158, 482]. Similarly, genomic re-sequencing efforts are now routinely employed to improve the quality of existing assemblies and to discover genomic variations, often linking them to phenotypic effects and disease[383, 482].

Other interesting applications include the immunoprecipitation of protein-bound RNA (RIP-seq / CLIP-seq / HITS-CLIP)[317, 616, 634], the identification of miRNA targets (Argonaute HITS-CLIP)[78], sequencing of ribosome-protected mRNA (ribosome profiling)[218, 643] and the profiling of DNA methylation (Methyl-seq / Bisulfite-seq)[324, 360].

1.2.3 Applications to Stem Cell Biology

The majority of early work in next-generation sequencing has focused on the evaluation of the technology as a tool for gene expression analysis, the discovery of TF binding sites and the analysis of chromatin signatures[340, 367, 506, 526, 585], but since then the number of publications

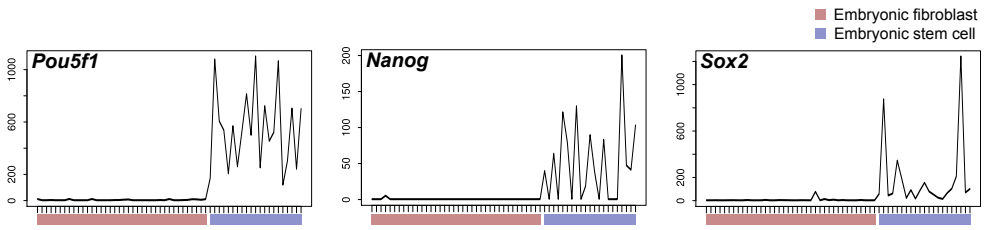

Figure 1.7: Single-cell expression of pluripotency markers. Expression levels of the stem cell-related TFs *Pou5f1*, *Nanog* and *Sox2* (from left to right) in single embryonic fibroblasts (red) and embryonic stem cells (blue) from a reanalysis of reference[223]. The variability is partly explained by technical differences, although for this plot only fairly high coverage libraries have been chosen (number of aligned reads $> 250,000$). Expression values are given as reads per kilobase-million (RPKM[367]).

making use of HTS technology has grown exponentially up to a point where it would make little sense to enumerate all of them. Instead, I will focus on the most influential literature making use of the technology to its best potential.

1.2.3.1 Gene Expression

Early adopters of HTS in the area of gene expression profiling focused mostly on establishing the technology as an alternative to microarrays. One study[465] measured gene expression levels in ESCs using DeepSAGE and Illumina microarrays and found a satisfactory degree of concordance between the measurements, but reported a higher dynamic range for the HTS-based assay. This observation was also confirmed by another study that compared expression levels in ESCs and embryoid bodies (EBs)[84], where the RNA-seq was able to detect the expression of almost $4,000$ genes that had previously been considered not expressed. Moreover, the authors stressed that they found evidence for a considerable degree of transcription ($31 - 37\%$ of all reads) outside annotated exons, that is, either from intronic regions or from intergenic regions of the genome. Most of this unexplained signal might not have been picked up before, because it was at a very low level. With this work, the researchers demonstrated that HTS is able to deliver a profile of mammalian transcription with an until then unseen level of coverage and accuracy.

More recently, several groups have begun to exploit the sensitivity of HTS for studying **gene expression in single cells**. In a number of pioneering studies, Tang and colleagues have first developed optimised methodologies and demonstrated their feasibility for the study of mRNA expression in single mouse blastomeres and oocytes[294,534–536] and then used this technique to follow up on transcriptional changes observed during the transition from blastocysts from the inner cell mass (ICM) to pluripotent ESCs *in vitro*[532]. They discovered an increasing expression of repressive epigenetic regulators coupled with a drop in the expression of activating regulators in the course of the transition. They also identified several differentially expressed miRNAs that were predicted to target differentiation- and pluripotency-related

genes, consistent with the change in cell state. Taking single-cell expression analysis to the next level, researchers are now trying to exploit the power of multiplex sequencing (**Section 1.2.2.2**). Initially, RNA from 48 single ESCs and 44 single embryonic fibroblasts (EFs) was subjected to this approach[223]. My own reassessment of this data showed that signal intensities were not consistently reliable across all sequenced cells: Even markers of pluripotent stem cells failed to be detected at all in some ESC samples (**Figure 1.7**). I believe this is mostly due to massively variable depths of sequencing of individual libraries and is likely to be resolved with use of the latest equipment and better balancing of the independent, barcoded samples (that is, by achieving a uniform split of cluster formation and sequencing depth across all libraries). As a result it is difficult to say which differences are due to actual biological variation within cell populations. Nevertheless, the study demonstrated impressively the feasibility in principle and laid the path for exciting future studies that will help to better understand transcriptional differences between individual cells.

Yet another use of the technology is the **assembly of transcriptomes**. One group demonstrated that it was not only possible to accurately reconstruct established transcriptomes from RNA-seq data[179], but that the transcriptomes of ESCs, lung fibroblasts and neural precursors were remarkably variant in the use of transcription start and termination sites and of alternatively spliced exons. Moreover, the authors identified a large number of cell type-specific large intergenic noncoding RNAs (lincRNA). Several lincRNAs were later on shown to have a major effect on the expression of pluripotency and differentiation genes[178], which – together with other studies[59,262,427,428,582] – has brought a new class of key regulatory elements to the attention of the research community. Similar observations were reported in a study following gene expression changes during neural differentiation of ESCs[602]. The researchers noted an astounding complexity in gene expression going beyond simple differential expression of genes: While ESCs were reported to express a wide variety of different isoforms of the same gene, it had been observed that many genes expressed a more restricted range of isoforms in increasingly committed stages of the differentiation process ("isoform specialisation").

1.2.3.2 Transcription Factors

In the past years, **ChIP-seq experiments targeting transcription factors** have expanded our knowledge about the transcriptional circuitry of ESCs[49,75,268,327,342,497]. In two of the most well-known studies to date, Chen and colleagues[75] and Marson and colleagues[342], investigated the binding profiles of the TFs *Pou5f1, Sox2* and *Nanog*, as well as several other important genes in ESCs. The ChIP analyses not only revealed potential downstream targets of important stem cell-related TFs, but additionally showed that some of them co-occupy

binding sites forming genomic clusters that might act as enhancers[¶]. Many clusters were also found to be associated with H3K4me3, generally believed to be a mark of active elements of the genome (**Section 1.1.5**). Interestingly, clusters formed by *Pou5f1*, *Sox2* and *Nanog* were also noted to associate with the transcriptional co-activator *Ep300*, further supporting their relevance[75]. Apart from the "core pluripotency cluster", formed by *Pou5f1*, *Sox2* and *Nanog* as well as *Smad1* and *Stat3*, a second set of TFs were found to frequently cluster together: *Myc*, *Mycn*, *Zfx* and *E2f1*[75]. It was also demonstrated that many miRNAs in ESCs seem to be controlled by ES-specific TFs[342]. miRNA promoters that were co-occupied by TFs and by polycomb group proteins were not active and could thus be believed to be in a poised state "ready" for expression[35,50]. Indeed, they were shown to be selectively activated in different cell types (tested with embryonic fibroblasts and neural precursors). On these grounds, miRNAs are believed to support stem cell pluripotency by fine-tuning the expression of differentiation-related regulators with the effect of suppressing differentiation signals while maintaining genes in a poised state. Many subsequent studies integrated further elements into the TF network of ESCs. For instance, two independent studies addressed the binding of the factors *Nr5a2*[198] and *Prdm14*[332], both of which have recently emerged as genes blocking differentiation (**Section 1.1.4**). The findings from all these studies have helped to augment our insight into the complex interactions of the heterogeneous factors controlling many aspects of the biological state of cells.

Of course, there are also numerous surveys of **DNA-protein interaction profiles in human cells**. For instance, one group of researchers[284] studied the TFs *POU5F1* and *NANOG* in human ESCs and compared their findings to the binding of corresponding proteins in mouse[75]. Surprisingly, they discovered that only a small fraction of *POU5F1* and *NANOG* binding sites were conserved across both species (about 4% and 5% of high-confidence binding sites for *POU5F1* and *NANOG*, respectively). In contrast, 50% of *CTCF* binding sites were conserved between both species. Other noteworthy experiments in human include the tracing the differentiation of human ESCs into definitive endoderm in an *in vitro* model[539] and the investigations into the TF network behind murine haematopoietic development[595,596]. Teo and colleagues identified *EOMES* as a candidate TF driving differentiation-specific expression events[539]. Overexpression of *EOMES* activates target genes that initiate spontaneous differentiation in self-renewal conditions. Many functional binding sites were further found to be shared with *SMAD2/3* (effectors of Activin/Nodal). Interestingly, ChIP-seq analysis by another group in mouse ESCs revealed dose-dependent binding (and effects) of *Smad2* directing cells to different fates[300], suggesting that *Eomes*-guided differentiation might also be present

[¶]All ChIP-seq experiments have been performed on populations of cells and from the data presented in the paper it is not possible to conclude whether the apparent co-occupancy of TFs does ever occur at the enhancer elements of the very same cell. This caveat applies to all ChIP-seq datasets presented throughout this thesis. One way of resolving the question whether two proteins do indeed physically co-occupy binding sites is the use of sequenctial ChIP[71,141,355,554].

in mouse. The studies by Wilson and colleagues[595,596], on the other hand, looked at TFs involved in haematopoietic specification, finding groups of cooperatively acting TFs similar to what had been found in mESCs[75]. It appears that the combinatorial control of expression by groups of TFs is a recurrent and conserved pattern of transcriptional regulation across species and cell types.

1.2.3.3 Polymerase Activity

RNA polymerase II (Pol II) is required for the expression of mRNA precursors of all protein-coding genes and many short RNAs. It is generally believed that Pol II is recruited to the promoters of genes by an integral network of TFs[203,431].

Utilising a protocol developed earlier[91], Min and co-workers quantified **RNA polymerase that was actively engaged in transcription** in ESCs and embyronic fibroblasts (EFs)[362]. The technique is now referred to as global run-on sequencing (GRO-seq) and may be used to investigate nascent transcription. Observed differences in GRO-seq density across gene bodies, in general, agree with microarray and RNA-seq measurements, but the authors also noticed a large number of genes with significant accumulations of polymerase in their promoter regions (with a peak approximately 30*bp* downstream of the TSS), both in ESCs and in EFs. They reasoned that this was indicative of paused polymerase and that entry into productive elongation was a rate-limiting step for the transcription of many genes. They further report that genes with an activating H3K4me3 mark exhibit higher and those with a repressing H3K27me3 mark lower levels of nascent transcription than the average. Strikingly, genes that have both marks ("bivalent genes") tend to have a high 5'-proximal density of aligned GRO-seq reads representing paused polymerase supporting the notion of transcriptionally poised genes[35,282,361]. Going even further, it had been noted that this holds in particular for genes targeted by *polycomb recruiting complex 2 (PRC2)*, but not *PRC1*. Conversely, genes bound by both showed neither active nor paused polymerase. These findings were thought to support the argument that PRC2 blocks transcription post-initiation, while PRC1 blocks it pre-initiation. Tackling polymerase pausing from a different angle, Rahl *et al.* generated ChIP-seq data for Pol II and related proteins[437]. Pol II occupancy at the TSS correlates highly with *NelfA* and *Supt5h* ("pause factors"). *Ctr9*, a subunit of PAF1, which is involved in elongation, on the other hand, was found inside gene bodies. It was shown that the TF *Myc* might be actively releasing Pol II from its pause. Loss of *Myc* arrests many genes in the paused state, whereas loss of *Pou5f1*, for example, disrupts transcription of target genes at an earlier stage such that in many cases even the promoter-proximal accumulation of Pol II disappears. In summary, both studies have greatly helped to advance our understanding of the transcriptional machinery and, in particular, of the role of the TF *Myc* in allowing transcription elongation to occur.

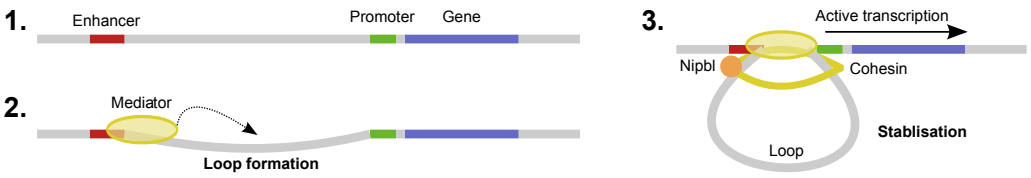

Figure 1.8: DNA loop formation by mediator / cohesin. Distal enhancer elements can be located tens or even hundreds of kilobases from a gene's promoter. Binding of mediator to the enhancers initiates loop formation further supported and stabilised by cohesin and *Nipbl*, bringing the enhancer close to the promoter in order to activate gene expression[245]. Figure inspired by reference[397].

While much research has gone into the workings of RNA polymerases I and II, our **understanding of RNA polymerase III** (Pol III) is less well developed. Pol III is responsible for the transcription of non-protein coding DNA into ncRNA, e.g. tRNA, 5S rRNA[24,108] and miRNAs[46,409]. In doing so, Pol III is necessary for cell viability and forms an essential part of the larger transcriptional machinery providing "ingredients" required, in the end, also for protein-coding gene expression. Encouraged by previous findings for Pol I and II[25,361,576], a team of researchers led by R.J. White and K. Zhao investigated the interplay of chromatin structure and Pol III using a combination of RNA-seq and ChIP-seq[26] in matched cell types (CD4+ T and HeLa cells). In brief, H3K4 methylation and H3K4, K9, K27 and K36 acetylation were all linked to active Poll III, while H3K27 and K9 methylation were associated with inactive Pol III (these mechanisms are the same for Pol II). Unlike Pol II, Pol III sites, however, lack H3K79me2 and H3K36me3. Even more surprisingly, Pol II was present at many Pol III sites, with one possible explanation being that some TFs might recruit both polymerases (e.g. *MYC*). The study provides much additional detail on Pol III and Pol III activity and their epigenetic landscape and will certainly serve as an important resource for future research.

In other research into the workings of the transcriptional machinery, one group was interested in the **role of mediator and cohesin**[245]. It has been suggested that enhancers driving active expression of genes are physically brought closer to the target genes promoter by the formation of DNA loops[235,359]. Cohesin is a candidate that can form such loops and mediator, which is a transcriptional co-activator interacting with the transcriptional machinery and can be found at enhancer sites, interacts with cohesin giving one potential explanation for this observation and may therefore link distal TF binding functionally to the activation of transcription[89,277,334,461] (**Figure 1.8**). In-depth ChIP-seq analysis revealed that mediator (*Med1* and *Med12*) is located at promoters and enhancers regions of $> 60\%$ of all actively transcribed genes in ESCs. Cohesin complex proteins (*Smc1a* and *Smc3*) co-occupy most of these regions and cohesin-mediator co-bound region (CMCRs) do also associate with Pol II, while cohesin-CTCF co-bound regions (CCCRs) did not show an enrichment for Pol II (CTCF is another factor involved in DNA-loop formation). In further experiments, the researchers then went on to demonstrate that the three proteins physically interact and that DNA-looping

does indeed occur between enhancers and active genes in ESCs (e.g. *Pou5f1* and *Lefty1*), but not for inactive genes (tested for the same genes in EFs). The model of transcriptional regulation put forward by this study is likely to inspire and profoundly influence much future research.

1.2.3.4 Epigenetics

One of the areas that probably received most attention and has greatly benefited from HTS technology is the genomic survey of epigenetic regulatory mechanisms.

In one early study, HTS was utilised to **analyse histone modifications** (HMs; here, H3K4me3 and H3K27me3) and DNA methylation in fully reprogrammed iPS cells[360]. The authors were able to show an impressive degree of similarity between the chromatin states of iPS and ES cells with a number of differentiation-related genes being bivalently enriched for both, H3K4me3 and H3K27me3, whereas they were monovalent in somatic cells or lose their enrichment for both chromatin marks completely. The same genes also show DNA hypermethylation in differentiated cells and the loss of this methylation was found to be a crucial step in the reprogramming process. The authors hypothesised that de-methylation might be inefficient and managed to show that addition of DNMT encourages reprogramming by helping cells to escape from a state in which they were still trapped on a partially differentiated level due to methylation of pluripotency-related genes. Further research into HMs, investigated how HMs contribute to ESCs, trophoblast stem cells (TSCs) and extraembryonic endoderm stem cells (XENs)[472]. They found that trimethylation of H3K4 (H3K4me3) exhibits a largely similar distribution across all cell types with a similar number of enriched regions generally located near the TSS of known genes. The repressive trimethylation of H3K27, on the other hand, displayed distinct patterns depending on the lineage: TSCs and XENs had about 7- to 5-fold lower number of sites enriched for H3K27me3 than ESCs and of those substantially fewer were located near the TSS of genes. Concordantly, bivalent domains in TSCs and XENs were also rare; evidently, as the authors point out, alternative epigenetic mechanisms must regulate expression in extraembryonic lineages. The authors identify H3K9me3 as one candidate. In summary, the study presents evidence for the importance of epigenetic modifications for early development and suggests that some of these modifications might indeed be crucial for the establishment of different lineages.

Setting out to build up a **map of DNA methylation states** in human ESCs (H1 cell line) and fetal lung fibroblasts (IMR90), one group of researchers also performed MethylC-seq experiments[324]. Briefly, the technique uses sodium bisulfite to convert unmethylated cytosines to uracil; uracil does not usually occur in DNA, so this information can be used to distinguish methylated and unmethylated cytosines. Comparison to TF binding data, revealed a marked decrease in methylation at the sites bound by one or more TF. Overall,

the fibroblast genome was more strongly methylated at *P300* and *SOX2* sites (in comparison to H1 ESCs), but showed no global difference at binding sites of the other factors. Markedly, non-CG methylation was limited almost entirely to ESCs. The authors further discovered that most non-CG methylation was more prevalent in gene bodies rather than their promoter regions and that higher methylation favoured stronger transcriptional activity. Finally, non-CG methylation, which appeared to have been lost in fibroblasts, was efficiently restored in iPSCs. This study is one of the first examples of a genome-wide, base-pair resolution examination of a mammalian methylome – with a sequence coverage, provided by the HTS technology, allowing to measure in an unbiased manner 94% of all cytosines in the human genome.

Numerous studies made use of ChIP-seq technology to investigate the **proteins that bestow these epigenetic profiles** on ESCs: Ho *et al.* investigated the chromatin remodelling complex esBAF by targeting its core component *Smarca4* (also known as *Brg*) and found functional interactions with *Pou5f1* and *Sox2*[200]. Another chromatin remodelling factor, *Chd7*, was found to co-localise with the same factor at enhancer elements[485]. Walker *et al.* identified and studied the *polycomb repressive complex 2* (*PRC2*)-member *polycomb-like 2* (*PCL2*; official name: *Mtf2*) and link its loss to differences in histone methylation and impaired differentiation (coupled with stronger self-renewal) in mouse ESCs[574]. Two groups found that *Jarid2* associates with *PRC2* and mediates the repression of its target genes, e.g. impairing the down-regulation of *Pou5f1* and therefore ESC differentiation[308, 417]. Lastly, research into the epigenetics of DNA methylation by Wu and colleagues[600], looked closely at *ten-eleven translocation protein 1* (*Tet1*), which converts 5-methylcytosine to 5-hydroxymethylcytosine and shed light on its role in DNA methylation, promoting pluripotency TFs and its involvement in the repression of polycomb targets.

1.2.4 High-Throughput Sequencing Paves the Way for Functional Genomics Research

In the previous section, a short overview of just a small selection of recent research has been presented. None of this work would have been possible without the use of HTS technology. The sheer pace with which the biological research community has embraced this new method is truly amazing and I have always been excited to be a part of this movement. In the next chapter, I will take the reader back in time to when I started working with HTS data with the aim to explore its potential for stem cell biology. The discussion about the advantages (and challenges) connected with this technology shall therefore be postponed until after this next chapter.

Chapter 2

Exploring the Potential of High-Throughput Sequencing

In late 2008, at the outset of the work described in this dissertation, high-throughput sequencing technologies (**Section 1.2**) were still in their infancy. Several suppliers had now started to actively market their individual platforms to the mass-market and initial reports from the literature reported impressive and promising results in terms of accuracy, coverage, flexibility and cost-efficiency[526, 585, 597]. I sought to assess the potential of this emerging technology for the study of gene expression and regulation in stem cell research and therefore carried out a series of exploratory studies in collaboration with various other research labs, which shall be portrayed in this chapter.

2.1 Global Expression Analysis of *Nanog*-Deficient Embryonic Stem Cells

In an initial effort, we conducted a pilot study in collaboration with Prof. Ian Chambers (Institute for Stem Cell Research / Centre for Regenerative Medicine, University of Edinburgh). Prof. Chambers' group studies ESCs, the molecular mechanisms of pluripotency and, in particular, the role of the transcription factor *Nanog*, a well-established member of the core transcriptional network of ESCs[67, 68, 70, 411] (**Section 1.1.4**). In their pursuit of a better understanding of the functional implications of *Nanog* activity, Chambers and colleagues have established numerous cell lines with experimentally modified levels of *Nanog* expression (stable and inducible), constituting a powerful system for the study of downstream targets of this transcription factor.

Gene expression in a selection of these cell lines was profiled using HTS. In this section,

I will first outline our motivation for doing so (**Section 2.1.1**), then explain in detail the experimental design and methodology (**Section 2.1.2**) and lastly discuss some of my findings and highlight conclusions drawn from this work (**Section 2.1.3**).

2.1.1 Motivation and Goals

As previously mentioned in the introduction to this chapter (**Chapter 2**), when I began working with HTS data, the technology was still poorly understood and initial reports, albeit promising, fell short of providing a convincing account of its value for actual biological research. Taking advantage of the fact that the University of Edinburgh's sequencing facility, the *GenePool*, had recently acquired a new *Illumina Genome Analyser* instrument, I sought to collaborate with a local research group to set up a pilot study.

I was fortunate enough to be situated in the same department with Prof. Ian Chambers, who shared my sceptical enthusiasm with respect to the emergent technology. Having previously attempted to quantify *Nanog*-dependent gene expression globally using a microarray assay from two of their cell lines, the Chambers lab had now undertaken to repeat this screen using the new technology. Several other research groups had also earlier sought to identify transcriptional effects on *Nanog* target genes using knock-down assays. Taken together, the existing data provided a good starting point for validation. Additionally, I was interested to see whether one could identify any further candidates.

2.1.2 Methodology

I shall now describe the experimental and analytical methodology employed in the execution of this pilot study.

2.1.2.1 Experimental Design

For the work in this pilot study, two cell lines were chosen, RCN(t) (short: NT) and RCNβH(t) (short: BT12)[68], representative of *Nanog*$^{+/-}$ and *Nanog*$^{-/-}$ mutant ESCs, respectively. These were the same cell lines for which also Affymetrix microarray was available (I. Chambers, unpublished data) thus making an ideal case for a validation study.

For both cell lines, two cultures were grown and total RNA was harvested independently, i.e. experiments were performed with two biological replicates each. Replication would make it possible to assess the variation in observed gene expression intensities, providing a first estimate of the reliability and repeatability of measurements and enabling the use of statistical tests to calculate metrics of significance for the differential expression of genes between the two cell lines.

Dataset	Cell Line	Genotype	Total Reads
NT-S	RCN(t)	$Nanog^{+/-}$	$3,265,654 \times 50bp = 163.3mb$
NT-L	RCN(t)	$Nanog^{+/-}$	$7,801,625 \times 50bp = 390.1mb$
BT12-S	RCNβH(t)	$Nanog^{-/-}$	$3,724,383 \times 50bp = 186.2mb$
BT12-L	RCNβH(t)	$Nanog^{-/-}$	$6,806,832 \times 50bp = 340.3mb$

Table 2.1: Cell-lines / datasets used in the pilot study. Overview of all cell lines datasets used in the pilot study.

The total RNA samples were submitted to the GenePool core facilities at the University of Edinburgh, who performed sample preparation and sequencing on an Illumina/Solexa Genome Analyser platform (first generation) according to the manufacturer's digital tag profiling protocol ("DeepSAGE", **Section 1.2.2.2**).

2.1.2.2 Development of an Analysis Pipeline

Much of my initial work focused on setting up an appropriate analysis environment by finding available tools, comparing and evaluating them and on filling in gaps by writing custom pieces of computer code. Inspired by some early publications[367, 465, 526], I identified as the key steps in the analysis process the assessment of the raw data quality, the alignment of short reads to a reference genome assembly and the quantification of gene expression and the comparison of expression patterns between different sample groups (**Figure 2.1**).

The data at hand was produced by following Illumina's digital gene expression protocol (**Section 1.2.2.2**), often also referred to as massively parallel signature profiling (MPSS), Tag-seq or DeepSAGE. As described earlier, this protocol targets well-defined short subsequences of transcripts. Although the sequenced libraries reported read sequences of a total length of $50bp$, actually only the first $17bp$ contained biologically meaningful information. On the other hand, it was known *a priori* that all tags neighboured a `CATG` sequence, that is, a *NlaIII* recognition site, so it was sensible to use this information to complete the tag sequences. Thus, the first processing step was the truncation of read sequences to a fixed length of $17bp$ followed by an extension using the nucleotide letters `CATG`, resulting in the final $21bp$ tag sequences subjected to further processing.

Next, the quality scores (discussed later in **Section 3.3.3.1**) of the tags were examined (**Figure 2.3**). I summed up all quality values for each 17bp-tag sequence individually and discarded those reads that had a cumulative score of less than a certain threshold T. I decided to use the cumulative quality score rather than the minimum quality score across the read as a quality control criterion specifically so to accept even reads in which a single base call might be incorrect. In the alignment strategy employed in the following step such errors are accounted for by accepting a limited number of mismatches between the bases in the read sequences and those in the reference. The quality score values corresponding to mismatched

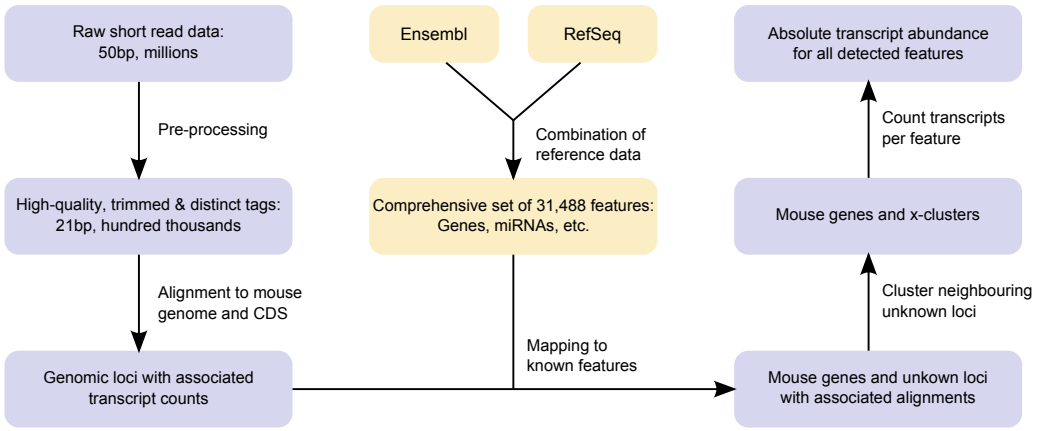

Figure 2.1: Alignment and mapping of DeepSAGE data. A schematic overview of the analysis of the Chambers lab DeepSAGE data.

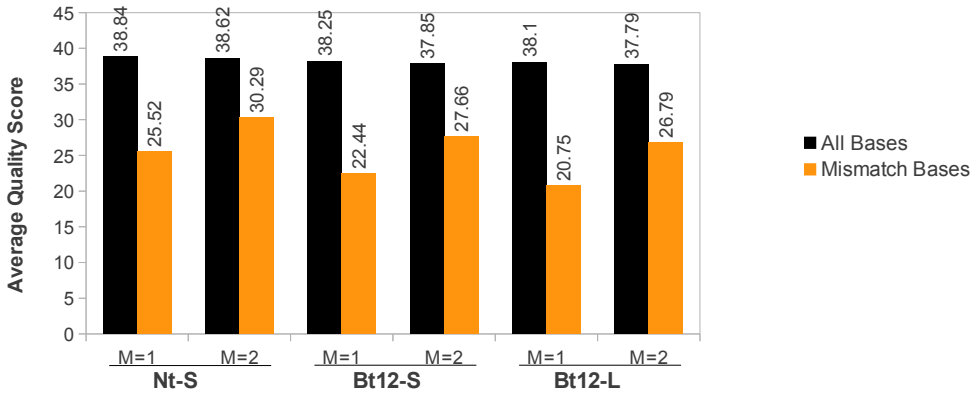

Figure 2.2: Quality of Mismatched Bases. Average quality scores across all bases of all aligned reads (black bars) and across only those bases that aligned with mismatches (yellow bars). This plot was created at a later point in time and uses, unlike the other data presented in this chapter, not a quality score scale ranging from 0-80, but instead one ranging from 0-40.

bases are markedly lower than the average across all aligned bases (**Figure 2.2**), suggesting that the primary reason for mismatches are indeed errors in the base calls during the sequencing process. Using the cumulative quality score threshold, either multiple bases need to be of a very low quality or the overall quality of all bases in a read would have to be rather bad. In both cases, alignments coming from sub-threshold reads could not be trusted and might obscure the signal measured.

The threshold was set to $T = 925$ (using the Illumina scale of quality scores, which ranges from about 0 to 80 per base) in order to discard an average of some 3% of all reads, which amounts to approximately the percentage of reads which would be expected to have an unreliable sequence[499]. All remaining reads were clustered into bins according to their sequence. A record of the total number of reads per cluster was kept and I then passed on only the distinct

sequence tags to the next processing step. Notably, the overall amount of data to be handled could be reduced to about a tenth on average by clustering identical tags.

The sequences of the distinct tag clusters were then aligned in several steps to the mouse reference genome using the Bowtie alignment software[292]:

1. Find all perfect matches (that is, alignments without mismatches) in the mouse genome (NCBI build 37).
2. Find perfect matches of all unaligned tags to known coding sequences from Ensembl[136].
3. Find perfect matches of all unaligned tags to known RefSeq mRNAs[430].
4. Repeat steps 1-3 allowing for one mismatch in the tag sequence.
5. Repeat step 1 with two mismatches in the tag sequence.

At each step I discarded all tags that were mapped to more than ten different genomic loci as highly repetitive and unlikely to yield any usable biological signal. Only a small percentage of tags remained completely unaligned. Those reads are generally considered to be due to erroneous sequencing, incomplete filtering or contaminations of the samples, unconventional splicing or other post-transcriptional modifications that result in transcripts not directly matchable to the genome. Steps 2 and 3, in which sequences were aligned to known transcripts rather than the genome, can be considered a measure to account for those reads that span exon-exon junctions which could not usually be aligned to the genome due to the presence of intronic sequence not present in the tag itself. I will discuss the alignment problem in a later chapter in more detail (**Section 3.3.3.2**).

Next, it was necessary to associate the genomic loci discovered by the sequence alignment program to known genes and other transcriptional units in the genome ("features"). I have built a comprehensive set of all known features by merging annotations from Ensembl (Release 54, 5 May 2009[136]) and RefSeq[430] (as obtained from the University of California, Santa Cruz, Genome Browser on 24 March 2009). All entries with overlapping exons were merged into one single entry yielding a total of $31,488$ features, most of which correspond to canonical, protein-coding genes (others include pseudogenes, mitochondrial, ribosomal and various kinds of short transcripts like miRNAs or snRNAs). I then tried to associate each genomic locus to the closest neighbouring feature by assigning them to one of seven classes:

1. Upstream: Up to $20kb$ upstream of the transcription start site (TSS) of the closest feature.
2. Exonic: Within an exon of a feature.
3. Intronic: Within a feature, but not in an exon.

4. Spliced: Spanning the junction between two (or more) exons.

5. Downstream: Up to $20kb$ downstream of the transcription termination site (TTS) of the closest feature.

6. Undecided: Equidistant to two features.

7. Unknown: No known feature within a $20kb$ window around the locus.

I also took the strand of each locus into consideration and, if the locus was on the opposite strand of the associated feature, assigned it to the aforementioned class anyway, but marked it as "putative anti-sense".

Many short read sequences seem to stem from regions of the genome nowhere near any known feature (the "Unknown" class from above). It has been reported that up to 99% of mammalian genomes show evidence for transcription at some level[39, 63, 601]. In the past, most low-level transcriptional event have been considered transcriptional noise, but with the discovery of more and more biologically functional short transcripts, it is now becoming increasingly clear that mammalian transcriptomes are vastly more complex than anticipated[177, 192, 356, 601]. I have therefore attempted to identify regions of the genome which exhibit coherent transcription likely to correspond to biologically meaningful transcripts. To find transcriptionally active units amongst the thousands of "Unknown" loci, all loci within a maximum distance of $1kb$ to each other were merged together. I will refer to the resulting pseudo-features as "x-clusters". In the next step, the x-clusters will be considered as one feature when calculating total transcript counts.

Finally, all tags aligning to the same feature were summed up (counting only tags in classes 2-4) to obtain a total transcript count and therefore an absolute intensity value for the expression level of each feature. At this point in time, most published studies relied solely on those transcripts that could be aligned uniquely to one location in the genome for this purpose. It is, however, desirable to also take non-uniquely mapped reads into account and since then many better approaches have emerged (see **Section 3.3.3.3**). I have therefore devised a formula that assigns reads to the most likely region of origin by assigning a part of the total read count proportionally to other reads mapping in the proximity of each possible mapping location. For this purpose, I first counted all mapped reads, spreading non-uniquely mapped reads equally about all possible locations. The read counts were then adjusted by assigning the counts of mapped reads proportional to each individual feature's contribution to the total sum of all possible feature mappings. This amounts to the following formulas:

$$C_{distr}(f) = \sum_{t \in tags(f)} \frac{w(t)}{|feats(t)|}, \tag{2.1}$$

is the auxiliary feature count of uniformly distributed reads, where $tags(f)$ is the set of all tags t mapping to feature f, $w(t)$ is the weight of tag t (the number of reads representing the same tag sequence) and $feats(t)$ is the set of all features that the tag t might map to*. The final maximum likelihood feature count is:

$$C_{ml\_distr}(f) = C_{distr}(f) \sum_{t \in tags(f)} \frac{w(t)}{\sum_{\hat{f} \in feats(t)} C_{distr}(\hat{f})}. \tag{2.2}$$

A similar approach to the utilisation of non-uniquely mapped reads has previously been employed in the ERANGE software package[367], however, the toolkit is not directly applicable to the digital transcriptomics data at hand, since – in addition to the proportional read-assignment of 'multi-mappers' – it furthermore normalises transcript counts proportional to the total length of the features. This normalisation step is sensible for randomly primed RNA-seq experiments, but is less appropriate for tag-based ones, where the length of the features does not necessarily correspond to the likelihood of discovering a suitable cleavage site in the feature's sequence (remember that ,in theory, sequenced tags should stem from the 3'-most *NlaIII* cleavage site of the transcript and hence be independent of the transcript length; **Section 1.2.2.2**). For comparison across different experiments, total transcripts counts were additionally transformed to reads per million (RPM; see **Section 3.3.3.3**).

2.1.2.3 Meta-Analytic Integration of External Data

In order to further leverage the information content of the experiment and to enable more advanced conclusions, I augmented our own data with material from other published studies. Where possible, I tried to map the results of these studies to the features identified in my analysis using the identifiers available. It is important to realise that such attempts are inherently flawed, because there is usually not a one-to-one mapping between different gene reference sets (the mapping function is not "bijective"). Therefore it is impossible to rule out the loss of certain information on the way.

External expression data: Loh *et al.*[327] and Ivanova *et al.*[227] had previously aimed to shed light on the downstream targets of *Nanog* by knocking down the expression of the gene by RNA interference (RNAi) using short hairpin RNA (shRNA). Sharov *et al.*[497] re-analysed and combined both datasets to identify a more reliable set of genes affected by the TF. I decided to use this improved dataset together with our own data to obtain an even more comprehensive set of *Nanog* targets. It should, however, be noted that a certain degree of discrepancy is to be expected. RNAi represents merely a knock-down of the target gene rather than a knock-out as given by the genetic deletion of the locus in *BT12*, which will aggravate differences in the cells

*Hence, given $\rho(t, f) = 1$ if and only if a mapping from tag t to feature f exists, then: $tags(f) = \{t \in T | \rho(t, f) = 1\}$ and $feats(t) = \{f \in F | \rho(t, f) = 1\}$.

Category	Distance	#TFBS	#Features
Distal	$30kb$ US – $5kb$ US	1,326	1,052
Proximal	$5kb$ US – $1kb$ US	399	363
Promoter	$1kb$ US – $1kb$ DS	260	246
Intragenic	$1.0kb$ DS – end of transcribed region	2,828	1,949
Unassigned	$> 30kb$ US and outside transcribed regions	2,959	0
Total	**any**	**7,772**	**3,229**

Table 2.2: High-confidence binding sites of *Nanog*. Binding sites independently discovered in at least two of four ChIP experiments[75, 327, 342, 497]. US = upstream, DS = downstream.

due to different biological background. Consequently, the effects might be less pronounced or even contradictory. Furthermore, off-target effects (i.e. effects on the transcription of genes other than the targeted *Nanog*) cannot be completely excluded although the authors made every effort to ensure and demonstrate the specificity of their constructs. Another interesting experiment was carried out by Singh *et al.*[507]. In this study, ESCs were sorted according to their *Nanog* expression level into two classes ($Nanog^{high}$ and $Nanog^{low}$) and the two sub-populations were examined for differences in their expression profiles using Illumina bead arrays. It has been reported that $Nanog^{high}$ cells express markers of pluripotent ESCs, while $Nanog^{low}$ cells express primitive endoderm markers, in particular *Gata6* which is said to be expressed mutually exclusively of *Nanog*. Similar trends should be observed between the cell lines in this experiment, however, one would not expect all measurements to agree: As for the previous knock-down studies, variable *Nanog* dosage does not necessarily have the same effect as the complete loss of *Nanog*. Moreover, Singh *et al.* cannot rule out that the cells in their cell populations have started to differentiate after sorting. Thus, their comparison might partially reflect differences between ESCs ($Nanog^{high}$) and differentiated progeny (parts of $Nanog^{low}$).

Chromatin immunoprecipitation (ChIP): In addition to external expression data, a large body of *Nanog* protein-DNA binding data obtained from four ChIP experiments was incorporated into the analysis[75, 327, 342, 497]. All four studies sought to identify *Nanog* binding sites using a combination of ChIP and subsequent sequencing of bound genomic regions. ChIP has been used extensively and successfully in the past to identify transcription factor binding sites (TFBS; **Section 1.2.2.3** and **Section 1.2.3.2**). I compiled a catalogue of all *Nanog* binding sites by overlaying the sites identified in the invividual experiments. After converting all TFBS coordinates to the latest assembly of the NCBI mouse reference genome (build 37) using the UCSC's LiftOver tool[283], merging the datasets yielded a total of 25,086 putative binding sites. I proceeded by considering only those TFBS with supporting evidence from at least two of the four studies to obtain a set of the most reliable binding sites. A TFBS was considered to be supported in multiple studies if they overlapped in at least $1bp$. The resulting set, which I call `NanogTFBS`, contains 7,762 TFBS.

Finally, all the sites in `NanogTFBS` were mapped to the closest feature in the combined set

of all features identified in any of the datasets of our study, including all clusters of unknown transcripts and recorded the distance of the centre of each TFBS to the transcription start site (TSS). In doing so, I allowed for a maximum distance of $30kb$ upstream the TSS or any distance within the feature itself downstream of the TSS and discarded all binding sites not falling within these bounds (**Table 2.2**). The distribution of TFBS with respect to gene targets agrees well with the one reported before[327].

2.1.3 Results

I will now discuss the results of the pilot study. After presenting the primary results of the analysis pipeline outlined in the previous section, I will address the comparison to microarray data and finally highlight some biological findings.

2.1.3.1 Quality and Genomic Coverage

The majority of the short reads from all four datasets could successfully be aligned to the mouse genome using the pipeline described in the previous chapter. Of the $96 - 99\%$ of reads that passed quality control in each sample (**Figure 2.3**), on average just above 60% could be mapped unambiguously and a further 30% with minimal repetitiveness (**Figure 2.4**). It is, however, necessary to remark that the large, wild-type sample (NT-L) constituted an exception in this case, with significantly less tags aligning to the genome – only 30% were aligned uniquely and about a quarter could not be aligned at all. I will point out a few more odd features of NT-L in this section and focus entirely on this sample in the **Section 2.1.3.2**.

Interestingly, filtering the reads according to their cumulative quality values, reduced the overall amount of tags more drastically than the entire read pool, e.g. while only 3.31% of all reads in *BT12-L* were discarded, a striking 12.36% $(83,301)$ of all distinct tags did not pass the quality control. In other words, many of the reads that are filtered out are those that are singletons or have only been reported a few times. I believe that it is more likely that these singletons arise from errors in the technology than sequences that have been read out many times, hence the removal of sub-threshold reads is thought to improve the overall quality of the data by removing erroneous signals.

Most regions were only covered by one or a few reads (**Figure 2.5**), but it should also be noted that a number of regions were detected that had several tens of thousands transcripts associated to them. This demonstrates the vast dynamic range of the sequencing technology for the detection of gene expression: Expression levels could be detected over almost five orders of magnitude.

As expected, the majority of tags appear to be transcripts from known protein coding genes (**Figure 2.6**). The overall distribution is remarkably similar for all samples (data not shown).

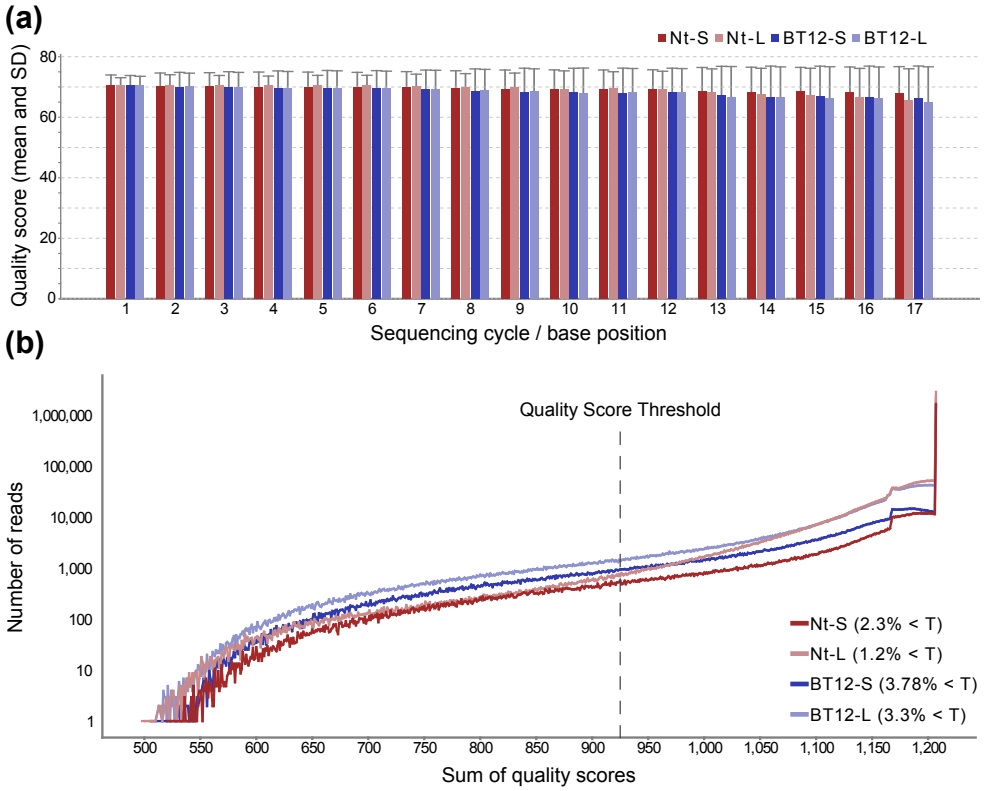

Figure 2.3: Quality scores of DeepSAGE libraries. The average quality score drops slightly with advancing read cycles (**a**), but remains at a very high level of confidence. Accordingly, the vast majority of all reads passes a cumulative quality threshold of $T = 925$ (**b**).

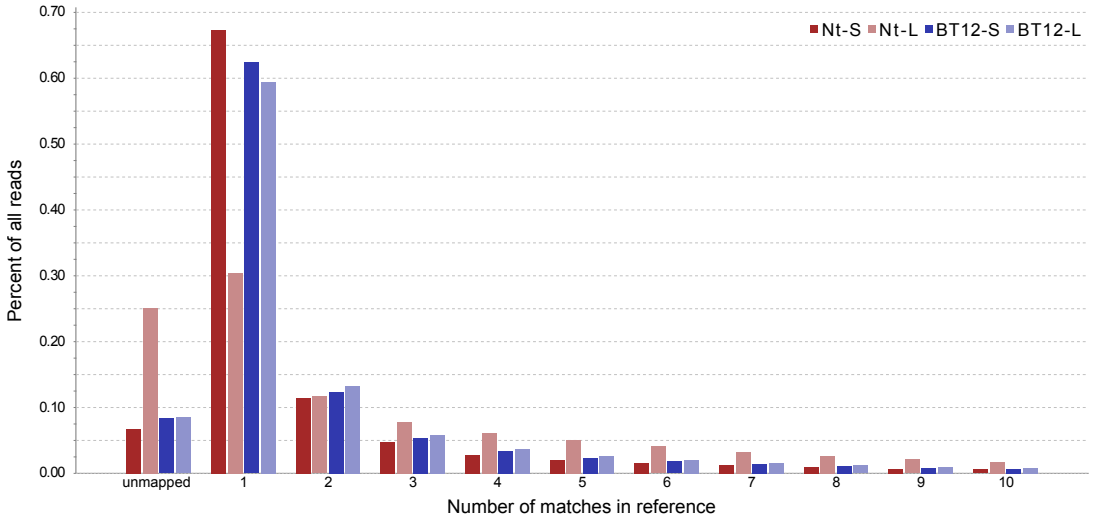

Figure 2.4: Alignment of DeepSAGE libraries. About $60 - 70\%$ of all reads could be aligned unambiguously to the reference genome and only about $4 - 8\%$ could not be aligned at all. Exceptionally, NT-L had an extraordinarily high number of unalignable and ambiguous reads.

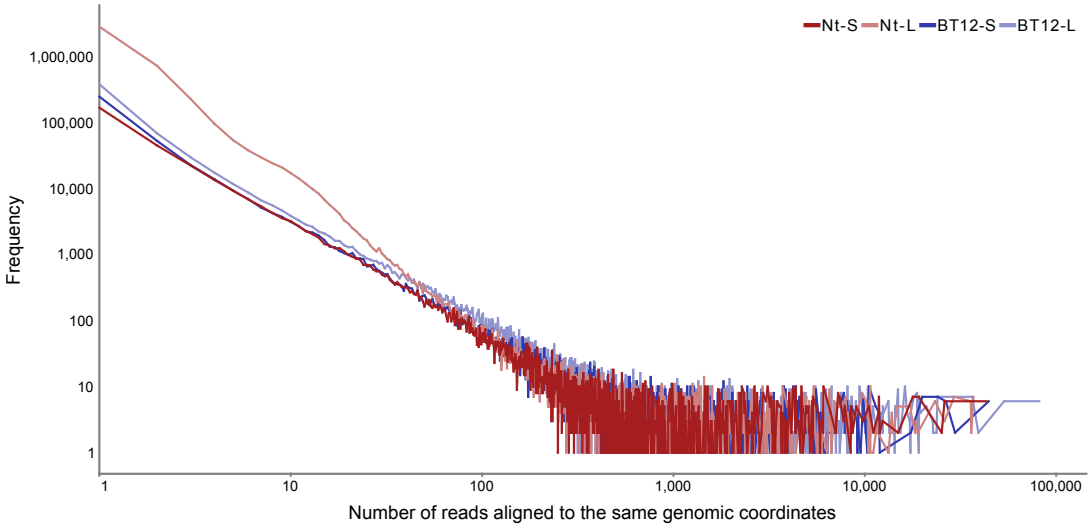

Figure 2.5: Tag frequency distribution of DeepSAGE libraries. The vast majority of genomic regions is only covered by 1-10 reads. Only a few regions have a coverage of several thousands of reads. NT-L has a particularly high number of genomic regions which are covered by 1-3 reads only.

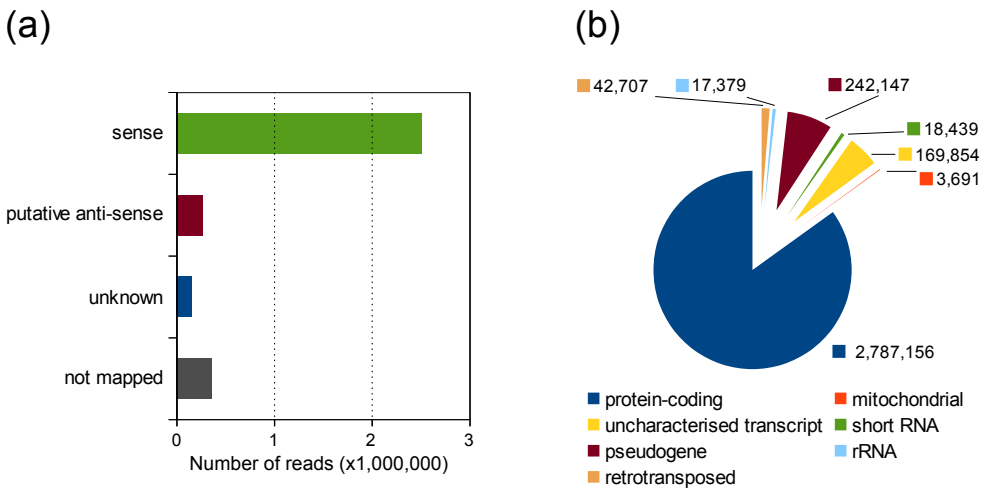

Figure 2.6: Categories of features measured. A representative example, BT12-S, for the categories of features detected in the pilot study. (**a**) Most transcripts were found to originate from regions corresponding to the sense strand of known transcripts. A considerable part appeared to belong to anti-sense transcripts. (**b**) The vast majority of transcripts comes from known protein coding genes, accompanied by some transcribed pseudo-genes and short RNAs (e.g. miRNAs).

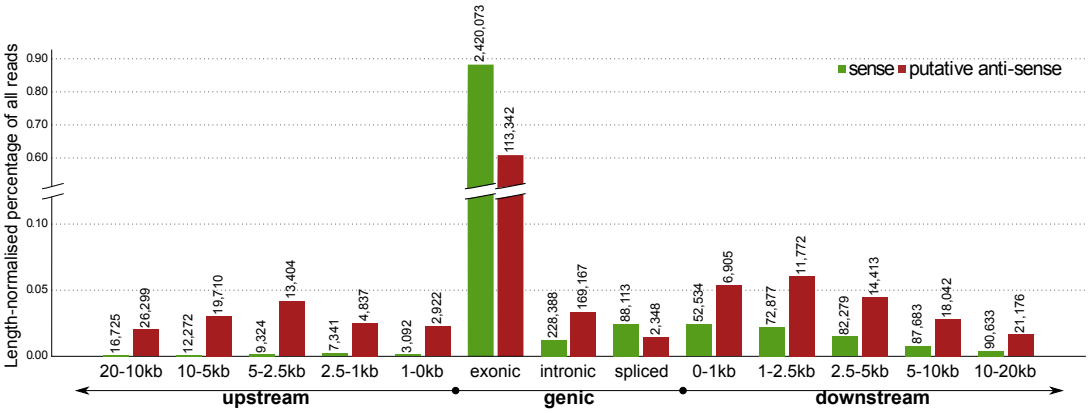

Figure 2.7: Distribution of transcripts across known gene models. All mapped tags were assigned to one bin with respect to their location to the next known feature. The total counts were summed up and normalised to a comparable measure by dividing by the length of each bin.

Besides protein-coding genes a further large group of transcripts were mapped to pseudo-genes, many of which have been known to be transcribed, but do not encode for proteins, others might indeed code functional proteins that have not yet been discovered. The remainder of the tags was mapped to either short RNAs, such as micro-RNAs and short nucleolar RNAs, mitochondrial genes or did map to regions of the genome with no known gene anywhere nearby. It is not clear whether these unknown transcripts arise purely from technical artifacts or if they actually correspond to unknown genes or other functional ncRNAs. Interestingly, many unknown transcripts were found near known genes, but on the opposite strand ("putative anti-sense transcription"). Anti-sense transcription might occur randomly as a bi-product of regular transcription, but might in other cases also serve a regulatory function like the suppression of sense transcription by binding of complementary transcripts[100, 194, 255, 295].

I had a closer look at the distribution of transcripts across known gene models (within a window of $20kb$ up- and downstream of the nearest known feature of each mapped tag) and divided the mapped tags into bins with respect to their location. The counts of each bin were normalised to account for any difference in size (**Figure 2.7**). Least surprisingly, the largest portion of sense transcripts was found within the exons or across the splice junctions of known genes. Some tags mapped into intronic regions (which might, in fact, be incorrectly annotated exons). The remaining sense transcripts spread across the neighbourhood of the feature, with a higher percentage falling in the downstream regions (gradually decreasing with distance from the gene). This might be partially due to incorrectly annotated 3' UTRs. The distribution of anti-sense transcripts by trend follows the distribution of sense transcripts, but is, in general, more evenly spread across the whole range, which indeed argues for a random and functionally inactive role of anti-sense transcription. It is, however, noteworthy that still the majority of putative anti-sense transcripts clustered in the exonic regions and just downstream of known

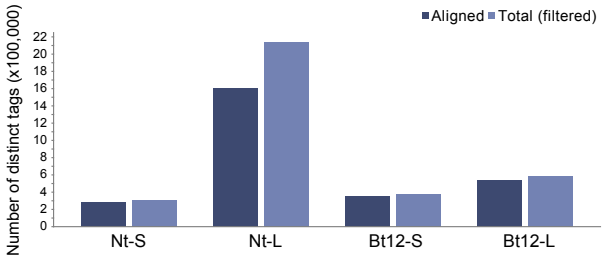

Figure 2.8: Number of distinct tags. The number of distinct tags found in each short read dataset increases slightly with library size. NT-L has extraordinarily many distinct tags, many of which are filtered out during quality control.

features (consistent with ref.[84]) where, at least some of it, might function in silencing sense transcription or suppressing incorrectly terminated transcription.

2.1.3.2 Detection of a Problematic Read Library

In the previous section, I have pointed out several times that one of the datasets, *NT-L*, exhibited somewhat different properties from the other datasets: An unexpectedly high number of reads could not be aligned to the genome and an extraordinarily high proportion of the remainder mapped ambiguously to multiple locations (**Figure 2.4**). One very striking difference between *NT-L* and the other samples can be seen in the ratio of distinct tags to overall reads: More than twice as many distinct tags were observed than expected (tag-to-reads ratio $\gamma = 0.278$; average ratio in the other samples $\gamma = 0.107$). The difference cannot be explained by the difference in library size alone. While one would naturally expect the number of distinct tags to grow with the overall number of short reads, the total number of distinct tags should approximate a plateau at a certain level. This trend is exemplified by the difference in the ratio between the smaller and the larger knock-out sample, with $\gamma = 0.119$ (*BT12-S*) and $\gamma = 0.099$ (*BT12-L*), respectively. The high tag-to-reads ratio can also explain the high number of low-coverage regions (**Figure 2.5**). Quite a large portion of the tags has been filtered out during quality control (**Figure 2.8**), but despite this measure the number of distinct tags is several orders higher than in the other samples.

In order to gain a better understanding of how *NT-L* differs from the rest, I have visualised the transcriptional activity across the entire genome using the UCSC Genome Browser[283] by converting the tag counts per genomic region to a custom user track. While the sequenced short reads usually clearly peak near the 3' ends of transcribed features, the *NT-L* sample appears to spread across the entire genome (**Figure 2.9**). The wide-spread distribution of transcripts is not only limited to genic regions, but spans the entire genome with the effect that almost 15% of the total transcript counts were assigned to uncharacterised regions as compared to an average of about 5% (data not shown).

What is the reason for the drastic differences between the datasets? The reported quality

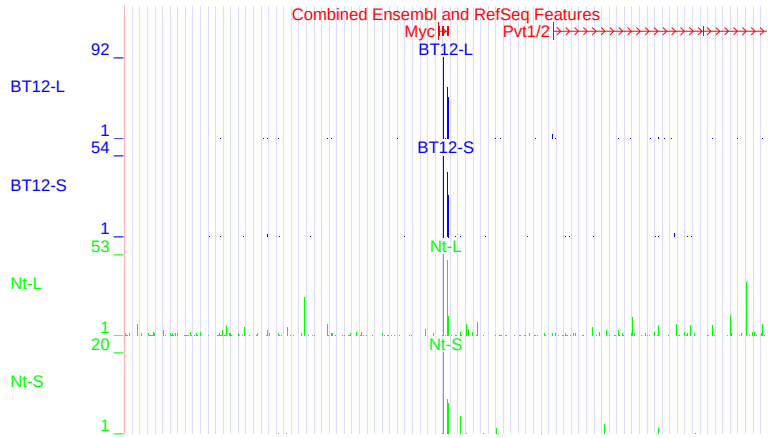

Figure 2.9: Unusual tag distribution. While the sequenced short reads usually clearly peak near the 3' ends of transcribed features, reads in the NT-L library appear to spread across the entire genome. Here, an example from the neighbourhood of the *c-Myc* locus (from UCSC Genome Browser[283]).

values for all sequenced reads are no worse than for the other samples (in fact, they are slightly better than the average; **Figure 2.3**). Of course, there is no way to tell for sure that the quality values are actually reliable in this particular case, however, they all come from the same lab and were processed in the same batch (in fact, they were most likely sequenced in the same machine run, on the same flow cell), so technical differences seem unlikely.

There are several steps in the sample preparation which might be prone to error. Since the RNA was not checked for its integrity prior to submission, I hypothesised that there might have been a contamination with genomic DNA. In the preparation of the sequencing library, transcripts are selected for poly-`A` using oligo-`d(T)` beads. Stretches of DNA might erroneously be selected by these beads if they contain a long stretch of `A/T`-rich sequence. I therefore investigated the nucleotide composition of the short read tag sequences (**Figure 2.10**) and, indeed, found a high number of `A` and `T` in the *NT-L* tag sequences. Oddly, the difference in nucleotide composition is just the opposite when looking at the absolute nucleotide counts across all sequenced reads (rather than only the distinct tags). Evidently, many of the poorly represented tags must be `A/T`-rich, which might concur with my hypothesis and hence explain the high tag-to-read ratio. Nevertheless, the question remains why the rest of the tags (which must be represented by a comparatively high number of reads each) is particularly `C/G`-rich and further investigation would be necessary to shed light on this question.

Other factors in the sample preparation might play a role in the special case of *NT-L*. Inconsistencies in the *NlaIII*-mediated cleavage of cDNAs might result in unexpected tag sequences, errors in the adapter-ligation, amplification and colony-formation steps can severely bias the read-out of sequence information and it cannot be ruled out that adverse conditions lead to a degradation or alteration of RNA – the consequences of which on the sequencing read-out would be unpredictable. Lastly, it remains possible (but, in my opinion, improbable)

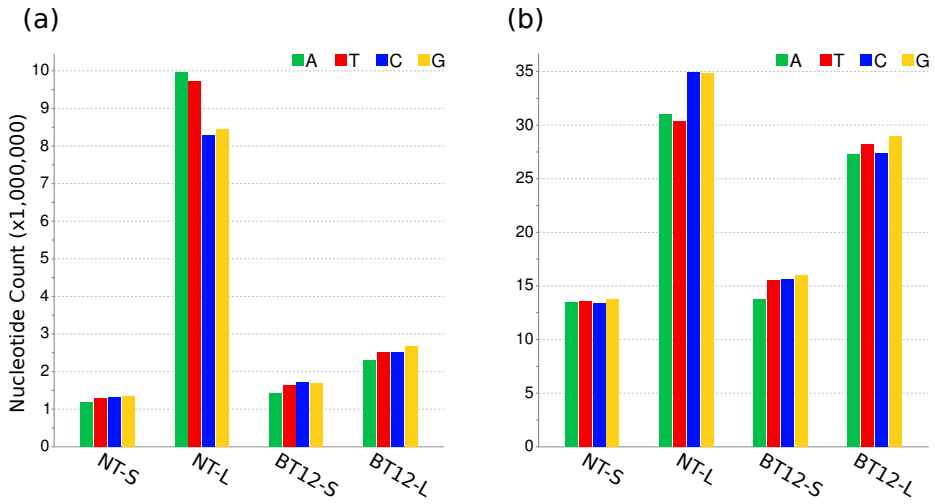

Figure 2.10: Nucleotide frequencies. (a) *Nt-L* exhibits an unusually high `A`- and `T`-content in its distinct tag sequences, this difference seems to be reversed when (b) looking at the sequences of all reads, where there is, in fact, an over-representation of `C` and `G`.

that the differences have an underlying biological reason.

Evidently, the discussion regarding the causes of the abnormal tag composition of *NT-L* remains speculative. At this point, I saw no other possibility, but to discard $NT-L$ from the further analysis. This special case emphasises the importance of quality control procedures prior to advanced processing of HTS data and highlights the need to look at measures beyond just the base-call quality scores, for instance, the nucleotide composition and tag frequency, to spot flaws in the data – a lesson I have taken into account during the later development of the GeneProf data analysis suite (see **Chapter 3** and, in particular, **Section 3.3.3.1**).

2.1.3.3 Differential Analysis and Comparison with Microarrays

In order to identify genes directly or indirectly linked to *Nanog* expression, I sought to assess differential gene expression between the two cell populations at hand, that is, I attempted to calculate a measure of statistical significance for differences observed between the two states to be due to actual biological mechanisms and *Nanog* dependence rather than attributable solely to chance. For this purpose, the *edgeR* package of the Bioconductor suite was used[458]. After applying a quantile normalisation to account for global, technical differences to the raw expression read counts, the version of edgeR used calculated moderated statistical tests assigning p-values to the observed differences in expression levels for each gene. These p-values were finally adjusted using the Benjamini-Hochberg method to correct for the expected false discovery rate due to multiple testing[32]. It should be noted that any measure of statistical significance is limited in its reliability by the availability of replicates. In this experiment, testing for statistical difference between two conditions with only 1 and 2 replicates each (due

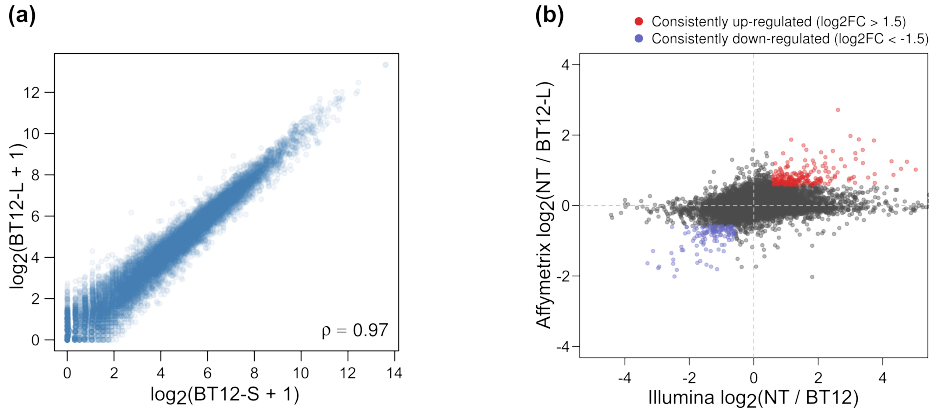

Figure 2.11: Consistency of DeepSAGE measurements between replicates and with microarrays. (**a**) The scatter plot demonstrates a very low degree of variability in expression values (reads per million) between two replicates (BT12-S and BT12-L) of the same cell type. (**b**) Logarithmic fold-change values between *NT* and BT12 cell line samples obtained using Illumina DeepSAGE (x-axis) and Affymetrix microarrays (y-axis). Consistently and strongly changing genes are highlighted in colour.

to the removal of NT-L) constitutes the bare minimum of replication necessary to make any reasonable judgement at all.

In some cases, very small changes in the expression level of a single gene can make a striking difference to the biology of a cell (cp. for example, the complex interactions of factors specifying the neural tube along morphogen gradients[341]), but in order to judge whether a small difference is meaningful, rather than a matter of random fluctuations, a large number of experimental observations is required. Hence, I decided to limit the analysis to those candidates that exhibit a quite drastic change in expression, which seem unlikely to be due to random fluctuations. Reassuringly, expression values in the two replicate datasets varied very little (BT12-S and BT12-L; **Figure 2.11**.a).

The list of detected features was filtered to only those which were deemed to change significantly according to edgeR (adjusted p-value ≤ 0.05) and which additionally changed at least 1.5-fold in either direction alongside *Nanog* ($log_2(1.5) \approx 0.585$). Additionally, I compared the fold-change values in our datasets to those obtained from experiments using the exact same cell lines assessed with Affymetrix microarrays and removed all those features from the further analysis in which the direction of change in the study at hand contradicted the ones observed previously. I considered the inconsistent changes in those features to be most probably independent of *Nanog* and thus negligible for the characterisation of *Nanog*-dependent transcription (**Figure 2.11**.b).

Further investigations into this matter revealed that expression signal intensities reported by the different platforms were most consistent for genes with a medium expression level (**Figure 2.12**). Microarrays work best for well-expressed known genes. For weakly expressed genes, probe fluorescence intensities can hardly be distinguished from the background level

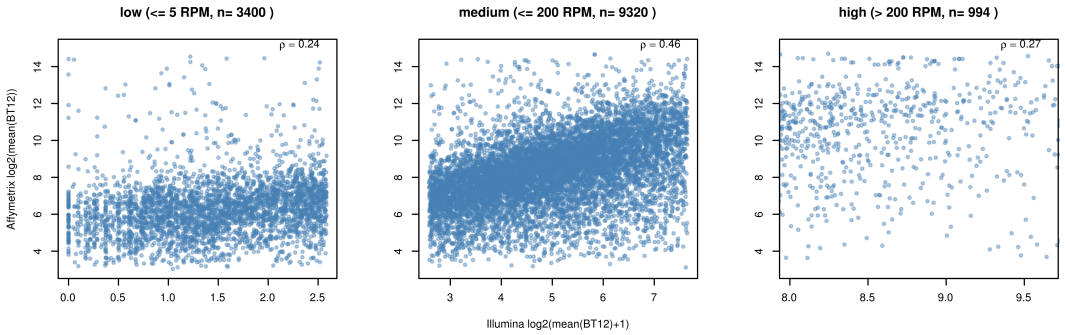

Figure 2.12: Gene expression intensities measured by Illumina sequencing and Affymetrix microarrays. Expression intensities are plotted on a logarithmic scale and are average over several replicates per technology (2x HTS, 3x Affymetrix). The plot is tiered into three panels by the overall mean expression level of the individual genes as shown in the plot label. The Spearman correlation coefficient ρ summarises the overall similarity of the signals reported by both platforms.

and are hence difficult to detect. Very high expression may saturate all probe sets present for a single gene on the array making accurate measurements above a certain level impossible. The lack of consistency between both technologies in the lower and upper expression range may therefore be explained by the inaccurate microarray measurements, rather than by a weakness of HTS. This observation is consistent with reports in the literature[340, 526].

2.1.3.4 Putative Downstream Targets of Nanog

Let us now focus on the biology of the system studied, specifically, I will look at a number of features identified as interesting downstream target candidates for *Nanog* and worthy of further investigation.

A total of 14,447 known genes was transcribed at a reliable level, that is, at least 5 tags were mapped into each gene's body. This corresponds to roughly half of all the features in the extended set of all mouse genes and short RNAs (**Section 2.1.2.2**). Additionally, evidence for the transcription of up to 11,132 anti-sense or novel features was detected. In this analysis, I initially focused on the first, well-defined portion.

Differential expression analysis yielded 1,176 genes ($adj.p \leq 0.01, |\log_2(Bt12/Nt)| \geq \log_2(1.5)$), which I filtered further according to the following criteria:

- **Consistent Expression Change.** The integration of external expression datasets (**Section 2.1.2.3**) afforded the opportunity to compare the trends in the *Nanog* knock-out at hand with other similar data. In particular, the in-house Affymetrix data would be expected to discover largely the same genes. Other studies, comparing cells with low *Nanog* expression (either due to manipulation or to cell sorting) to normal stem cells, should reveal similar trends in the expression patterns.

Bearing this in mind, I limited the list of candidates to only those genes that had been reported as differentially expressed in at least one of the other data sources (see **Section 2.1.2.3**). I furthermore excluded all genes whose direction of change (DOC) was inconsistent (i.e. those that had been found up-regulated in some, but down-regulated in other studies, or vice versa). It should be noted that this approach neglects the potential benefits of the sequencing technology employed (detection of previously not measurable genes) in favour of identifying the most reliable candidates (consistent between old and new technology).

- **Genes with TFBS.** Moreover, I sought to pinpoint direct targets of *Nanog* by eliminating all genes from the list that had no high-confidence binding site for the transcription factor (**Section 2.1.2.3**). Again, in doing so, I deliberately neglect second-order effects of *Nanog* and those whose binding sites have not been discovered yet.

Of the initial candidates, a total of 264 genes were supported by at least one other study with regards to differential expression and never contradicted in terms of DOC. The overlap of those genes with the 234 genes that had been found to have at least one reliable *Nanog* binding site amongst all differentially expressed genes, yielded 70 genes. **Table 2.3** shows the genes that appear to be directly activated ($n_{up} = 40$) or repressed ($n_{down} = 30$) by *Nanog*.

In order to assess the effects of the knock-out of *Nanog* on the biology of the cell, I attempted to analyse affected functional categories, transcriptional networks and signalling pathways. A number of free software tools for this purpose exist[101, 140, 476, 495], but in this instance I used a trial version of the commercial Ingenuity Pathway Analysis (`http://www.ingenuity.com`) software.

I first composed a transcriptional network of the candidate genes identified in the earlier analysis. Initially, I constructed a network of all genes with a known involvement in stem cell maintenance, pluripotency or, conversely, lineage commitment, differentiation and tissue/organ formation. This network was based on Ingenuity's literature-curated knowledge base. Subsequently, I extended this network by adding all high-confidence, direct *Nanog* targets (**Table 2.3**) and drawing an activating/inhibitory connection between each of them and *Nanog*. Lastly, I extended the network by adding all known, direct downstream-regulated targets (from the Ingenuity database) of the components of the network and plotted the network with respect to its localisation in the cell (**Figure 2.13**).

2.1.3.5 Discussion and Conclusions

Let us now try to summarise and discuss the outcomes of this analysis and speculate as to the implications of the observed results. It has previously been reported that the knock-out of the *Nanog* gene does not disrupt pluripotency *per se*, but rather pre-disposes ESCs to a

Name	Nt-S	Bt12-S	Bt12-L	Illumina Nanog-/-	Adj. P	Affy Nanog-/-	Singh Nanog high/low	Loh shNanog	Ivanova shNanog	Nanog TFBS
D630039A03Rik	11.7	0.0	0.0	-31.59	0.0003	-0.20		-0.59		1x proximal
Mras	81.5	8.8	7.1	-3.30	0.0000	-1.64		-1.60	-2.49	1x intragenic
Serpinb6c	39.4	3.1	4.7	-2.93	0.0000	-1.73				1x intragenic
Itga9	139.8	32.2	26.6	-2.25	0.0000	-1.17				4x distal, 1x promoter
Fut9	102.6	26.3	17.7	-2.20	0.0000	0.20		-1.47		1x intragenic
Ly75	159.6	38.1	31.7	-2.19	0.0000	-0.46		-0.70		1x intragenic, 1x distal
Nanog	659.8	158.4	154.5	-2.08	0.0000	-7.01		-1.81	-2.39	1x proximal, 1x promoter
Vegfc	133.3	36.9	28.4	-2.02	0.0000	-0.76	down		-1.23	1x promoter
Sorl1	137.2	32.7	37.8	-1.93	0.0000	-0.34		-1.18	-0.96	2x intragenic
Gpc3	576.8	151.7	156.4	-1.90	0.0000	-0.34		-1.15		2x intragenic
Kit	133.5	38.3	35.9	-1.84	0.0000	-0.93		-1.14		1x intragenic
Slc15a1	64.6	16.1	17.5	-1.84	0.0000	-0.70				1x promoter
Fhod3	90.9	29.8	24.1	-1.77	0.0000	-0.65		-1.36		3x distal
Manba	446.8	140.1	123.3	-1.77	0.0000	-1.14		-1.05		1x intragenic
Pdcl2	43.7	12.4	12.8	-1.76	0.0000	-1.39		-1.40		1x intragenic
Igfbp2	1,279.0	396.3	425.3	-1.63	0.0000	-1.04		-0.88		1x distal
Cisd3	74.2	26.4	22.3	-1.59	0.0000	-0.18	down		-0.86	2x distal
Chac1	109.7	41.6	33.4	-1.54	0.0000	-0.60		-0.89		1x distal
Tnfrsf21	68.8	23.2	26.8	-1.43	0.0000	-0.68				1x intragenic
Tex14	68.8	26.5	24.3	-1.42	0.0000	-0.39		-1.05	-1.06	1x intragenic
Zmat4	139.7	61.6	45.2	-1.38	0.0008	-0.51		-0.60		6x distal, 2x intragenic
Adam23	342.4	145.4	139.5	-1.26	0.0000	-0.72				1x proximal
Gpt2	237.7	96.5	105.4	-1.23	0.0000	-0.74		-0.87		1x distal, 1x intragenic
Igf2bp2	213.8	88.3	99.1	-1.17	0.0000	-0.58		-1.26		1x intragenic
Tet2	221.7	113.3	89.3	-1.13	0.0004	-0.42		-0.88		3x intragenic
Add3	138.9	71.0	62.9	-1.05	0.0000	-0.52		-0.72		1x distal, 1x intragenic
Lrrc2	135.8	66.7	66.8	-1.02	0.0000	0.17	down	-1.23		1x proximal, 1x promoter
5730419I09Rik	344.7	193.4	152.8	-1.00	0.0044	-0.24	down			1x intragenic
2310005N03Rik	101.9	54.0	48.5	-0.99	0.0001	-0.59				1x distal
Eras	1,194.5	611.7	699.8	-0.86	0.0000	-0.50		-1.30		1x intragenic
Dclk2	117.6	63.3	68.6	-0.82	0.0004	-0.14	down			3x intragenic
Dennd2c	174.2	109.8	92.9	-0.78	0.0022	-0.14			-1.13	1x distal
Slc38a4	486.8	283.0	304.4	-0.72	0.0000	-0.61				1x intragenic
Emb	1,586.0	940.8	989.6	-0.71	0.0000	-0.63		-0.97		1x proximal
Sntb2	108.0	72.2	65.0	-0.65	0.0093	-0.21			-0.96	1x distal, 1x intragenic
Lypla1	810.5	544.4	492.4	-0.65	0.0005	-0.23	down		-1.43	1x proximal
Rara	286.6	193.3	178.7	-0.63	0.0000	0.00			-1.36	1x proximal
Slc12a4	249.6	161.0	162.6	-0.63	0.0001	-0.64				1x promoter
Ptch1	303.9	184.2	209.2	-0.62	0.0000	-0.88		-0.63	-1.56	1x intragenic
Nampt	850.4	597.2	523.5	-0.61	0.0013	-0.15		-1.06		1x intragenic
Ppm1f	219.1	352.0	344.1	0.66	0.0000	0.11		0.62		1x distal
Pml	859.5	1,364.2	1,389.9	0.68	0.0000	0.74				2x distal
Lrp2	84.2	147.8	133.8	0.72	0.0012	0.37			1.49	1x intragenic
Ralgds	75.0	141.6	130.1	0.83	0.0002	-0.19			0.96	1x distal
Rnf12/Rlim	355.3	691.6	588.7	0.84	0.0015	0.63				1x proximal
Stx3	157.1	270.4	299.8	0.86	0.0000	0.82				1x intragenic
Adk	141.0	287.6	271.7	0.98	0.0000	0.61				1x distal
Ror2	43.2	97.9	83.7	1.03	0.0002	0.90		0.84		3x intragenic
Otx2	129.1	239.1	295.0	1.05	0.0029	1.23				1x intragenic
Top1	92.1	179.5	203.6	1.06	0.0000	0.64				1x intragenic
Axud1	80.3	172.8	176.3	1.10	0.0000	0.75		1.51		2x distal
Lats2	15.4	39.7	35.0	1.16	0.0083	1.87		0.79		1x intragenic
Urm1	224.4	540.7	479.4	1.17	0.0000	0.27		0.66		1x distal, 1x intragenic
Crlf2	26.5	53.5	66.5	1.21	0.0044	-0.22			1.06	1x intragenic
Pphln1	24.6	58.4	58.5	1.22	0.0012	0.44	up			3x intragenic
2210408I21Rik	12.8	35.8	31.9	1.30	0.0093	0.71				1x distal
Zfp771	65.7	172.3	158.9	1.31	0.0000	0.59				2x distal
Unc13b	31.2	96.9	74.1	1.40	0.0058	0.71				2x distal
Stk31	39.2	114.1	122.0	1.58	0.0000	0.96		1.18		1x intragenic
Afap1	94.7	258.9	361.6	1.70	0.0065	0.70		1.67	2.56	1x distal
1190005I06Rik	90.7	327.5	384.5	1.96	0.0000	0.46		0.74	0.76	1x intragenic
Ust	12.8	53.7	54.7	1.97	0.0000	-0.07	up			1x intragenic
Tdrd3	10.1	47.8	37.3	1.97	0.0001	0.71		1.09		1x distal, 1x intragenic
Plce1	6.1	36.9	25.4	2.06	0.0064	0.65				2x distal, 1x proximal
2610528J11Rik	14.0	65.1	55.8	2.06	0.0000	0.76			2.03	1x distal
Ets1	6.1	35.7	28.4	2.29	0.0001	0.82				1x intragenic
Snai1	5.3	23.3	29.9	2.40	0.0004	-0.02		1.16		1x distal
Rgs20	6.1	42.7	33.5	2.52	0.0000	0.81				1x intragenic
Ninj2	0.8	13.6	11.4	2.83	0.0083	0.19			1.36	1x distal
Hmga2	41.9	410.9	419.6	3.26	0.0000	1.88		0.98		2x intragenic

Table 2.3: *Nanog* target genes. The table shows the quantile-normalised tag count (`Nt-S`, `Bt12-S`, `Bt12-L`), $\log_2$ fold-change (`Illumina Nanog -/-`), FDR-adjusted P-value (`Adj. P`), $\log_2$ fold-change in the Affymetrix comparison libraries (`Affy Nanog -/-`), tendency of change in ref.[507] (`Singh Nanog high/low`), $\log_2$ fold-changes in ref.[227,327] (`Loh` and `Ivanova shRNA`). Differentially down- and up-regulated genes are high-lighted in red and green, respectively. The last column shows the type of binding site(s) found (see text). The list is limited to those genes that are (a) differentially expressed in our study ($p \leq 0.01, |\log_2(Bt12/Nt)| \geq \log_2(1.5)$)), (b) found differentially expressed in at least one other study (and never contradicted), and (c) have a binding site supported by at least two independent studies.

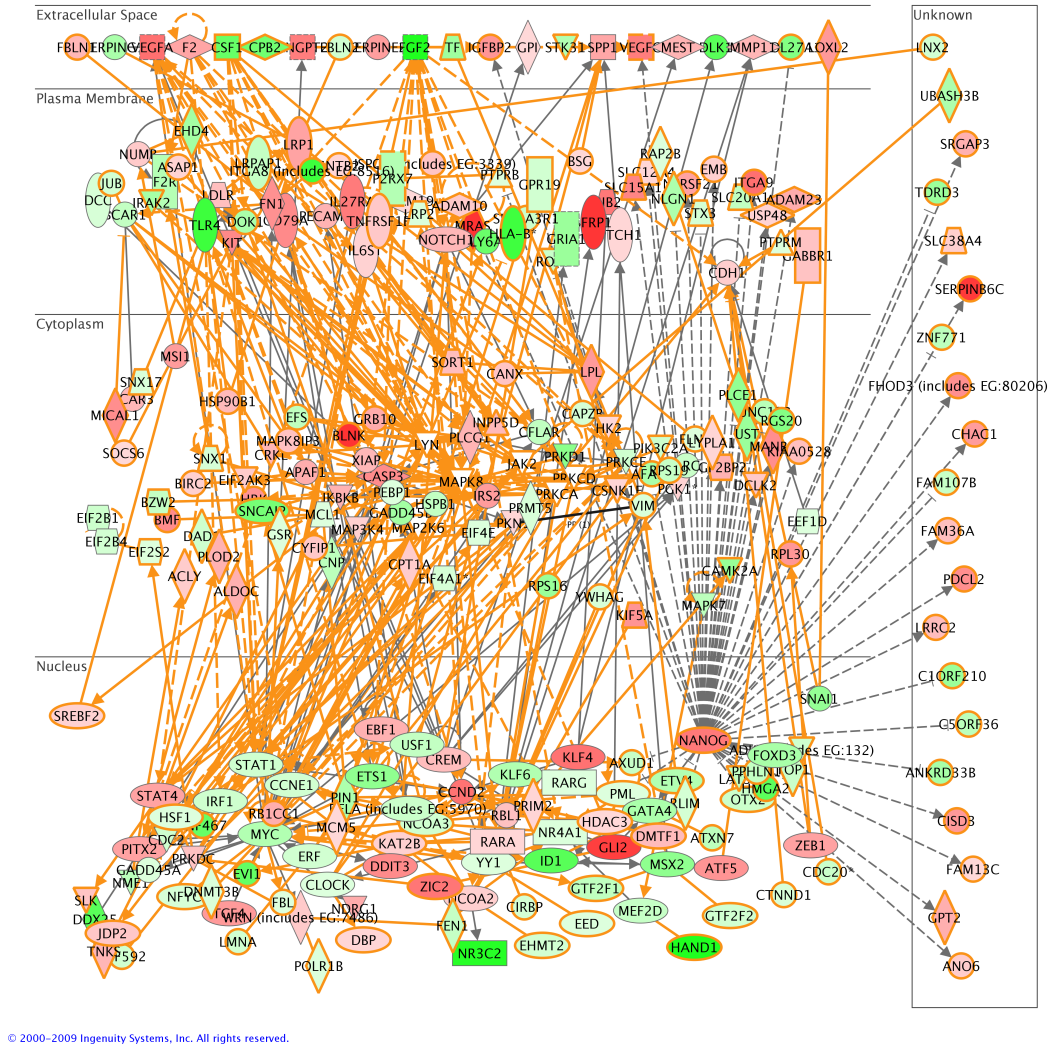

Figure 2.13: *Nanog* gene regulatory network. A regulatory network of direct and indirect *Nanog* activators, repressors and targets created using Ingenuity Pathway Analysis in combination with the data obtained in this study. Green = up-regulated (repressed by *Nanog*), red = down-regulated (activated by *Nanog*). Grey, dashed connections have been inserted manually, the others are curated by the manufacturer.

more "differentiable" state[68]. Consistent with this notion, cells were reported to maintain expression of major stem cell markers, for instance, *Pou5f1*, *Sox2* and *Zfp42*. These findings have largely been confirmed in this deep sequencing study, with most stem cell markers not showing any differential effect between both conditions, however, there are some exceptions (**Table 2.3**). Most strikingly, *Dnmt3l* and *Dppa4* appear to be considerably up-regulated, while *Eras* and *Tcl1* show a distinct drop in expression levels. Having a closer look at the RT-PCR results in reference[68] might consolidate this result for *Eras* and possibly *Dnmt3l* (although not to such an extent), but the other differences remain controversial. However, differential down-regulation of *Tcl1* upon *Nanog* depletion has been confirmed by both, our in-house microarrays and Loh *et al*[327]. Strangely, the drastic increase in *Dnmt3l* expression has not been detected in any other study, but at least the DOC seems to be confirmed by the microarrays.

I was expecting to observe a notable increase in *Gata6* expression in BT12. *Gata6* and *Gata4* are considered early markers of extraembryonic endoderm specification and their expression seems to be mutually exclusive from *Nanog* in the late blastocyst stage[466,507]. Both genes showed increased expression levels in the knock-out cell line ($log_2(\frac{Bt12}{NT})$: *Gata4*=1.78, *Gata6*=4.45). Taken together with an increased expression of *Cdx2* ($log_2(\frac{Bt12}{NT})$: *Cdx2*=1.01), which has a role in trophectoderm differentiation and may repress *Nanog* and *Pou5f1*[214,466], these findings support the notion that *Nanog*-deficient cells are prone to differentiation into extraembryonic lineages and that, in fact, a part of the cell population might have already undergone differentiation. The absolute expression level of all these genes appears to be very low (with a peak of 45 in 3.8 million, which amounts to approximately 3 transcripts per cell), but the changes have all been confirmed in at least one other study[227,327,507] and it seems likely that the low average expression levels stem not only from low abundance per cell, but from a selective expression from only those few cells that have undergone (or at least started) the differentiation process, which is levelled out by the majority of cells having remained in a pluripotent state. Interestingly, there is a high-confidence binding site for *Nanog* within the first intron of *Cdx2*, which might indicate that *Cdx2* is directly inhibited by *Nanog* in wild-type ESCs.

