THE UNIVERSITY *of* EDINBURGH

Distinct transcriptional signatures of aneuploidy in murine pluripotent cell populations

Stavroula Skylaki

Thesis presented for the degree of

Doctor of Philosophy

The University of Edinburgh

2011

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Stavroula Skylaki)

Acknowledgements

I would like to thank, first and foremost, my supervisor, Simon Tomlinson, for his guidance and knowledgeable comments throughout the completion of this thesis; for having his door always open and taking the time to discuss my concerns at any stage of this demanding journey. I would also like to thank the members of my PhD committee Val Wilson and Clare Blackburn for their assistance and advice; the members of our lab: Ed, Florian, Sofia, Aidan, Harsh, Will, Duncan and Hina that have been making our group a fun and supporting working environment; all the members of Thursday's developmental group meeting which invited me to participate and helped me improve my understanding of biology. I cannot express my gratitude enough to Val Wilson, Kei Kaji and Ian Chambers for their insights and comments for my manuscript based on this project, and Josh Brickman for his help and advice towards future career directions.

Many thanks to my friends and flatmates throughout these four years for keeping me optimistic during the hard times. I particularly want to thank my best friend, Filip Wymeersch, with whom we got through the whole PhD together and, among numerous coffee breaks, we managed to keep each other's spirits up (most of the time).

I'm grateful to Mike for his love, support and immense understanding that kept me going and looking forward to the future. And, of course, for the exceptional designing of the DI.S.C.O. logo. Finally, I will always be grateful to my family, my parents, Iakovo and Pepy, and my two sisters, Eleni and Maria, to whom I dedicate this thesis.

Table of Contents

Abstract

Genomic integrity in mouse embryonic and induced pluripotent stem cells can be compromised by factors such as extended time in culture and cellular reprogramming. Surprising, only a few studies have thus far examined the accumulation of chromosomal imbalances in mouse pluripotent populations upon prolonged propagation in vitro. It is presumed that specific recurring genetic changes can confer selective growth advantage and resistance to apoptosis and/or differentiation to the affected cells, although the genes that drive these processes remain elusive. The presence of these changes in published studies can confound the analysis of the data and hinder the reproducibility of the results.

At the transcriptional level, aneuploidy manifests as large chromosomal regions of aberrant gene expression. This thesis presents a method to identify these regions in large-scale datasets and interrogate for recurrent patterns. The present analysis shows that over half of the 315 mouse pluripotent samples examined carry whole or partial-chromosome spanning clusters of aberrant transcription. Furthermore, there are common gene expression changes across samples with any type of predicted aneuploidy and samples with chromosome-specific aberrations. These transcriptional signatures have been used to train classification models which can predict aneuploid samples with over 90% accuracy. This is an important step towards the development of a low-cost and reliable transcriptional validation assay for the presence of aneuploidy.

List of Publications

Journals:

- S. Skylaki, I. K. Sarantidis, S. R. Tomlinson, "DI.S.C.O. A tool for the DIscovery of Subtle Clustered Organisation in the genome." *In preparation.*

- S. Skylaki, S. R. Tomlinson, "Distinct transcriptional signatures of aneuploidy in murine pluripotent cell populations." *Submitted*.

Poster Presentations and Conferences:

- S. Skylaki, S. R. Tomlinson, "Large-scale integrated search for genomic clusters of aberrant transcriptional activity in mouse pluripotent stem cell populations", Stem Cell Biology Meeting, CSHL, September 20-24, 2011, Cold Spring Harbor. *Poster*.

- S. Skylaki, I. K. Sarantidis, S. R. Tomlinson, "Large-scale search for potential karyotypic instabilities in murine embryonic stem cell populations.", 8th Annual Meeting of the International Society for Stem Cell Research (ISSCR), June 16-19 2010, San Francisco, CA. *Poster*.

- S. Skylaki, I. K. Sarantidis, S. R. Tomlinson, "Sensitive Identification of Genomic Clusters of Differentially Expressed Genes", ADVANCES IN STEM CELL RESEARCH: Stem cells, systems & synthetic biology, June 15-17 2009, Cambridge, UK. European Molecular Biology Organization. *Poster*.

- S. Skylaki, I. K. Sarantidis, S. R. Tomlinson, "Sensitive Identification of Genomic Clusters of Differentially Expressed Genes", Hydra IV: Stem Cells & Regenerative Medicine Summer School, September 5-11 2008, Hydra, Greece. Organizers: EuroSystem Consortium. *Poster*.

List of Figures

List of Tables

Abbreviations

aCGH	array Comparative Genomic Hybridization
AUC	Area Under the ROC Curve
CIN	Chromosomal Instability
CNV	Copy Number Variation
CFSs	Common Fragile Sites
DI.S.C.O.	Discovery of Subtle Clustered Organization
EC	Embyonal Carcinoma
ESC	Embryonic Stem Cell
mESC/hESC	mouse/human Embryonic Stem Cell
ECM	Extra-Cellular Matrix
FISH	fluorescence *in situ* hybridization
FC	Fold Change
FDR	False Discovery Rate
GE(P)	Gene Expression (Profiling)
GUI	Graphical User Interface
HSPC	Hematopoietic Stem/Progenitor Cells
iPSC	induced Pluripotent Stem Cell
miPSC/hiPSC	mouse/human induced Pluripotent Stem Cell
mFISH	multicolour Fluorescent In Situ Hybridisation
MEFs	Mouse Embryonic Fibroblasts
MSC	Mesenchymal Stem Cell
MIN (MSI)	Microsatellite Instability
NSC	Neural Stem Cell
PAM	Prediction Analysis of Microarray
SAM	Significance Analysis of Microarrays
SKY	Spectral Karyotype
SNP	Single-Nucleotide Polymorphism
SOMs	Self-Organising Maps
SVs	Structural Variations
SVMs	Support Vector Machines
SVM-RFE	SVM with Recursive Feature Elimination
TV	Total Variation

UCSC	University of California, Santa Cruz
XCI	X Chromosome Inactivation

1. Introduction

1.1. Introduction

This thesis is concerned with the identification of aneuploidies and sub-chromosomal aberrations in mouse pluripotent stem cell populations by means of bioinformatics analysis of large collections of gene expression data. A number of recent studies have reported recurrent patterns of aneuploidy in human and mouse pluripotent cells that undergo culture for prolonged periods of time (Draper et al., 2004; Maitra et al., 2005; Baker et al., 2007; Amps et al., 2011; Laurent et al., 2011; Liu et al., 1997; Sugawara et al., 2006). These studies mainly perform conventional cytogenetic or array-based analysis to characterise the karyotype of the cells. However, the vast majority of published datasets in the field of stem cell biology consists of gene expression profiling data that are typically not accompanied with appropriate cytogenetic validation. When aneuploidies or other chromosomal alterations pass undetected in a sample, they may confound the experimental conclusions and hinder the reproducibility of the results. It has been shown that genomic changes can be detected at the transcriptional level as they can alter the expression levels of a high percentage of the genes present in the affected region (Pollack et al., 2002; Hyman et al., 2002; Schoch et al., 2006). It is therefore possible to use the transcriptional data for a first and quick test of the genomic integrity of the samples and several bioinformatics tools have been developed for this purpose ((Nilsson et al., 2008; De Preter et al., 2008; Callegaro et al., 2006) among others). Nonetheless, the application of these tools is still not common practise as indicated by the low number of stem cell related publications that report their use. This could be due to certain factors that might be limiting for the non-expert user such as the complexity of the computational model which they apply, lack of user-friendliness or requirement of prior familiarisation with statistical or computational packages. This thesis aims to overcome these limitations by providing a powerful and intuitive software application that can be used for the analysis of gene expression data and the identification of underlying patterns of chromosomal aberrations.

In addition, the present thesis is concerned with the discovery and quantification of genomic changes in large collections of murine pluripotent stem cell data. This is

particularly interesting since the obvious model for the recurrence of specific aneuploidies is their ability to confer a selective advantage to the carrier cells. Unveiling the commonly affected genomic intervals may reveal the candidate genes in the region, shed light to the mechanisms of growth advantage ascribed by aneuploidy and most importantly it may lead to a better understanding of normal and disease processes in stem cell biology. This thesis presents a large-scale integrated search methodology for the prediction of recurring patterns of aneuploidy in large collections of transcriptional data and reports the results of this type of analysis in mouse embryonic and induced pluripotent stem cells.

Finally, given the large quantity of analysed gene expression profiles, the present thesis explores the presence of distinct transcriptional signatures associated with different types of aneuploidy affecting specific chromosomes. In addition, the presence of a transcriptional signature linked to *any* type of aneuploidy discovered in the data, possibly as a downstream effect of the presence of any type of chromosomal aberration that can give a selective advantage to the cell, is also identified for the first time. This thesis uses these signatures to train classification models that can be used for the accurate prediction of genomic changes in uncharacterised samples.

To sum up, the scientific developments described in this thesis include the following:

- Development of a software application for the prediction of genomic aberrations at the transcriptional level for a single-experiment type of analysis.
- Development of a framework for the analysis of large collections of genome-wide expression data from mouse pluripotent stem cell populations for the prediction of chromosomal abnormalities
- Prediction of frequently recurring aneuploidies affecting specific chromosomes
- Discovery of distinct transcriptional signatures related with chromosome-specific aneuploidies
- Discovery of a global signature linked to the presence of any type of underlying aneuploidy
- Training and validation of classification models that can use a small number of diagnostic genes to predict the presence of aneuploidy in uncharacterised samples.

1.2. Thesis Structure

The structure of the thesis is as follows:

The Background presents an overview of the field of embryonic stem cell and induced pluripotent stem cell biology. A summary of the findings to date regarding the loss of genomic integrity in human and mouse pluripotent stem cell lines is also presented. In addition, a brief description of the transcriptomics tools and methods as well as the fundamental concepts of the machine learning techniques that are relevant to this thesis is provided.

Chapter 3 focuses on the software application DI.S.C.O. (DIscovery of Subtle Clustered Organisation). DI.S.C.O. can be used in order to identify clusters of concordant changes of gene expression levels that can be also diagnostic of underlying aneuploidies. The requirements, the development and the functionality of the tool as well as detailed description of the integrated computational methods are presented. In addition, extended validation of the computational methods in DI.S.C.O. with synthetic and biological data is performed.

Chapter 4 presents a methodology for the analysis of large collections of gene expression data for the prediction of patterns of aneuploidy. The application of the framework in a collection of mouse pluripotent stem cell samples and the results from the analysis are described.

Chapter 5 is concerned with the discovery of transcriptional signatures that are associated to frequently recurring chromosomal abnormalities in mouse pluripotent stem cells or to the presence of any type of aneuploidy in general. In addition, it discusses the training and the validation of classification models that can predict these aberrations in previously uncharacterised datasets using only a small number of diagnostic genes.

Chapter 6 contains the final discussion and the future directions of this work.

Finally, in the interest of brevity, supplemental material can be found in the attached

CD (referenced in the main text when appropriate). The list of supplemental files included in the CD can be found in the Appendix.

2. Background

This chapter presents the origin and the basic biological concepts of embryonic and induced pluripotent stem cells. It then explores the genomic integrity of human and mouse pluripotent stem cells and provides a summary of the findings of different studies about the type and the frequency of genomic changes observed in these cell populations. Finally, it gives an overview of the transcriptomics tools and the machine learning techniques that are relevant to the present study.

2.1. Pluripotent Stem Cells

2.1.1. Pre-implantation Mouse Embryo Biology

Mouse embryonic development starts with the union of the sperm and the oocyte to form the fertilised egg (*zygote*). The embryo travels along the oviduct and in 4.5 days after fertilisation it is implanted in the uterus. The period between *fertilisation* and *implantation* consists of two distinct stages of early-embryo development, the *cleavage* and the *blastulation.* As cleavage proceeds from the two-cell stage embryo to 16-cell stage, accompanied with the activation of embryonic genes and the rapid degradation of the majority of maternally expressed mRNAs, two distinct lineages arise: the *trophectoderm* (TE) and the *inner cell mass* (ICM). The segregation of the two lineages starts with compaction and polarisation of the cells to form the *morula* at the 16-cell stage. The morula consists of two cell populations, the inner cells that will eventually give rise to the ICM, and the outer cells that will subsequently give rise to the trophoblast cells (trophectoderm). This process is completed through *cavitation* where the outer trophoblast cells start secreting fluid into the morula in order to create the *blastocoel*. The resulting structure, shown in Figure 2.1, is called the *blastocyst* and it consists of the ICM and the trophectoderm (Gilbert, 2000). By the 64-cell stage, the two lineages are completely segregated (Dyce et al., 1987). Prior to implantation, the ICM cells form the *epiblast* which, after implantation, will progressively differentiate into the three germ-cell layers, *endoderm*, *ectoderm* and *definitive mesoderm* and eventually form the embryo proper after a series of successive lineage-commitment steps.

2.1.2. Embryonic Stem Cells

Embryonic stem cells (ESCs) are derived from the ICM cells of the pre-implantation embryo. They have the ability to self-renew through symmetric division *in vitro* (Smith, 2001), presumably indefinitely under the appropriate culture conditions (Suda et al., 1987), and to give rise to cells from any of the three primary germ layers *in vitro* and *in vivo* (Beddington and Robertson, 1989). The isolation of ESCs from two independent groups in 1981 is one of the most important achievements of mammalian developmental biology. ESCs were first isolated from mouse blastocysts and propagated in culture using feeder layers by Evans and Kaufman (1981) and Martin (1981). As it was later discovered, this had been possible because mouse fibroblast feeder cells express the leukaemia inhibitor factor (LIF) which blocks differentiation (Gearing et al., 1987; Smith et al., 1988). Mouse ESCs were subsequently derived with the use of LIF without feeders (Smith and Hooper, 1987).

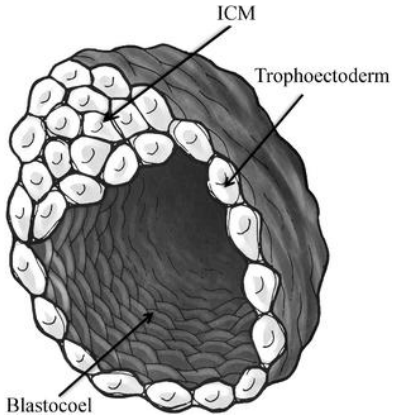

Figure 2.1 Schematic representation of the blastocyst of the mouse embryo. [Image from (De Miguel et al., 2010)]

The idea that the ICM might contain cells that could be kept in an undifferentiated state was based on the study of embryonic carcinoma cells (ECs). ECs are pluripotent cells derived from *teratocarcinomas*, a malignant neoplasm that contains undifferentiated cells as well as cells from all the three primal germ layers in various stages of differentiation. ECs can be maintained *in vitro* under defined conditions and demonstrate similar patterns of differentiation with normal ICM cells during development (Martin and Evans, 1974; Martin and Evans, 1975; Martin, 1975). When

ECs are injected in a host blastocyst they fail to contribute to the germ line and they often show aneuploidy (Chambers and Smith, 2004).

Contrary to ECs, ESCs can contribute to the germ line after blastocyst injection and re-implantation to a surrogate mother (Bradley et al., 1984). It was this important characteristic that paved the way for the genetic manipulation of ESCs (Robertson et al., 1986; Gossler et al., 1986), the generation of transgenic mice and their subsequent use for the study of specific gene mutations and as disease models. These findings highlighted the great number of potential applications of ES cells for the understanding of normal biological processes, drug discovery and regenerative medicine.

The isolation of human ESCs by Thomson et al. (1998) marked an important milestone that made the culture and directed differentiation of hESCs possible and enabled their applications in medical research as means of transplantation and drug testing. However, it has been suggested that human ESCs represent a later developmental stage than mouse ESCs and they depend on different extrinsic pathways that are more similar to the recently isolated mouse post-implantation epiblast-derived stem cells (EpiSCs) (Tesar et al., 2007; Brons et al., 2007). In addition, the process of X-chromosome inactivation (XCI) is an active field of research in both human and mouse ESCs and iPSCs. Current findings indicate that both human ESCs and mouse EpiSCs have undergone X-chromosome inactivation (XCI), contrary to mouse ESCs where both X chromosomes are active (Ng and Surani, 2011). It has been recently shown that ectopic expression of specific transcriptional factors in combination with defined culture conditions can revert human ESCs in a “naive” state of pluripotency more similar to mouse ESCs (Hanna et al., 2010).

2.1.3. Extrinsic pathways of Self-Renewal

As it has been mentioned in section 2.1.2, the initial derivation of mouse ESCs on feeder layers was possible because they were able to maintain self-renewal through the inhibition of differentiation that is conferred by the feeder-supplied LIF (Smith et al., 1988). LIF, a member of the IL-6 type of cytokines, and several related members of the IL-6 type cytokines can inhibit differentiation via the gp130 receptor (Yoshida et al., 1994). The activation of gp130 by homodimerisation or heterodimerisation with the

LIF receptor is required for the activation of the JAK/STAT signalling pathway and the SHP2/Erk mitogen-activated protein kinase cascade. These two pathways are involved in the fine tuning of the equilibrium between self-renewal and differentiation (Niwa et al., 1998; Burdon, 1999; Matsuda, 1999) (Figure 2.2).

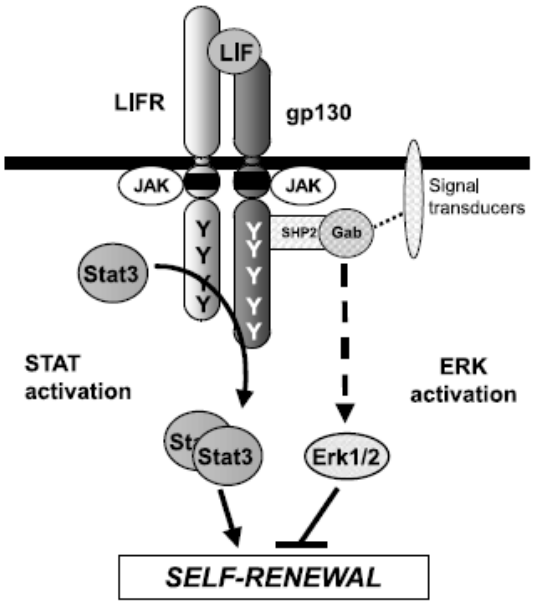

Figure 2.2 Extrinsic signalling pathways of ESC self-renewal and differentiation. ESC self-renewal and differentiation is tuned by the Lifr/gp130 mediated activation of the JAK/STAT and SHP2/Erk signalling pathways [Image from (Smith, 2001)].

In addition, Ying et al. (2003) found that the combination of LIF with BMP4 resulted to enhanced self-renewal and obliterated the need for serum during ESC derivation and culture. BMP4 blocks mainly neuronal differentiation by activating Smads which subsequently induce the transcription of Id (Inhibitors of Differentiation) genes.

Recently, it has been shown that ESCs can be also maintained in culture with the use of two small-molecules that inhibit the Fgf4-mediated Erk activation as well as an inhibitor of Gsk3β (Ying et al., 2008). This is termed the "3i" culture medium with the optional addition of LIF. (Figure 2.3).

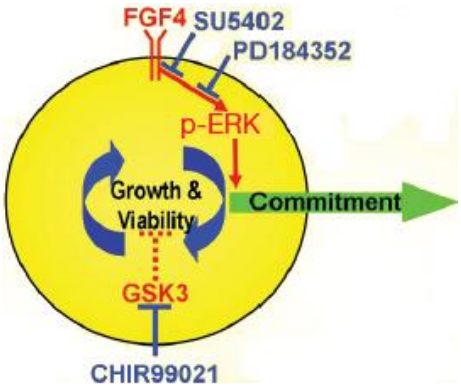

Figure 2.3 The "3i" medium inhibited pathways in mouse ESC cultures [Image adapted from (Ying et al., 2008)]

2.1.4. The Core Transcriptional Network of Pluripotency and Self-Renewal

Besides the extrinsic signalling pathways discussed in the previous section, a delicate intrinsic orchestration of a number of transcription factors is required for the maintenance of pluripotency and self-renewal. The ternary of *Oct4*, *Nanog* and *Sox2,* specifically, is found in the core of the transcriptional network that determines self-renewal and differentiation. During mouse development, both *Oct4* and *Nanog* are required for the establishment of the pluripotent identity, but not *Sox2*, presumably because of the high levels of lingering maternal *Sox2* (Chambers and Tomlinson, 2009).

Oct4 (a member of the POU class of transcription factors, also known as *Pou5f1*, *Oct3*, *Oct3/4*) is one of the most extensively studied intrinsic regulators of pluripotency and cell fate decisions in ESCs. Upon *Oct4* deletion *in vivo* the cells of the blastocyst differentiate to trophectoderm (Nichols et al., 1998). *Oct4* expression is also required for ESC maintenance *in vitro* in a LIF-independent way (Niwa et al., 2000). Niwa and colleagues used two ES cell lines to study the effects of *Oct4* gene dosage in self-renewal and pluripotency: (i) the ZHTc6 cell line where one *Oct4* allele is disrupted and *Oct4* expression from a tetracycline-suppressible *Oct4* transgene can be induced upon removal of tetracycline and (ii) the ZHBTc4 cell line where both *Oct4* alleles are inactivated and the tetracycline-suppressible *Oct4* transgene is the only source for *Oct4* expression. Niwa et al. (2000) has shown that there are specific *Oct4* gene dosage effects that control lineage specification. Less than 50% expression of *Oct4* results in differentiation to trophectoderm, while a two-fold increase induces differentiation to

endoderm or mesoderm. Nevertheless, expression of *Oct4* is not adequate to prevent differentiation upon LIF withdrawal (Niwa et al., 2000).

Nanog is a homeodomain protein that is specifically expressed in ESCs (Chambers et al., 2003; Mitsui et al., 2003). Over-expression of *Nanog* is sufficient to sustain pluripotency in the absence of STAT3 activation (Chambers et al., 2003; Mitsui et al., 2003). This indicates that *Nanog* is not downstream of the LIF signalling cascade. However, in the presence of LIF, *Nanog* over-expressing ESCs demonstrate enhanced self-renewal which could suggest that *Nanog* and STAT3 act independently but activate overlapping downstream transcriptional events (Chambers et al., 2003). Moreover, (Chambers et al., 2007) have shown that *Nanog* expression is heterogeneous in ES cell cultures, with cells expressing low levels of *Nanog* being more prone to differentiate. Interestingly, *Nanog*-null cells can still self-renew and contribute to all somatic tissues of chimaeras but not the germ line (Chambers et al., 2007).

Sox2 is expressed in ESCs (Avilion et al., 2003) and neural stem cells (NSCs) (Zappone et al., 2000). Deletion of *Sox2* in ESCs has similar characteristics with *Oct4* deletion and results to differentiation towards the trophectoderm lineage (Masui et al., 2007). *Sox2* and *Oct4* bind DNA cooperatively, as it has been shown for the activation of the *Fgf4* enhancer in mouse ES and EC cells (Ambrosetti et al., 1997; Ambrosetti et al., 2000).

Great efforts have been made towards the identification of downstream targets of these transcription factors (see (Chambers and Tomlinson, 2009) for a review). Understanding the exact molecular mechanism that the *Nanog/Oct4/Sox2* trio employs in order to maintain pluripotency is an ongoing goal for the stem cell community.

2.1.5. Induced Pluripotent Stem Cells

The differentiated cells of the adult organism can be forced towards an embryonic-like state by a process that reverts the unidirectional lineage specification: reprogramming. Different methods have demonstrated thus far that adult multipotent or terminally differentiated cells still retain the nuclear plasticity that allows them to reset to an ES-like state under specific conditions. This can be achieved by:

- Nuclear transfer (see (Hochedlinger and Jaenisch, 2006) for a review)
- Cell fusion ((Tada et al., 2001; Cowan et al., 2005))
- Direct reprogramming with ectopic expression of specific transcriptional factors (Takahashi and Yamanaka, 2006; Wernig et al., 2007; Maherali et al., 2007)

In a bench-mark experiment, (Takahashi and Yamanaka, 2006) showed that ectopic expression of the transcription factors *Oct4, Klf4, Sox2* and *c-Myc* by retroviral transduction can generate cells that resemble the blastocyst-derived ES cells in morphology and transcriptional profile. These cells were named induced pluripotent stem cells (iPSCs) by Yamanaka. These first iPSCs were selected by the activation of a drug resistance allele incorporated in the *Fbxo15* locus which is specifically expressed in ESCs. *Fbxo15*-based selection could only identify, however, partially reprogrammed iPSCs that did not demonstrate full demethylation of key developmental regulators such as *Oct4,* could not give raise to post-natal chimeras and contribute to the germ line (Takahashi and Yamanaka, 2006). Follow-up experiments refined the procedure by basing the selection on the expression of *Nanog* and *Oct4* (Okita et al., 2007; Wernig et al., 2007; Maherali et al., 2007), even by using integration-free mechanisms (Okita et al., 2008; Stadtfeld et al., 2008b).

Following these initial studies, numerous researchers have demonstrated that defined transcription factors can be introduced to different types of mouse lineage-restricted cells or terminally differentiated cells for iPSCs generation (i.e. liver and stomach cells (Aoi et al., 2008), bone marrow cells (Kunisato et al., 2010), terminally differentiated pancreatic β-cells (Stadtfeld et al., 2008a) and lymphocytes (Hanna et al., 2008)). In addition, iPSCs have been also generated from other species, such as human (Takahashi et al., 2007; Yu et al., 2007; Park et al., 2008), rat (Li et al., 2009) and rhesus monkey (Liu et al., 2008).

iPSCs could provide a cell source per individual with invaluable therapeutic applications as well as a model for understanding the mechanisms of normal development. Nonetheless, there are still many parameters that need to be investigated before such an application is possible. Safe and efficient generation of fully reprogrammed iPSCs can depend on many molecular and biological processes, including epigenetic modifications, micro-RNA regulation, cell cycle and DNA damage

control and apoptosis, that still need to be fully understood (Na et al., 2010).

2.2. Genomic Integrity of Pluripotent Cells

2.2.1. Genomic instability

Genomic instability refers to the processes that can increase the mutation rates in the genome and it is believed to contribute to carcinogenesis. Genomic instability mainly manifests by two mechanisms: *chromosomal instability* (CIN) and *microsatellite instability* (MIN or MSI). Microsatellites consist of multiple repetitive short sequences of DNA. These sequences also exist in normal cells but appear lengthier in certain tumours and are the result of faulty DNA repair pathways (Lengauer et al., 1997). Although, MIN may not necessarily affect the phenotype of the cell, it is by definition a type of genomic instability that has been observed in certain human tumours (Morgan et al., 1996).

Chromosomal instability affects whole or large parts of the chromosomes and includes events such as duplications, deletions, partial gain or losses and translocations. *Aneuploidy* can be defined as the deviation of the modal chromosomal number of an organism by means of gain or loss of whole or partial chromosomes. *Segmental aneuploidies* refer to the acquisition of extra chromosomal regions from amplification, deletions or rearrangements (Geigl et al., 2008). As shown in Figure 2.4, chromosomal abnormalities can be either *numerical* (when they affect the overall number of chromosomes in the cell) or *structural* (when they result in an alteration of the normal chromosome structure) (Atwood, 2011). Aneuploidy is closely related to CIN but it is not necessarily the case that aneuploid cells demonstrate higher rates of CIN (Geigl et al., 2008).

2.2.2. Tools for the Detection of Chromosomal Aberrations

There are various methods that can be applied for the identification of chromosomal abnormalities and the evaluation of the genomic integrity of a single cell or a cell population. These tools can be grossly divided in three categories: conventional cytogenetic analysis, molecular cytogenetics and array-based techniques (Catalina et al., 2007; Gijsbers and Ruivenkamp, 2011).

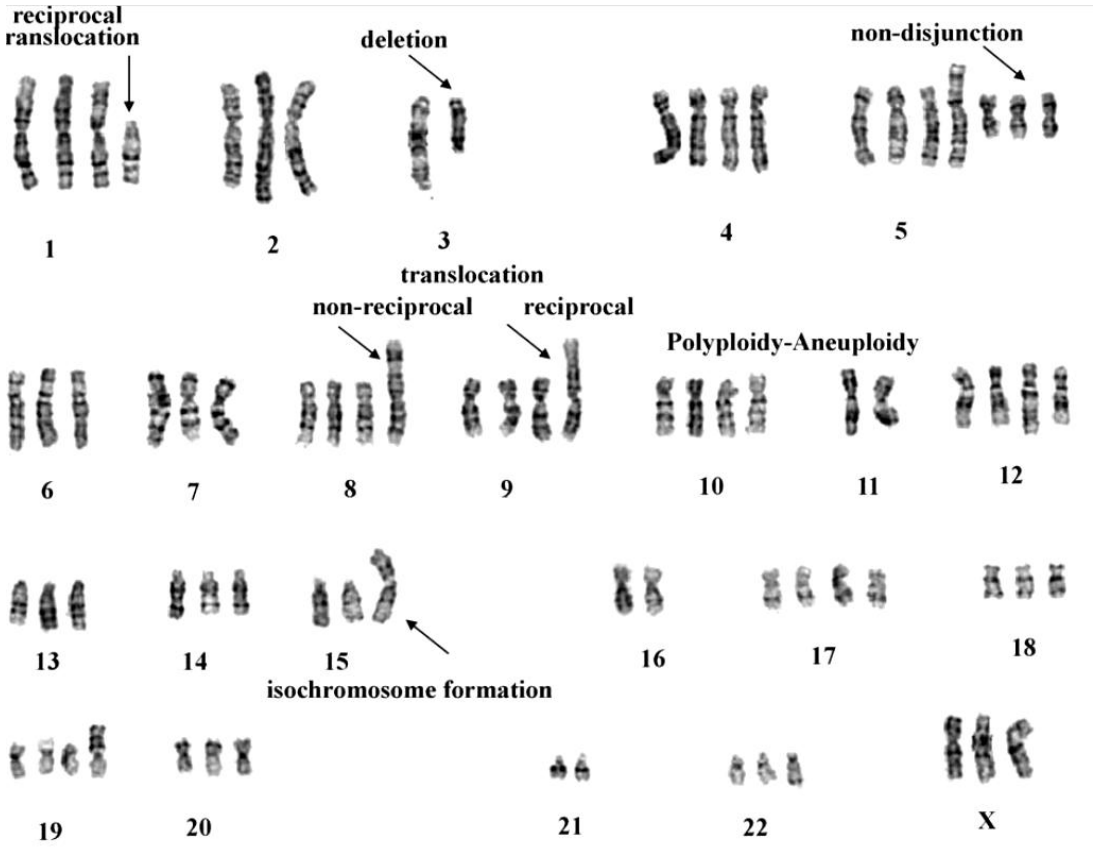

Figure 2.4 G-banding karyogram.
A G-banding karyogram of a HeLa cell (cervical tumour) consisting of various examples of numerical and structural chromosomal aberrations
[Image adapted from (Gagos and Irminger-Finger, 2005)]

Classical cytogenetic analysis examines the chromosome complement of a cell during metaphase where the condensed chromatin is visible in the form of chromosomes. The metaphase chromosomes can be stained producing distinct *banding* patterns. Chromosome bands arise from specific chromosomal regions that may vary in size and number of encompassing genes. Popular banding techniques include *G-banding* (trypsine Giemsa staining), *R-banding* (reverse bands of Giemsa staining with heat treatment) or *DAPI-banding* (fluorescent based 4'-6-Diamidino-2-Phenylindole (DAPI) staining) (Bayani and Squire, 2004). The above techniques produce ~300-400 stained bands and allow the identification of aneuploidies as well as gross chromosomal aberrations (gain, losses or rearrangements) (Rebuzzini et al., 2011). The achieved resolution is quite low ~5-10 Mb, nonetheless, the time and cost requirements of these methods are limited and thus they are widely adopted (Gijsbers and Ruivenkamp, 2011).

Fluorescent in situ hybridization (FISH) was the first cytogenetic technique based on principles of molecular biology, introduced in the early 80's (Van Prooijen-Knegt et al., 1982). In FISH fluorescent labelled DNA probes designed to hybridise to specific genomic regions of interest can be visualised under a fluorescent microscope. The resolution depends on the size of the probe, around 1-2 Mb. FISH is mostly used for the examination of the integrity of already known conspicuous regions i.e. recurrent aberrations in cancer. It is however a quite labour intensive technique (Gijsbers and Ruivenkamp, 2011). Two more molecular cytogenetic methods, similar to FISH, have been recently developed: *spectral karyotype* or SKY and *multicolour fluorescent in situ hybridization* (mFISH) (Schröck et al., 1996; Speicher et al., 1996). Both these methods output the whole karyotype and they have a much higher resolution than conventional cytogenetic techniques (~1-2 Mb). Since each chromosome is painted with a different fluorescent colour, inter-chromosomal rearrangements can be identified. Not all types of abnormalities can be however discovered. Small deletions or amplification as well as inversions are missed.

Microarray-based technologies offer much higher resolution for the detection of small genetic imbalances even at the level of single nucleotides. Two methods have been widely adopted: *array comparative genomic hybridization* (array-CGH) and *single nucleotide polymorphism array* (SNP array) (Solinas-Toldo et al., 1997; Peiffer et al., 2006). The aCGH technology uses DNA arrays that contain bacterial artificial chromosome (BAC) clones, P1-derived artificial chromosomes (PAC) or yeast artificial chromosomes (YAC) covering the whole genome (Solinas-Toldo et al., 1997). The resolution of these arrays is higher and depends on the size and the spacing of the probes used. Initially, a collection of 2,400 BAC clones was used to cover the human genome at a resolution of approximately one clone per 1.4 Mb (Snijders et al., 2001). Later designs, using 30,000 BAC clones and reaching a full coverage resolution have been also reported (Ishkanian et al., 2004). For mouse, the early 3K BAC arrays contained 3080 mouse BAC clones spanning across chromosomes 1 to 19 and X with a 1 Mb resolution (Chung et al., 2004). Additional improvements with the use of oligonucleotides instead of BAC clones have increased the resolution to ~100 Kb (van den IJssel, 2005). The technique is based on the competitive hybridisation of a test and control DNA sample differentially labelled with i.e. red and green fluorochromes. The relative fluorescence intensities of the two channels are calculated after the array is

scanned and imaged. The generated data can be then visualised in appropriate software packages and the intensity ratios can be analysed in order to identify regions of *copy number variations* (CNVs). In contrast to classical cytogenetic techniques, aCGH does not require the extra step of fixing the metaphase spreads and provides higher resolution. However, the equipment and reagents required are more expensive in comparison to conventional cytogenetic techniques.

SNP arrays consist of attached short fragments of DNA (probes), usually of a 25 to 50 nucleotides length, that are complementary to sequences around the SNP loci. Fragmented single-stranded DNA from the test sample is hybridised to the array and specialised equipment quantifies the signal intensity that results from the probe and the target DNA hybridization. Major commercial manufacturers of SNP arrays include Affymetrix and Illumina which employ slightly different methodologies but based on the same principles (see (LaFramboise, 2009) for a review). An advantage of SNP arrays is that they can be used for the identification of copy-neutral *loss of heterozygosity* (LOH) events, such as *uniparental disomy* (UPD), as well as CNVs. Of course, in the case of inbred organisms where all genomic loci are homozygous, SNP analysis can be only informative when performed across different inbred strains.

The downside of the array-based cytogenetics is that they assess the genomic integrity of a population of cells. As a result, they may not be able to discover subtler patterns present only in a few cells nor report the frequency of an aberration across the whole population. For example, a study has reported that an aberration present in 33% of the cells assessed by conventional karyotyping could not be detected by array-CGH, while the aberration became detectable when carried by 75% of the cells (Veltman et al., 2003). The lower limit of the method for the detection of mosaicism in the cell population is not clear and specific experiments are needed to address this question.

The techniques described in this section have different time, labour intensity, equipment and consumables requirements and thus they can be more appropriate for different types of experiments. To date, conventional cytogenetics and even simple metaphase counting is the method most routinely applied for the investigation of chromosome instability.

2.2.3. Culture Adaptation

In the general sense, culture adaptation refers to the ability to establish, expand and maintain a cell population *in vitro.* In the context of pluripotent stem cells, Baker et al. (2007) has proposed an alternative definition according to which *culture adaptation* refers to the phenomenon where pluripotent stem cells exhibit decreased population doubling times, increased cloning efficiency and accumulation of karyotypic abnormalities after prolonged time in culture (Baker et al., 2007). For the remaining of this thesis, the definition of Baker et al. (2007) is adopted.