Another interesting group of genes affected by *Nanog* is the *Zscan4*-family. Although, they do not appear to be direct targets of *Nanog*[†], three members of the family, *Zscan4f*, *Zscan4d* and *Zscan4c/d*, were deemed differentially up-regulated in our study (that is, more highly expressed in absence of *Nanog* than in its presence), which is somewhat surprising, since *Zscan4* has been found to be exclusively expressed in early developmental stages *in vivo* and its depletion hindered implantation of the blastocyst[125]. However, more recent research

[†]No TFBS has been found in any of the studies considered. Some new insights which I will present later in this thesis, however, hint towards a direct transcriptional control of *Zscan*-family genes by *Nanog* and other TFs: **Section 5.2.5** and **Figure C.2**.

points to an important role of *Zscan4* in the maintenance of genomic stability of ESCs[625] and the regulation of early embryonic genes[199]. Interestingly, *Zscan4* has been found expressed transiently only in a subset of ESCs, coinciding with telomerase repair[625]. Transient expression of the gene can promote reprogramming of fibroblasts to iPS cells[199].

Numerous other genes involved in pathways that are known to have an influence on stem cell differentiation into various lineages have been pinpointed, but it has proven difficult to summarise those into a coherent picture. Several growth factors (FGF, EGF, PDGF, TGF, VEGF) show changes in their transcript levels, which might lead to proliferation and differentiation, but conversely other members of the very same pathway give contradictory evidence. For example, *Fgf2* is up-regulated, while *Fgf4* is down-regulated at the same time. *Fgf2* has been reported to support the maintenance of human ESCs in culture[109, 165], whereas *Fgf4* has been found necessary for cells to commit to a lineage and undergo differentiation[285]. The changes in FGF levels therefore seem to counteract the loss in potential to maintain stem cell identity, by inhibiting differentiation. The effect of *Nanog* on *Fgf4* has been confirmed independently[227, 327] and there are two potential *Nanog* TFBS (one about $10kb$ upstream of the TSS and one in the 3' UTR), so it appears that *Nanog* promotes "differentiability" and hence pluripotency, partially via up-regulation of *Fgf4*.

Moreover, the expression of other major suppressors of cell differentiation and sustainers of pluripotency is lost (**Table 2.3** and **Figure 2.13**): *Klf4, Sfrp1, Mras, Trps1, Esrrb, Igfbp2, Tcf3, Gli2, Notch1, Ptch1* and *Smad7* are all involved in preventing differentiation and promoting stem cell proliferation. All of these genes have previously been pointed out as *Pou5f1* targets[497]. Surprisingly, markers of X-chromosome inactivation seem to indicate X reactivation, as *Xist* expression drops and *Eed* levels increase with the knock-out of *Nanog*[373, 374]. But since we were dealing with male cell lines, the effects might be misleading.

I also had a quick look at what I had termed "putative anti-sense transcription" earlier. Based on the suspicion that most of it would not be of any discernible biological relevance, I decided to look only for the most significantly changing anti-sense features (adjusted $p \leq 0.01, |\log_2(BT12/NT)| \geq 2$, maximum, normalised expression level > 20), comprising 26 down-regulated and 11 up-regulated features. The first list contained transcripts on the opposite strands of *Nanog, Zic2, Ifitm1, Pecam1, Klf4* and *Sall1*, the latter *Hmga2* and *Hs3st4*. The quality and quantity of change of all of these features were extremely similar to their sense-strand features. I think that this demonstrates that the anti-sense transcription is largely an artifact of the sequencing process. After the bridge-PCR amplification, cDNA fragments can essentially be present as replicas of both possible strands, but subsequent sequencing ought to only pick up those constructs identical to the original template of each cluster thanks to the specificity of the used sequencing primers. I suspect that in some cases constructs bind to the wrong flowcell-attached adapter and corrupted sequences might take over the cluster.

Alternatively, adapter sequences might have been inserted in the wrong direction in the earlier sample preparation equally leading to a transcript apparently emerging from the opposite strand. Whatever the source for potential errors, it stands to reason that these would be rather rare. Evidently, only about $7-8\%$ of all transcripts were assigned to anti-sense regions – if all anti-sense reads were erroneous and due to stochastic errors one would expect a roughly equal number of sense and anti-sense transcripts. One further source of erroneously annotated anti-sense transcription has yet to be mentioned: In a few cases, distinctly higher anti-sense transcription could be observed than for the opposing strand's actual feature. This was usually the case for novel and poorly characterised genes (e.g. *AL772393.11-2* or *Laptm4b*) and I believe that those features might actually be annotated to the wrong strand. In summary, while a lot of evidence of putative anti-sense transcription has been found in our study, I conclude that a large proportion of it might be due to flaws in the technique and to identify the real proportion of it is impossible using the current methods. A targeted approach to studying anti-sense transcription has been proposed by He *et al*[194]. They suggest to replace cytidine by uridine residues prior to sequencing, thereby making both strands more readily distinguishable. It would certainly be interesting to use this approach to study ES anti-sense transcription in more detail, in particular in the light of more recent findings which implicate RNA co-factors, including many anti-sense and extra-genic transcripts, in the regulation of PRC2-mediated gene silencing[634].

Lastly, I checked for potential novel features with a biological function. Some of the extra-genic transcription observed could be an artifact of ambiguous reads: If a read maps ambiguously to both a gene as well as an extra-genic region (because the respective bit of DNA is repetitive), a proportion of this read will be attributed to both possible locations. Therefore it will appear that there is an extra-genic signal, although it might actually have never originated from this extra-genic region. Given the data at hand, it is impossible to tell the difference with certainty. Other low-level extra-genic transcription could be explained by sequencing errors: A single misread nucleotide in a read could mean that this read mapped to a different region in the genome. In order to eliminate background transcription as well as the artifacts of repetitive sequences, I considered only *x-clusters* with at least 20 uniquely mapped tags in at least one of the samples and with a length greater than $21bp$, i.e. clusters constituted by more than one mapped region (remember that aligned reads were merged into clusters when they were within a maximum distance of $1kb$ to each other, **Section 2.1.2.2**). Amongst those *x-clusters*, I concentrated on the ones that were changing differentially with high significance (adjusted $p \leq 0.01$, $|\log_2(BT12/NT)| \geq 1$). Only 16 clusters satisfied these criteria: 7 down-regulated, 9 up-regulated. The maximum, normalised expression levels in those clusters ranged from 25 to 94, which I believe makes them unlikely to result from random expression as it is well in the range of expression levels from known sense transcripts

($median = 49$).

The most highly expressed, down-regulated cluster is located on chromosome 1, on the forward strand from base positions $138,587,227$ to $138,587,485$ (band $1qE4$). The region is highly conserved in rat, but lacks any conservation in other mammals. Transcripts from the same region have also been found in experiments within the Cancer Genome Anatomy Project (CGAP; `http://cgap.nci.nih.gov`). On the other site, the most highly expressed, up-regulated cluster can be found on chromosome five (forward strand) from position $63,808,344$ to $63,808,619$ (band $5qC3.1$). This region is partially conserved in higher mammals (human, rat and orang-utan).

These are just two examples and a larger-scale analysis in combination with external datasets and conservation scores might yield interesting new subjects for further research.

2.1.3.6 Supplementary Note

In 2011/2012, after the development of the GeneProf software (**Chapter 3**), I have repeated the analysis outlined in this chapter and augmented it further with additional high-throughput data. This work does now, together with many additional results generated primarily by Nicola Festuccia and Rodrigo Osorno (I. Chambers group), contribute to a manuscript which is currently being revised.

2.2 Identification of Pluripotency Genes in Plant Cells

To further investigate the potential of HTS, a second exploratory study in a non-model organism was undertaken. In collaboration with the research group of Prof. Gary Loake (Institute of Molecular Plant Sciences, University of Edinburgh), I participated in a study of global gene expression signatures in pluripotent plant cells, profiling two distinct cell types of the Japanese yew (*Taxus cuspidata*) using a DeepSAGE approach similar to the one employed before (**Section 2.1**).

2.2.1 Motivation and Goals

Plants are the source of a wide variety of chemicals of industrial and medicinal use[484]. Production-scale utilisation of full-grown plants is often not a cost-effective and feasible solution and consequently much effort has gone into deriving cells that may be grown in culture. Previous efforts focused on dedifferentiating cells into proliferating progenitor-like populations[546]. However, cultures of dedifferentiated plant cells (DDCs) are heterogeneous, grow slowly and inconsistently and, crucially, have been reported to return only low amounts of chemical products[15,97,163,523].

To avoid the flawed dedifferentiation process, my collaborators sought to establish a naturally undifferentiated, stem cell-like cell line from the cambium (**Figure 2.14**.a) of *T. cuspidata* (cambial meristemic cells, CMCs), which was expected to yield more stable growth properties and improve the efficiency of the biosynthetic production of taxol[97]. Taxol (also known by its commercial name, paclitaxel; Bristol-Myers Squibb, New York, USA) is a natural product of yew and is used as a mitotic inhibitor in cancer chemotherapy. Evidently, its large-scale production is therefore of great relevance.

From a data analysis point of view, what made this study different was the fact that, at the time of this work, no complete genome or transcriptome assembly was available for this organism and neither were there any commercial microarray platforms established that would have allowed us to carry out our investigations. I was therefore presented with an opportunity to gauge the potential of HTS to broach known frontiers and create novel insight.

2.2.2 Methodology

The derivation and study of *T. cuspidata* CMCs was a difficult and complex project and involved a great number of people. For the purposes of this dissertation, I shall focus mostly on the data processing and statistical analysis aspects of the study since these are most relevant for the remainder of this work. In order to enable a better understanding of the study as a whole, I will first briefly review the process that led to the establishment of the CMC populations and the assembly of a reference transcriptome for further analysis. Further

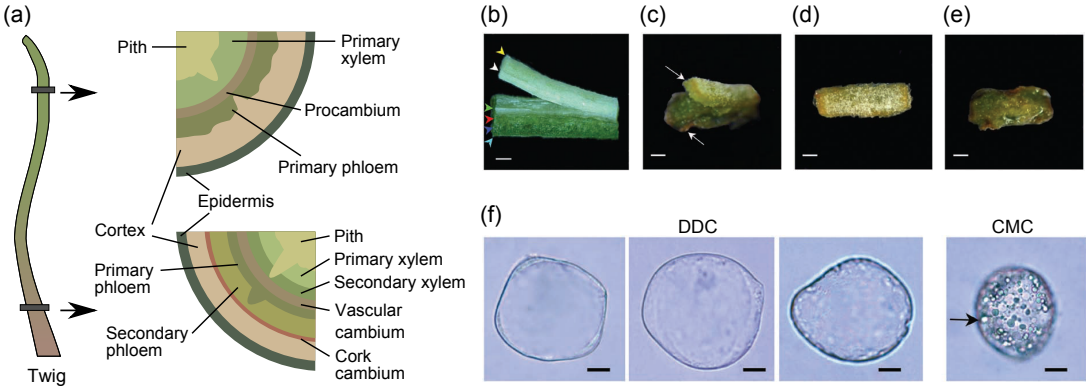

Figure 2.14: Derivation of plant cell lines. (**a**) Schematic representation of *Taxus* compartments relevant to this study. (**b**) Peeling off different layers of tissues. Yellow: pith, white: xylem, green: cambium, red: phloem, blue: cortex, turquoise: epidermis. Scale bar: 0.5mm. (**c**) Culture induces a visible split between DDCs (bottom) and CMCs (top). Scale bar: 1mm. (**d**) CMCs from cambium and (**e**) DDCs from phloem, cortex and epidermis. Scale bar: 1mm. (**f**) Micrograph of DDCs (left) and a CMC (right) demonstrating the presence of vacuole-like components in CMCs (black arrow). Figures (**b-f**) were reproduced with permission from reference[297].

details can be found in Lee *et al.*[297]. Afterwards, I will discuss the statistical analysis of the expression data at hand.

2.2.2.1 Derivation of Cambial Meristemic Cells

My collaborators decided to derive cells from the cambium of *T. cuspidata* (**Figure 2.14**.a), because they were believed to functionally resemble vascular stem cells and the cambial region targeted had been previously reported to produce high levels of taxol[521,615]. Briefly, to extract CMCs and DDCs, they peeled cambium together with cortex, phloem and epidermis from the xylem (**Figure 2.14**.b) and laid them on a suitable growth medium[297]. Initially (after 4-7d), cell division could only be observed in cambium (CMCs!), with DDCs emerging from phloem, cortex and epidermis after about 15d by dedifferentiation. A clear visual distinction between flat, uniformly spread CMCs and irregular DDCs was possible after 30d (**Figure 2.14**.c-e), thought to be due to inconsistent proliferation in DDCs. After separating the populations, both were cultured independently in slightly altered media resulting, finally, in CMC and DDC populations with distinct morphology and functional characteristics (**Figure 2.14**.f). DDCs were also derived from needles and embryos following optimised, previously established protocols[624,626].

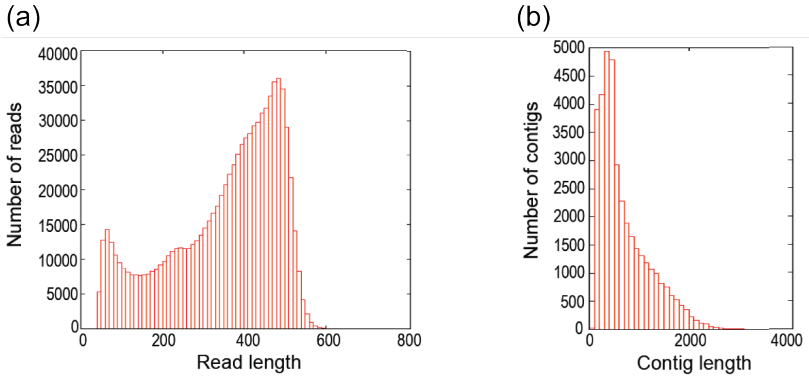

Figure 2.15: Transcriptome assembly. (**a**) Histogram of the read lengths in the pyrosequencing-based library used for transcriptome assembly. (**b**) Histogram of contig length in final assembly . All panels have been reproduced with permission from reference[297].

2.2.2.2 De-Novo Assembly of T. cuspidata Transcriptome and Digital Expression Analysis

Due to a lack of reference annotation, first a "transcriptome" needed to be assembled which could be used as a scaffold for alignment and a basis for the calculation and comparison of expression profiles. A (complete) transcriptome is a comprehensive set of the sequences of all mature transcripts of an organism. Methodologically, the assembly of a transcriptome nowadays typically involves the collection of total RNA from one or more cell types of the organism in question followed by HTS[40]. Sophisticated algorithms are then employed to put together partially overlapping sequences in order to construct full-length transcripts[40, 107, 455]. Naturally, the quality of a transcriptome assembly depends not only on the performance of this algorithm, but also on the depth of sequencing and the coverage of transcripts in the RNA sample provided. Consider, for example, a biased RNA sample from only one specific cell type will probably not contain all transcripts an organism is capable of producing – any transcriptome assembly based on such a sample would be inherently incomplete. More difficult to avoid, natural RNA samples are usually highly skewed towards strongly expressed genes and more rare transcripts might never be observed or sequenced if the coverage is not sufficient.

T. cuspidata RNA isolated from DDCs and CMCs by my collaborators was enriched for full-length sequences and rare transcripts and then submitted to the GenePool sequencing facility at the University of Edinburgh (`http://genepool.bio.ed.ac.uk`) for sequencing using a Roche/454 GS FLX instrument. A total depth of $860,800$ reads with an average length of $351bp$ per read was achieved (**Figure 2.15**.a). The GenePool assembled the reads into $36,906$ contigs[‡] using the Roche/454's own Newbler software (version 2.3; **Figure 2.15**.b). The contigs were annotated using BLAST[4] alignments against known protein and nucleotide

[‡]"Contigs" are continuous pieces of sequence build by assembling multiple reads into one. They may be thought of, with caution, as corresponding to transcript sequences.

sequences from similar plant species and Annot8r[483]. This procedure managed to successfully assign a putative function (by similarity) to about 62% of all contigs.

To quantify gene expression in CMCs and DDCs, purified RNA was prepared in triplicate (three samples each for CMCs and DDCs) for digital tag profiling / DeepSAGE with the *NlaIII* restriction enzyme according to Illumina's protocol (**Section 1.2.2.2**) at the GenePool and sequenced using a Illumina Genome Analyser II_x platform. The reads were truncated and extended to create meaningful tag sequences (as described in **Section 2.1.2**) and aligned to the previously assembled contigs using MAQ[310] (version 6.0.8). Only uniquely aligned tags were carried forward and taken into account for the calculation of tag counts per contig.

In summary, up to this point the primary computational work had been carried out by the GenePool core facility. The six datasets (3 CMC + 3 DDC) had been processed up to a stage where we had raw tag counts for 36,906 contigs, a large percentage of which had a putative function or homologous gene assigned to them.

2.2.2.3 Statistical Analysis of Differentially Expressed Genes

The final step in the data analysis was the identification of contigs that were differentially expressed between the two cell types, CMC and DDC. I decided to use, as previously (**Section 2.1.3.3**), the *edgeR* package[458] for this purpose, however, discovered after an initial trial using default parameters that many contigs had been called differentially expressed although their expression levels varied either (i) very little between samples groups or (ii) were inconsistent between replicates.

The first (i) was usually the case when the expression values in one class were very low or even zero. For these contigs, even a low expression in the other cell type was considered a strong change. This might very well be biologically relevant, but if the change was as low as from 0 to 1, I doubted it was distinguishable from the noise level in this assay.

The latter case (ii) was mostly due to only a single replicate exhibiting a drastic difference. Statistical methods, in general, are designed to account for such variation within groups, yet can sometimes fall victim to outliers. Notwithstanding a biological explanation, this phenomenon might well be due to a freak amplification of single tag sequences in some samples and one would not usually want to include the affected contigs in the candidate lists.

I sought to refine the analysis for the detection of highly-reliable candidate genes and to get rid of the suspected false positives (wrongly called differentially expressed genes) by optimising the parameter settings of *edgeR* and augmenting the analysis strategy with a pre- and a post-processing step.

Pre-processing: I first rescaled the raw tag counts in all libraries by dividing each count by the sum of the upper quartile of tag counts of the same library and subsequently multiplied the values with 1,000,000, effectively transforming the values into *reads per upper-quartile*

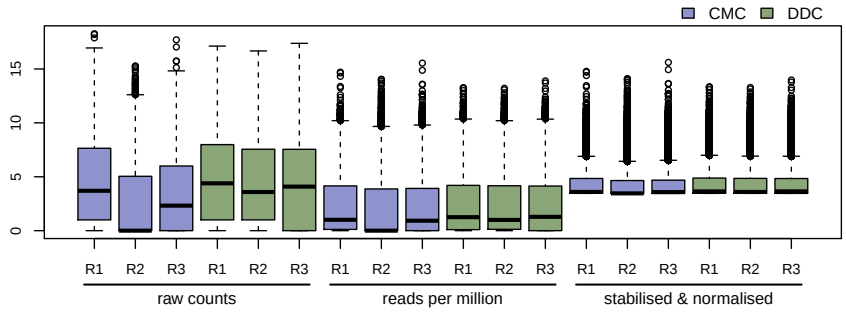

Figure 2.16: Normalisation of raw read counts. The boxplots show the distribution of expression values in their raw form (left), rescaled as reads per million (RPM, centre) and using the reads per upper-quartile million ($RPQ_{75}M$) normalisation followed by stabilisation described in this chapter (right).

million ($RPQ_{75}M$). This is meant to account for differences in library size by adjusting the counts in such a way that the most highly expressed contigs, which are also those usually most reliably detected, are on the same scale[57], reducing technical variability in library construction and sequencing. Next, a small stabilisation constant ($S = 10$) was added to each value, altering the signal to decrease the impact of difference between groups for very lowly expressed contigs, but leaving larger changes between more strongly expressed contigs largely untouched. Given $t(c_i)$ the raw tag count for an arbitrary contig c_i, Q_{75} the upper quartile of all tags counts, the full formula for the calculation of $RPQ_{75}M$ read counts amounts to:

$$RPQ_{75}M(c_i) = \frac{t(c_i) \times 1,000,000}{\sum_{c \in C} \rho(c) * t(c)} + S, \text{ where } \rho(c) = \begin{cases} 1, & \text{if } t(c) \geq Q_{75} \\ 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

Importantly, $RPQ_{75}M$ transformation alters the signal (raw reads counts) more strikingly than RPM (Equation 3.3), which essentially maintains the original distribution, deliberately neglecting contigs (or genes) with low detected expression estimates (**Figure 2.16**). For the analysis at hand, this was appropriate, since I was dealing with a poorly studied organism for which our transcriptomic assembly and annotations were likely to contain major flaws. The reduction of further sources of errors was therefore essential. For well-annotated model organisms, the same strategy might be less adequate and obscure weak, yet biologically relevant processes.

Statistical evaluation: To briefly recapitulate, *edgeR* uses an over-dispersed Poisson-distribution to model read counts after quantile normalisation in which the degree of overdispersion is moderated using an empirical Bayes procedure[458, 459]. A modified version of Fisher's exact test is employed to assess the probability that a gene or contig is differentially expressed.

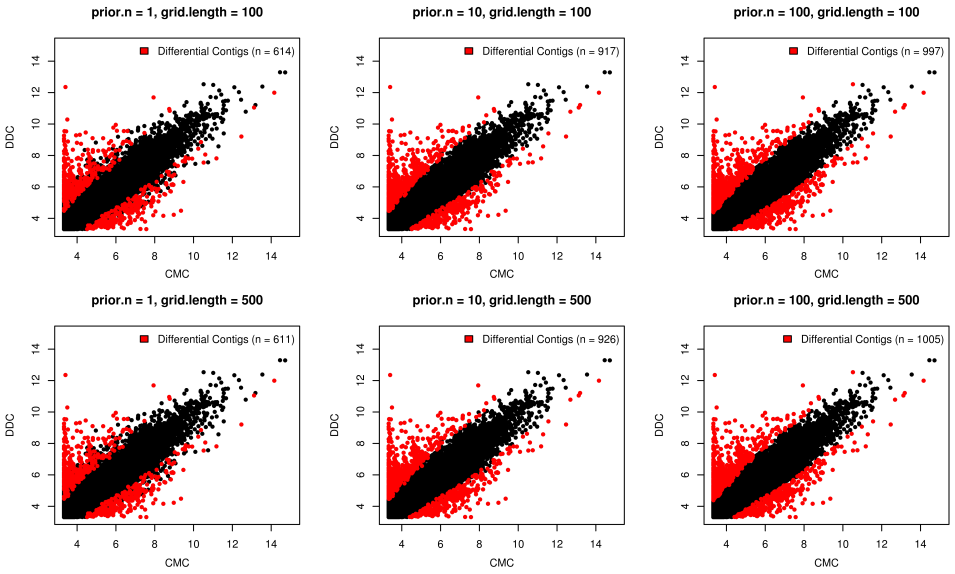

Figure 2.17: Optimisation of *edgeR* parameters. Scatterplots demonstrating the effect of the *edgeR* parameters *prior.n* and *grid.length* on the statistical assessment of differentially expressed genes. All values are log_2-scaled, quantile-normalised $RPQ_{75}M$ expression intensities averaged over three replicates. Differentially expressed contigs are highlighted in red ($FDR \leq 0.05$). N.B. The plots were created at a later time and with an updated version of *edgeR* (old version unavailable), which called, in general, fewer contigs as differential; the effect of all parameters remained equivalent.

I proceeded according to the steps outlined in the software's tutorials and experimented with the effects of the different parameters (**Figure 2.17**), finally setting on default values for all parameters but *prior.n* and *grid.length*, which I set to 10 and 500, respectively. Calculated p-values were corrected for multiple testing using the Benjamini-Hochberg method and I deemed a false discovery rate (FDR) threshold of $FDR \leq 0.05$ appropriate to detect differentially contigs, returning $1,229$ contigs as candidate factors for CMC/DDC identity.

Post-processing: Although all contigs detected by the statistical approach certainly merit attention, I decided to initially concentrate my investigations on contigs with particularly large and consistent changes, which were plausibly reasoned to have a notable effect on the morphological and functional differences observed between CMCs and DDCs. Thus, I filtered the candidates from the previous step ($n = 1,229$) by imposing a threshold on the minimum difference between any two replicates of both groups ($\Theta_{min.d} = 10RPM$) and retained only those candidates for which the direction of change (DOC) was consistent in all replicates, i.e. the replicates of one group (CMCs or DDCs) either had all higher or all lower values than those in the respective other group. A total of 563 high-confidence candidates were carried forward for further investigation.

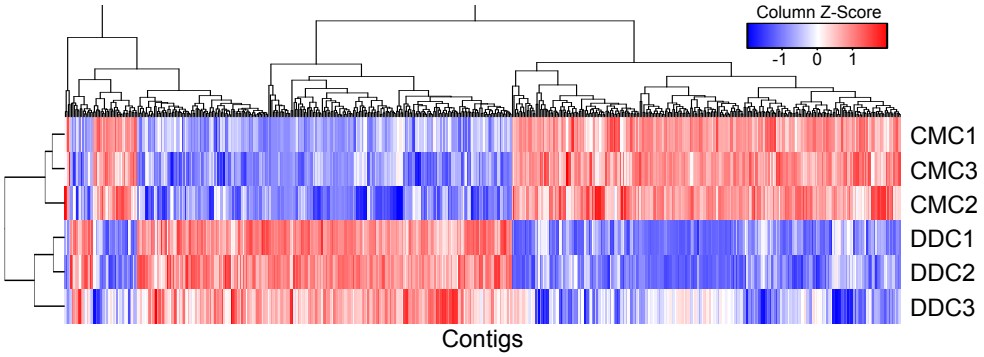

Figure 2.18: Candidate contigs for CMC/DDC identity. Hierarchically clustered heatmap of the expression levels of all assorted candidate contigs ($n = 563$). Colours are scaled per contig from lowest (blue) to highest (red) expression.

2.2.3 Results

I shall now discuss the results of the data analysis, briefly reviewing further insights gained by my collaborators during downstream investigations of the candidates discovered.

2.2.3.1 Candidate Factors for Cambial Meristemic Cell Identity

In the analysis, I identified several hundreds of high-confidence candidate contigs ($n = 563$; **Figure 2.18**), that, on the basis of the transcriptional data at hand, were considered likely to be implicated in the morphological and functional differences between CMCs and DDCs. Roughly an equal proportion of contigs were up- and down-regulated in CMCs with respect to DDCs ($n_{up} = 296, n_{down} = 267$). A selection of these contigs were validated using RT-PCR and qRT-PCR by my collaborators (**Figure 2.19**.a).

Interestingly, validated candidates included *contig01805*, which is highly similar (sequence similarity, see **Section 2.2.2.2**) to *Phloem intercalated with xylem (PXY)*, a member of a family of kinases that had previously been shown to be essential for the development of vascular tissue[134, 297]. Equally, *contig10710* had been found to be highly similar to *Wooden leg (WOL)*, known to be expressed in cambium of other plants and also believed to be affecting vascular development[333, 385].

These two contigs are merely examples of candidates that appeared reasonable targets for immediate follow-up study and many others exhibited similarity with proteins from other, better-studied organisms that were in line with stem cell-like properties of CMCs (**Figure 2.19**.b). Albeit my current results do not present any conclusive proof for the relevance of the candidates to proliferative and cell culture properties of CMCs nor for their role in the production of taxol (see next section, **Section 2.2.3.2**), this is a major first step towards this goal and demonstrates impressively how a combination of HTS approaches can be used to pinpoint biological factors with a putative functional role – even in poorly-studied organisms.

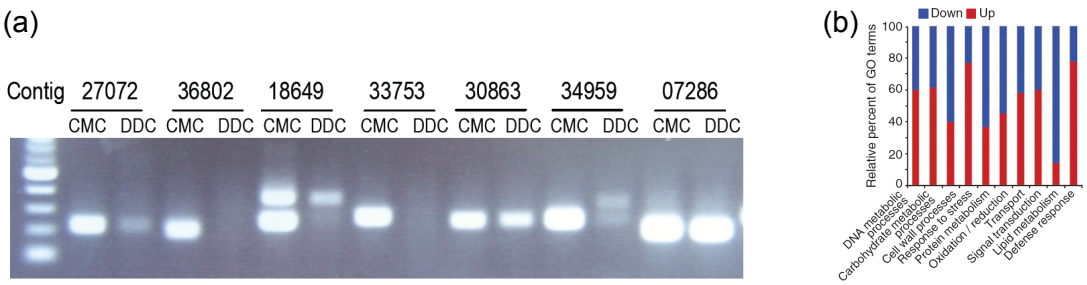

Figure 2.19: Validation and functional annotation of candidates. (**a**) Validation data for seven candidate contigs identified in the DeepSAGE differential expression screen. Ct07286 is a putative actin gene and was used as a control. (**b**) Relative frequency of gene ontology terms in groups of up- and down-regulated genes in CMCs with respect to DDCs. Validation data was generated by my collaborators and both figures have been reproduced from reference[297].

2.2.3.2 Clinical and Industrial Relevance of Findings

The transcriptional assays described before constituted only a minor part of the research project as a whole and in the further development of the investigations, my collaborators were able to produce convincing evidence of the different functional roles[297]. Firstly, CMCs clearly outperformed DDCs (either derived from embryos or needles) in terms of stable growth and proliferation potential on solid media (data not shown) and even more strikingly in suspension cultures in bioreactors of different sizes (ranging from 3 litres (**Figure 2.20**.a) to a 3 ton bioreactor suitable for industrial-scale production).

Measurements of the amount of taxol produced by CMCs in comparison to DDCs revealed an increased taxol biosynthesis potential of CMCs. Cells of both types that were cultured, again, on solid media (data not shown) or in bioreactor suspension cultures of various sizes and elicited to induce taxol biosynthesis by the addition of methyl-jasmonat, chitosan and a precursor phenylalanine. In all cases, CMCs produced consistently more taxol than DDCs (**Figure 2.20**.b). Assays of the production of abietanes, which also have been reported to suppress tumors[126], reported similar trends (**Figure 2.20**.c), suggesting that the phenomenon is not restricted to taxol biosynthesis only.

Preliminary experiments have also shown that CMCs in other plant species exhibit similar properties appealing for the production of natural plant products. CMCs extracted from ginseng (*P. ginseng*) and cultured in a bioreactor produced more than 20-fold higher amounts of ginsenosides – attributed, for instance, with neuroprotective and antioxidative effects – than ever reported[297] (**Figure 2.20**.d).

In conclusion, cultured CMCs might in future provide the means for the large-scale, cost-effective production of medicines, cosmetics and other chemicals from plant products. Cultures are largely independent of climate and at the same time require less space than full-scale plant cultivation making them a very sustainable and affordable platform for this purpose[297].

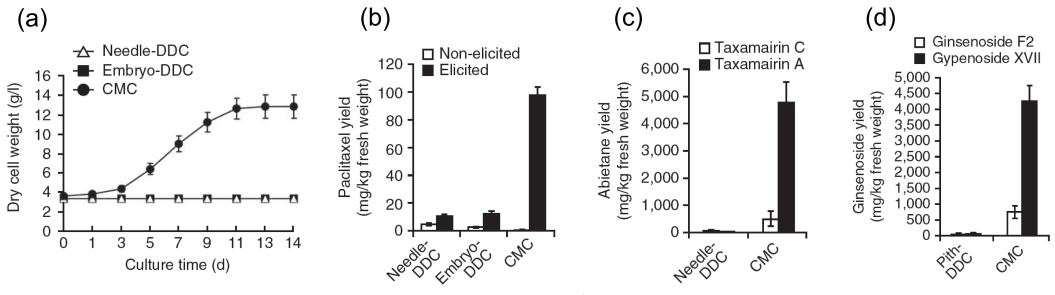

Figure 2.20: Growth potential and biosynthesis of CMCs. (**a**) Measured growth of DDCs derived from needles and embryos and CMCs in a 20 litre airlift bioreactor. (**b**) Taxol (paclitaxel) production in elicited 6-month old cell cultures after batch culture in a 3 litre air-lift bioreactor. (**c**) Production of the abietanes taxamairin A and C in DDCs and CMCs grown in a 3 liter air-lift bioreactor. (**d**) Production of the ginsenosides F2 and XVII in ginseng (*P. ginseng*) DDCs (pith-derived) and CMCs. Cultured in a 3 litre air-lift bioreactor. The data for these plots has been generated by my collaborators and all figures have been reproduced from reference[297].

2.3 Conclusions

High-throughput sequencing techniques have been much discussed in the recent years and many have predicted that they are soon to become the method of choice for transcriptional profiling on the large scale and may in the near future replace the still pre-dominant microarrays in this respect[246]. In other areas, for example, the study of DNA-protein binding or histone modifications (ChIP-seq), genome-wide methylation (Methyl-seq) or the discovery of genomic variations (resequencing), HTS has already surpassed its predecessors. In this last section of the current chapter I shall briefly discuss the major advantages and drawbacks of HTS with a particular focus on the conclusions I reached from my own exploratory pilot studies.

2.3.1 Unbiased Genome-Scale Assays of Gene Expression and Regulation

Many reports in the early HTS-related literature praised reproducibility, robustness and precision combined with the prospect of gaining a (largely) unbiased view of the whole transcriptome – even of unknown transcripts or in uncharacterised species – as the major advantage of the new technology over microarrays for assays of gene expression[340,526]. The pilot studies could confirm the applicability of deep sequencing platforms to the study of stem cells. The detected expression levels largely agreed with comparable intensities from Affymetrix microarrays and showed evidence for a wider dynamic range (**Figure 2.11** and **Figure 2.12**). Using HTS, I managed to detect features that could not previously have been found due to their limitation of microarrays to a fixed set of oligonucleotide probes, for instance, non-coding RNAs. I also found evidence for wide-spread anti-sense transcription and expression of genomic regions

outside the boundaries of known transcriptional features.

Since the sequencing technology can equally well be applied to non-transcriptional samples (**Section 1.2.2**) the prospect of using the same platform to investigate different aspects of the same biological samples offers an additional attractive bonus. One may expect highly consistent results from different perspectives on the same problem with a bare minimum of additional effort and costs. An example of such an holistic investigation of the genome is the ENCODE project[542], which has greatly helped our understanding of the general workings and regulation of transcription.

However, it has also become clear that at the current state of the art the costs associated with a sequencing project are still too high to be considered for routine use (being up to 10-fold higher than they would be using microarrays). In the beginning of 2012, after several years of research and development and despite early optimistic predictions, commercial HTS platforms have yet to rival the cost and processing times that make microarrays such an appealing technology. The application of HTS to transcriptome profiling therefore still remains a niche application for those that require the sensitivity (e.g. single cell studies[223, 534, 535]), seek to refine genomic annotations[59, 179, 552] or study alternative splicing events[45, 256, 413, 576].

A further increase in throughput combined with the possibility of multiplexing samples, also referred to as "bar-coding", which is now being made possible on most sequencing platforms, promises to soon lead to a massive drop in costs as several libraries can be read out in parallel (**Section 1.2.2**). This approach now becomes increasingly popular and wide-spread and offers exciting opportunities for future research[223] (**Section 1.2.3.1**).

2.3.2 High-Throughput Data Requires High-Throughput Analysis

The manifold applications of HTS make it necessary to incorporate, combine and juxtapose many heterogeneous kinds of data at once. Additionally, it was demonstrated that the integration of alternative functional genomics data, such as from microarray platforms (**Section 2.1.2.3**), can help to leverage an experiment's primary data even further creating additional insight and better understanding of the mechanisms under study[221, 555, 575, 632].

It may be expected that modern functional genomics technologies will in the coming years accumulate an amount of biological data unparalleled even by microarray technology (which, on January 23^{rd}, 2012, has amassed data from $27,858$ experiments or $686,135$ individual samples in the database of the Gene Expression Omnibus[22, 118]). An efficient use of these data is key to gaining a better understanding of biological functions, development and disease[632].

Currently, the advance of HTS is still hindered by data analysis challenges[338]. In order to harness the information that is now at our disposal, high-throughput data generation needs to be accompanied with high-throughput, integrative data analysis. The diverse tools that have

already been developed for several aspects of the HTS analysis pipelines (e.g. Bowtie[292] or edgeR[458,459], which I have used in this chapter), need to be made more widely accessible by all scientists and the burden of getting started with the data analysis must be reduced to allow more researchers to more rapidly exploit the data to its full extent. Additionally, I believe the community would greatly benefit from knowledge extracted from HTS experiments being more readily and quickly accessible.

With these conclusions in mind, I felt compelled to set out on the task of developing a new software system that would in future allow research to progress more smoothly and empower science by making experimental data, no matter how large and complex, accessible, interpretable and reusable at any time and from anywhere in the world. My efforts shall be described in detail in the following chapters (**Chapter 3** and **Chapter 4**).

Chapter 3

An Analysis Environment for RNA-seq and ChIP-seq Experiments

In this chapter, I shall describe the GeneProf software system, a graphical environment for the analysis of HTS experiments created in the course of my research project. Rather than just giving a description of the software itself, I will start by reiterating my motivation (**Section 3.1**) for developing this program, followed by a short account of the initial release version (**Section 3.2**) and then go into detail about the key challenges addressed in the software design process (**Section 3.3**). I will conclude the chapter with a brief evaluation, compare GeneProf with related software and highlight room for future improvements (**Section 3.4**).

3.1 Motivation and Goals

Why did I set out to write this new piece of software? In the recent years, novel HTS technologies have revolutionised the way in which biological researchers study the molecular mechanisms and effects of gene expression (**Section 1.2**). This impact is witnessed by an ever-increasing number of publications and by the unprecedented wealth of data that is now available. In late 2010, the Sequence Read Archive (SRA), the world's largest database of HTS data, boasted over 500 billion reads[306], a number which has almost tripled a year later (`http://www.ebi.ac.uk/ena/about/statistics`).

The huge volume and complexity of data produced by high-throughput sequencing (HTS) platforms make it difficult for many research labs, which may lack expertise and computing infrastructure, to fully harness the potential of HTS for the study of biological processes and

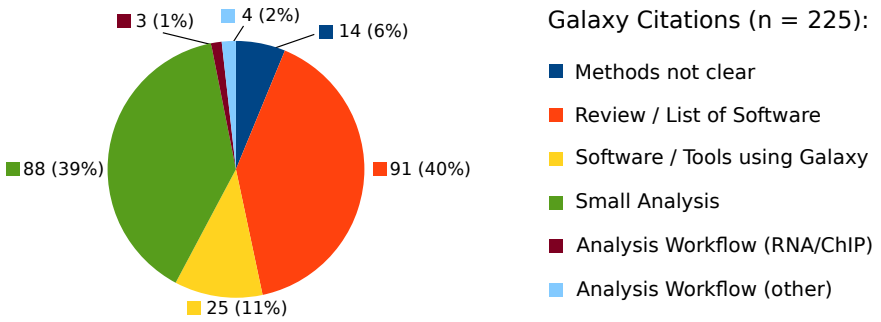

Figure 3.1: Citations of the Galaxy workflow engine. Galaxy is the most widely used environment for workflow-based analysis in biological research to date. Proportional distribution of citations of the three main *Galaxy* publications[43,153,160] obtained from Google Scholar (`http://scholar.google.com`, on August 29th, 2011) .

human disease. Although a large array of software has been developed to address individual aspects of the analysis process (such as the alignment of sequence reads to the genome or transcriptome or the detection of significant binding events or the quantification of gene expression; cp. `http://seqanswers.com/wiki/Software` or `http://www.stemdb.org/bioresources`), this software is at times difficult to set up, use and, especially, combine. As a consequence, we have now reached a point where the data processing rather than the data generation step may often become the bottleneck of biological experiments in terms of cost as well as time[338].

Workflow-based software suites, such as *Galaxy*[43,153,160] and *Taverna*[213,398], offer an attractive approach for dealing with complex data, because they allow visual combination of simple software components into large "workflows", enabling complex analyses without any need to write custom computer scripts. However, current workflow engines mostly focus on the computational processes involved, rather than on achieving particular biological goals. They usually attempt to provide an extremely flexible solution, often with the aim of being domain-independent, and therefore split up logical processes into many granular units. Setting up a workflow can be a daunting and time-consuming task for many life scientists, especially those without experience of the visual programming paradigm used. Thus, usage of workflow engines in biology has so far been mostly limited to small aspects of the analysis process or to only expert users with a computer-programming background (**Figure 3.1**). Existing tools are hence often not sufficient to make HTS fully accessible to the entire research community (**Section 3.4.1**).

Moreover, the effective reuse and integration of published research data from various sources is still a challenging task for most researchers. I believe that scientists would benefit greatly from a quick and easy way to look through published research data and to compare these with their own findings. To warrant a sensible comparison of data from different sources, it is essential that the entire process leading to the analysis results can be recapitulated and reproduced.

It was with these issues in mind, that I started working on a new software suite, which would

- integrate proven methods and tools into one coherent environment,
- make it easier for computational and experimental biologists alike to set up and run elaborate analyses workflows,
- boost the use of consistent and established methodologies by guaranteeing reproducibility and transparency,
- keep the biology at all stages at the heart of the system and facilitate interpretation of complex data with intuitive visualisations and helpful summaries
- and ease access to and reuse of public HTS data to avoid replication of efforts and costs.

3.2 The GeneProf System

To address the issues outlined in the previous section, I have created a software suite called GeneProf, which has recently been released to the general public. I will now first attempt to give a short overview of the software and system architecture (adapted from the supplementary material of reference[182]) before going into detail about design challenges and decisions in the next section (**Section 3.3**).

3.2.1 Overview

Foremost, GeneProf is a graphical software suite for the analysis of high-throughput sequencing data from RNA-seq and ChIP-seq experiments. Combining an array of well-established, popular algorithms and tools with an assortment of custom-developed functionality, researchers can channel arbitrarily complex analyses processes through the system taking them all the way from unprocessed, "raw" input data files to biologically meaningful results. At the same time, GeneProf acts as a comprehensive resource of integrated, readily interpretable findings by making the results of analysis performed within the system available via a user-friendly web interface (**Chapter 4**). Apart from searching, browsing and visualising these findings, all users may also reuse any data in their own analyses, broadening the impact and profitability of the original data and enriching new experiments to a scope otherwise not feasible (**Section 3.3.2.4**).

GeneProf simplifies the analysis workflow construction by providing assistive web forms ("wizards") that build elaborate workflows without exposing users to the underlying complexities of workflow programming (**Section 3.3.2.2**). These wizards abstract common, best

practice analysis steps into a series of logical stages, which researchers can customise quickly by answering only a few basic questions. The wizards provide a great entry point for new users and reduce the hands-on time required to perform analyses. Importantly, users may change all wizard-generated workflows later on to suit specialised requirements, so GeneProf does not sacrifice the full methodological flexibility offered by the workflow-based approach.

Data and analyses within GeneProf are tightly coupled by organizing both into "virtual experiments"[159]. The experiments are supplemented by all intermediate results and a history of the entire analysis procedure, not unlike a lab book. Researchers can link to these experiments in publications or share their analyses securely with collaborators prior to publication. All data and results remain the intellectual property of the user and are confidential until made public, at which point every visitor of the website can view the entire experiment and search, browse, visualise and export data. Importantly, registered users can easily import and reuse public data in other experiments.

The primary user interface for the application is completely web-based (**Figure 3.2** and **Section 3.3.2.1**), eliminating all setup costs for users: No additional software needs to be installed. GeneProf makes use of a dedicated, remote compute cluster (**Section 3.2.2** and **Section 3.3.4.2**), which carries out large-scale genomic analyses and dynamically balances the load between concurrently running processes over a network of computers. Given the vast amount of data produced by modern HTS platforms, this is of paramount importance to maintain the performance and scalability of the software as it gains a wider user base.

In a typical use-case, a researcher would upload her primary experimental data, e.g. short read sequences output by a HTS platform, to the GeneProf server or import published data from the Short Read Archive or the European Nucleotide Archive[305,306] using the built-in importer tool. One would then proceed to use one of GeneProf's wizards to set up a data analysis workflow. The constructed workflow will then be submitted for execution, which means it will be entered into a queue. A cluster of computers is constantly monitoring this queue and one node (that is, one computer in the network) will soon pick up the process and execute the analysis (**Section 3.3.4.2**). Once completed, the user will be notified by email and can then assess the outputs of the analysis following a link in the email. Primary analysis results (e.g. lists of binding sites for a transcription factor or differentially expressed genes) are automatically supplemented by a range of informative summary statistics and plots and researchers can use these to quickly gauge the outcomes of the analysis. At this point, more experienced users may decide to change parts of the workflow, e.g. by adjusting parameter settings or by adding additional components to the workflow, to deal with specialised requirements (**Section 3.3.2.3**).

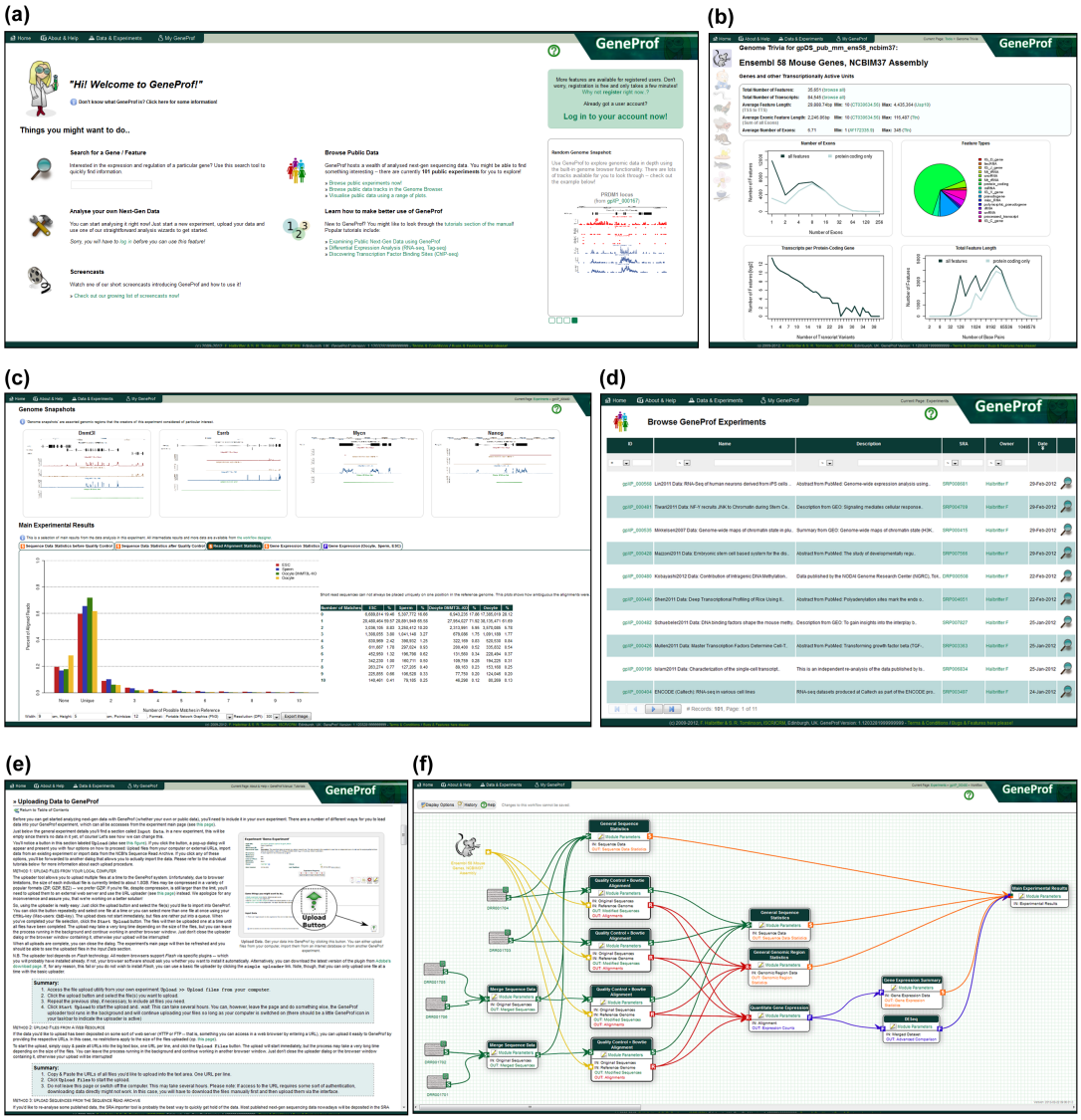

Figure 3.2: GeneProf web interface. GeneProf's primary user interface is completely web-based. (**a**) The GeneProf homepage is the primary access point to the application. (**b**) "Genome trivia" pages provide information about the genomes and genes in all supported reference datasets. (**c**) The experiment main pages provide an overview of all input data, the analysis workflow and main results for each experiment. (**d**) A large amount of public data is available for browsing via the website. (**e**) An extensive online manual is provided for all components of GeneProf. (**f**) The "workflow designer" shows a visual representation of a data analysis workflow and allows simple manipulation of the analysis via drag&drop of modular components.

3.2.2 System Architecture

GeneProf as a whole consists of three major components: A central web server, an assortment of databases and an arbitrary number of "job agencies and workers" (**Figure 3.3**).

3.2.2.1 Web Server

The GeneProf web server hosts all of the application's web pages and dynamic components and constitutes the only part of the system exposed to direct user interaction. The GeneProf web server handles all essential aspects of user management and the confidentiality of user data, acts as a primary interface between web front-end components and the GeneProf databases, converts data between different formats on demand and creates plots, data representations and summaries for the interface. Crucially, the web server acts as an intermediary between the experiment (processing job) queue and the user, allowing her to submit new jobs and track (or cancel) existing ones. Recently I have also added an alternative access layer, called the *GeneProf Web API*, which enables programmatic retrieval of data by computer programmers and data analysis experts for use in external web sites or programs.

3.2.2.2 Databases

GeneProf stores all its data in a combination of a relational database system and a file server (**Section 3.3.4.1**). Other than user-submitted scientific data, such as short read sequences and genomic data, which make up the core of what GeneProf is all about, these data comprise user records and other internal information such as, for example, the experiment (job) execution queue.

Smaller units of data and those information that require quick, random-access retrieval as well as dynamic filtering, sorting and the like can conveniently be stored in a relational database. In GeneProf, this means that all internal data as well as gene-centric data and reference annotations (called "Feature Data" and "Reference Data", respectively, throughout the GeneProf interface) are stored in this part of the database. Large chunks of data and data that does usually only require sequential access, on the other hand, ought to be stored on a file server. Here, I make use of a variety of compressed binary data formats to efficiently store and retrieve bulky data, such as short read sequences and genomic data (e.g. from alignments), effectively saving (disk) space and time (data access), which are both of major concern when dealing with the volume of data that we are presented with by modern functional genomics technologies.

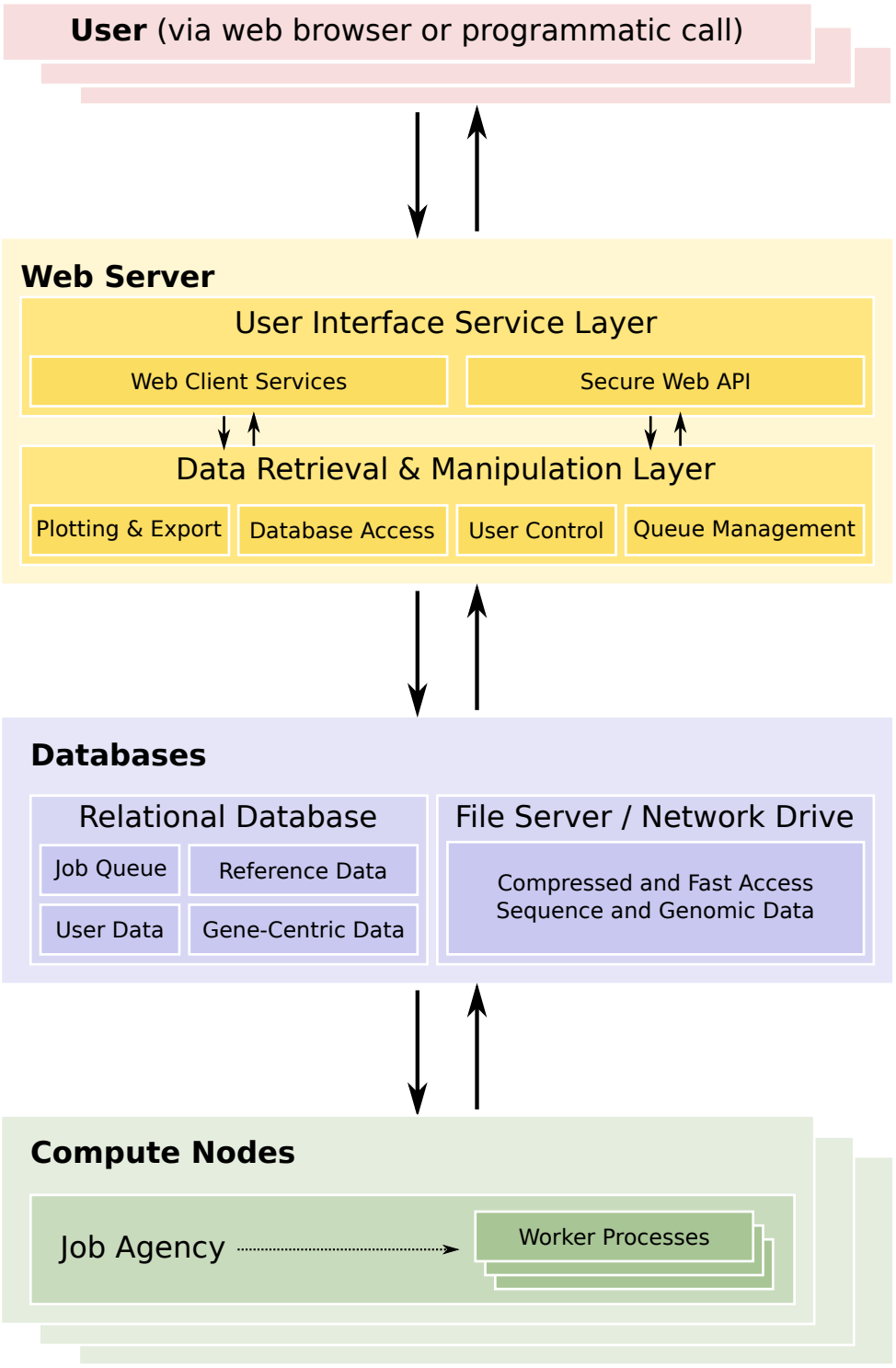

Figure 3.3: System architecture. GeneProf is split into three major components: A web server manages all client-side interactions, provides interface components and acts as the primary access point for job management. A combination of a relational database and a file server stores all experimental and internal data in a space- and time-efficient manner. Lastly, a flexible network of compute nodes ("job agencies" and "workers") deal with computationally demanding tasks.

3.2.2.3 Job Agencies and Workers

A powerful computer is of paramount importance to much of the data analysis performed in state-of-the-art bioinformatics workflows. It is not uncommon for individual processes to take several hours until completion and to require an amount of memory not currently available on most standard desktop workstations. GeneProf has therefore been designed to exploit a network of compute nodes to perform all processing steps required (**Section 3.3.4.2**).

I call these compute nodes "job agencies". Each job agency independently and constantly monitors the current experiment queue and waits for new jobs pending execution. When a new experiment is entered into the processing queue, one job agency will pick up this experiment and spawns a new "worker" process for this experiment's workflow. Each job agency may run several such worker processes in parallel and additional job agencies can be dynamically added to (or removed from) the computer pool to deal with changing data processing demand.

3.2.3 Availability

A public instance of the GeneProf web application, the primary interface to the GeneProf system detailed in the previous section, is hosted on infrastructure located at the Institute for Stem Cell Research / Centre for Regenerative Medicine of the University of Edinburgh. Funding for the purchase and maintenance of the hardware, which comes at no insignificant cost, was kindly provided from a combination of sources, foremost the European Commission Seventh Framework Programme 'EuroSystem' and the Centre for Regenerative Medicine.

The interface is now available to the general public at `http://www.geneprof.org` and academic researchers may use GeneProf free of charge for their own analysis projects.

3.3 Software and Algorithm Design and the Key Challenges Addressed

Let us now look in detail at some of the major concerns for the development of a software suite such as GeneProf and explain how these were addressed in the design and implementation of the software.

3.3.1 A Generic Framework for Executing Analysis Processes

A software suite for data analysis needs to be both comprehensive and flexible, while being easy to use. Striking the right balance can be a tricky task. Most bioinformatics tools and algorithms are being developed as command-line-based software only. Traditionally, computer programmers appreciate the flexibility of command-line programs, because, given the necessary

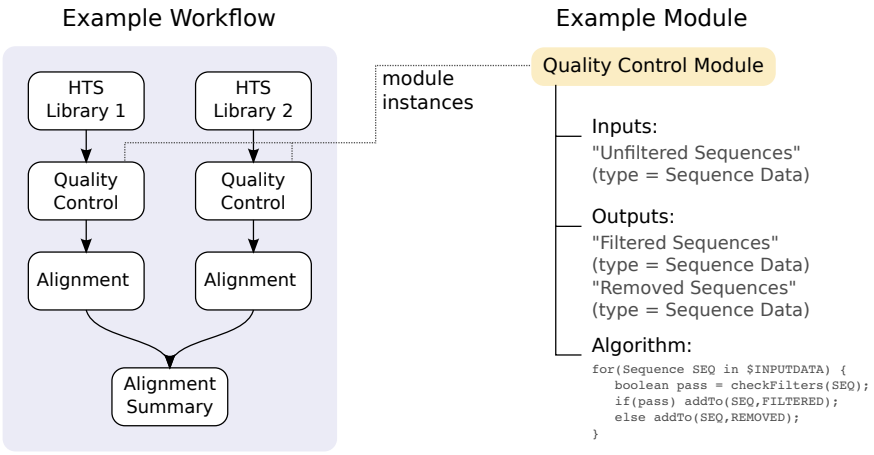

Figure 3.4: Workflows and workflow modules. A workflow (left) is made up of components, which are instances of one or more workflow modules (right). The outputs of one component may be used as inputs for another component. A workflow module is defined by its inputs, outputs and the algorithm that transform the former into the latter.

experience, it is possible to wire them together in arbitrary ways by writing custom computer scripts and inter-converting data formats between steps. This empowers experts to combine simple individual programs into pipelines (or "workflows") achieving complex outcomes.

Workflow-based software suites, such as *Galaxy*[43,153,160] and *Taverna*[213,398], offer an alternative approach for dealing with complex data, because they allow users to visually combine simple software components into ordered "workflows", enabling complex analyses without any need to write computer scripts. In effect, researchers need to spend less time working out how to use tools and can focus more on the actual analysis. However, existing workflow engines focus solely on the interconnection of individual programs. Their goal is to achieve computation, but not a particular biological goal.

I decided to use a workflow-based system at the heart of GeneProf, but to let users focus on achieving high-level analysis goals rather than low-level computational tasks. To do so, I added assistive tools that simplify workflow construction (see **Section 3.3.2.2**). GeneProf's workflows are made up of components that are instances of so-called "workflow modules". Modules are small pieces of computer code, that aim to achieve a certain goal by effectively transforming a set of input datasets into one or more output datasets. In earlier workflow software, these modules usually map directly to different command-line programs and the outputs of one process might have to go through additional modules in order to be converted to the right format for the next module's input. GeneProf's modules, on the other hand, correspond to logical stages in the analysis process, e.g. there will be one module for short read quality control (**Section 3.3.3.1**) or one for gene expression quantification (**Section 3.3.3.3**). Quite often the modules do indeed also map to an underlying (external) program, but this is by no means necessary: A module might well combine several programs into one unit, if that is

necessary to achieve a biological result. GeneProf makes use of internal data types and handles the conversions between formats automatically – effectively, shifting the responsibility of worrying about data formats from the user to the module programmer. These two key features, biology-focused modules and automated format conversion, make workflow construction substantially more straightforward and intuitive.

Importantly, from a programmer's point of view, GeneProf is still a workflow-based system, which offers some convenient advantages for developers: Benefiting from a comprehensive framework, bioinformaticians and algorithm developers can easily implement additional functionality without needing to worry about peripheral data processing requirements. For example, in order to develop a new alignment tool (**Section 3.3.3.2**) it should not be necessary to deal with issues of quality control or what could be done with the aligned reads afterwards. Developers can rely on GeneProf's framework to take care of these issues and only need to specify the particular types of inputs they require for their program and define which types of outputs are produced. Following a well-defined specification, additional functionality can be rapidly and efficiently implemented.

GeneProf currently (software version v1.1203282) features 80 workflow modules and many more are under development. For a complete list of all modules refer to **Section D.4**.

3.3.2 Making High-Throughput Sequencing Widely Accessible

There are now masses of HTS data published in the literature every week. Equally, every week sees the release of new software and tools refining methods for part of the analysis process and experts constantly improve the protocols and workflows dealt with. For many experimental biologists and bioinformaticians alike, it is practically impossible to keep track of all the latest algorithms and the expertise required for in-depth data analysis. This challenge holds back the optimal exploitation of HTS data to its full potential and hinders the progression of science. In the following sections, I will discuss how GeneProf attempts to ease access to HTS for researchers from all backgrounds without extensive training and without special equipment.

3.3.2.1 A User-Friendly Web Interface

The first step towards an accessible data analysis suite accessible is a user-friendly interface. As previously discussed, most bioinformatics software is delivered as command-line tools (**Section 3.3.1**). This is partly as a consequence of the publication-driven funding and partly due to the fact that good algorithm developers do not always make good interface designers. A graphical interface, though, helps to decrease the burden of getting used to a new piece of software. A good interface stands out by more than just the visual appeal – although the visual impression makes the overall user experience more pleasant: The interface helps novice

users to quickly discover the main functionality and guide the learning experience towards more advanced features. Experienced users benefit from interfaces that allow to speed up or even automate the handling of common tasks.

Vitally, interface design starts before the program is even started up the first time, at the installation process. Many potential users are (rightfully) scared off by complicated or poorly documented installation procedures, especially, if these include many external dependencies or even the operating system- or hardware-specific compilation of components that are not bundled with the main software.

An attractive approach to overcome the installation burden and present users immediately with a usable, graphical interface is the delivery of software via a web interface: Most researchers nowadays will be familiar with the use of a web browser and many will have experience with at least some of the successful web applications developed by others[135, 153, 259, 290]. The responsibility for the set up of software and dependencies lies with the provider of the service. Similarly, software updates can be managed centrally and users can always benefit from the latest release version without having to install updates themselves. Another advantage is, that users can access the software from anywhere, which might be of particular importance in a collaborative research environment with scientists accessing the same data analysis projects from their office or home computer or even from different sites across the world. Likewise data and results stored on the web server will be immediately available across sites.

With these considerations in mind, I chose to implement the primary user interface of GeneProf as a Java Enterprise web application (**Section 3.2.2.1**). Java technology has a proven track-record of delivering high-quality, stable and large-scale web applications and is one of today's most used and popular programming languages with a extensive set of publicly available extensions and software components allowing for rapid expansion of the system. A dedicated, high-performance compute cluster manages computationally demanding analysis processes in the back-end (**Section 3.2.2.3** and **Section 3.3.4.2**), so no special equipment will be required to use GeneProf: Any reasonably modern computer with a web browser will do (tested on Windows, Mac and Linux using Mozilla Firefox 3.5+, MS Internet Explorer 8+, Chrome, Safari and Opera).

The GeneProf homepage is a good example of how I attempted to make the application accessible to users with different levels of background knowledge. The page (**Figure 3.5**) summarises much of GeneProf's functionality at a glance: Apart from the navigation bar (shared between all pages, right at the top of the page, as will be familiar to most users from other web pages), the home page streamlines simple and rapid entry to some of the most common activities. Without further ado, users may search for data about genes of interest, start a new analysis project, browse public datasets or open the manual, tutorials and help pages. Furthermore, the page highlights some examples of analysis results and the latest

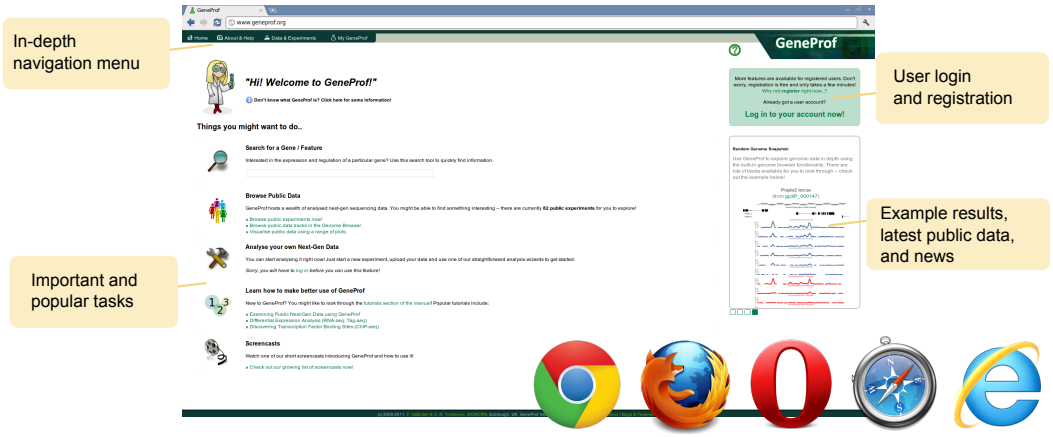

Figure 3.5: GeneProf homepage. The homepage of the GeneProf web interface is the primary entry point to the data analysis and search functionality of the software suite. Users may easily and quickly start new analysis projects, continue existing ones, browse data and results made public by other users or access advanced visualisations.

experimental data made public by users of the application. This allows new users to get a grasp of what the software is about and might help them to discover interesting findings relevant to their own research.

The whole application follows a tiered access model, starting with simple tasks and introducing users progressively to more advanced functions of the system: Novice users can start by looking through public data and analyses performed by others and then proceed to start a new experiment, upload their own data and use the built-in analysis wizards to set up a standard analysis workflow (**Section 3.3.1** and **Section 3.3.2.2**). As users become more experienced, they can start modifying the analysis workflows in detail or even set up completely new ones on their own using the dynamic workflow designer tool.

I have designed a number of step-by-step tutorials to help people get started. The tutorials cover topics such as how to make the best use of public data and the analysis of RNA-seq and ChIP-seq data. Additionally, all pages of the user interface, all analysis modules and important concepts are explained in detail in the online manual (`http://www.geneprof.org/help_and_tutorials.jsp`).

Lastly, GeneProf has a built-in bug and feature request tracking component. It can be very frustrating to get stuck at some point using a new software application due to technical fault or missing functionality. Such problems cannot always be foreseen and avoided, but a successful software system will be open for input and respond to feedback by the user community. For this purpose, I have wired a simple issue tracker tool into the GeneProf web interface. The advantage of a built-in solution over more feature-rich existing frameworks, e.g. Bugzilla (`http://www.bugzilla.org`) or Mantis (`http://www.mantisbt.org`), is the seamless integration into the GeneProf framework. There is no need to set up further user accounts or redirect to external pages, instead, users can issue reports directly from within

the application using their normal accounts.

3.3.2.2 Integration of Expert Knowledge

A user-friendly interface with good help and tutorials goes a long way when accessibility of a data analysis software is concerned, but even the best interface design cannot necessarily replace the expertise and experience that is often required to perform complicated data analysis tasks. As we will see later on (**Section 3.3.3**), HTS data analysis is a diverse process and involves numerous steps where informed choices need to be made about how best to proceed. Even if a software tool opens up all the possibilities and makes them easy enough to apply, new users will be baffled by the choice and find it difficult to proceed sensibly.

I sought to alleviate the problem by assisting users in their decision process. I established best practice protocols for common data analysis scenarios based on the literature and then built this knowledge into the GeneProf application by supplying assistive web forms, called "wizards", for these scenarios (**Figure 3.6**.a). Most users will be familiar with wizards from other applications such as installation wizards for programs of all sorts, office text processing products or the like. GeneProf's wizards abstract low-level analysis steps into a series of logical stages, replacing the manual construction of workflows as combinations of workflow modules (**Section 3.3.1**) with a few simple questions that need to be answered by the user. On the basis of the answers, the software will then automatically construct an analysis workflow by connecting together an appropriate series of workflow modules. Essentially, the wizards conceal one layer of additional complexity, which will be of particular benefit to novice users, but even expert data analysts benefit from the use of wizards for rapid, streamlined data analysis.

Importantly though, the wizard-created workflows are not static and can subsequently be adjusted manually to customise the workflow and suit specialised requirements. In the next section (**Section 3.3.2.3**), I will demonstrate why this is of great importance for actual, powerful data analysis.

At the moment, GeneProf features two wizards for constructing full-scale, start-to-finish analysis workflows:

- **RNA-seq Analysis.** This wizard combines GeneProf's custom-built quality control procedures (**Section 3.3.3.1**), with short read alignment (**Section 3.3.3.2**) using either the Bowtie[292] or Tophat[550] software, gene expression quantification (**Section 3.3.3.3**) and differential expression analysis (**Section 3.3.3.4**) using the DESeq algorithm[7] (**Figure 3.6**.b). In addition, informative summary statistics and plots will be created at all stages of the analysis process.

- **ChIP-seq Analysis.** Like the RNA-seq wizard, this wizard uses quality control and

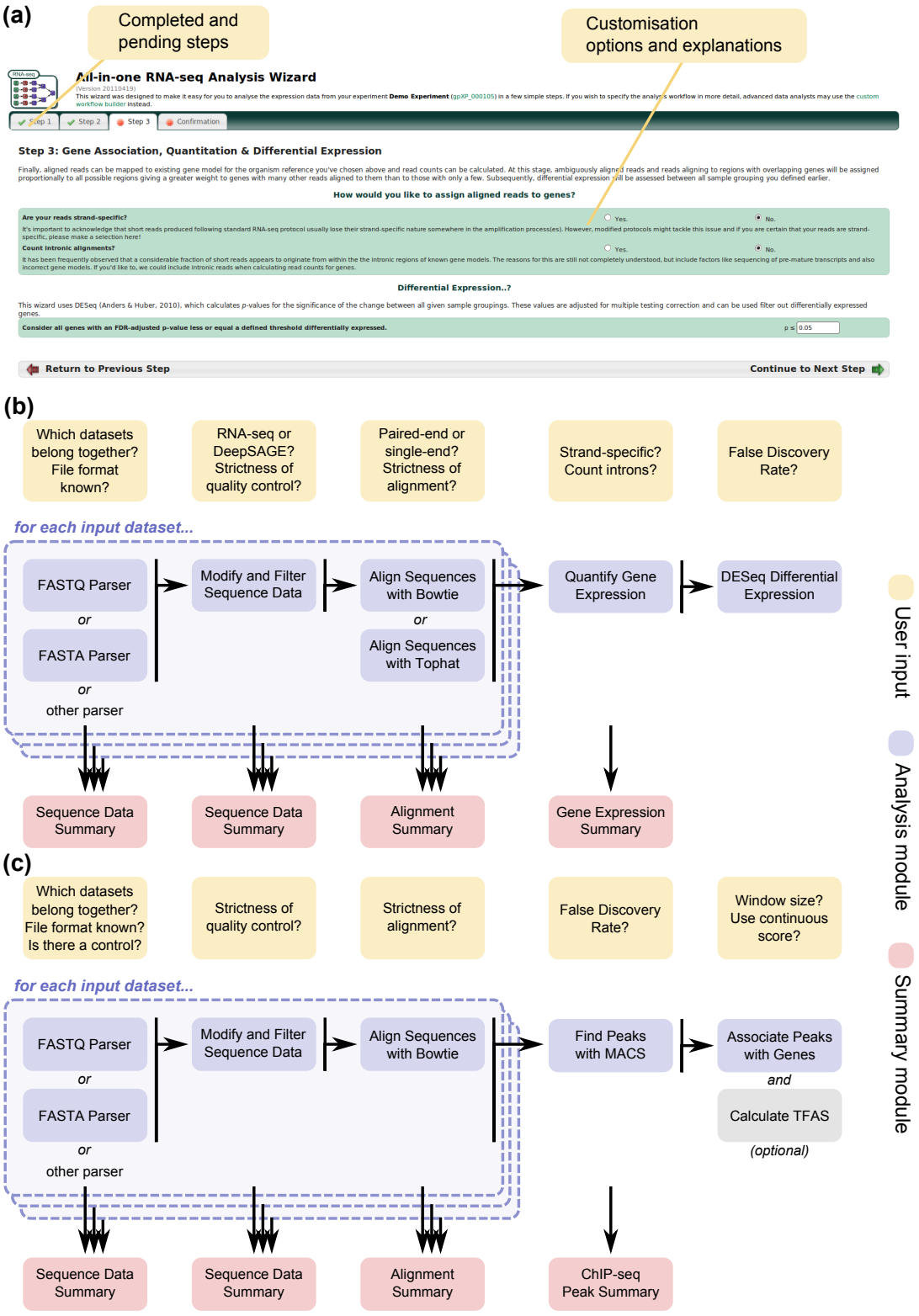

Figure 3.6: Analysis wizards. (**a**) Screenshot from the third step of the configuration of the RNA-seq wizard. The green ticks at the top-left indicate that steps 1 and 2 have already been completed. (**b**) Schematic representation the workflow created by the RNA-seq wizard by putting together workflow modules. (**c**) Workflow created by the ChIP-seq wizard.

alignment modules for the initial stages of the analysis. Alignment is followed by binding peak detection (**Section 3.3.3.5**) with the MACS software[631] and the association of those peaks with genes (**Section 3.3.3.5**) using custom-built and published methods[406] (**Figure 3.6**.c). The wizard has initially been designed for the analysis of transcription factor binding sites (TFBS), but I found it also useful for the analysis of other ChIP-seq data, e.g. for histone modifications.

In addition to the above-mentioned, there are three additional wizards simplifying aspects of the analysis process:

- **Quality Control.** This simple wizard streamlines the task of running many short read libraries through GeneProf's quality control modules.
- **Alignment.** If many datasets are to be aligned to the same genome, this wizard can speed up the workflow construction significantly by extending an existing workflow with the appropriately connected alignment modules (using Bowtie[292]).
- **Gene Expression.** Finally, this small wizard manages the quantification of gene expression intensities from a number of aligned short read datasets using custom-built modules.

I believe that GeneProf's wizards will in future help to improve the consistency of analysis protocols by providing tested and proven methodologies building a skeleton for further analysis. Existing wizards may be easily updated to take novel tools and methods into account and additional wizards (e.g. for specialised histone modification analysis, miRNA and short RNA data and the like; see **Section 3.4.4**) can be added as required.

3.3.2.3 Enabling Exploratory Data Analysis

The wide spectrum of applications made possible by HTS make it impossible to devise one solution that fits all analysis requirements. GeneProf's analysis wizards (**Section 3.3.2.2**) constitute a solid basis for advanced analysis by providing an established basic workflow for almost any type of analysis, but it will frequently be necessary to customise the workflows subsequently to achieve optimal results. Usually, the adjustments required are not very far-ranging and quite often the correction of just a few parameters might suffice. Also, it is not always possible to know at the outset of a data analysis project the best way to deal with the data at hand. For instance, how could one definitely decide on a way to deal with the quality control aspect of the analysis without knowing what the quality of the data is like?

GeneProf has been designed to support exploratory data analysis and make progressive adjustments straightforward and quick to deal with. Workflows constructed using the data analysis wizards will include special modules calculating informative summary statistics and

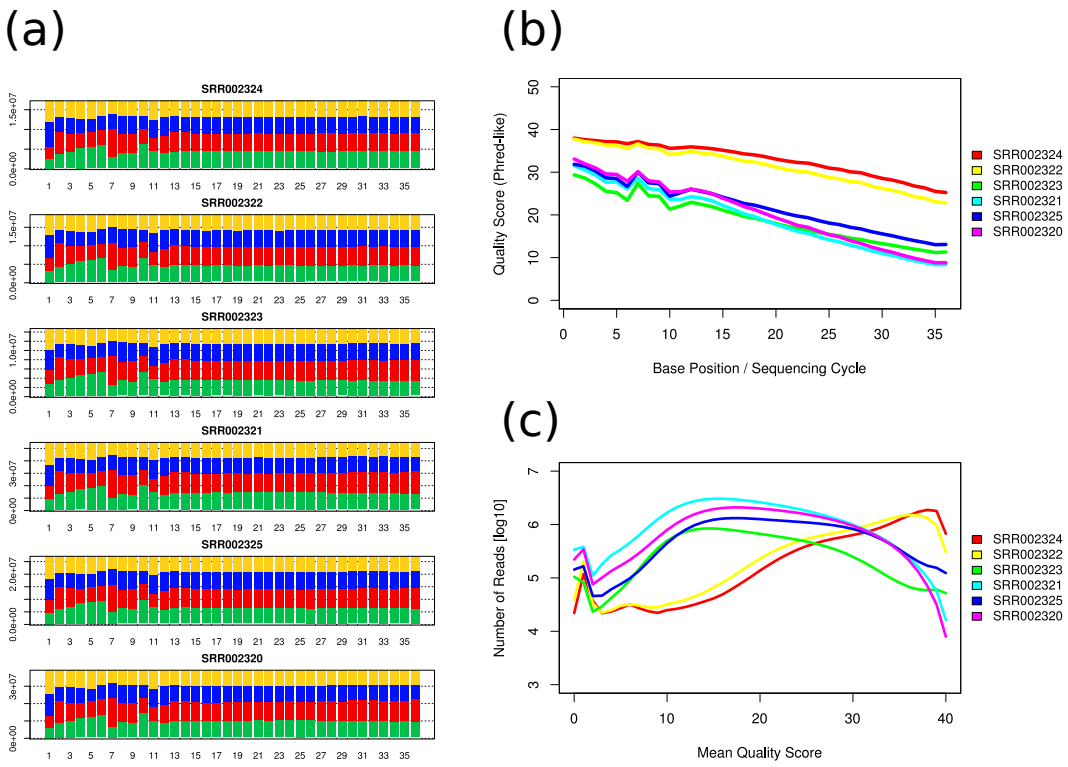

Figure 3.7: Sequence summary plots. Automatically created plots summarizing short read data of RNA-seq reads[340]. All plots in this example are from the GeneProf experiment `gpXP_000058`. **(a)** Nucleotide composition of short reads in all individual libraries across the length of the 36*bp* sequences. As is often observed, the distribution is slightly skewed in the first bases, but becomes more uniform towards the end of the reads. **(b)** Average Phred-like quality score per sequencing cycle and library. The quality drops notably with progressive sequencing cycle. Interestingly, the qualities are recovered in cycle 7 after an initial drop, probably thanks to an automated recalibration. **(c)** Frequency of reads with a certain average quality score. This plot can help to decide on appropriate thresholds for discarding low-quality reads.

plots at various stages of the process. The summaries make it easier to get a feel for the data and to spot flaws in the analysis procedure or data.

For example, I have often observed that the quality of short reads, especially in earlier HTS libraries where the technology was still quite new, declines rapidly with the length of the reads. That is, base calls at the end of a read are less reliable than those at the beginning, because errors accumulate in later sequencing cycles (**Section 1.2**). Such shortcomings are readily spotted in the pre- and post-quality control sequence summary statistics calculated by GeneProf alongside the primary analysis (**Figure 3.7**) and, if it turns out that the alignment of the sequences to the genome is hindered by the presence of too many erroneous bases, it might be advisable to trim off a portion of the read. GeneProf's quality control module can be customised to perform the trimming either statically, by cutting off a fixed number of nucleotides from the end of each read, or dynamically by trimming off the ends after the quality drops below a certain threshold (**Section 3.3.3.1**). After adjusting the parameters, GeneProf will automatically re-run all parts of the analysis that were dependent on the altered

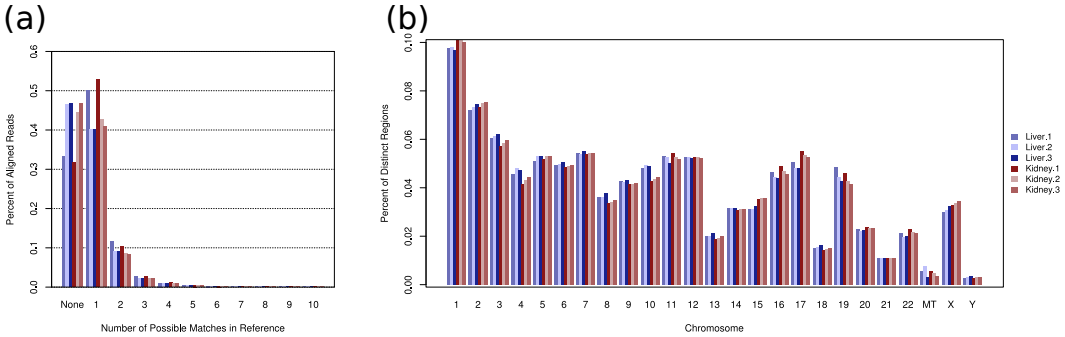

Figure 3.8: Alignments summary plots. Semi-automatically created plots detailing information about the alignment of RNA-seq reads[340]. All plots in this example are from the GeneProf experiment `gpXP_000058`. **(a)** Ambiguity of alignments is given as the number of possible matches in the genome identified for any one particular read. Unaligned reads or reads with more than 10 possible alignments are listed as "none". In two of the liver libraries over 45% of all reads could not be aligned, which might be problematic, but is not unusual in early HTS libraries. **(b)** The distribution of reads across all mouse chromosomes (including the mitochondrial pseudo-chromosome). The distribution is similar in all libraries and reflects the density with which genes are spread across the chromosomes.

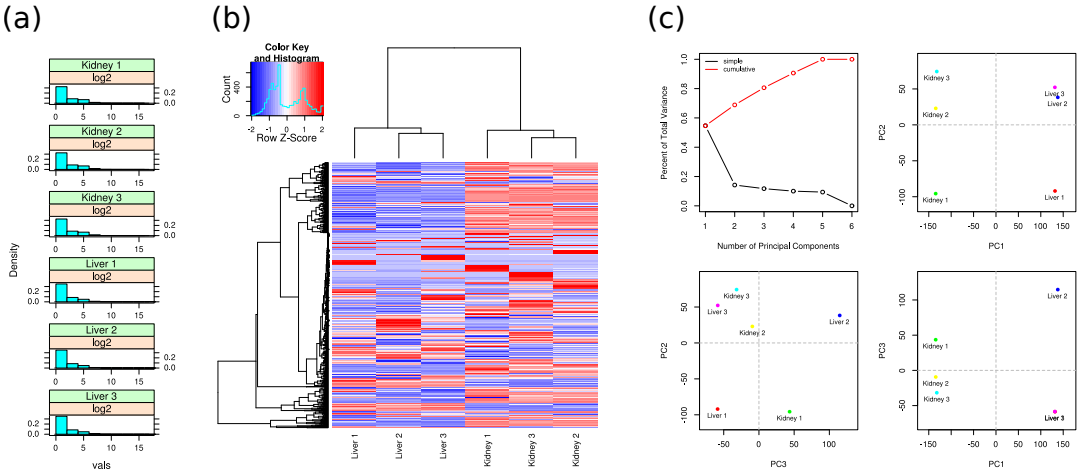

Figure 3.9: Gene expression summary plots. Semi-automatically created plots to support the interpretation of gene expression data. All plots in this example are from the GeneProf experiment `gpXP_000058` with data from a published RNA-seq study[340]. **(a)** Histograms of the log2-scaled expression values (reads per million) in the independent HTS libraries. **(b)** A heatmap of 1,000 randomly selected genes clearly demonstrating the similarity between libraries from the same tissue. Some genes which appear to be differentially expressed appear at the top of this heatmap. **(c)** Visualisation of the contribution of the individual libraries to the first three principal components (PCs). The first PC explains some 58% of the variance of the data and separates kidney nicely from liver.

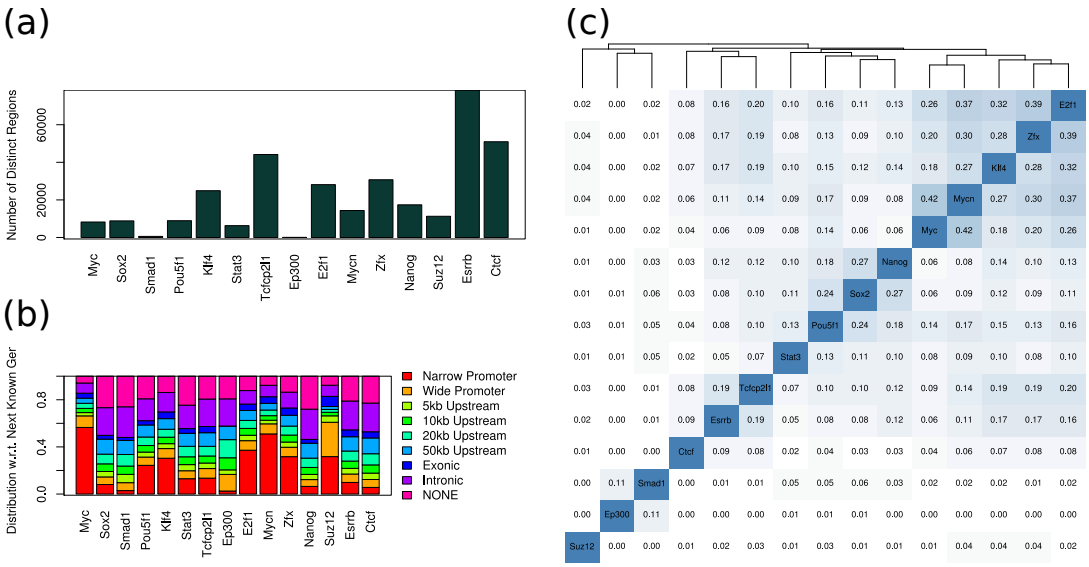

Figure 3.10: ChIP-seq peak summary plots. Semi-automatically created plots to support the interpretation of DNA-protein binding data. All plots in this example are from the GeneProf experiment `gpXP_000012` with data from a published ChIP-seq study[75]. **(a)** Number of putative binding sites (ChIP-seq peaks) detected for the 15 DNA-associated proteins studied. **(b)** Distribution of these binding sites with respect to known genes. Each binding site is assigned to one of the following categories: "Narrow / wide promoter" = within $0.5kb$ (narrow) or $2kb$ (wide) up- or downstream of the transcription start site (TSS) of a gene, "exonic" = anywhere within an exon of a gene, "intronic" = anywhere within an intron of a gene, "5 / 10 / 20 / 50kb upstream" = up to 5 / 10 / 20 or $50kb$ upstream of the TSS and "none" = none of the other categories. **(c)** Pair-wise overlaps of binding sites. The numbers (and colour intensity) report the percentage of binding peaks that appear in both libraries. Overlaps are calculated after extending the peaks by $500bp$ in both directions.

modules to make sure that results are consistent.

Thus, the combination of wizards with automated summary statistics and simple customisation of workflows empowers researchers with a novel path for rapid exploratory data analysis:

1. Create a basic workflow using an appropriate wizard.
2. Assess all relevant summary statistics.
3. If the statistics indicate any problems, adjust the analysis workflow and re-run, then return to step 2.
4. Proceed with downstream analysis, wet-lab work, etc.

There are four common types of data summaries used by the wizards (although they can, of course, also be employed in manually constructed workflows):

- **Sequence Data Summary.** Analysis of the composition of short read libraries in terms of the number, length and frequency of reads, their nucleotide composition and the base-

call quality scores, if available (**Section 3.3.3.1**). This information can be used to spot problematic sequencing runs or erroneous cycles. For an example, see **Figure 3.7**.

- **Alignment Summary.** Overview of the outcome of the alignment of one or more HTS libraries containing information about the number of aligned reads, the genomic distribution of alignments over chromosomes and alignment ambiguity (**Section 3.3.3.2** and **Figure 3.8**), useful as a gauge of alignment success rate and to spot genomic imbalance or bias.

- **Gene Expression Summary.** Statistics and plots describing the distribution of gene expression values in one or more libraries, supplemented by heatmaps, histograms and principal component analyses (**Section 3.3.3.3** and **Figure 3.9**). This information helps to get a feel for the genes expressed in datasets and visualises the similarity (or difference) between multiple libraries.

- **ChIP-seq Peak Summary.** An overview of the number and lengths of peaks in a dataset. The analysis will also look at the distribution of binding sites with respect to known gene models, e.g. by checking how many peaks fall within promoter, upstream or genic regions, and at the overlaps of peaks from different proteins (**Section 3.3.3.5** and **Figure 3.10**). Not only does this summary help to more quickly get an impression of the binding behaviour of one protein, but it is also highlights potential interactions of several factors.

3.3.2.4 Data Providence and Transparency

With the rapid rise of HTS, there was initially a distinct lack of established tools and methodologies for appropriate data analysis. As a consequence, many research labs had to come up with novel, *ad hoc* solutions to the problems they were facing. The methods sections of HTS-based publications (in particular the early ones) are most diverse and often riddled with "custom scripts" patching together analysis workflows. It has previously been observed that such cryptic methodologies lead to irreproducible results[220]. Publications with well-documented methods and readily available data, on the other hand, tend to be cited more often[426].

In order to critically assess published findings, it is essential that other researchers can evaluate and assess the primary research data, understand the way in which it was analysed and repeat the procedure. Successful approaches can serve as protocols for similar studies and is desirable that the methods are clear enough for others to exploit them for their own investigations. For this to work, two requirements need to be fulfilled: Firstly, unprocessed, "raw" experimental data needs to made publicly available. Most biological journals

do now require high-throughput datasets to be made available via public repositories such as the Gene Expression Omnibus (`http://www.ncbi.nlm.nih.gov/geo`) or the Sequence Read Archive (`http://www.ncbi.nlm.nih.gov/sra`). Secondly, the analysis procedure needs to be described in detail. While research journals, of course, expect the methods to be described for any accepted publication, the scope and format of the articles often make it impossible for scientists to include every minute detail such as, for example, which parameter settings have been used for algorithms run or which software versions were used – these will not be of interest for most readers, but may have a drastic effect on the outcome of the analysis. If custom software were used, this does additionally need to be made available, but external dependencies, that is, other programs installed on the developer's computer, can make it difficult to emulate the environment in which the analysis was originally performed, again potentially changing the results of the analysis.

In order to tackle these problems, it has been proposed to run data analysis in tracked environments keeping a record of the complete history of analysis steps and program executions[160, 357, 398]. GeneProf's workflow framework provides the ideal platform for the implementation of such a strategy: The analysis modules applied in GeneProf provide a good repertoire of advanced analysis functions. Every change to the analysis workflows as well as each execution of the individual modules is tracked via the system and presented to the user in the form of a complete, transparent analysis history, not unlike a lab book. Software versions are carefully controlled and legacy versions of outdated modules are kept to ensure the repeatability of previous analyses. Unlike in other systems, GeneProf's workflows incorporate all the scientific data. Existing software usually considers the analysis workflow a distinct entity of the data at hand: The workflow itself is a tool (or a protocol) that can be applied to different datasets. In GeneProf, however, each workflow is one instance of the combination of several tools to one set of data. In other words, a GeneProf workflow is one complete analysis experiment. I found that this helped experimental scientists to conceptualise complex analyses.

Analyses carried out within GeneProf can be made public in conjunction with the publication of research findings. They may be linked in articles to supplement the methods section and an automatically generated summary report covering the entire experiment from input data via analysis workflow to the results, may optionally be included as a supplementary document. This makes it more straightforward to include details about the data analysis methodology and helps scientists in future to easily recapitulate work carried out by others. We are making every effort to maintain public data in the system indefinitely and any GeneProf user may import public data into their own experimental workflows to enrich their analysis, effectively not only facilitating the reuse of established methodologies, but also of existing experimental data, helping to save costs and effort in data generation.

3.3.2.5 Visualization of Large-Scale Data

The interpretation of the outputs of large-scale functional genomics experiments is a challenging task. While it might be possible to look at, say, individual genes, the sheer mass and extent of data make it difficult to grasp the findings as a whole. Visualisations help to identify consistent patterns and derive advanced conclusions.

As far as genomic data is concerned, one of the most successful and useful methods for visualising large amounts of data has come in the form of genome browser software such as the UCSC Genome Browser[259], Ensembl[135] or IGV[457]. Genome browsers display a linearised version of the genome overlaid with a selection of annotation tracks, e.g. for known gene models or other regions of interest and alignment data (**Section 3.3.3.2**). Users can "browse through the genome" and examine particular regions, for instance, the surroundings of a gene implied in the regulation of a particular biological mechanism, to investigate expression patterns (RNA-seq data) or the binding of regulatory proteins (ChIP-seq data). This is a very quick and straightforward, yet incredibly efficient way to spot interesting patterns in genomic data (for an example, see **Figure 4.4**).

I decided to integrate a simple genome browser, making use of the *GenomeGraphs* package for *R*[115], directly into the GeneProf web interface to allow users to quickly get a feel for their own research data and to compare these with other genomic information available in the system. This browser is capable of juxtaposing up to 50 tracks based on GeneProf alignments of ChIP-seq and RNA-seq reads, binding peaks or other, arbitrary pieces of genomic information at once and without further processing by the user. The visualisations can be customised in a number of ways, e.g. by changing the colour, labels and plotting methods for individual tracks, and can be exported in various publication-quality image formats or as a set of *R* scripts to allow further customisation by experts. Examples of plots generated via GeneProf's genome browser will be shown later (**Figure 5.8** and **Figure 5.10**). For more advanced features and high-volume usage (GeneProf's genome browser cannot rival the speed of established, specialised software), users may export the genome annotation tracks in a variety of popular formats and use those files with another genome browser software of their choice.

Another powerful visualisation feature in GeneProf is the "Visual Data Explorer" (VDE), a hub for rapid creation of plots from large collections of datasets. The VDE accesses GeneProf's repository of public experimental data (**Chapter 4**) and offers selected techniques to plot data from many different experiments together. The data can be grouped by various annotations allowing users to look at the same data from many different angles. This opens innumerable ways to visualise the data. The VDE is currently still in an early development stage (**Section 3.4.4**), but already has three different visualisation techniques, namely correlation matrices, principal component analysis and histograms. These plot types have been chosen,

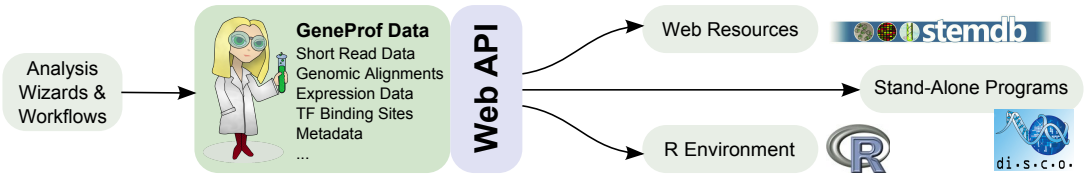

Figure 3.11: GeneProf WebAPI. The GeneProf WebAPI works as an interface between GeneProf's extensive databases of experimental data and results and external programs on the web or stand-alone software.

because they allow to concentrate large amounts of data into comparatively simple, easily interpretable plots and many scientists will be familiar with them[75,92,171]. Further advanced visualisation methods will be added in future releases. Please refer to **Figure 4.5** for examples of plots created using the current version of the VDE.

3.3.2.6 Integration with Other Software

The diverse range of applications and linked requirements for data processing of HTS (**Section 3.3.3**) make it impossible for any software developer (or software development team) to cover the entire field of algorithmic tools necessary. I have therefore never been under any illusion that GeneProf might be a universal solution for all researchers. That being said, I believe that the suggested software suite provides a solid foundation for most HTS-related research and should be sufficient to carry out the majority of tasks desired. Advanced downstream analysis, however, might at times benefit from the use of additional software not (or not yet) integrated into GeneProf. Rather than trying to outdo specialised tools, I have attempted to make GeneProf work together with them by providing functionality that makes it possible to transfer data from the GeneProf databases into external software.

The functionality in question has been summarised into a software component called the "GeneProf Web Application Programming Interface (API)" (**Figure 3.11**). The web API is a specification of web services by which advanced data analysts and computer programmers can retrieve data from the GeneProf web application via a well-defined set of hypertext transfer protocol (HTTP) requests, or, in other words, a set of universal resource locators (URLs) with parameters. Apart from the actual data, the web API can also be used to retrieve metadata about experiments and datasets.

I have specifically investigated the use of the web API for the integration with three software packages or environments and shall now briefly illustrate how the interaction will work:

- **R:** The *R*[435] framework for statistical computing is a powerful and popular platform for bioinformatics work, especially thanks to the availability of many add-on libraries via the Bioconductor repository[151]. GeneProf can export gene-centric and genomic data in a file format that can be loaded directly into *R* and by using the `Rcurl` package a direct

connection to the GeneProf Web API can be established, effectively allowing users to load data into an active *R* workspace as if the data was loaded from a local hard disk.

This mode of interaction facilitates highly-customised, in-depth downstream analysis with *R*, while benefiting from the rapid, visual and traceable data processing offered by GeneProf.

- **Unix command-line:** Specialists who are familiar with the use of Linux, Macintosh or other Unix-based operating systems, may concatenate Unix's command-line tools into rather complex chains. Using tools such as `wget`, it is straightforward enough to stream GeneProf data via the web API to any command-line tool in a Unix environment. Of course, IT-savvy users are not limited to basic Unix-tools, but can use the web API in conjunction with any command-line-based bioinformatics software that is set up on their computer.

- **DI.S.C.O:** As a prototypical application of the web API for the integration of GeneProf data with other advanced graphical tools, I have furthermore provided import and export functionality to support the use of short read alignment and RNA-seq gene expression data in DI.S.C.O. (Skylaki, L. & Tomlinson, S.R., *manuscript in preparation*), a graphical software tool for genomic clustering analysis developed in our group.

  Similar import/export functionality could be provided for many other tools with minimal effort.

The web API is fully documented on the GeneProf website and can be accessed from:
`http://www.geneprof.org/help_advancedtopics.jsp`.

3.3.3 Data Processing Requirements

The prospective uses of HTS technology for the study of diverse biological mechanisms are virtually unlimited and the ways in which data analysts deal with the data produced are certainly no less diverse. Nevertheless, certain set of tasks is pervasive to all analyses independent of the specific nature of the experimental setup and it is crucial for any software system targeted at HTS data analysis to support and streamline these processes.

3.3.3.1 Assessment and Control of Raw Data Quality

The success of any biological experiment stands or falls with the quality of the experimental data: Where data is flawed, unreliable or plainly wrong, researchers might easily be misled into drawing incorrect conclusions. It is therefore of paramount importance to assess and confirm the quality of input data prior to further processing and to take appropriate actions wherever doubts arise. Like any other large-scale assay, HTS data is subject to a multitude of steps

Quality Score Q	Error Probability $p_n(x \geq 1)$ $n = 1bp$	$n = 36bp$	$n = 100bp$
10	0.1000	0.9775	1.0000
20	0.0100	0.3036	0.6340
30	0.0010	0.0354	0.0952
40	0.0001	0.0036	0.0100

Table 3.1: Phred quality scores. The probability of reading out at least one incorrect base pair ($p_n(x \geq 1)$) in a read of length n, if all nucleotides were of quality score Q. Based on `http://en.wikipedia.org/wiki/Phred_quality_score`.

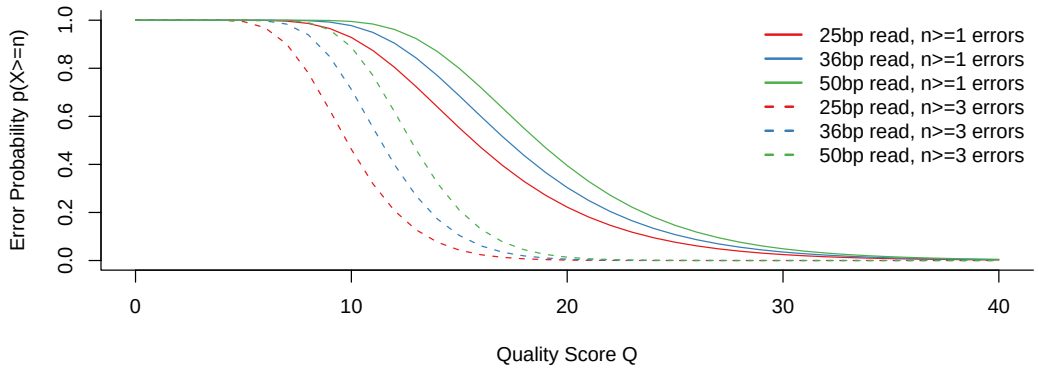

Figure 3.12: Read error probabilities. The probability (y-axis) of finding at least one (solid lines) or three (dashed) miscalled bases in a short read sequence of a given length, rises strongly with dropping read-quality (x-axis).

that might introduce artefacts, biases and errors into the process and, while it is not always possible to avoid those flaws completely, it is important to be aware of potential problems so one can account for them when interpreting the data.

Issues affecting HTS data can be broadly divided into two categories: (i) Problems introduced during materials handling and sample preparation, either due to procedural, human error or caused by technical faults, and (ii) errors in the sequencing process itself, that is, problems impairing the quality of the read-out of the correct nucleotide sequences.

Although the sequencing instruments and protocols have been considerably optimised over the last years to tackle both types of issues, it is still advisable to assure the quality of any new dataset produced. The de-facto standard format for delivering HTS data nowadays is FASTQ, a simple text-based file format, which, in addition to the nucleotide sequence of each read, stores a measure of data quality for each nucleotide in the read, the "quality score". The quality score reports, for each nucleotide, the probability that the respective read-out is correct and corresponds in scale to a Phred-like score between 0 and 50 (**Table 3.1** and **Figure 3.12**). The scores are provided by all major HTS platforms, although the technical details of how they might be estimated vary. In the FASTQ format, these numbers are encoded as characters

so that it is possible to represent each number as a single symbol. Unfortunately, due to a lack of standardisation in the early days of HTS, a number of variations of the format have emerged that differ slightly in the way these characters are encoded[85], that is, the characters in different versions of the file format will actually represent different numbers. The convention does now seem to converge increasingly to the use of the version of the format as championed by the Welcome Trust Sanger Centre (Hinxton, UK), but especially for older datasets it might sometimes be necessary to convert between different encodings – unfortunately, it is not always possible to determine automatically which encoding is being used. In GeneProf, I decided to always use the Sanger-style format and to make an attempt at automatically suggesting the correct format of uploaded datasets by looking at the range of values encoded by the characters and subsequently converting any non-standard data to the default format. This procedure usually works reasonably well, yet will at times require user input to correct mistakes. Thanks to the use of a system-wide default format, users do not usually have to worry about different formats any more and can focus on the interpretation and use of the data.

By using the quality scores as well as information about the nucleotide composition and distribution and the frequency of reads, one may draw conclusions about the overall quality of an HTS library and it might be possible to single out and remove or trim erroneous reads[96,489] or to correct them based on distribution assumptions and similarity to other reads in the library[216,257,352,475]. As described earlier (**Section 3.3.2.3** and **Figure 3.7**), GeneProf summarises raw short read data in a collection of informative plots detailing information about the quality scores and nucleotide composition of the reads. The information gathered from these reports can be used in conjunction with a special workflow module (**Section 3.3.1**) to efficiently handle problematic data by either filtering out reads that fail to pass user-defined criteria on the basis of average, minimum or cumulative quality score, sequence complexity, nucleotide content and length or by dynamically trimming leading or trailing erroneous fractions off otherwise good-quality reads. In order to make it easier for inexperienced users to choose sensible thresholds for this step, I have devised three levels of strictness that should generally achieve good results:

- Level 1 - "lenient": Only the very worst reads (average quality score $mean(Q) < 8$) will be removed from a dataset. This setting is currently the default, being the most conservative option, and might be the most advisable to use, in particular, for older datasets.

- Level 2 - "stringent": Reads will first be trimmed after the first occurrence of a uncertain nucleotide call (`N`). Any read which after trimming is shorter than $12nt$ or has an average quality score $mean(Q) < 15$ will be removed. The option will actively try to trim only

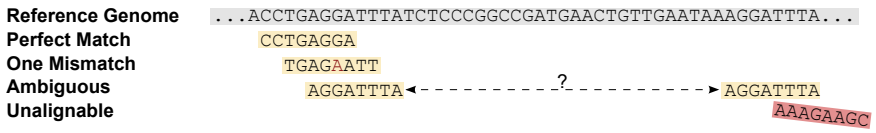

Figure 3.13: Short read alignment. Illustration of different, conceivable scenarios for short read alignment. Short reads, here 8bp in length, are aligned to a reference genome. Often, a unique match in the genome can be identified, especially, if permitting for mismatches. Some reads will align to multiple possible positions and others again will fail to align at all.

low-quality parts of reads, but ought to still maintain most reads in modern datasets.

- Level 3 - "draconian": Any read with a average quality score $mean(Q) < 20$ or which contains any base call with a quality $min(Q) < 10$ or which contains an uncertain base call (`N`) anywhere, will be discarded. This is the strictest version of the filter and might remove a sizeable fraction of some datasets, but will ensure that the remaining reads are of exceptional quality and reliability.

In summary, GeneProf encourages users to look into the quality of their raw data and provides the tools to filter out problematic reads. The quality control process is straightforward and quick and I hope that this will help to improve the awareness of potential issues in future applications.

3.3.3.2 Short Read Sequence Alignment

The area that has probably attracted most attention in the early days of HTS is the alignment of short read sequences. "Sequence alignment" is the process of arranging two nucleotide sequences (DNA or RNA) next to each other (the same principle applies to protein, i.e. amino acid sequences, but shall not be further discussed here). For HTS specifically, I am talking about the procedure by which sequenced reads are arranged on a reference genome or transcriptome assembly, effectively identifying the region of the genome where the fragment represented by the read originated from (**Figure 3.13**).

Although sequence alignment is not a new issue *per se*, with successful solutions having been in place for years, the sheer volume of data produced by HTS suddenly posed new challenges: Efficiency was now key. The established solutions (e.g. BLAST[4]) were quite simply not fast enough to make it feasible to routinely align millions of read sequences to a mammalian-sized genome. Consider this simple thought experiment: The most straightforward approach to sequence alignment is a simple lookup of the shorter sequence in the longer reference. Since one does not know *a priori* where the sequence might align, one would have to iterate the entire reference stepping through one base-pair at a time and check whether the two sub-sequences match. The haploid human genome, for example, is approximately 3 billion base-pairs in size. Assuming a $50bp$ read length, the exhaustive – that is, looking for *all* possible matches, rather

than just any *one* match – alignment of one sequence would then require some 150 billion comparisons of a pair of nucleotide letters. This amount of calculations can be performed on a modern high-end computer in just under a second and thus the alignment of just 10 million sequences (comparatively little with state-of-the-art HT sequencers) would take several months*. The situation is further complicated by the presence of sequencing and assembly errors and structural variations in genomes (SNPs, insertions, deletions and inversions), which necessitate allowing for mismatches between the two sequence strings. Finally, transcriptomic assays, that is, the sequencing of reverse-transcribed mRNA, can lead to short read sequences spanning the junctions of multiple exons. In order to be able to find a match for such a sequence it would thus be necessary to take known exon junctions into account (a strategy that stands and falls with the quality of the gene annotations) or to automatically discover likely junctions.

A number of more sophisticated algorithms have been proposed to deal with these issues (reviewed in[136,551]). Although the details of different implementations vary, nearly all algorithms work by the principle of first narrowing down the search space by applying heuristic methods and subsequently traversing the possible matches using sensitive, traditional sequence alignment methods. One way to quickly narrow down the search space is the use of a particularly efficient search "index". An index is essentially a structured lookup-table of some sort that makes it possible to quickly find matches to a search query. For short read alignment algorithms, it is possible to distinguish between two main approaches:

- Hash-based algorithms[310,312,320], define a so-called "hash function" which transforms a DNA-sequence into a numeric representation which may then be used to index a lookup-table. Hashes are well-established and popular tools in computer programming and very straightforward to implement. If the hash function is sufficiently simple to calculate, yet avoids conflicts (i.e. multiple DNA-sequences resolving to the same hash code), the method can be very efficient, but memory requirements can get out of hand: For long reads it will not be possible to store all possible matches in the genome in memory, so it is usually necessary to use a seed-based approach, which splits input reads into shorter fragments, which may be aligned independently and combined later on. Nevertheless, the memory requirements of hash-based alignment programs are often not trivial (several tens of gigabytes of memory may be necessary for mammalian-sized genomes). While it, of course, would be conceivable to further reduce the memory requirements by using smaller seeds, this would drastically impair the speed of the programs.

- Aligners based on a Burrows-Wheeler transformation (BWT)[292,308,313], typically use an Ferragina-Manzini-Index (FM index), an index based on a suffix array created from

*N.B. these estimates are deliberately left very vague since the precise measures depend on the implementation, exact hardware specification and load of the computer.

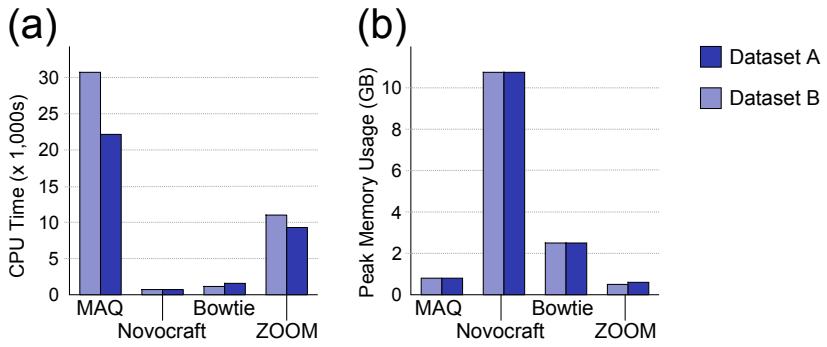

Figure 3.14: Short read aligners. Comparison of speed (**a**) and memory usage (**b**) of four selected short read alignment programs[292,310,320] applied to two test datasets (A = 3,724,383 reads, 21bp length; B = 3,265,654 reads, 21bp length) against the mouse reference genome (NCBIM37 assembly). The comparison was performed using default parameters and the latest version of each software available in December 2008.

BW-transformed input sequences. This particular index structure has proven to be very memory-efficient (typically less than 3GB even for the human genome), while allowing for rapid substring-queries as they are necessary for alignment. Thanks to the reduced memory footprint of the index, implementations of BWT-based alignment algorithms can focus on speed and as a consequence the corresponding programs are now typically orders of magnitude faster than hash-based algorithms.

For our purposes, I felt it was not necessary or even sensible to attempt to rewrite an entirely new solution to the alignment problem, but I rather decided to make use of a proven method from the literature. Based on my own evaluations (for instance, **Figure 3.14**), I chose to use Bowtie[291] for shorter sequences ($< 50bp$) and Tophat[550] (in itself based on Bowtie) for longer sequences and paired-end reads, since these appeared to offer the best trade-off between accuracy (correct alignments), flexibility (useful parametrisation options) and, in particular, speed (number of alignments per second). I have therefore installed both programs on the GeneProf servers and pre-built genome indices for all GeneProf-recommended reference datasets. I then implemented a workflow module that wraps these programs, thus making it possible to execute alignments within any GeneProf workflow.

3.3.3.3 Quantification of Gene Expression

One of the major prospective uses of HTS technology for stem cell biology and biology as a whole, is the accurate profiling and comparison of gene transcription in various cell types or treatment conditions via the sequencing of transcript fragments (RNA-seq or Tag-seq; **Section 1.2.2.2**). Going from raw nucleotide sequences to a measure of gene expression interpretable by domain experts, requires the quantification of the amounts of transcripts stemming from each individual gene. The fundamental idea is rather simple: The more reads one observes from any given gene, the more transcripts there were in the first place and thus

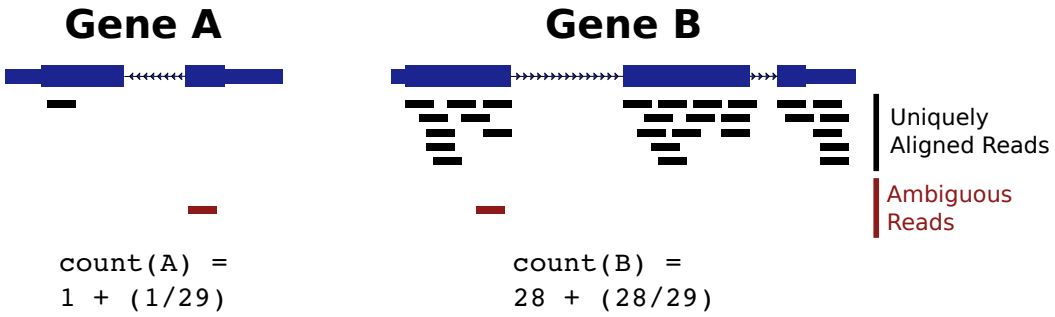

Figure 3.15: RNA-seq gene expression. A measure of gene expression intensity can be calculated by adding up all reads aligning to the exons of a specific gene model. Ambiguously aligned reads may be assigned proportionally.

the stronger a gene is expressed.

After aligning short reads to the genome, one knows from which region of the genome each read originated. In these days, we are fortunate enough to have a good annotation for the human genome and most model organisms, so it is then possible to compare the aligned read positions with the locations of annotated gene models and to sum up the number of reads aligning to the exons of each gene (**Figure 3.15**).

Naturally, the situation is slightly more complicated than that: Firstly, not all short reads from a typical RNA-seq library can be aligned uniquely to one position of the genome – this is due to the repetitive nature of genomes in general, and, in particular, the paralogous duplication of sub-sequences of genetic information for coding genes. In the simplest approach, one could just discard ambiguously aligned reads (often referred to as "multi-reads"), keeping only the most reliable fraction of the data. However, this approach may sacrifice important information, especially when one seeks to study differences between closely related genes or even transcript variants of the same gene. Other approaches try to make use of ambiguous information by either assigning ambiguously aligned reads to one random location, by spreading a fraction of the aligned read to all possible locations or by somehow spreading the read to possible locations proportional to the likelihood of a read originating from each spot[84, 367, 635]. In GeneProf, I decided to adapt a previously proposed approach from the latter category[367] (this is essentially the same strategy I employed earlier: **Section 2.1.2**): In a first round, the unique (that is, unambiguous) read counts for each gene are calculated. I make the assumption that an ambiguously aligned read is more likely to originate from a region belonging to a gene with strong evidence of other transcription and therefore use the unique read count to weigh the proportion of a multi-read that is assigned to each possible location. More precisely, the expression intensity for each gene will be calculated as:

$$count(g) = \sum_{r \in reads(g)} \frac{w(r)}{|align(r)|} \sum_{r \in reads(g)} \frac{w(r)}{\sum_{\hat{g} \in align(r)} \sum_{\hat{r} \in reads(\hat{g})} \frac{w(\hat{r})}{|align(\hat{r})|}}, \qquad (3.1)$$

where $count(g)$ is the expression count for an arbitrary gene g, $reads(g)$ are all reads aligning to gene g, $w(r)$ is the weight of read r (usually 1.0) and $align(r)$ are all possible alignments of read r.

But alignment ambiguity is not the only factor complicating expression quantification: Previous research[367, 552, 635] has highlighted structural attributes of genes confounding absolute gene intensity measures. In particular, it has been observed that longer transcripts tend to be overrepresented in sequencing libraries. It stands to reason that the length of a transcript has a direct and linear impact on the number of fragments sequenced from it and that therefore RNA-seq intensity measures are proportionally higher the longer a transcript. However, this bias can be easily accounted for by scaling intensities by the total length of a transcript[367] and I have adopted this strategy in GeneProf. Wherever absolute measures of expression intensity matter, GeneProf uses intensities expressed as "reads per kilo-base million" (RPKM), that is

$$rpkm(g) = \frac{count(g) * 1,000,000}{R} * \frac{1,000}{length(g)}, \tag{3.2}$$

with R the total number of aligned reads in a library and $length(g)$ the length of a gene in number of base-pairs. If the absolute intensity does not matter, GeneProf uses simpler "reads per million" (RPM) values instead:

$$rpm(g) = \frac{count(g) * 1,000,000}{R}. \tag{3.3}$$

Noteworthily, for matters of comparing gene expression counts the RPM values are usually sufficient, because the length bias (and other biases) will have an equivalent effect on the measured expression intensity in all investigated conditions. Other structural features of genes and their transcripts, such as GC content and other factors affecting the accessibility of the transcripts for random priming, are likely to play a role in the efficiency and uniformity with which transcripts can be detected and will therefore also contribute to the intensities calculated. However, these factors are inherently more difficult to account for and might even differ between specific sequencing platforms. Thus, I believe that effective normalisation methods to resolve these issues deserve further investigation in the future.

Of course, the quality of any intensity values calculated depends on the quality of the gene annotations available. For most model organisms, we now have a reasonably comprehensive and reliable database of protein-coding genes. In GeneProf, I decided to base internal gene models on the annotations from the Ensembl[135] database, which constitutes one of the most up-to-date, high-quality resources available. Despite state-of-the-art manual and automated curation, though, some gene models are still not perfectly well understood, often inaccurate and sometimes incorrect (**Figure 3.16**). Non-canonical units of transcription, such as short transcripts, miRNAs and pseudo-genes, in particular, do still undergo frequent updates. To

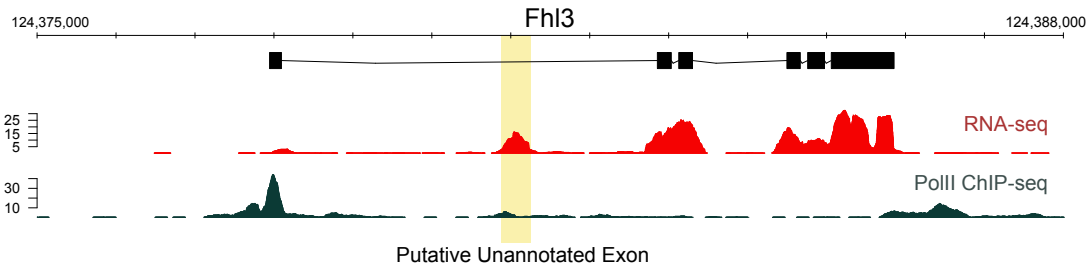

Figure 3.16: Incomplete gene models. HTS data can highlight inaccuracies in current gene annotations. Shown here is the *Fhl3* locus with a track showing the coverage of aligned RNA-seq reads[552] along the gene model. A putative novel, unannotated exon appears to be present falling into the intronic space between annotated exons 1 and 2. The presence of an exon is furthermore supported by weak, yet detectable, Polymerase 2 (*PolII*) binding at the same site (Richard Young lab, *unpublished data*). The data has been re-analysed and this plot created with the GeneProf software discussed in this thesis.

overcome this barrier, numerous approaches have been proposed that either augment existing annotations taking the data at hand into account, e.g. by altering or adding exon annotations[367,453,552], or even construct entire transcriptomes *de-novo*[40,107,455] before proceeding to the calculation of transcript counts. These approaches are particularly attractive to those dealing with non-model and poorly-studied organisms. Thanks to the high quality of the annotations for mouse and human, though, I found it not necessary to include them in the initial release of GeneProf and have focused solely on reference-based quantification of gene expression.

3.3.3.4 Assessment of Differential Gene Expression

Having calculated expression intensities for all genes profiled in an experiment (see previous section, **Section 3.3.3.3**), one might now want to compare several samples looking for genes that exhibit statistically significant differences in expression between different experimental conditions. Differentially expressed genes (DEGs) might play a role in the biological mechanism or function studied and make good candidates for further investigation.

It is tempting to believe that methods developed for microarray data analysis should be appropriate for this task, since the biological question in mind is similar, yet it has been noted that technical differences necessitate specialised statistical methods[459,460]. This is mostly due to the fact that microarrays, which measure the intensity of a fluorescent signal, produce continuous measures of gene expression, while sequencing-based assays produce inherently digital "counts" [†], and the distribution of expression values recorded behaves notably different in both approaches: Microarray intensities suffer from a background noise level leading to no absolute-zero measurements and most values being somewhat centred around a mid-range value; RNA-seq and DeepSAGE measurements, on the other hand, have a wider dynamic

[†]The numbers might not always be integers due to the way they have been normalised or ambiguity has been dealt with, however, this does not alter the fundamental difference in the nature of the signals

range and frequently record zero-values or very low measurements (in agreement with the common believe that most genes will not be active in normal conditions).

Proposed solutions try to tackle the problem by attempting to model the observed distribution in a more accurate way: Initially, a Poisson distribution was deemed appropriate (e.g. DEGSeq[583]), however, it has been noticed that variances are often underestimated and that the assumptions of the Poisson distribution are hence too restrictive[460]. The *edgeR* package[458] and *DESeq*[7] Bioconductor packages therefore use a negative binomial model with moderated gene-wise variances in order to further control the variance estimation (and test outcome). A problem impairing the accurate modelling of the actual distribution is a general lack of sufficient replication, which makes it necessary to estimate parameters for the models from incomplete data. The authors of the edgeR package have therefore decided to assume that there is a correlation between the variance and the mean ($\sigma^2 = \mu + \alpha\mu^2$) so that only one parameter needs to be estimated from the data. In other words, the assumption is that the variance for more highly-expressed genes is stronger than for genes with a lower expression level. The relationship is moderated by a single constant α which can either be assumed to be uniform across the dataset or may be estimated from genes with similar expression levels. *DESeq* takes a similar approach, but extends the model in such a way to allow for more general relationships between variance and mean, which can be calculated from the data. Similarly, *baySeq*[188] also assumes a negative binomial distribution, but additionally derives a prior distribution from the data using an empirical Bayes approach.

The methods highlighted above are by no means the only solutions, but probably represent the most popular tools for the purpose. Comparisons on the basis of simulated data[188] show one method superior in some cases, another in others and in real-world applications it will not usually be possible to choose the optimal approach, because one does not know the desired result beforehand. As a general rule of thumb, almost all methods agree on the most strongly changing genes (high fold-change) if a good amount of replication is given, but results vary more widely if one tries to assess smaller changes or fewer replicates per condition are available. I have found that a combination of either *edgeR* or *DESeq* with a simple fold-change threshold gave good results in terms of selecting genes with a convincingly changing signature. I have integrated both methods into the GeneProf framework by implementing workflow modules wrapping the original program code for the individual tools.

The chosen programs are both limited to pair-wise comparison between conditions and it appears desirable to extend GeneProf's repertoire of algorithms to methods capable of dealing with more complex experimental designs in future. Unfortunately, only the *baySeq* package has so far addressed this question at all and I am still awaiting further developments in the field. All of GeneProf's analysis are strongly gene-focused, yet modern RNA-seq data makes it possible to look more closely at transcription on a global scale and also to distinguish alternative

splicing events and thus the activity of different variants of the same gene. A number of methods for assessing alternative splicing have already been put forward[20, 167, 256, 413, 576] and it seems desirable to integrate those and others into future releases of the GeneProf software.

3.3.3.5 Binding Peak Detection and Peak-to-Gene Association

In the previous sections, I focused on data processing requirements for assays of gene transcription. Another popular and successful application of HTS to date has been chromatin immunoprecipitation followed by sequencing (ChIP-seq) for the study of DNA-protein interactions (**Section 1.2.2.3**). In stem cell biology, much has been learned about the fundamental core transcriptional network and the activity of the key transcriptional regulators by profiling the binding sites of important transcription factors genome-wide[75, 198, 342] and investigating the influence of other epigenetic factors such as histone modifications[342, 361].

In ChIP-seq experiments, the targets of the sequencing process are fragments of DNA which have been selectively enriched for those regions of the genome bound by or associated to a protein of interest (**Figure 3.17**), such as a transcription factor (TF). The DNA sequences strongly associated with the protein of interest will hence be preferentially sequenced. After alignment one can then trace back these sequences to the regions they have originated from. The end result is an enrichment of genomic regions that report putative binding events. When visualised appropriately, the regions in which many reads pile up resemble elevations in a broader binding landscape and are hence often called "peaks".

The identification of these peaks, marking their boundaries and distinguishing them from the background noise and technical artefacts has perhaps been one of the most researched areas in bioinformatics over the recent years[130, 234, 237, 242, 263, 393, 454, 562, 631]. Far from being a trivial problem, peak calling is obscured by weaknesses of the enrichment and sequencing procedures that plague the purity of the ChIP-seq signal. As discussed previously (**Section 1.2.2.3**), the quality of ChIP-seq data is vitally impacted by the quality of the antibody used for ChIP and it is important to acknowledge that even the most reliable antibody will never be able to pull down pure DNA (that is, only those actually bound by the protein of interest). In fact, it has been found that up to a third of commercially available antibodies are not of sufficient quality for large-scale ChIP experiments[415]. As a consequence, the signal is riddled with an omnipresent degree of background noise. Further complicating the situation, the noise level is not constant across the entire genome, but it has been observed that some chromosomal regions are systematically under- or overrepresented in ChIP-seq dataset. These regions might look like real binding peaks, although no specific binding event has happened[470]. There are several reasons for this, for instance, fragmentation of DNA is impaired by the accessibility of the chromatin and sequence composition and ChIP-antibodies might prefer certain fragments over others – either entirely non-specifically or because another, potentially

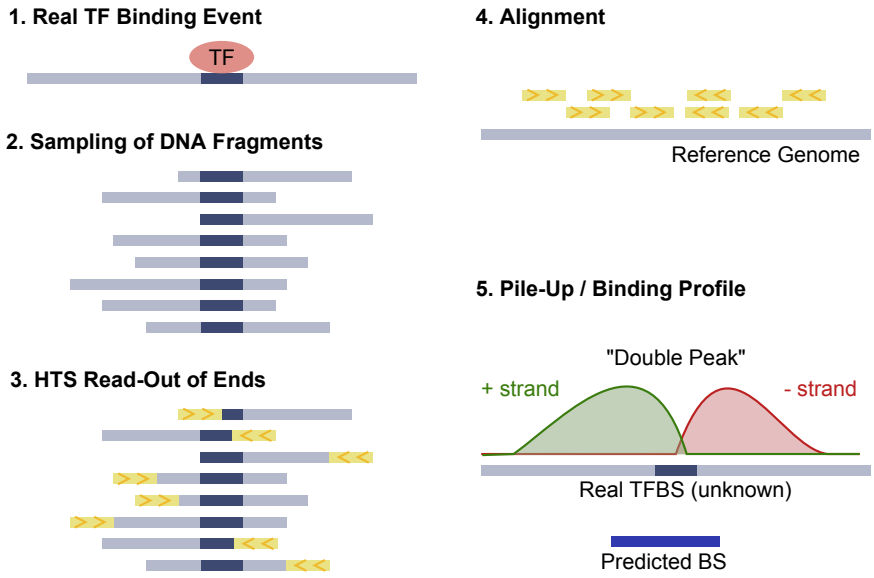

Figure 3.17: ChIP-seq for TFBS discovery. Illustrative summary of the ChIP-seq procedure for the discovery of transcription factor binding sites (TFBS). Random DNA fragments around a true TF-bound region are extracted and parts of the ends of those fragments are read out using HTS, then aligned to a reference genome, piled up into binding profiles and an algorithm is used to detect peaks in the profiles and determine likely boundaries of predicted TFBS.

similar protein is present[27,415,422]. For this reason, it is now common practice to perform an additional control experiment using either input DNA (DNA sheared prior to IP), mock IPs (without any antibody) or non-specific IPs (IP to a protein that does not bind to DNA, e.g. immunoglobulin or GFP). No consensus has yet been reached on which (or whether any) kind of control experiment is superior, but it appears that the use of input DNA is the most popular choice, probably owing to the ease of obtaining enough input material, which can be tricky to achieve when using a mock or non-specific IP that pulls down only very little DNA.

Perhaps even more difficult than the choice of the appropriate control mechanism, is the choice of a good peak calling algorithm. A great many tools have been put forward employing a wide variety of methodologies, ranging from simply imposing a threshold on the minimum height of a peak in the intensity profile, over those looking at fold change enrichments to the background sample, to more sophisticated solution using statistical models, the strandedness of "double peaks" (**Figure 3.17**) and peak shape. A number of attempts have been made to objectively compare different algorithms[288,422,590], but deciding on a universally applicable method of choice appears to be a futile task: Too different are the proteins to be studied and too diverse the experimental conditions and protocols. For example, while TFs would usually be expected to have very narrow peaks corresponding almost directly to the TFBS, histones tend to spread over larger portions of the chromosomes and hence have much broader peaks. But even distinguishing between such broadly different types of application (or protein) is not sufficient to automatically choose the best analysis approach and it might often

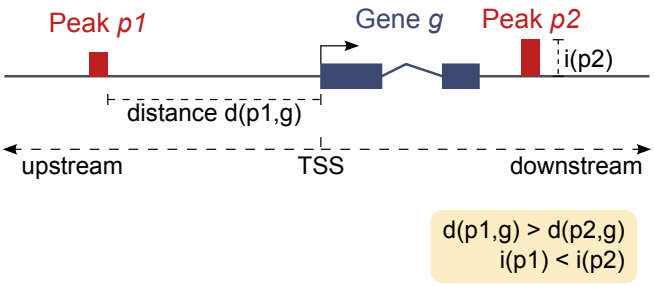

Figure 3.18: Peak-gene association. Illustration of terms and concepts used throughout the text, in particular, when talking about peak-to-gene association. Peaks p have a distance $d(p, g)$ to any given gene g and an intensity $i(p)$, for example, defined by the height.

be necessary to try several algorithms and manually examine a sample of the results to make a qualified expert call on the way forward. I have therefore chosen to integrate a selection of diverse algorithms[75,242,310,606,631] into GeneProf and aimed to support researchers in their exploratory analysis (**Section 3.3.2.3**). By default, GeneProf suggests using the MACS algorithm[631], which has emerged as one of the top choices in most comparisons and offers a flexible range of parametrisations.

The analysis of ChIP-seq data does not usually end with the calling of significant binding peaks: Typical downstream processing involves some form of peak-to-gene association, that is, in order to draw any actual biologically interesting conclusion from the datasets dealt with, it might often be necessary to find a way of telling which genes might be targeted by a TF or which are epigenetically active or repressed (assuming the biological purpose of DNA-binding is the regulation of target genes). It had traditionally been imagined that transcription factors were controlling their target genes directly by binding in their promoter region, but with the advent of genome-wide assays, it has become quickly evident that this is not always or, perhaps, mostly not the case. Much transcription factor binding happens up- or downstream of the promoter region, and it has been shown that, at least in some cases, even binding to distant enhancer elements as far apart as several tens of kilobases can have strong effects on the transcription of their targets[280].

The most straightforward way of associating a putative target gene with a binding factor is to use a windowed approach with a defined, static threshold. Although this would clearly miss many true targets and also include many false ones, in the absence of additional functional data this might often be the only real choice available to many researchers. GeneProf also takes this approach by default and associates binding peaks in a binary fashion (bound or not; true or false) with any gene for which the peak falls into a window of at most $20kb$ upstream or $1kb$ downstream of the transcription start site (TSS) – of course, the thresholds may be configured by the user (**Figure 3.18**). Since the definitions of the boundaries are rather arbitrary, it is important that it is easy for users to re-define and adjust all thresholds for

their own meta-analysis and GeneProf provides the means to do so easily enough, by allowing users to redefine the peak-to-gene association step of all peak datasets in a new meta-analysis experiment.

Current research investigates alternative approaches for linking observed binding activity (from ChIP-seq) to functional targets. In an attempt to avoid the setting of arbitrary thresholds some researchers seek to develop continuous scores of confidence for a functional linkage between a regulator and a target. Following this train of thought, Sharov and colleagues first introduced an *ad hoc* equation (score of potential function, SPF) for ranking putative targets of a peak p of the TF *Pou5f1* as a function of their distance $d(p, g)$ to the TSS of a gene and the height of the binding peak[497]. Interestingly, the researchers decided to factor in the binding intensity of another TF, *Nanog* (as $i_2(p)$), into the same score besides the intensity for *Pou5f1* (as $i_1(p)$), making it plausible to expect high-scoring genes to be tentative targets of both factors. Additionally, it was decided to score CpG-rich regions higher ($X(p) = 1$ if p is CpG-rich and $X(p) = 0$ otherwise), resulting into the following formula (with $\alpha, \beta, \gamma, \delta$ optimisable constants):

$$SPF(p, g) = (i_1(p)^\alpha + (\beta * i_2(p))^\alpha)(\frac{max(d(p, g), 1000)}{10000})^\gamma + \delta * X(p). \tag{3.4}$$

This scoring method worked well for the purposes of the study at hand and enabled the researchers to identify interesting candidate genes for further investigation. However, it is not suitable to serve as a generic function due to its dependency on fixed factors. Later, a simpler and more generic quantitative measure of TF-to-target association, the so-called transcription factor association strength (TFAS), was defined by others[406] as a function of binding intensity decreasing exponentially with distance from the TSS. Interestingly, all putative binding sites (peaks) within a large window ($1mb$) were factored into the formula, delivering one continuous number for each TF-target combination, that is:

$$TFAS(g) = \sum_{pinP} i(p) * e^{-d(g,p)/d_0}, \tag{3.5}$$

where P is the set of all peaks for the given TF, $i(p)$ is the intensity (height) of peak p and d_0 is a constant (usually set to $5,000$). In their paper[406], the investigators showed impressive correlations between the TFAS scores calculated and changes in expression levels. This supports the argument that the TFAS scores are meaningful and therefore also that (a) the distance of a peak to the TSS, (b) the height of a peak and (c) the number of peaks matter in determining functional targets of TFs. The approach has recently been extended in such a way to consider combinations of TFs to account for combinatorial control of expression by multiple factors[73].

Others have furthermore integrated the conservation of binding sites across species and shown association scores taking this information into account[196]. I do not consider this approach any further here, because this information is not currently available on a large scale that would allow it to be used in a generic framework project such as GeneProf. Thanks to its ease of calculation and general applicability, I decided to implement the TFAS[406] method into GeneProf and to offer it as part of the standard analysis pipeline for ChIP-seq data to all users.

3.3.3.6 Data Heterogeneity

In an earlier chapter (**Section 2.1.2.3**) I have demonstrated how data from multiple existing studies can be used to enrich the results of a new experiment. The integration of heterogeneous data from different sources requires the individual data points from all experiments to be mapped to a common reference framework. Arguably, the essential unit of understanding for most experimental, molecular biology and functional genomics is the "gene". Actually, the definitions of what a gene is are rather inexact and I generally prefer to refer to "features" instead, a broader term encompassing protein-coding genes, processed and unprocessed pseudo-genes, all sorts of ncRNAs and other genomic units actively transcribed. GeneProf's reference set of features is based on the Ensembl[135] genome database, one of the most well-maintained, high-quality resources for genomic information available at present and we have seen in the previous sections (**Section 3.3.3.3** and **Section 3.3.3.5**) how GeneProf summarises expression and DNA-binding data on a per-feature level, automatically bringing together information from diverse sources. Once summarised per-feature, the combination of arbitrarily many different datasets is straightforward and, in GeneProf, can be achieved within seconds. For users of the software system this means that they instantly have a wealth of information available at their disposal.

Ready-analysed data (**Chapter 4**) can be rapidly retrieved and compared by searching for individual genes of interest. The system automatically collects all information from published studies stored in its databases relevant to the queried gene and displays them together in one place (**Figure 4.3**). Apart from some generic information about the gene, e.g. gene symbols, accession numbers, genomic coordinates, protein structure and interactions and functional annotation, the feature-centric summary reports include expression data grouped by cell type, tissue of origin or other annotation data, information about the DNA-binding activity of the factor at hand, if it is a transcription factor or other DNA-associating protein. Similarly, information about DNA-binding activity of other factors in the proximity of this feature's TSS is also reported. Users can immediately dive deeper into any piece of information via the dynamic web interface, e.g. by browsing expression data in selected studies in detail or by examining genomic data, such as binding profiles of interesting factors near the studied gene,

using the built-in genome browser (**Section 3.3.2.5**).

3.3.4 Dealing with the Data Overload

Data from experiments using HTS technology is inherently difficult to process due to the sheer volume and size of the data itself. It is now not uncommon for a modern HTS platform to produce some 100 million short read sequences in a single run of the machine (see **Section 1.2**) and often one will be dealing with not just one sequencing library but dozens at once. This amounts to gigabytes of data files which need to be stored and processed necessitating vast disk storage arrays and powerful computing infrastructure. I will elaborate on these issues in the following pages.

3.3.4.1 Data Storage

To illustrate the immense volume of data involved, let us look at a real example: The dataset with the SRA accession number *SRR037952* comprises $18,567,994$ paired-end RNA-seq reads of length $153bp$[552]. That amounts to $2,840,903,082$ nucleotide characters that need to be stored in a file on some storage device. In a standard text file, a single character will occupy at least 1 byte, so this translates directly into some 2.8 gigabytes (GB) for the nucleotide sequences alone. As discussed earlier, HTS data is usually supplemented by an additional quality score character per nucleotide (FASTQ format, see **Section 3.3.3.1**), effectively doubling the required disk space. The files additionally require further formatting characters and identifiers for each read sequence, increasing the total file size to over $6.2GB$. Storing these amounts of data on the large scale requires (i) a large array of secure, but ideally low-cost disk storage and (ii) efficient data compression strategies to make long-term storage feasible and cope with the ever-increasing amounts of data.

Of course, standard file compression algorithms such as ZIP, GZIP or BZIP2 may be used to decrease the size of the data files and they do, indeed, help to drastically decrease the space requirements. For example, applying the GZIP algorithm to the dataset discussed above, reduced the size of the file to $2.0GB$ or about a third of the original size (see appendix, **Section D.3.1**). However, standard compression methods are agnostic of the inherent structure of the sequences at hand and therefore address the compression problem sub-optimally. For example, a standard text-file may use any one of $65,536$ symbols and usually a compression algorithm would have to be able to cope with all of them. DNA, however, is sampled from a much smaller "alphabet", only five characters – each corresponding to a nucleotide – need to be considered (`A`, `T`, `C`, `G` and `N`) and it should be possible to exploit this prior knowledge about the data to achieve an even better compression.

Before we look at how such compression may be achieved, I need to discuss another issue

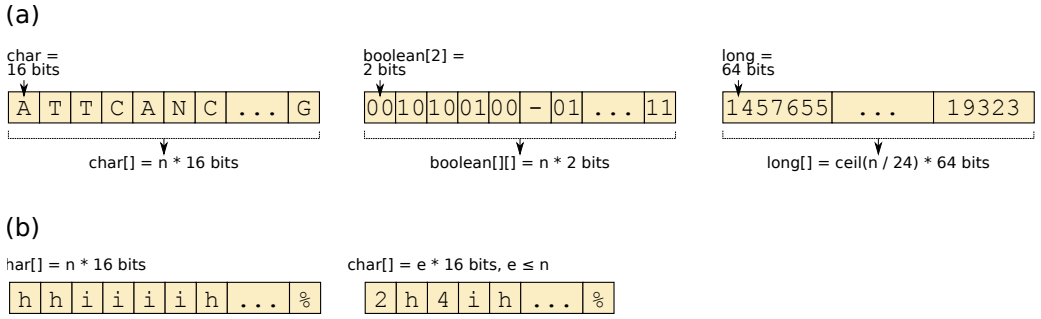

Figure 3.19: Short read compression methods. Illustration of different compression methods considered in GeneProf. **(a)** Space requirements and examples of nucleotide sequences in unencoded (left), 2-bit encoded (middle) and `long`-encoded (right) form. **(b)** An unencoded (left) and encoded base-call quality score string.

which is of paramount importance: Speed. Typically, compression algorithms aim to achieve some trade-off between the compression ratio (reduction in size) and processing speed (both for encoding and decoding compressed data). For example, while BZIP2 usually achieves a stronger compaction of files than GZIP (about 25% vs 33%), the latter may be up to five times faster decompressing files (see appendix, **Section D.3.1**). If the data will be in active use in GeneProf, i.e. it is being repeatedly read from, it is essential that one does not sacrifice too much of the processing performance for the sake of compression. It should, however, be noted that a greater reduction in size might have a secondary, beneficial effect on processing times: Not only does a smaller file size mean less disk read/write access (in exchange for more in-memory data access, which is orders of magnitude faster than disk access), but it will also reduce the amount of data which needs to be transferred between media in a multi-system (many computers), highly-parallel computing environment. All these factors need to be considered when choosing the optimal strategy for a specific application.

Published short read-specific compression schemes[102, 207, 278, 588, 613] attempt to make use of reference-based indices or exploit the redundancy within read libraries to compress entire libraries of short reads. These approaches achieve a great compression, but are rather time-intensive and inflexible. For example, reference-based methods could not compress HTS libraries from organisms for which no genome assembly is available *a priori* and others cannot compress data on the fly since they exploit the characteristics of the entire dataset to achieve the compression. I sought to find a straightforward encoding scheme that could be applied to data on the fly and placed more importance on optimal runtime performance rather than strong compression. I thus explored a number of different strategies for encoding nucleotide sequences (**Figure 3.19**.(a)), all based on the observation that there are only 5 nucleotide characters (including the "uncertain nucleotide" `N`), so standard encodings for characters on computers, which use either 1 byte (plain text files on the hard-disk) or 2 bytes (for in-memory representations in the Java programming language) per character, waste precious space since

they aim to be able to store a much wider range of symbols. I considered encoding each nucleotide character using two `boolean` values, each of which measures only 1 bit. Since two bits can hold four values, an entire nucleotide sequence could then be represented using a two-dimensional array of binary values, where the fifth nucleotide would be stored as an unassigned value. In theory this would give the best conceivable compression ratio (**Figure 3.20**.(a)), however, due to constraints imposed by the way variables are actually addressed on real computer systems[569] and due to the fact that arrays in Java, in addition to the data, store the length of the array, the practical memory consumption of this approach exceeds even the simple character representation (**Figure 3.20**.(b-e)). Instead, I devised a way to store the nucleotide sequences as 32- or 64-bit numbers (Java types `int` or `long`, respectively), which proved to be much more effective. To do so, a number from 1 to 5 was assigned to each nucleotide symbol and the algorithm adds up the nucleotides as a positioned sum (algorithm in appendix, **Section D.3.2**). It is thus possible to store 24 nucleotide symbols in a single 64-bit number (17% of the original size). My calculations and experiments proved that this representation is superior to the other schemes I tried. Unfortunately, the same strategy cannot be used for encoding the quality score characters for the base-calls, because there are too many different values to be encoded. I therefore decided to use a very simple approach here, in which repetitions of the same symbol are compacted into one symbol and a count ("run-length encoding"; **Figure 3.19**.(b)).

The `long`-encoding scheme allows for efficient in-memory representation of sequences. For long-term persistence, the sequences can be serialised in binary form and additionally compressed with a standard algorithm (I chose GZIP, since it offers a reasonable trade-off between file size and fast, decompression time). Sequence data does usually only have to be accessed serially, so no sophisticated indexing or querying methods are required.

Short read sequences, however, are not the only type of high-volume data of concern. Alignment of sequences to a reference genome (**Section 3.3.3.2**) produces a large quantity of genomic data, which may also benefit from special treatment. Unlike sequence data, genomic data does not only require good compression and fast serial access, but might also need to be queried and otherwise randomly accessed. This is because I intended to use this data for genomic data visualizations and also because it might be necessary to retrieve additional information about specific alignments at a later point in time without necessarily accessing all alignments. For example, for the calculation of gene expression counts (**Section 3.3.3.3**), GeneProf needs to find out how many and which reads aligned to the genomic region of a gene. Recently, the genomic data formats BIGWIG and BIGBED were introduced[260] to enable this kind of query on large genomic datasets, however, they are unsuitable for the purposes of the software system at hand, because they are unable to store additional annotation data alongside genomic coordinates. Another widely-used format for alignment data is the sequence

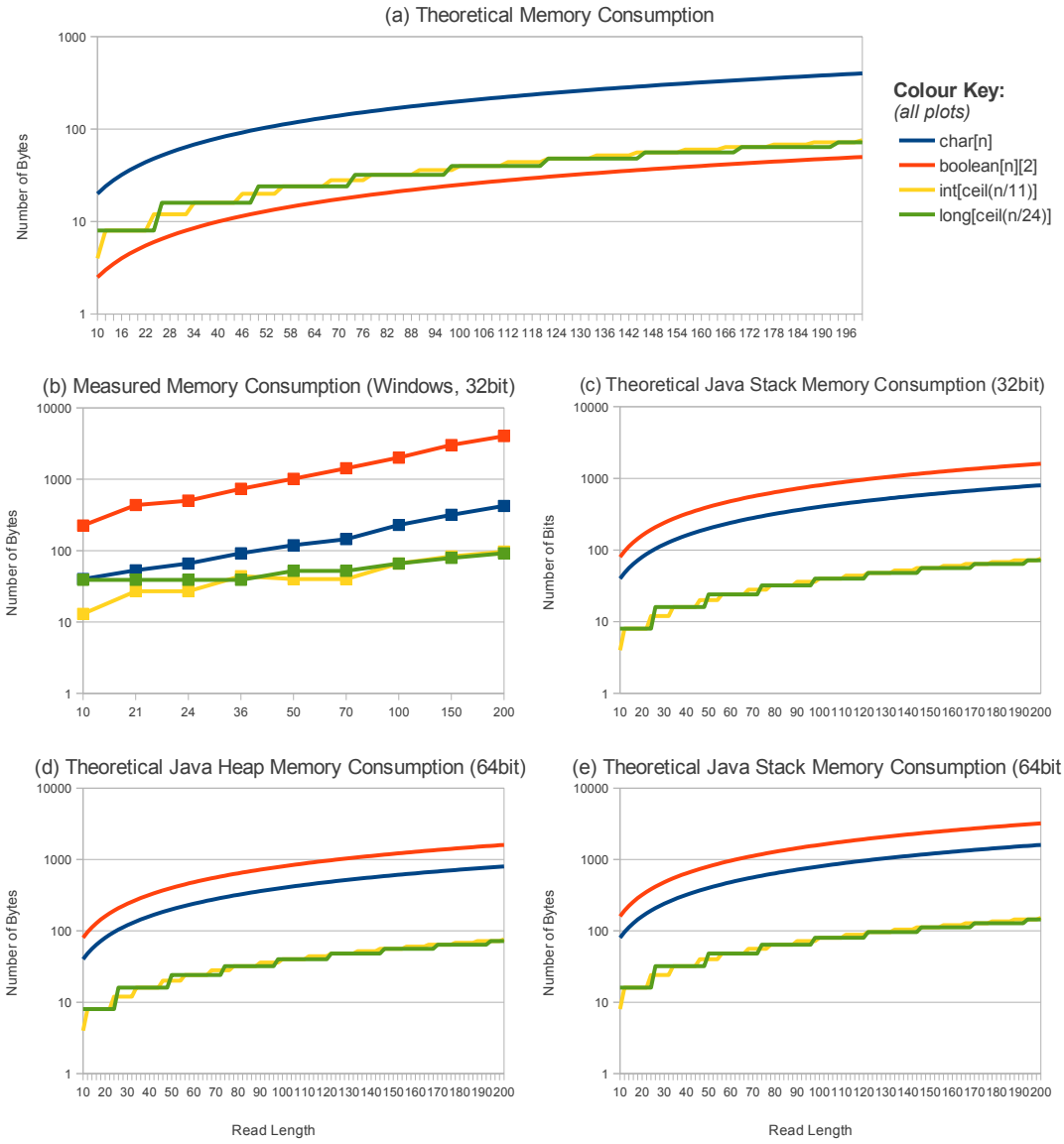

Figure 3.20: Memory consumption of short read data. **(a)** Theoretical memory consumption of the data-holding variables alone for one sequence with the given length (x-axis), see also **Figure 3.19**. **(b)** Actual, empirically measure memory consumption on a Windows Vista 32-bit operating system (averaged over 10,000 trials). **(c-e)** Theoretically expected memory consumption on a 32-bit operating system in the Java stack **(c)**, and on a 64-bit operating system on the heap **(d)** and the stack **(e)** using number from Venstermans *et al*[569].

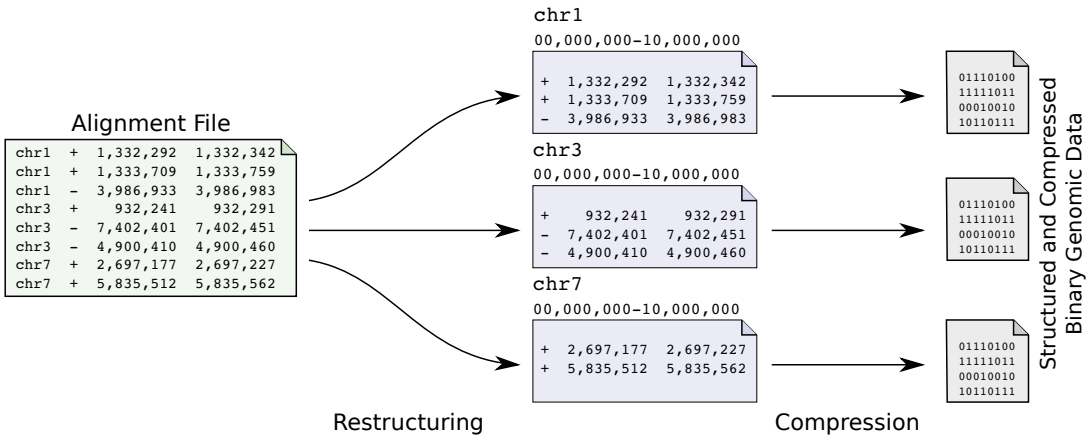

Figure 3.21: Large genomic region data. Illustration of the process for converting alignment data to high-performance, compressed binary data. The data is restructured into a well-defined set of files by the genomic coordinates represented and subsequently binarised and compressed.

alignment/map format (SAM/BAM)[309], which is now supported by most modern alignment programs. The format is more flexible, as it allows arbitrary annotational data to be added to alignments, but is sequence-focused rather than genomic *per se*, making it cumbersome to use for non-serial access. For ease of use and performance, I have therefore decided to use a simple, yet efficient storage format that I called *Large Genomic Region Dataset* (LGRDS). All coordinate data and additional annotations are serialised into a number of binary files that can be efficiently read in the Java programming language (**Figure 3.21**). The files are split by their genomic location, i.e. there would be one file per chromosome and segment, the chromosomes being split into $10mb$ segments. I found that the file structure itself suffices as a simple indexing pattern, enabling quick retrieval of alignment data from parts of the genome by iterating only those entries in a matching subset of the files. The combination of this file structure with a high-speed compression algorithm, Snappy (`http://code.google.com/p/snappy`), allows GeneProf to save the data from genomic alignments at a third of the original size with, indeed, a six-fold higher (serial) access speed than the equivalent uncompressed, text files (*data not shown*).

The remainder of the data concerned in HTS and genomic data analyses, e.g. gene-centric information such as expression data, is mostly rather reasonable in size and does primarily require rapid random access for filtering, sorting and searching. Thanks to the rather manageable file size, I did not consider it necessary to use any special data format for these kinds of information and decided to store them in a relational database system. Relational databases are established tools optimised for quick retrieval or arbitrary pieces of well-structured information and are the *de-facto* standard for storing all sorts of information in enterprise-scale environments.

3.3.4.2 Scalability and Efficiency

The last points that should be addressed when discussing general data processing requirements for a powerful and flexible data analysis suite are scalability and efficiency. "Scalability" is the ability of a software system to cope with increasing amounts of work or data. I use the term "efficiency" to refer to the ability of a system to use resources for maximum effect and, more specifically in computing terms, to handle tasks quickly and within the bounds of the hardware available. Efficiency, thus, is one of the cornerstones of a scalable software system.

It would be a pointless and uninteresting exercise to list all the specific algorithmic implementation details I have taken to ensure the best possible efficiency of the software. Let it suffice, instead, to discuss general concerns and approaches in an abstract manner. Computational efficiency comes in two flavours: Time and memory. Let us address one at a time.

As demonstrated in the context of short read alignment (**Section 3.3.3.2**), the sheer volume of HTS data can often transform tasks, which have rather trivial solutions on a smaller scale into problems difficult to cope with within feasible time limits. This is a well-known issue in computer science and generally referred to as computational complexity: The question is, how does an algorithm scale with growing size of the inputs? Take, for example, the comparison of n nucleotide sequences. If one was to compare each of these sequences with one particular other sequence, a total of n comparisons would be required. The problem is said to scale "linearly" with the size of the input. If, on the other hand, each sequence was compared with each other sequence, the total number of comparisons would grow to n^2 ("quadratic" complexity). This might not be a problem if n was rather small, but what if one wanted to do this with an entire HTS library? Even if a single comparison could be performed in a fraction of a second, say $1ns$ (that is 1 billionth of a second), comparing $1,000,000$ reads would take almost 17 minutes and the comparison of $100,000,000$ reads would take almost four months. While algorithms with non-linear runtime are therefore best avoided, if this is not possible, it might be necessary to come up with heuristic (that is, approximative) ways to solve real-world problems and much bioinformatics work focuses on finding innovative means to tackle these issues.

Algorithmic improvements, however, go only so far and in a real-world application it might be necessary to cut down the runtime of programs further in whatever way possible. In the end, even linear complexity, can be problematic when the input is large enough. Fortunately, many tasks that are difficult to deal with purely due to the size of the data, can be split into units that can be easily parallelised. Modern computers are now usually equipped with multiple processing cores and can hence deal with multiple sets of calculations simultaneously without impairing the performance of work executed in parallel. In the Java programming language, it is reasonably straightforward to implement parts of a program in such a way

that multiple calculations are executed in parallel. This is generally referred to as "multi-threading" – a "thread" being one process executed in parallel with other threads. GeneProf has been designed to make use of this technique wherever possible.

Apart from time-constraints, limited availability of system memory is one of the factors making HTS data analysis difficult to deal with. I have previously shown that the amounts of data to be worked with pose non-trivial challenges both for the long-term storage on disk and, random access memory being substantially more expensive than persistent storage, even more so for in-memory data handling (**Section 3.3.4.1**). GeneProf makes use of memory-efficient data structures, like, for instance, the collection framework provided by the GNU Trove code library (`http://trove.starlight-systems.com`) and, more importantly, attempts to avoid having too much data in memory at any given point in time by accessing data in a serial manner whenever this is possible. This means that, instead of retrieving an entire dataset at once, GeneProf will read only a few records at a time, perform its calculations and then continue iterating over the dataset until no more data is available. I have implemented the data accessors to make this "streamed" mode of handling datasets very straightforward. In consequence, large parts of the GeneProf system are entirely independent of and robust to the size of the input datasets, e.g. all functionality dealing with raw short read sequences will be able to deal with the ever-growing output size of improved HTS platforms.

Despite all measures taken to ensure efficient handling of tasks, in a multi-user environment a system might at times exceed the capacity of its resources. Therefore, a truly scalable system needs to be able to balance its workload carefully. In computing, one often refers to one unit of processing, e.g. one data analysis process, as a "job". Conversely, the process of distributing jobs over available resources is called "job scheduling". Jobs that can not immediately be allocated to a specific compute resource will typically remain in a queue while waiting for other processes to finish. A number of business-scale job scheduling frameworks exist, many of which have been developed for large, high-performance compute grids that semi-automatically split up extremely large processes into more manageable units. Perhaps the most successful framework is the Sun Grid Engine (`http://wikis.sun.com/display/GridEngine/Home`, now Oracle Grid Engine) and its numerous open-source derivatives. The functionality of these systems by far exceeds the requirements posed by GeneProf and the comprehensiveness comes at the cost of a difficult setup and high maintenance effort. I therefore decided to implement a simple job scheduling framework specifically tailored to the needs of GeneProf. The requirements to be addressed were:

- Analysis workflows are to be submitted via the web interface as jobs to the scheduler.
- Each job is to be allocated to one processing node – that is, one computer in a cluster.
- New nodes must be easy to add or remove from the cluster, so that computers may be

used for other purposes, if required.

- If a computer can not process a workflow component due to missing software, another node should take over the job.
- Updates of the GeneProf software should be easily, ideally automatically, distributed to all computers.
- Each node should process more than one workflow at a time, if it has sufficient resources.

The requirements were rather basic and could be easily addressed by queue-based "pull"-strategy: A database table is used to store the current status of all open, executing or completed jobs. When a user decides to submit a new workflow for processing, the web application server updates the database information to mark the job as pending. Each processing node runs a script (which I call the "job agency") that constantly monitors the database table. When a new job is submitted, one job agency will claim it by marking the corresponding database entry as executing (the job is "pulled" from the queue, hence the name of strategy), no other node will then pick up the same job (**Figure 3.22**). The executing job agency spawns a new separate program (a "worker") that then executes all the computer code necessary for the processing of the analysis workflow. Should, at any point, a worker not be able to deal with a sub-process of the workflow, e.g. because an external software is not installed on that specific computer, the worker "surrenders" on the job and marks it as pending again, so another job agency can allocate a worker for it. Each job agency has a limited number of workers available (as per computer-specific configuration) and may hence deal with several jobs at once. If no more workers are available, the job agency will seize claiming additional jobs. After an analysis workflow has been processed completely, the job agency marks the job as finished and frees the worker for other jobs. In addition to the job scheduling, each job agency monitors another database table that contains information about the setup of the web application server it is connected to. These information include the GeneProf software version currently in use. If a new version is to be deployed and the web application is restarted, the job agency will detect the version difference and stop looking for new jobs. After all currently executing analysis processes have been completed, the agency will shut-down, retrieve the latest software version from the server and restart automatically.

This simple architecture is absolutely sufficient for the successful operation of a job scheduling system for GeneProf. All processing nodes are completely agnostic and independent of each other and there is no central hub controlling them. New nodes can be dynamically wired into the system with no more effort than starting the job agency script. If a node needs to be taken out of the cluster for maintenance or other purposes, it can be shut-down via the web interface.

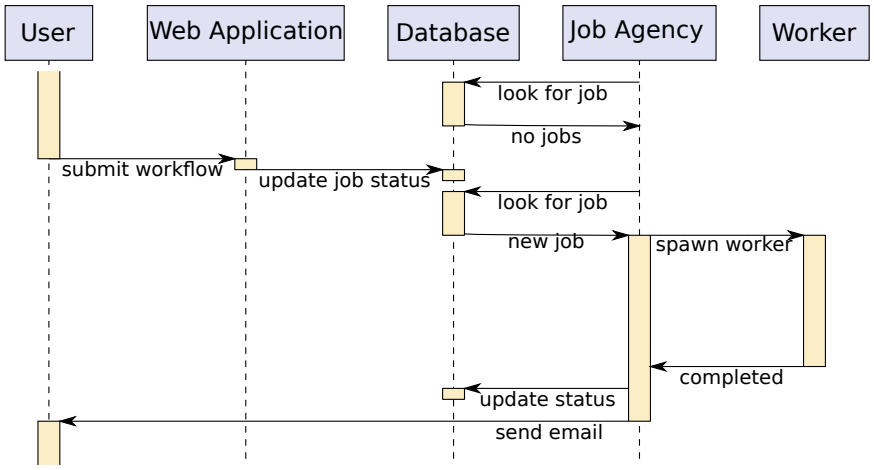

Figure 3.22: Job scheduling example. Illustrative example of the job scheduling process. After a user submits a workflow for processing, the web application updates the corresponding record in the database queue. A job agency will then discover and claim this job and spawn a worker to handle it. After the worker finishes, the agency updates the queue again and notifies the user, who may then continue working on this project.

3.4 Evaluation

Having presented my reasoning behind the development of the GeneProf software, I shall now evaluate the results of my efforts. I will first compare GeneProf with relevant other software packages, then give a short report about the usage of the public instance of the system during the first weeks of its public release and then conclude with some directions for future work and improvements.

3.4.1 Comparison with Existing Data Analysis Software

GeneProf is by no means the first software application for biological, large-scale data analysis; with the rise of microarray technology, the bioinformatics community has developed numerous tools to tackle the high-throughput data at hand. Many of these tools have been integrated as add-on packages into the *R* framework for statistical computing[151] , but a good number of graphical solutions have also been put forward, e.g. the Multi-Experiment Viewer (MeV[474]) or GenePattern[442]. Of course, numerous commercial products also seek to claim their share of the market. With HTS technology now becoming ubiquitous, the developers of many existing software solutions have attempted adapting their tools for the new data types, but they struggle with the demanding data processing requirements (**Section 3.3.3**) and often focus solely on assays of transcriptomic data (traditionally, the stronghold of microarrays) and can therefore not be considered comprehensive enough.

For these reasons, we could recently witness the development of a great number of novel tools specifically addressing the users of HTS platforms. The first software releases typically targeted specific application areas of the technology, e.g. Myrna[291], DSAP[210], miR-

NAkey[464], SeqBuster[414], RSEQTools[180], GENE-counter[99] and ArrayExpressHTS[161] for transcriptomic data (RNA-seq or shortRNA-seq) or W-ChIPeaks[290], CisGenome[234], ChIPseeqer[152] and CASSys[3] for downstream analysis of ChIP-seq data or others for metagenomic[62] analysis. Over time an increasing number of generic framework solutions addressed the HTS field either by providing command-line scripting environments, e.g. GATK[351], HTSeq (Simon Anders, unpublished) or components of Bioconductor[151], or via graphical interfaces, e.g. Taverna[213, 398], Galaxy[43, 153, 160] or KNIME[231, 323]. I present an overview comparison of the latter category of tools and GeneProf in **Table 3.2** with the criteria chosen and designated as follows:

Software has been compared in terms of their analysis capabilities for transcriptomic and regulatory next-generation sequencing data and their general usability. I have only included workflow-enabled software that is free for academic use and that I thought was addressing these issues. Some software might have additional features, which have not been considered for the sake of this comparison. I have made every effort to be objective, but unfortunately comparisons of this type are inherently biased and I acknowledge that this table might be subject to differences in opinion. Some software is constantly being updated and extended, so the list of supported features might have changed since I composed this comparison in June 2011 (extended to include KNIME in December 2011).

General properties: As a first point for comparison we shall concern ourselves with the overall interface of the software, dependencies on other software and the ease of setup. All these factors contribute strongly to the ease of use and therefore on how likely a system is to be adopted by the research community (see also **Section 3.3.2.1**). I distinguish two primary types of user interfaces: The first, command-line based scripting environments, are traditionally only appreciated by expert computer personnel, while the majority of scientific users would usually prefer a graphical interface design, which may be either in the form of a stand-alone desktop application or web-based, that is, accessible via the world-wide web as a web page. Web-based software has the advantage of not depending on any unusual software to be installed and does not require any installation themselves. Stand-alone software, on the other hand, frequently depends on other external programs, which can be very difficult and time-consuming for people to set up and manage, especially if no graphical, assistive installer is provided. I consider Galaxy and GeneProf to stand out by these criteria thanks to their independence of installation and use-immediately kind of nature, closely followed by the two graphical tools, KNIME and Taverna, which can be easily and quickly set up using install wizards and both provide user-friendly interfaces. I would anticipate that many users might struggle installing the other software, that require the compilation of operating system-dependent code and dependencies requiring a level of IT-expertise that cannot usually be expected of lay users.

Core functionality: Evidently, the usefulness of any data analysis software in the end

	GATK[351]	HTSeq	R/Bioconductor[151]	KNIME[231,323]	Taverna[213,398]	Galaxy[43,153,160]	GeneProf[182]
				General Properties			
Interface				++	++	++	++
Dependencies	+	+	+	++	++	++	++
Installation				+	+	++	++
				Core Functionality			
Quality Control	++	++	++	++		++	++
Alignment	+		+			++	++
RNA-seq	+	++	++	++		+	++
ChIP-seq			++	++		++	++
Downstream Analysis	++	++	++	+	++	++	++
Organism Support	++	++	++	++	++	++	++
				Workflow Design			
Design Methodology	+	+	+	++	++	++	++
Assisted Workflows					+		++
Exploratory Analysis	+	+	+	+			++
				Presentation of Results			
Interactive Results			+	+		+	++
Graphs & Plots	+	+	++	+		+	++
Genome Browser	+	+	+	+		+	++
				Data Providence & Integration			
Integration of Public Data			+			+	++
Gene-centric Summaries							++
Meta-Analysis	+	+	+			+	++
Linkable Workflows & Results				+	+	++	++
Transparent Analysis				+	+	+	++
Secure Data Sharing						++	++

Table 3.2: Comparison of assorted HTS analysis software. Software tools are rated on a scale from "missing / unsatisfactory" (empty) through "incomplete / insufficient" (one plus symbol: +) to "good / advanced" (two pluses: ++). See text for further details.

boils down to the core functionality supported. The best interface design and periphery does not do much good if no useful data analysis can be carried out with the system. For the types of applications I addressed and that I know are important at least for researchers in stem cell biology (**Section 3.3.3**), the software needs to support quality control, short read alignment, gene expression quantification and differential expression analysis (RNA-seq), peak finding and feature association (ChIP-seq) and, ideally, further functional downstream analysis. It is certainly preferable if all analysis steps can be performed from the same environment and do not require the manual execution of external programs. Importantly, all the software systems presented are to some degree or another flexible environments that could, in theory, be extended pretty much to any field of application desired. In practice, though, it is mostly not feasible to write additional code or wait for the implementation of new features, so I carry out this comparison on the basis of whatever version was publicly available at the time of assessment. While all systems examined support a reasonable degree of quality control measures, only Galaxy and GeneProf supported direct alignment of raw read data within the main system. Others either require execution of externally installed alignment software or have no documented support for alignment at all. This is probably due to the fact that alignment is a computationally demanding task (**Section 3.3.3.2**) and not feasible to support on a standard desktop computer. For the web-based systems, i.e. Galaxy and GeneProf, this is not a fundamental problem since all analyses are being executed remotely on high-performance compute nodes. HTseq and GATK currently focus on transcriptomic applications and Taverna's HTS-specific functionality was, at the time of comparison, not yet available. Galaxy did support expression quantification, but lacked support for normalisation and differential expression analysis. Overall, KNIME, Galaxy and GeneProf all offered a good and comparable range of functionality that should be sufficient for the majority of users. R/Bioconductor probably provides the most flexible and versatile framework, but requires expert skills to install, manage and use its full functionality.

Workflow design: In terms of workflow design, the main distinction, again, is between the graphical solutions and command-line frameworks: In both cases, workflows are made up of small programs, each of which is responsible for a particular sub-task. Command-line frameworks chain together these tools using custom computer scripts, which requires a advanced understanding of programming techniques in order to use them efficiently. All graphical suites evaluated, on the other hand, make use of a visual programming paradigm allowing users to combine different programs, represented by boxes, in a graphical manner using drag and drop of arrow connectors. In GeneProf, the individual tools are called "modules" (**Section 3.3.1**) and might combine several independent programs into one logical unit, making workflow creation even easier to understand.

I have learned from experience that novice users find it a bit difficult to draw up complex

workflows from scratch, especially in the beginning. GeneProf is the only software that actively assists users in the creation of common workflows by supplying a range of wizards for popular types of analysis (**Section 3.3.2.2**). This simplifies the entry to the program for novices users and allows them to learn over time from the automatically created workflows and apply the knowledge gained to more specialised workflows in future. The only other tool providing similar functionality is Taverna. In fact, Taverna does have a number of wizards that are provided via "portals" (websites that use Taverna at the back-end). However, there are currently no usable wizards for HTS analysis. A major drawback of Taverna's portal-based wizards is that the workflows created cannot be modified subsequently, which severely limits the flexibility of the system. GeneProf's wizards set up complex workflows within seconds, but impose no limits on later adjustments of the processing steps.

Another point discussed earlier is concerned with the support for exploratory data analysis (**Section 3.3.2.3**): Often, it is not possible to know beforehand exactly which programs (and parameters) will be best suited for a particular dataset at hand. It is therefore beneficial to have an easy means to adapt certain steps of the analysis without losing track of what one has done before and (ideally) without having to run all (time-consuming) processes again. The concept is well-supported in GeneProf, but less so in other tools. Script-based workflows can be adjusted easily given enough experience with programming and can be designed in such a way that not the entire analysis needs to be re-run, however, it quickly becomes difficult to keep track of different versions of the scripts (and the associated data and results). Workflows in Galaxy and Taverna can be adjusted easily enough, but they are distinct from the data dealt with (they themselves are tools that are applied to data), which means that in order to change only one parameter and examine the impact on the outcome of the analysis, the entire analysis needs to be repeated, which is a time-consuming process. Additionally, outdated analysis results accumulate and need to be manually removed otherwise one runs the risk of losing the overview over all results.

Presentation of results: Next, the way in which results are presented to the users will have a major influence on how useful the analysis actually is for biological research. All command-line programs as well as Taverna and, in most parts, KNIME and Galaxy produce static text files or custom file-format outputs. These files are not always immediately useful and might first need to be converted to other formats or opened in other programs so that researchers can examine the outputs. A few recent additions to the tool sets of KNIME and Galaxy introduce hyper-linked pages to the output results that start to address this issue making it easier to browse and examine datasets. GeneProf's output data is presented in the form of dynamic tables that can be browsed, searched, filtered and sorted instantly.

Most tools can produce a limited set of plots, which can, however, hardly be customised. The user is limited to whatever plots the software designers implemented and has no way

of changing them. The exceptions to this limitation are R/Bioconductor and GeneProf. R offers an impressive range of plotting capabilities and can create virtually any graphics conceivable, many types of plots were even specially designed for biological research use and are well-established and -understood in the community. GeneProf benefits from R's plotting functionality and provides it to users via an easily configurable, graphical interface. In addition to the standard customisation features provided via the interface, all of GeneProf's plots can alternatively be saved as a set of R-scripts (with supplementary data) so they can be adjusted further, for example, to use specific colour combinations, change labels and so on.