The indefinite self-renewal observed in ESC cultures *in vitro* does not occur *in vivo* since during normal embryonic development the ES cell population is transient and rapidly gives rise to lineage-restricted progenies. It has been shown that pluripotent ICM cells and ESCs in culture, although broadly very similar, display different epigenetic silencing marks across key pluripotency regulators (O'Neill et al., 2006). These findings suggest that specific selection mechanisms may exist *in vitro* that favour growth in culture.

An integral part of the culture adaptation process, in the context of Baker and colleagues' definition, is the accumulation of chromosomal aberrations. Indeed there are numerous reports of such aberrations present in both human and mouse ESCs as well as iPSCs which are discussed in detail in the following sections (2.2.4 and 2.2.5). Chromosomal abnormalities may arise randomly upon cell division but specific recurring aberrations that have the ability to overtake the normal cell population premise a mechanism that confers selective advantage to the cells carrying them. This is more likely to occur through processes that regulate cell cycle control, resistance to apoptosis and differentiation (Herszfeld et al., 2006; Harrison et al., 2009). Some of the potential mechanisms are discussed in more detail in section 2.2.6.

2.2.4. Genomic Integrity of Mouse Pluripotent Cells

2.2.4.1. Mouse ES cells

Although mouse ESC lines have been extensively used since they were first established in 1981 (Evans and Kaufman, 1981; Martin, 1981), surprisingly few studies have thus

far systematically assessed their genomic integrity in culture (Liu et al., 1997; Longo et al., 1997; Sugawara et al., 2006; Rebuzzini et al., 2008a; Rebuzzini et al., 2008b; Liang et al., 2008). From these studies, only Liu et al. (1997) and Sugawara et al. (2006) reported the exact karyotype of a high number of cell lines (35 and 88 respectively).

Liu et al. (1997) examined the relationship between the growth rate of ESCs, their ability for germ line transmission and the karyotypic integrity of the cells. It has been observed that higher passage cell lines have hindered efficiency to contribute to the germ line (Nagy et al., 1993). Liu et al. (1997) confirmed that colonies that demonstrate high growth rate often coincide with the presence of trisomy 8 which also reduces the germ line transmission rate. Clones from three independently derived ESC lines were analysed using G-banding and trisomy 8 was found to spontaneously arise in all three lines and eventually over-grow the cell population after six passages in culture. Trisomy 11 was also identified in 3 out of the 22 abnormal clones as well as other chromosomal changes. Longo et al. (1997) showed that somatic and germ line contribution highly correlates with the percentage of euploid metaphases in an ES clone and becomes limited when this percentage drops below ~40% (at about passage number 20). The study of Longo et al. (1997) comprised of three independently derived ESC lines from which 32 different ESC clones were cytogenetically analysed with the G-banding method. Unfortunately, the authors did not provide detailed information of the chromosomal mapping of the reported aberrations with the exception of a single ES clone whose karyotype was 41, XY, -13, -13, +10, +11, +19.

An additional study from Guo and colleagues validated four different ES cell lines (E14, W9.5, 2A and KPA cells) at different passage points using multicolour FISH analysis (Guo et al., 2005). The authors examined 9-15 metaphases per cell line. Even though the E14.1 cells exhibited normal morphology and expression of ES markers (*Oct4*, *Nanog*), they had acquired a duplication of 14q (10/10 metaphases) and a trisomy 8 (3/10 metaphases). Extending their analysis to the three additional cell lines W9.5, 2A and KPA, Guo et al. (2005) also identified a deletion of 6q in the W9.5 cell line (10/10 metaphases) as well as a trisomy 8 (7/11 metaphases) and trisomy 14 (5/11 metaphases) while the 2A and KPA cell lines were karyotypically normal.

The study of Sugawara et al. (2006) is the most comprehensive cytogenetic study to date and includes the karyotypic validation of 540 mouse embryonic stem (ES) cell

lines obtained through the course of three years (2001 to 2004) from 20 different institutions in Japan. From the cell lines examined, 66.5% showed normal chromosome number while the rest displayed different modal chromosome numbers (assessed by fifty metaphases counting per cell line). The authors randomly selected 88 cell lines for in-depth karyotypic analysis by a modified Q-banding technique (Nesbitt and Francke, 1973) and found that 60.2% of them were normal. The chromosomal changes of the rest 35 abnormal ES cell lines were predominantly trisomy 8 (51.4%), trisomy 8 with loss of a sex chromosome (14.3%) or combination of trisomy 8 and 11 (11.4%). In total, trisomy 8 in combination with other aberrations was present in 88.6% of the cell lines, and trisomy 11 in combination with other aberrations was present in 17.1% of the cell lines. In addition, 25.7% of the ES cell lines examined demonstrated a loss of a sex chromosome. Finally, the overall percentage of normal karyotypes was similar between the two different mouse strains analysed (58.3% for strain 129 animals and 58.8% for C57BL/6J × CBA) although a bit higher in the case of strain C57BL/6J mice (72.7%).

Rebuzzini et al. (2008a) examined the chromosomal make-out of the UPV04 ES cell line (Neri et al., 2007) during different time-points in culture by C-banding (Sumner, 1972) or DAPI banding (Schnedl et al., 1980). The authors analysed metaphases from passages 6 to 34 (93 metaphases on average per passage) and found that the overall percentage of euploid metaphases was constant through the three month culture period (~55% to 65%). For the aneuploid metaphases, there was a broad spectrum of chromosomal abnormalities that varied across the different passages. Contrary to Liu et al. (1997) and Sugawara et al. (2006), Rebuzzini et al. (2008a) did not identify trisomy 8 as the predominant type of aneuploidy (in fact, only a single case of trisomy 8 was discovered). These results could indicate that random patterns of aneuploidy may arise frequently during culture in a dynamic fashion but only specific aneuploidies actually confer selective advantage to the carrying cells which are, as a result, able to out-grow the normal cell population in a selective environment.

Importantly, the studies that have been discussed in this section refrain from providing any mechanistic hypothesis on the occurrence and prevalence of the genomic changes that they report and adopt a rather descriptive approach of the type of patterns they observe and their effect mostly on the ability for germ line transmission of the aneuploid cells. It is essential, therefore, to take these initials observations forward by delineating the processes that are involved in the accumulation of the observed

genomic aberrations and the selective advantage that it is believed they confer to the cells.

To conclude, chromosomal instability of mouse ES cell lines appears to be a widespread problem. It has been suggested, however, that embryonic stem cells are less permissive to mutations and demonstrate a much lower spontaneous mutation rates than somatic cells (Cervantes et al., 2002). Cervantes and colleagues used the *Aprt* mouse model which is derived from a cross between the 129 strain mice homozygous for the targeted null *Aprt* allele and C3H mice with wild-type *Aprt* (Vrieling et al., 1999). By examining the *Aprt* loss of function in mESCs and MEFs, this study identified that in mESCs the most frequent spontaneous mutation resulting to *Aprt* loss is the loss of chromosome followed by reduplication resulting to uniparental disomy (UPD) and apparent euploidy. Intriguingly, the *Aprt* gene is found on mouse chromosome 8. This could suggest a possible mechanism for the high frequencies of chromosome 8 specific aberrations in mouse ESC studies. Cervantes and colleagues also hypothesised that the observed UPD is a result of a series of nondisjunction events, where the cells acquire a trisomic chromosome 8 and subsequently return to isodisomy. Consistent with this, they were able to identify trisomy 8 in the population of the heterozygous *Aprt* ESCs. While it can be argued that the observed rates of UPD are specific to chromosome 8, another study has found high rates of isodisomy also on chromosomes 2, 5, 10 and 17 in mESCs (Lefebvre et al., 2001).

2.2.4.2. Mouse iPS cells

Since the first report in 2006 that mouse somatic cells can be reprogrammed to an ESC-like state through the transfection by four transcriptional factors (*Oct4, Sox2, Klf4* and *c-Myc*) (Takahashi and Yamanaka, 2006), the iPSC field has been the focus of great interest and intensive research. In the years that followed this important finding, many scientists have contributed to the significant advance of the field by improving the derivation process or generating iPSC from different somatic tissues (also refer to (Boué et al., 2010) for a list of pivotal publications). However, there seems to be a surprising lack of studies that investigate the genomic integrity of mouse iPSCs. Even though genetic changes in mESC lines have been reported by several independent laboratories, as reviewed in the previous section, and similar analysis has been performed extensively for their human counterparts, there is only a single study to date

that investigates the genomic integrity of mouse iPSCs during reprogramming (Quinlan et al., 2011) and no study that explores the effects of prolonged propagation in culture.

Quinlan and colleagues used high-resolution whole genome paired-end DNA sequencing to investigate the presence of chromosomal structural variations (SVs) such as inversions, micro-duplications or deletions, complex rearrangements and retroelement transpositions, during reprogramming. Paired-end sequencing is a sequencing approach that provides information about the physical distance of two reads in the genome in addition to the actual sequence of the reads. The authors used MEFs from a mixed mouse strain (Pcdh21/Cre - Z/EG transgenic mouse lines originally derived from C57BL/6CrSlc and CD1 mice respectively)(Boland et al., 2009) to obtain three iPSCs lines and indentified SVs by comparing them to the C57BL/6J reference genome. Two of the three iPSC lines (iMZ-9 and iMZ-21) were derived from the same MEF donor cell after viral transduction while the third (iMZ-11) was obtained from a different cell. By performing *pooled multi-sample data* analysis, meaning by analysing simultaneously the reads from all three cell lines, the authors were able to identify germline SVs that were present in all cell lines and the MEFs as well as de novo SVs that were the result of cellular reprogramming and were not present in the parental MEFs. Intriguingly, the authors identified only one de novo SVs present in each iPSC cell lines that had probably arisen during reprogramming. Two of the SVs affected two genes (*Plxna4* and *Cssp1*) while the third was in a nongenic region. In addition, no retroelement transpositions could be identified in any of the lines. The cells were fully reprogrammed and could generate viable mice. In the study of Quinlan et al. (2011), the iPSCs analysed were of a very low passage, suggesting that the process of reprogramming itself is not detrimental to the genomic integrity and careful selection of iPSCs could assure a normal karyotype. Of note, the mixed strain used in this experiment is not representative of the majority of mESC and miPSC lines currently in use and more experimental data are needed to investigate the effect of genetic background in the predisposition of certain pluripotent stem cell lines to demonstrate a compromised genomic stability (Catalina et al., 2008).

2.2.5. Genomic integrity of Human Pluripotent Cells

2.2.5.1. Human ES cells

Rigorous karyotypic validation of human pluripotent cell lines during *in vitro* propagation is critical in order to assure their genomic integrity. Since the necessary prerequisites for normal mouse ESCs and iPSCs, namely contribution to chimaeras and germ line transmission, cannot be assessed in human ES and iPS cell population for obvious reasons, the only way to establish that the cells have not been transformed is through karyotypic analysis. As a result, there are numerous studies that report the presence of aneuploidy and structural chromosomal aberrations in human pluripotent cells.

After their initial derivation in 1998 (Thomson et al., 1998), studies have shown that hESCs can be cultured for prolonged periods of time retaining a normal karyotype (Amit et al., 2000; Reubinoff et al., 2000). However, in the recent years increasing evidence has demonstrated that this is not necessarily a global characteristic of all cell lines. In 2004, two independent studies were the first to report the presence of karyotypic abnormalities in human ESCs. Draper et al. (2004) investigated the chromosomal make-out of four different cell lines using FISH and G-banding: three of them were sublines of the H7 line (H7.S0, H7.S6 and H7.S9) and two of them from H14 line (H14.S9 and H14.S14). The H7.S0 display normal karyotype and differentiation capacity for several months. However, the two sublines, H7.S6 and H7.S9, which were recovered from frozen batches, both acquired whole or partial chromosome 17 amplifications after several months in culture. The H7.S6 line had a 46,XX, der(6)t(6;17)(q27;q1) karyotype after 6 months in culture (60 passages) with a translocation of 17q to 6q, resulting to a trisomic 17q. A percentage of the cells also displayed trisomy 12. The H7.S9 line acquired a trisomy 17 within four months in culture. A trisomy 17 was also observed in the H14.S9 subline. For the H14.S14 subline, although initially karyotypically normal, after 22 passages in culture 76% of cells acquired a trisomy 17 and after 39 passages the culture was relatively homogeneous in trisomic 17 cells (95%), suggesting a selective growth advantage of the specific abnormality. Interestingly, 12p and 17q aberrations are often observed in the typically aneuploid human EC cells suggesting parallel mechanisms in these two closely related cell types (Atkin and Baker, 1982; Rodriguez et al., 1993; Skotheim et al., 2002). The

second study was the one of Inzunza et al. (2004) which used FISH and CGH to assess the genomic integrity of three female lines, HS181, HS235 and HS237, at passages 35-39. All three lines seem to have a normal karyotype initially, but line HS237 showed an isodicentric X chromosome at passage 61. However, Buzzard et al. (2004) reported that persistent changes similar to the ones identified by Draper and colleagues were not present in the six hES cell lines (hES1–2, and hES3–6) of the study. The lines of Buzzard and colleagues had been frequently karyotypically analysed, for between 34 and 140 passages, but only sporadic aberrations that could not overtake the cell population had been observed.

Mitalipova et al. (2005) observed that the method of propagation may be highly connected with the occurrence of aneuploidy. For example, the hESC lines BG01 and BG02 maintained a normal karyotype for passages 41 to 105 when propagated by manual dissection of the hESC colonies. The authors then used two methods that can disaggregate hES cell colonies for bulk passaging, a nonenzymatic (cell dissociation buffer, CDB) and an enzymatic (collagenase/trypsin, CT). The change of propagation method resulted to the appearance of aneuploidy: the BG01 cell line developed trisomy 12 and 17 (and sporadic appearance of trisomy X) as soon as 23 passages with the CDB method, while the BG02 line developed trisomy 12, 14, 17 and an extra copy of the X chromosome (and sporadic appearance of trisomy 20), after 56 passages with the CT method. Mitalipova and colleagues hypothesised that the difference in propagation methodology adopted by Draper et al. (2004) and Buzzard et al. (2004), that is enzymatic versus manual, could be the reason that the lines in the study of Draper and colleagues were more susceptible to karyotypic abnormalities.

The first study using SNP-arrays to identify chromosomal aberrations in hESC was performed by Maitra and colleagues in nine pairs of early and late passage hESC lines (Maitra et al., 2005). The authors reported a range of subchromosomal aberrations, including a 17q amplification and a whole chromosome 13 deletion in two different lines. Imreh et al. (2006) also examined the HS181 cell line and found that after 11 passages the 80% of the examined cells have acquired an abnormal karyotype (47, XX, del(7) (q11.2), +i(12) (p10)), again implicating chromosome 12. In addition, these cells showed impaired differentiation capacity when assessed by teratoma formation in immunodeficient mice.

Another study investigated the process of culture adaptation by comparing early passage normal samples and later passage adapted cell from a single cell line, H7 (Enver et al., 2005). This study is particularly interesting because it examined the expression profiles of normal and adapted cells by further dividing the two types of cells into two subpopulations according to the expression of the surface marker SSEA3. SSEA3 is a marker of undifferentiated hESCs (Draper et al., 2002). The authors demonstrated that the karyotypic abnormalities of the H7 adapted hES cell line (46,XX, der(6)t(6;17)(q27;q1)) can be readily identified at the transcriptional level by using Affymetrix expression arrays. However, not all the genes in the affected region show altered gene expression levels nor the ones that are indeed up-regulated have the same fold-change in mRNA expression levels (i.e. the notch ligand, DLK1, showed a 7-fold up-regulation). In addition, the majority of the differentially expressed genes were within the aberration coordinates and only a few were found in other genomic loci. As revealed by hierarchical clustering, both the SSEA3+ and SSEA3- adapted cells clustered closely to the SSEA3+ normal cells suggesting a shift of the adapted cells towards self-renewal. In addition, the adapted cells showed an enhanced cloning efficiency when compared to the normal cells (6-fold higher for the SSEA3+ adapted cells and 3-fold higher for the SSEA3- adapted cells) and reduced differentiation as assessed by colony morphology.

Besides the recurring chromosome 12, 17 and X aberrations (Baker et al., 2007), other studies have revealed the occurrence of a 20q11.21 amplicon. Lefort et al. (2008) examined five hESC lines (SA01, H9, H1, VUB01 and VUB05-HD) that have been propagated by manual dissection, using BAC array-CGH, SNP arrays and FISH. They found that the 20q11.21 duplication arose as an inverted repeat during long-term culture in the SA01 (p83), H9 (p41) and VUB05-HD (p103) lines. In addition, an extra copy of the 20q11.21 was found inserted in the 1p36.3 region at the H1 cell line (p24, 24% of the cells and p64, 47% of the cells). Similar results were reported by Spits et al. (2008), where 17 hES cell lines were tested by array-CGH and five out of the 17 carried a 20q11.21 duplication after several passages in culture. The 20q11.21 amplicon was also confirmed using array-CGH analysis by Wu et al. (2008) and by Werbowetski-Ogilvie et al. (2009). The same region was identified by yet another study in the CCTL-14 hESC line (Narva et al., 2010). Using high resolution SNP arrays, this study reports a range of CNVs present in otherwise karyotypically normal hESC lines, implicating chromosomes 4, 5, 10, 15, 18 and 20 (for larger CNVs of ~1-3 Mb). Collectively, these

findings suggest that the amplification of the 20q11.21 locus confers selective advantage to the affected cells. Since the specific aberration is quite small in length (~1.5 Mb) the number of encompassing candidate genes is also limited. All the above studies have proposed that two conspicuous candidates in the specific region may be driving the selection: the *BCL2L1* gene, a member of the Bcl2 pro-survival family that inhibits apoptosis under some types of cytotoxic stress (Cory and Adams, 2002) and *ID1*, a member of the ID proteins family that regulate cell differentiation and cell cycle (Prabhu et al., 1997; Barone et al., 1994) and have been also found implicated in cancer (Cheng et al., 2011; Pillai et al., 2011).

Importantly, two recent large-scale integrated analysis studies have investigated the largest collections to-date of human ES and iPS cell lines (Mayshar et al., 2010; Amps et al., 2011). Mayshar and colleagues performed transcriptional karyotyping to study the impact of underlying aneuploidies at the transcriptional profile of the affected cells using gene expression microarrays in combination with validation by SNP arrays. This study is closely related to the methodology adopted here and it's discussed in more detail in section 4.1.2. Briefly, Mayshar et al. (2010) identified chromosomal aberration in 32% of the examined hESC lines, involving whole chromosomes or chromosome segments, and specifically implicated chromosomes 12 and 17 that were both observed with a frequency of 3 out of 38 lines.

Finally, the study of Amps and colleagues is the most comprehensive study to date, consisting of 125 independent hESC lines and 11 reference iPSC lines derived from 38 laboratories worldwide. These lines were examined in earlier and later passages using high-resolution SNP-arrays. The authors provide a comprehensive resource of karyotypic imbalances accompanied with structural variation of normal and adapted hESC lines, as well as methylation analysis.

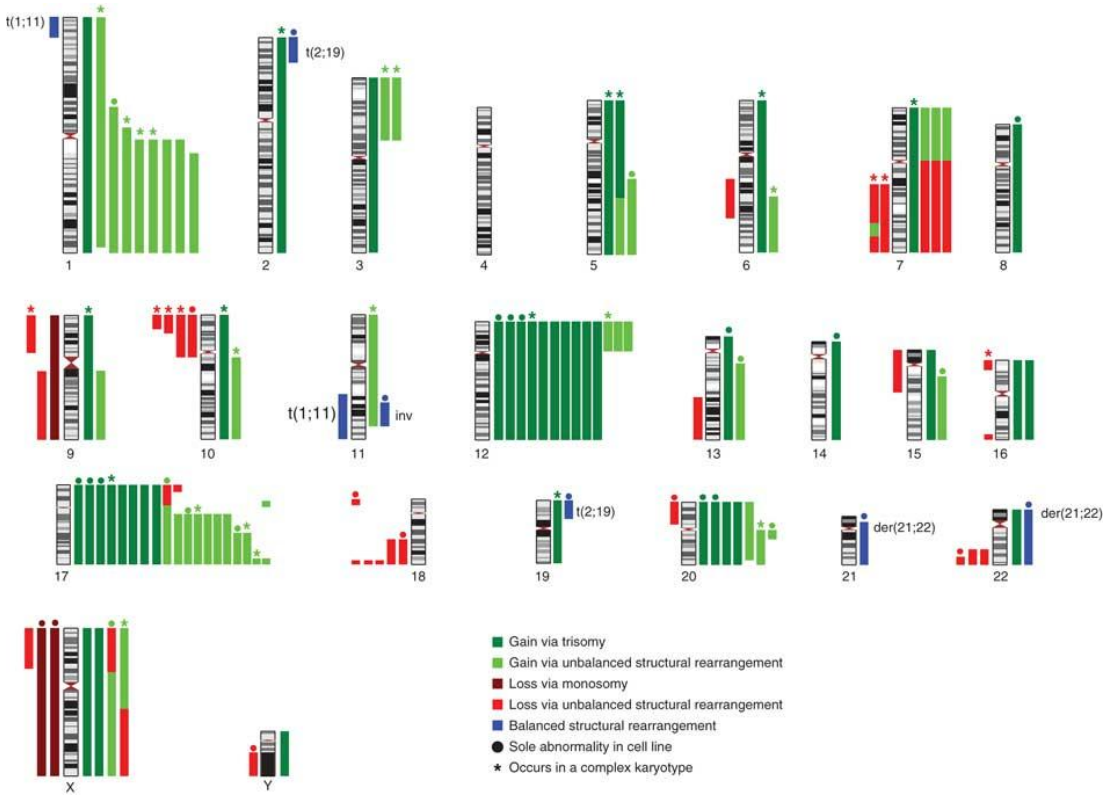

Figure 2.5 Genomic mapping of hESCs chromosomal aberrations.
Chromosomal mapping of identified aberrations in 125 hESC lines of which 42 were found abnormal specifically implicating chromosomes 1, 12, 17 and 20.
[Image from (Amps et al., 2011)]

Interestingly, from the 125 examined lines 42 (34%) had abnormal karyotypes. These included changes primarily in chromosomes 1, 12, 17 and 20 (almost half of the adapted lines), consistent with previous reports (Figure 2.5). In addition, late passage lines had a two-fold higher probability to carry chromosomal abnormalities. These changes were significantly more frequent in lines that have been enzymatically propagated for many passages. In the case of the 20q11.2 amplicon specifically, the pattern was present in 7 abnormal lines but more interesting, also in 22 of the karyotypically normal hESC lines as a structural variant. Among these 22 lines, five displayed the variant at an early passage and the rest 17 acquired it upon prolonged passaging. In addition, from the 17 lines that acquired the 20q11.21 amplicon in a later passage, three displayed extended karyotypic abnormality suggesting that the specific aberration may promote chromosomal instability. The authors identified the minimal overlapping region of the 20q11.21 gain to contain three genes expressed in hESCs. These were the *HM13*, *ID1* and *BCL2L1,* and at least *ID1* and *BCL2L1* have been repetitively reported in literature as putative candidate genes that drive the selection

behind the specific aberration (Maitra et al., 2005; Lefort et al., 2008; Spits et al., 2008).

2.2.5.2. Human iPS cells

Several recent reports have uncovered genetic and epigenetic changes in human iPSCs raising an alarm for the scientific community (Mayshar et al., 2010; Gore et al., 2011; Laurent et al., 2011; Hussein et al., 2011; Ben-David et al., 2011; Amps et al., 2011). Mayshar and colleagues used a large-scale gene expression meta-analysis approach to analyse the genomic integrity of 66 iPSC lines from independent laboratories. The authors found that approximately one fifth of the lines contained a large-scale aberration (whole or partial-chromosome spanning, > 10 Mb). Chromosome 12 was found trisomic or partially duplicated in four of the adapted lines and the overlapping aberrant region spanned the *GDF3-NANOG* locus. Both these genes also demonstrated significantly elevated expression levels in the abnormal cell lines. In addition, cells that carry the specific duplication rapidly overtook the cell population, as assessed by karyotyping of early and later passages, suggesting that the specific aberration confers selective advantage to the cells. The authors also proposed that the rapid selection for the chromosome 12 specific aberration may be due to the enrichment of the specific chromosome in cell-cycle genes that could potentially contribute to accelerated growth. Similar genetic changes was also identified by Amps et al. (2011) in three out of the 11 iPSC lines (27%) included in the study. Two of them carry a trisomy 12 and the third an inversion of chromosome 5.

In another study, Laurent et al. (2011) performed SNP analysis in a large collection of human pluripotent ESC and iPSC lines as well as non-pluripotent cell populations. The overall frequency of subchromosomal aberrations was significantly more frequent in pluripotent cell lines. Interestingly, chromosomal gains were more prevalent in hESCs while deletions in hiPSCs. Large chromosomal aberrations implicated chromosomes 12, 17, 20 and X as it has been previously reported. Furthermore, several lines carried small CNVs spanning pluripotency genes such as *NANOG* and *OCT4* as well as related pseudogenes whose function remains unknown. The study also linked extended passaging of hiPSCs with the presence of amplifications of oncogenes while the process of reprogramming was associated with deletions of tumour-suppressors (Laurent et al., 2011).

A separate study by Hussein and colleagues, using again high resolution SNP arrays, reported that a large number of CNVs appears de novo during reprogramming (two-fold higher frequencies than the levels present in the parental fibroblasts or ESCs). Perhaps unexpectedly, the majority of these changes are selected against during culture and later passage iPSC lines demonstrated lower frequencies of mutation. In addition, recurring deletions was mostly observed in common fragile sites (CFSs) suggesting that the novel CNVs may be the result of replicative stress during cellular reprogramming (Hussein et al., 2011).

Finally, Gore et al. (2011) performed exome sequencing on 22 human iPSC lines from seven different laboratories which have been derived using different reprogramming methodologies. The authors also used the parental fibroblasts to quantify the mutational load during reprogramming. They were able to validate 124 point mutations present in the iPSC lines but not the fibroblasts which averages to six coding mutations per iPSC line. The mutations were found to be enriched in cancer-related genes but no enrichment of specific pathways was present. Gore and colleagues suggest that the mutational load may not be the intrinsic effect of the reprogramming process but rather a result of the way colonies are picked and cells are propagated in culture.

2.2.6. Mechanisms of culture adaptation and tumorigenicity

Several studies have linked the events occurring during culture adaptation of pluripotent stem cells with tumorigenic transformation (Herszfeld et al., 2006; Baker et al., 2007; Harrison et al., 2007; Yang et al., 2008; Werbowetski-Ogilvie et al., 2009; Harrison et al., 2009; Blum and Benvenisty, 2009; Hovatta et al., 2010; Ben-David and Benvenisty, 2011). The “best” ESCs in culture, which most researchers would pick for propagation, are the cells that show higher proliferative rates, decreased differentiation and apoptosis and enhanced clonogenicity. The paradox is that the cells that demonstrate the above characteristics are also the ones that are most likely to carry karyotypic changes similar to the ones observed in tumorigenic transformation.

Recurring chromosomal aberrations hint towards an underlying selective mechanism rather than random acquisition. For hESCs in particular, frequent genomic changes are similar to changes observed in cancer cell types. For example, the common

chromosome 12, 17 and X amplifications in human pluripotent cells, discussed in the previous section, are also present in germ cell tumours and EC cells (Baker et al., 2007; Blum and Benvenisty, 2009). In addition, the recurrent 20q11.21 amplicon has been also reported in lung cancers and giant-cell tumour of bone (Beroukhim et al., 2010). When analysed by teratoma formation, adapted hESCs gave rise to tumours containing more primitive types of tissues and *OCT4*-positive cells in contrast to teratomas from normal hESCs that showed a more differentiated phenotype (Herszfeld et al., 2006; Yang et al., 2008). Finally, Herszfeld et al. (2006) observed that the expression of the CD30 surface marker, an EC-specific surface antigen, is restricted to the adapted hESCs and it can be used to distinguish them from normal hESCs. However, this finding was not confirmed by Harrison et al. (2009), suggesting that CD30 expression is not a universal characteristic of adapted hES cell lines and different mechanisms can contribute to culture adaptation.

Interestingly, initial findings suggest that ESCs demonstrate much lower spontaneous mutation rates than their somatic counterparts (10^{-6} in ESCs versus 10^{-4}, assessed by the number of spontaneous mutations at the *Aprt* reporter gene) (Cervantes et al., 2002). Since ESCs are endowed with the task to give rise to the whole developing embryo, a high mutation frequency cannot be tolerated. In fact, stringent mechanisms of DNA damage repair are present in order to assure their genomic integrity and cells with DNA damage are eliminated through apoptosis or differentiation (Tichy, 2011). For example, mESCs lack a G1 checkpoint following DNA damage and thus they are directly routed towards apoptosis (Hong and Stambrook, 2004). However, it has been found that the G2 decatenation checkpoint, which delays entry into mitosis if the chromosomes have not been sufficiently decatenated or disentangled, is error-prone in mESCs which could result to severe aneuploidy (Damelin et al., 2005).

Nonetheless, genomic aberrations occur in pluripotent stem cells and more importantly in a non-random fashion. For iPSCs particularly, there are three potential types of aberrations: (i) mutations in the parental somatic cells, (ii) mutations arising during the reprogramming process and (iii) genomic aberrations linked to prolonged culture propagation (Mayshar et al., 2010; Ben-David and Benvenisty, 2011). The sequencing of three mouse iPSC lines and their parental MEFs showed no evidence of high mutation rates during reprogramming (Quinlan et al., 2011). Studies in human iPSCs reported different types and rates of mutation associated with reprogramming (Laurent et al.,

2011; Gore et al., 2011; Hussein et al., 2011) but one of these studies proposed that the majority of these changes is actually selected against during culture (Hussein et al., 2011).

2.3. Transcriptomics

The *transcriptome* is the total of all RNA molecules present in a cell, including mRNAs, non-coding RNAs and small RNAs. The study of the transcriptional state of a cell is referred to as *transcriptomics* in contrast to the term *genomics* which describes the analysis of the genetic content of the cell. While the genome is relatively constant for the cells of a given organism, the transcriptome is widely varying from cell to cell depending on the specific type of tissue and the specific circumstances. The variations of the transcriptional state of a cell are depicted through different *gene expression* patterns. Gene expression is the process that begins from the transcription of a gene to produce a primary transcript (RNA molecule), continues with the processing of the primary transcript to obtain a mature RNA (mRNA) after intron removal (splicing) and finishes with the production of a functional protein.

Recent technological advances have generated a set of methods that can be used for the analysis and the quantification of the transcriptome. These technologies can be grossly divided into two categories: hybridisation- and sequencing-based approaches. Hybridisation-based techniques mainly rely on the use of microarrays. The application of microarray technology for the identification of genomic abnormalities has been already introduced in section 2.2.2. In the following sections a brief introduction of the gene expression microarrays and the RNA-seq technologies is provided. In addition, section 2.3.3 describes the application of transcriptomics analysis for the validation of the genomic integrity of a cell population, a concept that is fundamental to this study.

2.3.1. Gene Expression Microarrays

Gene expression microarrays can quantify the expression levels of thousands of genes of an organism in a single reaction. In a typical microarray experiment, mRNA molecules are hybridised to DNA template molecules which are fixed on a solid surface (glass slides, silicon chips or nylon membranes). The DNA template molecules, referred to as probes, can be cDNA or synthesised oligonucleotides of known sequences that complement the target transcript and known location on the chip and thus allow the identification of the expression levels of a gene whose mRNA product hybridises to a

specific probe (Schena et al., 1995; Schena et al., 1996; Lockhart et al., 1996). Typically the sample mRNA is reverse transcribed into the more stable cDNA and mRNA or cDNA is labelled by the incorporation of a fluorescent nucleotide. Hybridisation to the array is then performed and complementary sequences quantitatively bind depending on the abundance of each mRNA or cDNA molecule. The resulting fluorescent signal from each spot of the array gives a measure of the relative abundance of the target transcript in the test sample. These fluorescent signals are recorded by a laser scanner or CCD camera and computer image analysis outputs the final raw intensities for each transcript.

In cDNA microarrays, PCR-amplified DNAs (usually longer than 1 Kbp) are immobilised on the array chip. In oligonucleotides arrays the probes are shorter. Affymetrix GeneChip® arrays made by photolithography in situ include nucleotides of about 25 bp length (Lipshutz et al., 1999). In the Affymetrix GeneChip technology, multiple independent probes are designed to hybridise to different complementary sequences of the 3' end of each transcript in an attempt to minimise the signal-to-noise ratios and cross-hybridisation effects. Arrays based on this design are called 3' expression arrays and include some of the most popular arrays such as the Human U133 Plus 2.0 or the Mouse 430 2.0 array. In 3' expression arrays, probes are designed in pairs that differ only in one base pair in the middle of the sequence. The perfect match (PM) probes have a perfect complementary sequence to the target, while their partner mismatch (MM) control probes differ in the middle base of the sequence. Thus, the MM probes are used to control for background and cross-hybridisation noise. Recently, Affymetrix has released a new technology, the Exon arrays, that specifically interrogate gene expression at the exon level. Contrary to the traditional 3' expression arrays which use the MM probes to account for non-specific hybridisation, Exon arrays do not include MM probes but use instead a set of probes designed to measure cross-hybridisation noise due to pure background.

Another popular manufacturer, Agilent Technologies, produces arrays of longer oligonucleotides (~60 bp) manufactured in situ by use of ink jet printing technology (Hughes et al., 2001). Oligonucleotides arrays have the advantage over cDNA microarrays that they do not require the generation and management of clone libraries. In addition, it has been reported that the longer oligonucleotides provided in the Agilent platform give better sensitivity due to the larger surface available for

hybridisation (Hardiman, 2004).

Besides the actual abundance of a given transcript, several other factors can influence the intensity measurements from a microarray experiment. These include the sample preparation protocol, the array type, the sensitivity of the hybridisation process and the signal quantification (Hartemink et al., 2001). Because of the presence of this technical variation between different microarray samples, it is necessary to apply sample normalisation before performing data analysis. One of the most widely-used methods for Affymetrix GeneChip arrays normalisation is the Robust Multi-array Average (RMA) method described by Irizarry et al. (2003a). The RMA method includes three basic steps: an array-specific background adjustment, quantile normalization to ensure that the distribution of the expression values of each array in the comparison is the same and, finally, summarization to generate the normalised measurement for each probe (Bolstad et al., 2003; Irizarry et al., 2003b). The RMA method does not incorporate the PM/MM information which is available in the 3' expression arrays since it has been shown that subtraction of the MM signal from the PM signal can be problematic in the lower quantile of expression values due to noise (Irizarry et al., 2003b). One downside of the method is that it can be applied for the normalisation of samples from the same type of gene chip only. In the present study the RMA method has been used for the normalisation of samples from the Mouse 430 2.0 array. In addition, the PM/MM information has been used for the generation of the Present/Absent flags (further discussed in section 3.5) using the MAS5.0 algorithm (Hubbell et al., 2002).

The microarray methods are at the moment one of the most widely adopted tools for transcriptomics analysis due to their relative low cost and the high throughput they provide. Their popularity is depicted in the ever increasing number of publications to date using this technology. Accordingly, there is a wealth of readily available datasets deposited in public data repositories such as the NCBI's Gene Expression Omnibus (GEO) (Edgar et al., 2002) and the EBI's ArrayExpress (Brazma et al., 2003). Researchers can easily access this data in order to re-analyse or compare results and perform large-scale integrated bioinformatics analysis.