One of the most powerful ways of visualising genomic data is via the use of genome browsers (**Section 3.3.2.5**) and a number of great, user-friendly and quick solutions exist[135, 259, 457]. All tools other than Taverna provide means to export genomic data in formats compatible with the standard genome browsers and to modify existing tracks. Galaxy even has a simple built-in browser, however, this browser does not support plots summarising the coverage of alignments as densities (known as "wiggle-plots" or "wig-plots"), which is one of the most useful types of visualisation and thus cannot be considered sufficient. GeneProf also features a simple built-in genome browser, providing all essential functionality necessary to allow users to very quickly get a feel for their data. To support advanced genome browsing in external, fully-featured applications, GeneProf can also export all genomic data generated in a variety of popular data formats.

Data providence and integration: Lastly, I want to look at the topic of data transparency and providence (**Section 3.3.2.4**). For scientific data and the results of analyses to be really useful to the maximum possible extent, there needs to be a way to reuse the results and data from previous analyses. For stand-alone programs it is difficult to import public data since they would inherently depend on an external database or warehouse to store this data. There is some functionality in R/Bioconductor that facilitates import from public repositories, but this only concerns raw data. Conversely, Taverna and KNIME allow the reuse of analysis workflows made public via myExperiment[159], but do not store the data alongside. Galaxy offers the facilities to make both data and analysis public, and even provides means to describe both together in customisable summary pages (Galaxy Pages[160]), however, this process is time-consuming and has been used only very rarely. Making data publicly available is a matter of a few clicks in GeneProf. Public data then becomes immediately available for import into new projects. Meta-analysis of potentially large collections of diverse datasets is rapid and straightforward.

In addition to this, and unlike any other tool, GeneProf is backed by a large database of ready-analysed results and makes these available via gene-centric summary reports (**Figure 4.3**), which allows experimental biologists to quickly benefit from the use of the software and the insights gained from high-throughput functional genomic experiments without even

any need for their own HTS data.

Importantly, even when analysis workflows can be shared (as in KNIME, Taverna or Galaxy), this does not necessarily warrant transparency and reproducibility. Reproducibility is jeopardised as soon as data from external resources is used, but not integrated in the workflow, because it cannot be guaranteed that the data will still be available at a later point and that it will not change. Since only GeneProf directly integrates the data with the workflow all other programs run risk of inconsistencies. This is even more so true for the stand-alone programs that do not have an associated database for storing results: Scripts might be made available with publications or, at least, upon request, but experimental data needs to be uploaded to an external database – and this will only include the primary experimental data, e.g. short read data, but not include supplementary data, e.g. from genome assemblies or gene annotations, which are inherently prone to change frequently or become unavailable. In GeneProf all data used in the analysis is stored inside the internal databases and is frozen at the point of analysis, avoiding loss of primary and auxiliary data in future.

3.4.2 Higher-Order Analysis Systems and Long-Term Maintenance

The software packages discussed in the previous section are solutions for data analysis challenges. GeneProf also addresses these issues and I have compared the functionality of all systems on the grounds of how well they perform. However, GeneProf goes beyond this level: It has always been my aim to provide a platform for scientists that would enable them to expand their knowledge and gain new insights into biology, by making it possible for them to exploit state-of-the-art large-scale data resources that would otherwise be beyond the reach of most researchers. The analysis component of GeneProf is an essential necessity to establish the data at the heart of this platform, however, in a way, this component is peripheral to the higher-order functions of the system: In fact, it is not inconceivable that parts of or even the entire analysis framework could be replaced, if that was to help the development and maintenance of the system.

An example of a system that takes such an approach is the Stem Cell Discovery Engine (SCDE)[201]: The SCDE is a database of reanalysed experiments from the field of stem cell research that have been brought together under one roof. Much care has been taken to annotate the data in the system appropriately, so to make it possible to resuse the data and to compare datasets within the system. The SCDE utilises Galaxy[43,153,160] as an underlying analysis engine to provide users with the facilities to carry out advanced analyses in the system. However, unlike in GeneProf, the processing steps leading from the raw data to interpretable results are not directly part of SCDE and not carried out using the Galaxy-powered analysis system. The analysis component focuses instead only on the downstream comparison of pre-

analysed results, for instance, the intersection of target gene lists.

The SCDE is a good example of how an underlying analysis system can be leveraged in conjunction with a comprehensive and useful database to create additional functionality. Amongst the tools presented in the previous section (**Section 3.4.1**), Galaxy stands out as the most powerful and flexible framework available (apart from GeneProf). The software enjoys immense popularity, in particular, with software and algorithm developers who appreciate the ease with which they can integrate their own tools into the Galaxy framework (cp. **Figure 3.1**). As a result, Galaxy now has a large and active community building up comprehensive set of flexible data analysis tools. It is for this reason, that I see it as a desirable future development to integrate Galaxy into the GeneProf analysis framework. This could be achieved in two ways: Either the Galaxy workflow engine could replace the existing GeneProf framework and GeneProf modules could be rewritten to be compatible with Galaxy. Alternatively, individual Galaxy tools could be wrapped in GeneProf modules to make it possible to run them from within the existing framework. The GeneProf development team is currently investigating both options. In either case, this integration would happen in the background in such a way that the users of GeneProf would hardly notice any difference.

A potential integration with Galaxy opens an interesting avenue for a simplified long-term maintenance of GeneProf, because it would effectively allow GeneProf development to focus entirely on maintaining and expanding the higher-order components of the system: In this model, state-of-the-art analysis tools are contributed by the community to Galaxy and thus indirectly to GeneProf. The GeneProf team, on the other hand, wires these tools into GeneProf workflows and analysis wizards. This would remove a substantial part of the maintenance burden, leaving only issues related to the continuation of GeneProf data itself: I have previously discussed the importance of providing transparent and reproducible research data and results (**Section 3.3.2.4**) and therefore committed GeneProf experiments to maintaining the analysis and results in exactly the state they were when they were first generated. So long as it is feasible, GeneProf will therefore keep public experiments unaltered. However, this creates problems in terms of the comparability of older experiments with the results of new ones, in particular, if the reference genome annotations might have been updated since. In order to resolve this issue, I intend to implement a revision system into GeneProf whereby the system can maintain two parallel versions of each GeneProf experiment: The version as it was at the time of publication and an automatically updated version using the latest reference data. In addition to this, the creators of an experiment will be able to create derivative, manual revisions of experiments, so they can utilise the latest methodologies to extract additional findings from previously published data.

Lastly, it should be mentioned that the "maintenance" of GeneProf as a useful resource depends not only on keeping whatever data is already in the system, but also to constantly

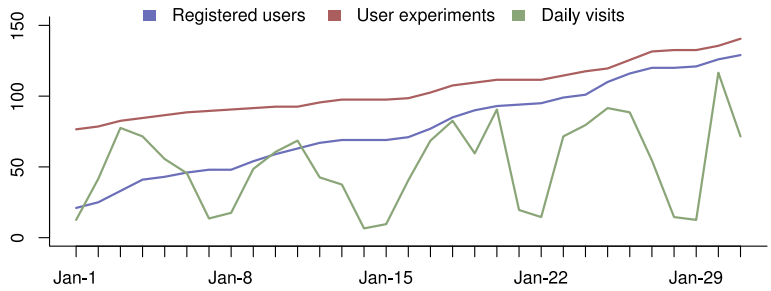

Figure 3.23: GeneProf usage. Anonymous usage records for the month of January, 2012, obtained from Google Analytics and internal measures. The number of user-created experiments excludes those set up by the GeneProf team (F. Halbritter, H.J. Vaidya and S.R. Tomlinson). Labels on the x-axis correspond to Sundays.

expand the repertoire with the latest research data. I myself and other members of the Tomlinson group will keep on analysing the data from the latest publications and add those to the GeneProf databases and as more users start using the software we hope they, too, will contribute to the database by making their published data and analyses public.

3.4.3 Usage Report

GeneProf has officially been launched with the publication of the paper presenting the software[182] in the beginning of January, 2012, but had previously been used extensively by a selected circle of testers. Looking back at the first month of usage (**Figure 3.23**), I can report a constantly increasing amount of interest in the software. With an initial usage peak coinciding with the online publication (December 28, 2011) and the release of the hard-copy of the January issue of Nature Methods (January 3, 2012), the daily number of visitors has further increased and is now beginning to stabilise at about 70 on peak days (Monday to Wednesday).

The majority of visitors come from the United Kingdom as well the United States ($n = 576$ and $n = 491$, respectively; **Figure 3.24**). Most users choose to browse the GeneProf website as a database looking at gene-specific information or public experiments from their field of interest (source: Google Analytics anonymous usage statistics). A sizeable fraction (7.5%) of visitors has further registered for an user account and started creating their own experiments (**Figure 3.23**). It may be expected that the active use of the software will increase in future, when previous visitors, now familiar with the software, generate new HTS data that can be analysed within the system.

GeneProf is currently actively being used for a number of ongoing cross-site collaborations. Our collaborators appreciate, in particular, that GeneProf allows them to browse through analysis results themselves in a way that enables them to closely examine the findings. Moreover, the software will be used as the basis of a future grant application and contributes a substantial part to another that is currently in its final stages of review.

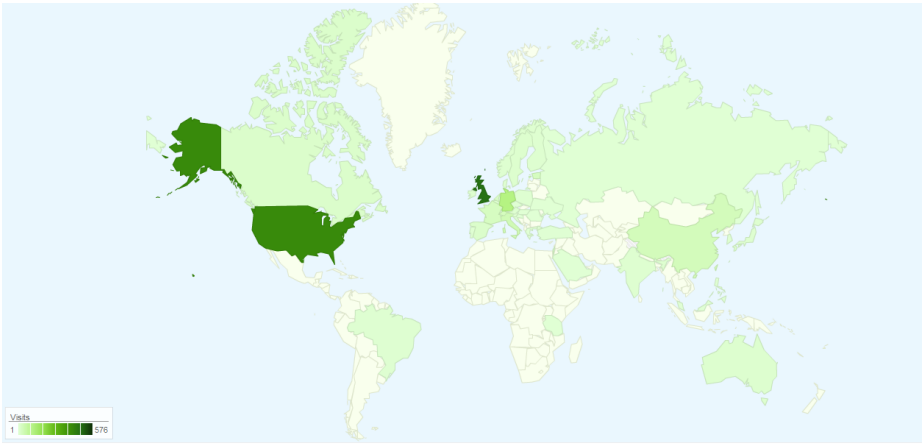

Figure 3.24: Visitor map. Geographic region of origin for all visits to the GeneProf website recorded in January, 2012. Colour intensity is proportional to the number of visitors. Source: Google Analytics.

3.4.4 Future Improvements

The GeneProf software, as it stands today, is a flexible tool and resource for biological research. Yet, I realise that there is a lot of potential for growth and expansion and many areas of the system merit improvement. I shall list here only a few assorted directions for future improvements (which come in addition to those mentioned in **Section 3.4.2**):

- The **Visual Data Explorer** (VDE; **Section 3.3.2.5**) is already a powerful tool for near-instant meta-analysis of large collections of datasets. Not only will the VDE benefit from increasing amounts of public data, but I also plan to add additional plotting methods and to simplify the way in which relevant data is going to be selected. Similar to the data analysis wizards, the VDE will in future suggest popular plots types and guide the user through the customisation steps including data selection. In this way, the user will be able to create plots of correlation matrices or principle component analyses (PCA) between expression values or binding profiles and histograms of those values, but there will also be additional types of visualisations depicting the scatter (and relationships) between properties like expression and binding of different factors, heatmaps augmented with additional annotations (presence of binding sites, function annotation, ..) and many more.

- The addition of personalisation features like **gene lists and favourite regions** will simplify repetitive tasks and make the use of the web interface more convenient and pleasant. Users will be able to save lists of genes of relevance to their research or that have been identified in their earlier analyses as potential candidates for a process or condition and use these lists to quickly filter datasets throughout the interface or to highlight corresponding data points in plots. Similarly, users can store "favourite

genomic regions" which can then be used for rapid navigation in the genome browser or to identify particularly relevant genomic events, e.g. the binding of transcription factors in an enhancer region under study.

- I will try to expand GeneProf into a **collaborative platform**: Already, users can share unpublished analysis and results with collaborators world-wide, but it is not yet possible to collaboratively work on and modify an ongoing analysis. In future, I will investigate ways to ensure data consistency live between multiple user sessions in parallel, which is a necessity for collaborative editing of workflows. Improvements to the interface representation will make it possible to communicate between sites and to see changes made by others in quasi-real time.

- In the near future, GeneProf will be expanded towards the field of **proteomics**. Large-scale, quantitative proteomics assays are becoming increasingly affordable and popular, yet to date there is no user-friendly, integrated software solution available to unify data processing steps and streamline data analysis. From a computational biologist's point of view, however, the analysis requirements are similar to the ones dealt with already for the purpose of HTS analysis in GeneProf and it is a straightforward exercise to extend the functionality of GeneProf to utilise proteomics-specific algorithms and software for advanced workflow-based data analysis. The benefits of integrating this kind of data to the system are immense and are promising to further expand our understanding of biological functions in stem, progenitor and mature cell populations. This part of the project will in future be addressed primarily by Duncan Godwin under supervision of Simon R. Tomlinson and in collaboration with myself.

- Finally, **additional modules and wizards** will be developed to extend GeneProf's data analysis functionality. Specifically, I want to address the issues of sequence motif discovery, transcriptome assembly and the analysis of histone states and methylation by either identifying suitable existing software and wiring it into the GeneProf workflow environment or by developing custom algorithms for the purpose. Moreover, I intend to add support for the processing of microarray data and for the integration of these data with the other data already in the system. GeneProf will then be able to benefit from the wealth of data that has previously been generated, substantially expanding the value of the GeneProf databases.

It should also be noted that I recognise the importance of a vibrant and active research community and do hope that the GeneProf user base will actively contribute ideas and suggestions to the future development of the application and to support community input I have implemented a feature request component (**Section 3.3.2.1**) directly into the web interface, so

that users can share and discuss their thoughts. Furthermore, I plan to improve the advanced programming interface (API) for module development and web access (WebAPI) and expect that bioinformaticians and computer programmers will start to develop additional functionality independently, which will eventually contribute to the repertoire of tools available in GeneProf workflows.

Chapter 4

Creation of a Comprehensive, Integrated Resource of High-Throughput Experiments

In this chapter I shall describe the creation and population of a comprehensive, integrated database resource of readily attainable and interpretable findings derived from a large-scale re-analysis of published HTS data. I will start by outlining the motivation behind this part of the project (**Section 4.1**) and explain the methodology for acquiring and analysing the data (**Section 4.2**). To conclude, I will then give a summary report of the data in the system (**Section 4.3**).

4.1 Motivation and Goals

The amounts of data produced by modern HTS technologies are unparalleled in the history of biology. Over the past years, the member projects of the International Nucleotide Sequencing Database Collaboration, namely the Sequence Read Archive (SRA; National Center for Biotechnology Information, USA), the European Nucleotide Archive (ENA; European Bioinformatics Institute, UK) and the DNA Data Bank of Japan (DDBJ; National Institute of Genetics, Japan).[275,304–306,503], have established electronic archives around the globe that have now accumulated hundreds of terabytes of data and attract more at an ever-increasing rate. News of an impeding shut-down of the SRA due to a lack of funding for the high maintenance costs shocked the genomics community in early 2011. Only later in the year the decision had been revoked – to a certain degree: It had been decided that the storage of certain types of data was no longer cost-effective. The data affected was the output of large-scale re-sequencing

projects; functional genomics data remained untouched.

But why is it that this data was deemed not worthy of being kept? The huge volume of data creates not only storage issues, but also concerns the processing of the data and, ultimately, the interpretation of the data. Although a desirable goal, an effective re-use of public data is an extremely challenging task. Most scientists do not have the expertise and time to download masses of data, identify software solutions and learn to apply them, just to re-examine the data from one dataset. The raw data in itself, however, is of little use and thus it vanishes only too easily into oblivion.

The consequence of this gap between the storage of high-throughput data and its use by the scientific community goes beyond funding bodies deciding against the feasibility of maintaining the archives of the data. Previous scientific findings become inaccessible to future scrutiny, it is impossible to derive further knowledge from incompletely analysed data and this, in turn, leads eventually to a duplication of efforts and the repetition of the same experiments – wasting time and money. If, on the other hand, the data and results were fully accessible in the first place, only the analysis might have to be repeated, altered or extended, which is typically a process running at a fraction of the cost and time. Additionally, the combination of data from multiple different sources can lead the way to new insights and this kind of analysis depends heavily on the availability of heterogeneous data.

With these considerations in mind, I set out to use the GeneProf software described in the previous chapter (**Chapter 3**) for a large-scale re-analysis of published transcriptomic and epigenomic HTS data, to bring the results of these analyses together in one integrated database and to make the results available in an interpretable and reusable manner to the scientific community.

Others have previously re-analysed and integrated collections of public data and made them available together[74, 133, 138, 589], but the usefulness of these efforts unfortunately was limited by the scope of the project: The resource needs to keep on growing and to be constantly updated with newly published findings, users need to be able to recapitulate and modify the analysis and combine data with their own. GeneProf provides an unprecedented opportunity to make this work.

4.2 Methodology

I will now describe the methodology employed to create a consistent and integrated repository of heterogeneous functional genomics data using the streamlined, large-scale analysis facilities of the GeneProf software suite (**Chapter 3**). I will first detail the strategy for the selection and acquisition of relevant datasets and afterwards outline the way in which the data was analysed.

4.2.1 Acquiring Raw Data from Published Studies

In order to select data relevant to stem cell research for inclusion in the GeneProf databases, I searched the literature for high-profile studies with associated RNA-seq and ChIP-seq data. I focused primarily on data from stem cell and progenitor populations in mouse and human, but also wanted to include some from other cell types and systems for comparison purposes. Other than in gene expression, I was also interested in the interplay of transcription factors and the epigenetic landscape of cells defined by histone states, so I directed the search further towards any ChIP-seq data that might be relevant for this purpose, particularly, if the factors had a known or putative involvement in stem cell maintenance or the differentiation into certain lineages.

In the initial phase, I selected 72 published and unpublished studies (42 mouse: **Table 4.1**, 25 human: **Table 4.2**, one each of chicken, fruitfly, thale cress, zebrafish and C. elegans: **Table 4.3**), all of which were to be re-analysed in a consistent manner (see analysis strategy described in **Section 4.2.2**), integrated and provided via GeneProf. Some of the data analysis work was carried out with the help of Simon R. Tomlinson and Harsh J. Vaidya – details of the specific contributions are given associated with each analysis record itself. Of course, more data will be added to the database in future and I expect that GeneProf users will contribute further data and analyses, too.

As mentioned earlier, most publicly available, raw HTS data is now available in the SRA and other sources[275, 304–306, 503] and can be freely downloaded from their websites. The archives store the datasets either in compressed FASTQ format (**Section 3.3.3.1**) or have developed custom file formats in order to store the data in a more disk space-efficient manner. Such is the case for the SRA's *sra-lite* format, which will – after download and decompression – have to be converted to FASTQ to make it possible to use the data with available software. The SRA provides a special software toolkit for the format conversion.

In order to facilitate the speedy and easy acquisition of many public datasets, I have added special data import tools for SRA and ENA data to GeneProf. These tools can be used to search the respective databases by terms of interest or accession numbers (usually provided alongside publications) and will then handle the entire download, decompression and conversion process for the user. Downloads, which are potentially very time-consuming since great amounts of data need to be transferred, will be executed on the processing compute cluster (**Section 3.2.2**), so users do not need to keep their computer running while downloads are in progress. In addition to the raw experimental data, GeneProf will attempt to discover relevant sample annotations from the source database to ease recognition and interpretation of the individual datasets later on. For instance, it is in most cases possible to find information about the names (labels), cell types or tissues, organism and the technology platform used to

Identifier		Type				
Accession	**Experiment Name**	**R**	**T**	**H**	**P**	
gpXP000012	Integration of external signalling pathways in ESCs		X			75
gpXP000023	Mapping and quantifying mammalian transcriptomes by RNA-Seq	X				367
gpXP000027	Control of ESC State by Mediator and Cohesin		X			245
gpXP000028	Connecting microRNA genes to the ESC transcriptional circuitry		X	X		342
gpXP000030	ChIP-Seq in secondary fibroblast with inducible cassettes for OSK		X	X	X	RY
gpXP000031	esBAF is an essential component of the core pluripotency network		X			200
gpXP000032	ChIP-seq accurately predicts tissue-specificactivity of enhancers		X			571
gpXP000042	Combinatorial transcriptional control in blood stem/progenitorcells		X	X		595
gpXP000043	Hippocampal transcriptome of DCLK-short over-expressing mice	X				526
gpXP000048	Genome-wide mapping of Nr5a2 in mESCs		X			198
gpXP000052	Jarid2 and PRC2, partner in regulating gene expression		X			308
gpXP000056	Genome-wide mapping of SCL/DNA interactions in erythroid cells		X			254
gpXP000059	High resolution analysis of genomic imprinting in the mouse brain	X				166
gpXP000067	Transcriptional programme controlled by Scl/Tal1 during early embryonic haematopoiesis		X			596
gpXP000068	CHD7 targets enhancers to modulate ESC-specific gene expression		X			485
gpXP000071	ATAC and Mediator coactivators form a stable complex and regulates a set of non-coding RNA genes		X			279
gpXP000072	Genome-wide mapping of EBF1 binding sites in murine pre B-cells		X			553
gpXP000073	Discrete roles of STAT4 and STAT6 TFs in tuning epigenetic modifications and transcription during helper T cell differentiation		X	X		586
gpXP000074	A global network of transcription factors, involving E2A, EBF1 and FOXO1, that orchestrates the B cell fate		X	X		322
gpXP000084	KLF1/EKLF regulatory networks in primary erythroid cells		X			530
gpXP000085	SC transcriptome profiling via massive-scale mRNA sequencing	X				84
gpXP000086	Promoter proximal pausing and its regulation by c-Myc in ESCs		X		X	437
gpXP000087	GC-rich sequence elements recruit PRC2 in mammalian ES cells.		X			354
gpXP000101	Role of Prdm14 in mouse ESCs: ChIP-seq and RNA-seq analyses	X	X	X		332
gpXP000102	Transcript assembly and abundance estimation from RNA-Seq	X				552
gpXP000103	Histone marks in MEFs before and after ectopic expression of reprogramming factors			X		RY
gpXP000114	LIM domain binding protein 1 regulates a transcriptional program essential for hematopoietic SC maintenance		X			311
gpXP000117	Hoxc9 ChIP-seq in differentiating motor neurons		X			244
gpXP000121	Expression and ChIP-seq analyses of ESCs, XSCs and TSCs			X		472
gpXP000125	ChIP-Seq for REST, MCAF1, Ring1b and H4K20me3 in mESCs		X	X		RY
gpXP000127	Graded Nodal/Activin signaling governs ESC fate decisions via differential recruitment of Phospho-Smad2 to Oct4		X			300
gpXP000147	Genome-wide profiling of PPARgamma: RXR and RNApol2		X		X	384
gpXP000151	Genome wide mapping of Jarid2 and Suz12 binding sites in mESCs before and after Jarid2 depletion		X			417
gpXP000156	Regulating RNApol pausing and transcription elongation in ESCs	X				362
gpXP000168	Ab initio reconstruction of transcriptomes of pluripotent and lineage committed cells reveals gene structures of lincRNAs	X				179
gpXP000169	Genome-wide map of PCL2 enrichment in undifferentiated ESCs		X			574
gpXP000175	Deletion of Tardbp down-regulates Tbc1d1 and alters fat metabolism	X				80
gpXP000178	A SNF2 protein targets variable copy number repeats and thereby influences allele-specific expression		X			296
gpXP000191	RNA-Seq of mouse dendritic cells	X				162
gpXP000194	Dual functions of Tet1 in transcriptional regulation in ESCs		X			600
gpXP000195	Global deterministic and stochastic allelic specific gene expression in single blastomeres of mouse early embryos	X				533
gpXP000203	Genome-wide binding of STAT3 and STAT5 under Th17 conditions		X			612

Table 4.1: List of mouse experiments. Overview of studies with Mouse data in the first release of GeneProf (n = 42). Type: R = RNA-seq / DeepSAGE / GRO-seq, T = TF ChIP-seq, H = HM ChIP-seq, P = Pol2 ChIP-seq. RY = R. Young Lab, unpublished.

Accession	Experiment Name	Type R	T	H	
gpXP000003	FoxA1 ChIP-seq		X		631
gpXP000040	Genome-wide mapping of OCT4, NANOG and CTCF in hESCs		X		284
gpXP000041	Distinct epigenomic landscapes of pluripotent and lineage-committed		X		325
gpXP000161	human cells				
gpXP000047	A general mechanism for transcription regulation by Oct1 and Oct4		X		251
	in response to genotoxic and oxidative stress				
gpXP000053	ChIP-Seq of Oct4 in Human ESCs		X		RY
gpXP000057	DNA specificity determinants associate with distinct transcription		X		204
	factor functions				
gpXP000058	RNA-seq: an assessment of technical reproducibility and comparison	X			340
	with gene expression arrays				
gpXP000065	Sex-specific and lineage-specific alternative splicing in primates	X			45
gpXP000107	Multiplexed massively parallel SELEX for characterization of TF		X		240
	binding specificities				
gpXP000109	Densely interconnected transcriptional circuits control cell states in		X		394
	human hematopoiesis				
gpXP000116	Histone methylation and TF binding during intestinal differentation		X	X	570
gpXP000133	Pol II and its associated epigenetic marks are present at Pol III-	X			24
	transcribed noncoding RNA genes				
gpXP000135	Pluripotency factors regulate definitive endoderm specification		X		539
	through Eomesodermin				
gpXP000136	Altered antisense-to-sense transcript ratios in breast cancer	X			345
gpXP000145	Dynamic transcriptomes during neural differentiation of ESCs	X			602
gpXP000153	RNA sequencing reveals the role of splicing polymorphisms in regu-	X			289
	lating human gene expression				
gpXP000160	Nascent RNA sequencing reveals widespread pausing and divergent	X			91
	initiation at human promoters				
gpXP000167	Genome-wide analysis of histone methylations in memory CD8+ T			X	10
	cells				
gpXP000178	A SNF2 protein targets variable copy number repeats and thereby		X		296
	influences allele-specific expression				
gpXP000181	Mapping of ETV1 genomic binding sites in gastrointestinal stromal		X		77
	tumor				
gpXP000182	Mediation of CTCF transcriptional insulation by DEAD-box RNA-		X		614
	binding protein p68 and steroid receptor RNA activator SRA				
gpXP000222	Identification of Beta-catenin binding regions in colon cancer cells		X		48
	using ChIP-Seq				
gpXP000255	Analysis of E2F1 mutant proteins reveals that N- and C-terminal		X		60
	protein interaction domains do not participate in targeting E2F1				
gpXP000265	Functional analysis of Kap1 genomic recruitment		X		228
gpXP000377	Mapping and analysis of chromatin state dynamics in nine human		X	X	122
gpXP000389	cell types (ENCODE project, split across 3 GeneProf experiments)				
gpXP000390					

Table 4.2: List of human experiments. A complete overview of all studies with Human data chosen for inclusion in the initial release of GeneProf (n = 25). Type: R = RNA-seq / DeepSAGE / GRO-seq, T = transcription factor / regulator ChIP-seq, H = histone ChIP-seq. RY = R. Young Lab, unpublished data.

Accession	Experiment Name	Organism	Type R	C	
gpXP000049	Sequencing of small RNAs from C. elegans embryos	C. elegans	X		519
gpXP000060	RNA-Seq of Drosophila cell line Dmel2	D. melanogaster	X		205
gpXP000062	Traf6 function in the innate immune response of ze-	D. rerio	X		518
	brafish embryos				
gpXP000108	Deep sequencing of small RNAs in transgenic wild type	A. thaliana	X		252
	plant and IWR1-type TF mutant				
gpXP000188	Shox ChIP-seq in chicken micromass cell cultures	G. gallus		X	105

Table 4.3: List of other experiments. A complete overview of all studies with data from organisms other than Human or Mouse, which were chosen for inclusion in the initial release of GeneProf (n = 5). Type: R = RNA-seq / DeepSAGE / GRO-seq, C = ChIP-seq.

create the datasets.

Utilising the import tools, acquiring the data from the studies outlined in the tables (**Table 4.1**, **Table 4.2** and **Table 4.3**) was rather straightforward and achieved with a minimum of hands-on time. I chose to subsequently manually augment, correct and standardise the sample annotation in order to support the intelligibility of what experiments are about and, ultimately, to make it possible to easily and meaningfully compare and juxtapose datasets from various sources later on.

As a bare minimum, I tried to always provide information about the organism, technology platform, meaningful dataset labels, groupings of datasets, cell types, tissues, cell lines and the targets of ChIP-seq antibodies, wherever applicable. This information was derived either from the full-text descriptions of the data in the source databases or by consulting the methods sections and supplementary material of the corresponding research publications.

4.2.2 Using GeneProf for High-Throughput Analysis

I will now explain how the GeneProf data analysis suite has been employed to streamline a large-scale reanalysis of published RNA- and ChIP-seq data to build up an integrated HTS-based resource of functional genomics data.

4.2.2.1 Wizard-Based Analysis

In order to create a fully integrated database of analysed experimental data that can be compared in a meaningful manner it is of paramount importance that all data must be processed in a consistent manner. However, it is equally important to acknowledge that it is not appropriate to analyse every single dataset in exactly the same way – too different are the protocols employed in various labs across the world and, even more so, too varied the biology underlying the experiments. "Consistent" does therefore not necessarily mean identical, but following equivalent principles and guidelines that ensure that the data will, on the one hand, be analysed in the most appropriate way for the dataset at hand and, at the same time, ensure the comparability of the results obtained.

I decided to use GeneProf's data analysis wizards with the default settings for all analysis in the first place (**Section 3.3.2.2**). After an initial run, I examined the automatically created summary reports manually in detail and, if necessary, adjusted the analysis procedure to deal with datasets for which the default procedure was not sufficient (see exploratory analysis: **Section 3.3.2.3**).

Most commonly, adjustments to the analysis pipeline only necessitated the truncation of reads to a certain length. As discussed before (**Section 3.3.3.1**), the quality of short read sequencing datasets does tend to decline towards the end of the reads due to the accumulation

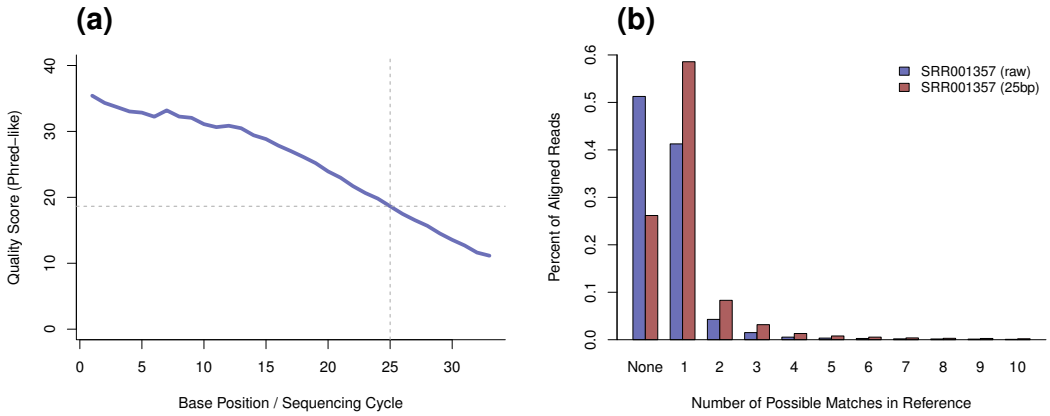

Figure 4.1: Accounting for problematic HTS data. The alignment of the dataset SRR001357[367] could be improved by trimming reads to a fixed length of 25bp. **(a)** After the first 25 sequencing cycles the quality scores drop below 20. **(b)** Approximately half of all unaligned reads could be aligned after trimming.

of sequencing errors. This effect is particularly pronounced for older datasets, which were using the early generation sequencing platforms, or for particularly long reads, stretching the capabilities of the technology. It is usually impossible to know *a priori* whether this phenomenon has any significant impact on a new dataset, but the plots provided by GeneProf help to quickly spot any trouble caused: If the alignment success rate falls below a certain level (any less than 60% of all reads aligned uniquely to the genome might be a reason for concern) a likely reason might be the suboptimal quality of the reads, which can be examined by looking at quality scores and nucleotide distribution across read cycles (**Figure 4.1**). If the plots revealed a clear break point beyond which the quality of the data seemed unacceptable, I would usually trim the reads to this length. Otherwise I tried to use a dynamic filtering strategy and truncated each read dynamically from the point onwards, where the quality dropped below a certain threshold (between $Q = 5$ or $Q = 10$ depending on the average quality score of the dataset; cp. **Section 3.3.3.1**), discarding any reads that were subsequently shorter than $12bp$. In rare cases, even these measures did not suffice to give a satisfactory alignment success rate, which prompted me to use an iterative alignment procedure[84] (implemented in a single GeneProf module): After initial quality control, I would attempt to align the entire library. Those reads that could not be aligned in the first step would then be truncated by $1 - 5bp$ and aligned again. The procedure was repeated up to ten times or until (a) no unaligned reads remained or (b) reads were too short to proceed with.

To further improve data processing, I also considered using the Tophat alignment tool[552] instead of the default option, Bowtie[292], whenever paired-end / mate-pair or long-read ($\geq 50bp$) RNA-seq data was concerned. The reason for this is simply that, for longer reads, the probability that a read might span the junction between multiple exons rises ("spliced

read"). Ungapped alignment programs, like Bowtie, cannot find a match for these reads in the genome, where the exonic sequences are interleaved with intronic DNA that is not present in the transcript sequence (**Section 3.3.3.2**). Tophat, on the other hand, has been developed to discover potential splice junctions automatically and does hence offer a better sensitivity for these datasets. Note that, even in datasets with short reads, some of the transcript fragments read out will span splice junctions, however, the proportion of coverage lost by missing the alignment of these reads is usually negligible and Bowtie has a clear advantage over Tophat in terms of speed (up to ten times faster), which makes it a more attractive default choice for a large-scale, generic and public data analysis system.

4.2.2.2 ChIP-seq Analysis

I employed the "All-in-one ChIP-seq Analysis Wizard" for the reanalysis of all transcription factor (TF)-binding and histone-modification ChIP-seq experiments alike (**Section 3.3.2.2**). Just to recapitulate, the wizard will create an experiment-specific data processing pipeline consisting of the following steps:

1. Merge raw read datasets belonging to the same ChIP-seq experiment. For instance, if multiple sequencing lanes have been used to increase coverage for the same DNA-associated protein, all corresponding datasets will be merged into one before proceeding.

2. Create summary reports for the quality and nucleotide composition of all datasets and apply basic quality control measures by filtering out all reads with a very low average quality score ($mean(Q) < 8$) (**Section 3.3.3.1**).

3. Align all libraries individually to the reference genome of the organism they belong to using the Bowtie algorithm[292] (**Section 3.3.3.2**). Discard all non-unique alignments.

4. Create summary reports for the alignment success rate and chromosomal distribution of alignments.

5. Use the MACS peak finding algorithm[631] to detect significantly enriched binding events ("peaks") corresponding to putative DNA-protein binding sites (**Section 3.3.3.5**).

6. Create summary statistics and plots describing the number and genomic distribution of binding sites. If multiple factors have been studied in the same experiment, the summary will also compare the binding sites for all these factors.

7. Associate the binding sites with nearby genes either in a binary fashion ("has a binding site" or "has no binding site") by considering a gene a target of a factor, if it has a binding site anywhere in the region up to $20kb$ upstream or $1kb$ downstream of the transcription start site (TSS) of the gene (**Section 3.3.3.5**).

8. Additionally, consider the transcription factor association strength (TFAS)[406] between all genes and each factor studied in the experiment to gain a good ranking criterion for interesting candidates (**Section 3.3.3.5**).

The wizard has been designed primarily for TF data and the algorithms chosen are optimised for this kind of data, however, I found that the methods could also be used reasonably well for a basic analysis of other ChIP-seq data even if it does not exhibit the characteristic binding patterns of TFs, which typically have well-defined narrow binding sites. Histones occupy larger regions of the genome and the "peaks" (**Section 3.3.3.5**) are less well defined than for TFs, but are nevertheless mostly detected using the MACS-algorithm[631] used by the wizard (MACS recommends certain parameter settings for histone modifications). More sophisticated analyses and comparisons of histone modifications can be performed at a later point on the basis of the alignment coverage reported in these experiments (see **Chapter 5**).

4.2.2.3 RNA-seq Analysis

For transcriptomic assays, that is RNA-seq and DeepSAGE experiments, I used the "All-in-one RNA-seq Analysis Wizard" in turn, creating workflows consisting of the following steps:

1. If applicable, merge raw read datasets for technical replicates.

2. Create summary reports for the quality and nucleotide composition of all datasets and apply basic quality control measures by filtering out all reads with a very low average quality score ($mean(Q) < 8$) (**Section 3.3.3.1**).

3. Align all libraries individually to the reference genome of the organism they belong to using the Bowtie algorithm[292] (**Section 3.3.3.2**). Accept alignments with up to 10 possible matches in the genome. For paired-end read datasets, I changed an option of the wizard with the effect that, instead of Bowtie, the Tophat program[552] was to be used, which is capable of dealing with gapped alignments. For datasets produced using the SOLiD platform, I used the iterative alignment strategy as described above (**Section 4.2.2.1**).

4. Create summary reports for the alignment success rate and chromosomal distribution of alignments.

5. Quantify gene expression by calculating the genomic coverage of reads with respect to known gene models using GeneProf's custom algorithms (**Section 3.3.3.3**). For short RNA datasets, a special analysis module was used that considered only shortRNA-features in the reference dataset.

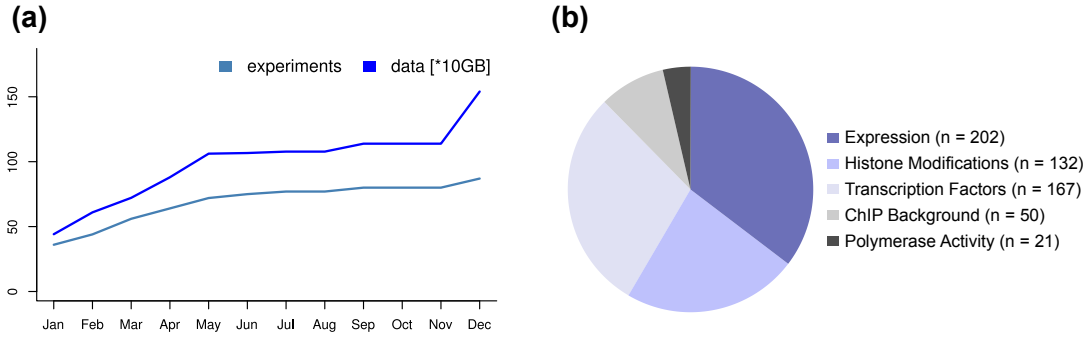

Figure 4.2: Publicly accessible experiments in the GeneProf database. (a) The amount of public data stored in the GeneProf databases has been growing constantly over the last year (January to December 2011). **(b)** Many genomic datasets are available as tracks for the built-in genome browser.

6. Summarise gene expression in all investigated datasets, compare the data and create heatmaps, correlation matrices, principal component analysis and other plots.

7. Use the DESeq algorithm[7] to assess differential gene expression between all groups of datasets in the experiment, i.e. between different biological conditions, cell types or tissues (**Section 3.3.3.4**).

8. Created filtered tables of genes found differentially expressed in each comparison ($FDR < 0.05$).

In this way, I could very quickly analyse gene expression patterns in a wide variety of biological systems and conditions. Importantly, the results include, apart from experiment-specific assays of differential expression, reusable measures of gene transcription (raw read counts per gene as well as intensities normalised as reads-per-million (RPM) and reads-per-kilobase-million (RPKM); **Section 3.3.3.3**), which will allow users to integrate data from multiple experiments straightforwardly in a useful manner.

4.3 A Knowledge-Base for Functional Genomics Experiments

At the time of the first public release of GeneProf in the beginning of January, 2012, the GeneProf databases had accumulated data from 72 independent experiments or 937 different HTS runs, amounting to more than 12,217,419,081 (12.2 billion) short reads and approaching 2 terabytes of public data. In addition to this, more than an equal amount of data was yet in the progress of being analysed and awaiting inclusion in the public databases. This is a vast amount of data not usually at the disposal of even the largest research labs (**Figure 4.2**.a).

In order to give the reader a better impression of what sort of information GeneProf offers to its user, I will now give four illustrative examples:

1. **Gene-centric information retrieval.** GeneProf automatically compiles all data relevant for the gene of interest into one concise summary page by cross-matching assorted data from many public experiments (**Section 3.3.3.6**). **Figure 4.3** shows the gene summary page for the transcription factor *Nanog* (in mouse) as an example. The page first provides generic information about the gene (collected from other databases), e.g. the name, external identifiers, transcript variants (all from Ensembl[136]), protein structure (Protein Data Bank[34]), functional annotation (Gene Ontology[11]) and known protein-protein interactions (BioGRID[516]). The following sections summarise information about (i) the expression of the gene in different conditions and cell types (based on RNA-seq data in GeneProf), (ii) genes potentially targeted by *Nanog* and (iii) TFs with enriched binding activity near *Nanog* (based on ChIP-seq data in GeneProf).

2. **Dissemination of genomic data.** Much of GeneProf's genomic data is available in the form of customisable tracks that can be displayed and juxtaposed in the built-in genome browser (**Section 3.3.2.5**) in order to visually disseminate the mechanisms of genome biology (**Figure 4.2**.b). **Figure 4.4** shows a screenshot of an active genome browser session in which I have visualised the genomic environment of *Nanog*, including tracks for three RNA-seq datasets[179] as well as ChIP-seq data for the TFs *Pou5f1, Nanog* and *Sox2* from two studies[75,342].

3. **Discovering patterns in large data collections.** With the Visual Data Explorer (VDE; **Section 3.3.2.5**), gene expression data and information about DNA-protein binding sites from many different experiments can be integrated and plotted together within seconds. To illustrate the use of the VDE, I picked human RNA-seq datasets from various publications[321,345,524,602] via the VDE interface and used two different plot types to compare their gene expression patterns: (i) Correlation matrix: A simple, graphical representation of the pair-wise Pearson correlation coefficients calculated between all datasets (**Figure 4.5**.a) and (ii) Principal component analysis (PCA): A mathematical method that extracts descriptive variables from the expression data (**Figure 4.5**.b). Both plots show how functionally related cell types cluster closely together, because their expression profiles are similar.

4. **Scrutinisation of public experiments.** Transparency and reproducibility of scientific data have been one of the main driving forces in the development of the GeneProf software (**Section 3.1** and **Section 3.3.2.4**). In order to avoid the obfuscation of results, I have therefore decided to not only make the final outcomes of GeneProf analyses available, but to also complement those with the entire analysis workflow, so that it may be subjected to the critical assessment of our peers. Every user can now browse through all public experiments, find out in detail how every step of the analysis was done and

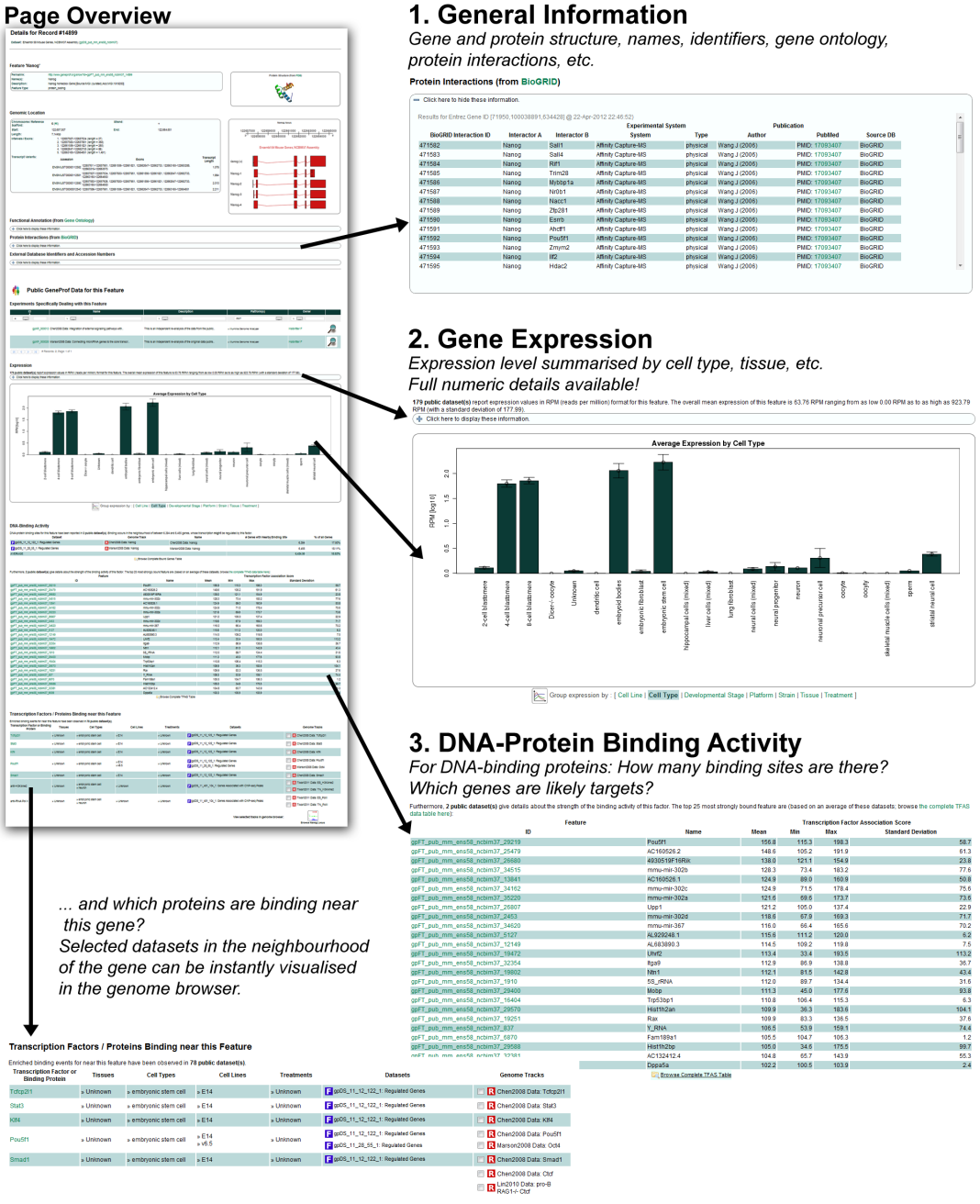

Figure 4.3: Gene-centric data summary. Overview of the gene-centric summary page for the gene *Nanog* with assorted sections highlighted. Retrieved 22 April 2012; `http://www.geneprof.org/show?id=gpFT_pub_mm_ens58_ncbim37_14899`.

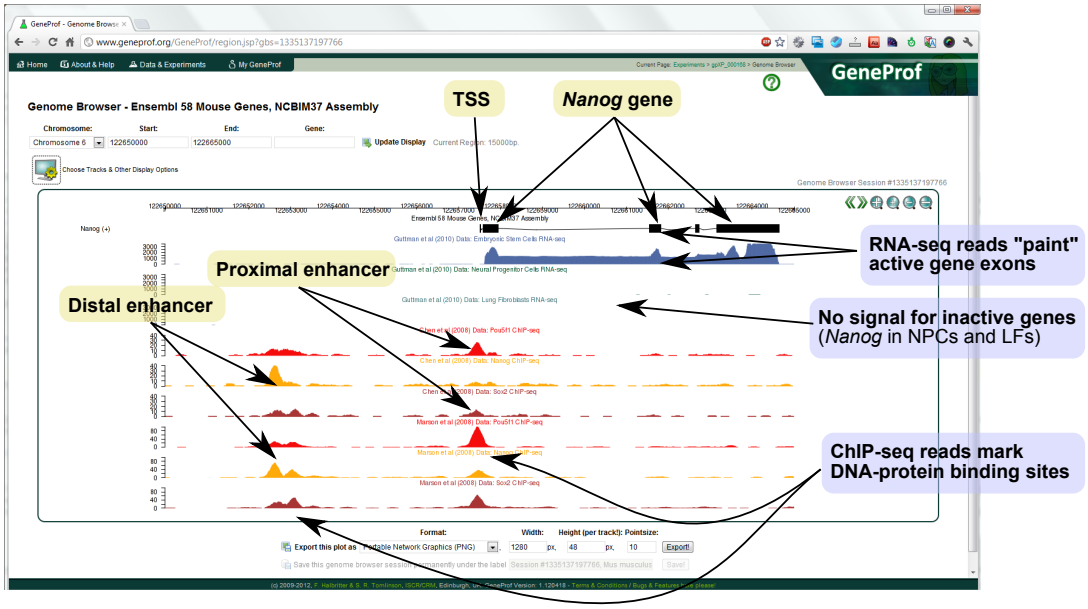

Figure 4.4: Genome browser: *Nanog*. This is an annotated screenshot showing the genomic landscape made up of aligned RNA-seq data from ESCs, neural progenitor cells and lung fibroblasts[179] and ChIP-seq data for the factors *Pou5f1, Nanog* and *Sox2* from two studies[75,342]. Shown here is the *Nanog* locus.

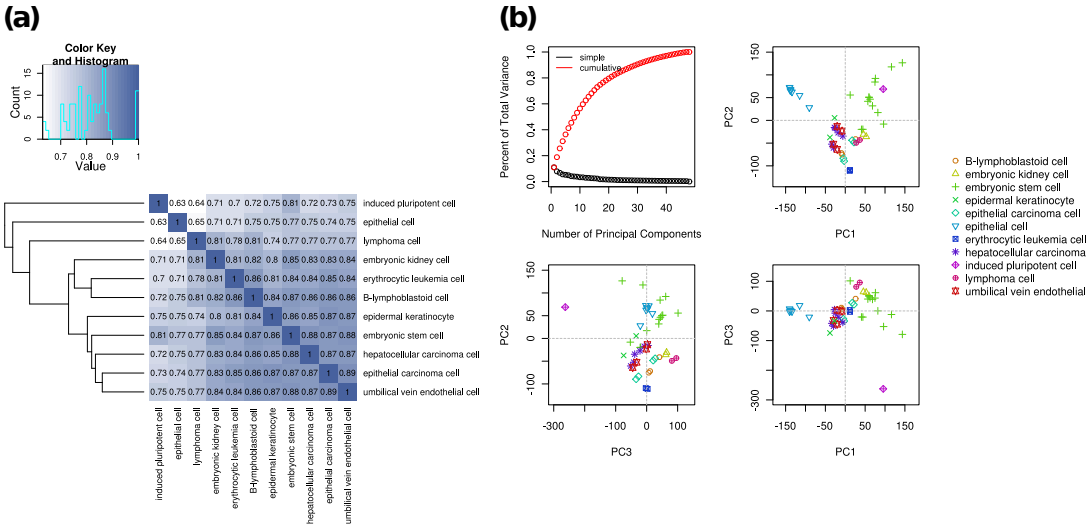

Figure 4.5: Visual data exploration. Example plots exported directly from GeneProf's Visual Data Explorer. (**a**) Visualisation of a Pearson correlation matrix between RNA-seq datasets summarised by cell type (correlation between arithmetic means). (**b**) Principal component (PC) analysis of the same datasets. The plot on the top left shows the percentage of variation explained by the individual PCs (block circles) and the sum of all PCs up to this point (red circles). The remaining plots show the contribution of each individual dataset to the first, second and third PC. Datasets clustering together are characteristically similar to each other in their gene expression profile.

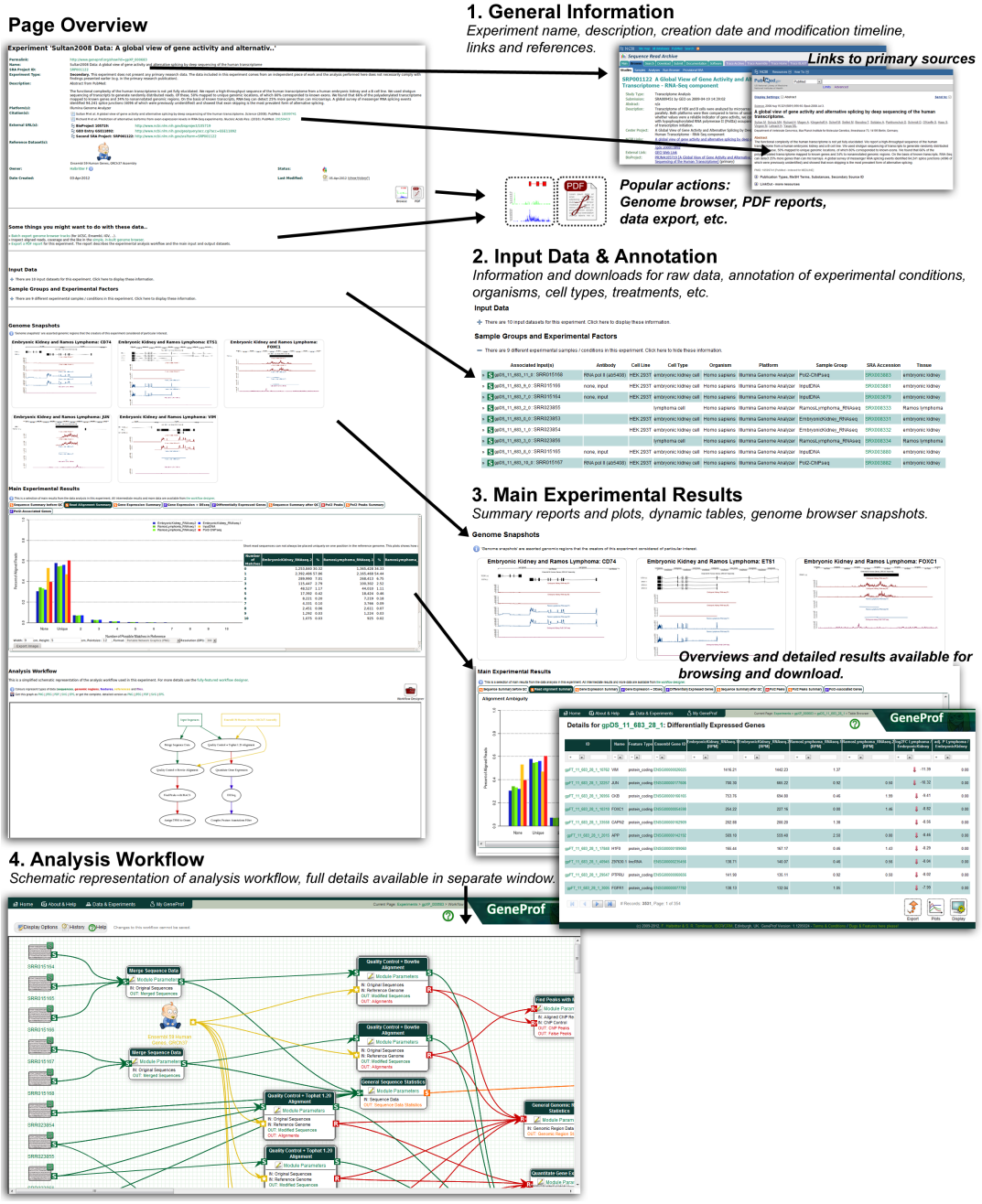

Figure 4.6: Experiment main page: *gpXP_000683*. The experiment main page of the GeneProf record with accession number *gpXP_000683* (`http://www.geneprof.org/show?id=gpXP_000683`). Selected sections have been highlighted and enlarged. This page summarises the most important information about a data analysis experiment in GeneProf. Many additional details are available via the other pages linked from this page.

decide whether the results are trustworthy – if not, the user can repeat questionable parts of the analysis herself. As an example, I show here the GeneProf experiment *gpXP_000683*, which is based on RNA-seq and ChIP-seq data from Sultan *et al.*[524]. From the main page for this experiment (**Figure 4.6**), GeneProf users have immediate access to the original publications and data sources (link-out to PubMed, SRA and GEO), the raw input data and the analysis workflow. The page also shows the main analysis results, including summary reports about raw data quality, alignment, gene expression and DNA-protein binding peaks.

4.4 Conclusion

The combination of the GeneProf software with the results of the data analyses described earlier (**Section 4.2**) and the advanced knowledge retrieval mechanisms outlined in the previous section (**Section 4.3**), make GeneProf more than a classic data analysis suite and yet more than a traditional static online database: The combination of all features offers the potential to serve as a truly useful and comprehensive resource for a wide range of scientists and to have a long-lasting impact on research by promoting knowledge transfer, exchange and exploitation. It was with the benefit of this plethora of operative data that I was able to address the questions investigated in the following chapter.

Chapter 5

An Integrative View of the Core Transcriptional Circuitry of Stem Cells

The development of the GeneProf software (**Chapter 3**) and database (**Chapter 4**) provided me with the ideal tool to tap the vast amount of genomic and epigenomic data accumulated by the scientific community over the past years. The aim was to extract relevant knowledge and derive novel insight into the workings of the core transcriptional circuitry of embryonic stem cells and about how single genetic factors fit into a large network able to shape a complex biological entity that will eventually give rise to life in all its splendid variety.

Many genes paramount to the establishment and maintenance of stem cell state have been identified over the last years (**Section 1.1.4**) and much attention has recently been paid to the regulatory mechanisms that influence their expression. Still, little is known about how the complex interplay of multiple regulatory signals can drive gene expression in such a precise way as it is required to distinguish the manifold types of cells of the developing and adult body. In this work, I was asking the question as to whether there was indeed a defining regulatory code (made up of a signature of DNA-binding proteins and histone modifications) that was able to separate genes that are specifically expressed only in stem cells from the remainder of the genes in the transcriptome (including those that might be active in stem cells and other cell types).

Protein	Code	Experiment	Line	Protein	Code	Experiment	Line
Myc	Myc	gpXP000012^{75}	E14	*Smarca4*	Sma4	gpXP000031^{200}	E14Tg2a
Tcfcp2l1	T2l1	gpXP000012^{75}	E14	*Ep300*	P3-2	gpXP000068^{485}	R1
Ep300	P3-1	gpXP000012^{75}	E14	*Chd7*	Chd7	gpXP000068^{485}	R1
E2F1	E2f1	gpXP000012^{75}	E14	*Jarid2*	Jd2	gpXP000052^{308}	V6.5
Zfx	Zfx	gpXP000012^{75}	E14	*Mtf2*	M2-1	gpXP000052^{308}	V6.5
Mycn	Mycn	gpXP000012^{75}	E14	*Nr5a2*	N5a2	gpXP000048^{198}	E14
Nanog	Ng-1	gpXP000012^{75}	E14	*Luzp1*	Luz	gpXP000071^{279}	E14
Sux12	Sz-1	gpXP000012^{75}	E14	*Spt5*	Spt5	gpXP000086^{437}	V6.5
Esrrb	Esrb	gpXP000012^{75}	E14	*NelfA*	NlfA	gpXP000086^{437}	V6.5
Ctcf	C-1	gpXP000012^{75}	E14	*Ctr9*	Ctr9	gpXP000086^{437}	V6.5
Sox2	Sx-1	gpXP000012^{75}	E14	*Yy1*	Yy1	gpXP000087^{354}	V6.5
Smad1	Smd1	gpXP000012^{75}	E14	*Prdm14*	Prdm	gpXP000101^{332}	LF2
Pou5f1	Po-1	gpXP000012^{75}	E14	*Ring1b*	R1b	gpXP000125 RY	V6.5
Klf4	Klf4	gpXP000012^{75}	E14	*REST*	Rest	gpXP000125 RY	V6.5
Stat3	S3	gpXP000012^{75}	E14	*MCAF1*	Mcaf	gpXP000125 RY	V6.5
Med1	Md1	gpXP000027^{245}	V6.5	*ATRX*	Atrx	gpXP000178^{296}	E14
Med12	Md12	gpXP000027^{245}	V6.5	*Mtf2*	M2-2	gpXP000169^{574}	R1
Smc3	Smc3	gpXP000027^{245}	V6.5	*Tet1*	Tet1	gpXP000194^{600}	E14Tg2A
Smc1	Smc1	gpXP000027^{245}	V6.5	*Ctcf*	C-2	gpXP000445^{512}	?
Nipbl	Nipb	gpXP000027^{245}	V6.5	*Ctcf*	C-3	gpXP000445^{512}	?
Nanog	Ng-2	gpXP000028^{342}	V6.5	*Ctcf*	C-4	gpXP000445^{512}	?
Suz12	Sz-2	gpXP000028^{342}	V6.5	*Smad3*	Smd3	gpXP000426^{368}	V6.5
Pou5f1	Po-2	gpXP000028^{342}	V6.5	*Jnk1/3*	Jnk	gpXP000481^{548}	?
Sox2	Sx-2	gpXP000028^{342}	V6.5	*Nfya*	Nfya	gpXP000481^{548}	?
Tcf3	Tcf3	gpXP000028^{342}	V6.5				

Table 5.1: Selected DNA-protein binding ChIP-seq datasets. ChIP-seq datasets assaying DNA-binding proteins (TFs, co-factors, ..) selected for further analysis. For the sake of brevity, dataset names are abbreviated in plot labels (column "code"). References refer to the study in which the data was originally released, RY = Richard Young, unpublished data.

5.1 Materials and Methods

I manually traversed the GeneProf database (**Chapter 4**) for experiments profiling the DNA-protein association of transcription factors, co-factors, epigenetic marks and elements of the transcriptional apparatus previously implicated in the control of pluripotency and self-renewal (DNA binding proteins: DBPs). I also looked for datasets with gene expression profiling and histone modification (HM) data in ESCs and other cell types.

Doing so, I collected 49 ChIP-seq datasets for DBPs (**Table 5.1**), 27 ChIP-seq datasets for HMs (**Table 5.2**) and 49 gene expression (RNA-seq) datasets (**Table 5.3**). For an overview of the putative function of these DBPs and HMs see **Section 1.1.4** and **Section 1.1.5**. For a few target proteins, I found more than one ChIP-seq dataset, e.g. there were multiple ChIP-seq datasets for the three core-factors. Similary, there were multiple RNA-seq datasets for most cell types. I expect that these data can give us an idea of the biological variability and believe that, by considering all results across laboratories and biological variants (e.g. different cell lines), one might be able to disseminate true core mechanisms from random (or non-targeted) variation.

For all the analyses presented in this chapter I used GeneProf to prepare and process the data and *R* to refine and customise plots and visualisations exported from GeneProf. GeneProf experiments with data analysis workflows and primary results are accessible via the web interface (**Section D.1**).

	In Embryonic Stem Cells			In Embryonic Fibroblasts	
HM	**Code**	**Experiment**	**Line**	**Code**	**Experiment**
H4K20me3	E20m3-1	gpXP_000125 RY	v6.5		
	E20m3-2	gpXP_000535^{361}	v6.5		
H3K27me3	E27m3-1	gpXP_000445^{512}	?	F27m3-1	gpXP_000103 RY
	E27m3-2	gpXP_000445^{512}	?	F27m3-2	gpXP_000535^{361}
	E27m3-3	gpXP_000445^{512}	?		
	E27m3-4	gpXP_000121^{472}	R1		
	E27m3-5	gpXP_000121^{472}	R1		
	E27m3-6	gpXP_000535^{361}	v6.5		
	E27m3-7	gpXP_000481^{548}	?		
H3K36me3	E36m3-1	gpXP_000028^{342}	v6.5	F36m3	gpXP_000535^{361}
	E36m3-2	gpXP_000535^{361}	v6.5		
H3K4me1	E4m1	gpXP_000445^{512}	?		
H3K4me2	E4m2-1	gpXP_000445^{512}	?		
	E4m2-2	gpXP_000445^{512}	?		
	E4m2-3	gpXP_000481^{548}	?		
H3K4me3	E4m3-1	gpXP_000121^{472}	R1	F4m3-1	gpXP_000103 RY
	E4m3-2	gpXP_000121^{472}	R1	F4m3-2	gpXP_000535^{361}
	E4m3-4	gpXP_000535^{361}	v6.5		
H3K79me2	E79m2	gpXP_000028^{342}	v6.5		
H3K9me3	E9m3	gpXP_000535^{361}	v6.5	F9m3-1	gpXP_000103 RY
				F9m3-2	gpXP_000535^{361}

Table 5.2: Selected histone modification ChIP-seq datasets. ChIP-seq datasets assaying histone modifications (HM) selected for further analysis. For the sake of brevity, dataset names are abbreviated in plot labels (column "code"). References refer to the study in which the data was originally released, RY = Richard Young, unpublished data.

Cell Type	**Code**	**Experiment**	**Cell Type**	**Code**	**Experiment**
Blastomere, 2-cell	B2-1*	gpXP_000195^{533}	ESC, *Prdm14* RNAi	ESC_P14	gpXP_000101^{332}
Blastomere, 2-cell	B2-2*	gpXP_000195^{533}	ESC, *Tardbp*$^{-/-}$	ESC_T-1	gpXP_000175^{80}
Blastomere, 2-cell	B2-3*	gpXP_000195^{533}	ESC, *Tardbp*$^{-/-}$	ESC_T-2	gpXP_000175^{80}
Blastomere, 2-cell	B2-4*	gpXP_000195^{533}	ESC, *Tardbp*$^{-/-}$	ESC_T-3	gpXP_000175^{80}
Blastomere, 2-cell	B2-5*	gpXP_000195^{533}	ESC	ESC-1	gpXP_000101^{332}
Blastomere, 2-cell	B2-6*	gpXP_000195^{533}	ESC	ESC-2	gpXP_000480^{274}
Blastomere, 2-cell	B2-7*	gpXP_000195^{533}	ESC	ESC-3	gpXP_000482^{512}
Blastomere, 2-cell	B2-8*	gpXP_000195^{533}	ESC	ESC-4	gpXP_000482^{512}
Blastomere, 4-cell	B4-1*	gpXP_000195^{533}	ESC	ESC-5	gpXP_000085^{84}
Blastomere, 4-cell	B4-2*	gpXP_000195^{533}	ESC	ESC-6	gpXP_000085^{84}
Blastomere, 4-cell	B4-3*	gpXP_000195^{533}	ESC	ESC-7	gpXP_000085^{84}
Blastomere, 4-cell	B4-4*	gpXP_000195^{533}	ESC	ESC-8	gpXP_000168^{179}
Blastomere, 4-cell	B4-5*	gpXP_000195^{533}	ESC	ESC-9	gpXP_000175^{80}
Blastomere, 4-cell	B4-6*	gpXP_000195^{533}	ESC	ESC-10	gpXP_000175^{80}
Blastomere, 8-cell	B8-1*	gpXP_000195^{533}	Neural Progenitor	NPC-1	gpXP_000482^{512}
Blastomere, 8-cell	B8-2*	gpXP_000195^{533}	Neural Progenitor	NPC-2	gpXP_000482^{512}
Blastomere, 8-cell	B8-3*	gpXP_000195^{533}	Neural Progenitor	NPC-3	gpXP_000168^{179}
Blastomere, 8-cell	B8-4*	gpXP_000195^{533}	Oocyte *Dicer*$^{-/-}$	Ooc_D-1*	gpXP_000195^{533}
Blastomere, 8-cell	B8-5*	gpXP_000195^{533}	Oocyte *Dicer*$^{-/-}$	Ooc_D-2*	gpXP_000195^{533}
Blastomere, 8-cell	B8-6*	gpXP_000195^{533}	Oocyte *Dnmt3l*$^{-/-}$	Ooc_D3	gpXP_000480^{274}
Embyoid Body	EB-1	gpXP_000085^{84}	Oocyte	Ooc-1	gpXP_000480^{274}
Embyoid Body	EB-2	gpXP_000085^{84}	Oocyte	Ooc-2*	gpXP_000195^{533}
Embyoid Body	EB-3	gpXP_000085^{84}	Oocyte	Ooc-3*	gpXP_000195^{533}
Embyoid Body	EB-4	gpXP_000085^{84}	Sperm	Sperm	gpXP_000480^{274}
Lung Fibroblast	LF	gpXP_000168^{179}			

Table 5.3: Selected gene expression RNA-seq datasets. RNA-seq datasets assaying gene expression selected for further analysis. For the sake of brevity, dataset names are abbreviated in plot labels (column "code"). References refer to the study in which the data was originally released. Datasets marked with an asterisk (*) are from single-cell studies.

5.2 Results

In order to drill down on the mechanisms that make stem cells what they are, I proceeded sequentially by first establishing a list of genes with an ESC-specific expression pattern (**Section 5.2.1**). I then looked on a broad scale at the wider genomic landscape of ESCs made up of histone marks and various types of DNA-associating proteins (**Section 5.2.2**) and then studied each of these in more detail (**Section 5.2.3** and **Section 5.2.4**, respectively). Lastly, I used the combination of all three types of measurements (gene expression, HMs and DBPs) to discriminate different groups of stem cell-related genes and to identify their regulatory markup (**Section 5.2.5**). **Figure 5.1** shows an overview of the entire analysis pipeline.

5.2.1 Identification of Members of the Core Transcriptional Circuitry

I first sought to identify genes and possibly other transcriptional features that were integral to the maintenance of stem cell identity. A number of groups have attempted to track down lists of "stem cell genes" by computational analysis before[12, 157, 276, 363] and I did not expect any groundbreaking revelations at this point. Rather the aim was to determine an updated and extended list of known key players, whose transcriptional patterns could be integrated in the subsequent analysis.

To do so, I used GeneProf to quantify the expression level of each gene in each of the assorted expression datasets (**Section 3.3.3.3**). To improve comparability of the calculated intensities (as RPKM), the expression values across all datasets were quantile-normalised*. Not unsurprisingly, I found striking differences between datasets other than explained by biological variation alone: While the bulk of all expression in most datasets could be attributed to protein-coding genes (as it would usually be expected in standard RNA-seq experiments), some datasets had a considerable skew towards miRNA and ncRNA transcription (**Figure 5.2**.a).

It should be noted that this drastic non-uniformity is, at least in part, due to the RPKM normalisation used (**Section 3.3.3.3**), which tends to inflate expression intensities recorded for very short transcripts (such as miRNAs and ncRNAs), making the effects of elevated short RNA expression levels more pronounced. Nevertheless, there is an apparent imbalance in the initial genome-wide distribution of reads, which I believe is due to technical differences between sequencing platforms and, in particular, differing protocols in the way the input material (RNA) was treated. Specifically, ESC-3, ESC-4, NPC-1 and NPC-2, all samples from the same study[512], have been prepared using depletion strategy for ribosomal RNA rather than by using the "standard" poly-A selection strategy employed in the other studies.

*Where necessary, I will in the following refer to the quantile normalised RPKM expression values as X_{qRPKM}

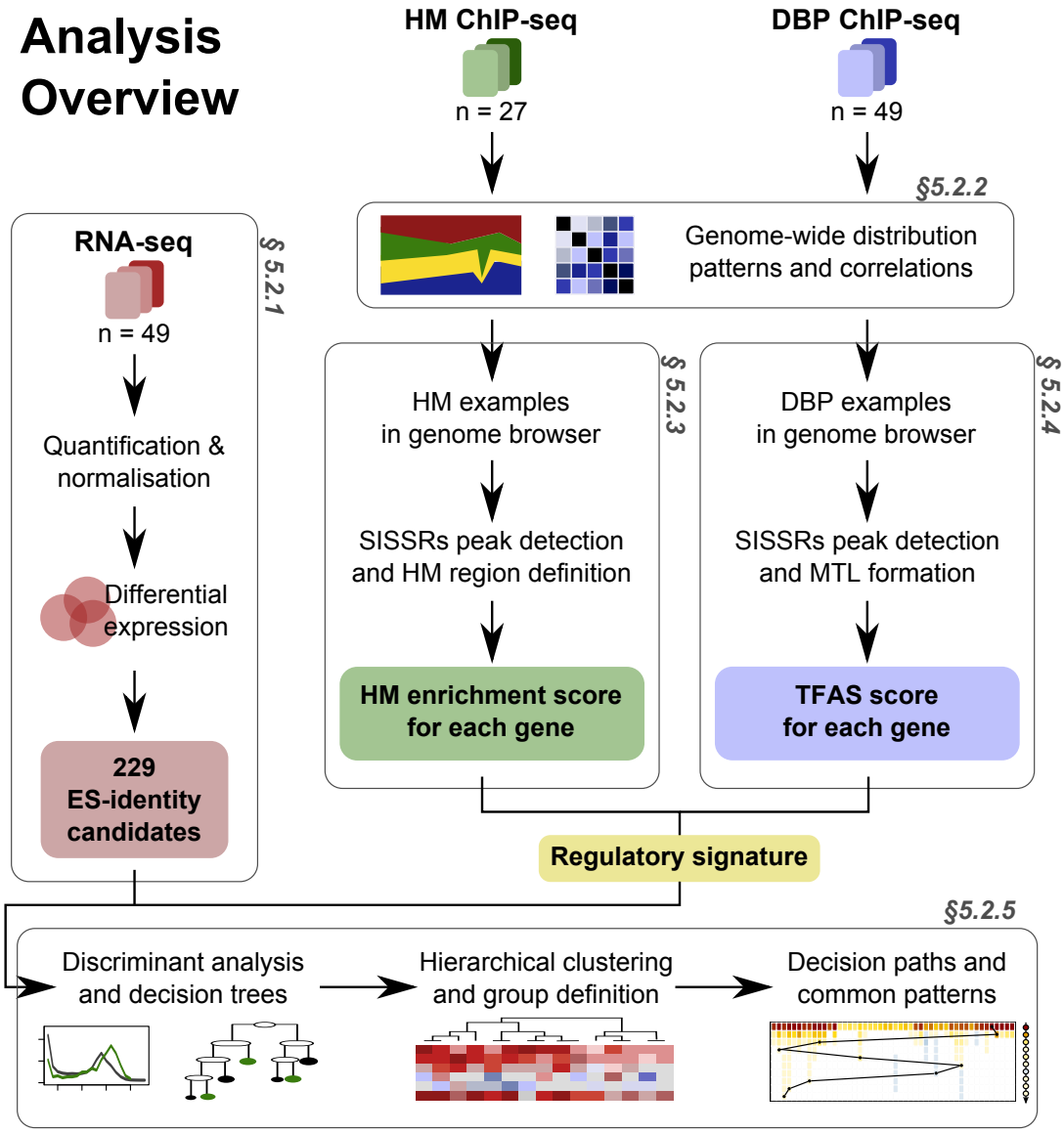

Figure 5.1: Overview of analysis pipeline. The analysis presented in this chapter consists of three converging branches: RNA-seq expression data is used to establish a list of genes specifically expressed in ESCs (**Section 5.2.1**). ChIP-seq data for HMs and DBPs is first analysed independently to calculate gene-centric HM enrichment scores and TFAS scores for DBPs, which are then combined into a regulatory signature for each gene (**Section 5.2.2**, **Section 5.2.3** and **Section 5.2.4**) . Using this signature, I employ machine learning methods to cluster the ES-identity candidate genes identified in the first step into groups and study the regulatory signature of one of these subgroups in detail (**Section 5.2.5**).

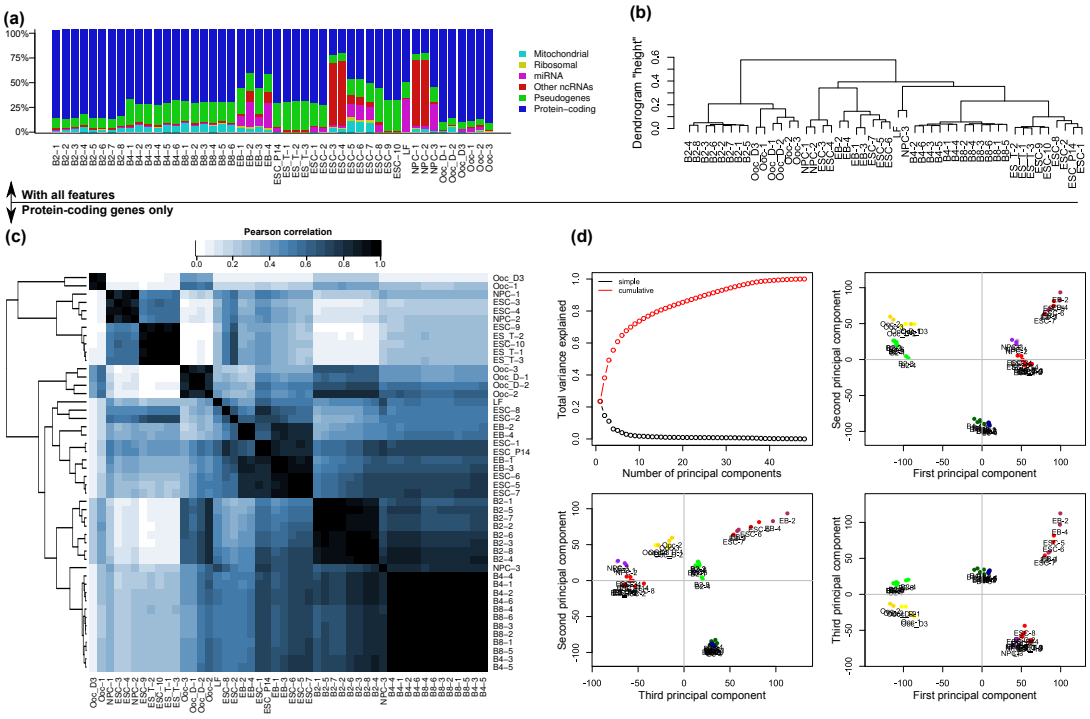

Figure 5.2: RNA-seq gene expression data. (**a**) Percentage of all quantile-normalised gene expression values (X_{qRPKM}) per feature type and dataset. (**b**) Dendrogram of correlation distances for all features clustered hierarchically by complete linkage. (**c**) Pearson correlation matrix clustered hierarchically by complete linkage for only the protein-coding features. (**d**) Contribution of individual datasets to the first three principal components (PCs). The first three PCs explain about 50% of the variation in the data.

This approach has been shown to be much more sensitive to non-coding RNAs, which might often be missed by conventional RNA-seq[98,211], explaining the distributional difference.

As a result, cluster analysis of the expression intensities obtained was strongly governed by "experiment-of-origin" rather than the "cell type-of-origin" (**Figure 5.2**.b). I expected that this imbalance would impair the latter analysis and therefore decided to focus only on the protein-coding portion of the genes annotated in the GeneProf reference dataset ($n_{protein-coding} = 22,806$ out of $n_{total} = 35,529$). I would like to stress that this is not due to a difference in the quality of the datasets *per se*, but rather due to a fundamental difference in the nature of the data studied. This difference makes it infeasible to compare both types of datasets across the board with the same measure without the use of some specialised normalisation technique – which is not within the scope of the current study.

Thus, I took from all datasets only the protein-coding genes and then repeated the quantile normalisation. Expression values (X_{qRPKM}) obtained in this way were generally better correlated between different samples representing the same cell type, although experiment-specific effects were still strong (**Figure 5.2**.c). Interestingly, though, principal component analysis of the signatures was able to distinguish the individual cell types rather well, regardless of technical differences (**Figure 5.2**.d).

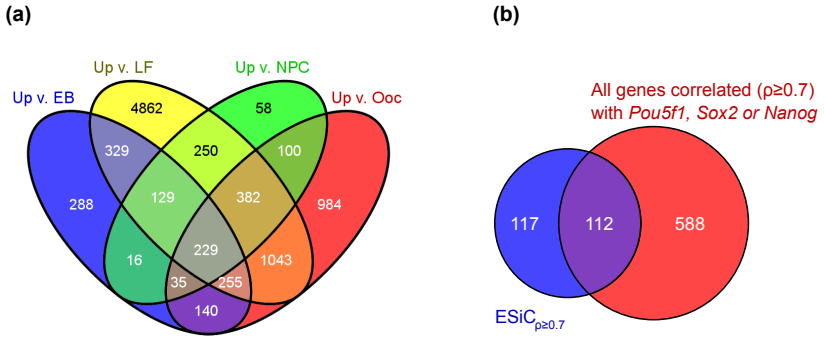

Figure 5.3: Overlaps of candidate genes. Venn diagrams demonstrating (**a**) the overlap between up-regulated genes in the different cell types as compared to ESCs and (**b**) the overlap between candidate genes (*ESiC*-1) highly correlated with either *Pou5f1, Sox2* or *Nanog* and all highly correlated protein coding genes.

I chose not to pursue these issues much further and instead decided to focus purely pragmatically on those genes that, despite all differences, could be clearly associated with ESCs. At this point, I was not really interested in an exhaustive list of all elements involved, but, on the contrary, preferred solely the strongest candidates, which I could be most confident about for the further analysis.

Therefore, the edgeR algorithm[458] was used to assess differential expression between

- all ESC samples and all lung fibroblasts (LF; 1 dataset),
- all ESC samples and all embryoid bodies (EB; 4 datasets),
- all ESC samples and all neural progenitor cells (NPC; 3 datasets),
- and all ESCs and all oocytes (Ooc; 3 datasets).

I called genes differentially expressed if they had an FDR-corrected p-value of $p \leq 0.1$ for EBs and oocytes and $p \leq 0.2$ for NPCs. A more permissive threshold was used for NPCs since I expected both undifferentiated cell types to be rather similar and to share candidate genes. For instance, *Sox2* is known to be expressed in NPCs, although at lower levels than in ESCs. The lack of replicates for LFs did not allow for meaningful statistical comparison, so I decided to use a fold change threshold of $|log_2 FC| >= log_2(1.5)$ for this comparison. I then took the overlap (intersection) of all gene lists obtained (**Figure 5.3**.a). It should be noted that only genes which were consistently up- or down-regulated in all comparisons were accepted.

I reasoned that genes discerned in such a way would be those that were involved in ESC-specific functions and not solely in the maintenance of generic progenitor states or early developmental mechanisms. Since I was primarily interested in genes closely associated with the core factors *Pou5f1*, *Sox2* and *Nanog*, I also calculated the Pearson correlation coefficient

between the expression signature of each of these genes and all other genes in the reference dataset and used these as a ranking criterion.

Not a single gene was expressed significantly higher in all other cell types as compared to ESCs. However, a number of genes was consistently over-expressed in ESCs throughout all comparisons ($n = 229$). I call those genes "ES-identity candidate genes" ($ESiC$). To confirm that the list did indeed contain genes relevant to stem cells, I characteristed the candidates in three ways:

- Almost half of all candidate genes (112 of 229, 48.9%) were strongly correlated ($\rho \geq 0.7$) with at least one of the core factors (**Figure 5.3**.b). This is a significantly higher proportion than in the entire dataset (700 out of $22,806$, 3.1%; hypergeometric $p(X \geq 112) \sim 7.9 \times 10^{-108}$). **Figure 5.4** shows all candidate genes with a high correlation to at least one core factor ($ESiC_{\rho \geq 0.7}$).

- Consistent with previous reports (cp. **Section 1.1.4**), the selected candidates include, besides the core factors *Pou5f1*, *Sox2* and *Nanog* themselves, genes such as *Zfp42*, *Nr0b1*, *Klf2/4/5/9*, *Lefty1/2*, *Tet1*, *Phc1*, *Fgf4/17*, *Esrrb*, *Dppa4/5a* and *Utf1* (**Section 1.1.4**). The list also includes many less well-studied genes, which will be discussed later (**Section 5.2.5**, **Section 5.3** and **Section 5.3.2**).

- Functional enrichment analysis with GOseq[620] yielded only four biological processes highly enriched ($FDR \leq 0.01$) in the candidates: "Stem cell maintenance" (GO:0019827, $FDR \sim 0$), "response to retinoic acid" (GO:0032526, $FDR \sim 0.0018$), "transcription" (GO:0006350, $FDR \sim 0.0095$) and "cellular zinc ion homeostasis" (GO:0006882, $FDR \sim 0.0095$).

As a side note to this analysis, I found it interesting to observe that there was globally a strong correlation between the transcriptional patterns of single oocytes and blastomeres of the 2-cell embryo, however, this global similarity appeared to be largely lost as early as at the 4-cell stage, so after one additional cell division. This observation is based solely on measurements from the same experiment[533] and using the same techniques, so is unlikely to be a mere artefact. On the other hand, 4- and 8-cell stage blastomeres became increasingly more similar to ESCs.