2.3.2. RNA-seq

The recent advances in the next-generation sequencing technology have also been used in the context of transcriptomics. Providing the highest resolution achieved to date, *RNA-seq* (RNA sequencing) can generate hundreds of millions of reads that can be analysed to provide an even more comprehensive image of the transcriptome. Briefly, the RNA-seq protocol consists of the collection of the sample RNA, which is subsequently used to generate a collection of cDNA fragments with adaptors ligated at one or both ends of the short sequence. The cDNA collection is typically amplified and high-throughput sequencing is performed to obtain reads from either the one end of each fragment (single-end sequencing) or from both ends (paired-end sequencing). Sophisticated algorithms are then used to perform quality control, align the reads to the reference transcriptome and identify differentially expressed transcripts or investigate other types of non-coding RNAs and microRNAs. Different sequencing technologies exist (i.e. Illumina (Solexa), Applied Biosystems, 454 Life Sciences) and each of them has specific limitations that render it more suitable for different types of applications. Nonetheless, RNA-seq presents several advantages over microarray technology including the wider range of expression it can capture (avoiding the caveat of decreased sensitivity in the lower and higher expression range that is present in microarrays) as well as the important ability to detect novel transcripts. A detailed discussion of the method, its advantages and limitations is provided by Wang et al. (2009).

2.3.3. Identification of Aberrant Genomic Regions at the Transcriptional Level

Concordant changes in gene expression levels across whole chromosomes or subchromosomal regions could be the result of underlying genomic or epigenetic alterations. Several studies from the field of cancer biology, where genomic aberrations are frequently observed, have demonstrated that the expression levels of a substantial number of genes in regions of chromosomal amplifications or deletions are altered in accordance with the DNA copy number of these regions (Pollack et al., 2002; Hyman et al., 2002; Kahlem et al., 2004; Myers et al., 2004; Schoch et al., 2006). For example, studying the correlation between gene expression levels and DNA copy numbers in

human breast tumours, Pollack et al. (2002) have shown that 62% of the genes in amplified regions have corresponding elevated levels of expression (at least 2 fold change) and a two-fold change in DNA copy numbers manifests as an average 1.5 fold change in gene expression levels. In addition, Mayshar et al. (2010) has successfully applied this approach in human pluripotent stem cells (discussed in detail in section 4.1.2).

Of note, the analysis of microarray data for the inference of regions of genomic abnormalities has several major benefits. First of all, there is a plethora of gene expression microarray data that is not accompanied with the equivalent genomic data. Ideally the analysis of genomic and transcriptional data should be combined; nonetheless, the mere amount of available transcriptional profiles provides a wealthy source of information that can be used to infer recurrent chromosomal abnormalities in large-scale meta-analysis approaches. Furthermore, it is to be expected that not all the genes in the affected genomic region, in fact only very few of them, will be functionally related to the mechanisms that confer selective advantage to the aneuploid cells. These key regulators need to be distinguished for genes in the affected region that also frequently have an aberrant copy number due to linkage. These genes would appear frequently altered in meta-analysis approaches, albeit as a result of their physical proximity with the key regulators rather than their functional consequences. The use of transcriptional data could thus potentially help in distinguishing the driver genes from the functionally unrelated bystanders, at least in the case where they do not demonstrate significant changes in their expression levels. Finally, this approach allows the simultaneous examination of the downstream effects of an aberration in a genome-wide scale by providing information about putative targets that are not necessarily located in the aberrant region.

The downside of the approach is the fact that it is not always clear if the observed pattern is a result of underlying genomic abnormality, or epigenetic alteration or even transcriptional regulation, especially in the case of genomic regions that only contain a small number of genes with altered gene expression. In addition, sophisticated bioinformatics analysis must be applied in order to infer the precise coordinates of the affected regions because of the noise introduced by genes whose expression is not necessarily changed by the presence of aneuploidy. For example, genes whose regulatory elements are not present, or are not generally expressed in the specific

tissue won't appear differentially expressed even if they are found in the amplified/deleted region. In addition, some genes may even appear differentially down-regulated while in an amplified region (or over-expressed while in a deleted region) if they demonstrate dramatically different expression levels in the control sample (i.e. dramatically over-expressed genes in a control without a gain may appear differentially down-regulated in the test sample carrying the gain due to the relative differences in the microarray measurements).

Several tools have been developed for the discovery of genomic clusters of differentially expressed genes that can be diagnostic of underlying aneuploidies. However, the use of these tools is not often reported in literature especially in the field of stem cell research. In section 3.1.2, a detailed presentation of the available methods and their current limitations is provided.

2.4. Machine Learning and Pattern Recognition in Gene Expression Data

2.4.1. Clustering Techniques

Cluster analysis consists of a set of unsupervised machine learning techniques that seek to group data that fall in distinct categories (clusters), with members within each group being similar to each other according to some scoring function and dissimilar from members of other groups (Hand et al., 2001).

Clusters analysis has been widely applied in biological studies where the researcher aims, for example, to group types of tissue samples based on the expression levels of n sets of genes. (Eisen et al., 1998) have demonstrated how cluster analysis can be used to group genes with similar expression profiles across a range of samples, visualise the results in an intuitive way and potentially identify functional categories related to specific clusters. Given that the original publication of Eisen et al. (1998) has been cited by ~11,400 authors (source: Google scholar), the popularity of clustering among the biology scientific community cannot be doubted. In some cases, however, it has been regarded with scepticism since clustering results of microarray experiments are often not reproducible, especially where the number of analysed samples is small (<50) (Garge et al., 2005).

Different clustering techniques have been developed thus far that can identify different kinds of clusters. Some of the widely used clustering methods are hierarchical clustering (Sokal and Michener, 1958), k-means clustering (MacQueen, 1967) and self-organising maps (SOMs) (Kohonen, 2001). Among them, the hierarchical clustering is one of the most (if not the most) widely adopted method for gene expression microarray analysis (Allison et al., 2006). Here, a brief description of the agglomerative hierarchical clustering approach is provided. Detailed description of other methods that can be used for clustering is not in the scope of this thesis.

Hierarchical clustering gradually merges similar points (*agglomerative*) or divides super-clusters (*divisive*). In addition, it offers an intuitive graphical representation of

the entire merging (or splitting) of the clusters in a tree-like plot, called the *dendrogram*. Clustering can be also simultaneously performed for both the samples in the study and the genes under measurement (Figure 2.6). In the agglomerative hierarchical clustering samples and, subsequently, clusters of samples are merged according to some *distance* measure. Initially, samples that have the smallest distance are merged and then each two clusters that are nearest are merged until the whole dataset is represented by a single top-level cluster (also discussed in section 4.3.2).

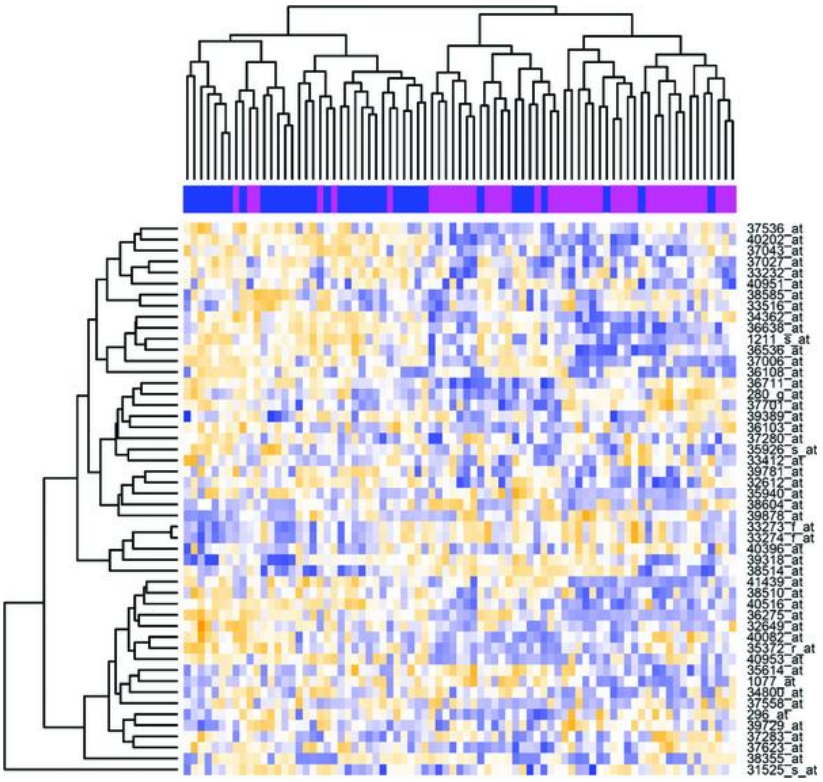

Figure 2.6 Heatmap of ALL data after filtering.
The rows of the heatmap correspond to genes and the columns to samples. Here, hierarchical clustering has been performed simultaneously to genes and to samples as depicted by the plotted dendrograms on the left and on the top of the heatmap. Class members are indicated by blue or purple under the samples dendrogram [figure from (Tarca et al., 2007)].

Common distance metrics for agglomerative hierarchical clustering include the:

- Euclidean distance, $\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$
- Manhattan distance, $\|a-b\|_1 = \sum_i |a_i - b_i|$
- Mahalanobis distance, $\sqrt{(a-b)^T S^{-1}(a-b)}$ where S is the covariance matrix, and
- Pearson correlation coefficient, $r_{a,b} = \frac{\sum_{i=1}^{N}(a_i-\bar{a})(b_i-\bar{b})}{(N-1)S_a S_b}$, which is the metric of

choice in the present study and it is presented in detail in section 4.3.2.

Euclidean and Pearson correlation-based distance are the most widely applied metrics in gene expression microarray analysis. There are however differences between the two metrics. Given three genes A, B and C, if we assume that the vector of expression values for gene A is closer to that of gene C but the expression of genes A and B changes in a similar way, then using the Euclidean distance will result to merging genes A and C, while using the Pearson correlation will result in merging genes A and B. Different distance metrics can, therefore, differentiate but there is not a consensus on which method is most appropriate for a specific dataset (Allison et al., 2006).

Another concept that arises in hierarchical clustering is the type of *linkage* to be applied for cluster merging. There are three typical variants to calculate the distance between two clusters while forming the dendrogram (Figure 2.7):

- *Average linkage*

  The distance between two clusters is the average/median pair-wise distance among all the members of the clusters.
- *Single linkage*

  The distance between two clusters is the smallest pair-wise distance between all the cluster members.
- *Complete linkage*

  The distance between two clusters is the maximum pair-wise distance between all the cluster members.

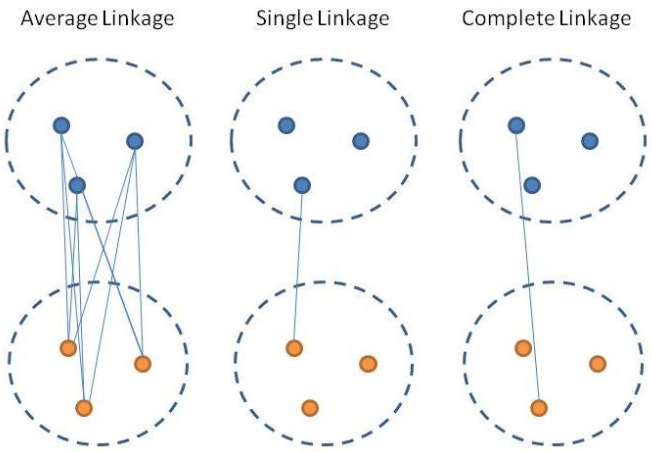

Figure 2.7 Three different types of clusters distance definitions used in hierarchical clustering.

2.4.2. Classification Techniques

Predictive models for classification are a set of supervised machine learning techniques that allow us to identify a mapping from a vector of measurements to some categorical value. The categorical value to be predicted is typically called the class of the data. A typical classification task consists of the creation of a prediction rule based on the measurements of a variable (attribute, feature) in samples with a known class label and the estimation of the performance of this rule in samples that has not been used to construct the rule (Hand et al., 2001). A classification problem can be perceived, therefore, as the construction of a class assignment function. This function will map the input space X to the categorical class space C. In *deterministic classification*, examined here, the sample is assigned to the correct class (contrary to the *probabilistic classification* where the sample is assigned to the most probable class). The class is known for the samples used in training the model (creating the prediction rule) and the class of the rest of the sample space needs to be estimated by applying the most accurate model.

The classification problem can be formulated as following: each sample is a pair (x_i, y_i) where x is a p-dimensional vector of the measured attributes (features), usually but not necessarily real-valued variables, and y is the class label of the sample, typically taken from a definite set of categorical labels $C = \{c_1, \dots, c_m\}$. The input of the classifier is a set of (x_i, y_i) pairs, namely the training set, which is used to construct the prediction rule. After the training phase is over, the classifier can be used to predict the class label of a new test vector x.

A closely related concept is the *decision boundary* (or decision region) of the prediction rule. If, for the sake of simplicity, we consider a classification model with only two input variables X_1 and X_2 and two output classes C_1 and C_2, the prediction rule will identify a piecewise constant surface in the (X_1,X_2) space that is partitioned according to C_1 and C_2. That is each point in the space will only belong to one of the two classes. So another way of defining the classification task is the learning of these decision boundaries (Hand et al., 2001). Given how different classification methods have different ways of identifying these boundaries, they can be less or more suitable for specific types of data.

The majority (if not all) of classification models have one or more *free parameters*. A free parameter is a variable of the model that can be used to tune the model towards a more accurate prediction. Ideally, models with the lowest number possible of free parameters that can still produce accurate predictions should be preferred. That is because in models with a small number of free parameters (equal or less than the attributes used as input), *over-fitting* is less likely to occur. The term over-fitting refers to the ability of a model to perfectly describe the training data but lacks *generalisation* capacity to data that have not been used in the training process. Of course, this is exactly the requirement of a good classification model: its ability to perform well in unseen data. In the case of datasets where the number of input features is so much greater than the number of available samples (which is normally the case in microarray gene expression experiments for example), over-fitting can be a common and limiting problem.

Classification has been widely used in many fields of biology and especially in microarray gene expression analysis. For microarray gene expression classification, each sample consists of thousands of measured features (gene transcripts) and belongs to a specific class (i.e. different tumour types) (Tarca et al., 2007). The great dimensionality p of the input vector x in the case of a gene expression microarray (that is the high number of input features measured in each microarray sample) and the low number of samples that is commonly available in a biological experiment makes the danger of over-fitting prominent in this type of analysis. In order to minimise this problem, it is required to select a small sub-set of features whose expression best correlates with the class of the sample and, therefore, is more likely to be informative during the creation of the prediction rule. This process is referred to as *feature selection*.

2.4.3. Feature selection for classification

Feature selection is commonly applied in order to identify a subset of features that can most efficiently distinguish between the two (or more) classes of the samples. Feature selection can be performed in two quite distinct ways: either by choosing the most appropriate features or by eliminating the most irrelevant features (Cristianini and

Shawe-Taylor, 2000).

Feature selection methods broadly fall into two categories: *filter* models and *wrapper* models. In filter models, feature selection is a pre-processing step, independent of the type of the classifier used. The computational cost of this approach is less but it can result to loss of information through the removal of features that could prove valuable for the creation of the prediction rule under a specific classification model. An example of feature selection through filtering was performed by Golub *et al.* (1999) where only genes with a high correlation with a specific cancer class were used as input. A repository of different filter methods for feature selection can be found in (Zhao et al., 2010). In wrapper approaches, the feature selection step is highly dependent on the classification model. Subsets of features are iteratively chosen and the accuracy of the classifier is evaluated. In this scenario, the classifier can be perceived as a "black box" that is being used to evaluate the predictive power of a specific feature subset (Guyon and Elisseeff, 2003). The set of features that results to the highest accuracy levels is subsequently chosen. The downside of these methods is that the identification of the optimal subset of features is an NP-hard problem: a problem that it is believed that it cannot be solved in polynomial time and, thus, it is extremely difficult to solve (Amaldi, 1998). A detailed discussion of different wrapper methods and definitions of the optimal feature set can be found in (Kohavi and John, 1997).

2.4.4. Validation in classification

There are different techniques that can be used in order to estimate the accuracy of a classifier and different measures that can be used to report it. From the methods that give an estimation of how well a classifier can generalise to new datasets, the following have been most commonly used in literature:

Hold-out method

The original dataset is split into two subsets: the training data and the test data. The training data are used to generate the prediction rule, while the test data are used to estimate the error rates of the classification. The downside of this method is that part of the information that can be used to train the classifier is lost as it is being used to test it.

In the case where the number of available samples is limited it may not be desirable to set aside part of the dataset for testing. It is also possible that the error rates obtain are specific to the way the split has been performed and the resulting test data.

Cross-validation

Cross-validation is another commonly applied technique. It can be divided in three categories: random subsampling, K-fold cross-validation and leave-one-out cross validation (LOOCV). In the random subsampling, a fixed number of samples are randomly selected to constitute the test data. The remaining data are used to train the classifier. This process is repeated N times and the average error rate obtained from each run represents the true error estimate. In the K-fold cross-validation, the $K-1$ first sub-samples are being selected for training in the first fold, the next $K-1$ sub-samples for the second fold and so on until all samples have been used for training. It is, in reality, very similar to random subsampling with the exception that the use of all samples as part of the test data is guaranteed. The true error rate is again calculated as the average of all error rates obtained at each fold. Finally, the LOOCV is a special case of K-fold cross-validation where $K = 1$. Typically, a good choice for K is $K = 10$. In the case of datasets where few samples are available LOOCV may be a better choice since the classifier is trained with as many examples as possible.

Bootstrapping

Bootstrapping is random resampling with replacement. This means that from a dataset of N samples, N samples are chosen with replacement for training and the remaining are used for testing. The process is repeated for K folds and the true error rate is again estimated by the average of the error rates obtained from each test subset.

Three-way data partitioning

In the three-way data partitioning method, data are divided in three subsets: the training, the validation and the test group. The training data is used to train the classifier and the validation data is used to choose the classifier with the lowest error rate. Next, the test data is used to report the true error rate of the model. In this way,

the true error rate is less biased since it has been calculated on data that have not been used in any way for the selection of the best-scoring classifier. Given that depending on the availability of samples may not be affordable to limit the number of samples used for the training of the classifier, another approach is to merge the training and validation datasets after choosing the best performing classifier and re-train the classifier with the merged dataset. The true error rate of the test dataset is then reported. The latter is the approach adopted for the training and validation of the SVM classifiers discussed in section 5.3.3.

Finally, as mentioned in the beginning of this section different measures can be used to perform error estimation. Most commonly in literature, the F-score and the Area Under the ROC Curve (AUC) have been applied. The AUC measure is used in the case of cross-validation experiments. In the present study, the F-score and the accuracy measures have been used to describe the error rates of the classification models. These are defined as following:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \qquad Precision = \frac{TP}{TP + FP}, \qquad Recall = \frac{TP}{TP + FN}$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Where:

$TP = True\ Positive, TN = True\ Negative,$

$FP = False\ Positive, FN = False\ Negative$

F1 score gives a balanced estimation between *Precision* and *Recall* and it is the most popular metric in cases of unbalanced classes (when instances of the one class are far more abundant than instances of the other). Accuracy, on the other hand, tends to under-estimate the ability of the classifier to predict smaller classes (Forman and Scholz, 2010). In the present study, both metrics are reported.

2.4.5. Some considerations for the classification of microarray data

In a typical microarray experiment replicate samples are collected from each condition. The common practice is to represent each replicate group with a single averaged instance. Since the number of available samples can be limited this results to a loss of important information that could have been used to better train the classification model. An alternative to this problem has been proposed by Karpievitch et al. (2009) with the use of *subject-level bootstrapping* in Random Forest (RF) classification. In any case, it is evident that the expression profiles of replicate samples are highly correlated and, thus, including replicate samples from the same study in both the training and the test datasets will introduce significant bias. It is a prerequisite, therefore, to partition the data in such a way that samples from the same study belong to only one of the partitioned groups, namely the training, the validation or the test data.

An additional consideration, highlighted in the study of Simon et al. (2003), is that the data used for testing and estimating the error rate of the classification model should not be included in the process of feature selection. For example, when cross-validation is used for the estimation of the error rates, feature selection has to be performed anew for each fold. If this important consideration is overlooked the estimated error-rates might be highly biased due to over-fitting to the specific dataset. Detailed description of the *selection bias* problem can be also found in the work of (Ambroise and McLachlan, 2002).

2.5. Research Objectives

The recurrence of specific abnormalities in pluripotent stem cells and their potential link to mechanisms that promote selective growth in culture raises three important questions: (i) how the presence of such aberrant patterns can hinder the validity and reproducibility of experimental results, (ii) what is the transcriptional footprint of these genomic changes and (iii) what information can we extract from these signatures to get insights into the normal biological functions of stem cell biology.

In order to address the above questions, this study will be focusing on:

- The development of sensitive bioinformatics tools that can be used for the identification of chromosomal aberrations from gene expression profiling data (i) at a single experiment level (DI.S.C.O. application) and (ii) in a large-scale integrated type of analysis,
- The application of the proposed methodologies for the assessment of a large collection of published transcriptional profiles from mouse pluripotent stem cell studies in order to quantify the overall frequency of aneuploidy and the recurrence of specific patterns,
- The identification of transcriptional signatures that are linked to aneuploidy, and, finally,
- The exploration of different classification models that can predict underlying aneuploidies using only a small number of diagnostic genes.

3. DI.S.C.O.: A Genomic View of Transcription

Figure 3.1 The DI.S.C.O. logo

This chapter introduces a novel computational tool for the analysis of gene expression data in the context of genomic position. DI.S.C.O. (*DIscovery of Subtle Clustered Organization*) can be used to identify regions of non-random transcriptional activity in the genome by a combination of visualisation and statistical techniques. The DI.S.C.O. application, an example dataset and the application's user manual is provided as part of the supplemental material in the attached CD.

3.1. Introduction

3.1.1. Background

As it has been highlighted by many independent studies (see section 2.2), the accumulation of chromosomal abnormalities in human and mouse pluripotent stem cells upon prolonged propagation in culture is a frequent phenomenon. These findings stress the need for karyotypic testing of pluripotent cell lines which is most commonly performed by chromosome counting, a quick and inexpensive technique that lacks however the ability to capture subtle chromosomal abnormalities such as partial chromosome gains or losses or complex events such as translocations. G-banding or chromosome painting by fluorescence *in situ* hybridization (FISH) or microarray-based

techniques, such as array comparative genomic hybridization (aCGH) and single nucleotide polymorphism (SNP) arrays, have higher resolution but they are not routinely performed since their cost becomes prohibitive when used for continuous testing of many cell lines over many passages.

Importantly, it is also possible to identify chromosomal imbalances by studying the genomic localization of gene expression levels. Chromosomal intervals of non-random transcriptional activity can be diagnostic of underlying aneuploidies which concordantly affect the expression patterns of the genes in the aberrant region. This approach has been already applied by two recent studies in order to identify patterns of chromosomal aberrations in human hESCs, hiPSCs and other multipotent cell types (Mayshar et al., 2010; Ben-David et al., 2011).

In the field of cancer biology, where genomic instability is prevalent, some of the developed methods specifically examine the relationship between gene expression levels and DNA copy numbers, using paired datasets, and have shown the high correlation between these two types of genomic data in the affected regions. These studies have examined a variety of disease and tumour profiles varying from breast cancers (Pollack et al., 2002; Hyman et al., 2002; Myers et al., 2004; Buness et al., 2007), head and neck squamous cell carcinoma (Masayesva et al., 2004), renal cell carcinomas (Cifola et al., 2008), trisomy 21 in Down syndrome patients and its equivalent mouse model (Kahlem et al., 2004; Lyle et al., 2004) or acute myeloid leukaemia datasets (Schoch et al., 2006; Hertzberg et al., 2007) among others. Alternatively, the emergence of chromosomal patterns of aberrant transcription could also be related to the perturbation of normal epigenetic regulation, such as the failure of XCI or even large region epigenetic silencing, present in cancer (Stransky et al., 2006; Frigola et al., 2006).

Given the relative abundance of gene expression microarray data compared to aCGH and SNP data, assessing genomic integrity at the transcriptional level can provide an easy and efficient way to validate pluripotent cell lines and identify problematic genomic regions which could potentially otherwise passed unnoticed and, therefore, confound the analysis of the experimental results. Furthermore, it can lead to the identification of candidate genes that have a causal role in the emergence and recurrence of specific aneuploidy patterns.

A number of methods have been thus far made available for this type of analysis, typically applied in the field of cancer biology (Toedling et al., 2005; Levin et al., 2005; Callegaro et al., 2006; De Preter et al., 2008; Nilsson et al., 2008). Even though the developed methods could also be used in a broader spectrum of applications, such as the karyotypic validation of pluripotent stem cells, their use has not been widely adopted by the biological community, as shown by the small percentage of published stem cell studies that include this type of test. This could be potentially due to a number of limitations in the currently available tools that render them inaccessible to biologists. For example, algorithms or scripts that require previous familiarization with specific statistical environments or programming experience can be limiting to the non-expert user. In the following sections, a detailed analysis of the related methods and their limitations is presented and how each one of them was dealt with during the development of DI.S.C.O.

It should be also noted that the usefulness of the approach is not necessarily limited to the identification of chromosomal abnormalities. Other applications include the discovery of a broad range of general structural genomic features such as the non-random genomic organisation in the eukaryotes (Michalak, 2008), the regional clustering of "housekeeping" genes (Lercher et al., 2002), of genes belonging to the same functional pathways (Lee and Sonnhammer, 2003) and of genes with spatiotemporal expression patterns in specific tissues or organs (Yamashita et al., 2004). Finally, concordant changes in the transcriptional levels of genes within a genomic region have also been identified around transgenic insertions within cell lines (Valor and Grant, 2007).

3.1.2. Related Tools and Limitations

The computational tools currently available for the identification of genomic clusters of differentially expressed genes can be roughly divided in three main categories:

(1) segmentation/ clustering algorithms or scripts,
(2) visualisation tools with no computational detection integration,
(3) visualisation tools with integrated computational detection.

It is immediately evident that the first two categories both suffer from the lack of either the visualisation or the computational component. The importance of the graphical representation cannot be overstated and it bears a great significance for the user since it gives a quick and intuitive first level overview of the data that is complementary to the algorithmic analysis. On the other hand, the acceptance or rejection of any scientific hypothesis and the reproducibility of the results should be based on proper statistical significance which can only be measured with the generation of appropriate p-values or defined computational scores. Thus, any tool that fails to incorporate both these aspects of the analysis immediately limits its general usability.

In Table 3.1, a detailed list of related visualization tools and/or algorithms for the identification of regional enrichment is presented. Highlighted are aspects of each tool that can restrict broader application. For example, implementations that are available through R (Ihaka and Gentleman, 1996) require a specialised knowledge of the R statistical environment which can be challenging for the non-specialised biologist user (such as the R scripts E-CHaRGE, MACAT, LAP and PREDA in Table 3.1). In addition, complex algorithms with a high number of parameters that need to be adjusted for individual datasets require an in-depth understanding both of the algorithm itself and the type of patterns that might be present in the dataset, which is not, in the majority of cases, *a priori* known (such as the TV minimization method) (Nilsson et al., 2008).

Another potential drawback under specific experimental hypothesis is the type of data that the method can process. Tools that typically accept only gene lists could potentially pose a limitation to the detection power of the methodology since they heavily rely on the method and the assumptions used for the gene list generation and do not reflect the continuous change of the transcriptional levels across the chromosomes (examples of such applications are the tools REEF, PGE and CROC in Table 3.1).

Another issue of importance is the capacity of the method to process high-throughput sequencing data such as RNA-seq data. Since next generation sequencing technologies are currently becoming increasingly popular and a large number of datasets is being constantly generated, there is an immediate need for tools that are able to examine data derived from this type of technologies. Where indicated, the methods presented in Table 3.1 could be potentially adjusted for RNA-seq data, relatively effortlessly by the

user without a need to adjust the algorithm itself.

As it is evident in Table 3.1 there are two popular approaches for the identification of regional enrichment: (i) the use of sliding genomic window of either fixed (i.e. the ChroCoLoc, REEF, CROC and TRAM methods), or adaptive size (PGE method), and (ii) the use of a kernel regression smoother (i.e. the E-CHaRGE, MACAT, LAP and PREDA scripts).

Almost all the methods of the first type of approach (genomic window) measure the enrichment within the window boundaries with the hypergeometric distribution. The hypergeometric distribution test gives the probability of having as many enriched genes in a specific window size and is defined as:

$$P(X = x > q) = \sum_{x=q}^{m} \frac{\binom{m}{x}\binom{t-m}{k-x}}{\binom{t}{k}}$$

where t is the total number of genes, k the genes in the specific window, q the number of differentially expressed genes within k and m the total number of differentially expressed genes in t.

In addition, the methods that use the sliding window-hypergeometric distribution strategy either test enrichment only in predefined genomic intervals such as the cytobands (ChroCoLoc) or require from the user to define the size of the window and the offset between subsequent windows (REEF, CROC, TRAM). The PGE method uses an improved concept of an adaptive size sliding window which overcomes the need for user-defined window properties (also see section 3.7.3 for detailed description).

DI.S.C.O. is addressing the limitation present in current available tools by providing a powerful and intuitive graphical user interface and a platform for the analysis of transcriptional data in terms of regional enrichment. DI.S.C.O. is particularly designed to facilitate the non-expert user. It includes three different computational methods and provides a framework for the integration and comparison of additional methods in the future. Furthermore, it can be easily used with next-generation sequencing RNA-seq data and has been integrated to the workflow engine GeneProf (Halbritter et al., 2012).

Table 3.1: Comparison to related tools.
List of relevant software applications, visualization tools or algorithms/ scripts that can be used for the identification of regional enrichment of differentially expressed genes and distinguishing features.

Tool	Graphical Output	GUI	Statistical Test or Algorithm	Data input	NextGen applicable	Part of complete analysis engine	Platform	Publication
Caryoscope	Yes	Yes	N/A	GEP[1]	No	No	Java	(Awad et al., 2004)
E-CHaRGE	Unknown	No	Compound Poisson Process model-based scan statistic	GEP	No	No	Requires R	(Levin et al., 2005)
MACAT	Yes	No	Kernel functions	GEP	No	No	Requires R	(Toedling et al., 2005)
ChroCoLoc	Yes	Yes	Hypergeometric Cytobands	GEP	No	Yes (Expression Profiler, EBI)	Web-based (Expression Profiler, EBI)	(Blake et al., 2006)
LAP	Yes	No	Local variable bandwidth kernel estimators	GEP	No	No	Requires R	(Callegaro et al., 2006)
REEF	Yes	Yes	Hypergeometric Fixed size sliding window	Gene lists	Yes	No	Requires Python	(Coppe et al., 2006)
PGE	Yes	Yes	Hypergeometric Adaptive size sliding window	Gene lists	Yes	No	Web-based	(De Preter et al., 2008)
TV	Yes	No	Total variation minimization	GEP	Yes	No	Standalone	(Nilsson et al., 2008)

[1] Gene Expression Profiling

Tool	Graphical Output	GUI	Statistical Test or Algorithm	Data input	NextGen applicable	Part of complete analysis engine	Platform	Publication
minimization			(segmentation method)					
GeneSpring GX	Yes	Yes	N/A	GEP	Yes	N/A	Standalone	Agilent Technologies
CROC	Yes	Yes	Hypergeometric Fixed size sliding window	Gene lists	Yes	No	Web-based	(Pignatelli et al., 2009)
TRAM	Yes	Yes	Hypergeometric Fixed size sliding window	GEP	No	Yes	Windows, Macintosh	(Lenzi et al., 2011)
PREDA	Yes	No	Non-linear kernel regression smoothing with adaptive bandwidth – Improvement of LAP method	GEP	No	Yes	Requires R	(Ferrari et al., 2011)
DI.S.C.O.	**Yes**	**Yes**	**Includes PGE, TV minimization and Nearest Neighbour (clustering approach) methods**	**GEP**	**Yes**	**Yes (also in GeneProf)**	**Standalone (Java Web Start)**	**Skylaki *et al.*, In preparation**

3.1.3. Previous Work

The fundamental concepts upon which DI.S.C.O. was based were first defined during my MSc Thesis in Bioinformatics (Skylaki, 2007). The aim of the project then was to develop a software tool that would plot gene expression data from microarray experiments as a function of genomic position. It was greatly focused on the visualisation aspect and aimed to overcome the main drawbacks of the tools that were available at the time. Even though the field has progressed a lot since then, and new methods have emerged, many of the limitations identified at that stage are still prominent as previously mentioned.

This initial stage of development (hereafter called DI.S.C.O. v1.0) resulted to the creation of a prototype application. By the end of the sort time available for the completion of an MSc thesis, however, the essential future steps for the transition from a prototype application to a complete computational solution became clear. The current version of the DI.S.C.O. application, presented here, has indeed moved a long way forward in terms of enhanced visualisation functionalities, incorporating multiple data analysis options and finally, including statistical detection methods. During this process, it has given rise to three more complementary MSc projects whose contributions would be also stated where appropriate (Sarantidis, 2008; Sveinbojornsson, 2010; Matzavinos, 2011).

The following sections present the fundamental concepts and functionality in the current DI.S.C.O. version. In the interest of facilitating the reader, a brief description of the original implementation, DI.S.C.O. v1.0, is given in section 3.10, after the introduction of the necessary basic principles of the application.

3.2. System Requirements

The following sections list the requirements that must be met by the DI.S.C.O. application during the development phase.

3.2.1. Technical requirements

The programming language chosen for the implementation of the application must be supported by all widely used operating systems in order to facilitate the development phase and maximise the effectiveness of the tool. The tool must be platform agnostic and should require only the minimal effort from the user to install and run. In addition, DI.S.C.O. should be a stable and reliable application with reasonable memory requirements and quick response time.

3.2.2. Visualisation

The visualisation component must offer an intuitive and powerful representation of the data which shall be on its own a valuable aspect of the data analysis process. It is essential that the initial plotting of the data must be automated and the optimal parameters for the visualisation must be data-driven in a sample-specific manner. This will enhance the pattern discovery ability of the user and establish the optimal parameters for effective visualisation at least at the very first stage. Subsequently, the user should be encouraged to explore different visualisation options and tune the visualisation parameters at will.

3.2.3. Detection Power

There are certain significant criteria that need to be met to ensure the high detection power of the integrated statistical methods:

Detection performance is the accurate identification of clusters of differentially expressed genes when they are present in the genome.
Response time is the required processing time for the identification of the regions of

interest which is of particular importance in the case of large data collections.

3.2.4. Usability

The end users of the application are presumed to be mainly biologists from the academic community with no expert knowledge of computational or statistical analysis. Therefore, DI.S.C.O. must be freely accessible through the internet and should not dependent upon the availability of specialised resources. The installation and set-up procedure must be straight forward and the user's ability to interact with the tool should require no previous familiarisation with any programming or statistical analysis interface.

The GUI component has to be clear and intuitive and it should facilitate the navigation through different options and features. This can be achieved by the use of comprehensive and contextual menus and actions as well as the integration of detailed documentation and help files. At any given step, the analysis must be transparent and the underlying data structures accessible to the user for further examination and exporting. In addition, the user must have easy access to summary tables for the different analysis steps performed and the type and attributes of the data processed.

Finally, the statistical analysis should require the minimum amount of user configuration with initial suggestion of optimal parameters that the user can subsequently choose to adjust if desired.

3.2.5. Extended Functionality for the Analysis of Gene Expression Data

The objective of the application presented here is to implement a genomic visualisation of the transcriptome while providing computational tools for the discovery of statistically significant chromosomal regions of aberrant gene expression levels. Along with the realisation of this goal, it is also recommended to enhance the user experience by supplying extensive functionality for the analysis of gene expression data in the context of genomic position.

Among the desirable features is the ability to perform basic data manipulation tasks

commonly applied during gene expression analysis. This includes data transformation and normalisation techniques, identification of differentially expressed genes, generation of gene lists with specific characteristics, generation of concise reports after the completion of the data analysis or data exports in different formats, both tabular and image, and finally, direction and representation of the results in independent genomic browsers of choice. Furthermore, the user should be able to import custom data tracks for the visualisation of different data types, such as array-CGH or SNP data, in combination with gene expression data.

3.3. System Architecture & Implementation

3.3.1. High-level System Architecture

An overview of the high-level system architecture can be found in Figure 3.2. A breakdown of the separate components is as follows:

Data Input Required and optional data files for the initialisation of the analysis. Further discussion in section 3.4.

Data Processing The core processing machinery of DI.S.C.O. The principal data analysis tasks and additional functionalities are presented in section 3.5.

Data Plotting The main visualisation component of DI.S.C.O., namely the **Genome Canvas**. Different options to improve the visualisation quality. Detailed discussion can be found in section 3.6.

Cluster Analysis This component consists of the automatic calculation of the appropriate thresholds for the identification of differentially expressed genes as well as the three statistical methods integrated in DI.S.C.O. It will be further discussed in section 3.7.

Data Output Export of the results and saving of the data structures used in the application.

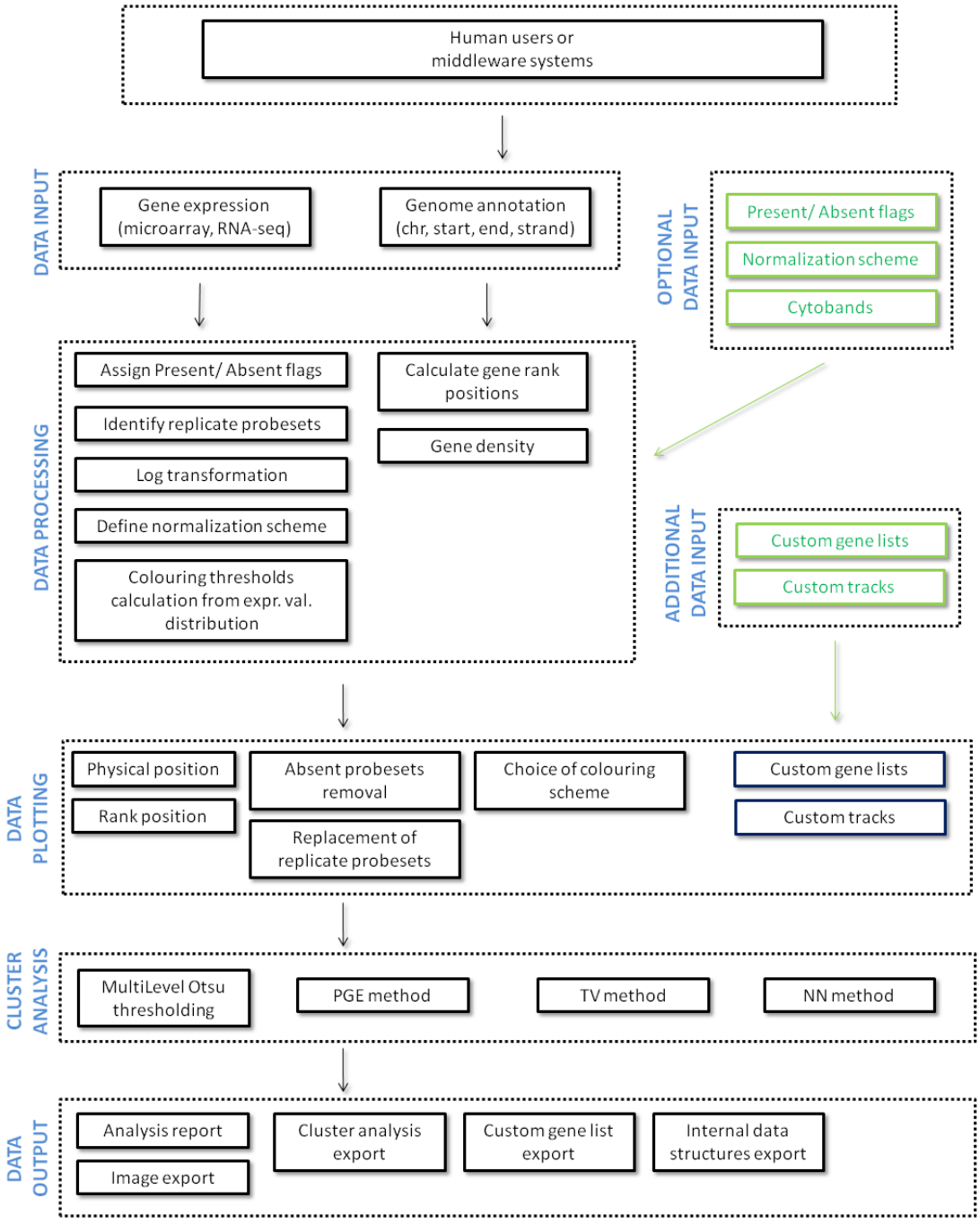

Figure 3.2 High-level system architecture overview.

3.3.2. Implementation

DI.S.C.O. has been implemented as a Java Web Start application in order to allow easy (cross-platform) access online, using Java 2 SDK standard edition version 1.6.0. Java Web Start can be used by any client system that supports the Java 2 platform and virtually all browsers. Java Web Start can be run on Windows 98/NT/2000/ME/XP, Linux, and the Solaris™ Operating Environment. Macintosh provides a version for their OS X release[2]. The current DI.S.C.O. implementation has been tested on Windows XP SP3 and Linux (CentOS release 5.7).

There are a total of 124 classes which are organized in 15 packages as shown in Figure 3.3. A detailed description of each class relationships, operations, attributes, and interfaces will not be attempted here in the interest of brevity. The functionality performed through the implementation will become however clear in the following sections. A brief explanation of each package and corresponding components is:

Disco.Reader contains the classes that read and validate the input files either of the user or of a middleware component. It performs error checking and ensures the consistency between different data types.

Disco.Organizer contains the classes that organise the input data into a hierarchical logic that follows the order chromosome, genome, experiment, normalisation and viewer.

Disco.Viewer is responsible for all the visualisation operations.

Disco.Analyzer consists of classes that perform the main analytical tasks. The calculation of Present/Absent flags, the replacement of replicate probe sets, the gene density estimation and the normalisation of the data are executed here.

Disco.Plotter contains classes responsible for the creations of supplementary plots and graphs.

Disco.Writer is responsible for all data exporting operations.

[2] Source: http://docs.oracle.com/javase/tutorial/deployment/webstart/

Disco.Tools	contains various help classes utilised in specific tasks.
DISCOhelp	implements the DI.S.C.O. help system.
Clustering.common	contains the classes that are commonly used from all clustering methods.
Clustering.stats	includes all the necessary statistical operations for the clustering methods.
Clustering.PGE/NN/TV	implements each one of the three clustering methods.

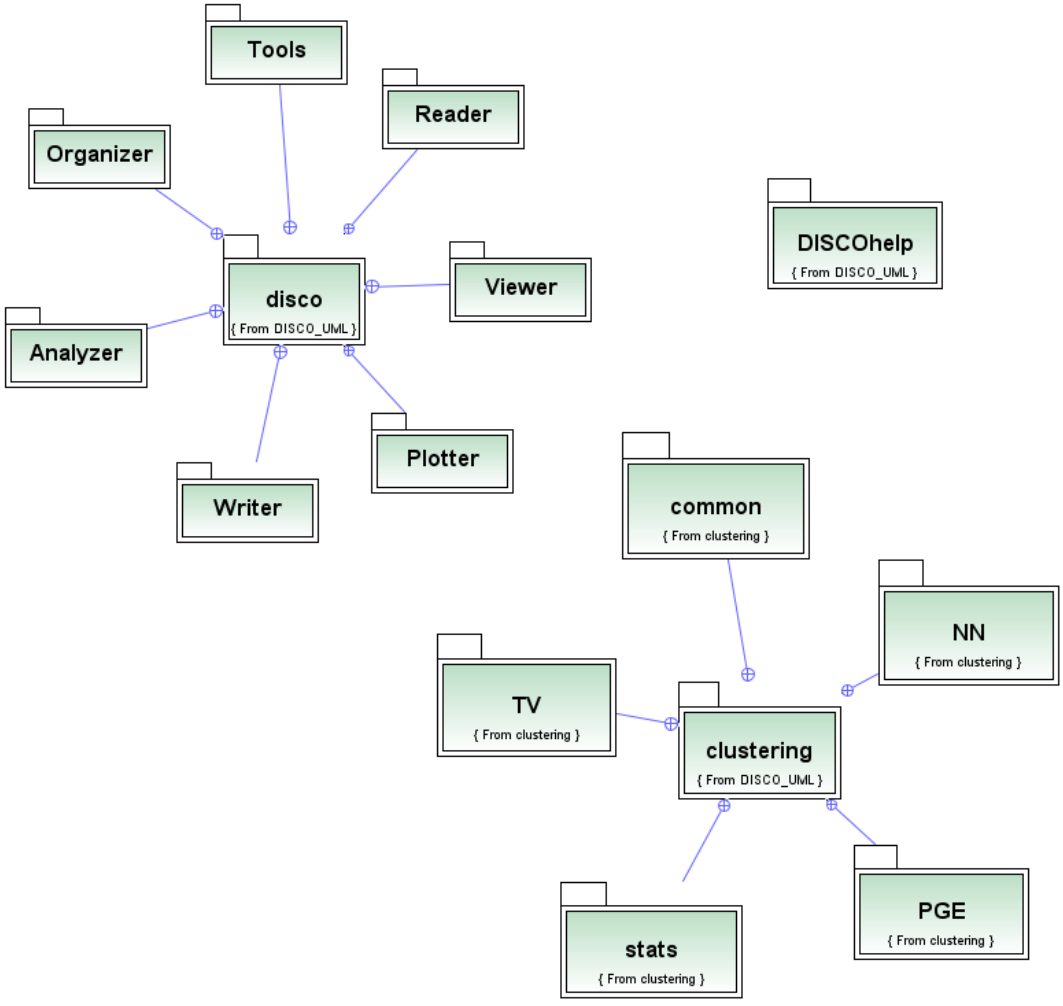

Figure 3.3 The DI.S.C.O. application package architecture

3.4. Data Input

The data files supported by DI.S.C.O are tab delimited text files. There are five data files that can be imported at the initiation of the DI.S.C.O. analysis: (i) the genome annotation file, (ii) the gene expression file, (iii) the Present/Absent flags file, (iv) the cytobands file and finally (v) the normalisation scheme file. The genome annotation and the gene expression files are essential for the analysis while the cytoband, the Present/Absent flags and the normalisation scheme files are optional. The content and format of each of these files is described as follows:

The Genome Annotation file

The genome annotation file provides information about the genes that are included in the analysis and their physical mapping on the genome. It consists of three columns: the transcript identifier, the common name of the corresponding gene and the mapping position in the genome and as many rows as the genes included in the experiment. The transcript identifier can be i.e. the probe set ID provided in Affymetrix microarray chips, the Illumina Probe ID in Illumina BeadChip arrays or any custom identifier (key) defined by the user as long as it is unique for each row.

When two (or more) mapping positions are assigned to the same transcript, only the first one will be considered for plotting. In addition, transcripts not assigned to a specific chromosome will be excluded from the analysis.

The Gene Expression file

The gene expression file provides the information about the expression levels of each gene in each sample included in the analysis. There is an 1-to-1 correspondence between each row of the gene expression file and the genome annotation file. The first column represents the transcript identifier while any subsequent column (as many as the samples included in the analysis) contains the normalised gene expression values for each transcript in each sample. The transcript identifier is matched between the gene expression file and the genome annotation file. When the two files are largely inconsistent with each other the user is prompted to correct the provided annotation.

The P/A flags file

This file is particularly relevant for the analysis of Affymetrix microarrays. As it has been briefly mentioned in section 2.3.1, the MAS5.0 algorithm (Hubbell et al., 2002) uses the measurements from the pairs of PM and MM probes in Affymetrix GeneChips to calculate a p-value reflecting the confidence that a gene is present in a sample. It then attributes a flag (or call) for each gene that can either be Present (P), Absent (A) or Marginal (M) (Liu et al., 2002). A probe set with an Absent flag has an expression intensity near to zero which can be due to the absence of expression of the corresponding transcript or problematic hybridisation. If, in the control sample, the same probe set has a higher expression intensity but still close to zero, it might appear differentially expressed. In reality, there is no confidence that the gene is present and it should be filtered out of the analysis.

The Present/Absent (P/A) calls file format is similar to the expression values file format where for each probe set **P** stands for Present and **A** for Absent for each sample. In the case where neither P or A is present in a column (for example M for Marginal) the gene is considered present for that sample. The user can generate this file using the MAS5.0 algorithm. In the case where the experiment is not performed with the Affymetrix platform, a custom threshold can be used to exclude Absent transcripts (i.e. transcripts in the range of the lower 50% of intensities per sample could be called Absent) (McClintick and Edenberg, 2006) or any other measure of choice (i.e. variance; Hackstadt and Hess, 2009). In addition, for RNA-seq data a user-defined threshold of minimum number of reads per transcript can be applied for the generation of the P/A flags file.

The Cytoband file

The cytoband file follows the BED format (standard UCSC Browser format) and can be directly downloaded from the UCSC site or from our server for human, mouse or rat (http://disco.stembio.org/). It describes the Giemsa stain bands of each chromosome in the organism.

The Normalisation Scheme file

At this point, it is necessary to make the distinction between the normalisation of the raw gene expression intensities using an appropriate normalisation method such as RMA (as described in section 2.3.1) and the normalisation of samples in order to identify differentially expressed genes, which is the focus of this section. The initial normalisation of the raw expression values must be performed by the user before data input. It is essential because it corrects for sample-specific background noise and brings the distribution of the expression values of each sample in the same range rendering them comparable.

The normalisation scheme, described here, refers to the way the samples are grouped and compared in order to identify the set of genes that show significant difference in their expression levels between the groups. Each sample can be assigned a set of attributes as for example replicate id, normal or tumour, time collected or type of strain. These parameters can be used to group the samples and subsequently compare them in a meaningful way, i.e. normal versus tumour, in order to find the genes that change between the two conditions.

The normalisation scheme file consists of as many rows as the number of samples for the specific experiment and as many columns as the number of parameters the user wishes to include. The sample names must match the sample names used in the genome annotation, gene expression and P/A flags files.

3.5. Data Processing

3.5.1. Typical workflow

After the data input step, the internal DI.S.C.O. data processing functionality begins. A typical data processing is presented in Figure 3.4.

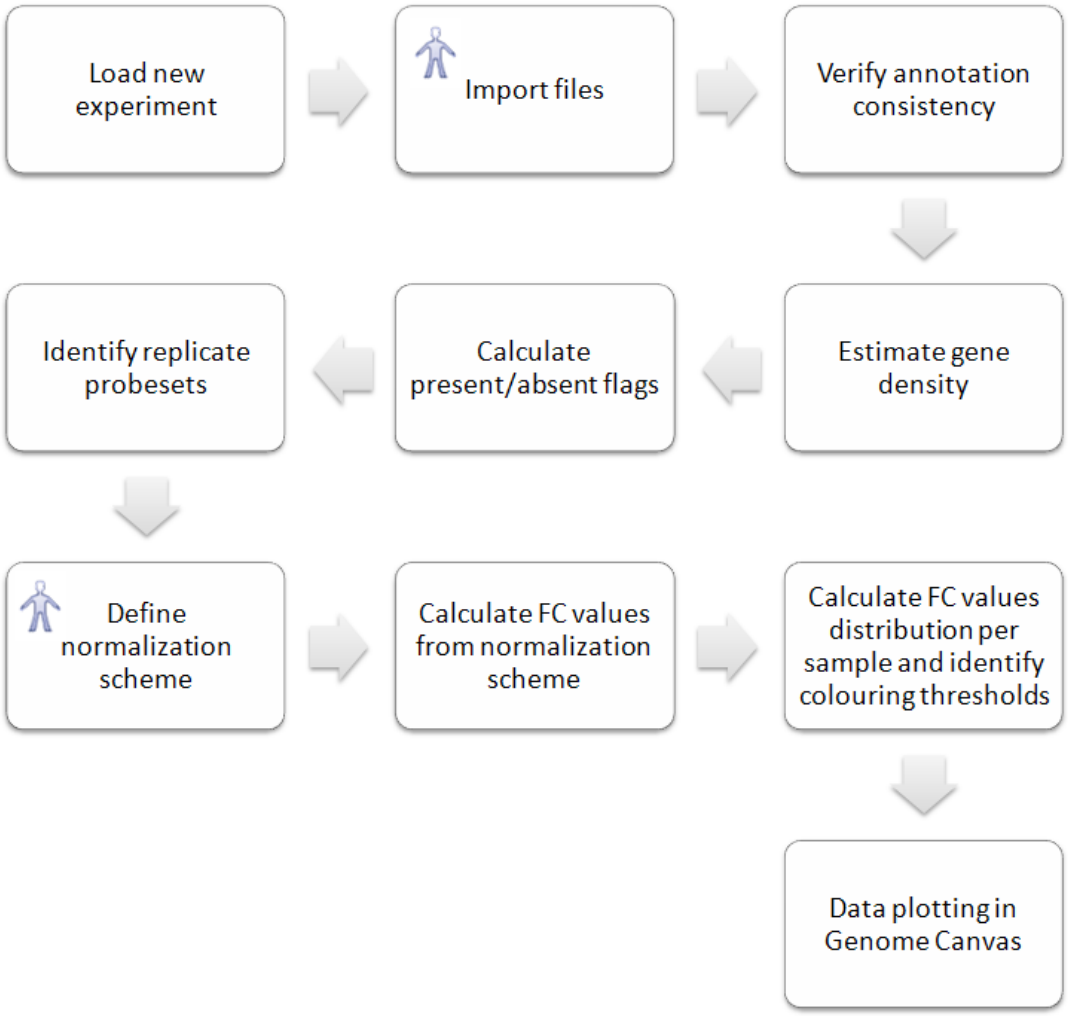

Figure 3.4 A typical workflow of the initial data processing of a new experiment.
It starts at the top left corner with the creation of a new experiment and finishes at the bottom right corner with the visualization of the data in the Genome Canvas window. The human icon indicates the steps where user input is required.

A detailed breakdown of each step follows:

a. **Load new experiment**

b. **Import files** (as described in section 3.4)

c. **Verify annotation**

The fold change (FC) of each gene is calculated based on the measurements provided in the gene expression file. The information of the physical position of the gene is found in the genome annotation file. Therefore, in order to assign the fold change (FC) of each gene at its respective genomic position, an 1-to-1 relationship must exist between the gene identifiers of the genome annotation file and the gene expression file. This step verifies the consistency between these two files. Transcripts that are not represented in both files are excluded and the processing continues as long as at least 80% of the genes are matched between the two files. Otherwise an error exception is thrown and the user is prompted to review the files. In addition, transcripts with no genomic mapping available are removed from the analysis.

d. **Estimate gene density**

Gene density across the chromosomes is calculated by a fixed size sliding window of 1 Mbp and a step size of the same size. Gene density is defined by the number of genes at any given genomic window divided by the expected number of genes if they were equally distributed on the chromosome.

e. **Calculate Present/Absent flags**

As it has been mentioned in the previous section, the P/A flags file can be imported by the user if it is already available. If, however, this file is not available to the user, P/A flags can be internally calculated in DI.S.C.O. They are assigned according to the lower percentile of the raw expression values that are considered to be expressed at background noise levels, as defined by the user during the data input step. The filtering out of the lower 50% of genes has been shown to eliminate most of the unreliable probe sets (McClintick and Edenberg, 2006).

f. Identify replicate probe sets

In the Affymetrix GeneChip arrays, a single gene may be represented with multiple probe sets on the chip. In the present study, probe sets that measure the expression levels of the same gene are referred to as *replicate probe sets*. As an example, in the mouse moe430_2 chip, 40% of the genes are represented by more than one probe set and almost 20% of the genes are represented by more than two probe sets per chip, in some cases reaching the number of ten replicate probe sets per gene (Li et al., 2008). Replicate probe sets exist for three reasons: (i) improved annotation has revealed that some cDNAs that were initially attributed to different loci are in reality the same gene, (ii) additional probe sets have been designed in case where the cross-hybridisation quality of a probe set was poor and (iii) different probe sets have been designed to capture alternative splicing of a single gene (Li et al., 2008). However, there is not a consensus on how replicate probe sets should be treated and different studies have chosen different approaches such as arbitrary selection of a single probe set per gene (Liao and Zhang, 2006) or selection of the probe set with the maximum intensity (Jordan et al., 2005).

The identification of replicate probe sets that correspond to a single gene is an important step in the data processing pipeline. At the Data Input step, DI.S.C.O. groups identifiers that are assigned to the same gene under this gene's name. At the Data Visualisation step, the user can choose to replace replicate probe sets (in case of Affymetrix data) or replicate genes (for any other type of data) by the maximum, minimum, average or median FC value of all the transcripts in the group. This replacement is quite important as it can eliminate clusters (visual or statistical) that simply correspond to the probe sets of a single gene being simultaneously up- or down-regulated and, of course, all mapping to the same genomic position. The downside of this replacement is that it introduces some information loss in the case where replicate probe sets actually represent different isoforms of the same gene. As a result, none of the proposed replacement options (max, min, average or median) is an ideal metric of the different isoforms' transcriptional activity.

g. Define normalisation scheme

The aim of a typical microarray experiment it to examine the relative changes of gene expression at a genome-wide scale and identify genes whose expression levels are significantly altered between two different states. For this purpose, samples are collected that correspond to different types of tissue, different drug treatments, knock-out and wild-type cell lines etc. In addition, to allow sufficient power for the subsequent statistical analysis, microarray experiments should include at least 3 replicates per condition/comparison. This information is provided by the user in the normalisation scheme. At this step, it is defined how the samples are to be grouped and compared to each other. An example normalisation scheme is presented in Figure 3.5.

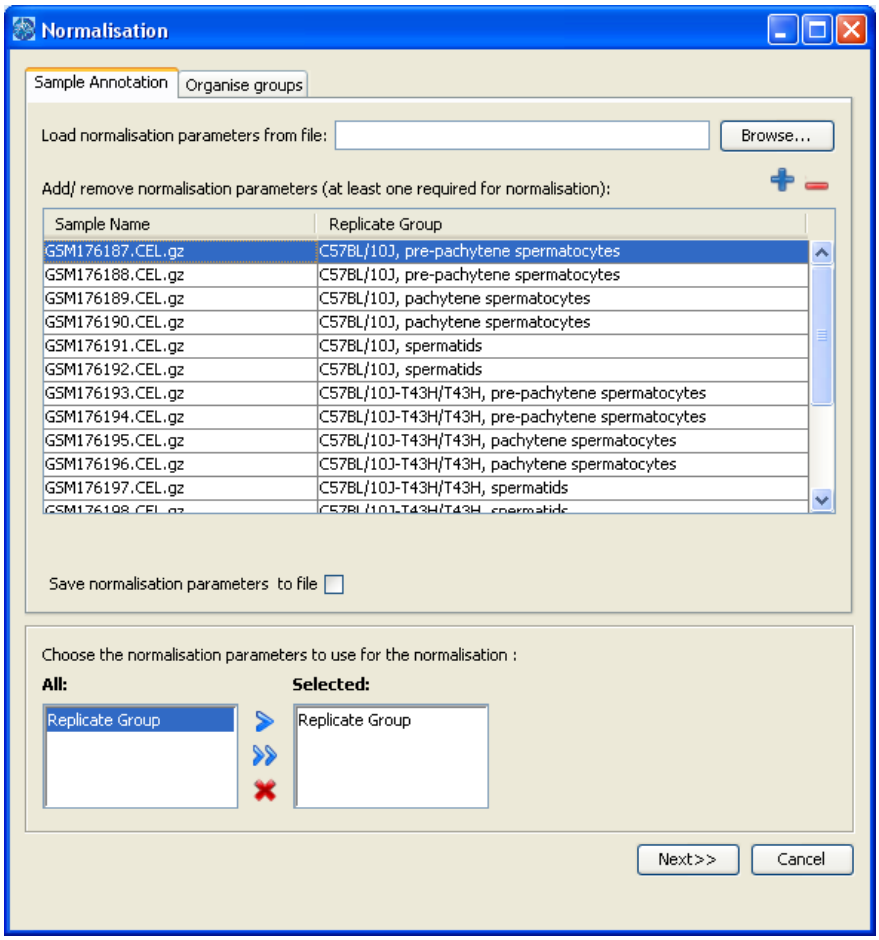

Figure 3.5 An example normalisation scheme.
The Sample Name column contains the samples that are present in the gene expression file. Each additional column of the table represents a parameter that can be used for the grouping and the normalisation of the samples (in this example there is only one parameter: the Replicate Group). The user can add or subtract parameters by clicking on the +/- buttons on the top right corner of the table. The available parameters can be found in the list with the indication "All" and the user can use the arrows to select the ones that will be used for the grouping of the samples in the list "Selected".

h. Calculate FC values from normalisation scheme

After the normalization scheme has been designed by the user, the FC values between the groups of samples to be compared can be calculated. Fold change can be defined as the gene expression ratio between two groups of samples, named hereafter the condition and the control. In DI.S.C.O., the expression levels of each group can be summarised by either the average or the median of the group. Alternatively, if the user chooses to omit the normalisation scheme designing, each sample will be compared to either the average or the median of all samples in the experiment.

i. Calculate FC values distribution per sample and identify colouring thresholds

The FC values that have been generated at the previous step are next log*e* transformed and modelled as a normal distribution as shown in Figure 3.6. FC coloured red signify over-expressed genes, yellow no change and blue down-regulated. The gradient from red to yellow to blue represents different levels of FC values varying from high to low. For the initial distribution and colour gradient, the maximum red and blue values correspond to ±1.5 standard deviations from the mean (86.6% of FC fall in the gradient). These thresholds can be also manually changed by the user after the initial data plotting.

j. Data plotting in Genome Canvas (as described in section 3.6)

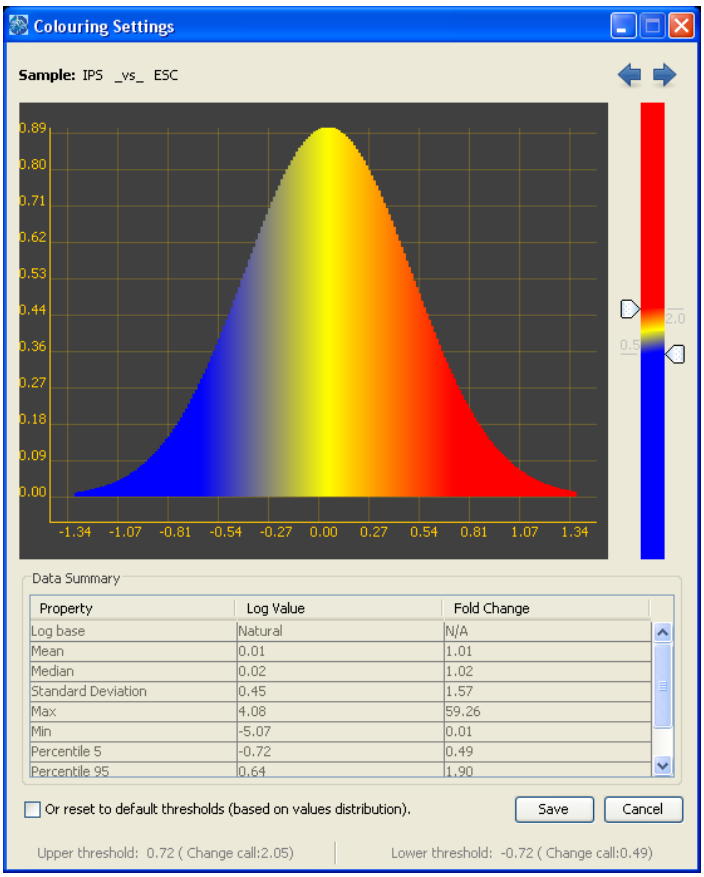

Figure 3.6 Colouring thresholds from FC distribution.

3.6. The DI.S.C.O. Visualisation

3.6.1. The DI.S.C.O. main window

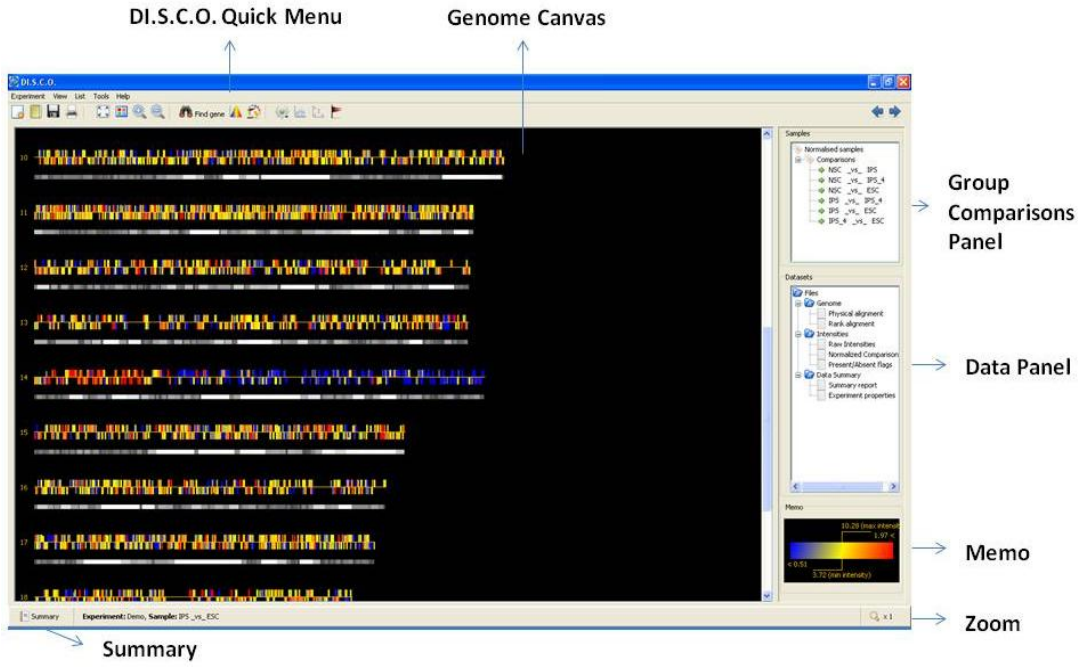

Figure 3.7 The main window of the DI.S.C.O. application

Figure 3.7 presents the main parts of the DI.S.C.O. application main window after data loading. In the **Genome Canvas**, a visual representation of the transcriptome is plotted. The Genome Canvas of Figure 3.7 is at the Full View mode, meaning that all the chromosomes of the organism are plotted on the screen in a way that the user can have a full overview of the genome. More details on the different view modes of the **Genome Canvas** will be given in section 3.6.9. In Figure 3.7, the 21 tracks representing the 21 mouse chromosomes are displayed. The **Group Comparisons** panel contains the list of comparisons between different sample groups as it has been defined during the normalisation process. In the **Data** panel, the core datasets used by DI.S.C.O. during and after data loading as well as a summary of the input data can be viewed and exported. The **Memo** panel gives a brief summary of the colouring scheme and the underlying FC distribution used to generate it. The **Summary** icon, at the bottom left corner of the main window, directs the user to obtain a summary of the datasets used for plotting. Finally, on the bottom right corner, the indication of the current zoom level of the **Genome Canvas** is displayed.

3.6.2. Memo and Genome Canvas interpretation

The **Memo** panel guides the user to interpret the visualization of the gene expression in the **Genome Canvas**. In the **Genome Canvas**, each line represents a chromosome and each rectangular, either over or under the chromosome line (forward or reverse strand respectively), represents a gene. There are three types of information regarding each gene: its genomic position in terms of physical coordinates on the chromosome, the fold change of the gene relative to the control as depicted by the colour assigned to the gene and, finally, the raw expression value of the gene in the condition group as depicted by the height of the gene. Figure 3.8 and Figure 3.9 give an example of how the **Memo** panel can be used to understand the plotting conventions of the **Genome Canvas**.

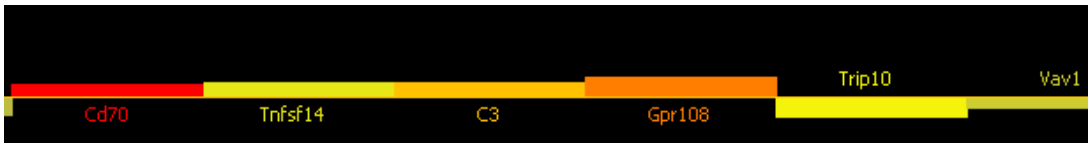

Figure 3.8 An example of a genomic region with zoom level x9 in the Genome Canvas.
Each gene, represented by a rectangular, is assigned a colour that corresponds to its fold change value. The height of the gene represents its expression value in the condition group (averaged across all samples in the condition group). The gene is plotted on its physical (or rank) position on the chromosome and over or under the chromosome line depending on the forward or reverse strand. In this example, *Cd70* is the most highly differentially expressed gene.

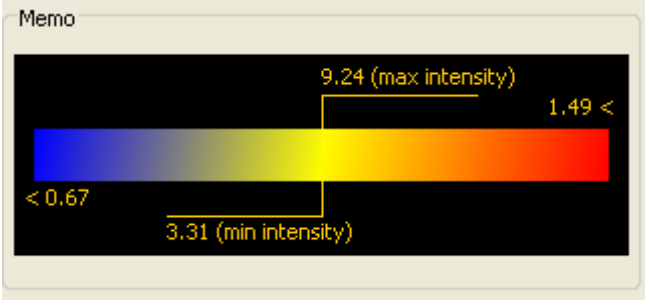

Figure 3.9 The Memo panel for the comparison presented in Figure 3.7.
The horizontal axis corresponds to fold change levels. The gradient represents the colour of the genes in that range. All genes with fold change values less than 0.67 are coloured blue while all genes with fold change values higher than 1.49 are coloured red. The remaining genes are assigned a colour from the gradient as depicted in the Memo panel. The vertical axis represents the intensity value (average of the condition group). The biggest height is assigned for genes that have an average intensity value equal or higher than 9.24 and accordingly the lowest for genes that have an average intensity value lower than 3.31.

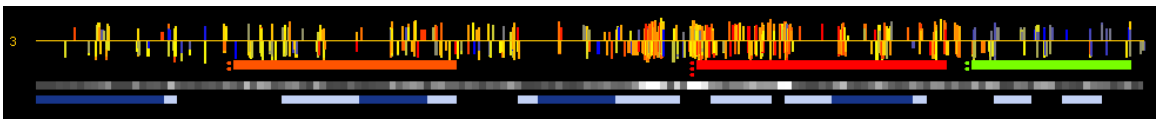

Figure 3.10 The physical view of a single chromosome

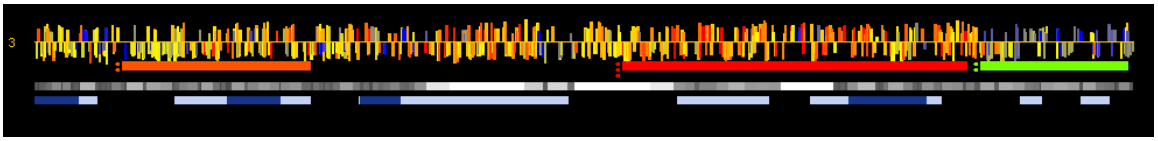

Figure 3.11 The rank view of a single chromosome

There are two different ways of plotting each gene on its mapping chromosome: the Physical view and the Ranking view. In the Physical view each gene is plotted on its actual physical chromosomal coordinates (from start position to end position) as indicated in the genome annotation file. In the ranking view, however, each gene is plotted according to its rank position on the chromosome. In this case, an arbitrary gene size of 100bp is assigned to all genes; the first gene on the chromosome occupies the position 1 to 100 bp, the second the position 101 to 200 and so on, until all genes are plotted on the chromosome. Therefore, the length of the chromosome is also converted to ranking coordinates depending on the number of genes that have been mapped to the chromosome. In this way, nongenic regions (gene deserts) can be removed and the distribution of genes on each chromosome is identical. The exclusion of gene deserts achieves a more compact visualisation for the identification of interesting patterns. In addition, there is no need to make assumption in order to model the gene distribution on the chromosomes for the computational processing of significant clusters.

In Figure 3.10 and Figure 3.11, there are four tracks:

i. The chromosome track
ii. The clustering output track
iii. The gene density track
iv. The cytoband track
v. Custom tracks (not shown here)

3.6.3. The Chromosome track

The first track next to the chromosome number is the chromosome track. The length of

the chromosome track is relative to the actual size of the chromosome in the genome (the size is defined from the genome annotation file by the base pair of the last gene on each chromosome). Each rectangular plotted on the track represents a gene. The height of the rectangular corresponds to the raw intensity value for this specific gene and the colour to the fold change of its expression level when compared to the control (after the normalisation). The fold change magnitude colour scale ranges between deep blue to deep red (with red being over-expression and blue under-expression). When zooming in the chromosome track after a certain level, the common names of the genes are also displayed below/above the gene depending on their orientation (forward/reverse strand).