5.2.2 Genome-Wide Distributions Patterns of Regulatory Proteins and Histone Modifications

Before further investigating the regulatory dynamics described by DNA-binding proteins (DBPs) and histone modifications (HMs), I first looked on a broader scale and in an unbiased manner at the global binding activity of the different proteins. To do so, I first calculated

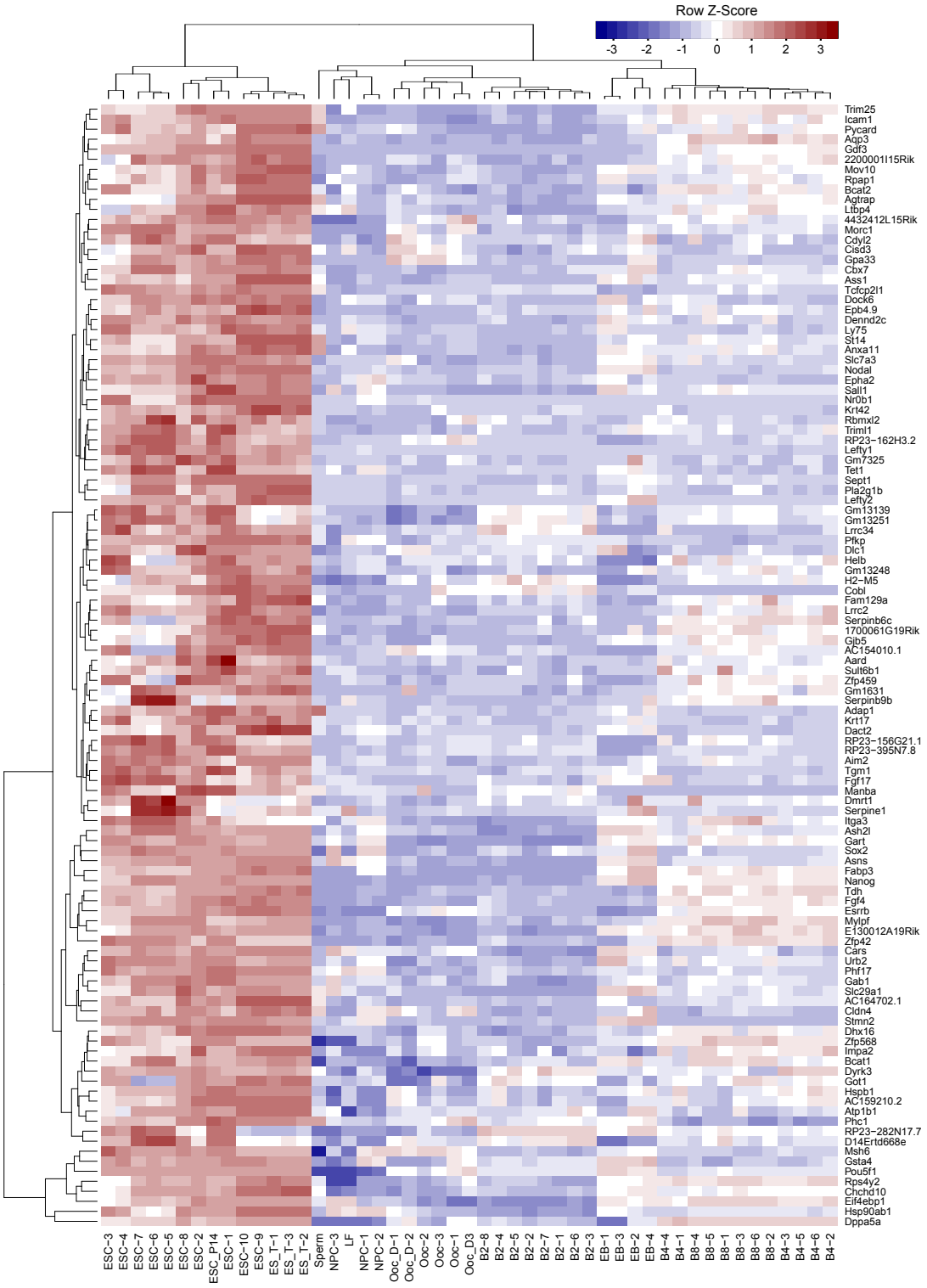

Figure 5.4: Clustered heatmap of ESC-identity candidate genes. The heatmap reports normalised (X_{qRPKM}), log_2-transformed gene expression values for assorted candidate ESC identity genes ($ESiC_{\rho \geq 0.7}, n = 112$) clustered hierarchically by complete linkage. Colours have been rescaled by row. Shades of blue indicate lower than average, shades of red higher than average expression.

the coverage of aligned ChIP-seq reads across the genome with respect to known genes in the GeneProf reference annotation, splitting reads into one of five categories:

- Intronic: In the intron of a gene.
- Exonic: In the exon of a gene.
- Promoter: Overlapping the promoter region of a gene, arbitrarily defined as the $1kb$ region surrounding the transcription start site (TSS).
- Near a gene: Within $50kb$[†] of the TSS or transcription termination site (TTS) of a gene.
- Not near any gene: None of the above.

The bulk of all aligned reads was, as expected, assigned to the largest categories, namely intronic and intergenic regions near genes (the majority of the mouse genome is in the proximity of at least one gene) and since the individual categories are of vastly variable size (number of bins: $exonic = 78,959; intronic = 947,372; promoter = 33,731; near\_gene = 1,466,149; not\_near\_gene = 769,599$), I normalised the counts for each category further by dividing them by the size of the category in order to get a better estimate of how the observed coverage relates to the expected coverage, if all regions of the genome were equally likely to be sampled (**Figure 5.5**.a).

One may conclude that a remarkably high number of all reads appeared to originate from genic regions and especially the promoters of known genes. *Ctr9* and *NelfA*, in particular, stood out from the profiles of the other proteins, since they seemed to be specifically enriched in exonic and promoter regions, respectively, which is in line with their expected function: *NelfA* (part of the NELF complex) coincides strongly with the initiation site of *PolII* transcription, where it prevents elongation when coupled with DSIF (containing *Spt5*)[437]. *Spt5* was also enriched at promoters, however, extended further into the gene. *Ctr9*, on the other hand, which is representative of PAF1, was enriched at the termination site of transcription and also present throughout the gene[437]. Several TFs were also enriched strongly at promoters: For example, *Myc* has been implicated in the same study in the release of *PolII* from the transcriptional pause[437]. The enrichment was less pronounced for other TFs, which might rather bind in distal enhancer elements, e.g. *Nr5a2*.

Next, I sought to look at the global similarity of the binding profiles of all proteins. I divided the genome into equally sized bins ($size = 1kb$) and summed up the number of reads falling into each of these bins. I then calculated the pair-wise Pearson correlation (ρ) between

[†]Note, I use a permissive window size of $50kb$ here first in order to get a coarse overview of the global binding patterns of all factors. In the following analysis I refine this initial impression by looking at the more detailed distribution of binding peaks with respect to the location of TSSs (**Figure 5.9** and **Figure 5.11**) and then finally decide to use a $20kb$ for the assignment of peaks to genes – a window size that attributes the majority of peaks to a target gene, but does not yet suffer too much from creating ambiguous assignments.

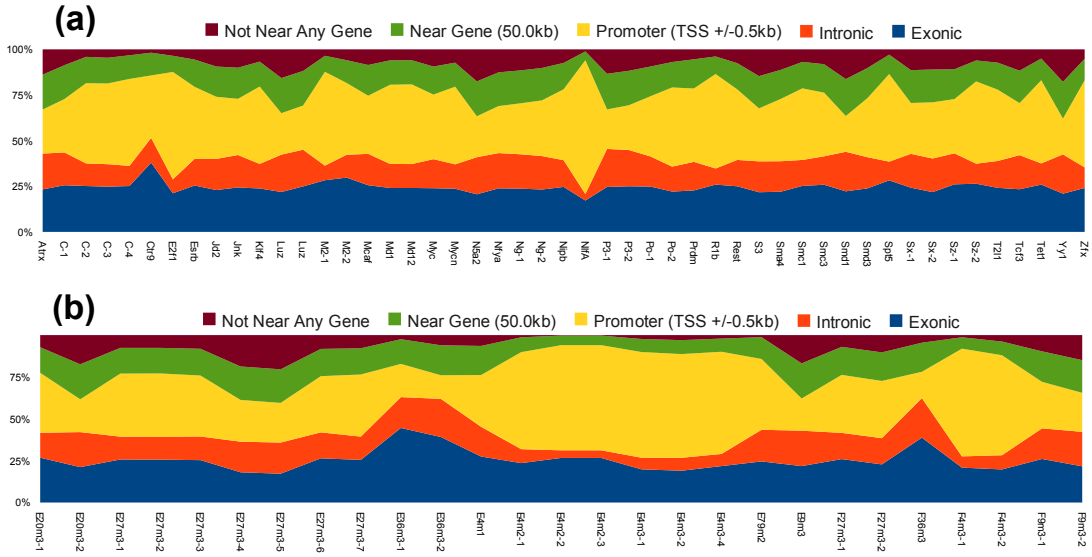

Figure 5.5: Global distribution of aligned short reads. Aligned read coverage was analysed with respect to known transcriptional features (genes, short RNAs, ..) and the total number of reads in each category depicted was summed up for each dataset. Shown are the d percentages normalised for variable category size for (**a**) various DNA-binding proteins in ESCs, (**b**) histone modifications in various embryonic cell types.

the bin counts of each combination of factors ("correlation matrix") and visualised the results as a heatmap (**Figure 5.6**). In order to more easily spot globally similar patterns, the heatmap was clustered hierarchically with average linkage defined on the Euclidean distance between correlation coefficients.

Generally speaking, the global patterns of all factors were positively correlated to some degree (average correlation $\hat{\rho} = 0.394$), indicating that probably a high fraction of genome-wide binding reported by ChIP-seq is due to genomic characteristics such as chromatin accessibility rather than the actual binding specificity of the protein in question. Datasets for the same or closely related proteins tended to cluster together (e.g. C-1 to C-4), although there were exceptions: Notably, datasets for the core pluripotency factors *Pou5f1* (Po-1 and Po-2) and *Sox2* (Sx-1 and Sx-2) did not cluster directly together, although their correlation was still reasonably high ($\rho_{Po-1/Po-2} = 0.586, \rho_{Sx-1/Sx-2} = 0.605$). Reassuringly, close clusters were also formed by different subunits of protein complexes: *Mtf2* and *Suz12* (PRC2) together with *Ring1b* (PRC1), *Med1* and *Med12* (mediator), *Smc1* and *Smc3* (cohesin). Interestingly, *Jarid2*, also PRC2-related, correlated more closely with a set of TFs rather than *Mtf2* and *Suz12*. TFs were in general closely linked in their genome-wide profile, with *Nanog, Tcf3, Sox2* and *Pou5f1* forming a particularly strong subunit. The last observation I would like to point out is, that while cohesin components *Smc1* and *Smc3* closely correlated with *Ctcf*, the strong correlation between cohesin was also detected for the promoter-linked mediator members *Med1* and *Med12*, but not so much for *Ctcf* and the mediator. It seems likely that a subset of genes might be occupied only by mediator and cohesin, which could be the active ones, while those

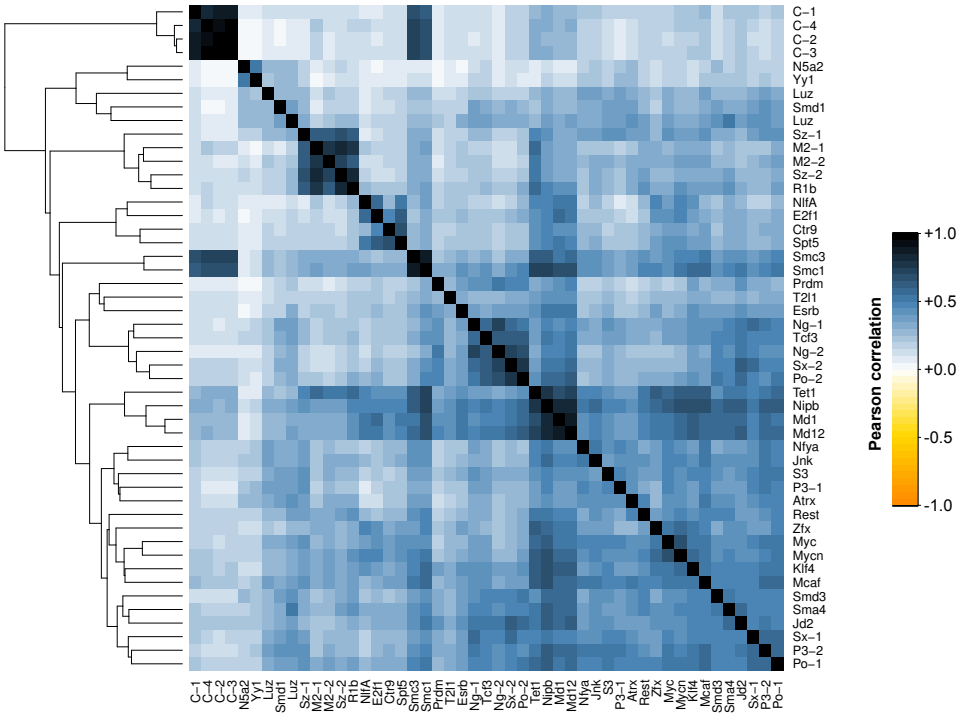

Figure 5.6: Correlation of genome-wide DNA-protein binding activity. Pair-wise Pearson correlation matrix of coverage counts across $1kb$-bins. Darker colors correspond to higher correlation. Factors were rearranged by hierarchical clustering with average linkage.

that also have *Ctcf* lack DNA-loop formation and mediator and are inactive[245].

I then repeated this analysis for the collections of HM data (**Figure 5.5**.b and **Figure 5.7**).

Trimethylation of lysine 27 as well as mono-, di- and trimethylation of lysine 4 of histone 3 appeared to be strongly enriched at the TSS of genes (**Figure 5.5**.b), consistent with their putative role in the activation and silencing of gene transcription (**Section 1.1.5.2**). This trend prevailed across both assayed cell types (ESCs and fibroblasts) and was largely consistent between datasets from different experiments. Methylation of lysines 9, 79 and 36 (especially the latter) and lysine 20 of histone 4, on the other hand, were less restricted to promoter regions and covered the entire gene body.

Clustering of the global distribution patterns confirmed that the major deciding factor for clustering is the type of HM profiled rather than the laboratory group that carried out the investigation (**Figure 5.7**). The distribution patterns of H3K4me2 and -me3 were generally closely correlated making up one major cluster together with H3K36me3 and H3K79me2 (the latter two forming a distinct subcluster). Monomethylation of H3K4, however, contributed to the other major cluster which was made up primarily of H3K27me3. Two H3K27me3 datasets, though, while still closely related with other data for the same HM, did not share the high similarity with modification patterns observed for other datasets. It is not clear whether this was due to technical differences or biological ones (e.g. due to the use of different cell lines).

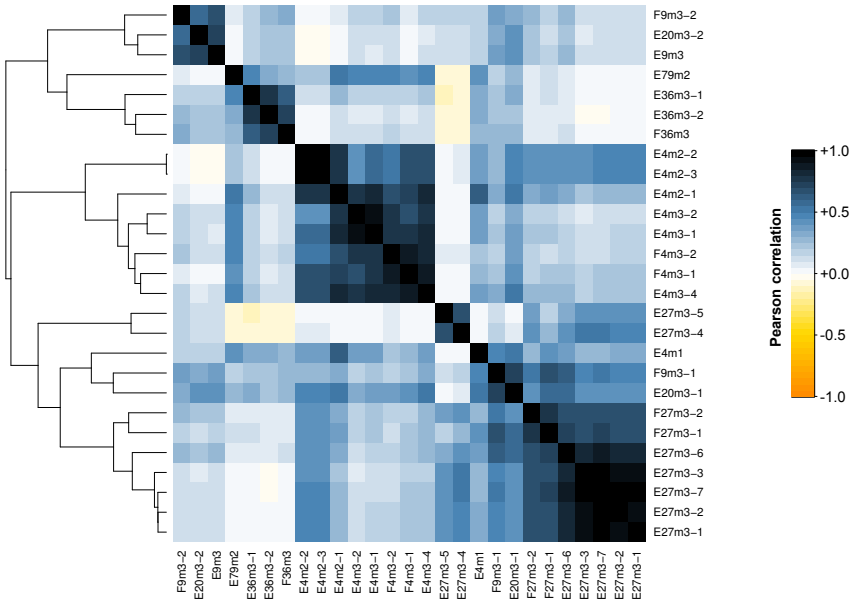

Figure 5.7: Correlation of genome-wide histone modification patterns. Pair-wise Pearson correlation matrix of coverage counts across $1kb$-bins. Darker colors correspond to higher correlation. Factors were rearranged by hierarchical clustering with average linkage.

The same two datasets showed evidence of a weak anti-correlation to the activating H3K36me3 and H3K79me2 marks.

The assignment of H3K9me3 and H4K20me2 was somewhat inconclusive since individual datasets were spread across the two major clusters. Nevertheless, I found it reassuring that the majority of related datasets clustered together and that the two main clusters corresponded to the functional distinction declared by the putative role of histone modifications marking active and inactive gene states, respectively.

5.2.3 Epigenetic State of Stem Cell Genes

I had noticed in the previous part of the analysis (**Section 5.2.2**), that certain HMs were mostly located at the TSS of genes, while others were spread more evenly across the entire gene body, that is promoter, exons and introns. As discussed in the introductory chapter of this thesis (**Section 1.1.5.2**), the presence of HMs is believed to correlate with, or even be causally involved with the activation and silencing of gene expression. I sought to examine HM patterns in more detail and decided to first have a closer look at the occupancy of the various modifications at some assorted gene loci (*Pou5f1, Sox2, Nanog, Fgf4* and *Cdx2*), where I examined the coverage of HMs within a genomic context of $14.0kb$ centred on the gene (**Figure 5.8**) using the genome browser built into GeneProf.

Interestingly, H3K36 trimethylation – thought to be a mark of active gene transcription – could be found strongly associated with chromatin around the genes *Pou5f1, Sox2, Nanog* and *Fgf4*, all of which are expressed in stem cells, but not near *Cdx2*, a differentiation marker not

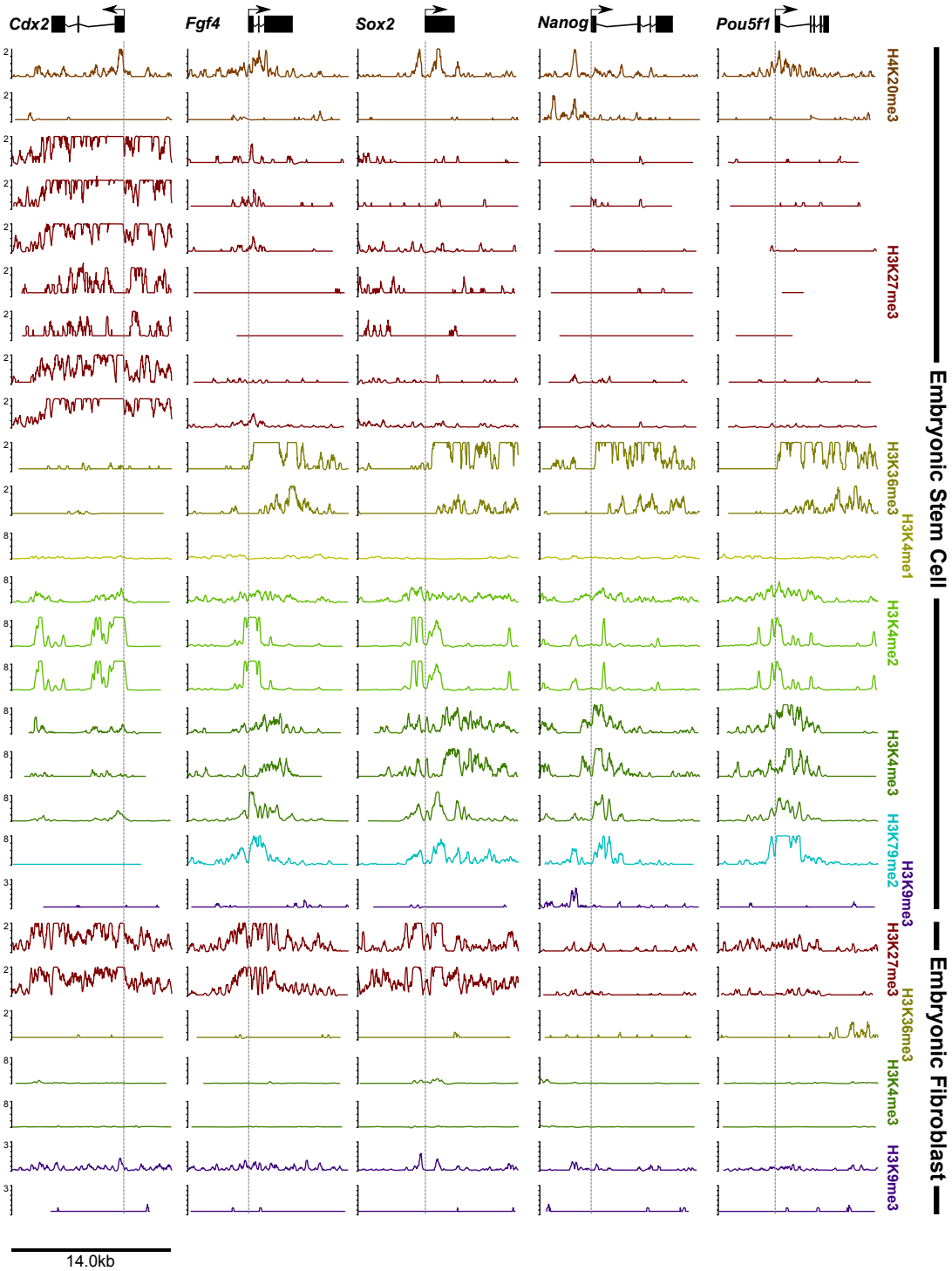

Figure 5.8: Detailed view of histone modifications in the genome. Post-processed graphics exported from GeneProf. Shown are coverage patterns of aligned ChIP-seq reads at five selected genomic loci. Track heights have been normalised in such a way that the height of each track corresponds to the number of reads per million aligned reads at each given position of the genome.

expressed in self-renewing ESCs (**Section 1.1.4**). H3K36me3 covered the entire gene body, starting from the promoter region and reaching often beyond the TTS. No H3K36me3 was detected for any of these genes in EFs, in which none of the genes are expressed.

Distinct differences in coverage between active and inactive genes could also be observed for H3K79me2 and H3K4me3, but not H3K4me2 or -me1. Trimethylation appeared to be strongly associated with active promoters, but sometimes reached far into the gene body (see *Sox2*: **Figure 5.8**) and it is lost entirely in EFs.

In contrast, repressive H3K27me3 was clearly preferentially associated with inactive genes (*Cdx2*) and was gained for genes silenced in differentiated cell types, in particular, *Fgf4* and *Sox2*, but less pronounced for *Pou5f1* and *Nanog*.

As in the previous analysis (**Section 5.2.2**), measurements for H3K9me3 and H4K20me3 were somewhat inconclusive on this small scale. H4K20me3 seemed to be associated with promoters (and possibly enhancers, see upstream of *Nanog*) in active as well as inactive genes and H3K9me3 signals were overall weak and no clear pattern stood out in this view.

In order to examine whether these observations held up on a global scale, I used the SISSRs peak detection algorithm[242] to search for "peaks", that is, regions of the genome that showed a statistically significant enrichment for any one HM in at least one of the datasets as compared to a control signal (**Section C.1**). Peaks for the same HM were then iteratively merged by joining together any peaks that were within $100bp$ of each other in order to define a comprehensive list of modification sites throughout the genome.

Examination of the number (**Figure 5.9**.a) and genomic location of these modification sites with respect to the closest annotated gene (**Figure 5.9**.b and **Figure 5.9**.c) showed that particularly many modification sites were found for H3K36me3, which covers broad regions across whole genes, and H3K4me2, which shows small, rather well-defined peaks that would not have been merged into larger groups by the iterative clustering strategy. Genome-wide, the majority of H4K20 and H3K4 di- and trimethylation was concentrated at the promoters of known genes, while H3K36 tri- and H3K4 monomethylation was observed throughout the gene body. A large part of H3K9me3 happened outside of genic regions in the upstream area of genes (possibly linked with enhancers) or in gene-remote regions.

A close look at the distribution of peaks with respect to the TSS of the next-closest gene (**Figure 5.9**.c), revealed distinct patterns for each modification: While all HMs were centred on the TSS[‡], I found it especially interesting to observe that H3K79me2 accumulated slightly downstream of the TSS with decreasing amounts detectable further into the gene body. H3K4 mono-methylation was slightly depleted at the TSS, probably due to an enrichment of di- and trimethylation of the same lysine (mutually exclusive with mono-methylation) at the same

[‡]This is partially due to a bias of the analysis that links peaks to the next TSS and will hence prefer assignments towards the TSS-centre of the plots. Nevertheless, true biology overrules this bias and distinct differences in patterns are clearly visible.

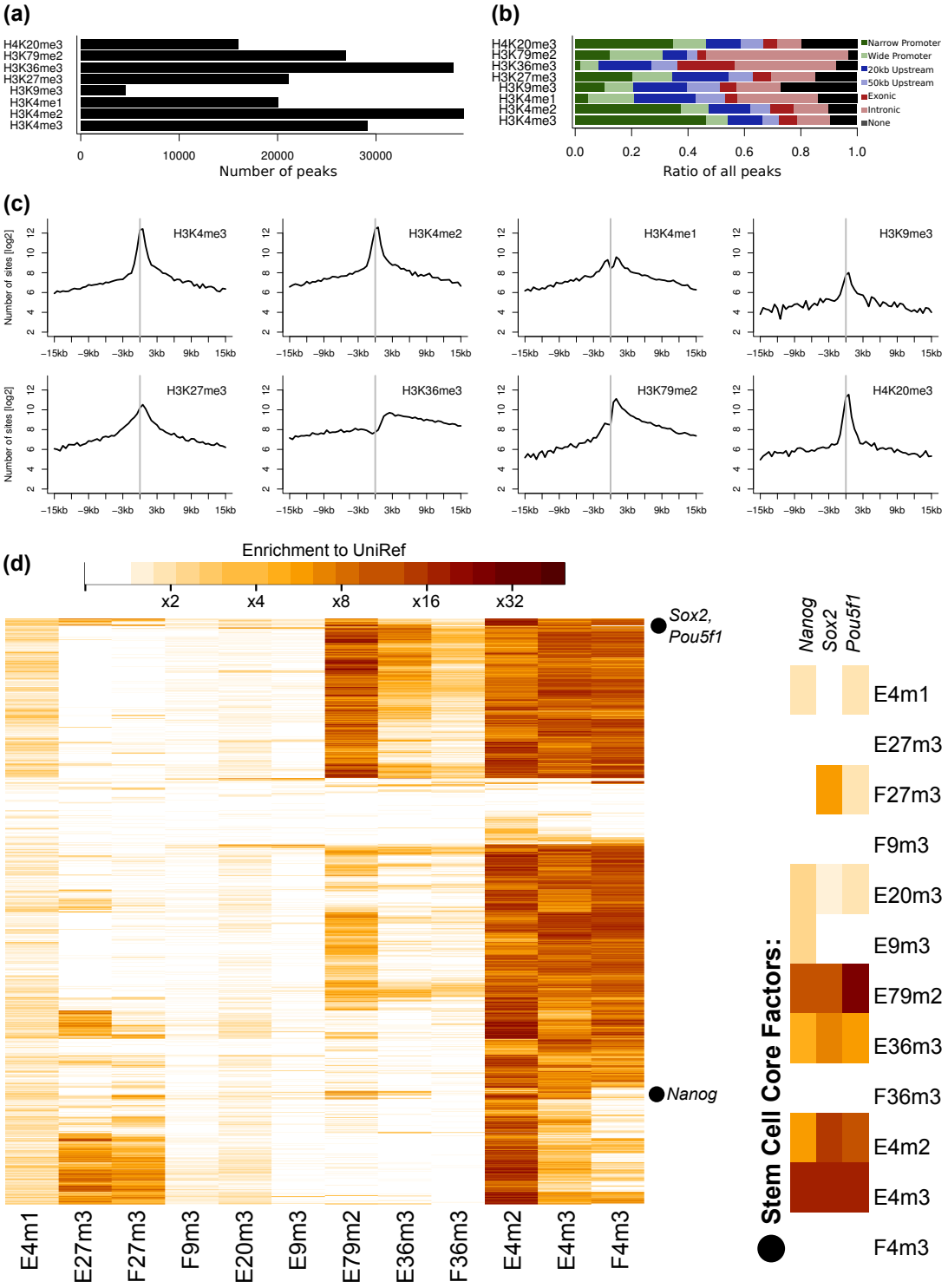

Figure 5.9: Genome-wide histone modification signatures. (**a**) The number of sites enriched for eight types of HMs. (**b**) Global distribution of HM sites with respect to the closest annotated gene. Narrow promoter: TSS ±0.5kb, wide promoter: TSS ±2kb, none = not near a known gene. (**c**) Detailed distribution of HM sites with respect to the TSS of the closest gene. (**d**) The heatmap reports log_2-fold changes (compared to UniRef) between normalised coverage intensities for HM clusters associated with all protein-coding genes with at least one cluster was assigned ($n = 16,871$). Rows and columns have been reordered by hierarchically clustering the intensities (with complete linkage) using the Euclidean distance for rows and Pearson correlation distance ($(1 - \rho)/2$) for columns.

1700067K01Rik	B4galnt4	Epcam	Insig1	Ndufs2	Rpl26	Slc7a7
2700062C07Rik	Bcar1	Esrrb	Lif	Nodal	Sall1	Spp1
Acy1	Calr	Fat1	Lin28	Notch3	Sall4	Trim71
Adam23	Ccnb1	Fgf4	Lipt1	Nphs1	Samd1	Tyro3
AL596446.7	Cdh1	Fgfr1op	Mast1	Prmt3	Sap25	Utf1
Anapc5	D10Wsu102e	Gm4767	Mcm3	Psmd7	Setd1b	Zfp64
Arhgap26	Ech1	Hjurp	Mrpl45	Pttg1	Setdb1	Zic3
Atxn2l	Elmo3	Hmgn2	Mycn	Rcor2	Slc12a7	

Table 5.4: Genes sharing common histone signature with *Pou5f1* and *Sox2*. Genes hierarchically clustering by HM signature together with the two core factors.

sites.

Peak calls are subject to many, largely arbitrary decisions and I sought to minimise the effect of thresholding by working with quantitative intensity values rather than qualitative peak calls alone. Using the previously defined modification sites, I quantified the amount of aligned reads falling into any one region and subsequently rescaled these counts to account for differences in sequencing library size (reads per million, RPM; **Section 3.3.3.3**). I then calculated the logarithmic (log_2) fold change with respect to the background signal ("enrichment"). Where multiple samples were available for the same HM in the same cell type, measurements were averaged for the sake of easier interpretation.

Generally speaking, I observed high levels of H3K4 methylation at a large proportion of all genes (**Figure 5.9**.d). I also noticed a strong concordance of intensity levels for H3K4 di- and trimethylations – in both cell types. For a subset of genes (including *Pou5f1, Sox2, Nanog*), H3K4me3 was abolished in fibroblasts and only a small number of genes appears to gain stronger H3K4me3 in fibroblasts. H3K4me1 was present at a similar set of genes as di- and trimethylation, however, at lower levels.

Those genes with the strongest H3K4me2, tended to be also strongly occupied with H3K27me3. However, H3K27me3 appeared to occupy a lower number of genes than H3K4. The methylation patterns of H3K27 for most genes were largely identical in ESCs and EFs, but several clusters of genes existed for which it was observed at either increased (including the stem cell core factors) or decreased levels in EFs.

Confirming my earlier observations, H3K36me3 was inversely correlated to H3K27me3, with genes that were highly trimethylated at H3K27 in ESCs being less strongly methylated in EFs and vice versa. Again, this held for the stem cell core factors as well as a cluster of other genes.

The genes with the strongest H3K36me3 in ESCs were also occupied by H3K79me2. Unfortunately, no data was available to confirm this trend in EFs.

H3K9me3 was rarer – or, at least, less often associated with genic regions and only a small number of genes showed noteworthy presence of this modification in ESCs, but levels were overall much higher in EFs in a pattern that appeared to be closely related to H4K20me3 and

also H3K4me1.

Shifting the focus to the genes (putatively) affected by the HM patterns described above, I noticed that the three core factors were somewhat separated in their epigenetic profile: *Pou5f1* and *Sox2* closely mirrored each other's profile, but *Nanog* showed a distinct pattern. Closer examination of other genes clustering alongside *Pou5f1* and *Sox2* yielded 57 candidates (**Table 5.4**), many of which had a known implication in stem cell characteristics, e.g. *Utf1*, *Sall1/4*, *Lin28*, *Nodal*, *Mycn*, *Notch3*, *Esrrb*, *Fgf4* and *Lif*. More genes showed a signature similar to *Nanog* ($n = 1,307$ at a similar clustering height) including *Dnmt3l*, *Chd7*, *Dppa2/3/4/5a*, *Eras*, *Kit*, *Lefty1/2*, *Nr0b1/2*, *Zfp42* and many more.

5.2.4 Control of Stem Cell Genes by Groups of Regulators

In the next step of the analysis, I applied a similar methodology as previously used for HMs to all DBP datasets. To get an impression of the nature of data I first examined the binding profiles of all proteins at a selection of genomic loci surrounding genes of particular interest using the GeneProf genome browser. As an example, I show here the binding profiles at five gene loci (*Pou5f1, Sox2, Nanog, Fgf4* and *Cdx2*; **Figure 5.10**). In the figure, only one sample (the first in **Table 5.1**) is shown for those DBPs where multiple datasets were available. The order and colouring of the individual tracks reflects the results of a similarity clustering performed at a later stage of the analysis (**Figure 5.14**).

Generally speaking, TFs tended to bind overlapping regions of the genome either near the promoter of (putative) target genes or at distinct regions that might serve as enhancers[75,245]. For instance, there are two rather well described enhancer regions upstream of *Nanog*, both of which clearly stood out in the binding profiles (**Figure 5.10**), with evidence of binding for *Pou5f1, Sox2, Nanog, Ep300, Nr5a2, Tcfcp2l1, Esrrb, Prdm14, Tcf3* and other TFs as well as elements of the transcriptional machinery (*Med1/12, Smc1/3, Nipbl, Spt5*).

Mtf2, Suz12, Jarid2 and *Ring1b* were associated with inactive genes (e.g. *Cdx2*), where they might facilitate the repression of the expression of those gene. In contrast, *Ctr9, Spt5, NelfA, Myc, Mycn* and others were closely linked to transcriptionally active genes. These are all proteins that are either components of the RNA polymerase machinery or crucial to its functioning, so they are indeed functionally linked to active transcription.

To study DBP profiles on a global scale, I looked for binding events that were enriched in comparison to the UniRef control (**Section C.1**) using SISSRs[242]. The individual DBPs occupied a vastly variable number of sites (**Figure 5.11**.a), ranging from only several hundred ($n_{Yy1} = 480$) to tens of thousands ($n_{Esrrb} = 76,727$).

I associated each of the detected peaks with the closest known gene and recorded the peak-to-TSS distance as a categorical value (**Figure 5.11**.b). This analysis confirmed the ob-

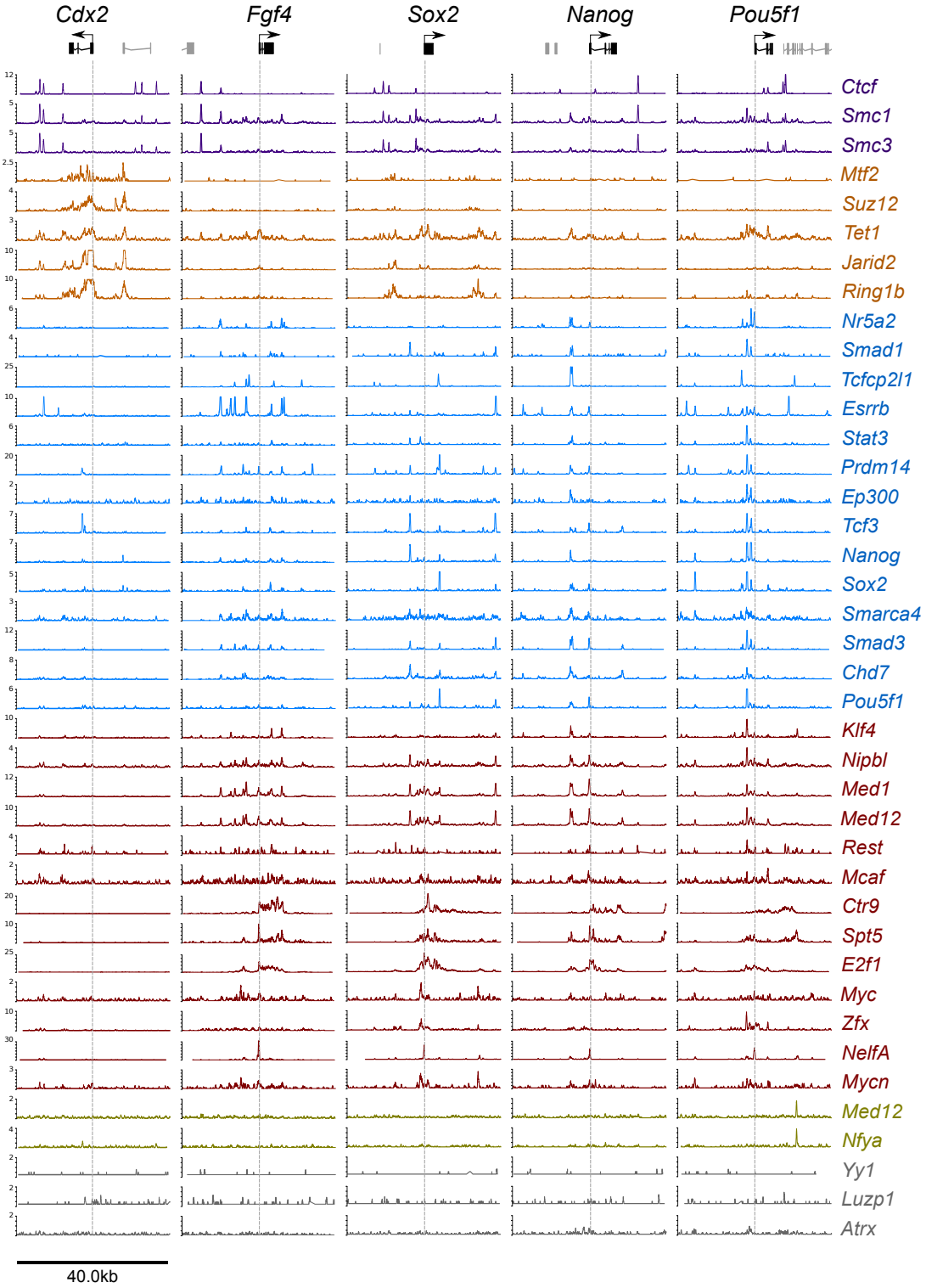

Figure 5.10: Detailed view of DNA-protein binding in the genome. Post-processed graphics exported from GeneProf. Shown are coverage patterns of aligned ChIP-seq reads at five selected genomic loci. Track heights have been normalised in such a way that the height of each track corresponds to the number of reads per million aligned reads at each given position of the genome.

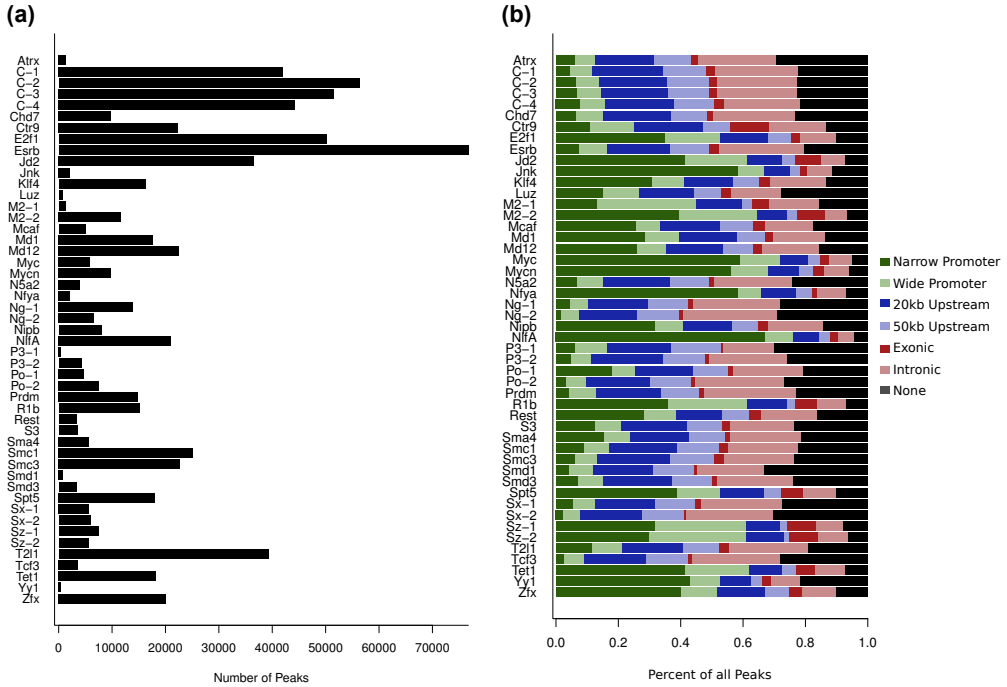

Figure 5.11: Number and distribution of DNA-protein binding sites. (**a**) The number of sites enriched for protein-to-DNA binding in 49 datasets examined. (**b**) Global distribution of these enriched binding sites with respect to the closest annotated gene. Narrow promoter: TSS ±0.5kb, wide promoter: TSS ±2kb, none = not anywhere near a known gene.

servations drawn from the previous small-scale examination of binding profiles (**Figure 5.10**). For instance, DBPs that are either direct members or functionally linked to the immediate control of polymerase activity were clearly clustered at the promoters of genes (e.g. *NelfA, Nipbl, Ring1b, Suz12, Myc, Mycn, Jarid2, Jnk1/3*). Subtle differences in the binding patterns were revealed by a closer look at the exact distance of the peaks with respect to the TSS of the closest genes (**Figure 5.12**): For example, one could see that, even though all peaks were centred on the TSS of this meta-genic profile, some proteins showed a preferential bias to the upstream region immediately adjacent to the TSS. This is consistent with the traditional model of how TFs might bind upstream of promoters to recruit polymerase or initiate transcription and held up most clearly for *Chd7, Nr5a2, Nanog, Pou5f1, Smarca4, Smc3* and *Sox2*. Proteins forming part of the transcriptional apparatus, *Ctr9* and *Spt5*, on the other hand, were clearly enriched downstream of the TSS where active transcription by polymerases was taking place (it appears that the data was capturing transcription as it happened at various places throughout the gene).

In order to examine the putative co-occupancy of DBPs genome-wide[§], I merged binding

[§]Since all ChIP-seq experiments have been performed on different populations of cells, one cannot say for certain that any of the DBPs or HMs mentioned in this analysis ever physically co-occur at the very same sites in the genome. One way of resolving the question whether two proteins do indeed physically co-occupy binding sites is the use of sequenctial ChIP[71,141,355,554], however, not enough large-scale data was available for me to use at the time when I performed the analysis presented here.

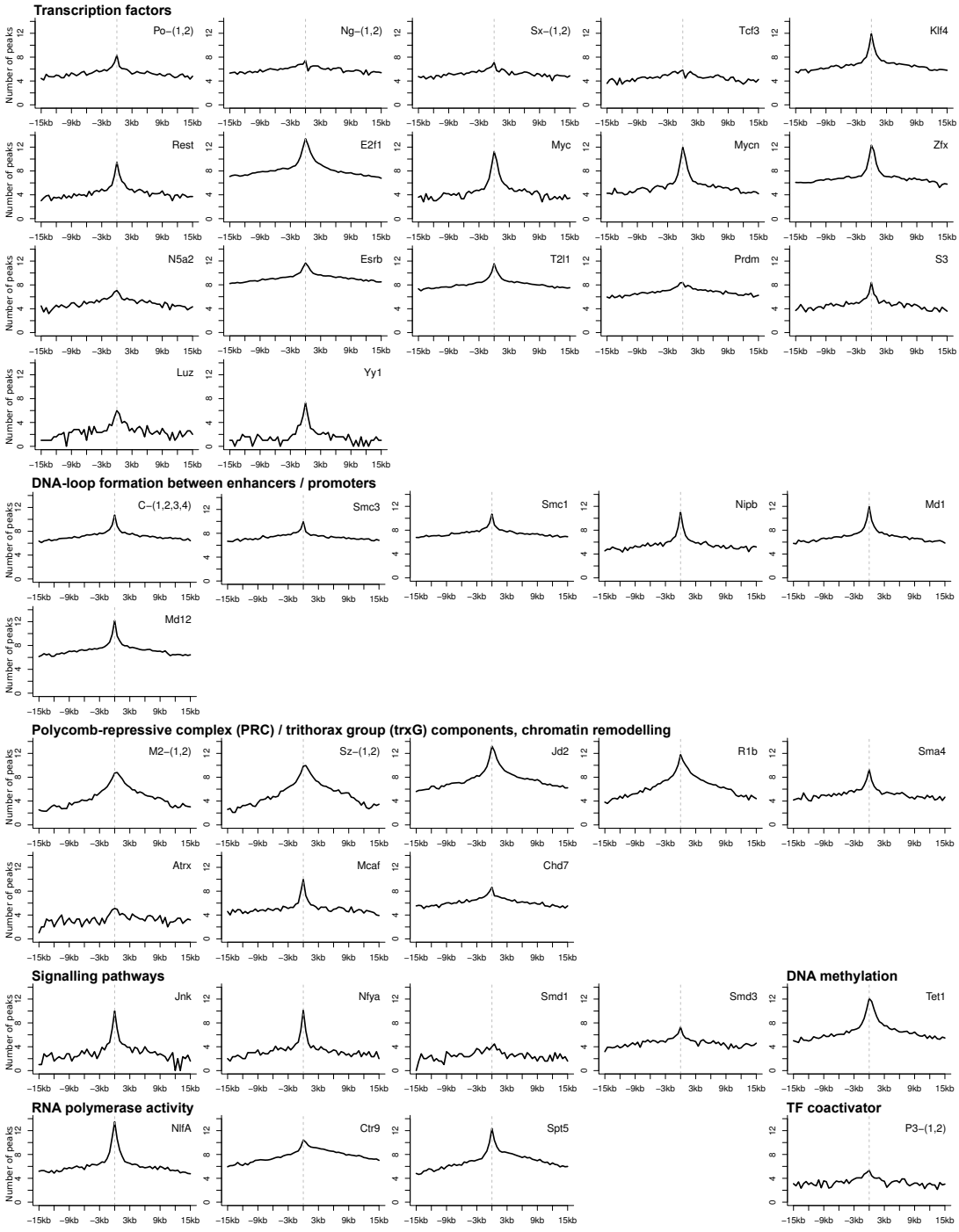

Figure 5.12: Distribution of DNA-protein binding sites near TSS. Binding sites were assigned to the closest neighbouring gene and the peak-to-TSS distance was recorded (rounded to $0.5kb$ accuracy). The plots show the frequency with which peaks were detected within a given proximity of the TSS (dashed line). Where multiple datasets for the same protein were available, the plot shows the average (arithmetic mean) of all measurements. The numbers on the y-axis are log_2-scaled.

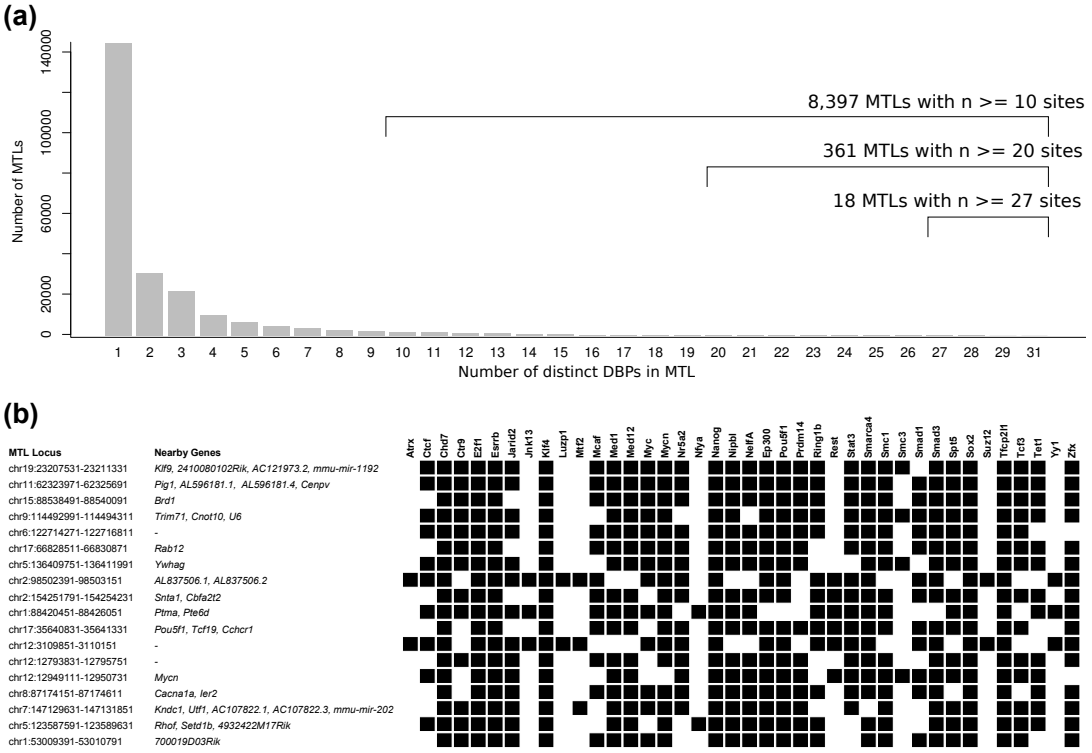

Figure 5.13: Frequency of putative DBP co-occupancy in MTLs. Peaks for all individual datasets were merged into "multiple transcription factor-binding loci" (MTLs)[75] if they were within 100*bp* of each other. (**a**) The plot shows the count of MTLs (y-axis) that incorporated a given number of binding peaks (x-axis). (**b**) A list of the 12 MTLs with more than 30 constituting peaks.

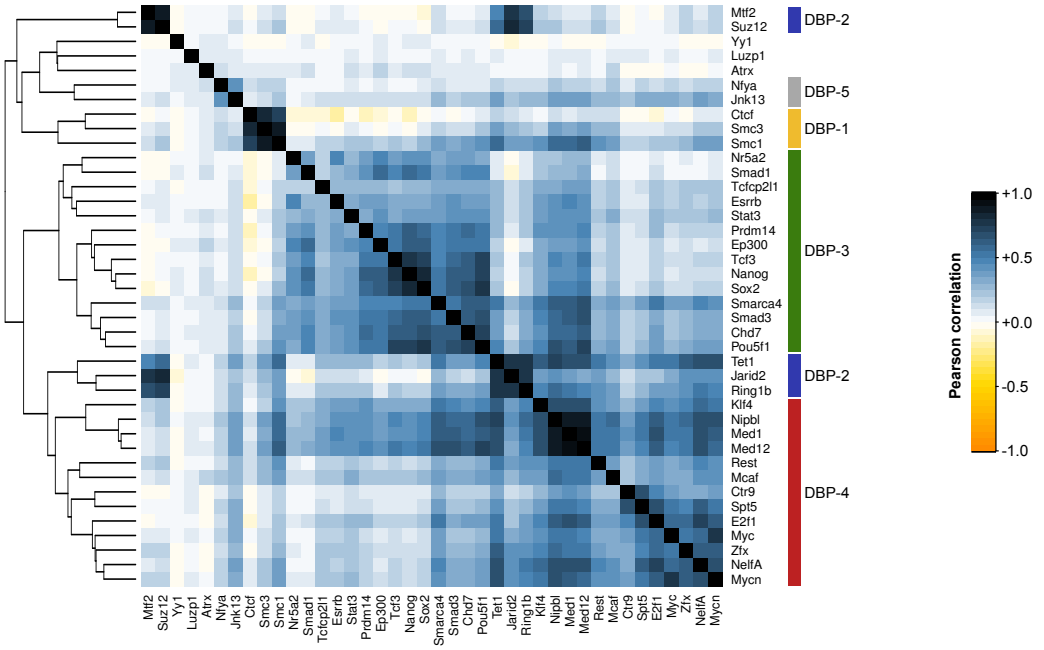

Figure 5.14: Cooccupancy patterns of DNA-binding proteins. Pair-wise Pearson correlation was calculated on the basis of the binding enrichment of each DBP across all MTLs. MTLs were generated by merging the peaks for all datasets that were within 100*bp* of each other. Rows and columns have been reordered by hierarchical clustering with complete linkage.

peaks located within at most 100*bp* of each other iteratively to form so called "multiple transcription factor-binding loci" (MTL)[75] and checked how frequently how many DBPs shared the same bound locus. Overall, the majority of loci was occupied by one factor alone, but a surprisingly high number of sites (38.7%) showed evidence for binding by several factors at once (**Figure 5.13**.a), with 8,397 MTLs bound by 10 or more different DBPs and 10 even by at least 28 DBPs. It seems plausible that much of this binding is indicative of cooperative (or antagonistic) functional relevance. Interestingly, one of the MTLs with the highest co-occupancy of DBPs was situated near *Pou5f1* (**Figure 5.13**.b).

Next, I went ahead to assess the global correlation of binding intensities across all detected peaks. During many previous analyses (in the process of populating the GeneProf databases; **Chapter 4**) I had noticed that the binding of individual factors at places enriched for the binding of another might sometimes not be sufficient to be called a "peak", yet the measured intensities tended to be stronger for related factors across the board. I therefore quantified the number of aligned reads for each dataset in each MTL, rescaled the intensities to account for differences in library size (reads per million), calculated the enrichment to the control (logarithmic fold change floored at 0) and calculated the global correlation between all datasets (**Figure 5.14**). The results of this analysis confirmed my suspicions: Nearly all DBPs were positively correlated to a considerable degree. The only notable exception to this phenomenon was *Ctcf*, which globally correlated strongly only with *Smc1* and *Smc3* and was actually anti-

correlated to the binding intensity of a number of TFs.

Interestingly, DBPs clustered in a way related to their functional similarity (**Section 1.1.4**). Five major groups were identified (although the boundaries are fuzzy!):

DBP-1: *Ctcf, Smc1, Smc3*: The Cohesin members are involved in DNA-loop formation connecting active enhancers to the core promoters[245], while *Ctcf* acts (amongst other roles) as a transcriptional insulator partly also via DNA-loop formation[280,424].

DBP-2: *Mtf2, Suz12, Tet1, Jarid2, Ring1b*: *Mtf2* and *Suz12* are both subunits of PRC2 and involved in the repression of gene expression[61,574]. Although visually distinct in the clustered heatmap, I also added *Jarid2* and *Ring1b* to this group (also PRC members), because they were very highly correlated to the former two. *Tet1*, which converts 5-mC to 5-hmC, although apparently not directly involved in PRC has previously been reported to bind to many of the same targets[592]. The latter three, and in particular *Tet1*, were also closely linked to the binding of proteins in group *DBP*-4 and some members of *DBP*-3.

DBP-3: *Nr5a2, Smad1, Tcfcp2l1, Esrrb, Stat3, Prdm14, Ep300, Tcf3, Nanog, Sox2, Smarca4, Smad3, Chd7, Pou5f1*: This group consists mainly of TFs, including the core factors. It also includes the co-activator *Ep300*, which has previously been reported to co-occupy many active enhancers[75] and *Smarca4*, which together with *Stat3* opens chromatin rendering the enhancers accessible to TF binding[200].

DBP-4: *Klf4, Nipbl, Med1, Med12, Rest, Mcaf, Ctr9, Spt5, E2f1, Myc, Zfx, NelfA, Mycn*: Mediator complex and polymerase-associated proteins that are present at active promoters[245,437]. The group also contains TFs that appear to be very closely linked to the presence of these proteins at the promoters, suggesting a more direct link to the regulation of transcription than those in group *DBP*-3. The first four DBPs were also correlated with the previous group (*DBP*-3), but the latter are less so.

DBP-5: *Nfya, Jnk1/3*: Both proteins are involved with chromatin remodelling and opening up chromatin at promoters. They have both been previously observed to cluster closely together at promoters with a role in differentiation[548]. The proteins clustered loosely with group *DBP*-1 and also showed a considerable correlation with promoter-associated components in *DBP*-2 and *DBP*-4.

The remaining proteins (*Yy1, Luzp1, Atrx*) did not closely correlate with any other, which I believe is mainly due to an overall weak binding intensity. It is not clear whether this lack of binding signal was due to a technical weakness (inability to detect the binding) or due to the biology of those proteins (genuinely low number of bound regions in the genome).

So far I have focused solely on the markup of the binding profile of a multitude of DBPs, but neglected how this binding relates to putative downstream effectors. Let us now shift the focus to a target gene-centric view. Associating enriched binding sites with potentially regulated genes is a matter of some controversy (cp. **Section 3.3.3.5**), but most published

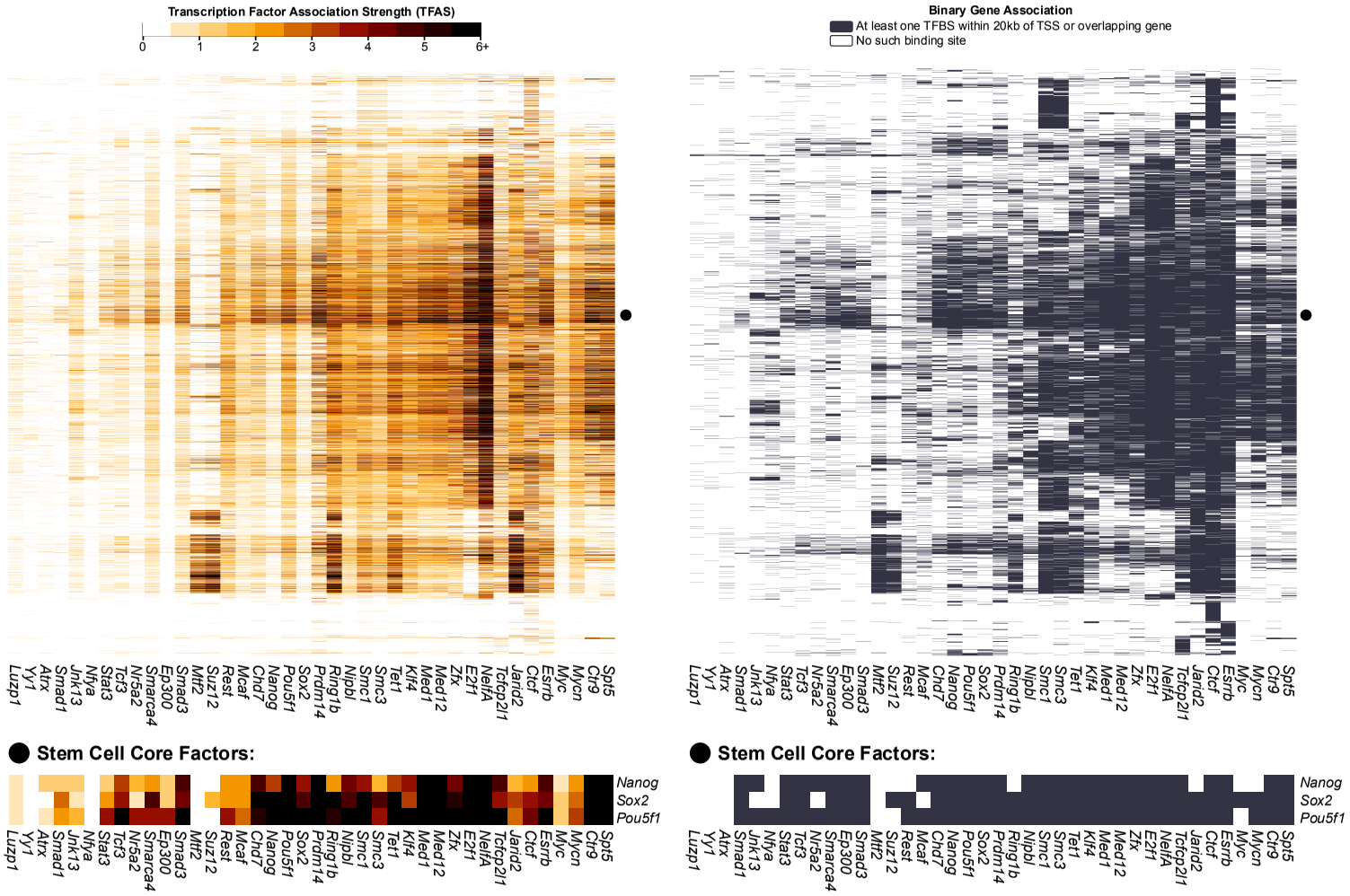

Figure 5.15: Co-occupancy patterns of DNA-binding proteins. (**a**) Heatmap of TFAS scores for 40 DBPs. Darker colours correspond to a higher score, i.e. stronger presence of the DBP in the proximity of the TSS of the said gene (one row per gene). Rows and columns have been reordered according to the clustering in (**b**), which shows the binary assignment of enriched binding sites to the same genes. Rows and columns have been reordered by hierarchical clustering with binary distance and average linkage.

research relies on a binary assignment of binding sites to genes. That is, a gene is called a target of a certain DBP, if there is a binding site for this factor somewhere within a fixed window surrounding the TSS of this gene. The choice of window size varies between publications from narrow ranges ($TSS \pm 0.5kb$) via those presuming binding matters only in the upstream region of genes ($TSS - 20kb$) to those that allow binding in a huge neighbourhood ($TSS \pm 100kb$). Some researchers associate only the closest known gene with a binding site, others choose to link all genes within the window to a binding site.

There is no definitive answer as to what is the best way of proceeding in this issue and it seems likely that thresholds indeed depend on the protein under study. For the purposes of this study, I chose to assign a peak to every gene for which it was either (a) within a window size of $20kb$ to either side of the outer-most TSS of a gene or (b) anywhere within the gene's body (introns or exons). Having examined the binding patterns of many TFs and other regulatory proteins throughout the genome (in the context of this analysis and of other work I have been doing before), this seemed a reasonable choice capturing the majority of characterised enhancer activity as well as those proteins exerting their function directly in the gene body. Nevertheless, I acknowledged that a simple binary assignment would miss certain functional links between DBPs and effector genes and I also applied an alternative strategy assigning to each gene-DBP combination a continuous score called the "transcription factor association strength"[406] (TFAS; **Section 3.3.3.5**). To recapitulate briefly, the TFAS takes all binding peaks within a huge range ($1mb$) surrounding the TSS into account and sums up the intensity of the binding observed in this peak weighted indirectly proportional to the distance of the peak to the TSS. Thus, genes that have many strong peaks close to their TSS will rank higher than those with only weak or remote peaks.

While the use of the TFAS scores overcomes the necessity for fixed cut-off thresholds, the issue remains that a binding site located in the proximity of a gene might not be regulating this target gene. Binding sites might instead be regulating genes that are much further away and possibly with other genes in between[280]. The combination of ChIP-seq experiments with targeted loss-of-function studies for the same DBPs may help to establish a better link between binding sites and the genes they regulate, but matched expression data is not yet widely available and even if it was, the assignment of peaks to target genes would still be hindered by second-order effects (the loss of expression of a transcription factor is likely to trigger a cascade of effects on the expression of other genes via intermediaries) and by the dependence of DBPs on co-factors and other influences (a binding site could be functionally regulating a target gene, but only if all other given determinants of regulation were available at the same time). Despite all given limitations, TFAS scores have been demonstrated to correlate reasonably well with gene expression[406], so overall this way of assigning binding events to their likely transcriptional targets appears to be valid.

2410080I02Rik	Bcat2	Gm7325	Klf2	Pycr2	Socs3	Zfp553
2410137M14Rik	Bcl3	Gpa33	Klf3	Rest	Spry2	Zfp57
4932422M17Rik	Capns1	Gpx4	Lefty1	RP23-117P3.3	Spry4	Zic3
6430527G18Rik	Cbx7	H2-M5	Macf1	Sall1	Tdh	Zscan10
9630014M24Rik	Cldn4	Hsd17b14	Mkrn1	Sall4	Wbscr27	
AC101915.1	Dusp27	Ifitm1	Mycn	Scd2	Zbtb45	
AC133494.1	Fam100b	Igfbp2	Mylpf	Sept1	Zbtb8a	
Agtrap	Fbxo36	Jam2	Nodal	Sgk1	Zfp13	
Arhgap26	Gemin7	Jarid2	Plekha4	Slc29a1	Zfp296	

Table 5.5: Genes sharing common regulator characteristics with *Pou5f1, Sox2* and *Nanog*. Genes hierarchically clustering by TFAS signature together with the three core factors.

The results of both analyses are summarised in **Figure 5.15**. Note that both heatmaps have been reordered by the similarity of rows and columns in the binary profile to facilitate comparability. There was a large number of genes with evidence for binding by many different factors. I was particularly interested in the small number of genes ($n = 61$), that had a signature of putative regulators very similar to *Pou5f1, Sox2* and *Nanog*. The list contained many genes previously implicated in ESC identity (in either a supporting or disrupting manner), e.g. *Jarid2*, *Klf2/3*, *Lefty1*, *Mycn*, *Nodal*, *Sall1/4* and *Rest* (**Table 5.5**). Those genes appear to be controlled by a shared set of regulatory inputs and it would be plausible to believe that they might also be functionally related, making even the less well-known members of the list interesting candidates for stem cell research.

5.2.5 Many Stem Cell Genes Share a Common Regulatory Signature

Finally, I meant to put the results of the previous analyses together to unravel regulatory signatures common to genes that are central to ESC identity (*ESiC* gene list, **Section 5.2.1**). In order to make it possible to compare values from DBPs and HMs in the same analyses, I first standardised all intensities calculated previously by subtracting the mean of each measurement and dividing by the standard deviation (zero-mean and unit variance normalisation).

I first meant to examine how the intensity of DBP and HM occupancy related to transcriptional activity in ESCs on the whole. Comparing the frequency with which all DBPs/HMs of a certain intensity occurred in all protein-coding genes that were transcriptionally active in ESCs ($X_{qRPKM} \geq 5, n_{active} = 7,375$) with those that were inactive ($X_{qRPKM} < 5, n_{inactive} = 15,431$), showed up notable differences in distribution, in particular, for various histone modifications (**Figure 5.16**.a): As fits well with our current model of their functional role, genes with a high level of H3K79me2, H3K36me3 and H3K4me3, were clearly enriched in the active subset of genes. Less pronounced, but still notable, the same held for H4K20me3 and H3K4me2, while high levels of H3K27me3 were only very rarely observed with active genes.

Many DBPs also showed differential patterns between active and inactive genes. Sensibly, these are proteins linked to polymerase and the transcriptional machinery (*Ctr9, Spt5, Smc1,*

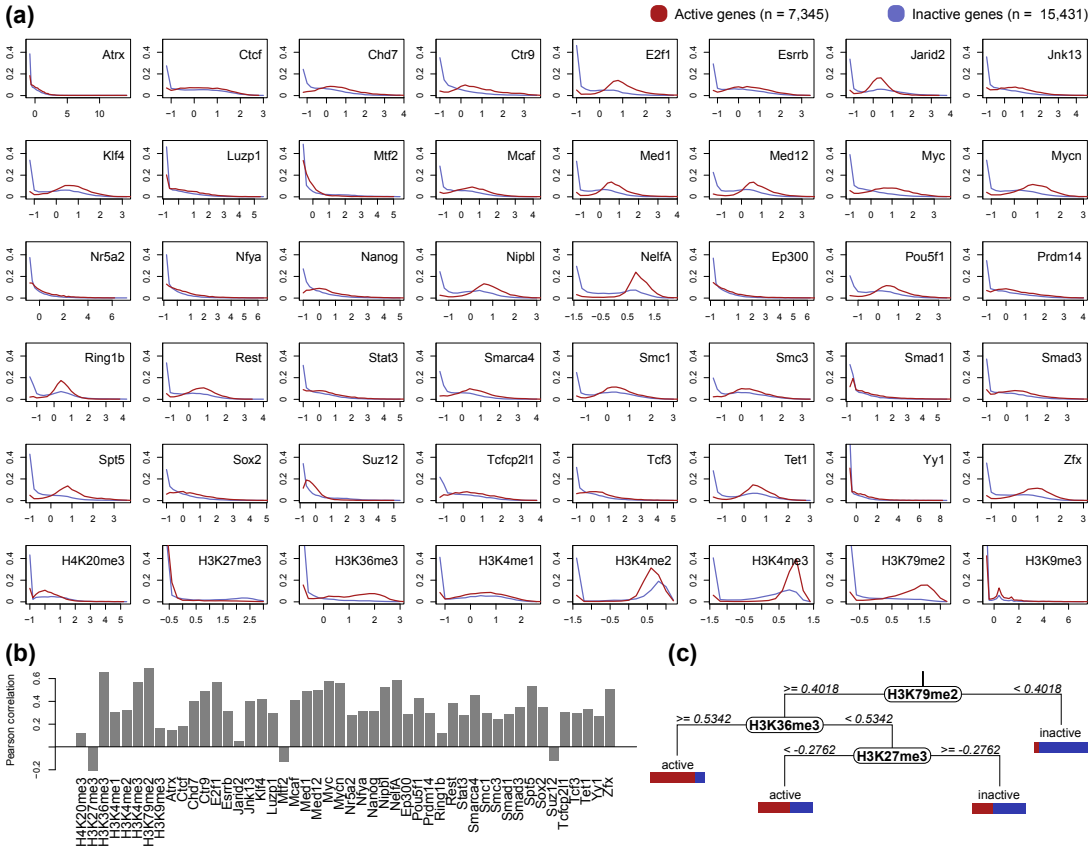

Figure 5.16: Regulatory signature of active and inactive genes. (**a**) Percentage of genes (y-axis) with a certain standardised intensity value (x-axis) for each DBP and HM. (**b**) Pearson correlation of HM and DBP intensities with average gene expression values (X_{qRPKM}) in ESCs. (**c**) Decision tree discriminating active from inactive genes based on the intensity of three HMs. The coloured bars report the percentage of genes from each category falling into the respective branch.

Smc3, Med1, Med12, Nipbl, NelfA): Above-average levels of binding for these proteins were preferentially associated with active genes. Many TFs, while present at all levels of intensity in active and inactive genes, were hardly ever found at the lowest observed intensity in the context of inactive genes (*Esrrb, Chd7, Klf4, Myc, Mycn, Nanog, Pou5f1, Prdm14, Smarca4, Tcfcp2l1, Zfx*).

The only DBPs clearly enriched in inactive genes were *Suz12* and *Mtf2*. Others with a supposed repressive function (*Ring1b, Rest, Jarid2*), were still preferentially associated with active genes, although only at modest (approximately average) levels of intensity. Global pair-wise Pearson correlation analysis of gene expression intensities with DBP/HM occupancy (**Figure 5.16**.b), also confirmed that most proteins were correlated positively to some degree with expression levels, however, only H3K36me3 and H3K79me2 at a strong level ($\rho > 0.6$). The only factors showing up a global negative correlation were H3K27me3, *Mtf2* and *Suz12*, but in all cases this correlation was rather weak. Importantly, this surprising observation does not necessarily contradict established models of the function of these proteins, but only goes

to show that the control of transcriptional activity depends on the complex interactions of a plethora of different factors.

I used linear discriminant analysis (LDA; as implemented in the `lda` function from the *MASS* package in R^{568}) to identify those variables (HM/DBP intensities) that were most conclusive for the distinction between active and inactive genes. This analysis returned H3K79me2 and H3K27me3 as the best discriminators (data not shown). However, none of these variables alone was sufficient to distinguish both classes: That is, even for H3K79me2 and H3K36me3, there were many inactive genes that had a high intensity.

I hypothesised that a combination of multiple variables might be able to discriminate active from inactive genes more successfully than a single factor alone. Therefore I tried to define a set of simple, (human-) understandable rules by which one could effectively distinguish both classes. A machine learning approach for determining such rules is given by so-called "decision trees" and I attempted to build such a tree using the *rpart* package in R^{543}. Based on the data at hand, the algorithm identified H3K79me2, H3K36me3 and H3K27me3 as the best discriminators (**Figure 5.16**) – consistent with my previous observations. Taking only the measurements for these three HMs into account, it was possible to distinguish active from inactive genes with high accuracy ($A = 0.844$) and precision ($P = 0.755$)[¶].

Evidently, the distinction made by this simple decision tree was still not perfect. I expect a large proportion of erroneous class predictions to be due to imperfect measurements and biological variation. That is, the HM, DBP and gene expression intensities used here are averages over a number of biological replicates (in themselves mixtures of heterogeneous cell populations). However, the unaveraged measurements within these classes are not always consistent – indeed, often they vary massively (**Figure 5.2**, **Figure 5.6**, **Figure 5.7**). This is due to technical measurement errors and the fact that the datasets were generated in different laboratories using a variety of cell lines, treatments and culture conditions. Hence the biology I am trying to model with this classifier is certainly not perfectly represented by the data that was available to me. Consequently, one could never expect a perfect discrimination to be achieved by the decision tree.

Given that there was a difference in regulatory and epigenetic markup between active and inactive genes in general, I now wanted to test whether there was a unique DBP/HM signature marking the 229 ES-identity candidate genes ($ESiC$) identified in **Section 5.2.1** on the basis of their gene expression patterns in different cell types. The majority of those candidates (226 of 229) were also "active" in ESCs according to the previously used criteria ($X_{qRPKM} \geq 5$)[‖],

[¶]The terms "accuracy" and "precision" are used in the sense in which they are generally defined in the field of machine learning. Accuracy is the ratio of correct classifications (true positives and true negatives) in the entire population. Precision is the ratio of true positives divided by all positive calls, here, the number of genes correctly called "active" divided by the number all genes predicted "active" (including those wrongly called "active").

[‖]The three genes that did not satisfy the "active" criterion were: *Olfr957, Sult6b1, Ankrd3*.

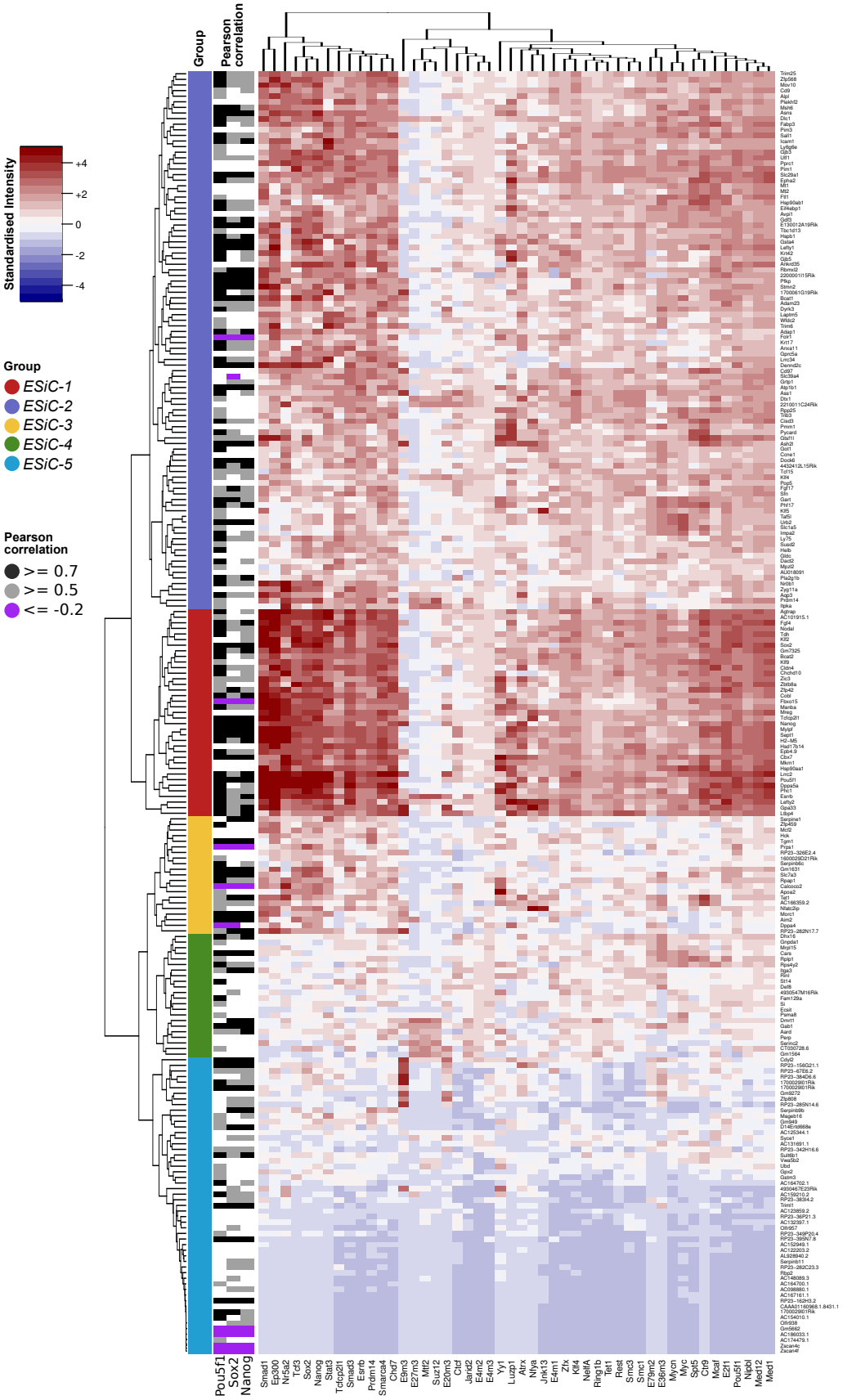

Figure 5.17: Regulatory signatures of ESC candidate genes. The heatmap reports standardised intensity values for HMs and DBPs for 229 selected ES-identity candidate genes (*ESiC*). Rows and columns have been reordered by hierarchical clustering with complete linkage.

so the task was not only to find a signature that marked active genes in ESCs, but distinguish a certain subgroup of active genes from rest of the transcriptome. If such a signature existed, it would be suggestive of common regulatory mechanisms driving the expression of an ESC gene transcriptional network. Even if no unique signature was shared across all candidate genes, there might be a subgroup of tightly co-regulated core elements of this network.

In order to look for a common regulatory code shared across ESC genes, I integrated all sources of data from the previous analyses (gene expression: **Section 5.2.1** , HMs: **Section 5.2.3** and DBPs: **Section 5.2.4**). Within these datasets, I concentrated on only the predefined candidate genes (*ESiC*). Hierarchical clustering of the regulatory signatures distinguished five distinct subgroups of genes within the candidate set (**Figure 5.17**):

ESiC-1 contained the three core factors as well as many other genes with a definite implication in ESC function, for instance, *Klf2, Tcfcp2l1, Phc1* and *Lefty2*. The cluster was marked by a high binding intensity across most DBPs and all "active" histone marks. On the other hand, intensities for the repressive histone mark H3K27me3 and PRC-members *Suz12* and *Mtf2* were low, with the exception of *Esrrb*, which also exhibited a comparatively high signal for these proteins.

ESiC-2 contained further stem cell genes (e.g. *Klf4/5, Nr0b1* and *Utf1*) and, like *ESiC*-1, showed evidence for binding of most HMs and DBPs. However, signal intensities were generally weaker and repressive influences were not always absent. Further subgroups might be distinguished in this large group, but I chose to leave this for later investigations.

ESiC-3/4/5: *ESiC*-3 still showed medium-intensity binding for many TFs (*Sox2, Nanog, Tcf3, Nr5a2, Smad1* and the co-factor *Ep300*), but weaker intensities for the other DBPs. It contained some previously characterised genes like *Dppa4* and *Tet1*, but also others which still need further investigation. *ESiC*-4 had even weaker signals for most DBPs and *ESiC*-5 had hardly any noteworthy evidence of binding – neither by DBPs nor by associating HMs. Interestingly, the last group contained almost only badly studied transcripts and during further investigations I have found that the lack of DBP/HM-signals for these genes might be explained by the repetitiveness of the genomic regions they are situated in (**Section C.2**).

I have noticed that many HMs and DBPs tended to show a higher propensity of strong signals in the candidates as compared to the rest of the transcriptome (**Figure 5.18**.a) and even in comparison to all active genes (data not shown for the sake of brevity). *Chd7, Esrrb, Klf4, Mcaf, Med1, Med12, Nipbl, Pou5f1, Prdm14, Smc1* and *Smad3* had visually clearly distinguishable patterns in both populations, however, again no single epigenetic or regulatory signal was powerful enough to discriminate all 229 *ESiC* genes from all other genes. I attempted to create a decision tree that would support the understanding of the separation between *ESiC* genes and other genes active in ESCs, but found the results too complex to give any insight into the biological nature of the difference (data not shown). This is not

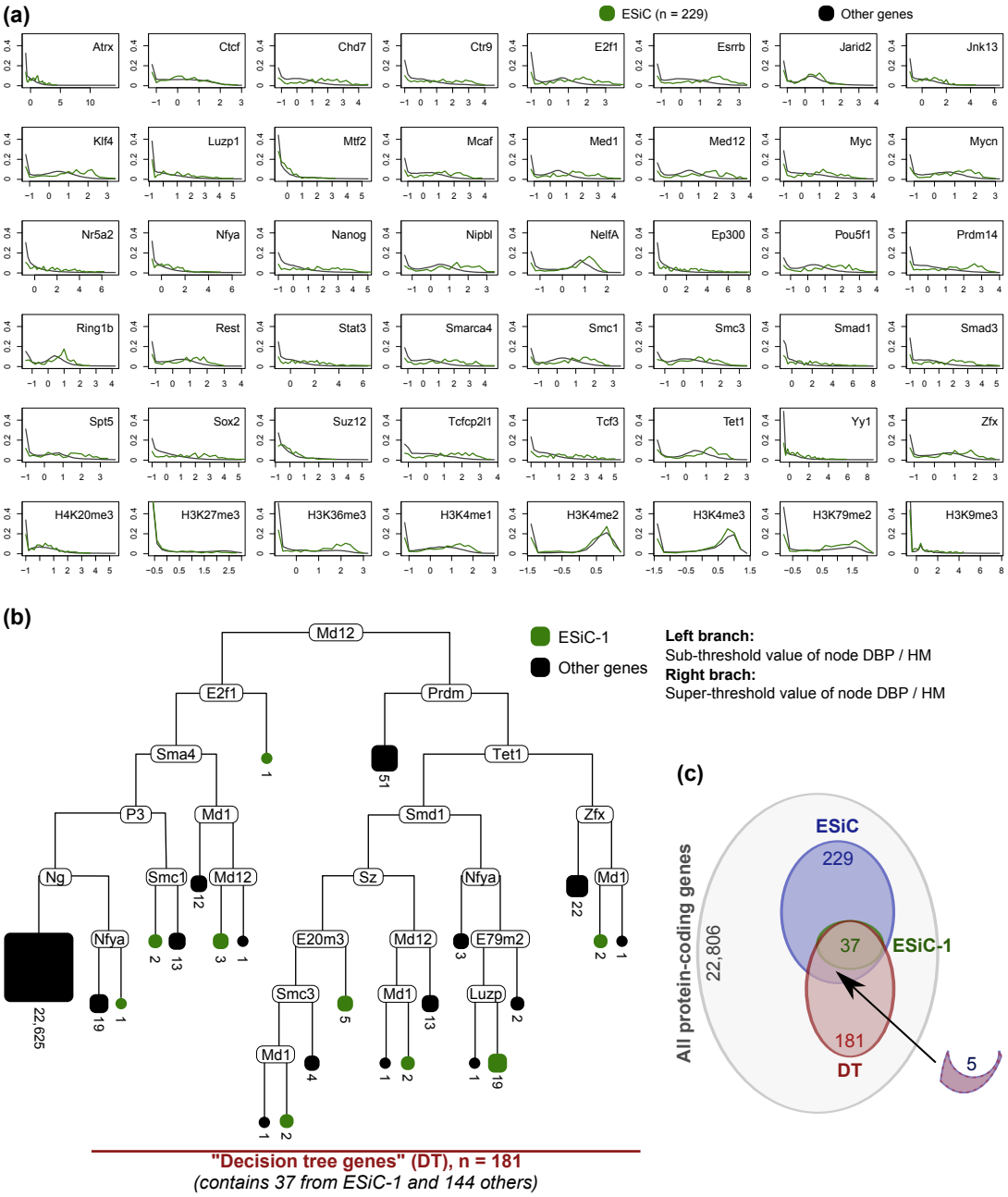

Figure 5.18: Regulatory signature of ESC genes. (**a**) Percentage of genes (y-axis) with a certain standardised intensity value (x-axis) for each individual DBP and HM. (**b**) Decision tree discriminating candidate genes from gene list *ESiC*-1 (see text) from all other genes on the basis of the intensity of DBPs and HMs. The decision tree has been restructured for easier interpretation in such a way that the left branch always means "low value of the decision node" and the right branch "high value", respectively. The endpoint correspond in colour to the group called (green: *ESiC*-1, black: other) and are in size roughly proportional to the number of genes included. (**c**) Venn diagram illustrating the overlap of the core group of the highest-confidence candidate genes (*ESiC*-1, green) with all ES-identity candidates (*ESiC*, blue) and all genes in the right branches of the decision tree in the previous panel (DT = decision tree, red). Five genes are in *ESiC* and *DT*, but not in *ESiC*-1.

so surprising given the drastic observable disparity between the groups identified from the hierarchical clustering (**Figure 5.17**) and the lack of signal for a subset of candidate genes (*ESiC*-5 and possibly more: **Section C.2**). Instead, I decided to focus, for the time being, entirely on *ESiC*-1, that is, those 37 genes for which I thought it might be most reasonable to expect an important, direct functional role linked to stem cells.