3.6.4. The Clustering Output track

As soon as one of the clustering algorithms has been executed and has identified clusters of differential gene expression levels on a chromosome, its output will be plotted in the clustering track right below the chromosome track. This is represented by a coloured bar (red for clusters of over-expressed genes and green for clusters of under-expressed genes) that spans from the starting base pair of the first gene in the cluster to the ending base pair of the last gene in the cluster. The intensity of the colour, either red or green, corresponds to the confidence level of the particular cluster as indicated by the significance metric of each algorithm (such as p-value or significance score). For the NN algorithm the confidence level is also indicated by three (high), two (medium) or one (low) dots next to the bar.

3.6.5. The Gene Density track

The next track to be plotted is the gene density track with a grey-to-white colour range (white for highly dense chromosomal areas and dark grey for sparse chromosomal areas). For the physical view, the gene density track does not provide additional information as the density of genes on the chromosomal locus can be directly identified by the user. In the ranking view, however, it is a very important feature because it highlights the gene “deserts” and “islands” and it links back to the physical view.

3.6.6. The Cytoband track

The Cytoband track gives additional information about the observed patterns in

relation to the chromosome bands (i.e. Giemsa stain bands obtained from the USCS genome browser) and the centromere location (centromeric and pericentromeric aberrations have been often described in tumours, for example see (Deng et al., 2010)).

3.6.7. Custom tracks

Custom tracks can also be imported by the user. The custom track can contain any type of genomic feature that can be plotted along the chromosomes and it will appear under the Cytoband track when available or under the Gene Density track when the Cytoband track is not available.

3.6.8. DI.S.C.O. Quick Menu

The DI.S.C.O. **Quick Menu** is located on top of the **Genome Canvas.** It represents a selection of the most common functionalities of the application for quick access by the user. In Figure 3.12, a short description of each button that appears in the **Quick Menu** can be found.

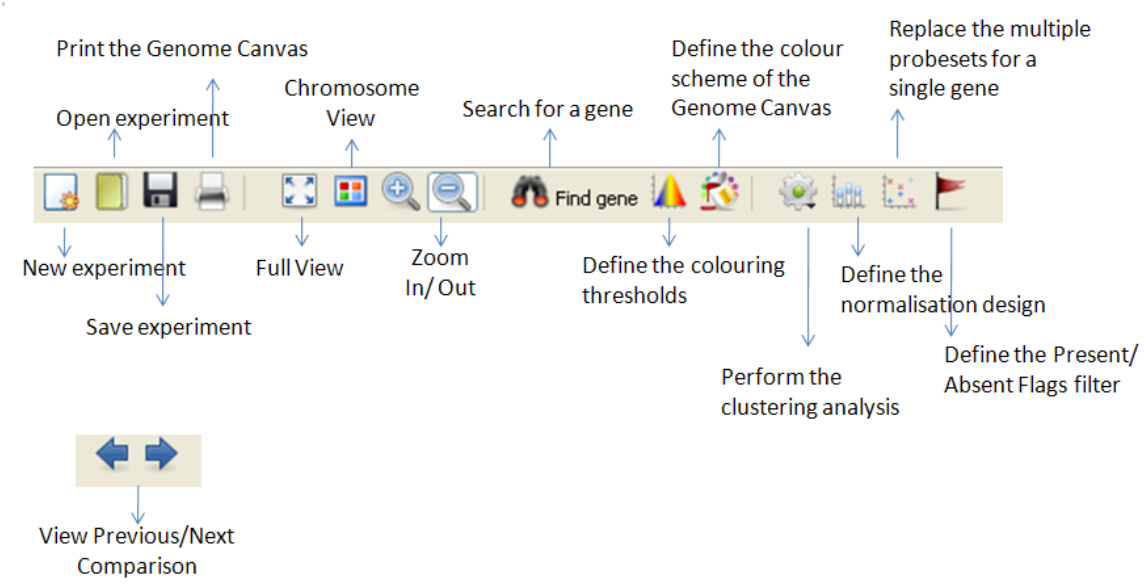

Figure 3.12: The DI.S.C.O. Quick Menu.
Brief description of each button included in the Quick Menu. These buttons represent the most common functionalities for quick access.

3.6.9. Different Views

In Figure 3.13 to Figure 3.16, the different views implemented in DI.S.C.O. are presented. The Full Genome view (Figure 3.13) displays the whole genome (all chromosomes) in a single window. It offers a comprehensive overview of the whole transcriptome in a genomic context. The Chromosome view (Figure 3.14) gives a more detailed view of each chromosome including the additional tracks which are not visible in the Full Genome view. In Figure 3.14, the Chromosome View is plotted in terms of physical position of the genes (Physical view). In Figure 3.15, the same data are plotted in terms of Rank view. The Rank view helps improving the visualisation since it effectively removes any genomic regions with no mapping genes. Finally, in Figure 3.16, the same data are presented with two extra plotting options: i) the Absent probe sets are no longer plotted on the chromosome and ii) replicate probe sets have been replaced with their median value. The succession of the following figures clearly demonstrates how the visualization efficacy can be greatly improved with the introduction of these steps.

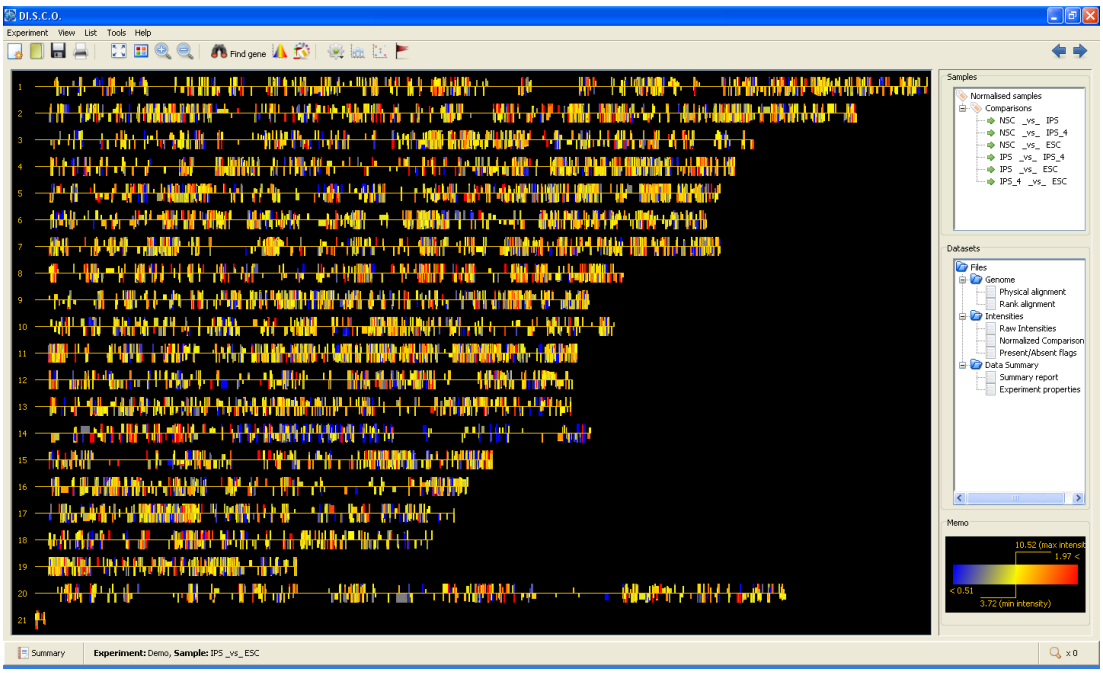

Figure 3.13 The Full Genome view.

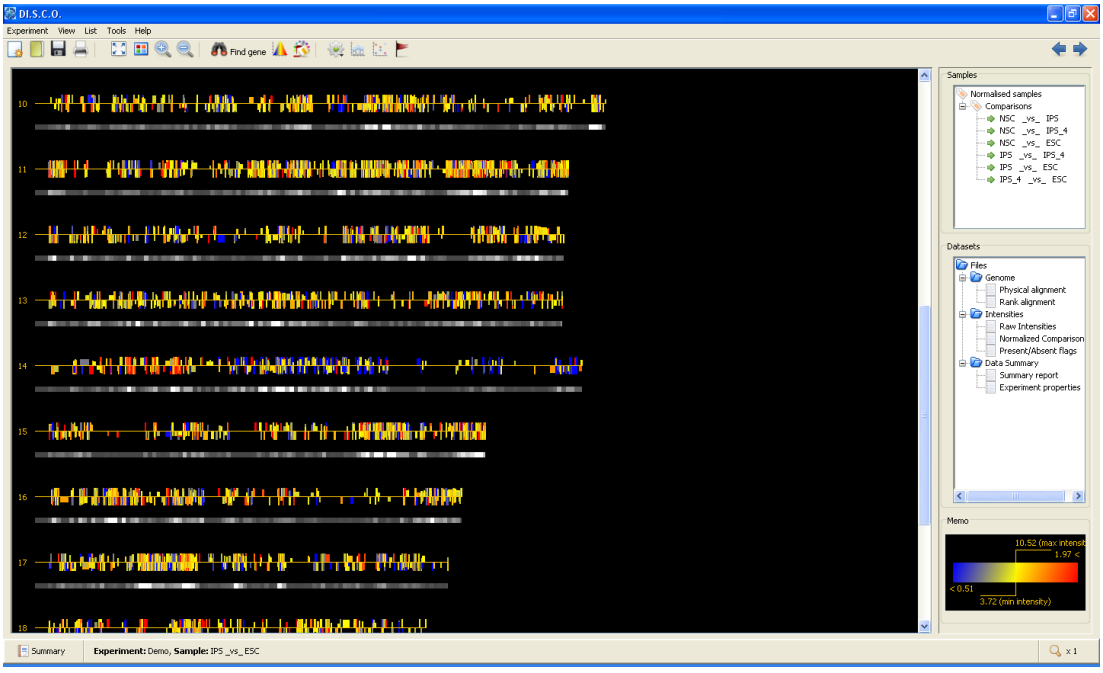

Figure 3.14 The Chromosome view in Physical view mode.
Chromosomes 10 to 17 displayed in Physical view.

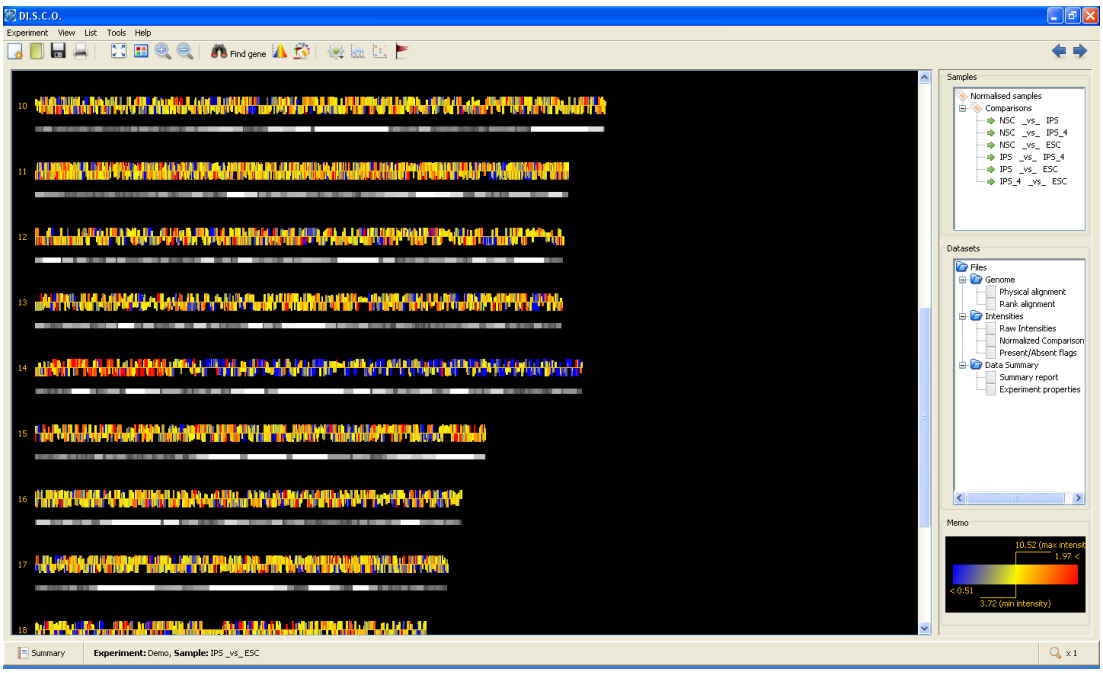

Figure 3.15 The Chromosome view in Rank view mode.
Chromosomes 10 to 17 displayed in Rank view.

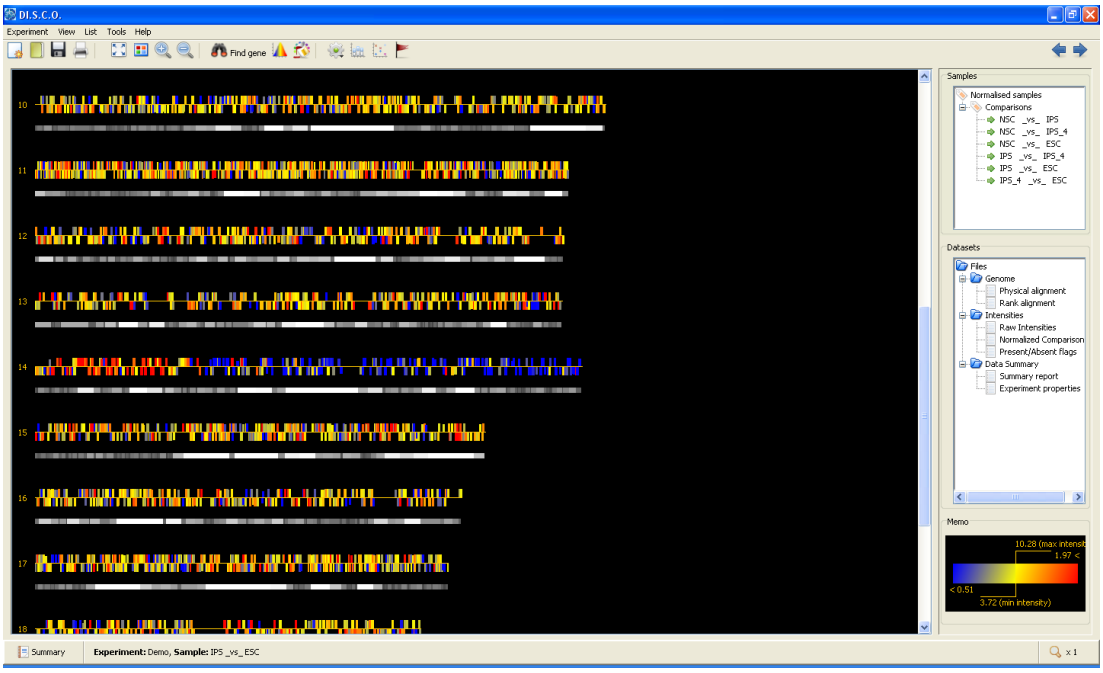

Figure 3.16 The Chromosome view in Rank view mode with Absent flags removal.
Rank view mode with additional absent flags removal and replacement of replicate probe sets with their median.

3.7. Integrated Algorithms for the Identification of Genomic Regions of Non-Random Transcriptional Activity

This section provides the description of the three integrated computational methods that constitute the cluster analysis module: the PGE method (De Preter et al., 2008), the TV minimisation method (Nilsson et al., 2008) and the NN method (Tomlinson, unpublished). All of the three methods were originally implemented in different programming languages and translated in Java by Ioannis Sarantidis as part of his MSc thesis (Sarantidis, 2008). The methods were then integrated in DI.S.C.O. in collaboration with I. Sarantidis. The histogram shape-based thresholding method was developed by Liao et al. (2001) based on the work of (Otsu, 1979). I have adapted the code provided by Yasunari Tosa (version Feb. 19th, 2005) that can be found in the plug-ins collection of the Image Processing and Analysis in Java (ImageJ) program[3].

3.7.1. Histogram shape-based thresholding of FC values

A typical task in computer vision and image analysis is the separation of the image in regions of interest, which represent objects, and regions that bear no relevant information and represent the background (*image segmentation*). *Intensity-based* image segmentation is commonly performed by examining the intensity of pixels and identifying thresholds that can distinguish between regions with high and low pixel intensity. During this process, it is presumed that relevant regions occupy a different range of intensity values from the background. The image can be then transformed into a binary representation where each pixel that falls below the identified threshold is assigned a value of zero and each pixel that is above the threshold a value of one. Alternative ways for image segmentation include clustering methods (k-means), entropy maximization methods, mixture modelling and region growing. A detailed review can be found here (Pal and Pal, 1993).

[3] http://rsbweb.nih.gov/ij/plugins/multi-otsu-threshold.html

A way to find the appropriate thresholds for image segmentation is to examine the histogram of pixel intensities in order to determine whether two or more distinct modes are present – corresponding to the foreground and the background respectively. In DI.S.C.O., the distribution of FC values is represented by a histogram which is also used to define the colouring thresholds. Therefore, an intuitive way of identifying the thresholds for gene expression levels that correspond to up- and under-regulation is by taking advantage of the properties of the FC values histogram distribution as well (also see section 3.5.1i). Contrary to grey-scale image processing where the pixels are separated in background and "relevant", in this case there are two "relevant" classes, namely over- and under-expressed while the no-change genes represent the background.

The Otsu method is a non-parametric method that identifies the optimal thresholds for grey-scale image segmentation by calculating the thresholds that maximise the between-class variance using an exhaustive search (Otsu, 1979). It has been adapted for multi-level thresholding using a faster recursive algorithm by Liao et al. (2001). The implementation proposed by the latter has been integrated in DI.S.C.O. to perform multi-level threshold identification for the three classes of gene expression previously described. A brief description of the original Otsu method follows:

The Otsu Method

Given a grey-scale image with N number of pixels each one having an intensity value of i where $i \in (1, L)$. The number of pixels having an intensity level of i is denoted by f_i and the probability of an intensity value i is given by

$$P_i = f_i / N$$

Let us denote the two classes, foreground and background, C_1 and C_2 with C_1 taking intensity levels from $(1, k)$ and C_2 from $(k + 1, L)$. The weights of the classes, W_1 and W_2, are defined the number of pixels per intensity level over the total number of pixels N:

$$W_1 = \sum_{i=1}^{k} p_i$$

And

$$W_2 = \sum_{i=k+1}^{L} p_i$$

While the class mean μ_C and variance σ_C are defined by:

$$\mu_1 = \sum_{i=1}^{k} i \frac{f_i}{\sum_{i=1}^{k} f_i}$$

$$\mu_2 = \sum_{i=k+1}^{L} i \frac{f_i}{\sum_{i=k+1}^{L} f_i}$$

$$\sigma_1^2 = \sum_{i=1}^{k} (i - \mu_1)^2 \frac{f_i}{\sum_{i=1}^{k} f_i}$$

And

$$\sigma_2^2 = \sum_{i=k+1}^{L} (i - \mu_2)^2 \frac{f_i}{\sum_{i=k+1}^{L} f_i}$$

The within-class variance can be computed as:

$$\sigma_W^2 = W_1 \sigma_1^2 + W_2 \sigma_2^2$$

The within-class variance needs to be calculated for all possible values of k and the value of k that yields the minimum σ_W^2 is chosen as the optimal threshold by this method.

The algorithm can be easily extended to multi-level thresholding in a similar way as described in (Liao et al., 2001). The implementation available in DI.S.C.O. is largely based on the code of Yasunari Tosa (Feb. 19th, 2005) from the algorithm of Liao et al. (2001) and further adapted to work with gene expression data. In the proposed implementation, the histogram of FC values is used to identify the optimal thresholds for up- and down-regulation and no change. The histogram consists of all FC values after filtering of outliers (FC values that lie after three standard deviations of the mean).

3.7.2. The NN algorithm

The Nearest Neighbour (NN) algorithm was originally implemented by Dr Simon R. Tomlinson in the programming language C++. In the NN algorithm, clusters are defined by using the distance between two neighbouring differentially expressed genes along the chromosome. For this implementation, every gene (or probe set) is assigned a change-call: up-, down-regulated or no change according to the specified thresholds. Two differentially expressed genes of the same change-call are in a cluster as long as (i)

the distance between them is smaller than a predefined threshold, namely the allowed gap (G), (ii) the number of genes with a different change-call within that distance is also smaller than G. In addition, the distance between neighbouring genes is defined in terms of the rank rather than the physical position of the genes in order to eliminate any positional bias from the underlying genomic distribution of genes on the chromosomes.

A graphical representation of the algorithm can be found in Figure 3.17. Initially, each chromosome is scanned in a linear fashion and adjacent genes with the same change-call are grouped together. The distance between each pair is calculated and pairs that fail the gap criteria are discarded. The remaining pairs are forming the clusters and each gene that falls within the cluster boundaries is assigned the change-call of the specific cluster. Clusters are always examined in an ordered fashion depending on the distance between the included genes. Thus, in case of conflict, where a gene of a pair has already been assigned a different change-call in a cluster of a smaller distance/gap but falls within the boundaries of a cluster of its original change-call with a larger distance/gap, the second cluster will be discarded. After this initial grouping, chromosomes are again scanned and clusters of genes with the same change-call, rather than pairs, are now grouped after considering the distance G between them. Each cluster is assigned a score C_s:

$$C_s = \left| r - g\frac{R}{G} \right|$$

where: $r, g: number\ of\ up-,\ down-\ regulated\ probesets\ in\ cluster$
and $R, G: total\ number\ of\ up-,\ down-\ regulated\ probesets\ in\ dataset$

In order to establish the null distribution of the score statistic, a user-defined number of random permutation is performed (proposed at least 1000). Permutations are performed by randomly shuffling the chromosomal position of the genes (using the MersenneTwister pseudorandom number generator (Matsumoto and Nishimura, 1998)) and run the algorithm in the resulting random dataset. The empirical p-value of score C_s can then be computed by observing how often a random score exceeds C_s. A user-defined p-value $p \in [0,1]$ is used as the threshold for significant clusters. As an additional step, the hypergeometric distribution (see section 3.1.2) is used to further filter clusters that have a hypergeometric p-value lower than p.

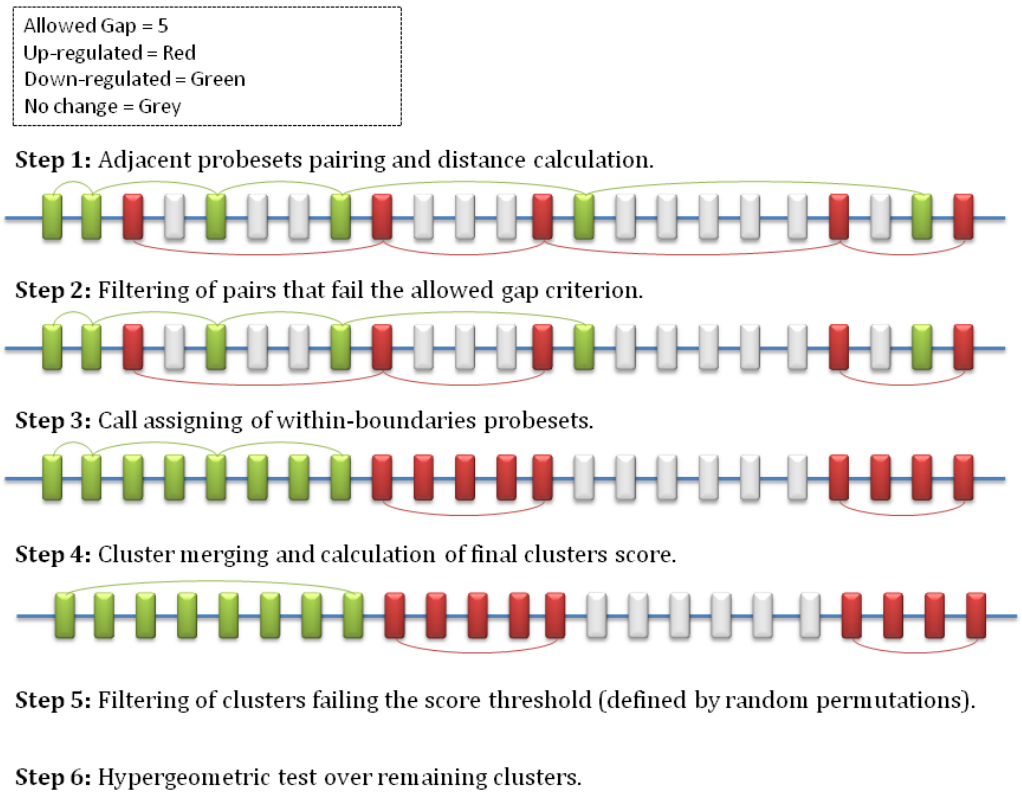

Figure 3.17 The NN algorithm

Finally, the algorithm assigns each significant cluster from the original data a confidence level: high, medium or low. A cluster is high confidence if its score exceeds the highest obtained score from the permutation data, medium confidence if its score lies between the median score and the top score obtained from the permutation data, and finally low if its score is below the median score and over the lower score obtained from the permutation data for significant clusters of the same change-call.

3.7.3. The PGE method

The Positional Gene Enrichment method (PGE) was first implemented as a Perl script, publicly accessible through a web interface by De Preter et al. (2008)[4]. The original implementation typically accepts gene lists, namely genes of interest, from five organisms (Homo sapiens, Mus musculus, Rattus norvegicus, Arabidopsis thaliana and

[4] http://homes.esat.kuleuven.be/~bioiuser/pge/

Saccharomyces cerevisiae). A gene list may contain i.e. over-expressed genes between two conditions, differentially expressed genes in an experiment or genes sharing a common characteristic such as members of the same pathway. In the DI.S.C.O. implementation of PGE, two gene lists are submitted, the one containing all over-expressed genes and the other consisting of all down-regulated genes.

PGE uses the hypergeometric distribution to identify significantly enriched regions (and in case where the total number of genes in the dataset is significantly larger that the size of the query, the cumulative binomial distribution). Again, the rank position of the genes is considered for the cluster formation. The candidate regions are defined using an adaptive size sliding window and must comply with the following set of rules (De Preter et al., 2008):

1. The region must contain at least two genes of interest
2. The region must not include a smaller region with the same number of genes of interest
3. There is not a bigger region that contains more genes of interest and the same number of genes of no interest
4. There is not a bigger region, containing the region in question, that includes a higher percentage of genes of interest
5. The region must not include a smaller region with a lower p-value
6. The region does not include any smaller region with less than expected genes of interest

These rules address the problem of region redundancy resulting by the use of the adaptive size sliding window. Specifically, rules 1 to 3 are enforcing the pertinence definition described in (Barriot et al., 2007) for any type of gene set enrichment analysis. The three additional rules proposed by De Preter et al. (2008) are introduced in order to further reduce redundancy especially in the context of chromosomal gene enrichment.

Figure 3.18 demonstrates how the PGE method initially tests all possible windows of all possible sizes ($\left(\frac{n(n-1)}{2}\right)$ windows where n is the number of genes of interest). Each node represents a candidate region but only the dark-green nodes meet the

aforementioned rules. In Figure 3.18 the input gene list consists of down-regulated genes while the list of over-expressed genes is submitted separately and processed in a similar way.

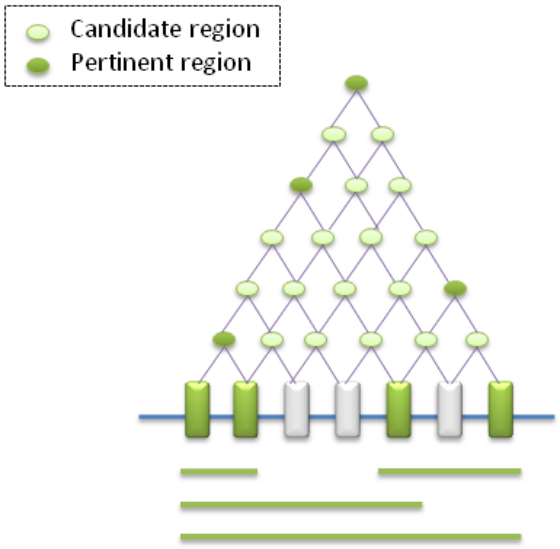

Figure 3.18 The PGE algorithm
(adapted from (De Preter et al., 2008))

3.7.4. The TV method

The Total Variation (TV) minimization method (Nilsson et al., 2008) is based on a technique that has been classically applied to the task of signal restoration or image recovery in the presence of noise. Contrary to the NN and PGE algorithms, TV uses the FC values (or alternatively the raw expression values), rather than the gene's change-call, to segment the chromosomal regions with biased gene expression. The original motivation for the development of this method was the discovery of regions of transcriptional deregulation as a result of underlying genomic or epigenetic events, such as aneuploidies or DNA methylation of large chromosomal domains.

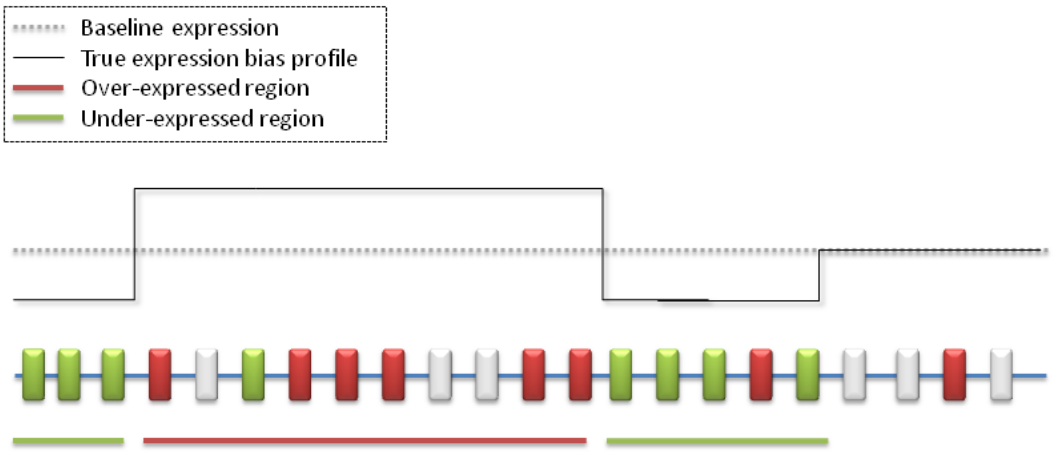

Figure 3.19 The TV algorithm.
Graphical representation of a chromosomal region with biased gene expression levels. The TV method assumes that the expression profile of a region with an increased/decreased average transcriptional activity can be represented as a constant signal as indicated by the black line. Genes whose expression levels deviate from the bias profile of the region are regarded as noise.

In the context of transcriptional data, the TV method considers gene expression levels as a piece-wise constant signal corrupted by noise. The noise corresponds to the inherited variability of gene expression and it is also linked to the observation that not all genes in a region of a chromosomal gain or loss will be necessarily affected by the underlying genomic change. The purpose of the method is to recover the constant profile by removing the signal-corrupting noise (Figure 3.19).

Problem formulation

Let $f(x)$: $I \rightarrow R$ be the expression level of genomic position x in a chromosomal interval I. The TV method regards $f(x)$ as the sum of two signal components: the piece-wise constant signal $u(x) = \mu_i, x \in I_i$ where μ_i represents a plateau level for the interval I and a high-frequency component $v(x)$ that represents the noise. The original signal can then be recovered by minimizing:

$$u = \int_I |u'| + \lambda(u - f)^2 dx$$

The first term, the L^1 norm of u', prevents over-segmentation while the second term is an L^2 fitting term that corrects for under-segmentation. The regularisation parameter λ balances the effects of these two counteracting terms. Therefore, for $\lambda = 0$ the constant bias profile will simply represent the average expression levels of the interval I, while for $\lambda \rightarrow \infty$ the solution will converge to $f(x)$. Nilsson et al. (2008) show that the

discrete-form objective can be represented by the following formula:

$$J_n^* = min_{n'}\{J_{n'-1}^* + |\mu(n'', n-1) - \mu(n', n)| + \lambda v(n', n)\}$$

where J_n^* is the optimal segmentation value for the objective function of a closed interval $[1, n]$ where n is the left boundary of the interval (after discretisation). In addition, $n' \in [1, n]$ is the starting point of the last segment in the optimal segmentation and $n'' \in [1, n']$ is the starting point of the last interval of the optimal segmentation. In addition, $\mu(a, b)$ is the average gene expression levels of a region $[a, b]$ while $v(a, b)$ is the sum-of-squares about the average for $[a, b]$. The proposed solution also follows the dynamic programming formulation since the optimal segmentation of a region can be explicitly calculated given the optimal segmentations of all previous regions.

The performance of the algorithm largely depends on the selection of the regularisation parameter λ. It is impossible, however, to define a single value for λ that will effectively reflect the optimal degree of segmentation for all genomic regions simultaneously. Therefore, Nilsson et al. (2008) propose a scheme that uses a range of λ values in order to capture small regions of altered gene expression levels and still keep larger regions intact from over-segmentation. In other words, the algorithm implementation starts by $\lambda = 0$ where a whole chromosome is considered as a relevant segment (a relevant segment is defined as a segment whose plateau μ_i exceeds a predefined threshold). Then, the λ value is iteratively increased in order to obtain higher resolution, where

$$\lambda_k = \frac{2}{\mu_i * s_k}$$

With s_k being the size of candidate segments to be examined for relevancy and $k > 0$. For example, $\lambda = \frac{2}{10}$ allows for the consideration of segments with average gene expression levels around 10, which can reflect a segment of 10 genes with plateau level around 1.0, or alternatively, a segment of 20 genes with plateau level around 0.5.

3.7.5. Typical workflow for cluster analysis

Typically, the initial data plotting is followed by cluster analysis. Figure 3.20 demonstrates the necessary steps for cluster analysis. It should be noted that for any of the clustering techniques, only the rank position of the genes/ probe sets is considered and genes that have an Absent flag in the condition group are filtered out.

Initially, the user can either accept the automatically calculated thresholds for over- and under-expression, as identified by the Multi-Level Otsu algorithm, or manually provide constant thresholds for all comparisons. Secondly, the user has to choose the preferred cluster analysis method and provide the necessary parameters for the specific method (or proceed with the default values). Table 3.2 provides a short description of the set of parameters required from each one of the three clustering methods. DI.S.C.O. is launched with the default set of values that can be used as an initial guideline. Of course, different datasets have different characteristics and the user is encouraged to explore different set of values in order to improve the performance of the algorithm of choice for the specific dataset. The meaning and the effect of each parameter has been already described in the sections presenting each algorithm. The effect of the parameters presented in Table 3.2 will be additionally described in section 3.11, where the validation of the methods is performed with synthetic and biological data. Ideally, a computational model that can achieve high prediction with a lesser number of parameters should be preferred.

Table 3.2 Short description of the required parameters of each one of the three clustering methods integrated in DI.S.C.O.

Parameter	Description
Nearest Neighbour method (NN)	
False positive rate	The desirable percent of false positives (0..1).
Number of permutations	Number of permutations for the randomisation step.
Blank spaces extent	The largest allowed gap between two over-expressed or under-expressed genes/probes.
Parameter	**Description**
Total Variation minimization method (TV)	
Smallest	Lowest number of probes expected in a segment times the average value of the segment.
Largest	Starting point of the regularization loop. The larger it is the larger are the initial clusters which the algorithm searches for. It should be set to the highest number of probes expected in a segment times the average value of this segment.
Threshold	Threshold for a cluster to be considered relevant.
Step	Parameter that controls the range of λ values to be considered and influences the execution time of the TV algorithm. Higher values mean lower execution times but also lower resolution to identify clusters.
Positional Gene Enrichment method (PGE)	
P-value threshold	Threshold for a cluster to be considered significant.
Average significance ratio	Threshold for p-value significance when compared to enrichment. The higher its value, the more important p-values are (compared to enrichment).