Again, I trained a decision tree on the distinction between this group (*ESiC*-1) and all other genes (**Figure 5.18**.b). The generated tree was able to achieve a perfect distinction between both classes ($A = 1.0, P = 1.0$). The vast majority of non-*ESiC*-1 genes ($22,620$ of $22,806$ protein-coding genes, 99.2%) could be distinguished from *ESiC*-1 by the application of just five decisions: High levels of *Med12, E2f1, Smarca4, Ep300* and *Nanog* were necessary for the inclusion in *ESiC*-1. These five decisions make up the left-most "branch" of the decision tree. The other decisions in the tree are only required to distinguish the remaining 144 "other" genes from *ESiC*-1 (**Figure 5.18**.c). These genes might be interesting in themselves thanks to their similarity to *ESiC*-1. In fact, five of them (*Epha2, Slc29a1, Utf1, Asns, Ftl1*) were also in the wider candidate list (*ESiC*), but had not been included in *ESiC*-1 by the hierarchical clustering (**Figure 5.17**).

The decision tree provides a way to distinguish *ESiC*-1 genes from others conclusively and without error with a minimum number of decisions. The tree basically states that genes with low levels of five key proteins (*Med12, E2f1, Smarca4, Ep300* and *Nanog*; left-most branch) are definitely not members of the *ESiC*-1 group. However, it would be a mistake to extrapolate this rule to say that all *ESiC*-1 genes had high levels of all those proteins. This is because second- and third-order decisions in the left branch of the tree do not pertain to the genes in the right-most branches (and vice versa). Consequently, for the genes in the right-most branch, one could not make any statement about *Nanog* levels, for example. Thus, the decision tree cannot help us to find a common DBP/HM profile for all *ESiC*-1 genes and to understand the co-regulatory mechanisms that coordinate those genes.

An alternative strategy was employed to find common characteristics of *ESiC*-1 genes: I defined the "discriminative power" $P(V|G_x)$ of a variable V (HM/DBP measurement) with respect to a group of genes $G_x \subset G$ as the percentage of non-G_x ($G_{other} = \overline{G}_x$) genes that could be discarded if a threshold Θ on the measurements for this variable ($m_V(x)$) was to be used. As a threshold, either the minimum measurement in G_x is used (Θ_{min}), implying that all passing genes need to be greater or equal to this threshold, or alternatively the maximum measurement in G_x (Θ_{max}), implying that all passing genes need to be less or equal to this threshold. The discriminative power is hence defined as:

$$P(V|G_x) = \frac{max(|\{x|x \in G_x \wedge m_V(x) \geq \Theta_{min}\}|, |\{x|x \in G_x \wedge m_V(x) \leq \Theta_{max}\}|)}{|G|}$$

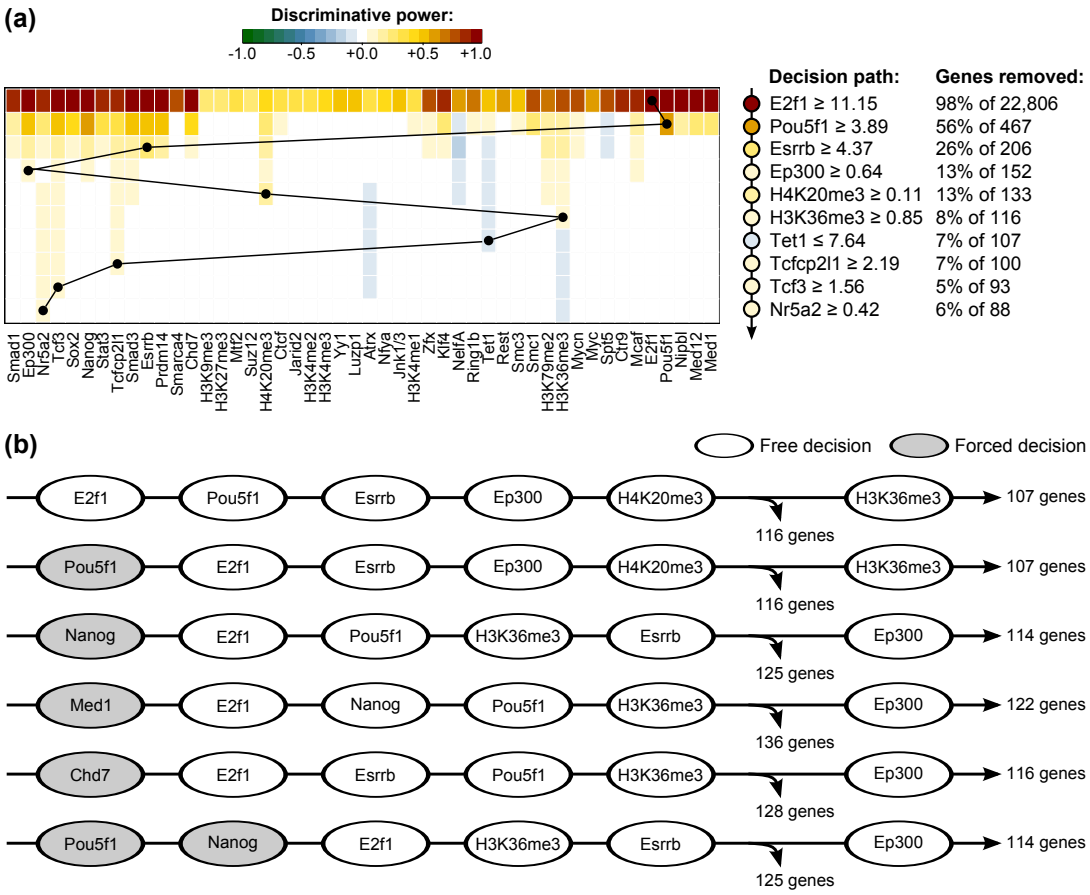

Figure 5.19: Common regulatory inputs of *ESiC-1* genes: (**a**) The plot visualises the potential of DBPs/HMs to distinguish $ESiC$-1 genes from others at 10 iterations of the decisions process (top to bottom). The absolute value of the discriminative power represents the percentage of non-$ESiC$-1 genes that can be removed by imposing a threshold on the given variable. Positive values mean that a high measurement of this variable is required for inclusion in $ESiC$-1, negative values require a low measurements. (**b**) Examples of how the decision path (first 6 steps) changes after manually forcing the use of certain variables (forced variables have gray background colour). Shown are also the numbers of genes after 5 and 6 decisions.

To distinguish "greater than" and "less than" decisions, I denote those decisions that require a measurement to be below the threshold as negative numbers. Using the discriminative power $P(V|ESiC\text{-}1)$, one can decide on a "decision path" discriminating $ESiC$-1 from other genes using an iterative strategy: For the first decision, the variable with the highest (absolute) discriminative power is chosen. After removing all genes that do not satisfy the threshold used for the first decision, the discriminative power for all variables is calculated on the remaining genes and again the most discriminative variable is chosen. The procedure will be repeated until no non-ESiC-1 genes are left or the selection of genes does not change any more.

The results of the iterative discriminative power analysis (IDPA), are shown in **Figure 5.19**.a. Interestingly, the very first decision already discards 98% of non-$ESiC$-1 genes. IDPA has revealed *E2f1* as the most decisive factor at this step: All $ESiC$-1 genes as well as 430 other genes have a very high level of this DBP associated with them. A number of alter-

native variables would be able to achieve a similar split: *E2f1* is closely followed by *Pou5f1, Chd7, Ep300, Med1, Nanog, Sox2, Tcf3* and *Med12* (in order), all of which could remove more than 95% of non-*ESiC*-1 genes in the very first decision. The second decision is based on the core pluripotency factor *Pou5f1*, that manages to further separate out 56% of the remaining non-*ESiC*-1 genes. Alternative decisions at this stage could be using *Nanog, Tcf3, Esrrb* or *Ep300*.

Why are there alternative decision variables? There is a certain level of redundancy between the genes marked by the different DBPs/HMs: For instance, genes that have a high level of *Pou5f1* also tend to have a high level of *Sox2*. I tried to examine the effects of changing the decision variables in the first step by enforcing the use of a sub-optimal variable, e.g. *Pou5f1, Nanog, Med1* and *Chd7* in the first stage or *Pou5f1* and *Nanog* in the first two stages (**Figure 5.19**.b). *Pou5f1* and *E2f1* were entirely interchangeable in the first two stages. Even changes to the other variables did not have any strong effect on the signature: The order in which some variables were chosen in the IDPA procedure changed, but the decision paths still maintained the same components and led to similar selections of genes. This demonstrates the robustness of the IDPA approach.

Evidently, the first two steps of the process are the most decisive, making it possible to rapidly reduce a list of $22,806$ protein-coding genes to 206, including the 37 core ES-identity candidate genes from *ESiC*-1. After another three decisions (*Esrrb, Ep300, H4K20me3*), the list of selected genes stabilises at just over a hundred genes ($n_{IDPA} = 116$). Any subsequent decision will cut off less than 10% of the remaining genes. I believe that the 79 genes ($n_{IDPA} - |ESiC\text{-}1| = 116 - 37 = 79$) that remain after five decisions together with the core stem cell genes from *ESiC*-1 make up another group of interesting candidates genes, because they share a core regulatory signature (high levels of *E2f1, Pou5f1, Esrrb, Ep300, H4K20me3*) that implicates them directly with the tightly co-regulated cluster of stem cell genes defined earlier (*ESiC*-1). I call this extended group of candidate genes *ESiC*-1^+.

The complete list of all "new" members of *ESiC*-1^+ (that is, those that were not in *ESiC*-1) along with their gene expression patterns and regulatory profiles is given in **Figure 5.20**. With only a few exceptions (*Arhgap26, Fbxo36, Plekha4, AC133494.1, AC142098.4, Gemin7, Setd1b, 4932422M17Rik*), these genes are expressed at an above-average level in ESCs. One gene, *Slc29a1*, was also part of the initial candidate gene list (*ESiC*), but the rest had been excluded by my strict analysis (**Section 5.2.1**), because they were either (i) not differentially expressed in at least one condition or (ii) more highly expressed in some other cell type than in ESCs. Interestingly, statement (ii) fits to a large number of histone-encoding genes that are indeed differentially expressed in ESCs in comparison to most other cell types, but are even more highly expressed in NPCs. It appears plausible that the expression of these histone genes was required in the genome-wide remodelling process that is necessary to enable

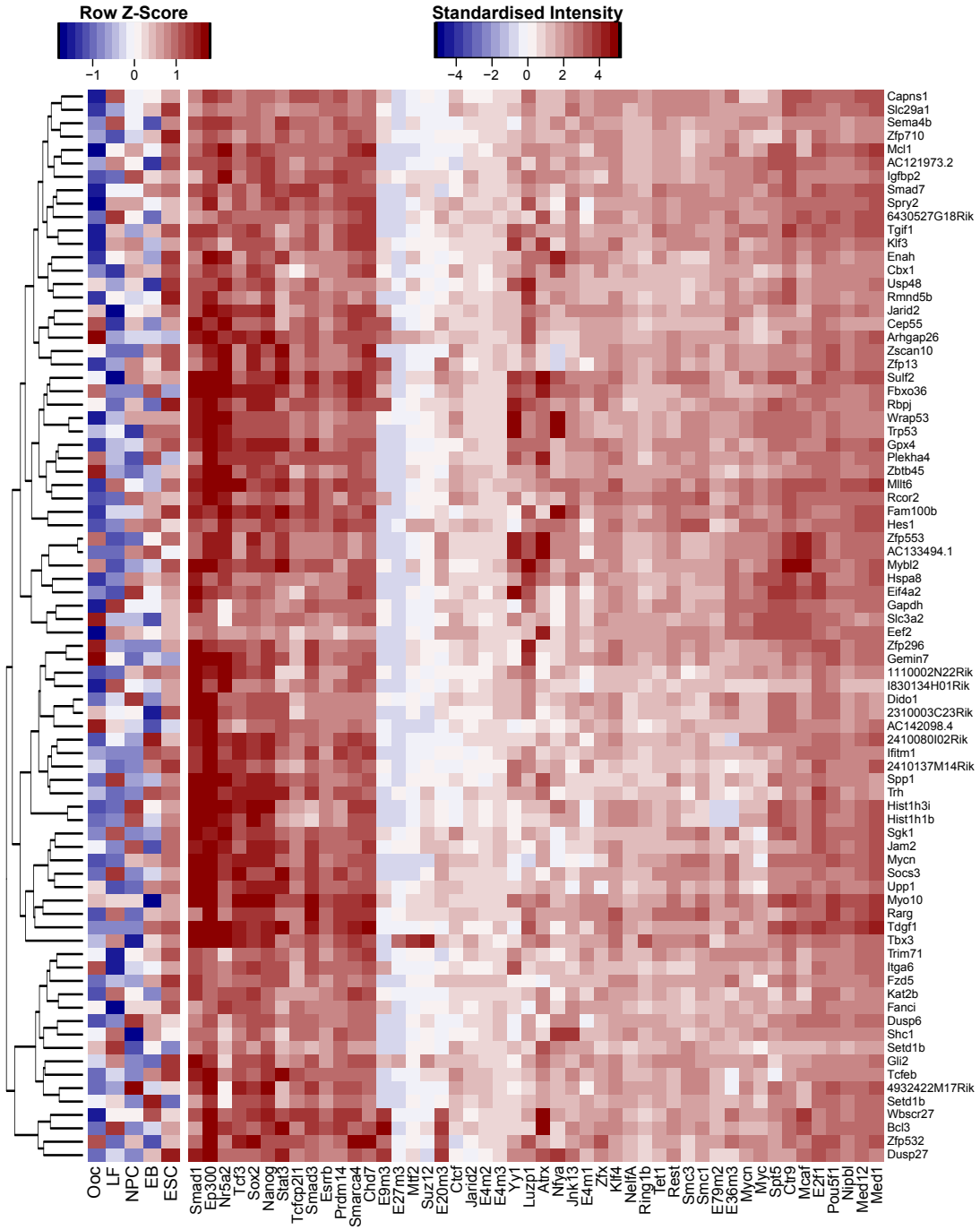

Figure 5.20: *ESiC-1*$^+$: Genes sharing a regulatory signature with *ESiC-1*. Combination of two heatmaps. Left: gene expression signature per cell type (mean of all samples). Colours have been scaled by row. Right: Standardised intensity values for DBPs and HMs. The rows in both heatmaps have been reordered by hierarchical clustering with complete linkage over the Euclidean distances in the right heatmap. The columns in the right heatmap have been reordered by the clustering in **Figure 5.17**.

the drastic epigenetic changes that enable pluripotency in stem cells and in the differentiation of those stem cells to cells with a more restricted state. Many other genes were still expressed in embryoid bodies to a degree that did not warrant calling differential expression towards ESC, which again seems reasonable in such a diverse aggregate of cells, many with a wide potency. The expression of those genes might mark the remnants of ES-identity in EBs.

Many of the genes in $ESiC\text{-}1^{+}$ have previously been implicated in ES-related functions, e.g. *Jarid2, Mycn* or *Tbx3*, but further work will be required to investigate the role of other members of the list. Curiously, the list also contains *Gapdh*, which is frequently used as a control for low-throughput expression assays (qPCR) and is generally considered to be a "house-keeping" gene. There are multiple strong binding sites for many DBPs and HMs in the proximity of *Gapdh* (warranting its inclusion in $ESiC\text{-}1^{+}$), but it is impossible to tell whether they are functionally linked *Gapdh*, because the gene, in its genomic context, is situated right in the middle of a cluster of overlapping genes (with *Iffo1* and *Ncapd2*). However, the expression of *Gapdh* is certainly not constant across cell types (see `http://www.geneprof.org/record.jsp?ds_id=pub_mm_ens58_ncbim37&id=24141`), shedding doubt on its use as a control gene. It will be interesting to see what future research will reveal about this gene.

5.3 Conclusions

There are two main conclusions to be drawn from the analysis presented in this chapter: Firstly, a rather small list of regulatory elements was defined that marks co-regulated genes with ESC-specific expression (**Section 5.3.1**). Secondly, based on common characteristics of this regulatory code between known key ESC genes and others, it was possible to identify a number of additional, *bona-fide* candidates for the core transcriptional circuitry of mouse ESCs (**Section 5.3.2**). I will now summarise and discuss these outcomes.

5.3.1 A Small List of Regulatory Elements is Sufficient to Define ESC Master Genes

It is of paramount importance not to misinterpret the results of the analysis presented in the previous section (**Section 5.2.5**) with respect to what they say about the regulation of stem cell genes: For example, one might be easily misled into thinking that the regulation of ESC genes (at least of $ESiC$-1) depended solely on the factors mentioned in the IDPA decision path (**Figure 5.19**.a) and that TFs previously considered important for ESC identity (e.g. *Sox2*) were insignificant just because they were not required for these decisions. This is most likely not the case. Indeed, many DBPs and HMs are strongly enriched in the proximity of $ESiC$-1 genes (and of $ESiC\text{-}1^{+}$) and I believe that this binding is functionally relevant. However, the

occurrence of these DBPs/HMs is non-informative with respect to the class distinction ("In *ESiC*-1" or not): This might be either because they also bind near many non-*ESiC*-1 genes, or because their binding intensity is redundant with respect to another DBP/HM. I use the term "redundant" here in a strictly statistical sense, as it gives no additional information. Biologically, the factors might not be redundant, but act cooperatively, antagonistically, or in some other way that causes them to frequently colocalise. Thus, what I am saying is not that the variables defined by the IDPA decision path describe the complete regulatory code of stem cells, but rather that these few regulatory elements are sufficient to discriminate ESC master genes from other genes in the stem cell transcriptome.

The composition of the list of regulatory elements in the decision path is very interesting: The very first decision disregards the majority of the mouse transcriptome, singling out genes with unusually high intensities for *Smad1, Ep300, Nr5a2, Tcf3, Sox2, Nanog, Stat3, Tcfcp2l1, Smad3, Esrrb, Prdm14, Smarca4, Zfx, Klf4, Mcaf, Pou5f1, Nipbl, Med1, Med12* and *E2f1*. Several factors achieve an almost equivalent split in the gene selection at this point (**Figure 5.19**), but the most optimal decision is made on the basis of *E2f1* intensity. *E2f1* (together with *E2f2* and *E2f3*) is believed to be important for normal cell cycle progression and survival, but can also function as a TF or by recruiting TFs to enhancers and promoters[41,82]. No interactions with any of the known ESC core TFs have been reported in the literature, however, *Gsk3β* has been shown to interact with *E2f1* promoting the ubiquitination of *E2f1*, blocking its activity[638]. *Gsk3β*-inhibitors have been used to maintain ESCs in an undifferentiated state[619] (**Section 1.1.3**). At this point, it is not clear whether said effect of *Gsk3β*-inhibitors might be, in part, due to increased *E2f1* activity in absence of *Gsk3β*. Furthermore, *E2f1* also interacts with several chromatin and histone modifiers, e.g. *Hdac1*[19] and *Dnmt1*[456]. I speculate that *E2f1* might act as a pioneering factor in stem cells, facilitating changes in chromatin structure favourable for active transcription and recruiting core regulatory elements to enhancer elements. In doing so, it marks a subgroup of genes accessible to TFs and to the elements of the transcriptional machinery. In a recent study, Cheng and Gerstein[76] have demonstrated that the binding intensity of *E2f1* is highly predictive of gene expression levels in ESCs lending further credibility to the importance of this gene in the transciptional network of stem cells.

All subsequent decisions would then single out ES-specific genes from these genes that are generally accessible to the transcriptional control by an assortment of TFs: As such it is not surprising that the core ESC regulator *Pou5f1* is the second key component of the decision path and that *Nanog*, another core element of the ESC transcriptional circuitry, could substitute as an alternative decision node (**Figure 5.19**). Of course, *Pou5f1* (and *Nanog*) also binds at many other genes, but amongst those "pre-filtered" by *E2f1*-intensity it might highlight those where it acts in concert with other TFs to establish a tightly regulated

control mechanism for cell state-critical genes. One of the factors that might be more crucial to these control mechanisms than previously expected could be *Esrrb*, which appears in the next step of the decision path. It has been known for some time that *Esrrb* is involved in ESC self-renewal[227,327] and can promote reprogramming of mouse embryonic fibroblasts to iPSCs[132]. The ways in which *Esrrb* exercises its function are still poorly understood, but recent research (personal communications and our own unpublished data: Festuccia, Osorno, Halbritter, Tomlinson & Chambers, *manuscript in preparation*) consolidates its importance at the heart of the ESC transcriptional circuitry.

The decision process is further helped along by the transcriptional co-factor *Ep300*, which has previously been observed to be present at many ES-specific enhancers[75]. *Ep300* is a versatile protein acting as a acetyl transferase for all histones[396] and other proteins[185,432]. In this role, it renders chromatin accessible for transcription factors and active transcription. Moreover, *Ep300* interacts with a plethora of proteins facilitating the binding of TFs. Amongst the DBPs included in my analysis, evidence has been demonstrated for interactions with *Yy1*[241] and *Smad3*[605] in mouse, and additionally for the orthologues of *Smad1*[419], *Tcf3*[53,369], *Stat3*[371,441], *Myc*[124,629] and *Klf4*[623] in human. It is not unlikely that the interactions observed in human also apply to the equivalent mouse proteins and it might even turn out that *Ep300* could interact with *Pou5f1* or *Sox2* directly, since interactions with *Sox4, Sox9* and *Pou3f2* – structurally similar proteins – have also been reported (source: `http://thebiogrid.org/108347/summary/homo-sapiens/ep300.html`). I speculate that *Ep300* might play a critical role in the enhancers and promoters of the genes in *ESiC*-1 by opening up chromatin and forming complexes with various TFs binding in these places. Without *Ep300* other key factors might not be able to exert their function correctly, explaining the presence of *Ep300* in the regulatory code of *ESiC*-1. Interestingly, the genetic deletion of *Ep300* in ESCs has been reported to affect *Nanog* expression (one of the members of *ESiC*-1) and impair the differentiation potential of the cells, but did not disrupt self-renewal[637].

Further studies will be required to scrutinise the significance of *Ep300* and other regulatory inputs in the context of *ESiC*-1 gene expression and to investigate in more detail how these factors colocalise, interact and cooperate to achieve their biological function.

5.3.2 New Candidates of the ESC Transcriptional Circuitry

In my analysis, I have used a "regulatory code" of DBP/HM inputs to define several lists of genes (*ESiC*-1 to *ESiC*-5: **Figure 5.17** and *ESiC*-1^{+}: **Figure 5.20**) that I consider high-confidence candidates with a likely role in important ESC-specific functions.

The list of genes with the most distinct regulatory signature (*ESiC*-1) contained many of the well-known members of the core ESC circuitry and genes previously implicated in the

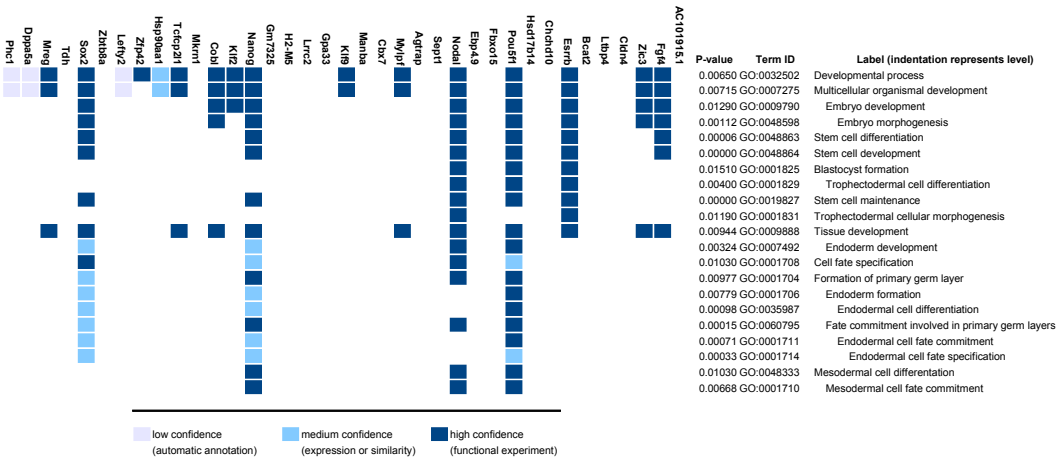

Figure 5.21: Functional annotation of *ESiC-1*. Gene ontology annotations for the 37 candidate genes from *ESiC*-1 as assigned by g:Profiler[446] (`http://biit.cs.ut.ee/gprofiler`; retrieved 26 April 2012).

maintenance of pluripotency and self-renewal, like *Pou5f1, Sox2, Nanog, Phc1, Lefty2, Nodal, Tdh, Klf2, Tcfcp2l1, Fbxo15, Zic3, Esrrb, Zfp42, Cbx7* or *Dppa5a* (**Section 1.1.4**). A quick overview of the putative function of all *ESiC*-1 genes is given in **Figure 5.21**. The findings in this study strengthen the link between those genes and ESC identity and call for further study of the less well known representatives of this list of genes. For instance, *Fbxo15* has been known to play some role in ESCs for quite a long time, however, little is known about what this role actually is. The gene is dispensable for self-renewal and pluripotency[549] and its use as a marker for the reprogramming of somatic cells to an induced pluripotent state cannot effectively distinguish fully from incompletely reprogrammed iPS cells[400]. Nevertheless, its expression is highly ES-specific and, as shown in this study, its genomic fingerprint closely reflects those of known key players in stem cells. Thus, I reason that *Fbxo15* might exert its role in ESCs redundantly with other factors and double knock-down studies might be required to delineate this function in detail.

The list also contains many candidates which have, to date, no known involvement in stem cells. Could their appearance alongside other established genes indicate some, as yet, unknown function specific to pluripotency or self-renewal? I have investigated these genes in the literature and public databases, singling out several particularly interesting candidates and will now further discuss their possible relevance in ESCs. Further experimental work will be required to validate any of these statements.

The gene *Zbtb8a* encodes a zinc-finger protein and presumed TF (source: UniProt, `http://www.uniprot.org/uniprot/Q96BR9`), that is included in the Kruppel-like family of TFs, but may have so far been overlooked, possibly due to its name. According to gene expression data from the ArrayExpress atlas (`http://www.ebi.ac.uk/gxa`), the gene is quickly down-

regulated during ESC differentiation[181] and is only weakly expressed in non-embryonic stem cells, e.g. haematopoietic stem cells[142], but is also spuriously found in other tissues and developmental stages, including adult cells. Several KLF's contribute to stem cell self-renewal, pluripotency and reprogramming[75,183,236,315], so this gene along with *Klf9*, which I have not mentioned before, but was also included in the list, might have a more central role in ESCs than previously believed.

The adhesion molecule *Claudin-4* (*Cldn4*) has recently been implicated with a differential role during the commitment of ESCs to endothelial and haematopoietic lineages[515]. The protein is involved in structural integrity and in tight junctions (source: EBI, `http://www.ebi.ac.uk`), which might possibly contribute to colony formation in ESCs. *Cldn4* is also linked to various cancers and its over-expression has recently been associated with derepression of epigenetically silenced genes[287], consistent with the concept of transcriptionally permissive ESC transcriptomes. Calling *Cldn4* "ES-specific", however, is certainly not warranted: Examination of transcriptional profiles from ArrayExpress does indeed confirm differential over-expression in ESCs in comparison to various other embryonic cell types, but there are somatic tissues in which it is much more strongly expressed[169]. It might be possible that *Cldn4* exerts a specific function in ESCs, perhaps via its properties as an epigenetic modifier, rather than in its role as a wide-spread membrane-associated molecule.

Cordon-bleu (Cobl) has been described as a nucleation factor involved in neuronal organisation[2], but was originally identified as a gene specifically regulated during early mouse development and patterning[150]. As for the previous genes, examination of a wider range of conditions via ArrayExpress shows that *Cobl* expression is not entirely ES-specific: *Cobl* expression in ESCs is quickly lost upon differentiation into EBs[181] and is higher in ESCs than in hematopoietic stem cells[142]. However, it is more highly expressed in various adult tissues than in ESCs. *Cobl*'s link to early mouse development could point to a possible role in ESCs.

Earlier computational meta-analysis has already pointed out the gene *Manba*[157], but no new findings with respect to ES-related functionality have been reported since. The gene has been found over-expressed in various adult organs, in particular, kidney[490], but not as high as in ESCs[563]. *Manba* encodes the enzyme *Beta-mannosidase* that works in the glycan metabolism pathway. It is not clear to me what role this particular gene might play in ESCs, but experimental evidence (including additional unpublished data not shown here) repeatedly links *Manba* to the ESCs.

*Septin-1 (*Sept1*)* is a protein involved in cytokinesis. In their hallmark study, Takahashi and Yamanaka[529] found *Sept1* as one of a group of genes that was up-regulated in ESCs in comparison to EFs and in some but not all iPS cells. Those cells that did not up-regulate *Sept1* also did not activate other stem cell markers like *Sox2* and *Dppa5a*, indicating that *Sept1* might mark a partially reprogrammed state and possibly be critical to whatever mechanisms

are required for a complete transition to pluripotency. Lastly, the gene is down-regulated upon inhibition of *Esrrb* by RNAi[406] and has also been reported to be a shared target of *Nr0b1* and *Nr5a2*[258], putting it further in line with *Pou5f1* and others.

The chaperone *Hsp90aa1* assists in the folding of target proteins (source: NCBI Gene, `http://www.ncbi.nlm.nih.gov/gene/3320`) and, most interestingly, has been linked to *Stat3* function in ESCs[480,494]. Knock-down of the co-chaperone *Hop* required for *Hsp90aa1* function results in extracellular accumulation of *Stat3*, decreased *Nanog* mRNA levels and loss in capacity for EB formation[328]. The human orthologue of the protein has been shown to interact with a large number of proteins (source: `http://thebiogrid.org/109552/summary/homo-sapiens/hsp90aa1.html`), for instance, *Tgfbr1/-2*, *Map3k3/-7* and *Fgfr3*, which might link the gene to several ESC-relevant signalling pathways (**Section 1.1.3**). It appears that this protein might play a central role in maintaining ESC pluripotency by supporting *Stat3*, *Nanog* and possibly others in their functioning.

An unrelated study also demonstrated that *Hsp90* suppression allowed efficient ubiquitination and degradation of a subunit of telomerase[270]. Sustained telomerase activity is required to maintain telomere length in continuously dividing stem cells. The same study also identified another gene in my candidate list, *makorin, ring finger protein, 1 (Mkrn1)*. *Mkrn1* encodes a ubiquitin ligase that might mediate this ubiquitination. Furthermore, the same protein has been linked to both cell survival and apoptosis via selective ubiquitination of *p53* and *p21*, respectively[298]. Recently, Emily Walker (University of Toronto) reported as part of her thesis that *Mkrn1* over-expression could support the maintenance of ESCs under differentiation conditions[573].

Glycoprotein A33 (Gpa33) is a gene formerly believed to be expressed "almost exclusively by intestinal epithelial cells"[239]. More recently it has been associated with colon cancer (in human) and is, in fact, used as a marker for this condition. Interestingly, its expression has been found to be *Klf4*-dependent[436]. Examination of global expression signatures from ArrayExpress, however, clearly shows that it is also expressed in ESCs and a number of other tissues[142,326]. I am unable to speculate about its involvement with stem cells, however, the gene, being a cell surface antigen might turn out to be useful as an ESC marker.

Lastly, *latent transforming growth factor beta binding protein 4 (Ltbp4)* is a protein that may have a role in the structure of the ECM (source: `http://ghr.nlm.nih.gov/gene/LTBP4`). The gene is also transcribed in various adult tissues, including heart, pancreas and lung (source: `http://www.copewithcytokines.de/cope.cgi?key=LTBP4`) and it is required for normal lung development[517], but it has also been confirmed independently as differentially expressed in ESCs and primordial germ cells[142,169]. More interestingly, *Ltbp4* has previously been correlated with *Pou5f1* when its expression levels were observed to drop quickly with *Pou5f1*-depletion[183]. *Ltbp4* can also bind to Tgfβ. Tgfβ signalling is important to many

developmental processes (**Section 1.1.3**) and the implication of its binding protein *Ltbp4* directly in the core transcriptional network of ESCs is certainly an interesting finding.

Before any further specific studies were to be carried out into any of these candidates or any of the other members of the candidate gene groups (*ESiC* or *ESiC*-1^{+}), I would suggest to investigate the genes further by investigating their functions using various databases in order to bring in complementary sources of knowledge and single out the most promising candidates for follow-up studies. I have done so here manually for a number of candidates, but the evaluations should be done in a more systematic manner using the following and similar resources:

- Loss-of-function phenotypes: The International Knockout Mouse Consortium[52] (IKMC; `http://www.knockoutmouse.org`) and its member projects, e.g. the Knockout Mouse Project (KOMP; `https://www.komp.org`), have started systematically investigating the phenotypic effects of the knock-out of all protein-coding mouse genes. The projects are still ongoing, but where phenotype information is available this would provide a valuable source of information for all candidate genes. In addition to the phenotype information from the IKMC, knock-out or knock-down microarray studies for several candidates are available and I would like to compare the effects observed in these studies to the effects observed after loss of the stem cell core factors (*Pou5f1, Sox2, Nanog*) to see whether there is any remarkable overlap (either by looking at the global correlation or at overlaps of differentially expressed target genes).

- Protein-protein interactions: Databases storing information about experimentally determined interactions between proteins, e.g. the BioGRID[516] (`http://thebiogrid.org`), can further help to support the candidacy of genes if interactions between those proteins and others with a relevance to stem cell function have previously been found.

- Evolutionary conservation: Genes with critical biological functions in essential developmental processes are likely to be conserved across species and thus it would certainly be a good idea to check how well the genes in my candidate lists are conserved, at least, across other mammalian species. This could be achieved either by looking only at the sequence conversation to closely matching homologs in human, rat and others (e.g. via Ensembl's BioMart; `http://www.ensembl.org`), or by summarising the sequence conservation score across multiple-species alignments per gene (multispecies conserved sequences, phastCons or regulatory potential scores; reviewed and compared by King *et al.*[271]). Multi-species conservation tracks for this purpose are available from the UCSC genome browser (`http://genome.ucsc.edu`).

- Other functional annotations: I have already used the Gene Ontology to annotate the

genes from *ESiC*-1 (**Figure 5.21**). However, many alternative sources of gene-centric functional annotation, pathway memberships, disease relevance and the like exist and there are a number of tools that can be used to annotate genes automatically, e.g. DAVID[208,209] (`http://david.abcc.ncifcrf.gov`). This information should be utilised to further annotate the candidate genes and to check whether the groups of co-regulated genes identified in the analysis are enriched for similar biological functions.

Chapter 6

Final Discussion

Approaching the end of this thesis, I shall now review the work that has been detailed before and put the primary research achievements in a broader context. Finally, some perspectives for future work shall be addressed and the work will be concluded with a few closing remarks.

6.1 Summary of Research Motivation and Achievements

When I started writing up this dissertation, I thought this would all be a rather short and concise affair. As it turns out, summarising the work of several years is anything but a trivial task – evidenced now by the extent of this document. I shall now try to summarise the main achievements of my work and reiterate how the different components described in the earlier chapters fit into that journey that is now soon to be concluded.

6.1.1 Motivation and Goals

From the outset, it had been my goal to investigate the fundamental mechanisms, the driving forces that make stem cells what they are. Decades of past research have elucidated a plethora of extrinsic and intrinsic contributors to the establishment and maintenance of the peculiar identity of these cells. A selection of these factors have been reviewed in **Chapter 1**, including a summary of the best-known signalling pathways that trigger cell-internal programs essential for this state. Many of the genes affected by these pathways have been studied extensively, although the way in which their expression is directed as a result of incoming stimuli is often poorly understood. Three genes stand out as the key regulators of stem cells: *Pou5f1* (also known as *Oct4*), *Nanog* and *Sox2*, although the latter is also expressed in a variety of neural cells in the embryo and even in some adult cells.

Virtually all other genes implicated in the ESC circuitry have in some way been linked to the activity of these core factors and it is now commonly believed that they orchestrate

the expression of many downstream effectors often in a cooperative manner possibly involving many other TFs and regulatory elements[75]. Simple binding by one factor alone is not sufficient to regulate transcription of target genes and the correlation of any one factor to the gene expression programme in changing cell states is generally poor. Additionally, the importance of non-genic influences on gene expression is becoming increasingly evident and it appears that it is only the right combination of TF binding activity with the presence of many transcriptional control elements like co-activators and polymerase-linked or -controlling elements as well as the epigenetic markup of a cell that allows productive transcription to occur[75,245,342]. Epigentic factors, in particular in the form of a multitude of histone modifications and DNA methylation, are believed to influence gene expression programmes beyond the life time of a single cell and are crucial for stable cell cultures. The importance of epigenetics has recently received much additional support with observations derived from the generation of iPS cells: It has been reasoned that a major epigenetic reset or remodelling is required to erase cell type-specific properties from differentiated cells in order to redefine their cell identity to one akin to that of ESCs[37,202,212,360].

It was my goal to expand our understanding of how heterogeneous regulatory inputs influence gene expression. I hypothesised that there were common regulatory mechanisms driving, if not all, then at least some of the functionally related members of the core transcriptional network of stem cells. Would it be possible to identify such a shared signature, a "regulatory code of stem cells"? Which genes were described by this code and which factors determined it?

6.1.2 Early Exploratory Data Analysis

With these questions in mind, I was thrilled to start my Ph.D. research at a time that coincided with the publication of the first large-scale applications of HTS to the study of regulatory and epigenetic mechanisms. Perhaps of most impact to my personal direction were those groundbreaking studies conducted in the laboratories of Ng[75], Young[342] and Meissner[360], which demonstrated the great potential this technology would have to offer for future functional genomics research.

6.1.2.1 Establishment of Data Analysis Workflows for High-Throughput Sequencing Data

At this time, I was keen to get an opportunity to try out the technology myself and was fortunate enough to have the chance to become involved in various collaborations, the two most extensive of which I have described in detail in **Chapter 2**. I valued these experiments, apart from their obvious relevance to stem cell research, as a vehicle to identify requirements

and issues with the data analysis of HTS and to establish effective, practical workflows for the processing of the large amounts of data generated.

This was not an easy task back then with many software tools still in their infancy and an overall lack of established methodologies. I therefore spent a lot of time looking for and evaluating existing software tools fit for the purpose and chained those together into a simple pipeline, filling in gaps where required, for instance, by writing custom scripts for assessing raw data quality and filtering out erroneous segments of the data or to quantify gene expression intensities from alignment coverage.

The pipeline developed and general expertise acquired were then applied in the context of of two collaborations, one of which has already resulted in a publication and another is currently under review. The primary results of these studies shall be briefly recapitulated in the next sections (**Section 6.1.2.2** and **Section 6.1.2.3**).

6.1.2.2 Identification of Putative Targets of the Transcription Factor *Nanog*

In this study, DeepSAGE expression profiling had been used to assay global gene expression signatures in wild-type ESCs and in a mutant in which the *Nanog* gene had been knocked out. This research was conducted in collaboration with Ian Chambers and various members of his group at the Institute for Stem Cell Research / Centre for Regenerative Medicine, aforemost Violetta Karwacki-Neisius, Nicola Festuccia and Rodrigo Osorno.

I applied my previously established analysis pipeline to the generated data and differential gene expression analysis yielded over a thousand genes. In-depth bioinformatics analysis allowed us to narrow down my initial results to a concise list of high-confidence candidate genes that I considered likely direct targets of the TF. I achieved this target refinement by integrating various external gene expression datasets as well as ChIP-seq and ChIP-on-chip binding data from published studies. These data were used to look for consistently observed expression changes associated with different levels of *Nanog* and also to find those genes with reliable binding sites in their proximity. Many of the candidate genes were subsequently studied by my collaborators resulting in promising future research – the gene *Rlim* (also known as *Rnf12*), for example, has already been studied further by the members of the Chambers group[375].

This demonstrates impressively how the meta-analytic integration of different datasets can help to enrich independent and otherwise isolated pieces of data and leverage existing knowledge to derive new insight, a philosophy which I have now very much taken to and try to advocate as part of all my ongoing work.

6.1.2.3 Determination of Transcriptional Characteristics of Stem Cell-Like Populations in Plants

In a second collaborative effort, I teamed up with the group of Gary Loake (Institute of Molecular Plant Sciences, University of Edinburgh), who are studying pluripotent and self-renewing cell populations in various plant species. In a remarkable piece of work they were able to isolate a population of cells from the cambium of the Japanese yew (*T. cuspidata*) that exhibited a proliferative potential exceeding that of other cells, in particular, dedifferentiated cell types which had previously been used for the derivation of various plant products. The use of these cells, called cambial meristemic cells (CMCs), opens up a new avenue for the effective, large-scale production of natural plant products with medicinal or cosmetic value, such as taxol, which is used in cancer treatments[297].

As a part of the larger study, comparative gene expression profiling was performed on two cell populations and I contributed to this work by comparing the data from both conditions by aligning the data to the newly assembled *T. cuspidata* transcriptome and statistically evaluating differences. A number of a contigs (basically, putative genes) were detected that were substantially over-expressed in CMCs and hence putatively involved in the stem cell-like properties of these cells. The existence of stem cells in plants in itself is not a new idea, but their genetic, epigenetic and regulatory properties have so far been poorly studied despite potential medical and commercial impact. The contigs discovered in this study are now being used by the Loake lab as markers for the most suitable cells for taxol production.

I am currently continuing my collaboration with the Loake lab to elucidate the role of similar cell populations in other plant species and to discover common properties.

6.1.3 Development of a Tool and Resource for the Study of Gene Expression and Regulation

Looking back at the effort it took me in the beginning to get started with HTS data analysis, the situation has definitely improved with a broad variety of rather mature tools available nowadays. Nevertheless, finding right tools and putting them effectively together remains a difficult task for those new to the matter.

After the initial round of pilot projects, I had set out to develop a new software tool, an environment for the execution of the kind of analysis workflows I had developed previously. The aim was to streamline common analysis tasks in a user-friendly, reproducible and transparent manner that would allow for the rapid analysis of large sets of experimental data.

The motivation for this project (**Section 3.1**) came from two angles: On the one hand, I meant to make HTS data analysis more accessible to all researchers. Commonly, the analysis process involves many largely repetitive tasks: Issues like quality control and alignment of

short read sequences to a reference genome are steps that are part of almost any experiment and from contact with other research groups I had learned that many researchers struggled even at this first hurdle. Why should it not be possible to provide the excellent openly available tools to a wider audience in a simple and usable manner?

On the other hand, I was motivated by my own research goals, of course. In order to effectively integrate the vast amounts of heterogeneous data generated by modern HTS instruments, it was critical to have a way to rapidly process them in a consistent manner, but with the ability to easily and quickly adapt standard pipelines for individual experiments. The latter is necessary, because although the analysis steps are largely the same, experimental techniques are variable and the exact same analysis approach does not always fit. Consequently, to make this work the software needed to be flexible and provide means to quickly assess the outcomes of each step of the process.

In response to these requirements, an analysis framework which I later called *GeneProf* was developed, which has been described in detail in **Chapter 3**. The main features may be summarised as:

- A web-based user interface presents an accessible and easy-to-use entry point for users. There is no need to install specialised bioinformatics tools or other software.
- Computationally complex genomics analysis tasks, that would usually necessitate powerful computer equipment, are being executed remotely on a network of high-performance, dedicated computing machines.
- The system integrates expert knowledge and assists users by providing best-practice data analysis approaches via simple data analysis "wizards". These wizards make it possible (even for novice users) to set up elaborate and sensible analysis workflows within minutes.
- Data analysis is powered by a flexible and adaptable workflow engine. In this workflow environment, all data analysis steps ("modules") can be combined in arbitrary ways to achieve highly specialised analysis goals. Wizards also create such workflows, so they can be adjusted later on as the user sees fit.
- All steps of the analysis are supplemented by a range of summary statistics and plots, which make it easy to assess the results at each stage and, if necessary, spot problems that can than be accounted for by amending the analysis workflow.
- The outputs and intermediate results of all steps of the analysis process are recorded, changes to the workflow tracked and all parameter settings are available through the workflow, making the analysis fully transparent and addressing the issue of reproducibility.

- Short read quality control measures and alignment are integral to all types of analysis and well-supported by the software. Several established, publicly available tools have been integrated into the system to provide a choice of methodologies.

- The system supports downstream RNA-seq analysis by providing means to quantify gene expression intensities from aligned read datasets and to normalise and compare the expression in different cell types, tissues or experimental conditions with the best available statistical methods.

- ChIP-seq analysis is also supported and the software can be used to identify sites of significant enrichment in binding profiles ("peaks") using multiple published algorithms. Peaks can also be assigned to putative target genes.

- Data from different experiments and different techniques can be juxtaposed and visualised together easily for comparison and meta-analysis.

I utilised this software to re-analyse a large amount of published data from studies relevant to stem cell research (**Chapter 4**). In this process, I soon realised that it would be most sensible to use the results of these analyses to build up an integrated database. Currently, most published HTS research data is submitted to public archives in raw format, which is commendable and a great step towards open science. However, the raw data in itself is of little immediate use to any researcher and requires laborious processing to be transformed to biologically meaningful findings.

Therefore, I have then extended GeneProf's functionality to combine the data analysis suite with a resource of all completed analysis experiments: All analysis projects that I (or others) run through the software can be made available (publicly, if desired) through the interface. Each project contains the complete input data and all analysis results in combination with the entire workflow that produced these results.

While most smaller research labs will probably rarely generate HTS data themselves, they can still benefit from the wealth of information that is available nowadays, thus boosting the effective sharing of knowledge. Data from experiments that have already been analysed in the system can be imported (within seconds) into other workflows, where it can be used to enrich primary experimental data and to leverage findings to another level – much like I did in my early data analysis projects described in **Section 2.1**. Of course, researchers may also choose to re-analyse individual pieces of data and to try out different methodologies to gain a better understanding of the nature of the data and the effects of different analysis steps.

The software has been released to the research community in the beginning of 2012[182] and has since attracted much interest. Thousands of people have visited the website and browsed the archives of data available and several hundreds have registered and started analysing their

own experiments (**Section 3.4.3**). I sincerely hope that this trend will keep up and I plan to generate further interest by implementing new features into the program and publicising its availability to the community.

6.1.4 A Step Towards Identifying Common Regulatory Mechanisms of Stem Cell Genes

Having developed the necessary tools, I could then return to the study of the regulatory mechanisms driving the expression of genes crucial for the establishment and maintenance of stem cell identity (**Chapter 5**). I hypothesised that functionally related genes in stem cells shared common regulatory mechanisms. To test this hypothesis, I proceeded in three steps: Firstly, I attempted to identify a list of genes that I considered likely to be important for ESCs (**Section 6.1.4.1**). Secondly, I gathered a large amount of data about the state of regulatory proteins in ESCs and objectively investigated the genome-wide characteristics of these signals (**Section 6.1.4.2**). Finally, I combined both collections of data to identify a regulatory signature shared between ESC-specific genes (**Section 6.1.4.3**).

6.1.4.1 Identification of Genes Expressed in Embryonic Stem Cells

I first wanted to establish a list of functionally related genes, so that I could later on look for common regulatory mechanisms within this group. I decided to focus on genes that were important for stem cells and reasoned that genes that were highly expressed specifically in ESCs would be likely candidates for this function. The idea was simply that genes that were phenotypically related (expressed in the same conditions) might serve similar or complementary functions. If these genes were expressed in ESCs, but not in other cell types, it would be plausible to expect that they were involved in conveying ESC-specific characteristics to cells.

To establish a list of candidate genes, I compared the global gene expression profile of mouse ESCs with those of four other cell types: Adult lung fibroblasts (LF), neural progenitor cells (NPC), embryoid bodies (EB) and totipotent oocytes (Ooc). For each comparison, I pulled out several datasets from the GeneProf database and looked for differentially over-expressed genes. I then took the intersection of the genes identified in all individual comparisons to pinpoint genes specifically expressed in ESCs. By including embryonic cell types in the comparison (NPC, EB, Ooc), it was possible to filter out genes that were important in early development, interesting in themselves, but not specific to the identity of ESCs.

The intersection of all comparisons contained 229 candidate genes (called "ES-identity candidates", *ESiC*; **Section 5.2.1**). This list was highly enriched for genes involved in the maintenance of stem cells, for instance, the core ESC regulators *Pou5f1, Sox2* and *Nanog*, supporting the notion that my methodology did indeed select genes relevant to stem cell

identity.

6.1.4.2 Investigation of the Genome-Wide Markup of Regulatory Signals

In the next step of the analysis, I wanted to examine different kinds of regulatory signals with respect to how they were distributed across the mouse genome and to discover relationships between them. Here, I call "regulatory" all those signals that might contribute towards alterations of the expression level of target genes. To get started, I chose to look at various types of histone modifications (HM) in ESCs and embryonic fibroblasts (EFs). Specifically, I looked at methylations of various lysine residues, which was the kind of HM with the most available data. Additionally, I collected all the datasets for DNA-binding proteins (DBP) in ESCs that were stored in the GeneProf database. These DBPs were either transcription factors (TFs), other proteins that were actively involved in shaping DNA in a way permissive for productive transcription or proteins directly involved in the transcriptional machinery around polymerase itself. There was data for 40 DBPs in total with several biological replicates for a number of them.

Using these datasets, I investigated the genome-wide patterns of DBP and HM distribution (**Section 5.2.2**, **Section 5.2.3** and **Section 5.2.4**): I quantified the occupancy levels of the surveyed proteins across the entire genome and checked how the occupancy related to known genes and to each other. Where did individual proteins bind? Were regions bound by one protein also enriched for the binding of another protein and were there any distinguishable groups of proteins binding in similar regions of the genome? Put briefly, the results of my investigations confirmed observations from previous research, but also revealed a few patterns that as such had not been described before:

Location of DBP/HM binding: I observed that many DBPs were preferentially enriched in the proximity of promoters, i.e. near the TSS of known genes. The TFs *Nanog, Sox2, Pou5f1, Nr5a2, Chd7* and others were specifically enriched upstream of the TSS, indicating that they might exercise their activity in distal enhancer elements. On the other hand, the *PolII*-interacting proteins *NelfA, Ctr9* and *Spt5* were occupying loci at promoters and within gene bodies consistent with the reports of others[437]. H3K36me3 and H3K79me2 were detected along the entire body of genes. All other HMs were preferentially enriched in promoter regions. This is coherent with their function: HMs modulate the accessibility of chromatin by DBPs and the transcriptional apparatus. Those HMs that were enriched at the TSS have an impact on the initiation of transcription, while those that are found throughout the gene open chromatin paving the way for transcriptional elongation. This is consistent with the general understanding of HM function[26, 76, 264, 343].

Similarity of DBP occupancy patterns: Genome-wide occupancy patterns of almost all DBPs were correlated to some degree. It is possible that this is a technical artefact caused

by preferential pull-down of certain DNA regions by ChIP regardless of actual protein binding. Another explanation might be a rather weak binding affinity of many DBPs resulting in all (accessible) DNA regions to be bound at a low level. Nevertheless, enrichment patterns beyond this background level of similarity successfully clustered functionally related proteins together. For instance, the mediator subunits *Med1* and *Med12* and the associated protein *Nipbl* or the PRC2 members *Suz12* and *Mtf2* frequently occured together at the same regions in the genome. It has been previously reported that many DBPs putatively co-occupy enhancer elements[75] and I can confirm this observation and say that it extends to more proteins than previously known. Some sites were occupied by as many as 31 distinct DBPs. One of the sites occupied by the most factors was in the proximity of the pluripotency gene *Pou5f1*. The TFs *Nanog, Sox2, Tcf3* and *Pou5f1* appeared to be particularly closely related. It is now commonly believed that *Pou5f1* and *Sox2* bind DNA cooperatively in many places by forming heterodimers[70]; such cooperative binding might also occur in other combinations of the mentioned factors. Another group of TFs was centred around *Myc, Mycn, Klf4, Zfx* and others. However, measurements from two independent studies for *Sox2* and *Pou5f1* were somewhat inconclusive as to whether there is indeed a global distinction between this group of TFs and the first. Binding of the insulator element *Ctcf* is (weakly) anti-correlated to the activity of many TFs, including *Pou5f1, Sox2*, and *Nanog*.

Similarity of HM occupancy patterns: On a global level, signals for the activating histone marks H3K4me2 and -me3 were closely correlated. Their profiles were also highly similar with H3K79me2. H3K36me3 distributed differently with respect to genes, but still clustered more closely with the other active marks than with the repressive ones. The repressive mark H3K27me3 also occupied similar regions as H3K4me2/-me3 (that is, regions overlapping the TSS of genes), but often at different genes. Interestingly, H3K4me1 was more closely correlated to H3K27me3 than to H3K4me2 or -me3. The global pattern of HM occupancy was highly correlated in ESCs and EFs (across all marks where data was available for both cell types), indicating that the majority of epigenetic signatures did not change between cell types. However, there was a subset of genes for which H3K27me3, H3K36me3 and H3K4me2/-me3 occupancy changed notably.

Common regulatory patterns pertaining to genes: The core stem cell genes *Pou5f1, Sox2* and *Nanog* shared a highly similar DBP profile, however, differed slightly in their HM markup. This was mostly due to a lack of (strong) *Nanog*- and *Pou5f1*-associated H3K27me3 in EFs and the presence of a *Nanog*-associated *H3K9me3* signal in ESCs. Genes sharing the same HM profile as *Pou5f1* and *Sox2* included many previously implicated in ESC state and developmental processes. A larger number of genes shared similarities in HM occupancy with *Nanog*. Similarly, genes sharing a DBP profile alike those of thee three co-factors also contained a substantial proportion of putative ESC regulators.

6.1.4.3 A Combination of Regulatory Signals Marks Phenotypically Related Genes in Stem Cells

In the final part of my analysis (**Section 5.2.5**), I attempted to combine the different measurements, HM and DBP signatures, to closely examine the regulatory code of the ESC candidate genes identified in the first stage (*ESiC* genes). Before I could discriminate an ESC-specific signature, though, it was necessary to find out whether there was a signature that distinguished active from inactive genes in general. Many DBPs and HMs showed differences in intensity levels between both groups of genes, however, no single factor alone would have been able to discriminate active and inactive genes reliably enough. That is, although H3K36me3, H3K4me3 and H3K79me2 were quite well correlated with expression levels, and *Mtf2* and *Suz12* quite anti-correlated to the same, their mere presence was not enough to say whether a gene was active or not. I found, however, that the combination of measurements for H3K36me3, H3K4me3 and H3K79me2 was fairly successful in predicting gene activity with 84.4% accuracy – much more could not not be expected given the variability in measurements between experiments and replicates.

It was not possible to define a single regulatory signature for all *ESiC* genes using the measurements at hand. However, I was able to identify five subgroups within the candidates, one of which was investigated in detail: The group of candidates termed *ESiC*-1 contained the three core factors and 34 other genes that were all marked by strong intensities for a large number of DBPs and HMs. It seems reasonable that these genes might make up a core of tightly regulated ESC-prototype genes. Indeed it was possible to perfectly discriminate *ESiC*-1 from the rest of the transcriptome by a computationally determined set of rules (**Figure 5.18**.b). In subsequent investigations, I discovered that the genes in *ESiC*-1 are marked primarily by an enrichment in the activity of four DBPs and one HM in the neighbourhood of their promoters (**Figure 5.19**): *E2f1, Pou5f1, Esrrb, Ep300* and H4K20me3. Based on measurements for these five regulatory inputs it was possible to distinguish *ESiC*-1 genes alongside 79 other genes with a similar regulatory markup (and many with known involvement in stem cell establishment and maintenance) from 99.5% of mouse transcriptome. The regulatory code defined in this part of the analysis has been discussed in **Section 5.3.1**.

Moreover, I have used the similarity of genes in terms of their regulatory inputs (DBPs and HMs) to define several lists of genes (*ESiC*-1 to *ESiC*-5: **Figure 5.17** and *ESiC*-1^{+}: **Figure 5.20**) that I consider high-confidence candidates with a likely role in important ESC-specific functions. These candidate lists contain many of the known champions of pluripotency and self-renewal, for instance, *Pou5f1, Sox2* and *Nanog*, but also include a number of genes of whose function little is known. I have discussed some particularly interesting candidates in **Section 5.3.2**. It will be exciting to see what future research will tell us about those genes.

6.1.5 Relation to Other Studies on Regulatory Elements

To date, most computational genomics research concerning itself with HMs and DBPs as regulatory mechanisms has focused on (i) how these are linked to transcriptional activity[112, 196, 641], (ii) whether the presence of regulatory proteins can be used to predict gene expression levels[76, 93, 143, 253, 267, 406, 492], (iii) on the identification of regulatory modules, that is, combinations of regulatory inputs that co-regulate target genes[1, 156, 493] and on (iv) how regulatory signatures differ between cell types, tissues or conditions[122].

I also address these kinds of questions in the beginning of my analysis, but eventually have a slightly different goal in mind: To identify common regulatory signatures that distinguish classes of genes and specifically, those genes that are important to stem cells. A better understanding of these regulatory mechanisms can complement our models of the transcriptional programme of stem cells, help to optimise the efficiency for the derivation and maintenance of stem cells (whether from the embryo or from somatic cells) and provide hypotheses for the perturbation of stem cell state and differentiation.

6.2 Future Work and Perspectives

Naturally, the work described in this thesis does not present the end of the line. The development of methods and the analysis described is very much an ongoing project and several future avenues shall be briefly outlined in the following paragraphs.

6.2.1 Expansion of the GeneProf Platform for Other Data and as a Rich Resource for the Research Community

I have previously pointed out several future improvements to the GeneProf system that I mean to implement in the future (**Section 3.4.4**). The improvement with probably most relevance to my future research will be an expansion to further types of data, support for DNA methylation being the most obvious candidate that could help to complete the regulatory signature I am trying to discover.

This improvement and further extensions will expand GeneProf's profile as a rich resource for the biological research community. GeneProf has already accumulated and processed a large amount of data and even during the process of writing this thesis the repertoire has further increased. I trust that many other researchers will benefit from this database.

6.2.2 Refining the Regulatory Code of Mouse Embryonic Stem Cells

As far as the data analysis is concerned, I have already pointed out several weaknesses in the current approach that need to be (and will be) addressed in the future.

First, one of the benefits of using HTS for expression profiling is that it can be used to study transcription as a whole and without any inherent bias. Limiting my analysis to the protein-coding fraction of all transcriptionally active units was thus a regrettable, yet in this particular context necessary, decision to make. Being more aware of the issues impairing comparability between studies, I will now be able to make a better-informed decision about which datasets to include in order to avoid discarding valuable information. Due to a lack of data, even the final candidate list contained many genes that, following further research, turned out to be expressed in many non-ES cell types. More high-quality RNA-seq datasets are now being published (including data from iPS cells) and I believe that using these in combination with new cross-experiment normalisation algorithms for HTS data[303], it will be possible to derive an (even) more accurate and complete candidate list.

Second, I have extensively used correlation measures to compare genome-wide similarities (and differences) between datasets. Several recent publications have proposed more sophisticated techniques to calculate such distance measures in a more precise and sensible manner by assessing similarity in peak profiles in an asymmetric fashion[81]. Additionally, a recent paper puts forward a novel way of normalising ChIP-seq intensities on the basis of shared binding peaks[496], which I also consider likely to further refine out results, possibly in a more appropriate manner than by the standardisation that I chose to apply to the final data.

Third, a major issue has been discovered with disregarding ambiguous alignments in ChIP-seq data (**Section C.2**). At present, almost all published research I am aware of is concentrating on uniquely aligned reads to avoid ambiguity. It is unlikely that the difference has any far-ranging impact on the global conclusions drawn from ChIP-seq studies, however, genes located in highly repetitive regions or those that are present in multiple (identical or near-identical) copies in the genome, will be substantially under-represented in all results. The ESC candidate genes in *ESiC*-5 are a striking example of this phenomenon and it is not unlikely that those in *ESiC*-4 are also affected to a lesser extent (**Section 5.2.5**). I plan to rigorously validate existing approaches[378,579] for dealing with alignment ambiguity in the context of ChIP-seq data and amend the data analysis methodology accordingly.

Fourth, the current selection of inputs represent only a small percentage of all known regulatory elements: Most notably, I have so far not considered any histone modifications other than methylations. More acetylation data is now becoming available and these datasets can be easily added on to the current data selection. Further, I have in the introduction briefly discussed the role of DNA methylation, but not yet integrated this kind of data in my analysis. With future improvements to the GeneProf software it will be possible to include the DNA methylation state of genes in the regulatory signature defined in the analysis. Finally, I have assessed 40 different DBPs, more than ever before in ESCs, yet this still represents only a tiny fraction of all proteins (it is estimated that there are somewhere in the range

of $1585 - 1727$ TFs in mouse[249, 440]). With the inclusion of other parameters it should, in principle, be possible to derive increasingly "clean" signatures for functionally related genes. But rather than including just more and more inputs, I consider it more promising to look for data that have already been implicated in pluripotency and self-renewal. For instance, I am keenly waiting for ChIP-seq data for *Zfp42, Nr0b1* and KLFs other than *Klf4*.

Lastly, many genome-scale datasets for human are currently being generated, primarily via the Encyclopedia of DNA Elements (ENCODE; `http://genome.ucsc.edu/ENCODE`) project and the Human Epigenome Project (HEP; `http://www.epigenome.org`). Consequently, there are now equivalent human datasets for many of the DBPs and HMs I have studied in mouse – in fact, there might be more by now and the data is consistently of excellent quality. It will be interesting to see how regulatory mechanisms translate from mouse to human.

6.3 Concluding Remarks

"It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity [..]"*, in short, it was a Ph.D. With the presentation of this thesis a long and laborious, yet often joyful and inspiring journey comes to a conclusion (or so I hope). The work I have been doing over the last years has driven me into the depths of biology, only to reemerge to light with ever more questions than I have had before. During all these years, I have spent much time developing and optimising methods, laying the groundwork that would enable me to ask those questions that had motivated me in the beginning. In the meantime, the field had moved forward a lot and I was excited to find more and more data being generated that I could use in my endeavours. We have now reached a point at which findings from many different aspects of stem cell biology can be fit into a larger, albeit certainly not yet complete picture and I have attempted to contribute just a little first step into this direction. The future is bright and new insight is close. I look forward, with excitement, to what the coming years will bring with them.

*From Charles Dickens, "A Tale of Two Cities"

Chapter 7

Bibliography

[1] S. Aerts, P. Van Loo, G. Thijs, Y. Moreau, and B. De Moor. Computational detection of cis -regulatory modules. *Bioinformatics*, 19 Suppl 2:5–14, Oct 2003.

[2] R. Ahuja, R. Pinyol, N. Reichenbach, L. Custer, J. Klingensmith, M. M. Kessels, and B. Qualmann. Cordon-bleu is an actin nucleation factor and controls neuronal morphology. *Cell*, 131(2):337–350, Oct 2007.

[3] M. Alawi, S. Kurtz, and M. Beckstette. CASSys: an integrated software-system for the interactive analysis of ChIP-seq data. *J Integr Bioinform*, 8:155, 2011.

[4] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, Oct 1990.

[5] D. C. Ambrosetti, C. Basilico, and L. Dailey. Synergistic activation of the fibroblast growth factor 4 enhancer by Sox2 and Oct-3 depends on protein-protein interactions facilitated by a specific spatial arrangement of factor binding sites. *Mol. Cell. Biol.*, 17:6321–6329, Nov 1997.

[6] D. C. Ambrosetti, H. R. Scholer, L. Dailey, and C. Basilico. Modulation of the activity of multiple transcriptional activation domains by the DNA binding domains mediates the synergistic action of Sox2 and Oct-3 on the fibroblast growth factor-4 enhancer. *J. Biol. Chem.*, 275:23387–23397, Jul 2000.

[7] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biol.*, 11:R106, 2010.

[8] F. Anokye-Danso, C. M. Trivedi, D. Juhr, M. Gupta, Z. Cui, Y. Tian, Y. Zhang, W. Yang, P. J. Gruber, J. A. Epstein, and E. E. Morrisey. Highly efficient miRNA-mediated reprogramming of mouse and human somatic cells to pluripotency. *Cell Stem Cell*, 8:376–388, Apr 2011.

[9] W. J. Ansorge. Next-generation DNA sequencing techniques. *N Biotechnol*, 25:195–203, 2009.

[10] Y. Araki, Z. Wang, C. Zang, W. H. Wood, D. Schones, K. Cui, T. Y. Roh, B. Lhotsky, R. P. Wersto, W. Peng, K. G. Becker, K. Zhao, and N. P. Weng. Genome-wide analysis of histone methylation reveals chromatin state-based regulation of gene transcription and function of memory CD8+ T cells. *Immunity*, 30:912–925, Jun 2009.

[11] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25:25–29, May 2000.

[12] S. Assou, T. Le Carrour, S. Tondeur, S. Strom, A. Gabelle, S. Marty, L. Nadal, V. Pantesco, T. Reme, J. P. Hugnot, S. Gasca, O. Hovatta, S. Hamamah, B. Klein, and J. De Vos. A meta-analysis of human embryonic stem cells transcriptome integrated into a web-based expression atlas. *Stem Cells*, 25(4):961–973, Apr 2007.

[13] J. Aubert, H. Dunstan, I. Chambers, and A. Smith. Functional gene screening in embryonic stem cells implicates Wnt antagonism in neural differentiation. *Nat. Biotechnol.*, 20:1240–1245, Dec 2002.

[14] V. Azuara, P. Perry, S. Sauer, M. Spivakov, H. F. J?rgensen, R. M. John, M. Gouti, M. Casanova, G. Warnes, M. Merkenschlager, and A. G. Fisher. Chromatin signatures of pluripotent cell lines. *Nat. Cell Biol.*, 8:532–538, May 2006.

[15] Š. Baebler, M. Hren, M. Camloh, M. Ravnikar, B. Bohanec, I. Plaper, R. Ucman, and J. Žel. Establishment of cell suspension cultures of yew (taxus× media rehd. and assessment of their genomic stability. *In Vitro Cellular & Developmental Biology-Plant*, 41(3):338–343, 2005.

[16] T. L. Bailey. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, 27:1653–1659, Jun 2011.

[17] R. Bajpai, D. A. Chen, A. Rada-Iglesias, J. Zhang, Y. Xiong, J. Helms, C. P. Chang, Y. Zhao, T. Swigut, and J. Wysocka. CHD7 cooperates with PBAF to control multipotent neural crest formation. *Nature*, 463(7283):958–962, Feb 2010.

[18] N. Ballas, C. Grunseich, D. D. Lu, J. C. Speh, and G. Mandel. REST and its corepressors mediate plasticity of neuronal gene chromatin throughout neurogenesis. *Cell*, 121:645–657, May 2005.

[19] C. Banchio, S. Lingrell, and D. E. Vance. Role of histone deacetylase in the expression of CTP:phosphocholine cytidylyltransferase alpha. *J. Biol. Chem.*, 281(15):10010–10015, Apr 2006.

[20] Y. Barash, J. A. Calarco, W. Gao, Q. Pan, X. Wang, O. Shai, B. J. Blencowe, and B. J. Frey. Deciphering the splicing code. *Nature*, 465:53–59, May 2010.

[21] N. Barker, A. Hurlstone, H. Musisi, A. Miles, M. Bienz, and H. Clevers. The chromatin remodelling factor Brg-1 interacts with beta-catenin to promote target gene activation. *EMBO J.*, 20(17):4935–4943, Sep 2001.

[22] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muertter, M. Holko, O. Ayanbule, A. Yefanov, and A. Soboleva. NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic Acids Res.*, 39:D1005–1010, Jan 2011.

[23] A. Barroso-del Jesus, G. Lucena-Aguilar, and P. Menendez. The miR-302-367 cluster as a potential stemness regulator in ESCs. *Cell Cycle*, 8(3):394–398, Feb 2009.

[24] A. Barski, I. Chepelev, D. Liko, S. Cuddapah, A. B. Fleming, J. Birch, K. Cui, R. J. White, and K. Zhao. Pol II and its associated epigenetic marks are present at Pol III-transcribed noncoding RNA genes. *Nat. Struct. Mol. Biol.*, 17:629–634, May 2010.

[25] A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129:823–837, May 2007.

[26] A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129:823–837, May 2007.

[27] A. Barski and K. Zhao. Genomic location analysis by ChIP-Seq. *J. Cell. Biochem.*, 107:11–18, May 2009.

[28] C. Baumann and R. De La Fuente. ATRX marks the inactive X chromosome (Xi) in somatic cells and during imprinted X chromosome inactivation in trophoblast stem cells. *Chromosoma*, 118(2):209–222, Apr 2009.

[29] G. M. Beattie, A. D. Lopez, N. Bucay, A. Hinton, M. T. Firpo, C. C. King, and A. Hayek. Activin A maintains pluripotency of human embryonic stem cells in the absence of feeder layers. *Stem Cells*, 23(4):489–495, Apr 2005.

[30] R. S. Beddington and E. J. Robertson. Axis development and early asymmetry in mammals. *Cell*, 96:195–209, Jan 1999.

[31] E. Ben-Shushan, J. R. Thompson, L. J. Gudas, and Y. Bergman. Rex-1, a gene encoding a transcription factor expressed in the early embryo, is regulated via Oct-3/4 and Oct-6 binding to an octamer site and a novel protein, Rox-1, binding to an adjacent site. *Mol. Cell. Biol.*, 18:1866–1878, Apr 1998.

[32] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[33] S. L. Berger. The complex language of chromatin regulation during transcription. *Nature*, 447:407–412, May 2007.

[34] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28(1):235–242, Jan 2000.

[35] B. E. Bernstein, T. S. Mikkelsen, X. Xie, M. Kamal, D. J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, R. Jaenisch, A. Wagschal, R. Feil, S. L. Schreiber, and E. S. Lander. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125:315–326, Apr 2006.

[36] T. H. Bestor. The DNA methyltransferases of mammals. *Hum. Mol. Genet.*, 9:2395–2402, Oct 2000.

[37] N. Bhutani, J. J. Brady, M. Damian, A. Sacco, S. Y. Corbel, and H. M. Blau. Reprogramming towards pluripotency requires AID-dependent DNA demethylation. *Nature*, 463:1042–1047, Feb 2010.

[38] A. Bird. DNA methylation patterns and epigenetic memory. *Genes Dev.*, 16:6–21, Jan 2002.

[39] E. Birney, J. A. Stamatoyannopoulos, A. Dutta, R. Guigo, T. R. Gingeras, E. H. Margulies, Z. Weng, M. Snyder, E. T. Dermitzakis, R. E. Thurman, M. S. Kuehn, C. M. Taylor, S. Neph, C. M. Koch, S. Asthana, A. Malhotra, I. Adzhubei, J. A. Greenbaum, R. M. Andrews, P. Flicek, P. J. Boyle, H. Cao, N. P. Carter, G. K. Clelland, S. Davis, N. Day, P. Dhami, S. C. Dillon, M. O. Dorschner, H. Fiegler, P. G. Giresi, J. Goldy, M. Hawrylycz, A. Haydock, R. Humbert, K. D. James, B. E. Johnson, E. M. Johnson, T. T. Frum, E. R. Rosenzweig, N. Karnani, K. Lee, G. C. Lefebvre, P. A. Navas, F. Neri, S. C. Parker, P. J. Sabo, R. Sandstrom, A. Shafer, D. Vetrie, M. Weaver, S. Wilcox, M. Yu, F. S. Collins,

J. Dekker, J. D. Lieb, T. D. Tullius, G. E. Crawford, S. Sunyaev, W. S. Noble, I. Dunham, F. Denoeud, A. Reymond, P. Kapranov, J. Rozowsky, D. Zheng, R. Castelo, A. Frankish, J. Harrow, S. Ghosh, A. Sandelin, I. L. Hofacker, R. Baertsch, D. Keefe, S. Dike, J. Cheng, H. A. Hirsch, E. A. Sekinger, J. Lagarde, J. F. Abril, A. Shahab, C. Flamm, C. Fried, J. Hackermuller, J. Hertel, M. Lindemeyer, K. Missal, A. Tanzer, S. Washietl, J. Korbel, O. Emanuelsson, J. S. Pedersen, N. Holroyd, R. Taylor, D. Swarbreck, N. Matthews, M. C. Dickson, D. J. Thomas, M. T. Weirauch, J. Gilbert, J. Drenkow, I. Bell, X. Zhao, K. G. Srinivasan, W. K. Sung, H. S. Ooi, K. P. Chiu, S. Foissac, T. Alioto, M. Brent, L. Pachter, M. L. Tress, A. Valencia, S. W. Choo, C. Y. Choo, C. Ucla, C. Manzano, C. Wyss, E. Cheung, T. G. Clark, J. B. Brown, M. Ganesh, S. Patel, H. Tammana, J. Chrast, C. N. Henrichsen, C. Kai, J. Kawai, U. Nagalakshmi, J. Wu, Z. Lian, J. Lian, P. Newburger, X. Zhang, P. Bickel, J. S. Mattick, P. Carninci, Y. Hayashizaki, S. Weissman, T. Hubbard, R. M. Myers, J. Rogers, P. F. Stadler, T. M. Lowe, C. L. Wei, Y. Ruan, K. Struhl, M. Gerstein, S. E. Antonarakis, Y. Fu, E. D. Green, U. Karaoz, A. Siepel, J. Taylor, L. A. Liefer, K. A. Wetterstrand, P. J. Good, E. A. Feingold, M. S. Guyer, G. M. Cooper, G. Asimenos, C. N. Dewey, M. Hou, S. Nikolaev, J. I. Montoya-Burgos, A. Loytynoja, S. Whelan, F. Pardi, T. Massingham, H. Huang, N. R. Zhang, I. Holmes, J. C. Mullikin, A. Ureta-Vidal, B. Paten, M. Seringhaus, D. Church, K. Rosenbloom, W. J. Kent, E. A. Stone, S. Batzoglou, N. Goldman, R. C. Hardison, D. Haussler, W. Miller, A. Sidow, N. D. Trinklein, Z. D. Zhang, L. Barrera, R. Stuart, D. C. King, A. Ameur, S. Enroth, M. C. Bieda, J. Kim, A. A. Bhinge, N. Jiang, J. Liu, F. Yao, V. B. Vega, C. W. Lee, P. Ng, A. Shahab, A. Yang, Z. Moqtaderi, Z. Zhu, X. Xu, S. Squazzo, M. J. Oberley, D. Inman, M. A. Singer, T. A. Richmond, K. J. Munn, A. Rada-Iglesias, O. Wallerman, J. Komorowski, J. C. Fowler, P. Couttet, A. W. Bruce, O. M. Dovey, P. D. Ellis, C. F. Langford, D. A. Nix, G. Euskirchen, S. Hartman, A. E. Urban, P. Kraus, S. Van Calcar, N. Heintzman, T. H. Kim, K. Wang, C. Qu, G. Hon, R. Luna, C. K. Glass, M. G. Rosenfeld, S. F. Aldred, S. J. Cooper, A. Halees, J. M. Lin, H. P. Shulha, X. Zhang, M. Xu, J. N. Haidar, Y. Yu, Y. Ruan, V. R. Iyer, R. D. Green, C. Wadelius, P. J. Farnham, B. Ren, R. A. Harte, A. S. Hinrichs, H. Trumbower, H. Clawson, J. Hillman-Jackson, A. S. Zweig, K. Smith, A. Thakkapallayil, G. Barber, R. M. Kuhn, D. Karolchik, L. Armengol, C. P. Bird, P. I. de Bakker, A. D. Kern, N. Lopez-Bigas, J. D. Martin, B. E. Stranger, A. Woodroffe, E. Davydov, A. Dimas, E. Eyras, I. B. Hallgrimsdottir, J. Huppert, M. C. Zody, G. R. Abecasis, X. Estivill, G. G. Bouffard, X. Guan, N. F. Hansen, J. R. Idol, V. V. Maduro, B. Maskeri, J. C. McDowell, M. Park, P. J. Thomas, A. C. Young, R. W. Blakesley, D. M. Muzny, E. Sodergren, D. A. Wheeler, K. C. Worley, H. Jiang, G. M. Weinstock, R. A. Gibbs, T. Graves, R. Fulton, E. R. Mardis, R. K. Wilson, M. Clamp, J. Cuff, S. Gnerre, D. B. Jaffe, J. L. Chang, K. Lindblad-Toh, E. S. Lander, M. Koriabine, M. Nefedov, K. Osoegawa, Y. Yoshinaga, B. Zhu, and P. J. de Jong. Identification and analysis of functional elements in 1human genome by the ENCODE pilot project. *Nature*, 447:799–816, Jun 2007.

[40] I. Birol, S. D. Jackman, C. B. Nielsen, J. Q. Qian, R. Varhol, G. Stazyk, R. D. Morin, Y. Zhao, M. Hirst, J. E. Schein, D. E. Horsman, J. M. Connors, R. D. Gascoyne, M. A. Marra, and S. J. Jones. De novo transcriptome assembly with ABySS. *Bioinformatics*, 25:2872–2877, Nov 2009.

[41] A. K. Biswas and D. G. Johnson. Transcriptional and nontranscriptional functions of E2F1 in response to DNA damage. *Cancer Res.*, 72(1):13–17, Jan 2012.

[42] K. Blair, J. Wray, and A. Smith. The liberation of embryonic stem cells. *PLoS Genet.*, 7:e1002019, Apr 2011.

[43] Daniel Blankenberg, Gregory Von Kuster, Nathaniel Coraor, Guruprasad Ananda, Ross Lazarus, Mary Mangan, Anton Nekrutenko, and James Taylor. *Galaxy: A Web-Based Genome Analysis Tool for Experimentalists*. John Wiley & Sons, Inc., 2010.

[44] H. M. Blau, C. P. Chiu, and C. Webster. Cytoplasmic activation of human nuclear genes in stable heterocaryons. *Cell*, 32:1171–1180, Apr 1983.

[45] R. Blekhman, J. C. Marioni, P. Zumbo, M. Stephens, and Y. Gilad. Sex-specific and lineage-specific alternative splicing in primates. *Genome Res.*, 20:180–189, Feb 2010.

[46] G. M. Borchert, W. Lanier, and B. L. Davidson. RNA polymerase III transcribes human microRNAs. *Nat. Struct. Mol. Biol.*, 13:1097–1101, Dec 2006.

[47] O. A. Botrugno, E. Fayard, J. S. Annicotte, C. Haby, T. Brennan, O. Wendling, T. Tanaka, T. Kodama, W. Thomas, J. Auwerx, and K. Schoonjans. Synergy between LRH-1 and beta-catenin induces G1 cyclin-mediated cell proliferation. *Mol. Cell*, 15:499–509, Aug 2004.

[48] D. Bottomly, S. L. Kyler, S. K. McWeeney, and G. S. Yochum. Identification of beta-catenin binding regions in colon cancer cells using ChIP-Seq. *Nucleic Acids Res.*, 38:5735–5745, Sep 2010.

[49] L. A. Boyer, T. I. Lee, M. F. Cole, S. E. Johnstone, S. S. Levine, J. P. Zucker, M. G. Guenther, R. M. Kumar, H. L. Murray, R. G. Jenner, D. K. Gifford, D. A. Melton, R. Jaenisch, and R. A. Young. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, 122:947–956, Sep 2005.

[50] L. A. Boyer, D. Mathur, and R. Jaenisch. Molecular control of pluripotency. *Curr. Opin. Genet. Dev.*, 16:455–462, Oct 2006.

[51] L. A. Boyer, K. Plath, J. Zeitlinger, T. Brambrink, L. A. Medeiros, T. I. Lee, S. S. Levine, M. Wernig, A. Tajonar, M. K. Ray, G. W. Bell, A. P. Otte, M. Vidal, D. K. Gifford, R. A. Young, and R. Jaenisch. Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature*, 441(7091):349–353, May 2006.

[52] A. Bradley, K. Anastassiadis, A. Ayadi, J. F. Battey, C. Bell, M. C. Birling, J. Bottomley, S. D. Brown, A. Burger, C. J. Bult, W. Bushell, F. S. Collins, C. Desaintes, B. Doe, A. Economides, J. T. Eppig, R. H. Finnell, C. Fletcher, M. Fray, D. Frendewey, R. H. Friedel, F. G. Grosveld, J. Hansen, Y. Herault, G. Hicks, A. Horlein, R. Houghton, M. Hrabe de Angelis, D. Huylebroeck, V. Iyer, P. J. de Jong, J. A. Kadin, C. Kaloff, K. Kennedy, M. Koutsourakis, K. C. Kent Lloyd, S. Marschall, J. Mason, C. McKerlie, M. P. McLeod, H. von Melchner, M. Moore, A. O. Mujica, A. Nagy, M. Nefedov, L. M. Nutter, G. Pavlovic, J. L. Peterson, J. Pollock, R. Ramirez-Solis, D. E. Rancourt, M. Raspa, J. E. Remacle, M. Ringwald, B. Rosen, N. Rosenthal, J. Rossant, P. Ruiz Noppinger, E. Ryder, J. Z. Schick, F. Schnutgen, P. Schofield, C. Seisenberger, M. Selloum, E. M. Simpson, W. C. Skarnes, D. Smedley, W. L. Stanford, A. Francis Stewart, K. Stone, K. Swan, H. Tadepally, L. Teboul, G. P. Tocchini-Valentini, D. Valenzuela, A. P. West, K. I. Yamamura, Y. Yoshinaga, and W. Wurst. The mammalian gene function resource: the international knockout mouse consortium. *Mamm. Genome*, Sep 2012.

[53] C. Bradney, M. Hjelmeland, Y. Komatsu, M. Yoshida, T. P. Yao, and Y. Zhuang. Regulation of E2A activities by histone acetyltransferases in B lymphocyte development. *J. Biol. Chem.*, 278(4):2370–2376, Jan 2003.

[54] T. Brambrink, R. Foreman, G. G. Welstead, C. J. Lengner, M. Wernig, H. Suh, and R. Jaenisch. Sequential expression of pluripotency markers during direct reprogramming of mouse somatic cells. *Cell Stem Cell*, 2:151–159, Feb 2008.

[55] I. G. Brons, L. E. Smithers, M. W. Trotter, P. Rugg-Gunn, B. Sun, S. M. Chuva de Sousa Lopes, S. K. Howlett, A. Clarkson, L. Ahrlund-Richter, R. A. Pedersen, and L. Vallier. Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature*, 448:191–195, Jul 2007.

[56] F. A. Brook and R. L. Gardner. The origin and efficient derivation of embryonic stem cells in the mouse. *Proc. Natl. Acad. Sci. U.S.A.*, 94:5709–5712, May 1997.

[57] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11:94, 2010.

[58] T. Burdon, C. Stracey, I. Chambers, J. Nichols, and A. Smith. Suppression of SHP-2 and ERK signalling promotes self-renewal of mouse embryonic stem cells. *Dev. Biol.*, 210:30–43, Jun 1999.

[59] M. N. Cabili, C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, and J. L. Rinn. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.*, 25:1915–1927, Sep 2011.

[60] A. R. Cao, R. Rabinovich, M. Xu, X. Xu, V. X. Jin, and P. J. Farnham. Genome-wide analysis of transcription factor E2F1 mutant proteins reveals that N- and C-terminal protein interaction domains do not participate in targeting E2F1 to the human genome. *J. Biol. Chem.*, 286:11985–11996, Apr 2011.

[61] R. Cao, L. Wang, H. Wang, L. Xia, H. Erdjument-Bromage, P. Tempst, R. S. Jones, and Y. Zhang. Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science*, 298:1039–1043, Nov 2002.

[62] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, 7:335–336, May 2010.