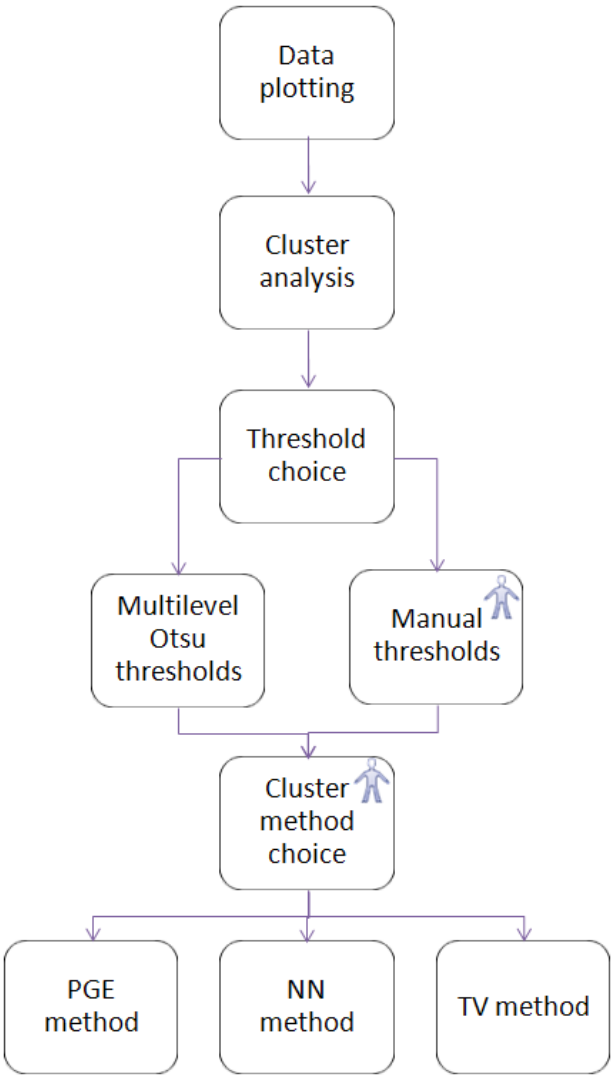

Figure 3.20 A typical workflow to perform cluster analysis. The human icon indicates steps where user input is required.

3.8. RNA-seq analysis

Thus far, the description of the application has been focusing on the analysis of gene expression data obtained using the DNA microarray technology. However, in the era of next-generation sequencing, it is essential to also account for data that have been generated using the emerging high-throughput sequencing technique of RNA-seq.

RNA-seq data analysis in DI.S.C.O. is quite straight-forward. It is presumed that the initial processing of the RNA-seq dataset is already performed before data input in DI.S.C.O. This includes i) obtaining the raw sequencing data, ii) performing quality control, iii) aligning to a reference genome or assembling a transcriptome *de-novo*, iv) identifying the expression levels of different genes or isoforms and v) normalization of different samples. There is a wealth of tools that can address each separate task which exceed the scope of this thesis (for a detailed review of current tools and challenges please refer to (Garber et al., 2011)).

In addition to the normalised raw data obtained from a typical RNA-seq analysis pipeline, the user can easily generate a custom Present/Absent flags file by deciding on a threshold of counts (reads) per gene over which the gene is considered to be expressed. Finally, the data can be imported, the FC values between the groups of samples to be compared can be calculated and the cluster analysis can be performed as previously described.

3.9. Accessing Remote Datasets

DI.S.C.O. is primarily designed to address the data analysis of a single experiment as a standalone client application. Recently, it has been implemented as a module for the integration in the GeneProf workflow engine (Halbritter et al., 2012) for the analysis of next-generation sequencing data. The development and integration of the GeneProf DI.S.C.O. module has been conducted in collaboration with Florian Halbritter and Panagiotis Matzavinos as part of the MSc thesis in Bioinformatics of the latter (Matzavinos, 2011).

GeneProf is a web-based workflow system that contains modules for the analysis of RNA-seq and ChIP-seq experiments. There is a selection of predefined workflows in the engine, while the user can also design custom workflows. In addition, the GeneProf database is a valuable resource of a great number of publicly available analysed datasets that can be further examined and re-processed by the user in order to gain biological insights.

Figure 3.21 demonstrates the DI.S.C.O. module in the GeneProf workflow designer (PGE algorithm implementation as described in section 3.7.3). The module typically accepts four input datasets (gene expression values, genome annotation, Present/Absent flags and normalisation scheme) from RNA-seq experiments processed as discussed in section 3.8. The output results of the analysis consist of the identified intervals (along with information such as size, p-value, start and end position of cluster) as well as an image representation of the Genome Canvas plot for the specific comparison. The integration of the DI.S.C.O. component guarantees an automated way of examining available datasets for positional enrichment of differentially expressed transcripts and highlights to the user experiments where clusters of high significance are present for further consideration. Finally, it offers a quick and easy incorporation of this type of analysis into analysis workflows.

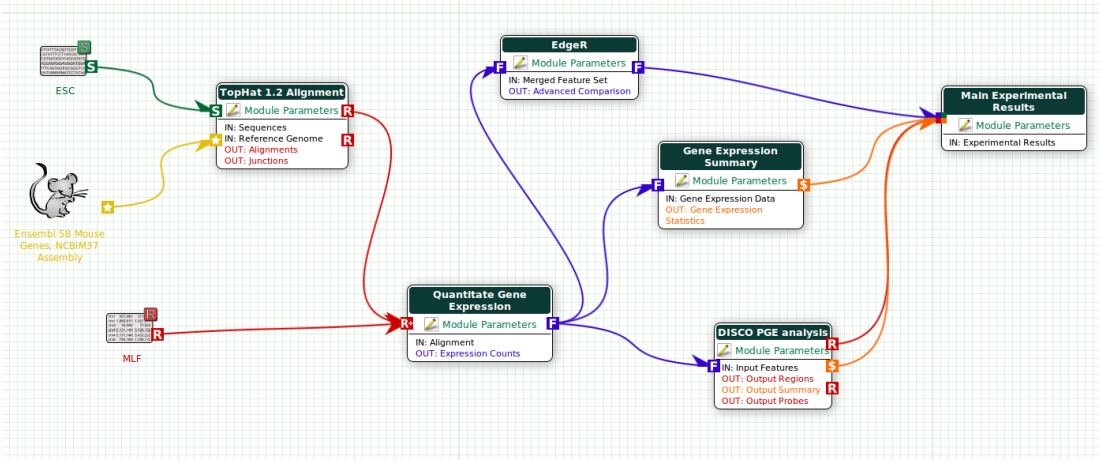

Figure 3.21 The DI.S.C.O. module in the GeneProf workflow designer.

3.10. DI.S.C.O. v1.0

As already discussed in section 3.1.3, the current version of the DI.S.C.O. application builds upon the work that was performed during my MSc project (Skylaki, 2007). Figure 3.22 presents a screenshot of the first prototype application that was submitted as part of the requirements for the completion of the MSc in Informatics. DI.S.C.O. v1.0 was a tool with functionality restricted to the visualisation of the transcriptional data on the genome. From the implementation described in the previous sections of this chapter, the features that were also available in this earlier version are as following:

- Display of the Physical View
- Display of the Ranking View
- Plotting of the Gene Density and the Cytoband tracks
- Display of significantly enriched intervals by importing the output of independent algorithms from a text file and plotting the results as an additional track.

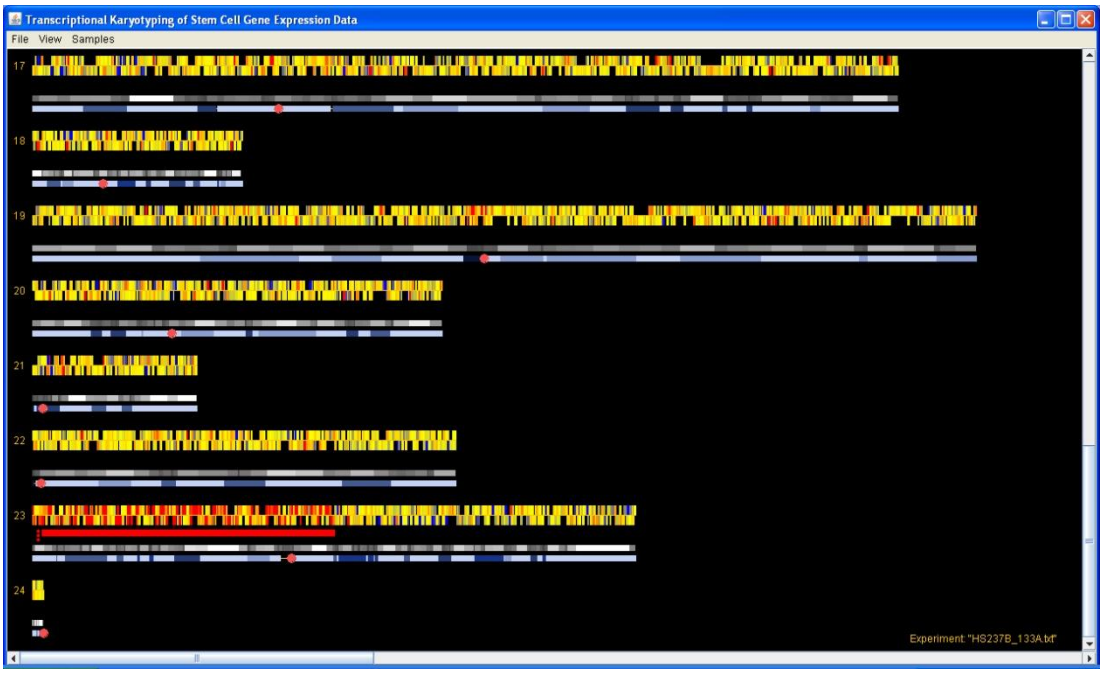

Figure 3.22 The DI.S.C.O. v1.0 visualisation tool in the Ranking view.

In summary, the improvements that have been introduced in the current DI.S.C.O. version are the following:

- Integration of computational methods for the identification of significantly enriched genomic regions,
- Extended functionality of the Genome Canvas, including zooming, available annotation on mouse-over and export of analysis in an image format,
- Automatic choice of thresholds for over- and under- expressed genes,
- Automatic choice of the colouring thresholds for the visualisation of the genes,
- Custom colouring schemes,
- Import and plotting of custom tracks,
- Creation, import and export of custom gene lists,
- Creation of line plots from the expression levels of a gene or a gene list across samples,
- Creation of analysis reports in Portable Document Format (PDF) format,
- Integrated functionality for the designing of the Normalisation scheme,
- P/A flags calculation and integration in the visualisation,
- Replicate probesets replacement,
- Extensive Help files and documentation,
- Extensive error detection and correction functionality during data input,
- Option to redirect the user to online Genome Browsers such as Ensembl[5] or UCSC[6].

For more information, see also the DI.S.C.O. Application User Guide, provided as part of the Supplemental Information in the attached CD.

[5] http://www.ensembl.org/index.html
[6] http://genome.ucsc.edu/

3.11. Validation of the Integrated Clustering Methods

This section discusses the methodology applied for the evaluation of the DI.S.C.O. application, specifically the performance of the integrated clustering methods. Two different types of data can be used for the validation: artificially generated data, where clusters are *a priori* positioned in selected intervals on the chromosomes, and real biological datasets, where the presence of a cluster has already been identified by classic cytogenetic techniques at the DNA level.

The validation of the algorithms requires benchmark data where the underlying distribution of clusters is known. This can be only achieved by the use of synthetic data that need, however, to adequately model the various properties of the gene expression profile of a real biological dataset. In reality, the use of artificial data for validation has been often criticised as synthetic data may represent an over-simplification of the real gene expression profile that prohibits effective evaluation. In addition, there is the danger that the artificially generated data implicitly match the assumptions that have been used to generate the model in the first place rendering the validation biased in favour of the specific model. On the other hand, there is a common problem with the use of experimental data for the evaluation of the performance of an algorithm: in most cases, this approach is limited to the confirmation of previously characterised features in the dataset in question. As a result, it is not feasible to penalise false positive clusters that have not been previously experimentally discovered.

The proposed validation methodology consists of a combined evaluation of both synthetic and experimental data and, therefore, aims to present a comprehensive measurement of the detection power of the methods. In addition, particular care has been taken in the designing of the artificial data in order to achieve the highest potentially accuracy in representing the complexity of a real biological gene expression profile. A detailed description follows in sections 3.11.1 and 3.11.2. This work builds upon and extends the initial validation performed in (Sarantidis, 2008).

3.11.1. Artificial data

The artificial data generated for the evaluation of the visualization and the computational methods integrated in DI.S.C.O. has been directly generated from real biological datasets in order to adequately model the complexity of a real biological transcriptional profile. For this purpose, the previously described dataset of (Kim et al., 2008b) has been used, which consists of neural stem cells (NSCs) reprogrammed to iPSCs using two factors (*Oct4* and *Klf4*) and four factors (*Oct4, Sox2, c-Myc* and *Klf4*) and the wild-type ESC line (three replicates per cell type, GSE10806). The samples have been normalised using the Multiple-Array Average (RMA) (Irizarry et al., 2003a) and genomic data mapping was based on the Affymetrix GeneChip Mouse Genome 430 2.0 Array mapping annotation.

In order to obtain the FC values for the validation of the methods the following normalisation scheme has been applied: i) a comparison between the iPSC-four factors (iPSC4) and ESC group of replicates and ii) a comparison between NSC and ESC groups of replicates. The specific normalization schema was chosen in order to model two different categories of comparisons: i) samples that are still quite similar in terms of transcriptional profiles, such as it would be expected for iPSC and ESC lines and ii) samples that are quite different, since they represent distinct cell types, and a higher number of differentially expressed genes is present. In these comparisons, each gene is represented with the median value of all its corresponding replicate probe sets when applicable. Annotation has been obtained from the Affymetrix website (version Mouse430_2 Release 25 (3/19/08)). From the total number of 45101 probe sets on the Mouse Genome 430 2.0 Array, 43109 had available genomic mapping annotation. After replacement of replicate probe sets the final dataset comprised of 26930 probe sets each one representing a single gene. Table 3.3 reports the statistical summary of the two real datasets ($\log_2 FC$) that have been used to generate the artificial data for the evaluation.

Table 3.3 A statistical summary of the two validation artificial datasets.
The iPSC vs ESC dataset is derived by comparing the median of the group of the three iPSCs samples to the median of the group of the three ESCs samples and the NSCs vs ESCs dataset is derived in a similar manner for the groups of NSC samples and ESC samples. Each gene is represented with a single value after replacement of its replicate probe sets with their median value. The table also presents the number of genes that have a FC value higher than (i) 1.5 FC, (ii) 2 FC or (iii) as defined by the Multi-Level Otsu Thresholds.

Statistical Properties of the three validation datasets		
Statistic metric	iPSC vs ESC	NSC vs ESC
Total number of probe sets	26930	26930
Average	0.012	-0.039
Standard deviation	0.386	1.054
Max value	5.577	10.188
Min value	-4.024	-9.301
FC $\geq$ 1.5	1110	3933
FC $\leq$ 1.5	1310	3978
FC $\geq$ 2	313	2475
FC $\leq$ 2	393	2224
FC $\geq$ Multi-Level Otsu Thr.	738	953
FC $\leq$ Multi-Level Otsu Thr.	533	459

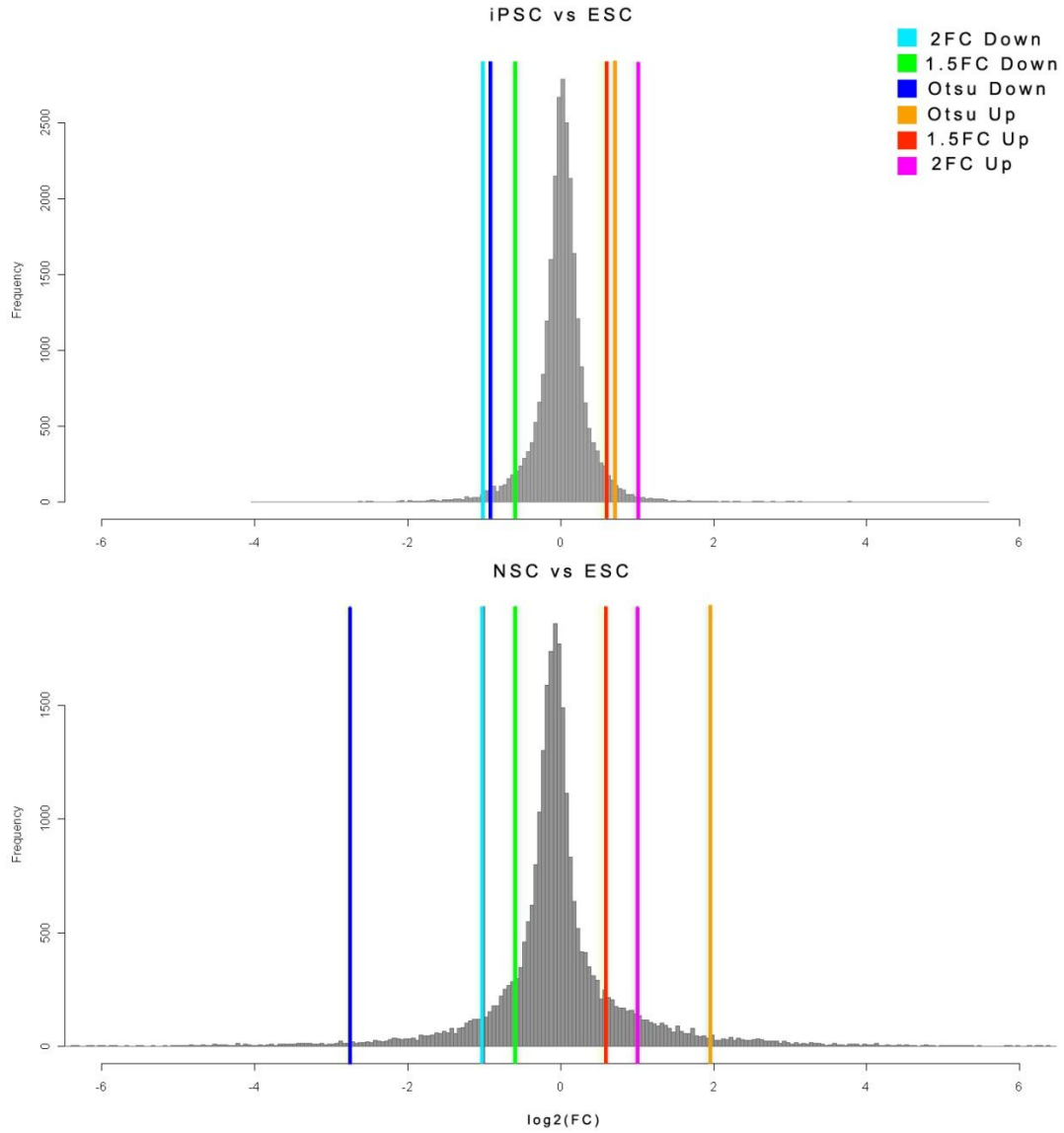

Figure 3.23 The histogram of the FC values of the two datasets used for validation. The coloured lines indicate the three different pairs of thresholds used in the evaluation process.

The generation of the artificial data has been performed as following: firstly, the FC values of the probe sets within a dataset were assigned randomly to probe sets of a different genomic location in order to assure that all biologically relevant clusters have been disrupted. Then, for each dataset, five random genomic locations (chromosomes 1, 14, 8, 17, 11, Table 3.4) were chosen as the starting position of artificial clusters of sizes 200, 100, 50, 20 and 5 probe sets. In every cluster, 85% of the probe sets in the cluster have an expression value higher than the specified threshold (Multi-Level Otsu

threshold, 1.5 FC and 2 FC). In this way, both clusters of relatively milder changes and greatly affected clusters can be modelled. Numerous studies, examining the correlation of gene expression levels and DNA copy numbers in regions of chromosomal aneuploidies, have shown that a high percentage of genes in the region are affected but not necessarily all. For example, 83% of genes in a region of gain showed more than 1.5 FC in yeast (Torres et al., 2007). By using a similar percentage (85%), it is possible to also model the "noise" introduced by genes that are present in a region but not affected by the underlying genomic change as a result of the gene dosage compensation mechanisms of the cell or the absence of expression of the regulatory inputs of the specific genes. In these comparisons the lower 45% of probe sets was considered to have a PMA flag of A (Absent) and multiple probe sets have been replaced with their median values.

Table 3.4 Genomic position of artificially generated clusters.

Chr	Start	Size
1	89110488	200, 100, 50, 20, 5
14	4437801	200, 100, 50, 20, 5
8	19990620	200, 100, 50, 20, 5
17	13117241	200, 100, 50, 20, 5
11	52209964	200, 100, 50, 20, 5

In each validation step, 10 artificial clusters were generated at the genomic positions presented in Table 3.4, with varying sizes (200, 100, 50, 20, 5 probe sets). In total, 50 artificial datasets were tested for each algorithm/parameter/threshold configuration as described in the following sections.

In each validation configuration, the following attributes were altered:

- Algorithm of choice
- Set of parameters for the specific algorithm
- Type of clusters (gain or loss)
- Thresholds used for differential expression

More specifically, to assess the detection power of each algorithm under each specific parameters configuration, three different types of clusters have been generated and tested under three different thresholds: clusters whose gene expression levels are

greater/lower than the Multi-Level Otsu thresholds (Figure 3.23), clusters whose expression levels are greater/lower than 1.5 FC and finally, than 2 FC. In addition, the detection ability of the algorithms was tested when the provided thresholds are 1.5 or 2 FC over clusters generated with the respective thresholds, or in the case of Multi-Level Otsu thresholds, over clusters generated with all three possible ways.

To access the performance of each validation configuration, the sensitivity statistic has been chosen, defined as:

$$Sensitivity = \frac{TP}{TP + FN}$$

with

$$TP = True\ Positives, FN = False\ Negatives$$

where TP is the number of genes correctly identified in a cluster, and FN the number of genes wrongly identified as not belonging to a cluster. Furthermore, the number of $FP = False\ Positives$, genes wrongly identified in clusters, was also recorded.

PGE method

The PGE algorithm uses the average significant ratio parameter (Table 3.2) to determine the balance between cluster with low p-values (which are biased towards larger clusters) and clusters which are smaller but have a better enrichment. Larger clusters with significant p-values may actually include smaller clusters with high enrichment (but not necessarily lower p-value) and this information may still be useful. In order to access the effect of the average significance ratio parameter of the PGE algorithm, validation was performed in 250 random datasets for a set of values ranging from 1 to 3 with a step of 0.5. The p-value threshold was kept constant at 0.01.

Figure 3.24A presents the average sensitivity of the different configurations for over-expressed clusters in the iPSC vs ESC dataset. Figure 3.24B displays the effects of the average significance ratio parameter in cluster fragmentation which reflects the degree that the cluster has been split in more than one sub-clusters. Fragmentation is denoted here as:

$$Fragmentation = \frac{1}{\#\ of\ sub - clusers\ within\ original\ cluster\ boundaries}$$

It is evident that the best sensitivity is achieved when the thresholds for differential expression match the average gene expression levels of the artificially generated clusters. The Multi-Level Otsu method's performance drops for clusters generated with an average of 1.5 FC (since the method estimates the thresholds at the 1.62 FC) and performs equally well in the case of clusters generated with 2 FC average gene expression within the cluster. In all configurations, the method's sensitivity drops at clusters with a small size of only 5 probe sets. Moreover, there is no great difference in the sensitivity of the algorithm for each examined threshold configuration for the average significant ratio parameter. However, this parameter does affect the fragmentation of the clusters. For values lower or equal to 1.5, the clusters are severely fragmented indicating that even the smallest amount of noise causes the method to stop extending the cluster boundaries. This could be problematic in cases of noisy clusters such as the ones encountered in the majority of real biological datasets. A value between 2 and 2.5 seems to perform better under all configurations tested.

In terms of false positive (FP) clusters (Figure 3.25B), all configurations have comparable performance but it improves when the average gene expression of the genes across the artificial clusters is in accordance with the thresholds used for detection. It is clear, nevertheless, that the number of FP decreases as the average significance ratio parameter increases with values between 2 and 2.5 showing the lowest FP rates. Thus, for the rest of the evaluation, the average significance ratio was chosen as 2.5.

In Figure 3.26, the same results are presented for the NSC vs ESC dataset. Again, best sensitivity is achieved by the Multi-Level Otsu thresholds using the Multi-Level Otsu method. However, the detection power of the method drops significantly when the 1.5 FC thresholds have been used for the generation of the clusters. This is due to the fact that for the NSC vs ESC dataset, where many genes are differentially expressed between the two cell types at the 1.5 FC level, the Multi-Level Otsu method has placed the thresholds for differential expression at the tails of the histogram distribution. Thus, the number of altered genes between the two configurations is quite different and the ability of the Multi-Level Otsu method to identify patterns is affected. For the case of clusters with gene expression over 2 FC, the sensitivity of the two methods is comparable for the NSC vs ESC dataset.

It is intuitive to consider a higher threshold for the NSC vs ESC dataset, since positional patterns will be obscured if a great percentage of the genome is differentially expressed independently of chromosomal position. By increasing the thresholds, as done by the Multi-Level Otsu method, it is possibly to identify regions that are enriched in differentially expressed genes more than a random genomic region. Finally, these results are similar for under-expressed artificial clusters (data not shown).

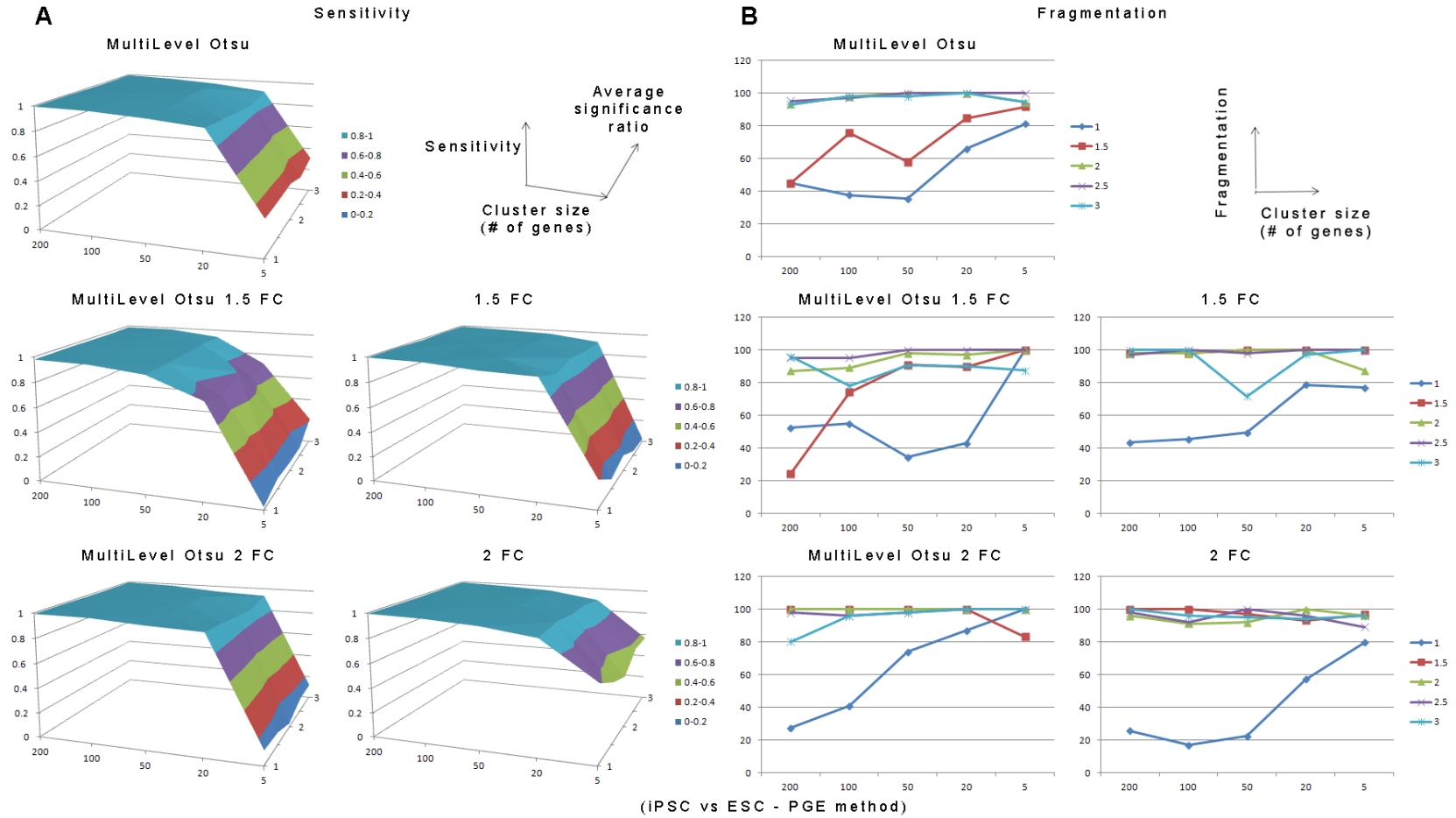

Figure 3.24 Sensitivity and fragmentation of the PGE algorithm with the average significant ratio parameter in a range from 1 to 3 (step 0.5).
A) Sensitivity of the five configurations used: Multi-Level Otsu defined thresholds with clusters of average gene expression varying from 1.5 to 2 FC (including the Multi-Level Otsu threshold which is placed at 1.62 FC). B) Cluster fragmentation denotes the number of sub-clusters included in a single cluster (a single cluster has a value of 100).

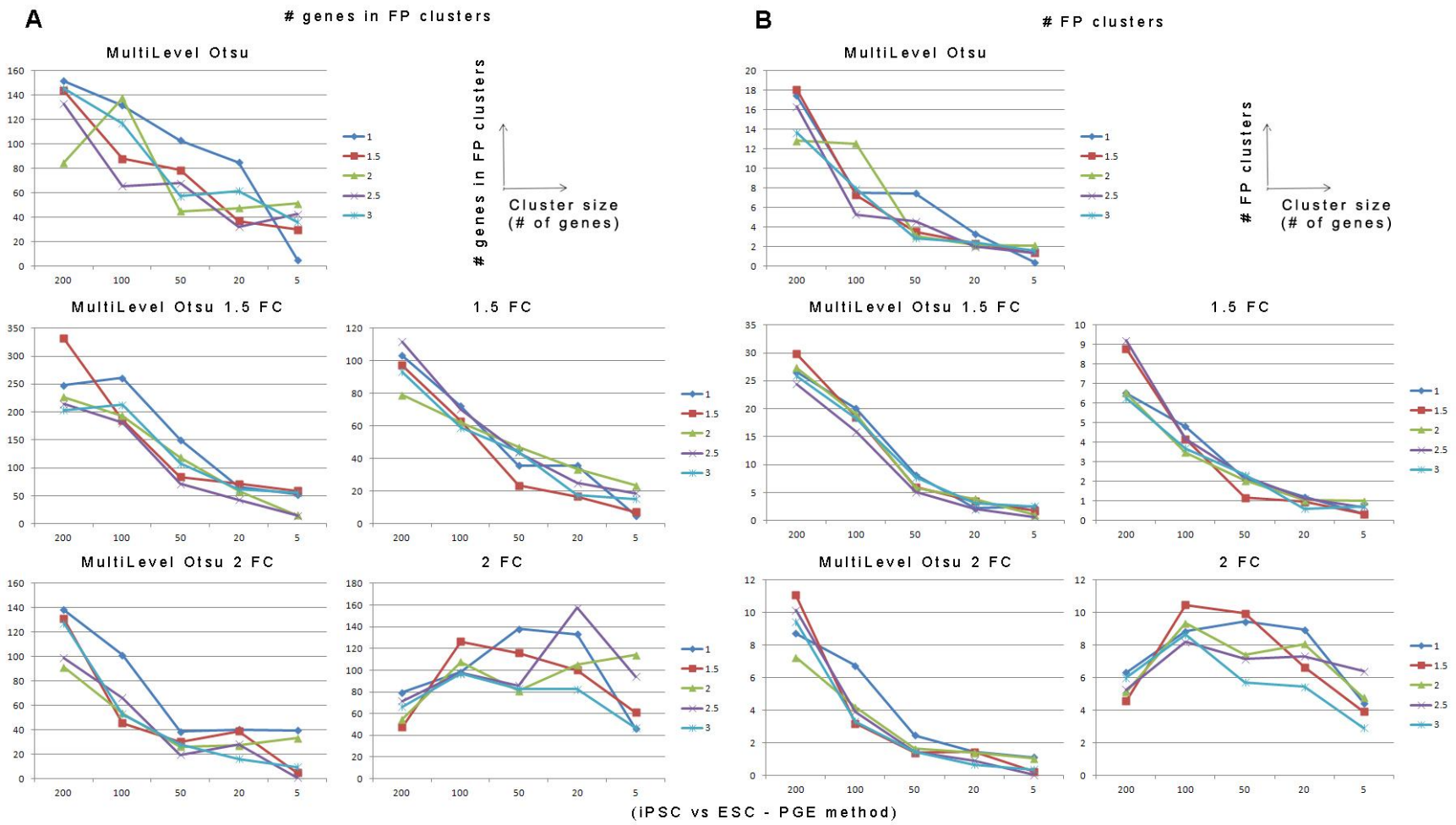

Figure 3.25 Number of genes in FP clusters and number of FP clusters with the average significant ratio parameter in a range from 1 to 3 (step 0.5). A) Number of genes in FP clusters of the five configurations used: Multi-Level Otsu defined thresholds with clusters of average gene expression varying from 1.5 to 2 FC (including the Multi-Level Otsu threshold which is placed at 1.62 FC). B) Number of FP clusters

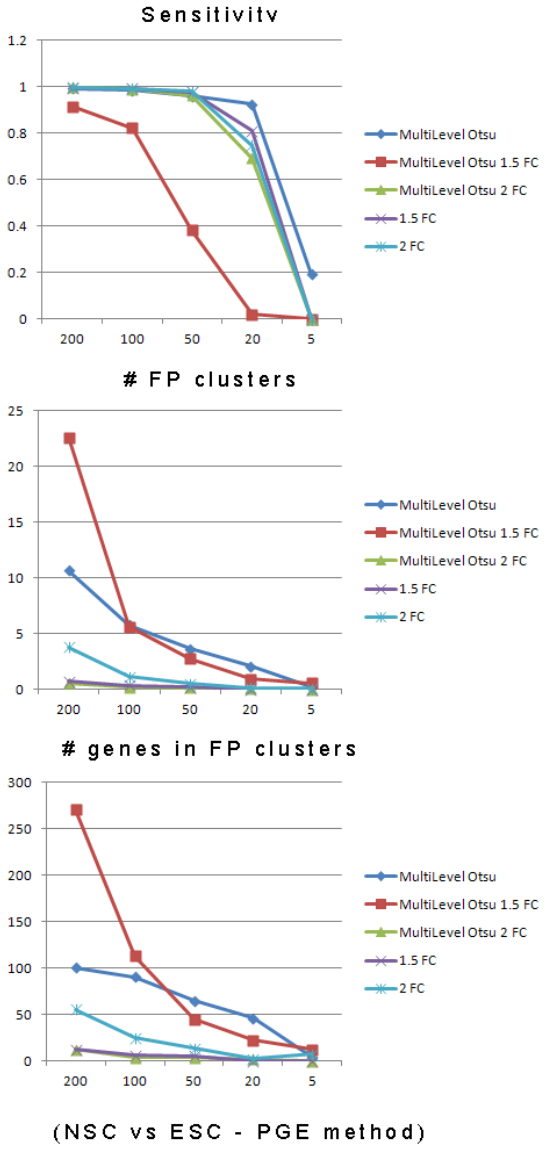

Figure 3.26 NSC vs ESC dataset validation for the PGE method.
Sensitivity, number of FP clusters and number of genes in FP clusters for the five threshold configurations for the PGE method in the NSC vs ESC dataset (average significant ratio=2.5).

NN Method

Firstly, the performance of the algorithm was assessed with various values for the gap parameter $g = \{5, 10, 50, 100\}$ with a constant significance p-value at 0.01 and 500 random permutations.

Figure 3.27 shows the sensitivity of the different threshold configurations which starts

to drop for clusters equal or smaller than 20 probe sets. This effect aggravates as the gap parameter augments. This is particularly prominent in Figure 3.28 where the number of genes in FP clusters and the number of FP clusters is displayed. The actual number of FP clusters is low but the clusters are quite large encompassing a high number of genes. Again, higher values of the gap parameter display higher rates of FP discovery and a gap value of 5 seems to demonstrate the best performance. In addition, the clusters are not fragmented in smaller sub-clusters (data not shown).

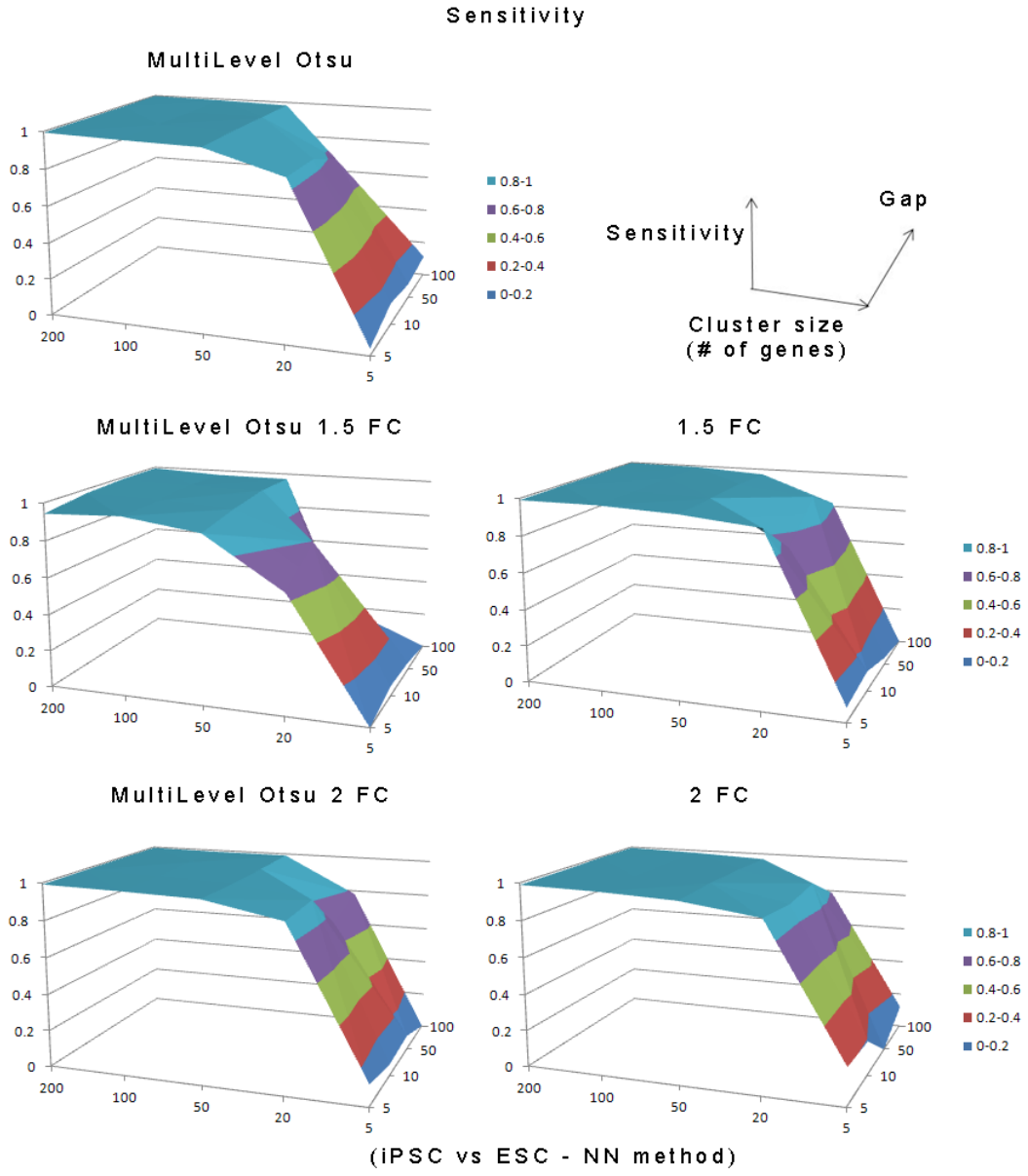

Figure 3.27 Sensitivity of the NN method with a range of values for the gap parameter (5, 10, 50, 100)

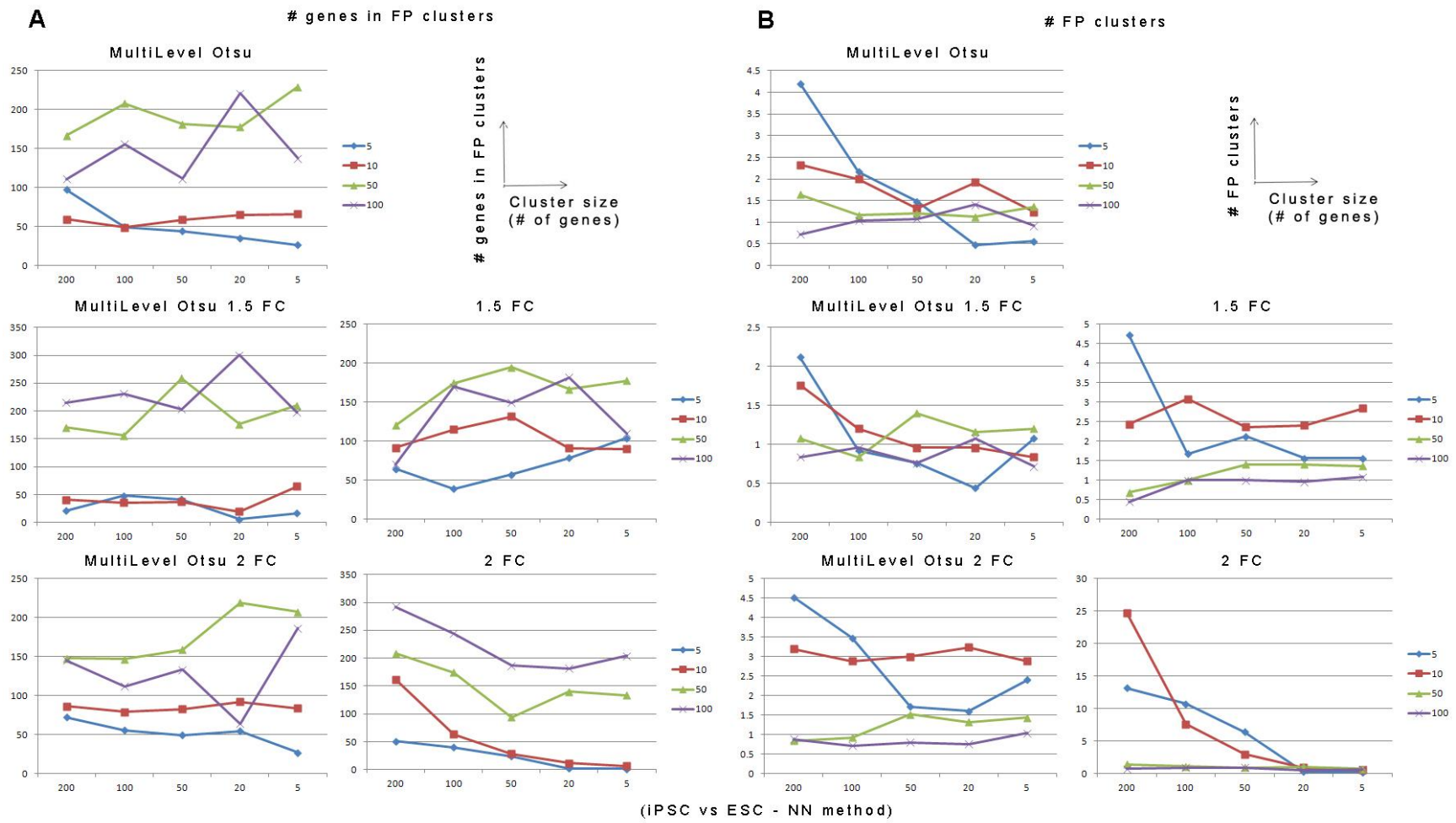

Figure 3.28 Number of genes in FP clusters and number of FP clusters with the gap parameter taking the values 5, 10, 50 and 100.
A) Number of genes in FP clusters of the five configurations used: Multi-Level Otsu defined thresholds with clusters of average gene expression varying from 1.5 to 2 FC (including the Multi-Level Otsu threshold which is placed at 1.62 FC). B) Number of FP clusters

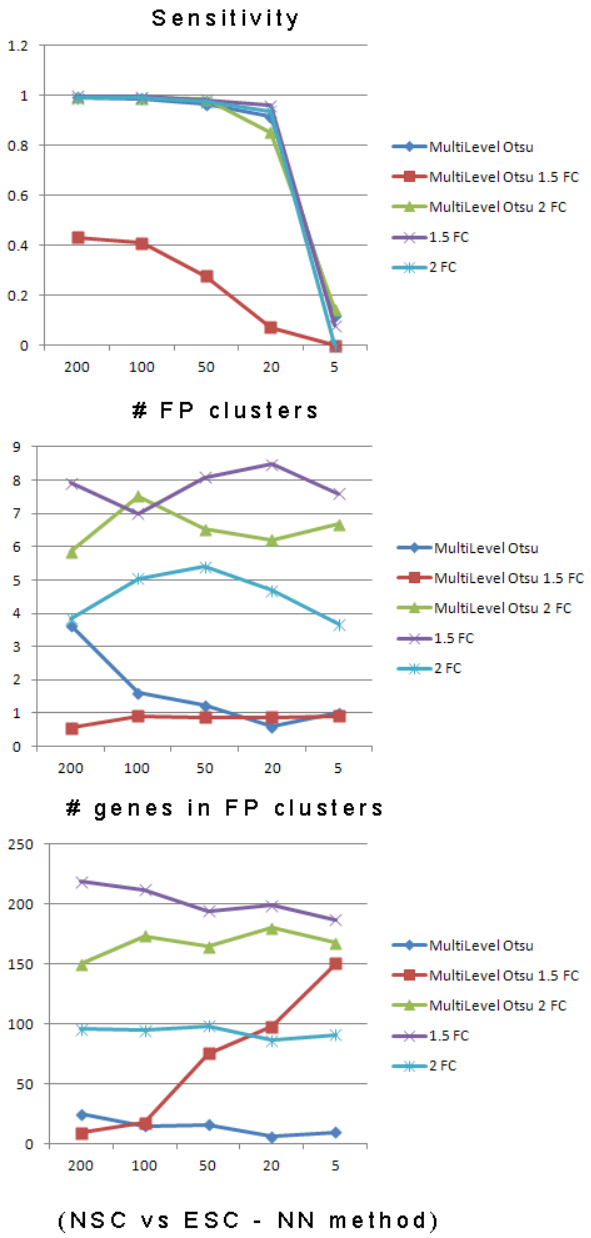

Figure 3.29 NSC vs ESC dataset validation for the NN method.
Sensitivity, number of FP clusters and number of genes in FP clusters for the five threshold configurations for the NN method in the NSC vs ESC dataset (gap=5).

It is possible however that a gap parameter value of 5 is in fact too stringent and does not reflect the noise encountered in actual biological datasets although it seems ideal in the artificially generated clusters. In reality, as it was discovered during the validation with biological datasets, a gap value of 20 is more realistic as a value of 5 fails to identify most of the clusters even when the other methods have successfully located them.

TV Method

As described previously, the TV method requires four parameters as user input. These control the maximum and minimum value of the denominator of the regularization parameter λ, the step for decreasing the denominator to allow searching for clusters of different sizes and finally, a threshold, the plateau μ, which denotes the average gene expression levels that a cluster should exceed in order to be regarded as over/under-expressed. Initially, the detection performance of the algorithm was accessed by using a range of values for the max and min λ denominator values (max, min) and the threshold (t) parameters while keeping the step constant with a value of 1. For the initial exploration of the parameters' space, the iPSC vs ESC dataset has been used with a range of min $= \{5,10,20\}$ and max $= \{50,100,150\}$.

Figure 3.30 shows the results for the three different thresholds (t): 1.2 FC that represents clusters of a subtle change in the expression levels, 1.5 FC for medium change, and 2 FC for highly affected clusters. The performance of the algorithm was measured in terms of sensitivity, number of FP clusters and number of FP genes in the FP clusters since it is important to examine both the number of FP clusters and their sizes.

From Figure 3.30 it becomes clear that the best prediction rate is achieved by the combination of $(t = 1.5, max = (100,150), min = 20)$. For the 2 FC threshold, the performance of the algorithm drops since the clusters are formed with the Multi-Level Otsu threshold (1.62 FC) and, as a result, the average expression level of the genomic clusters is below the detection threshold of the algorithm. It is still possible however to detect mild changes at the 1.2 FC level. For the remaining of the validation analysis the results from the combination of $(t = 1.5, max = 100, min = 20)$ will be presented. It should be noted that a value of 100 for the λ denominator can correspond to a cluster of 50 genes with 2 FC average gene expression, for example, or a cluster of 200 genes with 1.5 FC average gene expression.

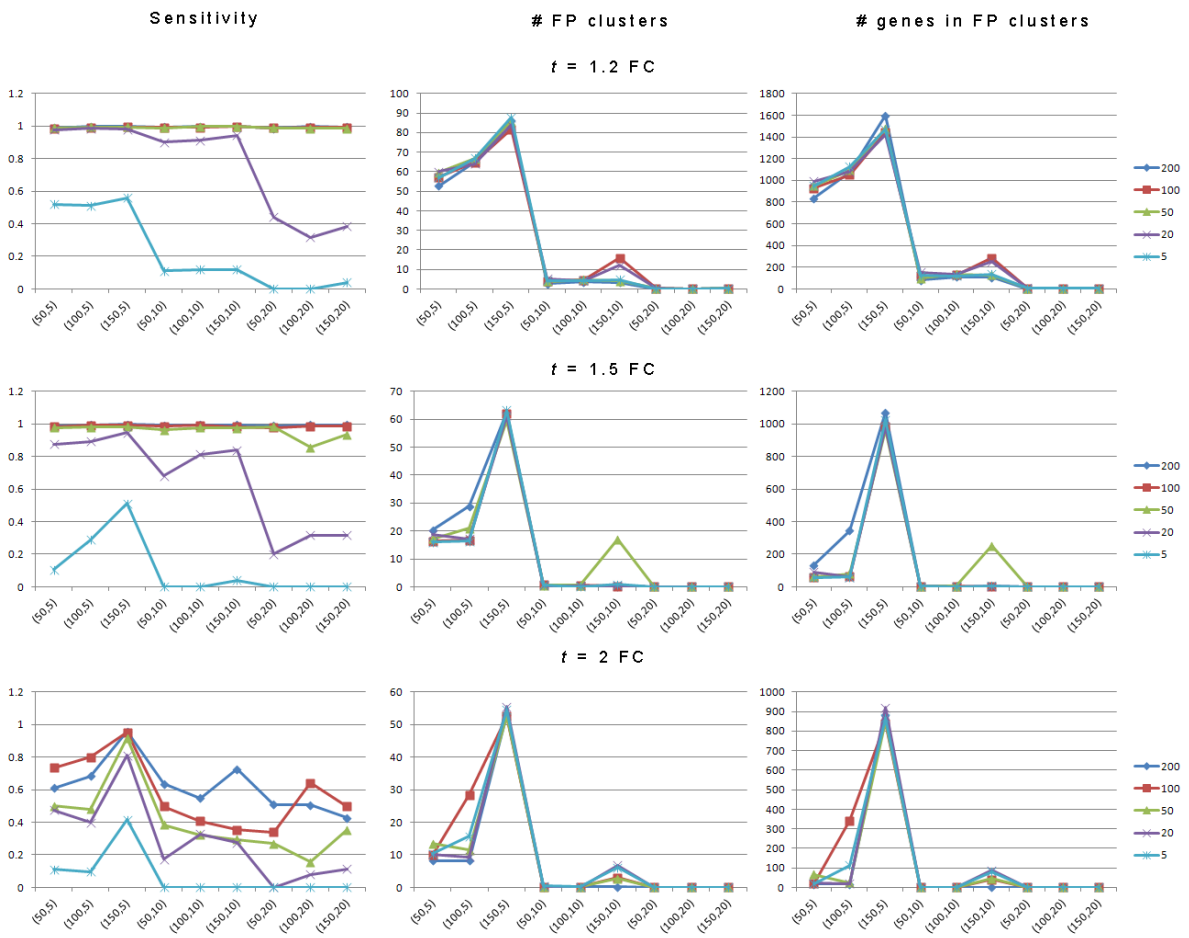

Figure 3.30 Detection performance of TV method under three thresholds (*t*) with a range of (max,min) λ denominator values for different cluster sizes.

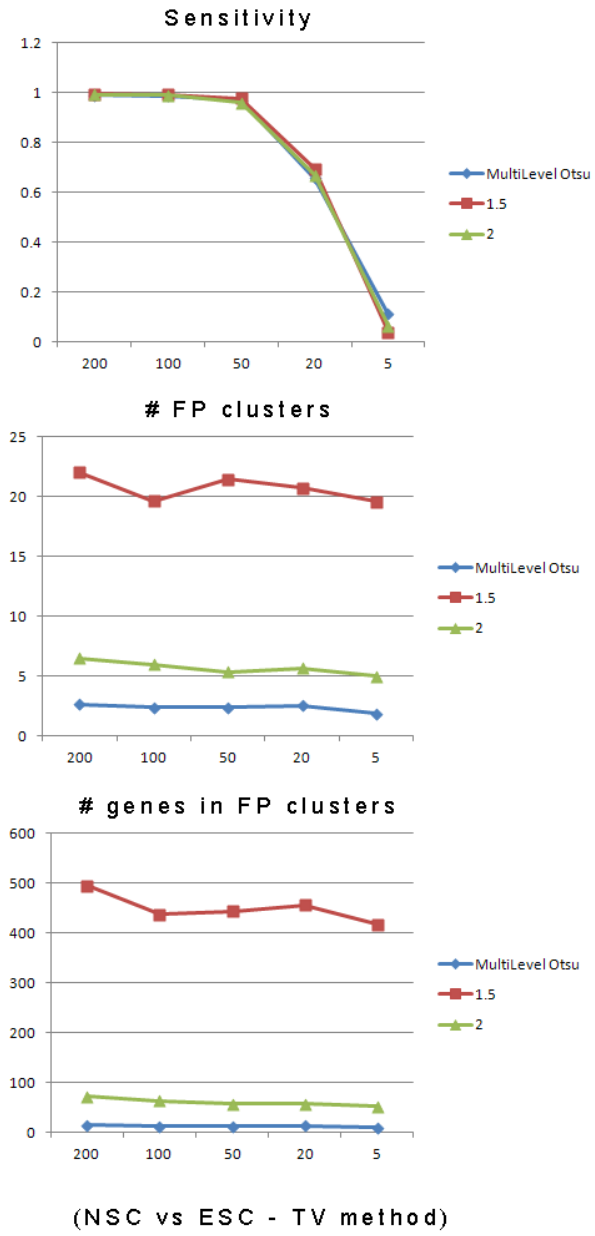

Figure 3.31 Sensitivity, number of FP clusters and number of genes in FP clusters for the five configurations for the TV method in the NSC vs ESC dataset ($t = 1.5, max = 100, min = 20$).

In addition, the method showed equal performance when tested at the NSC vs ESC dataset where the number of differentially expressed genes is dramatically different both for the 1.5 FC level and the 2 FC level. Figure 3.31 shows the three metrics for the NSC vs ESC data using the three different thresholds and highlights the effectiveness of the Multi-Level Otsu method in terms of reducing the FP rates both in terms of identified FP clusters and in terms of keeping their size minimal and thus producing less candidate genes. Finally, the sensitivity of the method starts to drop for clusters

smaller than 50 probe sets.

Response Times

A realistic response time is a necessary prerequisite for every effective method implementation. If Figure 3.32 the average response times of each algorithm under each threshold configuration are displayed for both the iPSC vs ESC and NSC vs ESC datasets. In the general scenario, the PGE method demonstrates the quickest performance, taking less than 2 sec when the number of differentially expressed genes is low. The execution time increases dramatically, however, when a large percentage of the genome is differentially expressed such as in the case of the NSC vs ESC dataset with relatively low thresholds. The NN method is comparably fast and its execution time does not greatly vary between different runs. Finally, the TV method is significantly slower in most cases although, like the NN method, its execution time does not correlate with the number of differentially expressed genes.

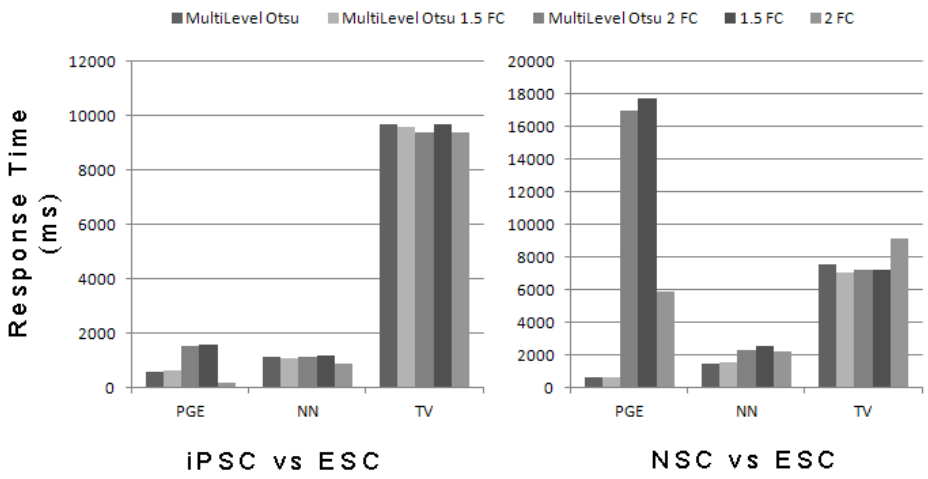

Figure 3.32 Average response times of the three algorithms for the two types of artificial datasets.

3.11.2. Characterised biological datasets

For the evaluation of the methods with real biological data, a range of datasets that have been previously characterised with conventional cytogenetic techniques have been used. The datasets included in this analysis have been selected so as they can reflect different types and sizes of genomic aberrations and the ability to identify these

patterns at the transcriptional level has been assessed. The mapping of gene expression levels was performed with the Affymetrix U133 _Plus2 chip for human and the Affymetrix GeneChip Mouse Genome 430 2.0 Array mapping annotation for mouse data.

Table 3.5 presents the results for the three methods across the different datasets. Where indicated by the symbol (¥), the method attributed more than one clusters to the region in question, resulting to a failure to identify the correct boundaries of the cluster and fragmentation.

According to this analysis, the PGE method was the most successful in identifying the underlying genomic patterns using the parameters defined at the validation step with artificial data (average significant ratio=2.5).

For the NN method, it was clear that the outperforming gap parameter of 5 at the artificial data validation could not adequately capture the complexity of real biological datasets and the parameter was set to 20. With a more relaxed gap threshold, the method identified most of the clusters but, also, suffered from over-fragmentation.

Finally, the TV method demonstrated the lower prediction rates, missing the localised expression patterns in 5 out of 16 datasets and only partially locating the cluster in 5 more. Again, the parameters had to be re-adjusted and the configuration that could identify the underlying signal was $(max = 100, min = 20)$ with a threshold varying from 1.3 to 2 FC. This adjustment was possible since the underlying pattern was known for the specific datasets. However, it reveals a serious drawback of the method that requires a prior knowledge of the type of cluster to be found. Since the model used by the TV method requires four different input parameters from the user, it is not possible to assume that the correct set of parameters could be easily identified in uncharacterised datasets. In addition, in order to exclude the possibility that this observation is related to the integration of the algorithm in DI.S.C.O., Sarantidis (2008) tested the current implementation against the original algorithm (Sarantidis, 2008).

Table 3.5 Method validation with real biological datasets, already characterized by cytogenetic analysis (¥ cluster is segmented to smaller sub-clusters)

Name	Description	Karyotype	PGE	NN	TV	Accession
chHES-3	hESCs (p=53)	46,XX,dup(1)(p32p36)	dup(1)(p32p36)¥	-	-	GSE7234
chHES-3	hESCs (p=172)	46,XX,dup(1)(p32p36) t(1;6;4)(q25;q23;p16) ins(4;1)(p16; q21q25)	dup(1)(p32p36) ¥ dup(4)(p16)	-	-	GSE7234
H14	hESCs (p=25)	48,XY,+12,+der(17) del(17)(p12p13.3)hsr(17)(p11.2)	Partial +12 ¥ dub(17)(p11.2-q25.2)	Dub(12)(p13q15q21q22q23.1) dub(17)(p11.2)	dub(17)(p11.2)	GSE6561
p-hiPS01	hiPSC line 1	Ts1, Ts9	Ts1, Ts9 ¥	Partial dub(1) and dub(9) ¥	-	GSE16093
p-hiPS02	hiPSC line 2	Ts1, Ts9	Ts1, Ts9 ¥	Partial dub(1) and dub(9) ¥	-	GSE16093
hfT18	cell free mRNA	Ts18	Ts18	Ts18	-	GSE25634
MEF89	MEFs	Ts16	Ts16	-	Partial dub(16)	GSE12501
MEF92	MEFs	Ts16	Ts16	Partial dub(16)	Ts16	GSE12501
MEF776	MEFs	Ts13	Ts13	Ts13	Partial dub(13) ¥	GSE12501
MEF780	MEFs	Ts13	Ts13	Ts13 ¥	Partial dub(13) ¥	GSE12501
MEF666	MEFs	Ts13	Ts13	Ts13 ¥	Ts13	GSE12501
MEF1113	MEFs	Ts19	Ts19	Ts19 ¥	Partial dub(19) ¥	GSE12501
MEF836	MEFs	Ts1	Ts1	Ts1	Partial dub(1) ¥	GSE12501
HS237	hESC line	46,X, idic(X)(q21), del(X)(q21-qter)	dub(X)(p22-q21) del(X)(q21-q23)	dub(X)(p22-q21) del(X)(q21-q23)	dub(X)(p11.23) del(X)(q21.1)	Skottman 2005*
C57BL/10 J-T43H/+	Pachytene spermatocytes	t(16;17) dub(17)(~20-40Mb) Interferes with XCI	dub(17) (20-47Mb) Up-regulated X	dub(17) (20-47Mb) Up-regulated X ¥	dub(17) (27-34Mb) ¥ Up-regulated X ¥	GSE7306
hIPS18	hIPS18 (p=12)	47(X,Y) +12	Ts(12) ¥	-	-	GSE21243

* https://services.btk.fi/index.php?id=1174

3.12. Conclusions

This chapter has presented the DI.S.C.O. software application for the identification of genomic clusters of differentially expressed genes. This part of the project was mainly focused on the development of a user-friendly and powerful tool that integrates intuitive visualization of the transcriptional data with computational methods that can identify statistically significant genomic regions of aberrant transcriptional activity.

DI.S.C.O. aims to overcome the limitations that are present in the related tools as described in section 3.1.2 by providing an automated and intuitive analysis platform for the user (or semi-automated when it cannot be avoided). It can be, therefore, easily used as a first level quick and inexpensive test for the validation of the genomic integrity of many published datasets, especially in biological fields where chromosomal aberrations are frequent, such as the case of pluripotent stem cell lines. The limitations addressed here could be potentially the reason why the scientific community has not yet routinely adopted this type of methods. In the following chapter, the extent of this widespread problem is going to be discussed in more detail.

Importantly, DI.S.C.O. is organism agnostic and platform independent as long as a genome annotation file is available for the specific organism and technology (tested with Affymetrix, Illumina and Agilent expression arrays as well as RNA-seq generated expression data). It can integrate and display different types of genomic features as additional custom tracks. It integrates and strengthens previously proposed methods (De Preter et al., 2008; Nilsson et al., 2008) and provides a framework for the comparison of novel methods, such as the NN method. The methods included in the tool have been extensively validated with both artificial and real biological datasets and their detection power and limitations have been discussed in section 3.11.

Finally, DI.S.C.O. is extended to the analysis of high-throughput sequencing data for the newly emerged next-generation sequencing technologies and it is implemented as a module in the GeneProf workflow engine (Halbritter et al., 2012).

4. Large-scale integrated search for aberrant transcriptional intervals in mouse pluripotent stem cells

This chapter presents an optimized methodology for the large-scale analysis of transcriptional data for the identification of recurrent chromosomal clusters of differentially expressed genes across a collection of mouse pluripotent stem cells.

4.1. Introduction

4.1.1. Background

As it has been already discussed in chapter 2.2, the accumulation of chromosomal imbalances in pluripotent stem cell populations is a frequent phenomenon, closely linked to the process of culture adaptation (Baker et al., 2007; Draper et al., 2004; Enver et al., 2005). The majority of the karyotypic validation of pluripotent stem cell lines has been performed by the use of conventional cytogenetic techniques or array-based methods such as aCGH and SNP genotyping. However, it has been already shown that this type of analysis can be also done at the transcriptional level due to the high correlation between gene dosage and gene expression that a high percentage of the genes in the affected region demonstrate (Pollack et al., 2002; Hyman et al., 2002; Hertzberg et al., 2007). Two recent studies were based on this principle in order to perform transcriptional karyotyping in human pluripotent and multipotent stem cells (Mayshar et al., 2010; Ben-David et al., 2011).

Interestingly, recurrent chromosomal aberration often map to specific genomic locations which correspond to chromosomes 12, 17, 20 and X in human (Draper et al., 2004; Baker et al., 2007; Inzunza et al., 2004; Mitalipova et al., 2005; Amps et al., 2011) and chromosomes 8 and 11 in mouse (Liu et al., 1997; Longo et al., 1997; Sugawara et al., 2006). Recently, it has been also shown that iPSCs tend to demonstrate a compromised genomic integrity during reprogramming and after establishment of the cell line in culture (Mayshar et al., 2010; Laurent et al., 2011; Hussein et al., 2011; Gore et al., 2011). It has been suggested that specific aneuploidies may tend to recur because

they confer a selective growth advantage to the cells *in vitro.* Selection for particular genomic changes could occur for many different reasons, namely because they confer resistance to apoptosis, enhanced self-renewal through loss of cell-cycle control or limited differentiation capacity (Harrison et al., 2009). When aneuploidies are present in a rapidly dividing self-renewing cell in a selective environment, the affected cells can potentially outgrow normal cells and eventually dominate the cell population. Consistent with this hypothesis, it has been demonstrated that mouse ESCs with a trisomy 8 outgrow normal cells with a diploid karyotype in competitive cultures (Liu et al., 1997).

This chapter presents a large-scale integrated approach for the identification of recurrent abnormalities in pluripotent stem cell populations using gene expression data. The significant advantage of using transcriptional profiling data for this type of analysis lies in the wealth of available datasets in public repositories that can be readily downloaded and re-analysed. In addition, the detection of these patterns at the transcriptional level can also potentially reveal candidate genes that drive the selection process. While gene expression analysis has been used to predict aneuploidies in human ESCs and iPSCs as well as other adult stem cell types this is the first comprehensive study in mouse to date. Therefore, it is important for both understanding data for this model system and also offers the opportunity for comparative analysis to existing human data.

4.1.2. Related Approaches and Challenges

Recently, Mayshar *et al.* (2010) used global gene expression meta-analysis in a collection of 66 hiPSC and 38 hESC samples from different passage numbers from 18 independent studies in order to assess the chromosomal integrity of human pluripotent stem cell lines. The study was also able to validate the results of the analysis by examining the corresponding DNA data that were available for 50% of the studies included in the meta-analysis.

At the first step of the analysis, Mayshar *et al.* (2010) obtained the FC values for each sample by dividing each gene with the median of the expression levels of the gene across the whole dataset. The expression of each gene was represented by a single

probe set (after random selection) that was Present in more than 80% of the samples. In addition, Mayshar and colleagues averaged profiles with similar expression levels after performing hierarchical clustering of the whole dataset. Then, for the identification of chromosomal enrichment of up-regulated genes only (>1.5 FC), the authors used two gene expression analysis software: Expander (Sharan et al., 2003) and EASE (Hosack et al., 2003) that both measure enrichment in chromosome cytobands or whole chromosomes. In order to further examine spatial patterns that do not necessarily overlap with predefined chromosomal intervals such as the cytobands, the authors used an application developed for aCGH data, CGH-explorer (Lingjaerde et al., 2004), using the program's piecewise constant fit (PCF) algorithm with a constant set of parameters (Least allowed deviation = 0.3; Least allowed aberration size = 80; Winsorize at quantile = 0.001; Penalty = 10; Threshold = 0.01).

Mayshar *et al.* (2010) identified high occurrences (6 hiPSC cell lines, 9% of samples examined) of chromosome 12 related aberrations (partial gains or full trisomies) and hypothesized that this could be due to the over-expression of the cluster of pluripotency genes *Nanog* and *Gdf3* at the 12q.

Ben-David *et al.* (2011) used the exact same methodology to interrogate human multipotent stem cells (208 samples of pluripotent stem cells, 144 samples of mesenchymal stem cells (MSCs), 97 samples of neural stem cells (NSCs) and 177 samples of hematopoietic stem/progenitor cells (HSPCs), from 58 independent studies). The authors were able to identify multipotent cell type-specific aneuploidies, recurring with a similar frequency (NSCs) or lower frequency (MSCs) than the one reported for pluripotent stem cells. These finding stress once more the necessity for rigorous karyotypic testing of pluripotent and multipotent cell lines.

Although, the results from the studies of Mayshar et al. (2010) and Ben-David et al. (2011) highlight an important problem, there are several limitations in their analysis pipeline. The use of a globally averaged expression profile can be limiting for the identification of subtler patterns of aberrations since it makes the assumption that a "normal" transcriptional stem cell state can be adequately represented by the global average of many different states. Given the heterogeneity of this type of data collections (derived from different labs, under different culture conditions and experimental designs) this assumption could potentially hinder the ability to reveal patterns of

interest. A detailed description of how the use of the global average expression profile may hinder the ability to predict aberrant regions follows in section 4.4.

Furthermore, both these studies assessed positional enrichment in predefined chromosomal intervals (cytobands or whole chromosomes) submitting gene lists of only up-regulated genes and using a constant threshold across the whole dataset, factors that could also be limiting for the analysis. The continuous change across the chromosome was investigated using an algorithm designed for aCGH data (CGH-explorer) which could identify highly affected regions but may lack the specificity to highlight milder changes. The CGH-explorer (Lingjaerde et al., 2005) is a moving window mean smoothing technique. The main criticism for mean smoothing is that it increases the signal-to-noise ratio with a cost of blurring the ends of the sequence which could result to over- or under- segmentation (Lai et al., 2005; Wineinger et al., 2008). In addition, the choice of an appropriate window size is not straight forward: larger windows result to smoother curves and possibly lower sensitivity, while smaller window sizes may introduce a large number of false positives (Chari et al., 2006). Finally, the method does not take into account the spatial distribution of probes on the chromosomes. It is not clear how the differences in gene density of different genomic regions could affect the prediction power of the approach.

The methodology proposed here addresses these potential problems by applying a sample-centric analysis with sample-specific thresholds in an iterative way across the whole dataset. In addition, the PGE algorithm (discussed in section 3.7.3) has been used, a method specifically designed for gene expression data and whose predictive ability has been demonstrated by extensive validation with both artificial and characterised biological datasets (section 3.11).

4.2. Data Collection and Data Normalization

The initial dataset of the present study consisted of 481 public domain gene expression data (373 ESC and 108 iPSC samples from 64 experimental designs) for the Affymetrix GeneChip Mouse Genome 430 2.0 Array from the Gene Expression Omnibus (GEO)[7] and ArrayExpress[8] public databases (the complete list of the analysed samples and their annotation is available as part of the supplemental material in the attached CD, Table S1).

The raw CEL files obtained from this table were globally normalized using the Robust Multiple-Array Average (RMA) (Irizarry et al., 2003a) and P/A flags were extracted by the MAS5.0 algorithm (Hubbell et al., 2002), both from the "affy" package of the Bioconductor suite[9] in the R statistical environment (Ihaka and Gentleman, 1996).