[63] P. Carninci, T. Kasukawa, S. Katayama, J. Gough, M. C. Frith, N. Maeda, R. Oyama, T. Ravasi, B. Lenhard, C. Wells, R. Kodzius, K. Shimokawa, V. B. Bajic, S. E. Brenner, S. Batalov, A. R. Forrest, M. Zavolan, M. J. Davis, L. G. Wilming, V. Aidinis, J. E. Allen, A. Ambesi-Impiombato, R. Apweiler, R. N. Aturaliya, T. L. Bailey, M. Bansal, L. Baxter, K. W. Beisel, T. Bersano, H. Bono, A. M. Chalk, K. P. Chiu, V. Choudhary, A. Christoffels, D. R. Clutterbuck, M. L. Crowe, E. Dalla, B. P. Dalrymple, B. de Bono, G. Della Gatta, D. di Bernardo, T. Down, P. Engstrom, M. Fagiolini, G. Faulkner, C. F. Fletcher, T. Fukushima, M. Furuno, S. Futaki, M. Gariboldi, P. Georgii-Hemming, T. R. Gingeras, T. Gojobori, R. E. Green, S. Gustincich, M. Harbers, Y. Hayashi, T. K. Hensch, N. Hirokawa, D. Hill, L. Huminiecki, M. Iacono, K. Ikeo, A. Iwama, T. Ishikawa, M. Jakt, A. Kanapin, M. Katoh, Y. Kawasawa, J. Kelso, H. Kitamura, H. Kitano, G. Kollias, S. P. Krishnan, A. Kruger, S. K. Kummerfeld, I. V. Kurochkin, L. F. Lareau, D. Lazarevic, L. Lipovich, J. Liu, S. Liuni, S. McWilliam, M. Madan Babu, M. Madera, L. Marchionni, H. Matsuda, S. Matsuzawa, H. Miki, F. Mignone, S. Miyake, K. Morris, S. Mottagui-Tabar, N. Mulder, N. Nakano, H. Nakauchi, P. Ng, R. Nilsson, S. Nishiguchi, S. Nishikawa, F. Nori, O. Ohara, Y. Okazaki, V. Orlando, K. C. Pang, W. J. Pavan, G. Pavesi, G. Pesole, N. Petrovsky, S. Piazza, J. Reed, J. F. Reid, B. Z. Ring, M. Ringwald, B. Rost, Y. Ruan, S. L. Salzberg, A. Sandelin, C. Schneider, C. Schonbach, K. Sekiguchi, C. A. Semple, S. Seno, L. Sessa, Y. Sheng, Y. Shibata, H. Shimada, K. Shimada, D. Silva, B. Sinclair, S. Sperling, E. Stupka, K. Sugiura, R. Sultana, Y. Takenaka, K. Taki, K. Tammoja, S. L. Tan, S. Tang, M. S. Taylor, J. Tegner, S. A. Teichmann, H. R. Ueda, E. van Nimwegen, R. Verardo, C. L. Wei, K. Yagi, H. Yamanishi, E. Zabarovsky, S. Zhu, A. Zimmer, W. Hide, C. Bult, S. M. Grimmond, R. D. Teasdale, E. T. Liu, V. Brusic, J. Quackenbush, C. Wahlestedt, J. S. Mattick, D. A. Hume, C. Kai, D. Sasaki, Y. Tomaru, S. Fukuda, M. Kanamori-Katayama, M. Suzuki, J. Aoki, T. Arakawa, J. Iida, K. Imamura, M. Itoh, T. Kato, H. Kawaji, N. Kawagashira, T. Kawashima, M. Kojima, S. Kondo, H. Konno, K. Nakano, N. Ninomiya, T. Nishio, M. Okada, C. Plessy, K. Shibata, T. Shiraki, S. Suzuki, M. Tagami, K. Waki, A. Watahiki, Y. Okamura-Oho, H. Suzuki, J. Kawai, and Y. Hayashizaki. The transcriptional landscape of the mammalian genome. *Science*, 309:1559–1563, Sep 2005.

[64] P. Cartwright, C. McLean, A. Sheppard, D. Rivett, K. Jones, and S. Dalton. LIF/STAT3 controls ES cell self-renewal and pluripotency by a Myc-dependent mechanism. *Development*, 132:885–896, Mar 2005.

[65] S. J. Chamberlain and C. I. Brannan. The Prader-Willi syndrome imprinting center activates the paternally expressed murine Ube3a antisense transcript but represses paternal Ube3a. *Genomics*, 73(3):316–322, May 2001.

[66] I. Chambers. The molecular basis of pluripotency in mouse embryonic stem cells. *Cloning Stem Cells*, 6:386–391, 2004.

[67] I. Chambers, D. Colby, M. Robertson, J. Nichols, S. Lee, S. Tweedie, and A. Smith. Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell*, 113:643–655, May 2003.

[68] I. Chambers, J. Silva, D. Colby, J. Nichols, B. Nijmeijer, M. Robertson, J. Vrana, K. Jones, L. Grotewold, and A. Smith. Nanog safeguards pluripotency and mediates germline development. *Nature*, 450:1230–1234, Dec 2007.

[69] I. Chambers and A. Smith. Self-renewal of teratocarcinoma and embryonic stem cells. *Oncogene*, 23:7150–7160, Sep 2004.

[70] I. Chambers and S. R. Tomlinson. The transcriptional foundation of pluripotency. *Development*, 136:2311–2322, Jul 2009.

[71] D. Chaya, T. Hayamizu, M. Bustin, and K. S. Zaret. Transcription factor FoxA (HNF3) on a nucleosome at an enhancer complex in liver chromatin. *J. Biol. Chem.*, 276(48):44385–44389, Nov 2001.

[72] C. Chen, D. Ridzon, C. T. Lee, J. Blake, Y. Sun, and W. M. Strauss. Defining embryonic stem cell identity using differentiation-related microRNAs and their potential targets. *Mamm. Genome*, 18(5):316–327, May 2007.

[73] G. Chen and Q. Zhou. Searching ChIP-seq genomic islands for combinatorial regulatory codes in mouse embryonic stem cells. *BMC Genomics*, 12:515, Oct 2011.

[74] L. Chen, G. Wu, and H. Ji. hmChIP: a database and web server for exploring publicly available human and mouse ChIP-seq and ChIP-chip data. *Bioinformatics*, 27:1447–1448, May 2011.

[75] X. Chen, H. Xu, P. Yuan, F. Fang, M. Huss, V. B. Vega, E. Wong, Y. L. Orlov, W. Zhang, J. Jiang, Y. H. Loh, H. C. Yeo, Z. X. Yeo, V. Narang, K. R. Govindarajan, B. Leong, A. Shahab, Y. Ruan, G. Bourque, W. K. Sung, N. D. Clarke, C. L. Wei, and H. H. Ng. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133:1106–1117, Jun 2008.

[76] C. Cheng and M. Gerstein. Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Res.*, 40:553–568, Jan 2012.

[77] P. Chi, Y. Chen, L. Zhang, X. Guo, J. Wongvipat, T. Shamu, J. A. Fletcher, S. Dewell, R. G. Maki, D. Zheng, C. R. Antonescu, C. D. Allis, and C. L. Sawyers. ETV1 is a lineage survival factor that cooperates with KIT in gastrointestinal stromal tumours. *Nature*, 467:849–853, Oct 2010.

[78] S. W. Chi, J. B. Zang, A. Mele, and R. B. Darnell. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 460(7254):479–486, Jul 2009.

[79] N. Y. Chia, Y. S. Chan, B. Feng, X. Lu, Y. L. Orlov, D. Moreau, P. Kumar, L. Yang, J. Jiang, M. S. Lau, M. Huss, B. S. Soh, P. Kraus, P. Li, T. Lufkin, B. Lim, N. D. Clarke, F. Bard, and H. H. Ng. A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. *Nature*, 468(7321):316–320, Nov 2010.

[80] P. M. Chiang, J. Ling, Y. H. Jeong, D. L. Price, S. M. Aja, and P. C. Wong. Deletion of TDP-43 down-regulates Tbc1d1, a gene linked to obesity, and alters body fat metabolism. *Proc. Natl. Acad. Sci. U.S.A.*, 107:16320–16324, Sep 2010.

[81] D.M. Chikina and G.O. Troyanskaya. An effective statistical evaluation of ChIPseq dataset similarity. *Bioinformatics*, 28:607–613, Mar 2012.

[82] J. L. Chong, P. L. Wenzel, M. T. Saenz-Robles, V. Nair, A. Ferrey, J. P. Hagan, Y. M. Gomez, N. Sharma, H. Z. Chen, M. Ouseph, S. H. Wang, P. Trikha, B. Culp, L. Mezache, D. J. Winton, O. J. Sansom, D. Chen, R. Bremner, P. G. Cantalupo, M. L. Robinson, J. M. Pipas, and G. Leone. E2f1-3 switch from activators in progenitor cells to repressors in differentiating cells. *Nature*, 462(7275):930–934, Dec 2009.

[83] L. S. Chuang, H. I. Ian, T. W. Koh, H. H. Ng, G. Xu, and B. F. Li. Human DNA-(cytosine-5) methyltransferase-PCNA complex as a target for p21WAF1. *Science*, 277:1996–2000, Sep 1997.

[84] N. Cloonan, A. R. Forrest, G. Kolle, B. B. Gardiner, G. J. Faulkner, M. K. Brown, D. F. Taylor, A. L. Steptoe, S. Wani, G. Bethel, A. J. Robertson, A. C. Perkins, S. J. Bruce, C. C. Lee, S. S. Ranade, H. E. Peckham, J. M. Manning, K. J. McKernan, and S. M. Grimmond. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods*, 5:613–619, Jul 2008.

[85] P. J. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.*, 38:1767–1771, Apr 2010.

[86] K. Cockburn and J. Rossant. Making the blastocyst: lessons from the mouse. *J. Clin. Invest.*, 120:995–1003, Apr 2010.

[87] M. F. Cole, S. E. Johnstone, J. J. Newman, M. H. Kagey, and R. A. Young. Tcf3 is an integral component of the core regulatory circuitry of embryonic stem cells. *Genes Dev.*, 22(6):746–755, Mar 2008.

[88] P. Collas. The current state of chromatin immunoprecipitation. *Mol. Biotechnol.*, 45:87–100, May 2010.

[89] R. C. Conaway, S. Sato, C. Tomomori-Sato, T. Yao, and J. W. Conaway. The mammalian Mediator complex and its role in transcriptional regulation. *Trends Biochem. Sci.*, 30:250–255, May 2005.

[90] R. M. Cook-Deegan. The Alta summit, December 1984. *Genomics*, 5:661–663, Oct 1989.

[91] L. J. Core, J. J. Waterfall, and J. T. Lis. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322:1845–1848, Dec 2008.

[92] L. S. Correa-Cerro, Y. Piao, A. A. Sharov, A. Nishiyama, J. S. Cadet, H. Yu, L. V. Sharova, L. Xin, H. G. Hoang, M. Thomas, Y. Qian, D. B. Dudekula, E. Meyers, B. Y. Binder, G. Mowrer, U. Bassey, D. L. Longo, D. Schlessinger, and M. S. Ko. Generation of mouse ES cell lines engineered for the forced induction of transcription factors. *Sci Rep*, 1:167, 2011.

[93] I. G. Costa, H. G. Roider, T. G. do Rego, and F. d. e. A. de Carvalho. Predicting gene expression in T cell differentiation from histone modifications and transcription factor binding affinities by linear mixture models. *BMC Bioinformatics*, 12 Suppl 1:S29, 2011.

[94] R. Cotterman, V. X. Jin, S. R. Krig, J. M. Lemen, A. Wey, P. J. Farnham, and P. S. Knoepfler. N-Myc regulates a widespread euchromatic program in the human genome partially independent of its role as a classical transcription factor. *Cancer Res.*, 68:9654–9662, Dec 2008.

[95] C. A. Cowan, J. Atienza, D. A. Melton, and K. Eggan. Nuclear reprogramming of somatic cells after fusion with human embryonic stem cells. *Science*, 309:1369–1373, Aug 2005.

[96] M. P. Cox, D. A. Peterson, and P. J. Biggs. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics*, 11:485, 2010.

[97] R. Croteau, R. E. Ketchum, R. M. Long, R. Kaspera, and M. R. Wildung. Taxol biosynthesis and molecular genetics. *Phytochem Rev*, 5:75–97, Feb 2006.

[98] P. Cui, Q. Lin, F. Ding, C. Xin, W. Gong, L. Zhang, J. Geng, B. Zhang, X. Yu, J. Yang, S. Hu, and J. Yu. A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. *Genomics*, 96:259–265, Nov 2010.

[99] J. S. Cumbie, J. A. Kimbrel, Y. Di, D. W. Schafer, L. J. Wilhelm, S. E. Fox, C. M. Sullivan, A. D. Curzon, J. C. Carrington, T. C. Mockler, and J. H. Chang. GENE-counter: a computational pipeline for the analysis of RNA-Seq data for gene expression differences. *PLoS ONE*, 6:e25279, 2011.

[100] D. Dahary, O. Elroy-Stein, and R. Sorek. Naturally occurring antisense: transcriptional leakage or real overlap? *Genome Res.*, 15:364–368, Mar 2005.

[101] K. D. Dahlquist, N. Salomonis, K. Vranizan, S. C. Lawlor, and B. R. Conklin. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat. Genet.*, 31:19–20, May 2002.

[102] K. Daily, P. Rigor, S. Christley, X. Xie, and P. Baldi. Data structures and compression algorithms for high-throughput sequencing technologies. *BMC Bioinformatics*, 11:514, 2010.

[103] M. De Gobbi, D. Garrick, M. Lynch, D. Vernimmen, J. R. Hughes, N. Goardon, S. Luc, K. M. Lower, J. A. Sloane-Stanley, C. Pina, S. Soneji, R. Renella, T. Enver, S. Taylor, S. E. Jacobsen, P. Vyas, R. J. Gibbons, and D. R. Higgs. Generation of bivalent chromatin domains during cell fate decisions. *Epigenetics Chromatin*, 4(1):9, 2011.

[104] M. de Hoon and Y. Hayashizaki. Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference. *BioTechniques*, 44:627–628, Apr 2008.

[105] E. Decker, C. Durand, S. Bender, C. Rodelsperger, A. Glaser, J. Hecht, K. U. Schneider, and G. Rappold. FGFR3 is a target of the homeobox transcription factor SHOX in limb development. *Hum. Mol. Genet.*, 20:1524–1535, Apr 2011.

[106] M. Dejosez, J. S. Krumenacker, L. J. Zitur, M. Passeri, L. F. Chu, Z. Songyang, J. A. Thomson, and T. P. Zwaka. Ronin is essential for embryogenesis and the pluripotency of mouse embryonic stem cells. *Cell*, 133:1162–1174, Jun 2008.

[107] F. Denoeud, J. M. Aury, C. Da Silva, B. Noel, O. Rogier, M. Delledonne, M. Morgante, G. Valle, P. Wincker, C. Scarpelli, O. Jaillon, and F. Artiguenave. Annotating genomes with massive-scale RNA sequencing. *Genome Biol.*, 9:R175, 2008.

[108] G. Dieci, G. Fiorino, M. Castelnuovo, M. Teichmann, and A. Pagano. The expanding RNA polymerase III transcriptome. *Trends Genet.*, 23:614–622, Dec 2007.

[109] S. Diecke, A. Quiroga-Negreira, T. Redmer, and D. Besser. FGF2 signaling in mouse embryonic fibroblasts is crucial for self-renewal of embryonic stem cells. *Cells Tissues Organs (Print)*, 188:52–61, 2008.

[110] J. T. Do and H. R. Scholer. Regulatory circuits underlying pluripotency and reprogramming. *Trends Pharmacol. Sci.*, 30:296–302, Jun 2009.

[111] D. Dominguez-Sola, C. Y. Ying, C. Grandori, L. Ruggiero, B. Chen, M. Li, D. A. Galloway, W. Gu, J. Gautier, and R. Dalla-Favera. Non-transcriptional control of DNA replication by c-Myc. *Nature*, 448:445–451, Jul 2007.

[112] D. Dong, X. Shao, and Z. Zhang. Differential effects of chromatin regulators and transcription factors on gene regulation: a nucleosomal perspective. *Bioinformatics*, 27:147–152, Jan 2011.

[113] T. Du and P. D. Zamore. microPrimer: the biogenesis and function of microRNA. *Development*, 132:4645–4652, Nov 2005.

[114] M. Dufva. Introduction to microarray technology. *Methods Mol. Biol.*, 529:1–22, 2009.

[115] S. Durinck, J. Bullard, P. T. Spellman, and S. Dudoit. GenomeGraphs: integrated genomic data visualization with R. *BMC Bioinformatics*, 10:2, 2009.

[116] J. D. Ebben, M. Zorniak, P. A. Clark, and J. S. Kuo. Introduction to induced pluripotent stem cells: advancing the potential for personalized medicine. *World Neurosurg*, 76:270–275, 2011.

[117] R. Eckner, M. E. Ewen, D. Newsome, M. Gerdes, J. A. DeCaprio, J. B. Lawrence, and D. M. Livingston. Molecular cloning and functional analysis of the adenovirus E1A-associated 300-kD protein (p300) reveals a protein with properties of a transcriptional adaptor. *Genes Dev.*, 8(8):869–884, Apr 1994.

[118] R. Edgar, M. Domrachev, and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, 30:207–210, Jan 2002.

[119] C. A. Edwards and A. C. Ferguson-Smith. Mechanisms regulating imprinted genes in clusters. *Curr. Opin. Cell Biol.*, 19:281–289, Jun 2007.

[120] J. A. Efe, X. Yuan, K. Jiang, and S. Ding. Development unchained: how cellular reprogramming is redefining our view of cell fate and identity. *Sci Prog*, 94:298–322, 2011.

[121] E. Engelen, U. Akinci, J. C. Bryne, J. Hou, C. Gontan, M. Moen, D. Szumska, C. Kockx, W. van Ijcken, D. H. Dekkers, J. Demmers, E. J. Rijkers, S. Bhattacharya, S. Philipsen, L. H. Pevny, F. G. Grosveld, R. J. Rottier, B. Lenhard, and R. A. Poot. Sox2 cooperates with Chd7 to regulate genes that are mutated in human syndromes. *Nat. Genet.*, 43(6):607–611, Jun 2011.

[122] J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, and B. E. Bernstein. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473:43–49, May 2011.

[123] M. J. Evans and M. H. Kaufman. Establishment in culture of pluripotential cells from mouse embryos. *Nature*, 292:154–156, Jul 1981.

[124] F. Faiola, X. Liu, S. Lo, S. Pan, K. Zhang, E. Lymar, A. Farina, and E. Martinez. Dual regulation of c-Myc by p300 via acetylation-dependent control of Myc protein turnover and coactivation of Myc-induced transcription. *Mol. Cell. Biol.*, 25(23):10220–10234, Dec 2005.

[125] G. Falco, S. L. Lee, I. Stanghellini, U. C. Bassey, T. Hamatani, and M. S. Ko. Zscan4: a novel gene expressed exclusively in late 2-cell embryos and embryonic stem cells. *Dev. Biol.*, 307:539–550, Jul 2007.

[126] J.M. Fang and Y.S. Cheng. Lignans, flavonoids and phenolic derivatives from taxus ma/re. *Journal of the Chinese Chemical Society*, 52:1999–811, 1999.

[127] C. R. Farthing, G. Ficz, R. K. Ng, C. F. Chan, S. Andrews, W. Dean, M. Hemberger, and W. Reik. Global mapping of DNA methylation in mouse promoters reveals epigenetic reprogramming of pluripotency genes. *PLoS Genet.*, 4:e1000116, Jun 2008.

[128] M. Fedurco, A. Romieu, S. Williams, I. Lawrence, and G. Turcatti. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res.*, 34:e22, 2006.

[129] A. P. Feinberg and B. Tycko. The history of cancer epigenetics. *Nat. Rev. Cancer*, 4:143–153, Feb 2004.

[130] A. P. Fejes, G. Robertson, M. Bilenky, R. Varhol, M. Bainbridge, and S. J. Jones. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, 24:1729–1730, Aug 2008.

[131] N. Feldman, A. Gerson, J. Fang, E. Li, Y. Zhang, Y. Shinkai, H. Cedar, and Y. Bergman. G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis. *Nat. Cell Biol.*, 8:188–194, Feb 2006.

[132] B. Feng, J. Jiang, P. Kraus, J. H. Ng, J. C. Heng, Y. S. Chan, L. P. Yaw, W. Zhang, Y. H. Loh, J. Han, V. B. Vega, V. Cacheux-Rataboul, B. Lim, T. Lufkin, and H. H. Ng. Reprogramming of fibroblasts into induced pluripotent stem cells with orphan nuclear receptor Esrrb. *Nat. Cell Biol.*, 11(2):197–203, Feb 2009.

[133] I. M. Fingerman, L. McDaniel, X. Zhang, W. Ratzat, T. Hassan, Z. Jiang, R. F. Cohen, and G. D. Schuler. NCBI Epigenomics: a new public resource for exploring epigenomic data sets. *Nucleic Acids Res.*, 39:D908–912, Jan 2011.

[134] K. Fisher and S. Turner. PXY, a receptor-like kinase essential for maintaining polarity during plant vascular-tissue development. *Curr. Biol.*, 17:1061–1066, Jun 2007.

[135] P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent, Y. Chen, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A. Kahari, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, P. Larsson, I. Longden, W. McLaren, B. Overduin, B. Pritchard, H. S. Riat, D. Rios, G. R. Ritchie, M. Ruffier, M. Schuster, D. Sobral, G. Spudich, Y. A. Tang, S. Trevanion, J. Vandrovcova, A. J. Vilella, S. White, S. P. Wilder, A. Zadissa, J. Zamora, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, T. J. Hubbard, A. Parker, G. Proctor, J. Vogel, and S. M. Searle. Ensembl 2011. *Nucleic Acids Res.*, 39:D800–806, Jan 2011.

[136] P. Flicek and E. Birney. Sense from sequence reads: methods for alignment and assembly. *Nat. Methods*, 6:S6–S12, Nov 2009.

[137] S. R. Frank, T. Parisi, S. Taubert, P. Fernandez, M. Fuchs, H. M. Chan, D. M. Livingston, and B. Amati. MYC recruits the TIP60 histone acetyltransferase complex to chromatin. *EMBO Rep.*, 4:575–580, Jun 2003.

[138] A. C. Frazee, B. Langmead, and J. T. Leek. ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*, 12:449, 2011.

[139] J. Fujikura, E. Yamato, S. Yonemura, K. Hosoda, S. Masui, K. Nakao, J. Miyazaki Ji, and H. Niwa. Differentiation of embryonic stem cells is induced by GATA factors. *Genes Dev.*, 16:784–789, Apr 2002.

[140] A. Funahashi, M. Morohashi, H. Kitano, and N. Tanimura. CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO*, 1(5):159 – 162, 2003.

[141] M. Furlan-Magaril, H. Rincon-Arano, and F. Recillas-Targa. Sequential chromatin immunoprecipitation protocol: ChIP-reChIP. *Methods Mol. Biol.*, 543:253–266, 2009.

[142] J. M. Galan-Caridad, S. Harel, T. L. Arenzana, Z. E. Hou, F. K. Doetsch, L. A. Mirny, and B. Reizis. Zfx controls the self-renewal of embryonic and hematopoietic stem cells. *Cell*, 129:345–357, Apr 2007.

[143] F. Gao, B. C. Foat, and H. J. Bussemaker. Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics*, 5:31, Mar 2004.

[144] R. L. Gardner. Origin and differentiation of extraembryonic tissues in the mouse. *Int Rev Exp Pathol*, 24:63–133, 1983.

[145] R. L. Gardner. The early blastocyst is bilaterally symmetrical and its axis of symmetry is aligned with the animal-vegetal axis of the zygote in the mouse. *Development*, 124:289–301, Jan 1997.

[146] R. L. Gardner. The initial phase of embryonic patterning in mammals. *Int. Rev. Cytol.*, 203:233–290, 2001.

[147] R. L. Gardner and F. A. Brook. Reflections on the biology of embryonic stem (ES) cells. *Int. J. Dev. Biol.*, 41:235–243, Apr 1997.

[148] P. Gariglio, J. Buss, and M. H. Green. Sarkosyl activation of RNA polymerase activity in mitotic mouse cells. *FEBS Lett.*, 44:330–333, Aug 1974.

[149] D. Garrick, J. A. Sharpe, R. Arkell, L. Dobbie, A. J. Smith, W. G. Wood, D. R. Higgs, and R. J. Gibbons. Loss of Atrx affects trophoblast development and the pattern of X-inactivation in extraembryonic tissues. *PLoS Genet.*, 2(4):e58, Apr 2006.

[150] S. Gasca, D. P. Hill, J. Klingensmith, and J. Rossant. Characterization of a gene trap insertion into a novel gene, cordon-bleu, expressed in axial structures of the gastrulating mouse embryo. *Dev. Genet.*, 17(2):141–154, 1995.

[151] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, 5:R80, 2004.

[152] E. G. Giannopoulou and O. Elemento. An integrated ChIP-seq analysis platform with customizable workflows. *BMC Bioinformatics*, 12:277, 2011.

[153] Belinda Giardine, Cathy Riemer, Ross C. Hardison, Richard Burhans, Laura Elnitski, Prachi Shah, Yi Zhang, Daniel Blankenberg, Istvan Albert, James Taylor, Webb Miller, W. James Kent, and Anton Nekrutenko. Galaxy: A platform for interactive large-scale genome analysis. *Genome Research*, 15(10):1451–1455, 2005.

[154] S.F. Gilbert. *Developmental Biology*. Sinauer Associates, Sunderland, MA, USA, 9e edition, 2010.

[155] W. Gilbert and A. Maxam. The nucleotide sequence of the lac operator. *Proc. Natl. Acad. Sci. U.S.A.*, 70:3581–3584, Dec 1973.

[156] H. Z. Girgis and I. Ovcharenko. Predicting tissue specific cis-regulatory modules in the human genome using pairs of co-occurring motifs. *BMC Bioinformatics*, 13:25, Feb 2012.

[157] C. H. Glover, M. Marin, C. J. Eaves, C. D. Helgason, J. M. Piret, and J. Bryan. Meta-analysis of differentiating mouse embryonic stem cell gene expression kinetics reveals early change of a small gene set. *PLoS Comput. Biol.*, 2:e158, Nov 2006.

[158] S. Gnerre, I. Maccallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe, G. Hall, T. P. Shea, S. Sykes, A. M. Berlin, D. Aird, M. Costello, R. Daza, L. Williams, R. Nicol, A. Gnirke, C. Nusbaum, E. S. Lander, and D. B. Jaffe. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. U.S.A.*, 108(4):1513–1518, Jan 2011.

[159] C. A. Goble, J. Bhagat, S. Aleksejevs, D. Cruickshank, D. Michaelides, D. Newman, M. Borkum, S. Bechhofer, M. Roos, P. Li, and D. De Roure. myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Res.*, 38:W677–682, Jul 2010.

[160] Jeremy Goecks, Anton Nekrutenko, James Taylor, and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8):R86, 2010.

[161] A. Goncalves, A. Tikhonov, A. Brazma, and M. Kapushesky. A pipeline for RNA-seq data processing and quality assessment. *Bioinformatics*, 27:867–869, Mar 2011.

[162] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, 29:644–652, Jul 2011.

[163] G. Grafi, H. Ben-Meir, Y. Avivi, M. Moshe, Y. Dahan, and A. Zemach. Histone methylation controls telomerase-independent telomere lengthening in cells undergoing dedifferentiation. *Dev. Biol.*, 306:838–846, Jun 2007.

[164] V. Graham, J. Khudyakov, P. Ellis, and L. Pevny. SOX2 functions to maintain neural progenitor identity. *Neuron*, 39(5):749–765, Aug 2003.

[165] B. Greber, H. Lehrach, and J. Adjaye. Fibroblast growth factor 2 modulates transforming growth factor beta signaling in mouse embryonic fibroblasts and human ESCs (hESCs) to support hESC self-renewal. *Stem Cells*, 25:455–464, Feb 2007.

[166] C. Gregg, J. Zhang, B. Weissbourd, S. Luo, G. P. Schroth, D. Haig, and C. Dulac. High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science*, 329:643–648, Aug 2010.

[167] M. Griffith, O. L. Griffith, J. Mwenifumbo, R. Goya, A. S. Morrissy, R. D. Morin, R. Corbett, M. J. Tang, Y. C. Hou, T. J. Pugh, G. Robertson, S. Chittaranjan, A. Ally, J. K. Asano, S. Y. Chan, H. I. Li, H. McDonald, K. Teague, Y. Zhao, T. Zeng, A. Delaney, M. Hirst, G. B. Morin, S. J. Jones, I. T. Tai, and M. A. Marra. Alternative expression analysis by RNA sequencing. *Nat. Methods*, 7:843–847, Oct 2010.

[168] A. Groth, W. Rocha, A. Verreault, and G. Almouzni. Chromatin challenges during DNA replication and repair. *Cell*, 128:721–733, Feb 2007.

[169] M. Grskovic, C. Chaivorapol, A. Gaspar-Maia, H. Li, and M. Ramalho-Santos. Systematic identification of cis-regulatory sequences active in mouse and human embryonic stem cells. *PLoS Genet.*, 3(8):e145, Aug 2007.

[170] P. Gu, B. Goodwin, A. C. Chung, X. Xu, D. A. Wheeler, R. R. Price, C. Galardi, L. Peng, A. M. Latour, B. H. Koller, J. Gossen, S. A. Kliewer, and A. J. Cooney. Orphan nuclear receptor LRH-1 is required to maintain Oct4 expression at the epiblast stage of embryonic development. *Mol. Cell. Biol.*, 25:3492–3505, May 2005.

[171] G. Guo, M. Huss, G. Q. Tong, C. Wang, L. Li Sun, N. D. Clarke, and P. Robson. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev. Cell*, 18(4):675–685, Apr 2010.

[172] G. Guo and A. Smith. A genome-wide screen in EpiSCs identifies Nr5a nuclear receptors as potent inducers of ground state pluripotency. *Development*, 137:3185–3192, Oct 2010.

[173] G. Guo, J. Yang, J. Nichols, J. S. Hall, I. Eyres, W. Mansfield, and A. Smith. Klf4 reverts developmentally programmed restriction of ground state pluripotency. *Development*, 136:1063–1069, Apr 2009.

[174] J. B. Gurdon. Adult frogs derived from the nuclei of single somatic cells. *Dev. Biol.*, 4:256–273, Apr 1962.

[175] J. B. Gurdon and D. A. Melton. Nuclear reprogramming in cells. *Science*, 322:1811–1815, Dec 2008.

[176] J. B. Gurdon and I. Wilmut. Nuclear transfer to eggs and oocytes. *Cold Spring Harb Perspect Biol*, 3, Jun 2011.

[177] M. Guttman, I. Amit, M. Garber, C. French, M. F. Lin, D. Feldser, M. Huarte, O. Zuk, B. W. Carey, J. P. Cassady, M. N. Cabili, R. Jaenisch, T. S. Mikkelsen, T. Jacks, N. Hacohen, B. E. Bernstein, M. Kellis, A. Regev, J. L. Rinn, and E. S. Lander. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 458:223–227, Mar 2009.

[178] M. Guttman, J. Donaghey, B. W. Carey, M. Garber, J. K. Grenier, G. Munson, G. Young, A. B. Lucas, R. Ach, L. Bruhn, X. Yang, I. Amit, A. Meissner, A. Regev, J. L. Rinn, D. E. Root, and E. S. Lander. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature*, 477:295–300, Sep 2011.

[179] M. Guttman, M. Garber, J. Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M. J. Koziol, A. Gnirke, C. Nusbaum, J. L. Rinn, E. S. Lander, and A. Regev. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.*, 28:503–510, May 2010.

[180] L. Habegger, A. Sboner, T. A. Gianoulis, J. Rozowsky, A. Agarwal, M. Snyder, and M. Gerstein. RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics*, 27:281–283, Jan 2011.

[181] K. Hailesellasse Sene, C. J. Porter, G. Palidwor, C. Perez-Iratxeta, E. M. Muro, P. A. Campbell, M. A. Rudnicki, and M. A. Andrade-Navarro. Gene function in early mouse embryonic stem cell differentiation. *BMC Genomics*, 8:85, 2007.

[182] F. Halbritter, H.J. Vaidya, and S.R. Tomlinson. GeneProf: analysis of high-throughput sequencing experiments. *Nat. Methods*, 9:7–8, Jan 2012.

[183] J. Hall, G. Guo, J. Wray, I. Eyres, J. Nichols, L. Grotewold, S. Morfopoulou, P. Humphreys, W. Mansfield, R. Walker, S. Tomlinson, and A. Smith. Oct4 and LIF/Stat3 additively induce Krppel factors to sustain embryonic stem cell self-renewal. *Cell Stem Cell*, 5:597–609, Dec 2009.

[184] H. Han, R. Nutiu, J. Moffat, and B.J. Blencowe. SnapShot: High-throughput sequencing applications. *Cell*, 146:1044, 2011.

[185] Y. Han, Y. H. Jin, Y. J. Kim, B. Y. Kang, H. J. Choi, D. W. Kim, C. Y. Yeo, and K. Y. Lee. Acetylation of Sirt2 by p300 attenuates its deacetylase activity. *Biochem. Biophys. Res. Commun.*, 375(4):576–580, Oct 2008.

[186] L. Handoko, H. Xu, G. Li, C. Y. Ngan, E. Chew, M. Schnapp, C. W. Lee, C. Ye, J. L. Ping, F. Mulawadi, E. Wong, J. Sheng, Y. Zhang, T. Poh, C. S. Chan, G. Kunarso, A. Shahab, G. Bourque, V. Cacheux-Rataboul, W. K. Sung, Y. Ruan, and C. L. Wei. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat. Genet.*, 43(7):630–638, Jul 2011.

[187] J. H. Hanna, K. Saha, and R. Jaenisch. Pluripotency and cellular reprogramming: facts, hypotheses, unresolved issues. *Cell*, 143(4):508–525, Nov 2010.

[188] T. J. Hardcastle and K. A. Kelly. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11:422, 2010.

[189] A. H. Hart, L. Hartley, M. Ibrahim, and L. Robb. Identification, cloning and expression analysis of the pluripotency promoting Nanog genes in mouse and human. *Dev. Dyn.*, 230:187–198, May 2004.

[190] N. Hattori, T. Abe, N. Hattori, M. Suzuki, T. Matsuyama, S. Yoshida, E. Li, and K. Shiota. Preference of DNA methyltransferases for CpG islands in mouse embryonic stem cells. *Genome Res.*, 14:1733–1740, Sep 2004.

[191] K. Hayashi and M. A. Surani. Resetting the epigenome beyond pluripotency in the germline. *Cell Stem Cell*, 4:493–498, Jun 2009.

[192] Y. Hayashizaki and P. Carninci. Genome Network and FANTOM3: assessing the complexity of the transcriptome. *PLoS Genet.*, 2:e63, Apr 2006.

[193] X. He, C. C. Chen, F. Hong, F. Fang, S. Sinha, H. H. Ng, and S. Zhong. A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome-wide binding data. *PLoS ONE*, 4:e8155, 2009.

[194] Y. He, B. Vogelstein, V. E. Velculescu, N. Papadopoulos, and K. W. Kinzler. The antisense transcriptomes of human cells. *Science*, 322:1855–1857, Dec 2008.

[195] E. Heitz. Heterochromatin, chromocentren, chromoren. *Ber. Dtsch. Bot. Ges.*, 47:274–284, 1929.

[196] M. Hemberg and G. Kreiman. Conservation of transcription factor binding events predicts gene expression across species. *Nucleic Acids Res.*, 39:7092–7102, Sep 2011.

[197] B. Hendrich and S. Tweedie. The methyl-CpG binding domain and the evolving role of DNA methylation in animals. *Trends Genet.*, 19:269–277, May 2003.

[198] J. C. Heng, B. Feng, J. Han, J. Jiang, P. Kraus, J. H. Ng, Y. L. Orlov, M. Huss, L. Yang, T. Lufkin, B. Lim, and H. H. Ng. The nuclear receptor Nr5a2 can replace Oct4 in the reprogramming of murine somatic cells to pluripotent cells. *Cell Stem Cell*, 6:167–174, Feb 2010.

[199] T. Hirata, T. Amano, Y. Nakatake, M. Amano, Y. Piao, H. G. Hoang, and M. S. Ko. Zscan4 transiently reactivates early embryonic genes during the generation of induced pluripotent stem cells. *Sci Rep*, 2:208, 2012.

[200] L. Ho, R. Jothi, J. L. Ronan, K. Cui, K. Zhao, and G. R. Crabtree. An embryonic stem cell chromatin remodeling complex, esBAF, is an essential component of the core pluripotency transcriptional network. *Proc. Natl. Acad. Sci. U.S.A.*, 106:5187–5191, Mar 2009.

[201] S. J. Ho Sui, K. Begley, D. Reilly, B. Chapman, R. McGovern, P. Rocca-Sera, E. Maguire, G. M. Altschuler, T. A. Hansen, R. Sompallae, A. Krivtsov, R. A. Shivdasani, S. A. Armstrong, A. C. Culhane, M. Correll, S. A. Sansone, O. Hofmann, and W. Hide. The Stem Cell Discovery Engine: an integrated repository and analysis system for cancer stem cell comparisons. *Nucleic Acids Res.*, 40(Database issue):D984–991, Jan 2012.

[202] K. Hochedlinger and K. Plath. Epigenetic reprogramming and induced pluripotency. *Development*, 136:509–523, Feb 2009.

[203] A. Hochheimer and R. Tjian. Diversified transcription initiation complexes expand promoter selectivity and tissue-specific gene expression. *Genes Dev.*, 17:1309–1320, Jun 2003.

[204] P. C. Hollenhorst, K. J. Chandler, R. L. Poulsen, W. E. Johnson, N. A. Speck, and B. J. Graves. DNA specificity determinants associate with distinct transcription factor functions. *PLoS Genet.*, 5:e1000778, Dec 2009.

[205] T. Horn, T. Sandmann, and M. Boutros. Design and evaluation of genome-wide libraries for RNA interference screens. *Genome Biol.*, 11:R61, 2010.

[206] H. B. Houbaviy, M. F. Murray, and P. A. Sharp. Embryonic stem cell-specific MicroRNAs. *Dev. Cell*, 5(2):351–358, Aug 2003.

[207] M. Hsi-Yang Fritz, R. Leinonen, G. Cochrane, and E. Birney. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.*, 21:734–740, May 2011.

[208] d. a. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, 37(1):1–13, Jan 2009.

[209] d. a. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4(1):44–57, 2009.

[210] P. J. Huang, Y. C. Liu, C. C. Lee, W. C. Lin, R. R. Gan, P. C. Lyu, and P. Tang. DSAP: deep-sequencing small RNA analysis pipeline. *Nucleic Acids Res.*, 38:W385–391, Jul 2010.

[211] R. Huang, M. Jaritz, P. Guenzl, I. Vlatkovic, A. Sommer, I. M. Tamir, H. Marks, T. Klampfl, R. Kralovics, H. G. Stunnenberg, D. P. Barlow, and F. M. Pauler. An RNA-Seq strategy to detect the complete coding and non-coding transcriptome including full-length imprinted macro ncRNAs. *PLoS ONE*, 6:e27288, 2011.

[212] D. Huangfu, R. Maehr, W. Guo, A. Eijkelenboom, M. Snitow, A. E. Chen, and D. A. Melton. Induction of pluripotent stem cells by defined factors is greatly improved by small-molecule compounds. *Nat. Biotechnol.*, 26:795–797, Jul 2008.

[213] D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. R. Pocock, P. Li, and T. Oinn. Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.*, 34:W729–732, Jul 2006.

[214] L. Hyslop, M. Stojkovic, L. Armstrong, T. Walter, P. Stojkovic, S. Przyborski, M. Herbert, A. Murdoch, T. Strachan, and M. Lako. Downregulation of NANOG induces differentiation of human embryonic stem cells to extraembryonic lineages. *Stem Cells*, 23:1035–1043, Sep 2005.

[215] T. Ichimura, S. Watanabe, Y. Sakamoto, T. Aoto, N. Fujita, and M. Nakao. Transcriptional repression and heterochromatin formation by MBD1 and MCAF/AM family proteins. *J. Biol. Chem.*, 280(14):13928–13935, Apr 2005.

[216] L. Ilie, F. Fazayeli, and S. Ilie. HiTEC: accurate error correction in high-throughput sequencing data. *Bioinformatics*, 27:295–302, Feb 2011.

[217] R. S. Illingworth and A. P. Bird. CpG islands–'a rough guide'. *FEBS Lett.*, 583:1713–1720, Jun 2009.

[218] N. T. Ingolia, L. F. Lareau, and J. S. Weissman. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, 147(4):789–802, Nov 2011.

[219] H. Inoue and S. Yamanaka. The use of induced pluripotent stem cells in drug development. *Clin. Pharmacol. Ther.*, 89:655–661, May 2011.

[220] J. P. Ioannidis, D. B. Allison, C. A. Ball, I. Coulibaly, X. Cui, A. C. Culhane, M. Falchi, C. Furlanello, L. Game, G. Jurman, J. Mangion, T. Mehta, M. Nitzberg, G. P. Page, E. Petretto, and V. van Noort. Repeatability of published microarray gene expression analyses. *Nat. Genet.*, 41(2):149–155, Feb 2009.

[221] J. P. Ioannidis and J. Lau. Pooling research results: benefits and limitations of meta-analysis. *Jt Comm J Qual Improv*, 25:462–469, Sep 1999.

[222] T. Isagawa, G. Nagae, N. Shiraki, T. Fujita, N. Sato, S. Ishikawa, S. Kume, and H. Aburatani. DNA methylation profiling of embryonic stem cell differentiation into the three germ layers. *PLoS ONE*, 6:e26052, 2011.

[223] S. Islam, U. Kjallquist, A. Moliner, P. Zajac, J. B. Fan, P. Lonnerberg, and S. Linnarsson. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.*, 21:1160–1167, Jul 2011.

[224] K. Isono, Y. Fujimura, J. Shinga, M. Yamaki, J. O-Wang, Y. Takihara, Y. Murahashi, Y. Takada, Y. Mizutani-Koseki, and H. Koseki. Mammalian polyhomeotic homologues Phc2 and Phc1 act in synergy to mediate polycomb repression of Hox genes. *Mol. Cell. Biol.*, 25(15):6694–6706, Aug 2005.

[225] S. Ito, A. C. D'Alessio, O. V. Taranova, K. Hong, L. C. Sowers, and Y. Zhang. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature*, 466:1129–1133, Aug 2010.

[226] T. Itoh, K. Miyake, T. Yamaguchi, M. Tsuge, H. Kaneoka, and S. Iijima. Constitutive expression of the brg1 gene requires GC-boxes near to the transcriptional start site. *J. Biochem.*, 149(3):301–309, Mar 2011.

[227] N. Ivanova, R. Dobrin, R. Lu, I. Kotenko, J. Levorse, C. DeCoste, X. Schafer, Y. Lun, and I. R. Lemischka. Dissecting self-renewal in stem cells with RNA interference. *Nature*, 442:533–538, Aug 2006.

[228] S. Iyengar, A. V. Ivanov, V. X. Jin, F. J. Rauscher, and P. J. Farnham. Functional analysis of KAP1 genomic recruitment. *Mol. Cell. Biol.*, 31:1833–1847, May 2011.

[229] A. K. Iyer and E. R. McCabe. Molecular mechanisms of DAX1 action. *Mol. Genet. Metab.*, 83:60–73, 2004.

[230] R. Jaenisch and R. Young. Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. *Cell*, 132:567–582, Feb 2008.

[231] B. Jagla, B. Wiswedel, and J. Y. Coppee. Extending KNIME for next-generation sequencing data analysis. *Bioinformatics*, 27:2907–2909, Oct 2011.

[232] R. Jauch, C. K. Ng, K. S. Saikatendu, R. C. Stevens, and P. R. Kolatkar. Crystal structure and DNA binding of the homeodomain of the stem cell transcription factor Nanog. *J. Mol. Biol.*, 376:758–770, Feb 2008.

[233] Y. Jeon and J. T. Lee. YY1 tethers Xist RNA to the inactive X nucleation center. *Cell*, 146(1):119–133, Jul 2011.

[234] H. Ji, H. Jiang, W. Ma, D. S. Johnson, R. M. Myers, and W. H. Wong. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.*, 26:1293–1300, Nov 2008.

[235] H. Jiang and B. M. Peterlin. Differential chromatin looping regulates CD4 expression in immature thymocytes. *Mol. Cell. Biol.*, 28:907–912, Feb 2008.

[236] J. Jiang, Y. S. Chan, Y. H. Loh, J. Cai, G. Q. Tong, C. A. Lim, P. Robson, S. Zhong, and H. H. Ng. A core Klf circuitry regulates self-renewal of embryonic stem cells. *Nat. Cell Biol.*, 10:353–360, Mar 2008.

[237] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316:1497–1502, Jun 2007.

[238] M. H. Johnson, B. Maro, and M. Takeichi. The role of cell adhesion in the synchronization and orientation of polarization in 8-cell mouse blastomeres. *J Embryol Exp Morphol*, 93:239–255, Apr 1986.

[239] C. N. Johnstone, S. J. White, N. C. Tebbutt, F. J. Clay, M. Ernst, W. H. Biggs, C. S. Viars, S. Czekay, K. C. Arden, and J. K. Heath. Analysis of the regulation of the A33 antigen gene reveals intestine-specific mechanisms of gene expression. *J. Biol. Chem.*, 277(37):34531–34539, Sep 2002.

[240] A. Jolma, T. Kivioja, J. Toivonen, L. Cheng, G. Wei, M. Enge, M. Taipale, J. M. Vaquerizas, J. Yan, M. J. Sillanpaa, M. Bonke, K. Palin, S. Talukder, T. R. Hughes, N. M. Luscombe, E. Ukkonen, and J. Taipale. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.*, 20:861–873, Jun 2010.

[241] M. Joo, J. G. Wright, N. N. Hu, R. T. Sadikot, G. Y. Park, T. S. Blackwell, and J. W. Christman. Yin Yang 1 enhances cyclooxygenase-2 gene expression in macrophages. *Am. J. Physiol. Lung Cell Mol. Physiol.*, 292(5):L1219–1226, May 2007.

[242] R. Jothi, S. Cuddapah, A. Barski, K. Cui, and K. Zhao. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.*, 36:5221–5231, Sep 2008.

[243] R. L. Judson, J. E. Babiarz, M. Venere, and R. Blelloch. Embryonic stem cell-specific microRNAs promote induced pluripotency. *Nat. Biotechnol.*, 27:459–461, May 2009.

[244] H. Jung, J. Lacombe, E. O. Mazzoni, K. F. Liem, J. Grinstein, S. Mahony, D. Mukhopadhyay, D. K. Gifford, R. A. Young, K. V. Anderson, H. Wichterle, and J. S. Dasen. Global control of motor neuron topography mediated by the repressive actions of a single hox gene. *Neuron*, 67:781–796, Sep 2010.

[245] M. H. Kagey, J. J. Newman, S. Bilodeau, Y. Zhan, D. A. Orlando, N. L. van Berkum, C. C. Ebmeier, J. Goossens, P. B. Rahl, S. S. Levine, D. J. Taatjes, J. Dekker, and R. A. Young. Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467:430–435, Sep 2010.

[246] A. Kahvejian, J. Quackenbush, and J. F. Thompson. What would you do if you could sequence everything? *Nat. Biotechnol.*, 26:1125–1133, Oct 2008.

[247] K. Kaji, I. M. Caballero, R. MacLeod, J. Nichols, V. A. Wilson, and B. Hendrich. The NuRD component Mbd3 is required for pluripotency of embryonic stem cells. *Nat. Cell Biol.*, 8:285–292, Mar 2006.

[248] K. Kaji, K. Norrby, A. Paca, M. Mileikovsky, P. Mohseni, and K. Woltjen. Virus-free induction of pluripotency and subsequent excision of reprogramming factors. *Nature*, 458:771–775, Apr 2009.

[249] M. Kanamori, H. Konno, N. Osato, J. Kawai, Y. Hayashizaki, and H. Suzuki. A genome-wide and nonredundant mouse transcription factor database. *Biochem. Biophys. Res. Commun.*, 322:787–793, Sep 2004.

[250] C. Kanellopoulou, S. A. Muljo, A. L. Kung, S. Ganesan, R. Drapkin, T. Jenuwein, D. M. Livingston, and K. Rajewsky. Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. *Genes Dev.*, 19(4):489–501, Feb 2005.

[251] J. Kang, M. Gemberling, M. Nakamura, F. G. Whitby, H. Handa, W. G. Fairbrother, and D. Tantin. A general mechanism for transcription regulation by Oct1 and Oct4 in response to genotoxic and oxidative stress. *Genes Dev.*, 23:208–222, Jan 2009.

[252] T. Kanno, E. Bucher, L. Daxinger, B. Huettel, D. P. Kreil, F. Breinig, M. Lind, M. J. Schmitt, S. A. Simon, S. G. Gurazada, B. C. Meyers, Z. J. Lorkovic, A. J. Matzke, and M. Matzke. RNA-directed DNA methylation and plant development require an IWR1-type transcription factor. *EMBO Rep.*, 11:65–71, Jan 2010.

[253] R. Karlic, H. R. Chung, J. Lasserre, K. Vlahovicek, and M. Vingron. Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, 107:2926–2931, Feb 2010.

[254] M. T. Kassouf, J. R. Hughes, S. Taylor, S. J. McGowan, S. Soneji, A. L. Green, P. Vyas, and C. Porcher. Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. *Genome Res.*, 20:1064–1083, Aug 2010.

[255] S. Katayama, Y. Tomaru, T. Kasukawa, K. Waki, M. Nakanishi, M. Nakamura, H. Nishida, C. C. Yap, M. Suzuki, J. Kawai, H. Suzuki, P. Carninci, Y. Hayashizaki, C. Wells, M. Frith, T. Ravasi, K. C. Pang, J. Hallinan, J. Mattick, D. A. Hume, L. Lipovich, S. Batalov, P. G. Engstrm, Y. Mizuno, M. A. Faghihi, A. Sandelin, A. M. Chalk, S. Mottagui-Tabar, Z. Liang, B. Lenhard, and C. Wahlestedt. Antisense transcription in the mammalian transcriptome. *Science*, 309:1564–1566, Sep 2005.

[256] Y. Katz, E. T. Wang, E. M. Airoldi, and C. B. Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods*, 7:1009–1015, Dec 2010.

[257] D. R. Kelley, M. C. Schatz, and S. L. Salzberg. Quake: quality-aware detection and correction of sequencing errors. *Genome Biol.*, 11:R116, 2010.

[258] V. R. Kelly, B. Xu, R. Kuick, R. J. Koenig, and G. D. Hammer. Dax1 up-regulates Oct4 expression in mouse embryonic stem cells via LRH-1 and SRA. *Mol. Endocrinol.*, 24:2281–2291, Dec 2010.

[259] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The human genome browser at UCSC. *Genome Res.*, 12:996–1006, Jun 2002.

[260] W. J. Kent, A. S. Zweig, G. Barber, A. S. Hinrichs, and D. Karolchik. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, 26:2204–2207, Sep 2010.

[261] O. Khalfallah, M. Rouleau, P. Barbry, B. Bardoni, and E. Lalli. Dax-1 knockdown in mouse embryonic stem cells induces loss of pluripotency and multilineage differentiation. *Stem Cells*, 27:1529–1537, Jul 2009.

[262] A. M. Khalil, M. Guttman, M. Huarte, M. Garber, A. Raj, D. Rivea Morales, K. Thomas, A. Presser, B. E. Bernstein, A. van Oudenaarden, A. Regev, E. S. Lander, and J. L. Rinn. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, 106:11667–11672, Jul 2009.

[263] P. V. Kharchenko, M. Y. Tolstorukov, and P. J. Park. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.*, 26:1351–1359, Dec 2008.

[264] S. P. Khare, F. Habib, R. Sharma, N. Gadewal, S. Gupta, and S. Galande. HIstome–a relational knowledgebase of human histone proteins and histone modifying enzymes. *Nucleic Acids Res.*, 40:D337–342, Jan 2012.

[265] B. L. Kidder, S. Palmer, and J. G. Knott. SWI/SNF-Brg1 regulates self-renewal and occupies core pluripotency-related genes in embryonic stem cells. *Stem Cells*, 27(2):317–328, Feb 2009.

[266] G. D. Kim, J. Ni, N. Kelesoglu, R. J. Roberts, and S. Pradhan. Co-operation and communication between the human maintenance and de novo DNA (cytosine-5) methyltransferases. *EMBO J.*, 21:4183–4195, Aug 2002.

[267] H. D. Kim, T. Shay, E. K. O'Shea, and A. Regev. Transcriptional regulatory circuits: predicting numbers from alphabets. *Science*, 325:429–432, Jul 2009.

[268] J. Kim, J. Chu, X. Shen, J. Wang, and S. H. Orkin. An extended transcriptional network for pluripotency of embryonic stem cells. *Cell*, 132:1049–1061, Mar 2008.

[269] J. Kim, A. J. Woo, J. Chu, J. W. Snow, Y. Fujiwara, C. G. Kim, A. B. Cantor, and S. H. Orkin. A Myc network accounts for similarities between embryonic stem and cancer cell transcription programs. *Cell*, 143:313–324, Oct 2010.

[270] J. H. Kim, S. M. Park, M. R. Kang, S. Y. Oh, T. H. Lee, M. T. Muller, and I. K. Chung. Ubiquitin ligase MKRN1 modulates telomere length homeostasis through a proteolysis of hTERT. *Genes Dev.*, 19(7):776–781, Apr 2005.

[271] D. C. King, J. Taylor, L. Elnitski, F. Chiaromonte, W. Miller, and R. C. Hardison. Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome Res.*, 15(8):1051–1060, Aug 2005.

[272] J. D. Klemm and C. O. Pabo. Oct-1 POU domain-DNA interactions: cooperative binding of isolated subdomains and effects of covalent linkage. *Genes Dev.*, 10:27–36, Jan 1996.

[273] W. P. Kloosterman and R. H. Plasterk. The diverse functions of microRNAs in animal development and disease. *Dev. Cell*, 11:441–450, Oct 2006.

[274] H. Kobayashi, T. Sakurai, M. Imai, N. Takahashi, A. Fukuda, O. Yayoi, S. Sato, K. Nakabayashi, K. Hata, Y. Sotomaru, Y. Suzuki, and T. Kono. Contribution of intragenic DNA methylation in mouse gametic DNA methylomes to establish oocyte-specific heritable marks. *PLoS Genet.*, 8(1):e1002440, Jan 2012.

[275] Y. Kodama, M. Shumway, and R. Leinonen. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res*, 40:D54–D56, Jan 2012.

[276] M. Koeva, E. C. Forsberg, and J. M. Stuart. Computational integration of homolog and pathway gene module expression reveals general stemness signatures. *PLoS ONE*, 6(4):e18968, 2011.

[277] R. D. Kornberg. Mediator and the mechanism of transcriptional activation. *Trends Biochem. Sci.*, 30:235–239, May 2005.

[278] C. Kozanitis, C. Saunders, S. Kruglyak, V. Bafna, and G. Varghese. Compressing genomic sequence fragments using SlimGene. *J. Comput. Biol.*, 18:401–413, Mar 2011.

[279] A. R. Krebs, J. Demmers, K. Karmodiya, N. C. Chang, A. C. Chang, and L. Tora. ATAC and Mediator coactivators form a stable complex and regulate a set of non-coding RNA genes. *EMBO Rep.*, 11:541–547, Jul 2010.

[280] I. Krivega and A. Dean. Enhancer and promoter interactions-long distance calls. *Curr Opin Genet Dev*, Dec 2011.

[281] P. Krupinski, V. Chickarmane, and C. Peterson. Simulating the mammalian blastocyst–molecular and mechanical interactions pattern the embryo. *PLoS Comput. Biol.*, 7:e1001128, May 2011.

[282] M. Ku, R. P. Koche, E. Rheinbay, E. M. Mendenhall, M. Endoh, T. S. Mikkelsen, A. Presser, C. Nusbaum, X. Xie, A. S. Chi, M. Adli, S. Kasif, L. M. Ptaszek, C. A. Cowan, E. S. Lander, H. Koseki, and B. E. Bernstein. Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet.*, 4:e1000242, Oct 2008.

[283] R. M. Kuhn, D. Karolchik, A. S. Zweig, T. Wang, K. E. Smith, K. R. Rosenbloom, B. Rhead, B. J. Raney, A. Pohl, M. Pheasant, L. Meyer, F. Hsu, A. S. Hinrichs, R. A. Harte, B. Giardine, P. Fujita, M. Diekhans, T. Dreszer, H. Clawson, G. P. Barber, D. Haussler, and W. J. Kent. The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res.*, 37:D755–761, Jan 2009.

[284] G. Kunarso, N. Y. Chia, J. Jeyakani, C. Hwang, X. Lu, Y. S. Chan, H. H. Ng, and G. Bourque. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.*, 42:631–634, Jul 2010.

[285] T. Kunath, M. K. Saba-El-Leil, M. Almousailleakh, J. Wray, S. Meloche, and A. Smith. FGF stimulation of the Erk1/2 signalling cascade triggers transition of pluripotent embryonic stem cells from self-renewal to lineage commitment. *Development*, 134:2895–2902, Aug 2007.

[286] C. H. Kuo, J. H. Deng, Q. Deng, and S. Y. Ying. A novel role of miR-302/367 in reprogramming. *Biochem. Biophys. Res. Commun.*, 417(1):11–16, Jan 2012.

[287] M. J. Kwon, S. H. Kim, H. M. Jeong, H. S. Jung, S. S. Kim, J. E. Lee, M. C. Gye, O. C. Erkin, S. S. Koh, Y. L. Choi, C. K. Park, and Y. K. Shin. Claudin-4 overexpression is associated with epigenetic derepression in gastric carcinoma. *Lab. Invest.*, 91(11):1652–1667, Nov 2011.

[288] T. D. Laajala, S. Raghav, S. Tuomela, R. Lahesmaa, T. Aittokallio, and L. L. Elo. A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. *BMC Genomics*, 10:618, 2009.

[289] E. Lalonde, K. C. Ha, Z. Wang, A. Bemmo, C. L. Kleinman, T. Kwan, T. Pastinen, and J. Majewski. RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. *Genome Res.*, 21:545–554, Apr 2011.

[290] X. Lan, R. Bonneville, J. Apostolos, W. Wu, and V. X. Jin. W-ChIPeaks: a comprehensive web application tool for processing ChIP-chip and ChIP-seq data. *Bioinformatics*, 27:428–430, Feb 2011.

[291] B. Langmead, K. D. Hansen, and J. T. Leek. Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biol.*, 11:R83, 2010.

[292] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10:R25, Mar 2009.

[293] F. Lanner and J. Rossant. The role of FGF/Erk signaling in pluripotent cells. *Development*, 137(20):3351–3360, Oct 2010.

[294] K. Q. Lao, F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, B. Tuch, J. Bodeau, A. Siddiqui, and M. A. Surani. mRNA-sequencing whole transcriptome analysis of a single cell on the SOLiD system. *J Biomol Tech*, 20:266–271, Dec 2009.

[295] M. Lapidot and Y. Pilpel. Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms. *EMBO Rep.*, 7:1216–1222, Dec 2006.

[296] M. J. Law, K. M. Lower, H. P. Voon, J. R. Hughes, D. Garrick, V. Viprakasit, M. Mitson, M. De Gobbi, M. Marra, A. Morris, A. Abbott, S. P. Wilder, S. Taylor, G. M. Santos, J. Cross, H. Ayyub, S. Jones, J. Ragoussis, D. Rhodes, I. Dunham, D. R. Higgs, and R. J. Gibbons. ATR-X syndrome protein targets tandem repeats and influences allele-specific expression in a size-dependent manner. *Cell*, 143:367–378, Oct 2010.

[297] E. K. Lee, Y. W. Jin, J. H. Park, Y. M. Yoo, S. M. Hong, R. Amir, Z. Yan, E. Kwon, A. Elfick, S. Tomlinson, F. Halbritter, T. Waibel, B. W. Yun, and G. J. Loake. Cultured cambial meristematic cells as a source of plant natural products. *Nat. Biotechnol.*, 28:1213–1217, Nov 2010.

[298] E. W. Lee, M. S. Lee, S. Camus, J. Ghim, M. R. Yang, W. Oh, N. C. Ha, D. P. Lane, and J. Song. Differential regulation of p53 and p21 by MKRN1 E3 ligase controls cell cycle arrest and apoptosis. *EMBO J.*, 28(14):2100–2113, Jul 2009.

[299] J. H. Lee, S. R. Hart, and D. G. Skalnik. Histone deacetylase activity is required for embryonic stem cell differentiation. *Genesis*, 38:32–38, Jan 2004.

[300] K. L. Lee, S. K. Lim, Y. L. Orlov, l. e. Y. Yit, H. Yang, L. T. Ang, L. Poellinger, and B. Lim. Graded Nodal/Activin signaling titrates conversion of quantitative phospho-Smad2 levels into qualitative embryonic stem cell fate decisions. *PLoS Genet.*, 7:e1002130, Jun 2011.

[301] M. W. Lee, A. C. Chang, D. S. Sun, C. Y. Hsu, and N. C. Chang. Restricted expression of LUZP in neural lineage cells: a study in embryonic stem cells. *J. Biomed. Sci.*, 8(6):504–511, 2001.

[302] T. I. Lee, R. G. Jenner, L. A. Boyer, M. G. Guenther, S. S. Levine, R. M. Kumar, B. Chevalier, S. E. Johnstone, M. F. Cole, K. Isono, H. Koseki, T. Fuchikami, K. Abe, H. L. Murray, J. P. Zucker, B. Yuan, G. W. Bell, E. Herbolsheimer, N. M. Hannett, K. Sun, D. T. Odom, A. P. Otte, T. L. Volkert, D. P. Bartel, D. A. Melton, D. K. Gifford, R. Jaenisch, and R. A. Young. Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell*, 125:301–313, Apr 2006.

[303] J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, and J. D. Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28:882–883, Mar 2012.

[304] R. Leinonen, R. Akhtar, E. Birney, L. Bower, A. Cerdeno-Tarraga, Y. Cheng, I. Cleland, N. Faruque, N. Goodgame, R. Gibson, G. Hoad, M. Jang, N. Pakseresht, S. Plaister, R. Radhakrishnan, K. Reddy, S. Sobhany, P. Ten Hoopen, R. Vaughan, V. Zalunin, and G. Cochrane. The European Nucleotide Archive. *Nucleic Acids Res.*, 39:28–31, Jan 2011.

[305] Rasko Leinonen, Ruth Akhtar, Ewan Birney, James Bonfield, Lawrence Bower, Matt Corbett, Ying Cheng, Fehmi Demiralp, Nadeem Faruque, Neil Goodgame, Richard Gibson, Gemma Hoad, Christopher Hunter, Mikyung Jang, Steven Leonard, Quan Lin, Rodrigo Lopez, Michael Maguire, Hamish McWilliam, Sheila Plaister, Rajesh Radhakrishnan, Siamak Sobhany, Guy Slater, Petra Ten Hoopen, Franck Valentin, Robert Vaughan, Vadim Zalunin, Daniel Zerbino, and Guy Cochrane. Improvements to services at the european nucleotide archive. *Nucleic Acids Research*, 38(suppl 1):D39–D45, 2010.

[306] Rasko Leinonen, Hideaki Sugawara, and Martin Shumway. The sequence read archive. *Nucleic Acids Research*, 2010.

[307] H. Leonhardt, A. W. Page, H. U. Weier, and T. H. Bestor. A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. *Cell*, 71:865–873, Nov 1992.

[308] G. Li, R. Margueron, M. Ku, P. Chambon, B. E. Bernstein, and D. Reinberg. Jarid2 and PRC2, partners in regulating gene expression. *Genes Dev.*, 24:368–380, Feb 2010.

[309] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25:2078–2079, Aug 2009.

[310] H. Li, J. Ruan, and R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, 18:1851–1858, Nov 2008.

[311] L. Li, R. Jothi, K. Cui, J. Y. Lee, T. Cohen, M. Gorivodsky, I. Tzchori, Y. Zhao, S. M. Hayes, E. H. Bresnick, K. Zhao, H. Westphal, and P. E. Love. Nuclear adaptor Ldb1 regulates a transcriptional program essential for the maintenance of hematopoietic stem cells. *Nat. Immunol.*, 12:129–136, Feb 2011.

[312] R. Li, Y. Li, K. Kristiansen, and J. Wang. SOAP: short oligonucleotide alignment program. *Bioinformatics*, 24(5):713–714, Mar 2008.

[313] R. Li, C. Yu, Y. Li, T. W. Lam, S. M. Yiu, K. Kristiansen, and J. Wang. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–1967, Aug 2009.

[314] W. Li, H. Zhou, R. Abujarour, S. Zhu, J. Young Joo, T. Lin, E. Hao, H. R. Scholer, A. Hayek, and S. Ding. Generation of human-induced pluripotent stem cells in the absence of exogenous Sox2. *Stem Cells*, 27:2992–3000, Dec 2009.

[315] Y. Li, J. McClintick, L. Zhong, H. J. Edenberg, M. C. Yoder, and R. J. Chan. Murine embryonic stem cell differentiation is promoted by SOCS-3 and inhibited by the zinc finger transcription factor Klf4. *Blood*, 105:635–637, Jan 2005.

[316] J. Liang, M. Wan, Y. Zhang, P. Gu, H. Xin, S. Y. Jung, J. Qin, J. Wong, A. J. Cooney, D. Liu, and Z. Songyang. Nanog and Oct4 associate with unique transcriptional repression complexes in embryonic stem cells. *Nat. Cell Biol.*, 10:731–739, Jun 2008.

[317] D. D. Licatalosi, A. Mele, J. J. Fak, J. Ule, M. Kayikci, S. W. Chi, T. A. Clark, A. C. Schweitzer, J. E. Blume, X. Wang, J. C. Darnell, and R. B. Darnell. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464–469, Nov 2008.

[318] L. S. Lim, F. H. Hong, G. Kunarso, and L. W. Stanton. The pluripotency regulator Zic3 is a direct activator of the Nanog promoter in ESCs. *Stem Cells*, 28(11):1961–1969, Nov 2010.

[319] L. S. Lim, Y. H. Loh, W. Zhang, Y. Li, X. Chen, Y. Wang, M. Bakre, H. H. Ng, and L. W. Stanton. Zic3 is required for maintenance of pluripotency in embryonic stem cells. *Mol. Biol. Cell*, 18(4):1348–1358, Apr 2007.

[320] H. Lin, Z. Zhang, M. Q. Zhang, B. Ma, and M. Li. ZOOM! Zillions of oligos mapped. *Bioinformatics*, 24:2431–2437, Nov 2008.

[321] M. Lin, E. Pedrosa, A. Shah, A. Hrabovsky, S. Maqbool, D. Zheng, and H. M. Lachman. RNA-Seq of human neurons derived from iPS cells reveals candidate long non-coding RNAs involved in neurogenesis and neuropsychiatric disorders. *PLoS ONE*, 6(9):e23356, 2011.

[322] Y. C. Lin, S. Jhunjhunwala, C. Benner, S. Heinz, E. Welinder, R. Mansson, M. Sigvardsson, J. Hagman, C. A. Espinoza, J. Dutkowski, T. Ideker, C. K. Glass, and C. Murre. A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate. *Nat. Immunol.*, 11:635–643, Jul 2010.

[323] P. Lindenbaum, S. Le Scouarnec, V. Portero, and R. Redon. Knime4Bio: a set of custom nodes for the interpretation of next-generation sequencing data with KNIME. *Bioinformatics*, 27:3200–3201, Nov 2011.

[324] R. Lister, R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar, and J. R. Ecker. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, 133:523–536, May 2008.

[325] R. Lister, M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q. M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A. H. Millar, J. A. Thomson, B. Ren, and J. R. Ecker. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462:315–322, Nov 2009.

[326] L. Liu, G. Z. Luo, W. Yang, X. Zhao, Q. Zheng, Z. Lv, W. Li, H. J. Wu, L. Wang, X. J. Wang, and Q. Zhou. Activation of the imprinted Dlk1-Dio3 region correlates with pluripotency levels of mouse stem cells. *J. Biol. Chem.*, 285(25):19483–19490, Jun 2010.

[327] Y. H. Loh, Q. Wu, J. L. Chew, V. B. Vega, W. Zhang, X. Chen, G. Bourque, J. George, B. Leong, J. Liu, K. Y. Wong, K. W. Sung, C. W. Lee, X. D. Zhao, K. P. Chiu, L. Lipovich, V. A. Kuznetsov, P. Robson, L. W. Stanton, C. L. Wei, Y. Ruan, B. Lim, and H. H. Ng. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.*, 38:431–440, Apr 2006.

[328] V. M. Longshaw, M. Baxter, M. Prewitz, and G. L. Blatch. Knockdown of the co-chaperone Hop promotes extranuclear accumulation of Stat3 in mouse embryonic stem cells. *Eur. J. Cell Biol.*, 88(3):153–166, Mar 2009.

[329] R. Lu, A. Yang, and Y. Jin. Dual functions of T-box 3 (Tbx3) in the control of self-renewal and extraembryonic endoderm differentiation in mouse embryonic stem cells. *J. Biol. Chem.*, 286(10):8425–8436, Mar 2011.

[330] J. S. Lunn, S. A. Sakowski, J. Hur, and E. L. Feldman. Stem cell technology for neurodegenerative diseases. *Ann. Neurol.*, 70:353–361, Sep 2011.

[331] C. Luzzani, C. Solari, N. Losino, W. Ariel, L. Romorini, C. Bluguermann, G. Sevlever, L. Baranao, S. Miriuka, and A. Guberman. Modulation of chromatin modifying factors' gene expression in embryonic and induced pluripotent stem cells. *Biochem. Biophys. Res. Commun.*, 410(4):816–822, Jul 2011.

[332] Z. Ma, T. Swigut, A. Valouev, A. Rada-Iglesias, and J. Wysocka. Sequence-specific regulator Prdm14 safeguards mouse ESCs from entering extraembryonic endoderm fates. *Nat. Struct. Mol. Biol.*, 18:120–127, Feb 2011.

[333] A. P. Mahonen, M. Bonke, L. Kauppinen, M. Riikonen, P. N. Benfey, and Y. Helariutta. A novel two-component hybrid molecule regulates vascular morphogenesis of the Arabidopsis root. *Genes Dev.*, 14:2938–2943, Dec 2000.

[334] S. Malik and R. G. Roeder. Dynamic regulation of pol II transcription by the mammalian Mediator complex. *Trends Biochem. Sci.*, 30:256–263, May 2005.

[335] L. Mamanova, R. M. Andrews, K. D. James, E. M. Sheridan, P. D. Ellis, C. F. Langford, T. W. Ost, J. E. Collins, and D. J. Turner. FRT-seq: amplification-free, strand-specific transcriptome sequencing. *Nat. Methods*, 7(2):130–132, Feb 2010.

[336] E. R. Mardis. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet*, 9:387–402, 2008.

[337] E. R. Mardis. The impact of next-generation sequencing technology on genetics. *Trends Genet.*, 24:133–141, Mar 2008.

[338] Elaine Mardis. The $1,000 *genome*, *the* 100,000 analysis? *Genome Medicine*, 2(11):84, 2010.

[339] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile,

R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437:376–380, Sep 2005.

[340] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, 18:1509–1517, Sep 2008.

[341] T. Marquardt and S. L. Pfaff. Cracking the transcriptional code for cell specification in the neural tube. *Cell*, 106:651–654, Sep 2001.

[342] A. Marson, S. S. Levine, M. F. Cole, G. M. Frampton, T. Brambrink, S. Johnstone, M. G. Guenther, W. K. Johnston, M. Wernig, J. Newman, J. M. Calabrese, L. M. Dennis, T. L. Volkert, S. Gupta, J. Love, N. Hannett, P. A. Sharp, D. P. Bartel, R. Jaenisch, and R. A. Young. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell*, 134:521–533, Aug 2008.

[343] C. Martin and Y. Zhang. The diverse functions of histone lysine methylation. *Nat. Rev. Mol. Cell Biol.*, 6:838–849, Nov 2005.

[344] G. R. Martin. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proc. Natl. Acad. Sci. U.S.A.*, 78:7634–7638, Dec 1981.

[345] R. Maruyama, M. Shipitsin, S. Choudhury, Z. Wu, A. Protopopov, J. Yao, P. K. Lo, M. Bessarabova, A. Ishkin, Y. Nikolsky, X. S. Liu, S. Sukumar, and K. Polyak. Altered antisense-to-sense transcript ratios in breast cancer. *Proc. Natl. Acad. Sci. U.S.A.*, Nov 2010.

[346] H. Masaki, T. Nishida, S. Kitajima, K. Asahina, and H. Teraoka. Developmental pluripotency-associated 4 (DPPA4) localized in active chromatin inhibits mouse embryonic stem cell differentiation into a primitive ectoderm lineage. *J. Biol. Chem.*, 282:33034–33042, Nov 2007.

[347] M. J. Mason, K. Plath, and Q. Zhou. Identification of context-dependent motifs by contrasting ChIP binding data. *Bioinformatics*, 26:2826–2832, Nov 2010.

[348] T. Matsuda, T. Nakamura, K. Nakao, T. Arai, M. Katsuki, T. Heike, and T. Yokota. STAT3 activation is sufficient to maintain an undifferentiated state of mouse embryonic stem cells. *EMBO J.*, 18:4261–4269, Aug 1999.

[349] S. Matsumoto, F. Banine, J. Struve, R. Xing, C. Adams, Y. Liu, D. Metzger, P. Chambon, M. S. Rao, and L. S. Sherman. Brg1 is required for murine neural stem cell maintenance and gliogenesis. *Dev. Biol.*, 289(2):372–383, Jan 2006.

[350] E. R. McCabe. DAX1: Increasing complexity in the roles of this novel nuclear receptor. *Mol. Cell. Endocrinol.*, 265-266:179–182, Feb 2007.

[351] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, 20:1297–1303, Sep 2010.

[352] P. Medvedev, E. Scott, B. Kakaradov, and P. Pevzner. Error correction of high-throughput sequencing datasets with non-uniform coverage. *Bioinformatics*, 27:i137–141, Jul 2011.

[353] A. Meissner. Epigenetic modifications in pluripotent and differentiated cells. *Nat. Biotechnol.*, 28:1079–1088, Oct 2010.

[354] E. M. Mendenhall, R. P. Koche, T. Truong, V. W. Zhou, B. Issac, A. S. Chi, M. Ku, and B. E. Bernstein. GC-rich sequence elements recruit PRC2 in mammalian ES cells. *PLoS Genet.*, 6:e1001244, 2010.

[355] M.A. Mendoza-Parra, S. Pattabhiraman, and H. Gronemeyer. Sequential chromatin immunoprecipitation protocol for global analysis through massive parallel sequencing (reChIP-seq). *Protocol Exchange*, 2012.

[356] T. R. Mercer, M. E. Dinger, and J. S. Mattick. Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.*, 10:155–159, Mar 2009.

[357] J. P. Mesirov. Computer science. Accessible reproducible research. *Science*, 327(5964):415–416, Jan 2010.

[358] M. L. Metzker. Sequencing technologies - the next generation. *Nat. Rev. Genet.*, 11:31–46, Jan 2010.

[359] A. Miele and J. Dekker. Long-range chromosomal interactions and gene regulation. *Mol Biosyst*, 4:1046–1057, Nov 2008.

[360] T. S. Mikkelsen, J. Hanna, X. Zhang, M. Ku, M. Wernig, P. Schorderet, B. E. Bernstein, R. Jaenisch, E. S. Lander, and A. Meissner. Dissecting direct reprogramming through integrative genomic analysis. *Nature*, 454:49–55, Jul 2008.

[361] T. S. Mikkelsen, M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T. K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S. Lander, and B. E. Bernstein. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–560, Aug 2007.

[362] I. M. Min, J. J. Waterfall, L. J. Core, R. J. Munroe, J. Schimenti, and J. T. Lis. Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes Dev.*, 25:742–754, Apr 2011.

[363] K. Mitsui, Y. Tokuzawa, H. Itoh, K. Segawa, M. Murakami, K. Takahashi, M. Maruyama, M. Maeda, and S. Yamanaka. The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell*, 113:631–642, May 2003.

[364] L. Morey, G. Pascual, L. Cozzuto, G. Roma, A. Wutz, S. A. Benitah, and L. Di Croce. Nonoverlapping functions of the Polycomb group Cbx family of proteins in embryonic stem cells. *Cell Stem Cell*, 10(1):47–62, Jan 2012.

[365] H. D. Morgan, W. Dean, H. A. Coker, W. Reik, and S. K. Petersen-Mahrt. Activation-induced cytidine deaminase deaminates 5-methylcytosine in DNA and is expressed in pluripotent tissues: implications for epigenetic reprogramming. *J. Biol. Chem.*, 279:52353–52360, Dec 2004.

[366] R. D. Morin, Y. Zhao, A. L. Prabhu, N. Dhalla, H. McDonald, P. Pandoh, A. Tam, T. Zeng, M. Hirst, and M. Marra. Preparation and analysis of microRNA libraries using the Illumina massively parallel sequencing technology. *Methods Mol. Biol.*, 650:173–199, 2010.

[367] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, 5:621–628, Jul 2008.

[368] A. C. Mullen, D. A. Orlando, J. J. Newman, J. Loven, R. M. Kumar, S. Bilodeau, J. Reddy, M. G. Guenther, R. P. DeKoter, and R. A. Young. Master transcription factors determine cell-type-specific responses to TGF-beta signaling. *Cell*, 147(3):565–576, Oct 2011.

[369] A. Murayama, M. S. Kim, J. Yanagisawa, K. Takeyama, and S. Kato. Transrepression by a liganded nuclear receptor via a bHLH activator through co-regulator switching. *EMBO J.*, 23(7):1598–1608, Apr 2004.

[370] E. P. Murchison, J. F. Partridge, O. H. Tam, S. Cheloufi, and G. J. Hannon. Characterization of Dicer-deficient murine embryonic stem cells. *Proc. Natl. Acad. Sci. U.S.A.*, 102(34):12135–12140, Aug 2005.

[371] K. Nakashima, M. Yanagisawa, H. Arakawa, N. Kimura, T. Hisatsune, M. Kawabata, K. Miyazono, and T. Taga. Synergistic signaling in fetal brain by STAT3-Smad1 complex bridged by p300. *Science*, 284(5413):479–482, Apr 1999.

[372] Y. Nakatake, N. Fukui, Y. Iwamatsu, S. Masui, K. Takahashi, R. Yagi, K. Yagi, J. Miyazaki, R. Matoba, M. S. Ko, and H. Niwa. Klf4 cooperates with Oct3/4 and Sox2 to activate the Lefty1 core promoter in embryonic stem cells. *Mol. Cell. Biol.*, 26:7772–7782, Oct 2006.

[373] P. Navarro and P. Avner. When X-inactivation meets pluripotency: an intimate rendezvous. *FEBS Lett.*, 583:1721–1727, Jun 2009.

[374] P. Navarro, I. Chambers, V. Karwacki-Neisius, C. Chureau, C. Morey, C. Rougeulle, and P. Avner. Molecular coupling of Xist regulation and pluripotency. *Science*, 321:1693–1695, Sep 2008.

[375] P. Navarro, M. Moffat, N. P. Mullin, and I. Chambers. The X-inactivation trans-activator Rnf12 is negatively regulated by pluripotency factors in embryonic stem cells. *Hum. Genet.*, 130(2):255–264, Aug 2011.

[376] P. Navarro, A. Oldfield, J. Legoupi, N. Festuccia, A. Dubois, M. Attia, J. Schoorlemmer, C. Rougeulle, I. Chambers, and P. Avner. Molecular coupling of Tsix regulation and pluripotency. *Nature*, 468:457–460, Nov 2010.

[377] T. J. Nelson, A. Martinez-Fernandez, and A. Terzic. Induced pluripotent stem cells: developmental biology to regenerative medicine. *Nat Rev Cardiol*, 7:700–710, Dec 2010.

[378] D. Newkirk, J. Biesinger, A. Chon, K. Yokomori, and X. Xie. AREM: aligning short reads from ChIP-sequencing by expectation maximization. *J. Comput. Biol.*, 18:1495–1505, Nov 2011.

[379] R. K. Ng and J. B. Gurdon. Epigenetic inheritance of cell differentiation status. *Cell Cycle*, 7:1173–1177, May 2008.

[380] J. Nichols, K. Jones, J. M. Phillips, S. A. Newland, M. Roode, W. Mansfield, A. Smith, and A. Cooke. Validated germline-competent embryonic stem cell lines from nonobese diabetic mice. *Nat. Med.*, 15:814–818, Jul 2009.

[381] J. Nichols and A. Smith. Naive and primed pluripotent states. *Cell Stem Cell*, 4:487–492, Jun 2009.

[382] J. Nichols, B. Zevnik, K. Anastassiadis, H. Niwa, D. Klewe-Nebenius, I. Chambers, H. Scholer, and A. Smith. Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell*, 95:379–391, Oct 1998.

[383] R. Nielsen, J. S. Paul, A. Albrechtsen, and Y. S. Song. Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.*, 12(6):443–451, Jun 2011.

[384] R. Nielsen, T. A. Pedersen, D. Hagenbeek, P. Moulos, R. Siersbaek, E. Megens, S. Denissov, M. B?rgesen, K. J. Francoijs, S. Mandrup, and H. G. Stunnenberg. Genome-wide profiling of PPARgamma:RXR and RNA polymerase II occupancy reveals temporal activation of distinct metabolic pathways and changes in RXR dimer composition during adipogenesis. *Genes Dev.*, 22:2953–2967, Nov 2008.

[385] K. Nieminen, J. Immanen, M. Laxell, L. Kauppinen, P. Tarkowski, K. Dolezal, S. Tahtiharju, A. Elo, M. Decourteix, K. Ljung, R. Bhalerao, K. Keinonen, V. A. Albert, and Y. Helariutta. Cytokinin signaling regulates cambial development in poplar. *Proc. Natl. Acad. Sci. U.S.A.*, 105:20032–20037, Dec 2008.

[386] H. Niwa. Molecular mechanism to maintain stem cell renewal of ES cells. *Cell Struct. Funct.*, 26:137–148, Jun 2001.

[387] H. Niwa. Open conformation chromatin and pluripotency. *Genes Dev.*, 21:2671–2676, Nov 2007.

[388] H. Niwa. Mouse ES cell culture system as a model of development. *Dev. Growth Differ.*, 52:275–283, Apr 2010. [DOI:10.1111/j.1440-169X.2009.01166.x] [PubMed:20148924].

[389] H. Niwa, T. Burdon, I. Chambers, and A. Smith. Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3. *Genes Dev.*, 12:2048–2060, Jul 1998.

[390] H. Niwa, J. Miyazaki, and A. G. Smith. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nat. Genet.*, 24:372–376, Apr 2000.

[391] H. Niwa, K. Ogawa, D. Shimosato, and K. Adachi. A parallel circuit of LIF signalling pathways maintains pluripotency of mouse ES cells. *Nature*, 460:118–122, Jul 2009.

[392] H. Niwa, Y. Toyooka, D. Shimosato, D. Strumpf, K. Takahashi, R. Yagi, and J. Rossant. Interaction between Oct3/4 and Cdx2 determines trophectoderm differentiation. *Cell*, 123:917–929, Dec 2005.

[393] D. A. Nix, S. J. Courdy, and K. M. Boucher. Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC Bioinformatics*, 9:523, 2008.

[394] N. Novershtern, A. Subramanian, L. N. Lawton, R. H. Mak, W. N. Haining, M. E. McConkey, N. Habib, N. Yosef, C. Y. Chang, T. Shay, G. M. Frampton, A. C. Drake, I. Leskov, B. Nilsson, F. Preffer, D. Dombkowski, J. W. Evans, T. Liefeld, J. S. Smutko, J. Chen, N. Friedman, R. A. Young, T. R. Golub, A. Regev, and B. L. Ebert. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell*, 144:296–309, Jan 2011.

[395] D. O'Carroll, S. Erhardt, M. Pagani, S. C. Barton, M. A. Surani, and T. Jenuwein. The polycomb-group gene Ezh2 is required for early mouse development. *Mol. Cell. Biol.*, 21:4330–4336, Jul 2001.

[396] V. V. Ogryzko, R. L. Schiltz, V. Russanova, B. H. Howard, and Y. Nakatani. The transcriptional coactivators p300 and CBP are histone acetyltransferases. *Cell*, 87(5):953–959, Nov 1996.

[397] R. Ohlsson. Gene expression: The coherent Mediator. *Nature*, 467:406–407, Sep 2010.

[398] Thomas Oinn, Mark Greenwood, Matthew Addis, Nedim Alpdemir, Justin Ferris, Kevin Glover, Carole Goble, Antoon Goderis, Duncan Hull, Darren Marvin, Peter Li, Phillip Lord, Matthew Pocock, Martin Senger, Robert Stevens, Anil Wipat, and Christopher Wroe. Taverna: lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice and Experience*, 18(10):1067–1100, August 2006.

[399] M. Okano, D. W. Bell, D. A. Haber, and E. Li. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99:247–257, Oct 1999.

[400] K. Okita, T. Ichisaka, and S. Yamanaka. Generation of germline-competent induced pluripotent stem cells. *Nature*, 448:313–317, Jul 2007.

[401] K. Okita, M. Nakagawa, H. Hyenjong, T. Ichisaka, and S. Yamanaka. Generation of mouse induced pluripotent stem cells without viral vectors. *Science*, 322:949–953, Nov 2008.

[402] K. Okita and S. Yamanaka. Intracellular signaling pathways regulating pluripotency of embryonic stem cells. *Curr Stem Cell Res Ther*, 1:103–111, Jan 2006.

[403] A. Okuda, A. Fukushima, M. Nishimoto, A. Orimo, T. Yamagishi, Y. Nabeshima, M. Kuro-o, Y. Nabeshima, K. Boon, M. Keaveney, H. G. Stunnenberg, and M. Muramatsu. UTF1, a novel transcriptional coactivator expressed in pluripotent embryonic stem cells and extra-embryonic cells. *EMBO J.*, 17:2019–2032, Apr 1998.

[404] A. O'Loghlen, A. M. Munoz-Cabello, A. Gaspar-Maia, H. A. Wu, A. Banito, N. Kunowska, T. Racek, H. N. Pemberton, P. Beolchi, F. Lavial, O. Masui, M. Vermeulen, T. Carroll, J. Graumann, E. Heard, N. Dillon, V. Azuara, A. P. Snijders, G. Peters, E. Bernstein, and J. Gil. MicroRNA regulation of Cbx7 mediates a switch of Polycomb orthologs during ESC differentiation. *Cell Stem Cell*, 10(1):33–46, Jan 2012.

[405] T. T. Onder, N. Kara, A. Cherry, A. U. Sinha, N. Zhu, K. M. Bernt, P. Cahan, O. B. Mancarci, J. Unternaehrer, P. B. Gupta, E. S. Lander, S. A. Armstrong, and G. Q. Daley. Chromatin-modifying enzymes as modulators of reprogramming. *Nature*, Mar 2012.

[406] Z. Ouyang, Q. Zhou, and W. H. Wong. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc. Natl. Acad. Sci. U.S.A.*, 106:21521–21526, Dec 2009.

[407] F. Ozsolak, A. Goren, M. Gymrek, M. Guttman, A. Regev, B. E. Bernstein, and P. M. Milos. Digital transcriptome profiling from attomole-level RNA samples. *Genome Res.*, 20(4):519–525, Apr 2010.

[408] F. Ozsolak, A. R. Platt, D. R. Jones, J. G. Reifenberger, L. E. Sass, P. McInerney, J. F. Thompson, J. Bowers, M. Jarosz, and P. M. Milos. Direct RNA sequencing. *Nature*, 461(7265):814–818, Oct 2009.

[409] F. Ozsolak, L. L. Poling, Z. Wang, H. Liu, X. S. Liu, R. G. Roeder, X. Zhang, J. S. Song, and D. E. Fisher. Chromatin structure analyses identify miRNA promoters. *Genes Dev.*, 22:3172–3183, Nov 2008.

[410] S. L. Palmieri, W. Peter, H. Hess, and H. R. Scholer. Oct-4 transcription factor is differentially expressed in the mouse embryo during establishment of the first two extraembryonic cell lineages involved in implantation. *Dev. Biol.*, 166:259–267, Nov 1994.

[411] G. Pan and J. A. Thomson. Nanog and transcriptional networks in embryonic stem cell pluripotency. *Cell Res.*, 17(1):42–49, Jan 2007.

[412] G. Pan, S. Tian, J. Nie, C. Yang, V. Ruotti, H. Wei, G. A. Jonsdottir, R. Stewart, and J. A. Thomson. Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. *Cell Stem Cell*, 1(3):299–312, Sep 2007.

[413] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, 40:1413–1415, Dec 2008.

[414] L. Pantano, X. Estivill, and E. Marti. SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells. *Nucleic Acids Res.*, 38:e34, Mar 2010.

[415] P. J. Park. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, 10:669–680, Oct 2009.

[416] D. Pasini, A. P. Bracken, M. R. Jensen, E. Lazzerini Denchi, and K. Helin. Suz12 is essential for mouse development and for EZH2 histone methyltransferase activity. *EMBO J.*, 23:4061–4071, Oct 2004.

[417] D. Pasini, P. A. Cloos, J. Walfridsson, L. Olsson, J. P. Bukowski, J. V. Johansen, M. Bak, N. Tommerup, J. Rappsilber, and K. Helin. JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells. *Nature*, 464:306–310, Mar 2010.

[418] S. Pauklin, R. A. Pedersen, and L. Vallier. Mouse pluripotent stem cells at a glance. *J. Cell. Sci.*, 124:3727–3732, Nov 2011.

[419] K. L. Pearson, T. Hunter, and R. Janknecht. Activation of Smad1-mediated transcription by p300/CBP. *Biochim. Biophys. Acta*, 1489(2-3):354–364, Dec 1999.

[420] S. Peng, L. L. Chen, X. X. Lei, L. Yang, H. Lin, G. G. Carmichael, and Y. Huang. Genome-wide studies reveal that Lin28 enhances the translation of genes important for growth and survival of human embryonic stem cells. *Stem Cells*, 29(3):496–504, Mar 2011.

[421] S. Peng, N. J. Maihle, and Y. Huang. Pluripotency factors Lin28 and Oct4 identify a sub-population of stem cell-like cells in ovarian cancer. *Oncogene*, 29(14):2153–2159, Apr 2010.

[422] S. Pepke, B. Wold, and A. Mortazavi. Computation for ChIP-seq and RNA-seq studies. *Nat. Methods*, 6(11 Suppl):22–32, Nov 2009.

[423] M. Pesce and H. R. Scholer. Oct-4: gatekeeper in the beginnings of mammalian development. *Stem Cells*, 19:271–278, 2001.

[424] J. E. Phillips and V. G. Corces. CTCF: master weaver of the genome. *Cell*, 137(7):1194–1211, Jun 2009.

[425] K. Phillips and B. Luisi. The virtuoso of versatility: POU proteins that flex to fit. *J. Mol. Biol.*, 302:1023–1039, Oct 2000.

[426] H. A. Piwowar, R. S. Day, and D. B. Fridsma. Sharing detailed research data is associated with increased citation rate. *PLoS ONE*, 2(3):e308, 2007.

[427] J. Ponjavic, C. P. Ponting, and G. Lunter. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.*, 17:556–565, May 2007.

[428] C. P. Ponting, P. L. Oliver, and W. Reik. Evolution and functions of long noncoding RNAs. *Cell*, 136:629–641, Feb 2009.

[429] A. Prokhortchouk, B. Hendrich, H. J?rgensen, A. Ruzov, M. Wilm, G. Georgiev, A. Bird, and E. Prokhortchouk. The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor. *Genes Dev.*, 15:1613–1618, Jul 2001.

[430] K. D. Pruitt, T. Tatusova, W. Klimke, and D. R. Maglott. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.*, 37:D32–36, Jan 2009.

[431] M. Ptashne and A. Gann. Transcriptional activation by recruitment. *Nature*, 386:569–577, Apr 1997.

[432] Y. Qiu, Y. Zhao, M. Becker, S. John, B. S. Parekh, S. Huang, A. Hendarwanto, E. D. Martinez, Y. Chen, H. Lu, N. L. Adkins, D. A. Stavreva, M. Wiench, P. T. Georgel, R. L. Schiltz, and G. L. Hager. HDAC1 acetylation is linked to progressive modulation of steroid receptor-induced gene transcription. *Mol. Cell*, 22(5):669–679, Jun 2006.

[433] M. A. Quail, I. Kozarewa, F. Smith, A. Scally, P. J. Stephens, R. Durbin, H. Swerdlow, and D. J. Turner. A large genome center's improvements to the Illumina sequencing system. *Nat. Methods*, 5(12):1005–1010, Dec 2008.

[434] M. A. Quail, H. Swerdlow, and D. J. Turner. Improved protocols for the illumina genome analyzer sequencing system. *Curr Protoc Hum Genet*, Chapter 18:Unit 18.2, Jul 2009.

[435] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[436] J. Rageul, S. Mottier, A. Jarry, Y. Shah, S. Theoleyre, D. Masson, F. J. Gonzalez, C. L. Laboisse, and M. G. Denis. KLF4-dependent, PPARgamma-induced expression of GPA33 in colon cancer cell lines. *Int. J. Cancer*, 125(12):2802–2809, Dec 2009.

[437] P. B. Rahl, C. Y. Lin, A. C. Seila, R. A. Flynn, S. McCuine, C. B. Burge, P. A. Sharp, and R. A. Young. c-Myc regulates transcriptional pause release. *Cell*, 141:432–445, Apr 2010.

[438] S. Ramchandani, S. K. Bhattacharya, N. Cervoni, and M. Szyf. DNA methylation is a reversible biological signal. *Proc. Natl. Acad. Sci. U.S.A.*, 96:6107–6112, May 1999.

[439] E. B. Rasmussen and J. T. Lis. In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. *Proc. Natl. Acad. Sci. U.S.A.*, 90:7923–7927, Sep 1993.

[440] T. Ravasi, H. Suzuki, C. V. Cannistraci, S. Katayama, V. B. Bajic, K. Tan, A. Akalin, S. Schmeier, M. Kanamori-Katayama, N. Bertin, P. Carninci, C. O. Daub, A. R. Forrest, J. Gough, S. Grimmond, J. H. Han, T. Hashimoto, W. Hide, O. Hofmann, A. Kamburov, M. Kaur, H. Kawaji, A. Kubosaki, T. Lassmann, E. van Nimwegen, C. R. MacPherson, C. Ogawa, A. Radovanovic, A. Schwartz, R. D. Teasdale, J. Tegner, B. Lenhard, S. A. Teichmann, T. Arakawa, N. Ninomiya, K. Murakami, M. Tagami, S. Fukuda, K. Imamura, C. Kai, R. Ishihara, Y. Kitazume, J. Kawai, D. A. Hume, T. Ideker, and Y. Hayashizaki. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, 140:744–752, Mar 2010.

[441] S. Ray, C. T. Sherman, M. Lu, and A. R. Brasier. Angiotensinogen gene expression is dependent on signal transducer and activator of transcription 3-mediated p300/cAMP response element binding protein-binding protein coactivator recruitment and histone acetyltransferase activity. *Mol. Endocrinol.*, 16(4):824–836, Apr 2002.

[442] M. Reich, T. Liefeld, J. Gould, J. Lerner, P. Tamayo, and J. P. Mesirov. GenePattern 2.0. *Nat. Genet.*, 38(5):500–501, May 2006.

[443] W. Reik. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, 447:425–432, May 2007.

[444] W. Reik, W. Dean, and J. Walter. Epigenetic reprogramming in mammalian development. *Science*, 293(5532):1089–1093, Aug 2001.

[445] W. Reik and J. Walter. Genomic imprinting: parental influence on the genome. *Nat. Rev. Genet.*, 2(1):21–32, Jan 2001.

[446] J. Reimand, T. Arak, and J. Vilo. g:Profiler–a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Res.*, 39(Web Server issue):W307–315, Jul 2011.

[447] J. Reinartz, E. Bruyns, J. Z. Lin, T. Burcham, S. Brenner, B. Bowen, M. Kramer, and R. Woychik. Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. *Brief Funct Genomic Proteomic*, 1:95–104, Feb 2002.

[448] A. Remenyi, K. Lins, L. J. Nissen, R. Reinbold, H. R. Scholer, and M. Wilmanns. Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. *Genes Dev.*, 17:2048–2059, Aug 2003.

[449] H. S. Rhee and B. F. Pugh. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, 147:1408–1419, Dec 2011.

[450] I. Rhee, K. E. Bachman, B. H. Park, K. W. Jair, R. W. Yen, K. E. Schuebel, H. Cui, A. P. Feinberg, C. Lengauer, K. W. Kinzler, S. B. Baylin, and B. Vogelstein. DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. *Nature*, 416:552–556, Apr 2002.

[451] K. T. Rigbolt, T. A. Prokhorova, V. Akimov, J. Henningsen, P. T. Johansen, I. Kratchmarova, M. Kassem, M. Mann, J. V. Olsen, and B. Blagoev. System-wide temporal characterization of the proteome and phosphoproteome of human embryonic stem cell differentiation. *Sci Signal*, 4:rs3, 2011.

[452] L. Ringrose and R. Paro. Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins. *Annu. Rev. Genet.*, 38:413–443, 2004.

[453] A. Roberts, H. Pimentel, C. Trapnell, and L. Pachter. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, 27:2325–2329, Sep 2011.

[454] G. Robertson, M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O. L. Griffith, A. He, M. Marra, M. Snyder, and S. Jones. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, 4:651–657, Aug 2007.

[455] G. Robertson, J. Schein, R. Chiu, R. Corbett, M. Field, S. D. Jackman, K. Mungall, S. Lee, H. M. Okada, J. Q. Qian, M. Griffith, A. Raymond, N. Thiessen, T. Cezard, Y. S. Butterfield, R. Newsome, S. K. Chan, R. She, R. Varhol, B. Kamoh, A. L. Prabhu, A. Tam, Y. Zhao, R. A. Moore, M. Hirst, M. A. Marra, S. J. Jones, P. A. Hoodless, and I. Birol. De novo assembly and analysis of RNA-seq data. *Nat. Methods*, 7:909–912, Nov 2010.

[456] K. D. Robertson, S. Ait-Si-Ali, T. Yokochi, P. A. Wade, P. L. Jones, and A. P. Wolffe. DNMT1 forms a complex with Rb, E2F1 and HDAC1 and represses transcription from E2F-responsive promoters. *Nat. Genet.*, 25(3):338–342, Jul 2000.

[457] J. T. Robinson, H. Thorvaldsdottir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. Integrative genomics viewer. *Nat. Biotechnol.*, 29:24–26, Jan 2011.

[458] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139–140, Jan 2010.

[459] M. D. Robinson and G. K. Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23:2881–2887, Nov 2007.

[460] M. D. Robinson and G. K. Smyth. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9:321–332, Apr 2008.

[461] R. G. Roeder. Role of general and gene-specific cofactors in the regulation of eukaryotic transcription. *Cold Spring Harb. Symp. Quant. Biol.*, 63:201–218, 1998.

[462] T. C. Roloff and U. A. Nuber. Chromatin, epigenetics and stem cells. *Eur. J. Cell Biol.*, 84:123–135, Mar 2005.

[463] T. C. Roloff, H. H. Ropers, and U. A. Nuber. Comparative study of methyl-CpG-binding domain proteins. *BMC Genomics*, 4:1, Jan 2003.

[464] R. Ronen, I. Gan, S. Modai, A. Sukacheov, G. Dror, E. Halperin, and N. Shomron. miRNAkey: a software for microRNA deep sequencing analysis. *Bioinformatics*, 26:2615–2616, Oct 2010.

[465] R. Rosenkranz, T. Borodina, H. Lehrach, and H. Himmelbauer. Characterizing the mouse ES cell transcriptome with Illumina sequencing. *Genomics*, 92:187–194, Oct 2008.

[466] J. Rossant and P. P. Tam. Blastocyst lineage formation, early embryonic asymmetries and axis patterning in the mouse. *Development*, 136:701–713, Mar 2009.

[467] J. M. Rothberg and J. H. Leamon. The development and impact of 454 sequencing. *Nat. Biotechnol.*, 26:1117–1124, Oct 2008.

[468] A. E. Rougvie and J. T. Lis. The RNA polymerase II molecule at the 5' end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged. *Cell*, 54:795–804, Sep 1988.

[469] M. Roussigne, S. Kossida, A. C. Lavigne, T. Clouaire, V. Ecochard, A. Glories, F. Amalric, and J. P. Girard. The THAP domain: a novel protein motif with similarity to the DNA-binding domain of P element transposase. *Trends Biochem. Sci.*, 28:66–69, Feb 2003.

[470] J. Rozowsky, G. Euskirchen, R. K. Auerbach, Z. D. Zhang, T. Gibson, R. Bjornson, N. Carriero, M. Snyder, and M. B. Gerstein. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.*, 27:66–75, Jan 2009.

[471] L. L. Rubin and K. M. Haston. Stem cell biology and drug discovery. *BMC Biol.*, 9:42, 2011.

[472] P. J. Rugg-Gunn, B. J. Cox, A. Ralston, and J. Rossant. Distinct histone modifications in stem cell lines and tissue lineages from the early mouse embryo. *Proc. Natl. Acad. Sci. U.S.A.*, 107:10783–10790, Jun 2010.

[473] M.B. Ruzinova and R. Benezra. Id proteins in development, cell cycle and cancer. *Trends in cell biology*, 13(8):410–418, 2003.

[474] A. I. Saeed, V. Sharov, J. White, J. Li, W. Liang, N. Bhagabati, J. Braisted, M. Klapa, T. Currier, M. Thiagarajan, A. Sturn, M. Snuffin, A. Rezantsev, D. Popov, A. Ryltsov, E. Kostukovich, I. Borisovsky, Z. Liu, A. Vinsavich, V. Trush, and J. Quackenbush. TM4: a free, open-source system for microarray data management and analysis. *BioTechniques*, 34(2):374–378, Feb 2003.

[475] L. Salmela. Correction of sequencing errors in a mixed set of reads. *Bioinformatics*, 26:1284–1290, May 2010.

[476] N. Salomonis, K. Hanspers, A. C. Zambon, K. Vranizan, S. C. Lawlor, K. D. Dahlquist, S. W. Doniger, J. Stuart, B. R. Conklin, and A. R. Pico. GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics*, 8:217, 2007.

[477] F. Sanger. The Croonian Lecture, 1975. Nucleotide sequences in DNA. *Proc. R. Soc. Lond., B, Biol. Sci.*, 191:317–333, Dec 1975.

[478] F. Sanger and A. R. Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.*, 94:441–448, May 1975.

[479] N. Sato, L. Meijer, L. Skaltsounis, P. Greengard, and A. H. Brivanlou. Maintenance of pluripotency in human and mouse embryonic stem cells through activation of Wnt signaling by a pharmacological GSK-3-specific inhibitor. *Nat. Med.*, 10:55–63, Jan 2004.

[480] N. Sato, T. Yamamoto, Y. Sekine, T. Yumioka, A. Junicho, H. Fuse, and T. Matsuda. Involvement of heat-shock protein 90 in the interleukin-6-mediated signaling pathway through STAT3. *Biochem. Biophys. Res. Commun.*, 300(4):847–852, Jan 2003.

[481] R. Scatena, P. Bottoni, A. Pontoglio, and B. Giardina. Cancer stem cells: the development of new cancer therapeutics. *Expert Opin Biol Ther*, 11:875–892, Jul 2011.

[482] M. C. Schatz, A. L. Delcher, and S. L. Salzberg. Assembly of large genomes using second-generation sequencing. *Genome Res.*, 20(9):1165–1173, Sep 2010.

[483] R. Schmid and M. L. Blaxter. annot8r: GO, EC and KEGG annotation of EST datasets. *BMC Bioinformatics*, 9:180, 2008.

[484] B. M. Schmidt, D. M. Ribnicky, P. E. Lipsky, and I. Raskin. Revisiting the ancient concept of botanical therapeutics. *Nat. Chem. Biol.*, 3:360–366, Jul 2007.

[485] M. P. Schnetz, L. Handoko, B. Akhtar-Zaidi, C. F. Bartels, C. F. Pereira, A. G. Fisher, D. J. Adams, P. Flicek, G. E. Crawford, T. Laframboise, P. Tesar, C. L. Wei, and P. C. Scacheri. CHD7 targets active gene enhancer elements to modulate ES cell-specific gene expression. *PLoS Genet.*, 6:e1001023, Jul 2010.

[486] S. Schneuwly, R. Klemenz, and W. J. Gehring. Redesigning the body plan of Drosophila by ectopic expression of the homoeotic gene Antennapedia. *Nature*, 325:816–818, 1987.

[487] H. R. Scholer, G. R. Dressler, R. Balling, H. Rohdewohld, and P. Gruss. Oct-4: a germline-specific transcription factor mapping to the mouse t-complex. *EMBO J.*, 9:2185–2195, Jul 1990.

[488] H. R. Scholer, S. Ruppert, N. Suzuki, K. Chowdhury, and P. Gruss. New type of POU domain in germ line-specific protein Oct-4. *Nature*, 344:435–439, Mar 1990.

[489] J. Schroder, J. Bailey, T. Conway, and J. Zobel. Reference-free validation of short read data. *PLoS ONE*, 5:e12681, 2010.

[490] B. Schumacher, I. van der Pluijm, M. J. Moorhouse, T. Kosteas, A. R. Robinson, Y. Suh, T. M. Breit, H. van Steeg, L. J. Niedernhofer, W. van Ijcken, A. Bartke, S. R. Spindler, J. H. Hoeijmakers, G. T. van der Horst, and G. A. Garinis. Delayed and accelerated aging share common longevity assurance mechanisms. *PLoS Genet.*, 4(8):e1000161, 2008.

[491] K. B. Scotland, S. Chen, R. Sylvester, and L. J. Gudas. Analysis of Rex1 (zfp42) function in embryonic stem cell differentiation. *Dev. Dyn.*, 238:1863–1877, Aug 2009.

[492] E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, and U. Gaul. Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature*, 451:535–540, Jan 2008.

[493] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, 34:166–176, Jun 2003.

[494] M. Shah, K. Patel, V. A. Fried, and P. B. Sehgal. Interactions of STAT3 with caveolin-1 and heat shock protein 90 in plasma membrane raft and cytosolic complexes. Preservation of cytokine signaling during fever. *J. Biol. Chem.*, 277(47):45662–45669, Nov 2002.

[495] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13:2498–2504, Nov 2003.

[496] Z. Shao, Y. Zhang, G. C. Yuan, S. H. Orkin, and D. J. Waxman. MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. *Genome Biol*, 13:R16, Mar 2012.

[497] A. A. Sharov, S. Masui, L. V. Sharova, Y. Piao, K. Aiba, R. Matoba, L. Xin, H. Niwa, and M. S. Ko. Identification of Pou5f1, Sox2, and Nanog downstream target genes with statistical confidence by applying a novel algorithm to time course microarray and genome-wide chromatin immunoprecipitation data. *BMC Genomics*, 9:269, 2008.

[498] X. Shen, Y. Liu, Y. J. Hsu, Y. Fujiwara, J. Kim, X. Mao, G. C. Yuan, and S. H. Orkin. EZH1 mediates methylation on histone H3 lysine 27 and complements EZH2 in maintaining stem cell identity and executing pluripotency. *Mol. Cell*, 32:491–502, Nov 2008.

[499] J. Shendure and H. Ji. Next-generation DNA sequencing. *Nat. Biotechnol.*, 26:1135–1145, Oct 2008.

[500] J. Shendure, G. J. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon, A. M. Rosenbaum, M. D. Wang, K. Zhang, R. D. Mitra, and G. M. Church. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309:1728–1732, Sep 2005.

[501] W. Shi, H. Wang, G. Pan, Y. Geng, Y. Guo, and D. Pei. Regulation of the pluripotency marker Rex-1 by Nanog and Sox2. *J. Biol. Chem.*, 281:23319–23325, Aug 2006.

[502] Y. Shi, C. Desponts, J. T. Do, H. S. Hahm, H. R. Scholer, and S. Ding. Induction of pluripotent stem cells from mouse embryonic fibroblasts by Oct4 and Klf4 with small-molecule compounds. *Cell Stem Cell*, 3:568–574, Nov 2008.

[503] M. Shumway, G. Cochrane, and H. Sugawara. Archiving next generation sequencing data. *Nucleic Acids Res.*, 38:D870–871, Jan 2010.

[504] J. Silva, J. Nichols, T. W. Theunissen, G. Guo, A. L. van Oosten, O. Barrandon, J. Wray, S. Yamanaka, I. Chambers, and A. Smith. Nanog is the gateway to the pluripotent ground state. *Cell*, 138:722–737, Aug 2009.

[505] J. Silva and A. Smith. Capturing pluripotency. *Cell*, 132:532–536, Feb 2008.

[506] S. A. Simon, J. Zhai, R. S. Nandety, K. P. McCormick, J. Zeng, D. Mejia, and B. C. Meyers. Short-Read Sequencing Technologies for Transcriptional Analyses. *Annu Rev Plant Biol*, Jan 2009.

[507] A. M. Singh, T. Hamazaki, K. E. Hankowski, and N. Terada. A heterogeneous expression pattern for Nanog in embryonic stem cells. *Stem Cells*, 25:2534–2542, Oct 2007.

[508] S. K. Singh, M. N. Kagalwala, J. Parker-Thornburg, H. Adams, and S. Majumder. REST maintains self-renewal and pluripotency of embryonic stem cells. *Nature*, 453:223–227, May 2008.

[509] A. G. Smith. Embryo-derived stem cells: of mice and men. *Annu. Rev. Cell Dev. Biol.*, 17:435–462, 2001.

[510] A. G. Smith, J. K. Heath, D. D. Donaldson, G. G. Wong, J. Moreau, M. Stahl, and D. Rogers. Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides. *Nature*, 336:688–690, Dec 1988.

[511] S. Y. Sokol. Maintaining embryonic stem cell pluripotency with Wnt signaling. *Development*, 138(20):4341–4350, Oct 2011.

[512] M. B. Stadler, R. Murr, L. Burger, R. Ivanek, F. Lienert, A. Scholer, C. Wirbelauer, E. J. Oakeley, D. Gaidatzis, V. K. Tiwari, and D. Schubeler. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, 480(7378):490–495, Dec 2011.

[513] M. Stadtfeld and K. Hochedlinger. Induced pluripotency: history, mechanisms, and applications. *Genes Dev.*, 24:2239–2263, Oct 2010.

[514] M. Stadtfeld, M. Nagaya, J. Utikal, G. Weir, and K. Hochedlinger. Induced pluripotent stem cells generated without viral integration. *Science*, 322:945–949, Nov 2008.

[515] B. L. Stankovich, E. Aguayo, F. Barragan, A. Sharma, and M. G. Pallavicini. Differential adhesion molecule expression during murine embryonic stem cell commitment to the hematopoietic and endothelial lineages. *PLoS ONE*, 6(9):e23810, 2011.

[516] C. Stark, B. J. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, J. Nixon, K. Van Auken, X. Wang, X. Shi, T. Reguly, J. M. Rust, A. Winter, K. Dolinski, and M. Tyers. The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.*, 39(Database issue):698–704, Jan 2011.

[517] A. Sterner-Kock, I. S. Thorey, K. Koli, F. Wempe, J. Otte, T. Bangsow, K. Kuhlmeier, T. Kirchner, S. Jin, J. Keski-Oja, and H. von Melchner. Disruption of the gene encoding the latent transforming growth factor-beta binding protein 4 (LTBP-4) causes abnormal lung development, cardiomyopathy, and colorectal cancer. *Genes Dev.*, 16(17):2264–2273, Sep 2002.

[518] O. W. Stockhammer, H. Rauwerda, F. R. Wittink, T. M. Breit, A. H. Meijer, and H. P. Spaink. Transcriptome analysis of Traf6 function in the innate immune response of zebrafish embryos. *Mol. Immunol.*, 48:179–190, 2010.

[519] M. Stoeckius, J. Maaskola, T. Colombo, H. P. Rahn, M. R. Friedlander, N. Li, W. Chen, F. Piano, and N. Rajewsky. Large-scale sorting of C. elegans embryos reveals the dynamics of small RNA expression. *Nat. Methods*, 6:745–751, Oct 2009.

[520] R. B. Stoughton. Applications of DNA microarrays in biology. *Annu. Rev. Biochem.*, 74:53–82, 2005.

[521] G.A. Strobel, A. Stierle, and WM Hess. Taxol formation in yew–taxus. *Plant Science*, 92(1):1–12, 1993.

[522] D. Strumpf, C. A. Mao, Y. Yamanaka, A. Ralston, K. Chawengsaksophak, F. Beck, and J. Rossant. Cdx2 is required for correct cell fate specification and differentiation of trophectoderm in the mouse blastocyst. *Development*, 132:2093–2102, May 2005.

[523] K. Sugimoto, Y. Jiao, and E. M. Meyerowitz. Arabidopsis regeneration from multiple tissues occurs via a root development pathway. *Dev. Cell*, 18:463–471, Mar 2010.

[524] M. Sultan, M. H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, D. Schmidt, S. O'Keeffe, S. Haas, M. Vingron, H. Lehrach, and M. L. Yaspo. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, 321(5891):956–960, Aug 2008.

[525] K. E. Szulwach, X. Li, Y. Li, C. X. Song, J. W. Han, S. Kim, S. Namburi, K. Hermetz, J. J. Kim, M. K. Rudd, Y. S. Yoon, B. Ren, C. He, and P. Jin. Integrating 5-hydroxymethylcytosine into the epigenomic landscape of human embryonic stem cells. *PLoS Genet.*, 7:e1002154, Jun 2011.

[526] P. A. 't Hoen, Y. Ariyurek, H. H. Thygesen, E. Vreugdenhil, R. H. Vossen, R. X. de Menezes, J. M. Boer, G. J. van Ommen, and J. T. den Dunnen. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res.*, 36:e141, Dec 2008.

[527] M. Tada, Y. Takahama, K. Abe, N. Nakatsuji, and T. Tada. Nuclear reprogramming of somatic cells by in vitro hybridization with ES cells. *Curr. Biol.*, 11:1553–1558, Oct 2001.

[528] M. Tahiliani, K. P. Koh, Y. Shen, W. A. Pastor, H. Bandukwala, Y. Brudno, S. Agarwal, L. M. Iyer, D. R. Liu, L. Aravind, and A. Rao. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, 324:930–935, May 2009.

[529] K. Takahashi and S. Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126:663–676, Aug 2006.

[530] M. R. Tallack, T. Whitington, W. S. Yuen, E. N. Wainwright, J. R. Keys, B. B. Gardiner, E. Nourbakhsh, N. Cloonan, S. M. Grimmond, T. L. Bailey, and A. C. Perkins. A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. *Genome Res.*, 20:1052–1063, Aug 2010.

[531] T. S. Tanaka, I. Lopez de Silanes, L. V. Sharova, H. Akutsu, T. Yoshikawa, H. Amano, S. Yamanaka, M. Gorospe, and M. S. Ko. Esg1, expressed exclusively in preimplantation embryos, germline, and embryonic stem cells, is a putative RNA-binding protein with broad RNA targets. *Dev. Growth Differ.*, 48:381–390, Aug 2006.

[532] F. Tang, C. Barbacioru, S. Bao, C. Lee, E. Nordman, X. Wang, K. Lao, and M. A. Surani. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. *Cell Stem Cell*, 6:468–478, May 2010.

[533] F. Tang, C. Barbacioru, E. Nordman, S. Bao, C. Lee, X. Wang, B. B. Tuch, E. Heard, K. Lao, and M. A. Surani. Deterministic and stochastic allele specific gene expression in single mouse blastomeres. *PLoS ONE*, 6:e21208, 2011.

[534] F. Tang, C. Barbacioru, E. Nordman, B. Li, N. Xu, V. I. Bashkirov, K. Lao, and M. A. Surani. RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nat Protoc*, 5:516–535, 2010.

[535] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao, and M. A. Surani. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*, 6:377–382, May 2009.

[536] F. Tang, K. Lao, and M. A. Surani. Development and applications of single-cell transcriptome analysis. *Nat. Methods*, 8:6–11, Apr 2011.

[537] Y. Tay, J. Zhang, A. M. Thomson, B. Lim, and I. Rigoutsos. MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. *Nature*, 455(7216):1124–1128, Oct 2008.

[538] Y. M. Tay, W. L. Tam, Y. S. Ang, P. M. Gaughwin, H. Yang, W. Wang, R. Liu, J. George, H. H. Ng, R. J. Perera, T. Lufkin, I. Rigoutsos, A. M. Thomson, and B. Lim. MicroRNA-134 modulates the differentiation of mouse embryonic stem cells, where it causes post-transcriptional attenuation of Nanog and LRH1. *Stem Cells*, 26:17–29, Jan 2008.

[539] A. K. Teo, S. J. Arnold, M. W. Trotter, S. Brown, L. T. Ang, Z. Chng, E. J. Robertson, N. R. Dunn, and L. Vallier. Pluripotency factors regulate definitive endoderm specification through eomesodermin. *Genes Dev.*, 25:238–250, Feb 2011.

[540] R. Terranova, S. Yokobayashi, M. B. Stadler, A. P. Otte, M. van Lohuizen, S. H. Orkin, and A. H. Peters. Polycomb group proteins Ezh2 and Rnf2 direct genomic contraction and imprinted repression in early mouse embryos. *Dev. Cell*, 15(5):668–679, Nov 2008.

[541] P. J. Tesar, J. G. Chenoweth, F. A. Brook, T. J. Davies, E. P. Evans, D. L. Mack, R. L. Gardner, and R. D. McKay. New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature*, 448:196–199, Jul 2007.

[542] The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447:799–816, Jun 2007.

[543] Terry M Therneau and Beth Atkinson. R port by Brian Ripley. *rpart: Recursive Partitioning*, 2011. R package version 3.1-50.

[544] M. Thomas-Chollier, C. Herrmann, M. Defrance, O. Sand, D. Thieffry, and J. van Helden. RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res*, Dec 2011.

[545] J. A. Thomson, J. Itskovitz-Eldor, S. S. Shapiro, M. A. Waknitz, J. J. Swiergiel, V. S. Marshall, and J. M. Jones. Embryonic stem cell lines derived from human blastocysts. *Science*, 282:1145–1147, Nov 1998.

[546] T. A. Thorpe. History of plant tissue culture. *Mol. Biotechnol.*, 37:169–180, Oct 2007.

[547] G. Tiscornia, E. L. Vivas, and J. C. Belmonte. Diseases in a dish: modeling human genetic disorders using induced pluripotent cells. *Nat. Med.*, 17:1570–1576, Dec 2011.

[548] V. K. Tiwari, M. B. Stadler, C. Wirbelauer, R. Paro, D. Schubeler, and C. Beisel. A chromatin-modifying function of JNK during stem cell differentiation. *Nat. Genet.*, 44(1):94–100, Jan 2012.

[549] Y. Tokuzawa, E. Kaiho, M. Maruyama, K. Takahashi, K. Mitsui, M. Maeda, H. Niwa, and S. Yamanaka. Fbx15 is a novel target of Oct3/4 but is dispensable for embryonic stem cell self-renewal and mouse development. *Mol. Cell. Biol.*, 23(8):2699–2708, Apr 2003.

[550] C. Trapnell, L. Pachter, and S. L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25:1105–1111, May 2009.

[551] C. Trapnell and S. L. Salzberg. How to map billions of short reads onto genomes. *Nat. Biotechnol.*, 27:455–457, May 2009.

[552] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, 28:511–515, May 2010.

[553] T. Treiber, E. M. Mandel, S. Pott, I. Gyory, S. Firner, E. T. Liu, and R. Grosschedl. Early B cell factor 1 regulates B cell gene networks by activation, repression, and transcription- independent poising of chromatin. *Immunity*, 32:714–725, May 2010.

[554] A. D. Truax and S. F. Greer. ChIP and Re-ChIP assays: investigating interactions between regulatory proteins, histone modifications, and the DNA sequences to which they bind. *Methods Mol. Biol.*, 809:175–188, 2012.

[555] G. C. Tseng, D. Ghosh, and E. Feingold. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res*, Jan 2012.

[556] N. Tsuneyoshi, T. Sumi, H. Onda, H. Nojima, N. Nakatsuji, and H. Suemori. PRDM14 suppresses expression of differentiation marker genes in human embryonic stem cells. *Biochem. Biophys. Res. Commun.*, 367:899–905, Mar 2008.

[557] G. Turcatti, A. Romieu, M. Fedurco, and A. P. Tairi. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Res.*, 36:e25, Mar 2008.

[558] B. M. Turner. Reading signals on the nucleosome with a new nomenclature for modified histones. *Nat. Struct. Mol. Biol.*, 12:110–112, Feb 2005.

[559] M. Vairapandi. Characterization of DNA demethylation in normal and cancerous cell lines and the regulatory role of cell cycle proteins in human DNA demethylase activity. *J. Cell. Biochem.*, 91:572–583, Feb 2004.

[560] L. Vallier, M. Alexander, and R. A. Pedersen. Activin/Nodal and FGF pathways cooperate to maintain pluripotency of human embryonic stem cells. *J. Cell. Sci.*, 118:4495–4509, Oct 2005.

[561] L. Vallier, D. Reynolds, and R. A. Pedersen. Nodal inhibits differentiation of human embryonic stem cells along the neuroectodermal default pathway. *Dev. Biol.*, 275(2):403–421, Nov 2004.

[562] A. Valouev, D. S. Johnson, A. Sundquist, C. Medina, E. Anton, S. Batzoglou, R. M. Myers, and A. Sidow. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods*, 5:829–834, Sep 2008.

[563] J. van Arensbergen, J. Garcia-Hurtado, I. Moran, M. A. Maestro, X. Xu, M. Van de Casteele, A. L. Skoudy, M. Palassini, H. Heimberg, and J. Ferrer. Derepression of Polycomb targets during pancreatic organogenesis allows insulin-producing beta-cells to adopt a neural gene activity program. *Genome Res.*, 20(6):722–732, Jun 2010.

[564] D. L. van den Berg, T. Snoek, N. P. Mullin, A. Yates, K. Bezstarosti, J. Demmers, I. Chambers, and R. A. Poot. An Oct4-centered protein interaction network in embryonic stem cells. *Cell Stem Cell*, 6:369–381, Apr 2010.

[565] V. van den Boom, S. M. Kooistra, M. Boesjes, B. Geverts, A. B. Houtsmuller, K. Monzen, I. Komuro, J. Essers, L. J. Drenth-Diephuis, and B. J. Eggen. UTF1 is a chromatin-associated protein involved in ES cell differentiation. *J. Cell Biol.*, 178:913–924, Sep 2007.

[566] P. van der Stoop, E. A. Boutsma, D. Hulsman, S. Noback, M. Heimerikx, R. M. Kerkhoven, J. W. Voncken, L. F. Wessels, and M. van Lohuizen. Ubiquitin E3 ligase Ring1b/Rnf2 of polycomb repressive complex 1 contributes to stable maintenance of mouse embryonic stem cells. *PLoS ONE*, 3(5):e2235, 2008.

[567] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler. Serial analysis of gene expression. *Science*, 270:484–487, Oct 1995.

[568] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[569] K. Venstermans, L. Eeckhout, and K. De Bosschere. 64-bit versus 32-bit virtual machines for java. *Software: Practice and Experience*, 36(1):1–26, 2006.

[570] M. P. Verzi, H. Shin, H. H. He, R. Sulahian, C. A. Meyer, R. K. Montgomery, J. C. Fleet, M. Brown, X. S. Liu, and R. A. Shivdasani. Differentiation-specific histone modifications reveal dynamic chromatin interactions and partners for the intestinal transcription factor CDX2. *Dev. Cell*, 19:713–726, Nov 2010.

[571] A. Visel, M. J. Blow, Z. Li, T. Zhang, J. A. Akiyama, A. Holt, I. Plajzer-Frick, M. Shoukry, C. Wright, F. Chen, V. Afzal, B. Ren, E. M. Rubin, and L. A. Pennacchio. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457:854–858, Feb 2009.

[572] P. A. Wade, A. Gegonne, P. L. Jones, E. Ballestar, F. Aubry, and A. P. Wolffe. Mi-2 complex couples DNA methylation to chromatin remodelling and histone deacetylation. *Nat. Genet.*, 23:62–66, Sep 1999.

[573] E. Walker. *Transcriptional Network Analysis During Early Differentiation Reveals a Role for Polycomb-like 2 in Mouse Embryonic Stem Cell Commitment*. PhD thesis, Institute of Biomaterials and Biomedial Engineering, University of Toronto, 2011.

[574] E. Walker, W. Y. Chang, J. Hunkapiller, G. Cagney, K. Garcha, J. Torchia, N. J. Krogan, J. F. Reiter, and W. L. Stanford. Polycomb-like 2 associates with PRC2 and regulates transcriptional networks during mouse embryonic stem cell self-renewal and differentiation. *Cell Stem Cell*, 6:153–166, Feb 2010.

[575] E. Walker, A. V. Hernandez, and M. W. Kattan. Meta-analysis: Its strengths and limitations. *Cleve Clin J Med*, 75:431–439, Jun 2008.

[576] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456:470–476, Nov 2008.

[577] H. Wang, L. Wang, H. Erdjument-Bromage, M. Vidal, P. Tempst, R. S. Jones, and Y. Zhang. Role of histone H2A ubiquitination in Polycomb silencing. *Nature*, 431:873–878, Oct 2004.

[578] J. Wang, P. Alexander, L. Wu, R. Hammer, O. Cleaver, and S. L. McKnight. Dependence of mouse embryonic stem cells on threonine catabolism. *Science*, 325(5939):435–439, Jul 2009.

[579] J. Wang, A. Huda, V. V. Lunyak, and I. K. Jordan. A Gibbs sampling strategy applied to the mapping of ambiguous short-sequence tags. *Bioinformatics*, 26:2501–2508, Oct 2010.

[580] J. Wang, S. Rao, J. Chu, X. Shen, D. N. Levasseur, T. W. Theunissen, and S. H. Orkin. A protein interaction network for pluripotency of embryonic stem cells. *Nature*, 444:364–368, Nov 2006.

[581] K. Wang, S. Sengupta, L. Magnani, C. A. Wilson, R. W. Henry, and J. G. Knott. Brg1 is required for Cdx2-mediated repression of Oct4 expression in mouse blastocysts. *PLoS ONE*, 5(5):e10622, 2010.

[582] K. C. Wang and H. Y. Chang. Molecular mechanisms of long noncoding RNAs. *Mol. Cell*, 43:904–914, Sep 2011.

[583] L. Wang, Z. Feng, X. Wang, X. Wang, and X. Zhang. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, 26:136–138, Jan 2010.

[584] W. Wang, C. Lin, D. Lu, Z. Ning, T. Cox, D. Melvin, X. Wang, A. Bradley, and P. Liu. Chromosomal transposition of PiggyBac in mouse embryonic stem cells. *Proc. Natl. Acad. Sci. U.S.A.*, 105:9290–9295, Jul 2008.

[585] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10:57–63, Jan 2009.

[586] L. Wei, G. Vahedi, H. W. Sun, W. T. Watford, H. Takatori, H. L. Ramos, H. Takahashi, J. Liang, G. Gutierrez-Cruz, C. Zang, W. Peng, J. J. O'Shea, and Y. Kanno. Discrete roles of STAT4 and STAT6 transcription factors in tuning epigenetic modifications and transcription during T helper cell differentiation. *Immunity*, 32:840–851, Jun 2010.

[587] P. Western, J. Maldonado-Saldivia, J. van den Bergen, P. Hajkova, M. Saitou, S. Barton, and M. A. Surani. Analysis of Esg1 expression in pluripotent cells and the germline reveals similarities with Oct4 and Sox2 and differences between human pluripotent cell lines. *Stem Cells*, 23:1436–1442, 2005.

[588] W. T. White and M. D. Hendy. Compressing DNA sequence databases with coil. *BMC Bioinformatics*, 9:242, 2008.

[589] E. Wijaya, M. C. Frith, K. Asai, and P. Horton. RecountDB: a database of mapped and count corrected transcribed sequences. *Nucleic Acids Res.*, 40:D1089–1092, Jan 2012.

[590] E. G. Wilbanks and M. T. Facciotti. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS ONE*, 5:e11471, 2010.

[591] D. C. Williams, M. Cai, and G. M. Clore. Molecular basis for synergistic transcriptional activation by Oct1 and Sox2 revealed from the solution structure of the 42-kDa Oct1.Sox2.Hoxb1-DNA ternary transcription factor complex. *J. Biol. Chem.*, 279:1449–1457, Jan 2004.

[592] K. Williams, J. Christensen, M. T. Pedersen, J. V. Johansen, P. A. Cloos, J. Rappsilber, and K. Helin. TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature*, 473:343–348, May 2011.

[593] R. L. Williams, D. J. Hilton, S. Pease, T. A. Willson, C. L. Stewart, D. P. Gearing, E. F. Wagner, D. Metcalf, N. A. Nicola, and N. M. Gough. Myeloid leukaemia inhibitory factor maintains the developmental potential of embryonic stem cells. *Nature*, 336:684–687, Dec 1988.

[594] I. Wilmut, A. E. Schnieke, J. McWhir, A. J. Kind, and K. H. Campbell. Viable offspring derived from fetal and adult mammalian cells. *Nature*, 385:810–813, Feb 1997.

[595] N. K. Wilson, S. D. Foster, X. Wang, K. Knezevic, J. Schutte, P. Kaimakis, P. M. Chilarska, S. Kinston, W. H. Ouwehand, E. Dzierzak, J. E. Pimanda, M. F. de Bruijn, and B. Gottgens. Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell*, 7:532–544, Oct 2010.

[596] N. K. Wilson, D. Miranda-Saavedra, S. Kinston, N. Bonadies, S. D. Foster, F. Calero-Nieto, M. A. Dawson, I. J. Donaldson, S. Dumon, J. Frampton, R. Janky, X. H. Sun, S. A. Teichmann, A. J. Bannister, and B. Gottgens. The transcriptional program controlled by the stem cell leukemia gene Scl/Tal1 during early embryonic hematopoietic development. *Blood*, 113:5456–5465, May 2009.

[597] B. Wold and R. M. Myers. Sequence census methods for functional genomics. *Nat. Methods*, 5:19–21, Jan 2008.

[598] K. Woltjen, I. P. Michael, P. Mohseni, R. Desai, M. Mileikovsky, R. Hamalainen, R. Cowling, W. Wang, P. Liu, M. Gertsenstein, K. Kaji, H. K. Sung, and A. Nagy. piggyBac transposition reprograms fibroblasts to induced pluripotent stem cells. *Nature*, 458:766–770, Apr 2009.

[599] L. H. Wong, J. D. McGhie, M. Sim, M. A. Anderson, S. Ahn, R. D. Hannan, A. J. George, K. A. Morgan, J. R. Mann, and K. H. Choo. ATRX interacts with H3.3 in maintaining telomere structural integrity in pluripotent embryonic stem cells. *Genome Res.*, 20(3):351–360, Mar 2010.

[600] H. Wu, A. C. D'Alessio, S. Ito, K. Xia, Z. Wang, K. Cui, K. Zhao, Y. E. Sun, and Y. Zhang. Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature*, 473:389–393, May 2011.

[601] J. Q. Wu, J. Du, J. Rozowsky, Z. Zhang, A. E. Urban, G. Euskirchen, S. Weissman, M. Gerstein, and M. Snyder. Systematic analysis of transcribed loci in ENCODE regions using RACE sequencing reveals extensive transcription in the human genome. *Genome Biol.*, 9:R3, Jan 2008.

[602] J. Q. Wu, L. Habegger, P. Noisa, A. Szekely, C. Qiu, S. Hutchison, D. Raha, M. Egholm, H. Lin, S. Weissman, W. Cui, M. Gerstein, and M. Snyder. Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, 107:5254–5259, Mar 2010.

[603] Q. Wu, X. Chen, J. Zhang, Y. H. Loh, T. Y. Low, W. Zhang, W. Zhang, S. K. Sze, B. Lim, and H. H. Ng. Sall4 interacts with Nanog and co-occupies Nanog genomic sites in embryonic stem cells. *J. Biol. Chem.*, 281:24090–24094, Aug 2006.

[604] S. M. Wu and K. Hochedlinger. Harnessing the potential of induced pluripotent stem cells for regenerative medicine. *Nat. Cell Biol.*, 13:497–505, May 2011.

[605] B. Xiong, Y. Rui, M. Zhang, K. Shi, S. Jia, T. Tian, K. Yin, H. Huang, S. Lin, X. Zhao, Y. Chen, Y. G. Chen, S. C. Lin, and A. Meng. Tob1 controls dorsal development of zebrafish embryos by antagonizing maternal beta-catenin transcriptional activity. *Dev. Cell*, 11(2):225–238, Aug 2006.

[606] H. Xu, L. Handoko, X. Wei, C. Ye, J. Sheng, C. L. Wei, F. Lin, and W. K. Sung. A signal-noise model for significance analysis of ChIP-seq with negative control. *Bioinformatics*, 26(9):1199–1204, May 2010.

[607] J. Xu, R. Sylvester, A. P. Tighe, S. Chen, and L. J. Gudas. Transcriptional activation of the suppressor of cytokine signaling-3 (SOCS-3) gene via STAT3 is increased in F9 REX1 (ZFP-42) knockout teratocarcinoma stem cells relative to wild-type cells. *J. Mol. Biol.*, 377:28–46, Mar 2008.

[608] S. Yamaguchi, H. Kimura, M. Tada, N. Nakatsuji, and T. Tada. Nanog expression in mouse germ cell development. *Gene Expr. Patterns*, 5:639–646, Jun 2005.

[609] M. Yamaji, Y. Seki, K. Kurimoto, Y. Yabuta, M. Yuasa, M. Shigeta, K. Yamanaka, Y. Ohinata, and M. Saitou. Critical function of Prdm14 for the establishment of the germ cell lineage in mice. *Nat. Genet.*, 40:1016–1022, Aug 2008.

[610] S. Yamanaka. Pluripotency and nuclear reprogramming. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 363:2079–2087, Jun 2008.

[611] S. Yamanaka and H. M. Blau. Nuclear reprogramming to a pluripotent state by three approaches. *Nature*, 465:704–712, Jun 2010.

[612] X. P. Yang, K. Ghoreschi, S. M. Steward-Tharp, J. Rodriguez-Canales, J. Zhu, J. R. Grainger, K. Hirahara, H. W. Sun, L. Wei, G. Vahedi, Y. Kanno, J. J. O'Shea, and A. Laurence. Opposing regulation of the locus encoding IL-17 through direct, reciprocal actions of STAT3 and STAT5. *Nat. Immunol.*, 12:247–254, Mar 2011.

[613] V. Yanovsky. ReCoil - an algorithm for compression of extremely large datasets of dna data. *Algorithms Mol Biol*, 6:23, 2011.

[614] H. Yao, K. Brick, Y. Evrard, T. Xiao, R. D. Camerini-Otero, and G. Felsenfeld. Mediation of CTCF transcriptional insulation by DEAD-box RNA-binding protein p68 and steroid receptor RNA activator SRA. *Genes Dev.*, 24:2543–2555, Nov 2010.

[615] Z. H. Ye. Vascular tissue differentiation and pattern formation in plants. *Annu Rev Plant Biol*, 53:183–202, 2002.

[616] G. W. Yeo, N. G. Coufal, T. Y. Liang, G. E. Peng, X. D. Fu, and F. H. Gage. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nat. Struct. Mol. Biol.*, 16(2):130–137, Feb 2009.

[617] R. Yi and E. Fuchs. MicroRNAs and their roles in mammalian stem cells. *J. Cell. Sci.*, 124:1775–1783, Jun 2011.

[618] Q. L. Ying, J. Nichols, I. Chambers, and A. Smith. BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3. *Cell*, 115:281–292, Oct 2003.

[619] Q. L. Ying, J. Wray, J. Nichols, L. Batlle-Morera, B. Doble, J. Woodgett, P. Cohen, and A. Smith. The ground state of embryonic stem cell self-renewal. *Nature*, 453:519–523, May 2008.

[620] M. D. Young, M. J. Wakefield, G. K. Smyth, and A. Oshlack. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.*, 11:R14, 2010.

[621] R. A. Young. Control of the embryonic stem cell state. *Cell*, 144:940–954, Mar 2011.

[622] J. Yu, M. A. Vodyanik, K. Smuga-Otto, J. Antosiewicz-Bourget, J. L. Frane, S. Tian, J. Nie, G. A. Jonsdottir, V. Ruotti, R. Stewart, I. I. Slukvin, and J. A. Thomson. Induced pluripotent stem cell lines derived from human somatic cells. *Science*, 318:1917–1920, Dec 2007.

[623] K. Yu, B. Zheng, M. Han, and J. K. Wen. ATRA activates and PDGF-BB represses the SM22 promoter through KLF4 binding to, or dissociating from, its cis-DNA elements. *Cardiovasc. Res.*, 90(3):464–474, Jun 2011.

[624] Y. Yukimune, H. Tabata, Y. Higashi, and Y. Hara. Methyl jasmonate-induced overproduction of paclitaxel and baccatin III in Taxus cell suspension cultures. *Nat. Biotechnol.*, 14:1129–1132, Sep 1996.

[625] M. Zalzman, G. Falco, L. V. Sharova, A. Nishiyama, M. Thomas, S. L. Lee, C. A. Stagg, H. G. Hoang, H. T. Yang, F. E. Indig, R. P. Wersto, and M. S. Ko. Zscan4 regulates telomere elongation and genomic stability in ES cells. *Nature*, 464(7290):858–863, Apr 2010.

[626] X. Zang, X.G. Mei, C.H. Zhang, C.T. Lu, and T. Ke. Improved paclitaxel accumulation in cell suspension cultures of taxus chinensis by brassinolide. *Biotechnology letters*, 23(13):1047–1049, 2001.

[627] C. Zhang, X. Ye, H. Zhang, M. Ding, and H. Deng. GATA factors induce mouse embryonic stem cell differentiation toward extraembryonic endoderm. *Stem Cells Dev.*, 16:605–613, Aug 2007.

[628] J. Zhang, W. L. Tam, G. Q. Tong, Q. Wu, H. Y. Chan, B. S. Soh, Y. Lou, J. Yang, Y. Ma, L. Chai, H. H. Ng, T. Lufkin, P. Robson, and B. Lim. Sall4 modulates embryonic stem cell pluripotency and early embryonic development by the transcriptional regulation of Pou5f1. *Nat. Cell Biol.*, 8:1114–1123, Oct 2006.

[629] K. Zhang, F. Faiola, and E. Martinez. Six lysine residues on c-Myc are direct substrates for acetylation by p300. *Biochem. Biophys. Res. Commun.*, 336(1):274–280, Oct 2005.

[630] X. Zhang, J. Zhang, T. Wang, M. A. Esteban, and D. Pei. Esrrb activates Oct4 transcription and sustains self-renewal and pluripotency in embryonic stem cells. *J. Biol. Chem.*, 283:35825–35833, Dec 2008.

[631] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nussbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, 9:R137, 2008.

[632] Z. Zhang, V.B. Bajic, J. Yu, K.H. Cheung, and J.P. Townsend. *Data Integration in Bioinformatics: Current Efforts and Challenges*. InTech Open Access Publishing, 2011.

[633] Z. Zhang and B. F. Pugh. High-resolution genome-wide mapping of the primary structure of chromatin. *Cell*, 144:175–186, Jan 2011.

[634] J. Zhao, T. K. Ohsumi, J. T. Kung, Y. Ogawa, D. J. Grau, K. Sarma, J. J. Song, R. E. Kingston, M. Borowsky, and J. T. Lee. Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol. Cell*, 40(6):939–953, Dec 2010.

[635] S. Zheng and L. Chen. A hierarchical Bayesian model for comparing transcriptomes at the individual transcript isoform level. *Nucleic Acids Res.*, 37:e75, Jun 2009.

[636] S. Zhong, J. G. Joung, Y. Zheng, Y. R. Chen, B. Liu, Y. Shao, J. Z. Xiang, Z. Fei, and J. J. Giovannoni. High-throughput illumina strand-specific RNA sequencing library preparation. *Cold Spring Harb Protoc*, 2011(8):940–949, Aug 2011.

[637] X. Zhong and Y. Jin. Critical roles of coactivator p300 in mouse embryonic stem cell differentiation and Nanog expression. *J. Biol. Chem.*, 284(14):9168–9175, Apr 2009.

[638] F. Zhou, L. Zhang, A. Wang, B. Song, K. Gong, L. Zhang, M. Hu, X. Zhang, N. Zhao, and Y. Gong. The association of GSK3 beta with E2F1 facilitates nerve growth factor-induced neural cell differentiation. *J. Biol. Chem.*, 283(21):14506–14515, May 2008.

[639] H. Zhou, S. Wu, J. Y. Joo, S. Zhu, D. W. Han, T. Lin, S. Trauger, G. Bien, S. Yao, Y. Zhu, G. Siuzdak, H. R. Scholer, L. Duan, and S. Ding. Generation of induced pluripotent stem cells using recombinant proteins. *Cell Stem Cell*, 4:381–384, May 2009.

[640] W. Zhou and C. R. Freed. Adenoviral gene delivery can reprogram human fibroblasts to induced pluripotent stem cells. *Stem Cells*, 27:2667–2674, Nov 2009.

[641] X. Zhou, P. Sumazin, P. Rajbhandari, and A. Califano. A systems biology approach to transcription factor binding site prediction. *PLoS ONE*, 5:e9878, 2010.

[642] B. Zhu, Y. Zheng, H. Angliker, S. Schwarz, S. Thiry, M. Siegmann, and J. P. Jost. 5-Methylcytosine DNA glycosylase activity is also present in the human MBD4 (G/T mismatch glycosylase) and in a related avian sequence. *Nucleic Acids Res.*, 28:4157–4165, Nov 2000.

[643] M. Zywicki, K. Bakowska-Zywicka, and N. Polacek. Revealing stable processing products from ribosome-associated small RNAs by deep-sequencing data analysis. *Nucleic Acids Res*, Jan 2012.

Appendix A

Abbreviations

API Application programming interface (sometimes "Advanced programming interface"). A specification by which computer programmers can use an externally-developed software.

bp Base-pair. One bond between the nitrogenous bases of complementary DNA. This is commonly used as a unit of length for DNA sequences (and also RNA sequences, even if not strictly applicable to most types of RNA).

ChIP Chromatin immunoprecipitation. An experimental technique used to investigate the interaction of certain proteins with DNA. Selective precipitation and purification of sheared DNA fragments bonded by the protein of interest, e.g. a transcription factor.

ChIP-seq ChIP-sequencing. Chromatin immunoprecipitation followed by sequencing of extracted DNA fragments. A technique which is now routinely employed to study the activity patterns of DNA-binding proteins.

CMC Cambial meristemic cell. A undifferentiated cell derived from the cambium of *T. cupsidata* (and other plants).

DBP DNA-binding protein. I use this term to collectively refer to transcription factors, transcriptional insulators and other elements of the transcriptional machinery that directly or indirectly bind or in any way associate with DNA.

DDC Dedifferentiated cell. A proliferating cell derived from either needles or embyros of *T. cupsidata* (and other plants).

DEG Differentially expressed gene. A transcriptional feature that exhibits statistically significant differences in expression levels between two or more conditions. Statistical significance may be assessed with numerous different methods.

DOC Direction of change. In fold change analysis, the sign of the logarithmic fold change, i.e. whether a feature was up- or down-regulated.

DNA Deoxyribonucleic acid. A macromolecule made up of a double-stranded chain of nucleotides. DNA encodes the genetic information constituting the basis for the development and operation of life.

EB Embryoid body. A cluster of cells originating from ESCs in which colony-formation has been prevented and a part of the cells has differentiated (or started to).

ECM Extracellular matrix. Structural components of animal tissue outside cells. The matrix gives support to cells and is involved in signalling, nutrition and other important functions.

EF Embryonic fibroblast. Fibroblasts are cells making up the ECM, collagen and connective tissue.

ES(C) Embryonic stem (cell). A pluripotent cell which can be maintained indefinitely *in vitro* and can differentiate into any cell of the body (but not into extraembryonic tissues).

ESiC Embryonic stem cell identity candidates. A list of candidate genes identified by my analyses. I consider these genes to be central to the establishment and maintenance of **ESC** identity.

FACS Fluorescence-activated cell sorting. A type of flow cytometry used for sorting cells into different populations on the basis of their fluorescent properties.

gb Gigabase. $1,000,000,000$ base-pairs, see **bp**.

GB Gigabyte. $1,024 * 1,024 * 1,024 = 1,073,741,824$ bytes, see **KB**.

GRO-seq Global run-on sequencing. An experimental approach using HTS to profile RNA polymerases in the state of active transcription.

HAT Histoneacetyltransferase. An enzyme that adds acetyl to histone tails.

HDAC Histone deacetylase. An enzyme that removes acetyl from modified histone tails.

HM Histone modification. Any sort of biochemical modification (methylation, phosphorylation, ..) to a histone tail.

HTTP Hypertext transfer protocol. A networking protocol most famous for its use in the world wide web.

HTS High-throughput sequencing. I use this term collectively referring to all modern, massively parallel sequencing platforms and their applications.

ICM Inner cell mass. A mass of cells occurring during early development (before implantation) in the blastocyst. ESCs are derived from these cells.

IDPA Discriminative power analysis. A method for finding common regulatory inputs of groups of genes.

iPS/iPSC Induced pluripotent stem cell. A somatic cell that has been reprogrammed to a stem cell-like pluripotent and self-renewing state.

kb Kilobase. $1,000$ base-pairs, see **bp**.

KB Kilobyte. $1,024$ bytes. A byte is a unit of digital information consisting of 8 bits (each bit is a binary value, 0 or 1).

LDA Linear discriminant analysis. A mathematical method that aims to identify descriptive variables that distinguish sets of data.

LF Lung fibroblast. Fibroblasts are cells making up the ECM, collagen and connective tissue.

mb Megabase. $1,000,000$ base-pairs, see **bp**.

MB Megabyte. $1,024 * 1,024 = 1,048,576$ bytes, see **KB**.

miRNA Micro-RNA. A very short species of non-coding RNA that can interact with mRNA, DNA and histones.

ncRNA Non-coding RNA. Any sort of transcript that is not translated into a protein, including miRNAs.

NGS Next-generation sequencing. Synonymous to high-throughput sequencing (**HTS**).

NPC Neural progenitor cell. An oligopotent progenitor of neural cell types.

PC(A) Principal component (analysis). A mathematical method aiming to identify descriptive variables in a set of values by projecting the data into a lower-dimensional space.

PE Primitive endoderm. An early developmental lineage.

PRC Polycomb repressive complex (divided into PRC1 and PRC2). Proteins involved in the mediation of epigenetic silencing.

ROI Region of interest. A genomic region deemed worthy of particular interest, for example, an enriched binding event in a ChIP-seq experiment.

RNA Ribonuleic acid. A macromolecule made up of a (single-stranded) chain of nucleotides. Various species of RNA exist, importantly messenger RNA (mRNA), which is transcribed from DNA, carries the information encoding synthesis of a wide range of proteins.

RPKM Reads per kilobase million. A unit denoting gene expression levels from RNA-seq experiments.

RPM Reads per million. A unit denoting gene expression levels from RNA-seq experiments.

RNA-seq High-throughput sequencing of messenger RNA (or more frequently of reverse-transcribed cDNA). A technique used for the study of gene expression and transcriptome assembly.

SNP Single nucleotide polymorphism. A variation in the genome sequence between individuals or paired chromosomes within the same individual. In this type of variation, only one single nucleotide differs between DNA sequences.

SRA Sequence Read Archive. A public database of raw high-throughput sequencing data. `http://www.ncbi.nlm.nih.gov/sra`.

TB Terabyte. $1,024 * 1,024 * 1,024 * 1,024 = 1,099,511,627,776$ bytes, see **KB**.

TE Trophectoderm. An early developmental lineage.

TF Transcription factor. A DNA-binding protein controlling the transcriptional activity of target genes.

TFBS Transcription factor binding site. A genomic locus enriched for the binding of a certain **TF**.

TSS Transcription start site. The genomic locus of a gene at which transcription is initiated. Many genes possess multiple, alternative start sites.

TTS Transcription termination site.The genomic locus of a gene at which transcription ends. Many genes possess multiple, alternative termination sites.

URL Universal resource locator. A character string referring uniquely to one particular internet resource.

Appendix B

List of Publications, Presentations and Posters

Peer-reviewed publications based on work carried out during the course of the work outlined in this thesis:

- Festuccia, N., Osorno, R., **Halbritter, F.**, Karwacki-Neisius, V., Navarro, P., Colby, D., Wong, F., Yates, A., Tomlinson, S.R. & Chambers, I. Esrrb Is a Direct Nanog Target Gene that Can Substitute for Nanog Function in Pluripotent Cells. *Cell Stem Cell* 11(4), 477-490 (2012).
- **Halbritter, F.**, Vaidya, H.J. & Tomlinson, S.R. GeneProf: analysis of high-throughput sequencing experiments. *Nature Methods* 9, 7-8 (2011).
- Lee, E.-K., Jin, Y.-W., Park, J.H., Yoo, Y.M., Hong, S.M., Amir, R., Yan, Z., Kwon, E., Alfick, A., Tomlinson, S.R., **Halbritter, F.**, Waibel, T., Yun, B.-W. & Loake, G.J. Cultured cambial meristematic cells as a source of plant natural products. *Nature Biotechnology* 28, 1213-1217 (2010).

Peer-reviewed publications pre-dating this thesis:

- **Halbritter, F.** & Geibel, P. Learning models of relational MDPs using graph kernels. *MICAI 2007: Advanced in Artificial Intelligence, Lecture Notes in Computer Science* 4827, 409-419 (2007).

Manuscripts in preparation:

- **Halbritter, F.**, Brandsma, J., van den Berg, D., Tomlinson, S.R. & Poot, R. Interactions of core pluripotency transcription factors. *Manuscript in preparation.*
- Tetelin, S., O'Neill, K., Bredenkamp, N., Vaidya, H.J., **Halbritter, F.**, Tomlinson, S.R. & Blackburn, C. Role of Foxn1 in thymus. *Manuscript in preparation.*
- **Halbritter, F.** & Tomlinson, S.R. The regulatory code of stem cells. *Manuscript in preparation.*

Conference presentations and posters (excluding internal talks and conference attendances without presentation):

- Talk: "ChIP-seq Data Analysis using GeneProf" (2011). EuroSyStem Workgroup on the Biology of Neural Systems, Milan, Italy.
- Poster: "GeneProf: Integrated Analysis of High-Throughput Sequencing Data" (2011). 19th International Conference on Intelligent Systems for Molecular Biology (ISMB) / 10th European Conference on Computation Biology (ECCB), Vienna, Austria.
- Talk / practical: "Analysis of Next-Gen Sequencing Data" (2009). Quantitative 'Omics Technologies Workshop, Edinburgh, UK.
- Poster: "Digital Transcriptomics for Stem Cell Bioinformatics" (2009). Hydra V Summer School: Stem Cells and Regenerative Medicine, Hydra, Greece.
- Talk / practical: "Finding Data on Stem Cells in StemDB" (2009). Computational Stem Cell Biology Workshop, Leipzig, Germany.

Appendix C

Additional Notes about Data Analysis Issues

C.1 Definition of a Universal Background Signal for Peak Detection Analysis

The analysis of ChIP-seq datasets, whether targeted at DBPs or HMs, usually boils down to the identification of enriched binding (or accumulation) events for the protein of interest in specific regions of the genome ("peak finding"; **Section 3.3.3.5**). It has been noted many times that local elevations in ChIP-seq binding profiles do not always necessarily correspond to "true" biological enrichment events, but that they might instead be caused by other factors such as the accessibility of chromatin, the general susceptibility of specific DNA regions to be pulled out by ChIP or fragmentation, and non-specific binding of the ChIP antibody[415, 470, 631].

For this reason, most researchers nowadays supplement their primary experiment with a negative control sample that can be used to distinguish "false positives" from real binding events. There are various different types of control samples that are being used, but no clear consensus exists as to which might be most appropriate in general. One possibility is the use of an antibody against a protein that is known not to bind DNA. For instance, anti-GFP (green fluorescent protein) or IgG (Immunoglobulin G) are commonly used. Any DNA fragments pulled out would hence be explainable by non-specific effects. Alternatively, other groups prefer to use randomly fragmented input DNA from whole cell extracts as a control. Regions of accessible chromatin and DNA stretches that preferentially come out of the screen can thus be identified and controlled for.

I do not attempt to give a justification or even a conclusion with respect to which type of control is best to use, however, I have noticed that differences between ChIP-seq experiments for the same proteins are often in part caused by the use of different controls. I have therefore hypothesised that the use of a common control might help to improve consistency between observations and went ahead to build a universal background signal dataset (called "UniRef") by combining control samples from six different experiments in ESCs (GeneProf accession codes: gpXP000012, gpXP000027, gpXP000028, gpXP000031, gpXP000048, gpXP000071).

Taking ChIP-seq datasets for *Pou5f1* from two independent studies[75, 342] as an example, it was possible to demonstrate that the use of the UniRef control dataset helped to increase the ratio of overlapping peak calls consistently between three different peak detection algorithms: MACS[631], SISSRs[242] and ChIPseqPeakFinder (CSPF)[75]. For this comparison, I have calibrated the stringency of the individual peak callers in such a way to approximate the estimated true number of binding peaks for *Pou5f1* ($n_{expected} = 4,407$; cp. supplementary material of reference[75]). The results of this experiment are summarised in **Figure C.1**.

Alas, it must be noted that the overlaps, while improved, are still rather poor. Some differences might be explained by actual biological diversity between the cells used in both experiments (E14 and v6.5), but I would expect others to be caused by differences in antibody

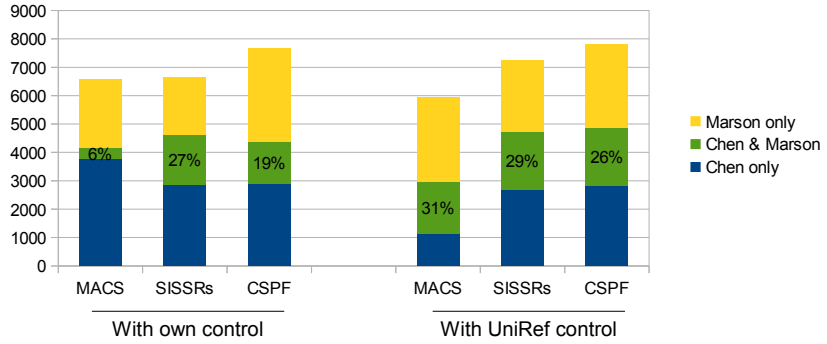

Figure C.1: UniRef control improves agreement of peak calls. The use of a cross-experiment control sample improves the agreement (percentage overlap) between peaks from different ChIP-seq experiments for the same factor. Chen data[75], Marson data[342].

specificity and other technical reasons which are not easily accounted for[70].

I do not want to suggest to replace experimental ChIP-seq controls with UniRef in general and only use it for the purposes of the meta-analysis presented in **Chapter 5** (in all cases and without any exceptions even if not explicitly mentioned), in order to reduce the effect of differences between experimental setups other than the factor under study.

C.2 Impact of DNA Repetitiveness and Short Read Mappability on ChIP-seq Analysis

During the course of the analysis presented in **Section 5.2.5**, I had noticed that a considerable number of genes were missing any sort of noteworthy regulatory signal. I hypothesised that this phenomenon was in part due to the fact that these genes were situated in highly repetitive regions of the genome or that they themselves were present in various copies throughout the genome.

In order to assess the validity of this hypothesis, I first used the GeneProf genome browser to examine the binding profiles in a wide window around three of the genes missing a ChIP-seq signal: *AC186033.1*, renamed *Zscan4f-ps* in the latest release of Ensembl, *Zscan4c* and *Zscan4f*. Looking at the surroundings of these genes in the GeneProf genome browser, revealed that there was indeed a distinct lack of aligned reads. The *Zscan4* family of genes, in particular, are highly similar in sequence.

It is common practice to accept only uniquely aligned reads for ChIP-seq data analysis in order to avoid ambiguity. For repetitive genomic regions this strategy might lead to fewer successfully aligned reads. The lack of binding signal in repetitive DNA regions might hence be an artifact of computational "mappability" rather than the consequence of genuinely low binding activity or an inability to capture binding events on the sample preparation-end of the experiment.

To check whether "mappability" had indeed an impact on the signal in this region, I realigned the raw data of six arbitrarily picked ChIP-seq datasets allowing for up to 10 ambiguous matches in the genome and compared the coverage profile with the unique alignment profile I was working with before. To my surprise, there was a striking difference in both profiles (**Figure C.2** left) at the locus of *Zscan4f* and other genes that were missing intensities across the board (*ESiC*-5). Hardly any difference was noticed for the majority of other genes (checked against (*ESiC*-1), e.g. *Nanog* (**Figure C.2** right). Thus, it appears that the repetitiveness of DNA does have an impact on the investigation of regulatory signatures by ChIP-seq in a subset of genes, but that it does not critically effect the conclusions I have drawn here with respect to the genes in *ESiC*-1.

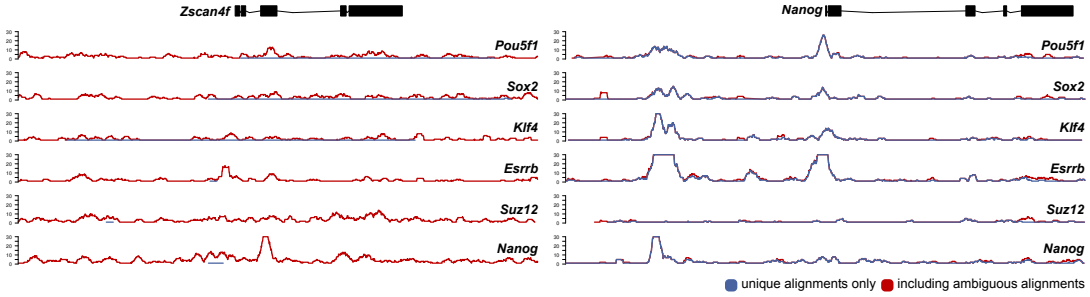

Figure C.2: Effect of repetitive DNA sequence on alignment. Genomic snapshot of a highly repetitive genomic locus (*Zscan4f*, left) and a less repetitive region for comparison (*Nanog*, right). Shown are alignments of six ChIP-seq datasets[75] allowing only uniquely mapped reads (blue) or all maps with up to 10 possible alignments (red).

Accounting for the problem of DNA repetitiveness (and hence alignment ambiguity) is not trivial. Ambiguity in alignments is an issue, to my knowledge, so far addressed almost exclusively for RNA-seq data, where approaches have been developed that take ambiguous alignments into account, including the method used in GeneProf (**Section 3.3.3.3**). All ChIP-seq peak-finding algorithms I have worked with so far, though, assume uniqueness of alignments. A literature search revealed two possible routes for addressing this issue: A post-alignment strategy for resolving ambiguously aligned reads based on the local genomic context has been proposed that would be compatible with existing ChIP-seq tools[579]. This is in principle not unlike my RNA-seq ambiguity resolution that makes use of information about other reads aligned to the same gene. Alternatively, the idea has been put forward to combine alignment and peak detection into a single step that could make use of ambiguity information in the statistical procedure[378]. The matter of which of these strategies is preferable or whether indeed either of them is capable of solving the issue at hand, certainly deserves further investigation that is beyond the scope of current study.

Appendix D

Additional Notes about the GeneProf Software and Algorithms

D.1 Access to Data and Analyses from this Thesis

All data and analyses presented in **Chapter 5** of this thesis can be accessed via the GeneProf software (`http://www.geneprof.org`) under the accession codes listed below. Please note, the corresponding experiments have not (yet) been made publicly available, therefore access is restricted to selected individuals ("collaborators"). If you wish to view these experiments, please register for a GeneProf user account and get in touch.

Accession	Experiment
gpXP_000557	Mouse ESC Universal ChIP-seq Background
gpXP_000558	Meta-Analysis of Transcription Factor Binding in ESCs
gpXP_000564	Meta-Analysis of Histone Modifications in ESCs
gpXP_000565	Meta-Analysis of Gene Expression in ESCs and other Cell Types
gpXP_000588	Meta-Analysis: Integration of Gene Expression, Transcription Factor Binding and Histone Modifications in ESCs
gpXP_000634	No signal due to mappability?

D.2 External Software and Algorithms Used

The GeneProf data analysis makes use of a great number of publically available third-party software. At the time of the first public release (coinciding with the writing of this thesis, GeneProf Version 1.1204041), the following is a comprehensive list of all relevant packages:

Basic Code Dependencies: Apache Commons software libraries, GNU Trove, JDOM, Java Secure Channel, JExcelAPI, JavaMail API, Legion of Bouncy Castle, Zehon File Transfer, Picard, SAMTools, Xstream, XPP3, Google Snappy.

Web Interace Dependencies: recaptcha4j, jQuery, jQuery UI, Adobe Spry, OpenJacob Draw2d, sprintf for JS, Swfupload, SACK, DHTMLGoodies Modal Dialog, jsplumb, snap2objects icons, FamFamFam icons.

External Programs Used: R, TexLive, ImageMagick, GraphViz, Bioconductor (various libraries), MACS, SISSRs, ChIPseqPeakFinder, Bowtie, Tophat, FASTX Toolkit, BEDTools, SRA Toolkit.

An up-to-date version of dependencies is maintained with the GeneProf software license online at `http://www.geneprof.org/terms_and_conditions.jsp`.

D.3 Data Compression

D.3.1 Performance of Assorted Compression Algorithms

This is an informal comparison of a number of widely-used, general-purpose compression algorithms on the dataset with the SRA accession number *SRR037952*[552]. Compression and decompression times were measured with the Unix tool `time` over one single trial and may hence differ slightly when repeated. All algorithms were tested using there implementation in Ubuntu Linux 10.04 with default compression level.

Algorithm	**File Size (bytes)**	**Ratio**	**Compression** (s)	**Decompression** (s)
None	6,613,373,443	1.0	0	0
GZIP	2,168,061,531	0.33	657.4	67.2
BZIP2	1,774,190,147	0.27	692.0	332.3
ZIP	2,168,061,785	0.33	918.0	73.8

D.3.2 Short Read Sequence Encoding

The following two algorithms are used to encode and decode nucleotide sequence for efficient in-memory storage in the Java programming language.

D.3.2.0.1 Encoding Algorithm

```
public final static long NUM_NUCLEOTIDES = 5;
public final static long ENCODING_MULTIPLIER =  NUM_NUCLEOTIDES + 1;
public final static int MAX_SEQ_LENGTH_PER_LONG = 24;

public static long[] sequence2longs(char[] nucs) {
        char[][] segments = splitStringInSegments(nucs,
                                                  MAX_SEQ_LENGTH_PER_LONG);
        long[] encoded = new long[segments.length];
        for (int i = 0; i < segments.length; i++) {
            encoded[i] = sequence2long(segments[i]);
        }
        return encoded;
}

private static long sequence2long(char[] nucs) {
        long pos = 1;
        long l = 0;
        for (char c : nucs) {
            l += getNucToInt(c) * pos;
            pos *= ENCODING_MULTIPLIER;
        }
        return l;
}

private static char[][] splitStringInSegments(char[] nucs, int len) {
        int l = (int) Math.ceil((double) nucs.length / ((double) len));
        char[][] segments = new char[l][];
        int start = Integer.MIN_VALUE, end = 0;
        for (int i = 0; i < l; i++) {
            start = end;
            end = Math.min(nucs.length, end + len);
            segments[i] = copyOfRange(nucs, start, end);
        }
        return segments;
}
```

D.3.2.0.2 Decoding Algorithm

```
public static String longs2sequence(long[] ls) {
        StringBuilder sb = new StringBuilder(ls.length
                                        * MAX_SEQ_LENGTH_PER_LONG);
        for (int i = 0; i < ls.length; i++) {
            sb.append(long2sequence(ls[i]));
        }
        return sb.toString();
}

private static String long2sequence(long l) {
        long tmp = l;
        StringBuilder sb = new StringBuilder(
                                SequenceEncoder.MAX_SEQ_LENGTH_PER_LONG
        );
        while (tmp > 0) {
            if (sb.length() > SequenceEncoder.MAX_SEQ_LENGTH_PER_LONG) {
                throw new RuntimeException("Error in encoded sequence.");
            }
            sb.append(getIntToNuc((int) (tmp % ENCODING_MULTIPLIER)));
            tmp /= ENCODING_MULTIPLIER;
        }
        return sb.toString();
}
```

D.4 Workflow Modules

All workflow modules available to all GeneProf users as of software version v1.1204041 are listed in the following tables ($n = 80$). In addition to these modules, another 28 are currently under development.

Modules marked with one asterisk (*) are so-called "meta-modules", that is, combinations of other modules that combine larger units of work into one simple and concise building block. Modules marked with two asterisks (**) are only available to administrators / super-users – they are being used to modify or augment the public database in GeneProf.

Name	Description
Add Annotations to Reference	Augment a reference dataset with annotations.
Align against cDNA with Bowtie	Align sequences to a transcriptome.
Align against DNA with Bowtie	Align sequences to a reference genome.
Align against Sequences with Bowtie	Align sequences to a arbitrary other sequences.
Assign TFBS to Genes	Assign ChIP-seq peaks to nearby genes.
Basic Features Filter	Filter feature data, e.g. by fold change or p-value.
Basic Genomic Regions Filter	Filter genomic regions, e.g. by FDR-values.
Basic Sequences Filter	Filter sequences on the basis of their annotations.
BEDTools: intersectBed	Return overlaps between genomic datasets.
Bowtie Output Parser	Parse genomic regions from Bowtie aligments.
Bowtie Output Parser (Mate-Paired)	Parse genomic regions from paired-end Bowtie aligments.
Calculate Additional Columns	Calculate new annotation columns for the given features.
Calculate Additional Columns (Region Data)	Calculate new annotation columns for the given genomic data.
Calculate TFAS	Calculate the TF association strength for each gene.
Center Peaks	Center peaks on their heighest point.
ChIP-seq Peak Summary	Summarise statistics about ChIP-seq peaks.
Compare Feature Data	Juxtapose multiple feature datasets.
Complex Features Filter	Filter feature data using complex criteria.
Complex Genomic Regions Filter	Filter genomic data using complex criteria.
Complex Sequences Filter	Filter sequence data using complex criteria.
Create Transcriptome-only Reference	Define new references by providing only a transcriptome assembly.
Define a new Reference Set	Define new references based on gene annotations, a transcriptome and a genome assembly.
DESeq	Assess differential expression with DESeq.
DESeq (for Region Data)	Assess differential expression with DESeq (for genomic data).
Calculate Fold Changes	Assess differential expression by fold change.
Calculate Fold Changes (Region Data)	Assess differential expression by fold change (for genomic data).
Drop Feature Annotation Columns	Drop annotatios from features.
Drop Region Annotation Columns	Drop annotations from a genomic data.
EdgeR	Assess differential expression with EdgeR.
EdgeR (for Region Data)	Assess differential expression with EdgeR (for genomic data).
Extract Regions from Reference	Extract genomic coordinates (e.g. promoters or exons) from a reference.
Extract Sequences from Regions	Extract the DNA sequences from genomic regions.
FASTA Parser	Parse sequence data from a FASTA-files.
FASTQ Paired-End Parser	Parse paired-end sequence data from a a single FASTQ-file.
FASTQ Paired-End Parser (2 Files)	Parse paired-end sequence data from two FASTQ-files.
FASTQ Parser	Parse sequence data from a FASTQ-file.
FASTX Toolkit: Artifacts Filter	Remove sequencing artifacts.
FASTX Toolkit: Clip Adapter Sequences	Remove adapter sequences.
FASTX Toolkit: Reverse Complement	Transform sequence to their reverse complement.
Feature Annotations Parser	Parse feature data (e.g. expression values) from a text file.
Find Peaks with CCAT	Find peaks ChIP-seq data using CCAT.
Find Peaks with ChIPSeqPeakFinder	Find peaks ChIP-seq data using ChIPSeqPeakFinder.
Find Peaks with MACS	Find peaks ChIP-seq data using MACS.
Find Peaks with SISSRs v1.4	Find peaks ChIP-seq data using SISSRs.
Gene Expression Summary	Summarise statistics about gene expression.
General Genomic Region Statistics	Summarise statistics for genomic datasets.
General Sequence Statistics	Summarise statistics for nucleotide sequence datasets.
Generic Sequence Parser	Guess the file format and parse sequence data.
Genomic Region Parser	Parse genomic data from text files (e.g. BED).

Name	Description
Genomic Region Parser	Parse genomic data from text files (e.g. BED).
GOSeq Enrichment Analysis	Gene ontology enrichment analysis with GOSeq.
MACS + Gene Association + Statistics	Use MACS to detect peaks, assign them to nearby genes and creates a summary report in one step. *
Main Experimental Results	Mark a selection of datasets in a workflow as the main results.
Make Annotations Public	Add datasets to the public collection of searchable data. **
Make Reference Public	Add a reference to the public collection of recommended reference datasets. **
Make Tracks Public	Add datasets to the public collection of browser tracks. **
Map Features to Another Reference	Map the features in a feature dataset onto another reference.
Map Regions to Genes	Assign genomic regions (e.g. ChIP peaks) to nearby genes.
MEME Motif Discovery	Find DNA motifs using MEME.
Merge Genomic Region Data	Merge multiple genomic datasets.
Merge Sequence Data	Merge multiple sequence datasets.
Modify and Filter Sequences	Trim, expand or alter sequences and apply permanent filters.
Modify Genomic Regions	Trim, expand or merge all regions in a dataset.
Parse Reference Set from GenBank	Define new reference sets by parsing GenBank files.
Put Aligned Reads into Bins	Split genome into bins and count reads aligned to each bin.
QC + Bowtie	Filter reads and align them using Bowtie. *
QC + Bowtie Iterative Trimming Alignment	Quality control and repeated cycles of alignment followed by read trimming. *
QC + Tophat	Filter reads and align them using Tophat. *
Quantile Normalisation	Apply a quantile normalisation.
Quantitate Coverage in Regions	Calculate the read count for each provided genomic region.
Quantitate Gene Expression	Calculate an expression value for each gene.
Quantitate Promoter Activity	Quantify the coverage intensity for each promoter.
Random Sample of Features	Select a random subset of features.
Random Sample of Genomic Regions	Select a random subset of genomic regions.
Random Sample of Sequences	Select a random subset of sequences.
Raw Sequence Parser	Parse raw sequences from a file containing one sequence per line.
SAM/BAM Region Parser	Parse genomic data from a SAM- or BAM-formatted file.
Select Regions for Regions	Select the regions whose IDs are in a genomic dataset.
Select Sequences for Regions	Select the sequences whose IDs are in a genomic dataset.
Select Sequences for Sequences	Select the sequences whose IDs are in another sequence dataset.
Separate Mate Sequences	Separate paired-end sequences into two independent sequences.
Split Sequences into Mate Pairs	Split single-end sequences into two separate sequences (mate-pairs).
SRA File Parser	Parse sequences from an SRA- or SRAlite-formatted file.
TopHat Alignment	Align sequences to a genome using the Tophat.