In order to identify differentially expressed genes, all the probe sets for which mapping positions were available have been considered (43109 probe sets). In addition, each gene was represented with a single value after averaging replicate probe sets with their median values, resulting to a final sum of 26930 probe sets.

[7] http://www.ncbi.nlm.nih.gov/geo
[8] http://www.ebi.ac.uk/arrayexpress
[9] http://www.bioconductor.org/

4.3. Establishing the appropriate baseline for FC values generation

4.3.1. The need for an appropriate baseline discovery method

Gene expression microarrays measure simultaneously the relative activity of every gene in the genome. In a typical microarray experiment, RNA is obtained from a control sample and a condition sample (which is by some criterion different from the control). The raw intensity values of each sample are then measured and compared in order to investigate the differences between these two conditions. The term baseline refers to the sample that is chosen to represent the control, that is the sample which establishes the cell state the deviation from which we wish to measure.

It is not uncommon that, in a single experiment type of analysis, the control sample is obtained from the parental cell line while the condition sample is derived from the same cell line after some kind of genetic manipulation. For example, a common experimental set-up is the creation of a conditional gene knock-out cell line in order to study the pathways affected by the lack of expression of the specific gene of interest. In this scenario, if the parental cell line relatively homogeneously carries a chromosomal aberration, there is a high chance that this aberration will be also present in the knock-out cell line. The relative change between the two samples will be unable to reveal the presence of the aneuploid chromosome. On the other hand, if the chromosomal imbalance only occurs in the derived knock-out cell line (either as a result of the genetic manipulation or because it is random but still counteracts and balances some process that is necessary for the cell and disturbed through the genetic manipulation), then it will be detectable with the above comparison.

So even if the validation of the samples at the transcriptional level is performed (with tools such as DI.S.C.O.), it is not necessarily the case that the underlying chromosomal imbalances would be discovered depending on the normalisation scheme used and the availability of karyotypically normal control samples. This problem can potentially deteriorate in large-scale integrated search approaches where a great number of

samples are analysed and it can lead to the under-estimation of the rate of recurrent aneuploidies in the examined dataset.

Ideally, samples from a cell line should be compared with a karyotypically normal instance of the same cell line in order to improve the detection power of any positional enrichment method. If such normal samples are not available, the next most transcriptionally similar sample can be used to reveal potential chromosomal imbalances manifested at the transcriptional level.

In order to address this problem, the present study uses a similarity measure of the transcriptional profiles of the different samples included in the analysis, and defines the comparisons to be performed using this similarity in an iterative way. The approach proposed is described in detail in the following section.

4.3.2. Hierarchical Clustering of Microarray Profiles

Firstly, the distance matrix of the different samples was calculated using the Pearson correlation. The Pearson correlation gives a measure of the similarity (linear dependence) of two variables, which is represented with a value ranging from -1 to 1. Typically, values between 0.7 and 1 denote high correlation (or -1 to -0.7 for high anti-correlation) while a value close to zero denotes no correlation between the two variables. Pearson's correlation coefficient is defined as the covariance of the two variables divided by the product of their standard deviations:

$$r_{X,Y} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{(N-1)S_X S_Y}$$

The distance matrix based on Pearson correlations has been used to perform agglomerative hierarchical clustering with average linkage in order to obtain a measure of similarity between samples and subsequently groups of samples (Figure 4.1).

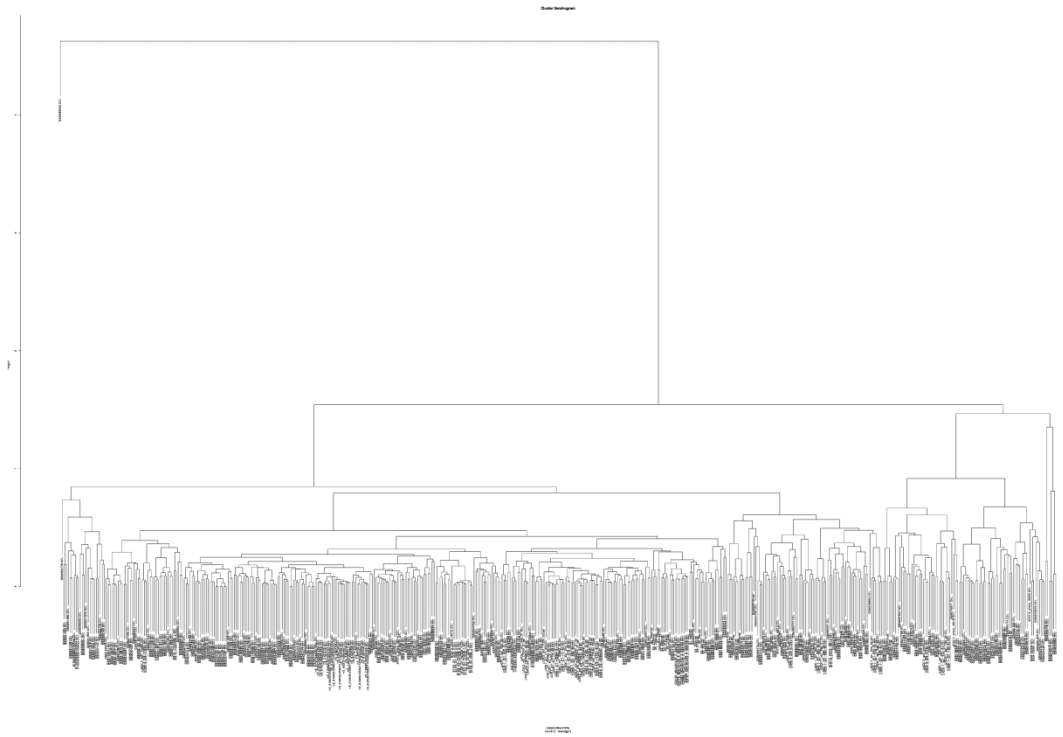

Figure 4.1 The resulting dendrogram after the agglomerative hierarchical clustering of the 481 samples included in the study.
The precise structure of the dendrogram of the present analysis can be found in Table S2 in the Supplemental Material provided in the attached CD.

The lower level nodes of the dendrogram in Figure 4.1 (referred to as leaf-nodes) represent the different samples, while the remaining nodes represent the clusters to which the samples belong according to the distance matrix. Clusters are formed by joining individual leaf-nodes or existing clusters and each joining point is a node. Each node (except the leaf-nodes), consists of a right and left sub-branch of clusters. In addition, the y-axis represents the distance between samples or clusters of samples in an incremental fashion. Therefore, samples that are highly correlated will have a low distance (close to zero) and will be represented by nodes nearer the bottom of the dendrogram. While clusters become gradually dissimilar, their distance is higher and they will appear in the upper levels of the graph.

4.4. Dendrogram-based Positional Enrichment Analysis

The dendrogram produced by the hierarchical clustering of the samples can be used to define the normalisation scheme (the comparisons to be performed), between samples and groups of samples. **Error! Reference source not found.**A shows a schematic representation of the branch-wise iterative comparisons logic. Firstly, comparisons are being performed at the first level of the dendrogram (at the leaf-nodes). The leaf-nodes consist of pairs of samples that are most similar to each other within the whole data matrix. In the majority of the cases, these will be replicate samples or samples from the same cell line cultured in a similar way. The median across all samples consists of the trim mean (0.05% of outliers) of the raw expression values of each gene. These values are used in order to identify the direction of the aberration (gain or loss). In the case of gain, for example, we would expect a higher number of genes in the affected sample to have higher raw expression levels than their trim mean values. This step is necessary since it is otherwise impossible to determine if the identified cluster is the result of a gain in the condition sample or a deletion in the control sample (Figure 4.2B**Error! Reference source not found.Error! Reference source not found.**).

Figure 4.2Bshows a scenario that will result in the identification of the enriched cluster. In this scenario, the identified cluster can be either a gain in Sample 1 (the condition, red) or a loss in Sample 2 (the control, green). The region will be identified but the type of the cluster (under the assumption that this is a genomic aberration) can be only assigned after comparing the trim mean expression of each specific gene in the cluster with the expression levels of the respective gene in the specific sample. For this assignment, the hypergeometric distribution (see section 3.1.2) is used in order to identify whether the cluster is enriched in up-regulated or down-regulated genes. The two probabilities are calculated and the type of cluster is assigned according to the lowest hypergeometric probability.

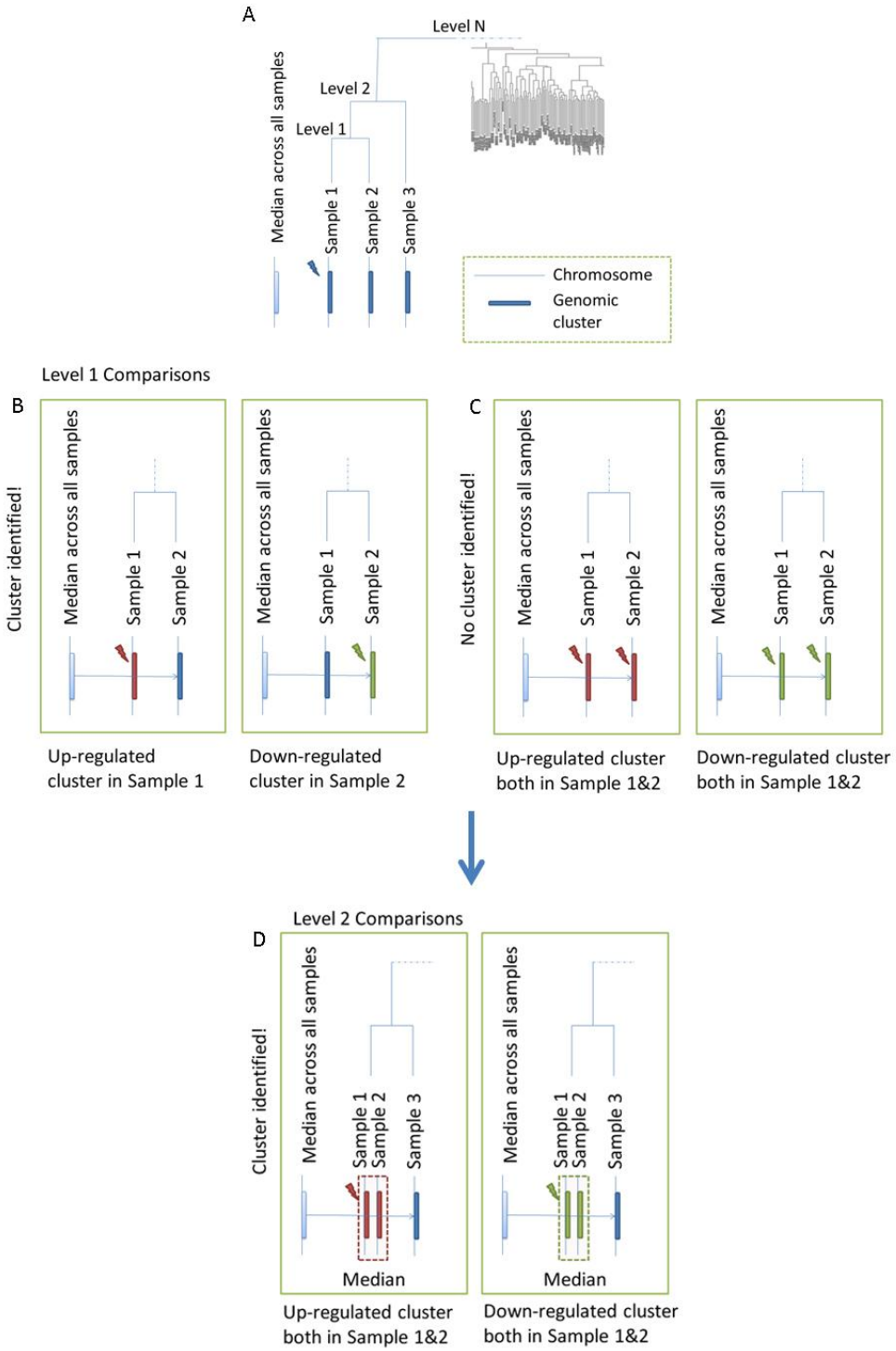

Figure 4.2 Dendrogram based positional enrichment. A) The branch-wise iterative comparison design. B) Level 1 comparisons of the dendrogram where a cluster is present at one sample. The chromosomal cluster can be identified since it appears in the one of the two samples being compared. C) Level 1 comparisons of the dendrogram where a cluster is present in both samples. The chromosomal cluster cannot be identified since it appears in both samples being compared. D) Level 2 comparisons of the dendrogram. The chromosomal cluster can now be identified since the median of the two abnormal samples is being compared with a normal sample.

Figure 4.2C, on the other hand, presents a scenario where the cluster is present in both samples (condition and control) and it cannot be identified at the first level comparisons step since it does not result in a relative change of the FC values of the samples being compared. However, as it is depicted inFigure 4.2D, the cluster will be discovered at the next level comparisons, where the median of the two samples is being compared to the next most similar sample that does not bear the specific aberration. As a result, it is possible now to correctly predict the presence of the underlying gain or loss in both samples 1 and 2. By iterating across the whole dendrogram in a branch-wise manner, it is feasible to obtain, in a sample specific manner, an accurate estimate of the patterns of aneuploidy present.

For this analysis, the PGE method has been used with an average significant ratio of 2.5 (discussed in section 3.7.3). The PGE algorithm is the method of choice since it was able to predict correctly the majority of patterns during the validation with previously characterised biological datasets (see section 3.11.2). In addition, the Multi-Level Otsu method has been used for the calculation of the sample-specific thresholds in each branch-wise comparison (average of thresholds used across the hierarchical tree: >1.56 FC and <0.66 FC). In Figure 4.3, the enhanced detection power of the methodology is presented. A single sample (GEO ID: GSM517044) is firstly compared to the global trim mean (panel A) and secondly, to the group of samples indicated by the dendrogram (GSM517041, GSM517042, and GSM517043 - series GSE20576) (panel B). The up-regulated cluster across chromosome 8 can be partially identified in (A) but it becomes much prominent in (B) while an up-regulation of the first part of chromosome 1 is also revealed which was not identified in (A). The distance of the transcriptional profiles between sample GSM517044 and the global trimmed mean obscures the ability to evaluate the transcriptional profile of GSM517044 in (A). Therefore, this approach can reveal the unique subtle changes of each sample that differentiate it from its most similar neighbour(s). It deviates from previous methodologies (Mayshar *et al.*, 2011; Ben-David *et al.*, 2011) in that it avoids the use of a globally averaged profile as a definition of a “normal” stem cell state to represent complex stem cell expression patterns (also discussed in section 4.1.2).

Figure 4.3 Improved detection of the proposed method.
An example of the enhanced detection power of the approach which is based on the "branch-wise comparisons" across the dendrogram. (A) Visualization of the transcriptional profile from sample GSM517044 (series GSE20576) when compared to the global Trimmed Mean (Chromosomes 1-10). (B) Visualization of the transcriptional profile from sample GSM517044 when compared to samples GSM517041, GSM517042, and GSM517043 (series GSE20576) as indicated by the hierarchical clustering derived dendrogram. The comparison is presented in Table S2 of the Supplemental Material in the attached CD (AHCL level = 99).

4.5. The large-scale integrated search workflow

Figure 4.4 summarizes the proposed methodology, which consists of the following steps:

(A) Global normalization of 481 public samples using RMA.
(B) Pearson correlation derived distance matrix and agglomerative hierarchical clustering with average linkage of the normalised data.
(C-D) PGE analysis with Multi-Level Otsu thresholding for identification of recurrent aberrant localized expression across the dendrogram.

In addition, for each comparison of the samples, the PGE algorithm corrects the enrichment p-value for multiple testing using False Discovery Rate (FDR) (Benjamini and Hochberg, 1995). An additional adjustment to the p-values to correct for multiple testing over the whole dendrogram was performed using 1,000 random permutations of data generated from randomised genomic mappings of the FC values. The resulting clusters were filtered for adjusted p-values lower than 0.05 and size greater than 10 genes.

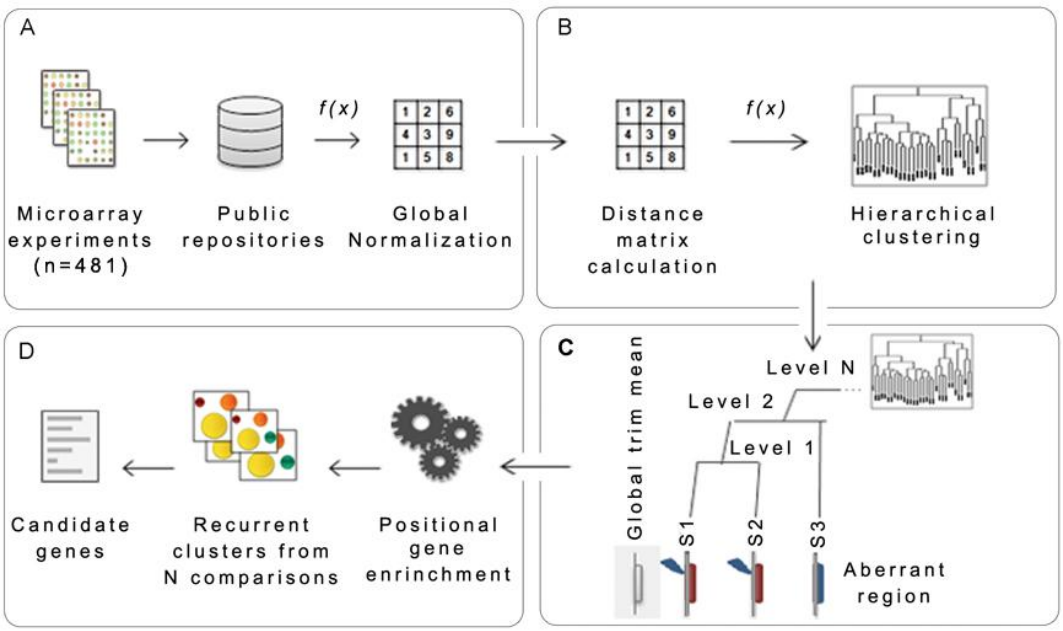

Figure 4.4 The large-scale integrated analysis workflow

4.6. Results

4.6.1. A Catalogue of Predicted Aberrant Intervals in Mouse ESCs and iPSCs

The PGE analysis of the clustered gene expression samples generated a large set of predicted regions of chromosomal change (the complete list is available as part of the supplementary files provided with the attached CD, Table S2). The most prevalent recurring intervals that have been observed map to chromosomes 6, 8, 11, 14 and most commonly in chromosome X (Figure 4.5, constructed using Circos (Krzywinski et al., 2009)). It is plausible that a percentage of the observed clusters on chromosome X correspond to varying states of X chromosome inactivation (XCI), while others to DNA copy number (CN) alterations. However, it should be noted that in mouse ESCs, all lines for which annotation was available (~70%) were annotated as male. Larger intervals covering extensive regions of the autosomes are likely to reflect underlying aneuploidies since co-regulation of large genomic regions is not commonly observed as the result of transcriptional regulation. Interestingly, 75.43% (binomial, p-value=3.8E-50) of the identified intervals are predicted gains, which implies that amplifications or activation events are much more frequent than deletions or coordinated down-regulation. A strikingly similar percentage of copy number variations (CNVs) in human ESCs has been reported to correspond to amplifications (72%) (Narva et al., 2010). The tendency towards a higher frequency of gains rather than losses in pluripotent stem cell lines has been also reported by Amps et al. (2011) where 39 ES cell lines (~31%) demonstrated gains of chromosomal material versus 20 lines (~16%) that showed a loss. In addition, Laurent et al. (2011) has suggested that time in culture correlates with duplications of oncogenes.

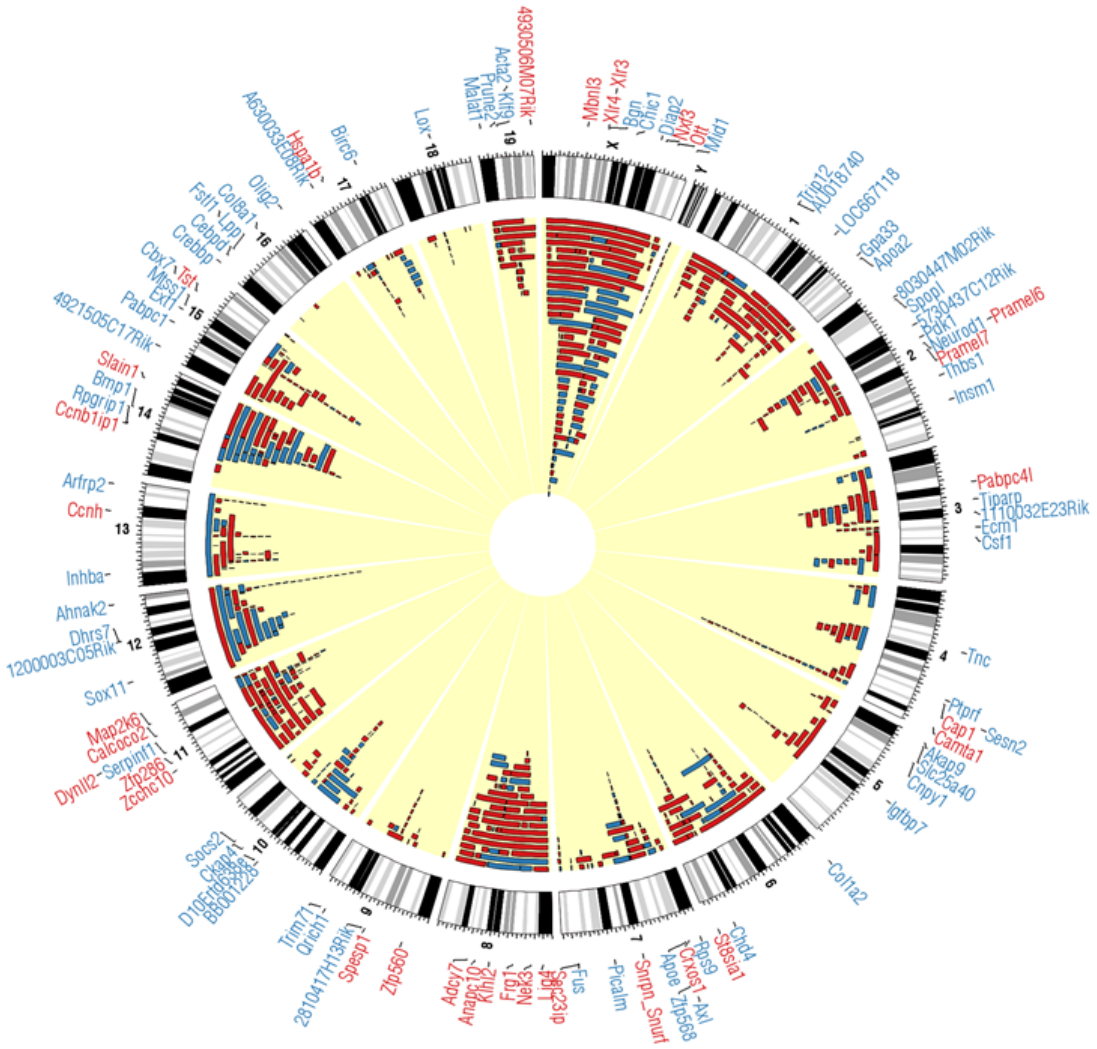

Figure 4.5 Identified aberrant intervals presented on a circular karyotype.
The circular karyotype of all predicted significantly over-expressed (red) and under-expressed (blue) intervals in the matrix and the genes that are differentially expressed between predicted normal and aberrant samples (red for up-regulated genes and blue for down-regulated). Larger effects observed in chromosomes 8, 11, 14 and X. Figure generated in Circos (Krzywinski et al., 2009).

In addition, Figure 4.6 (constructed using Circos (Krzywinski et al., 2009)) gives a graphical overview of the chromosomal intervals that seem to appear simultaneously aberrant. In the circular karyotype plot, each line connects two different genomic regions that have been predicted abnormal in the same sample. Therefore, with this type of representation, it is possible to summarize the complex patterns of recurrence of the concordantly aberrant chromosomal intervals across the whole dataset.

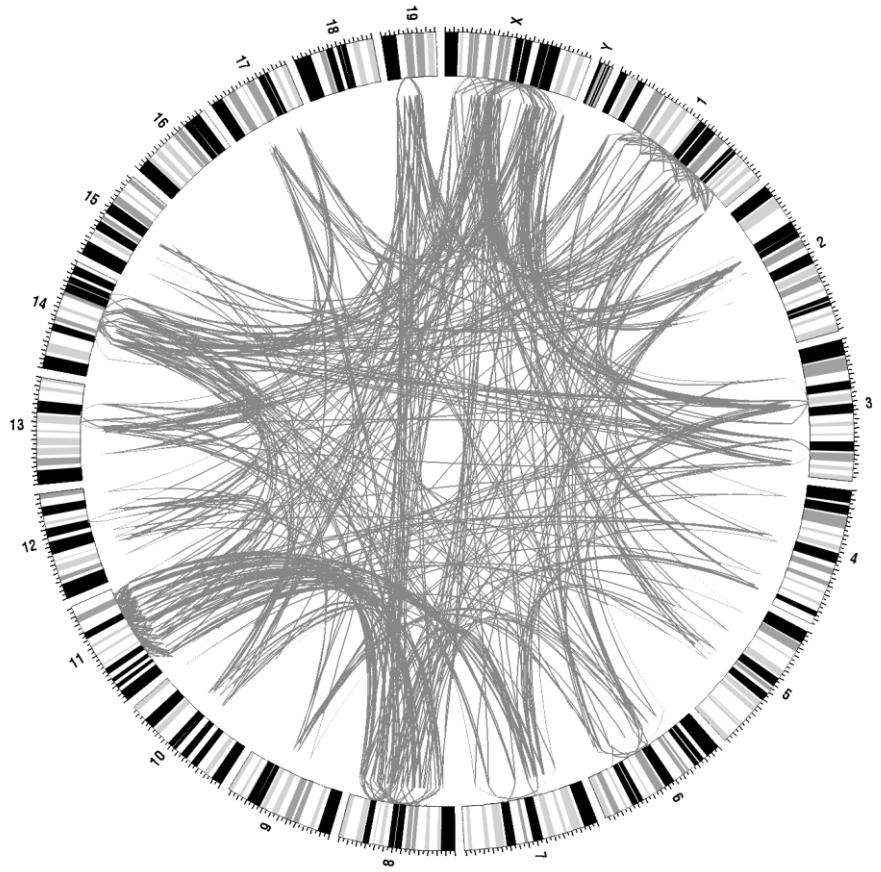

Figure 4.6 A connectivity map of jointly re-occurring clusters with major hubs at the chromosomes 1, 8, 11, 14, 19 and X. Figure generated in Circos (Krzywinski et al., 2009).

4.6.2. Expression of pluripotency markers

The PGE analysis resulting clusters were visually inspected using DI.S.C.O. and samples with whole- or partial-chromosome spanning clusters were classified as "Aberrant". The rest of the samples were categorized as "Normal" even if they included small clusters since the genomic or transcriptional origin of smaller aberrant intervals cannot be distinguished at the transcriptional level.

By examining the expression levels of hallmark pluripotency genes such as *Nanog, Pou5f1* (*Oct4*) and *Sox2* (Chambers et al., 2007; Silva et al., 2009; Nichols et al., 1998; Niwa et al., 2000; Schöler et al., 1990; Nichols et al., 1998; Avilion et al., 2003), it was possible to reveal distinct differences of the expression levels of these markers between the Normal and the Aberrant groups of samples.

In Figure 4.7, samples of the two groups (Normal and Aberrant) have been plotted according to the *Nanog* expression levels in an ascending order. It becomes clear that the Aberrant group largely consists of samples with *Nanog* high expression levels (86% of the samples) while the percentage drops to 47% in the Normal group. The expression of the pluripotency genes *Oct4* and *Sox2* correlates with the *Nanog* expression levels at the high expression range, with the exception of the group of samples at the far right of the graph where the *Oct4* suppression is induced by tetracycline after the targeted inactivation of the endogenous alleles (ZHBTc4 cell line) (Niwa et al., 2000).

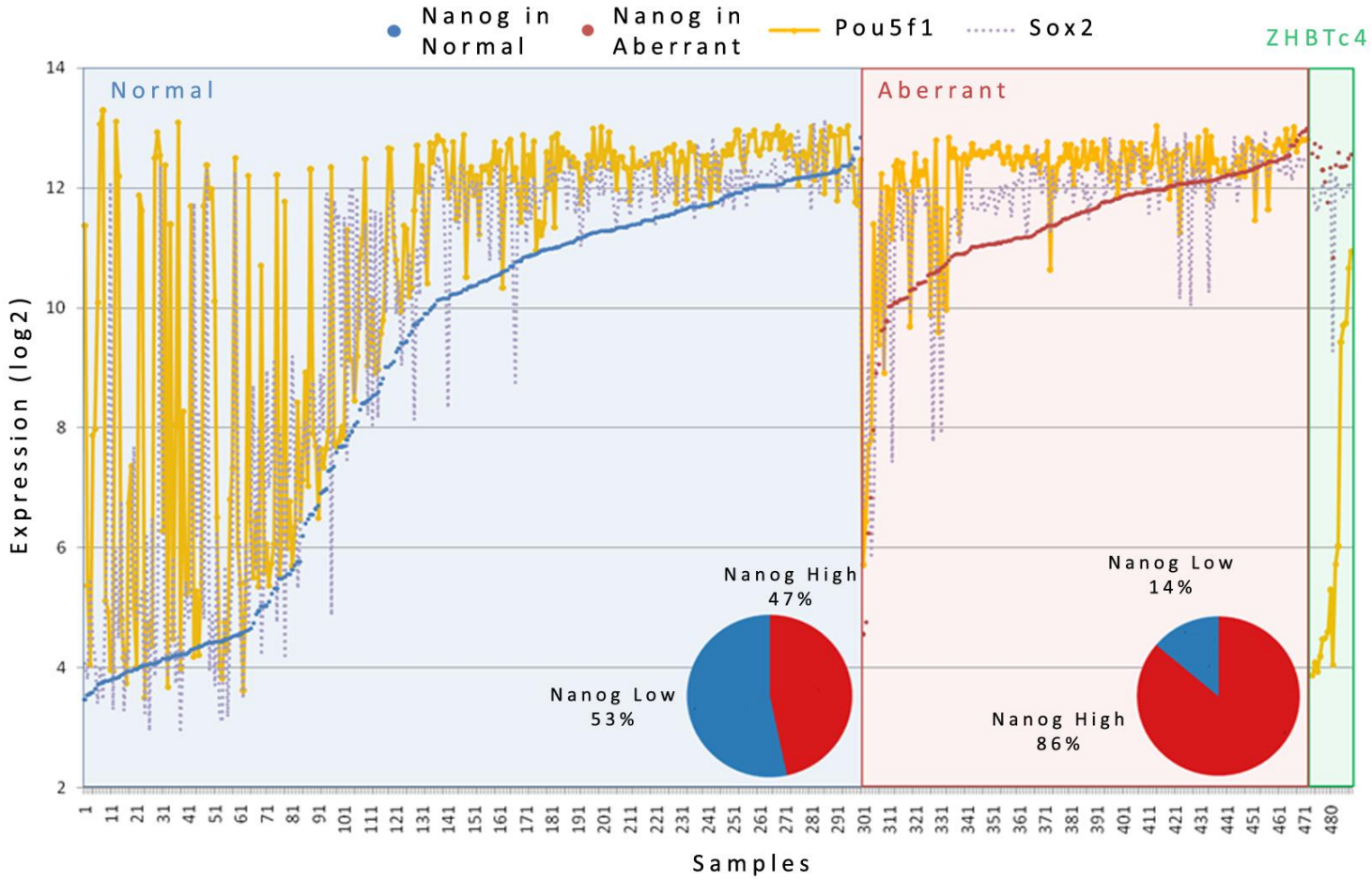

Figure 4.7 Expression levels of pluripotency markers (*Nanog*, *Pou5f1* (*Oct4*) and *Sox2*) in the normal and aberrant groups of samples.
Pie charts of the percentage of samples in each group that demonstrate high and low *Nanog* expression levels. The collection of samples at the far right section of the plot corresponds to samples where the *Pou5f1* (*Oct4*) expression was genetically engineered to be low (ZHBTc4 cell line).

By using a combination of the available annotation of the samples and the expression of the pluripotency markers *Nanog, Oct4* and *Sox2*, it was possible to identify that the majority of the samples with *Nanog* low expression levels are either differentiated cells, cells in various time-points of a differentiation protocol or partially reprogrammed cells. Even though *Nanog* expression is not necessary to sustain pluripotency (Chambers et al., 2007), it can be hypothesised that samples with low levels of *Nanog* expression potentially have higher percentages of differentiating or partially-reprogrammed cells. This was highly consistent with the obtained sample annotation. Given the highly unbalanced percentages of homogeneously pluripotent cell lines between the Normal and Aberrant groups of samples, the ability to detect the transcriptional changes due to aneuploidy can be obscured by the transcriptional changes linked to the differentiating signature in half the samples of the Normal group. In order to be able to distinguish between the differentiation signature and the aneuploidy signature, the samples included in the study were divided in two groups based on the criterion of *Nanog* expression levels: the *Nanog*-high group, which consists of pluripotent stem cell populations (315 samples) and the *Nanog*-low group, which consists of their differentiating or partially-reprogrammed counterparts (166 samples).

The extent of aneuploidy in the *Nanog*-high group and the *Nanog*-low group is presented in Figure 4.8. The percentage of aberrant samples in the *Nanog*-low subset is much lower (30%) than the ones in the *Nanog*-high subgroup (57%). This difference may reflect differences in the frequency of pluripotent cells in cultures or the ability to detect these subtle signatures in mixtures of differentiating cells.

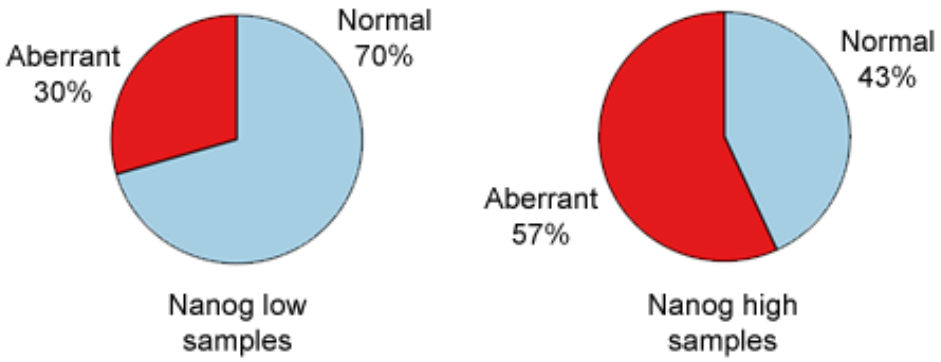

Figure 4.8 Percentages of aneuploid samples in the sub-groups of *Nanog*-high and *Nanog*-low samples.

4.6.3. Comparison to published cytogenetic studies in mESCs

A detailed discussion of the findings of the various studies that have assessed the genomic integrity of mouse ESCs in culture has been presented in section 2.2.4. The small number of available studies assessing systematically the phenomenon of culture adaptation in mouse pluripotent stem cells is however surprising. The majority of these studies have reported an alarming range of chromosomal abnormalities (Liu et al., 1997; Longo et al., 1997; Guo et al., 2005; Sugawara et al., 2006; Rebuzzini et al., 2008a; Rebuzzini et al., 2008b; Liang et al., 2008). From these studies, only (Liu et al., 1997) and (Sugawara et al., 2006) included a high number of cell lines (29 and 88 respectively) and provided the detailed karyotype of the analysed cells. Frequent genomic changes seem to be non-random and mainly involve chromosomes 8, 11 and X (Liu et al., 1997; Longo et al., 1997; Sugawara et al., 2006). The study of Guo et al. (2005) has also implicated chromosome 6 and 14.

Figure 4.9 presents a comparison between the present study and the studies of Liu et al. (1997) and Sugawara et al. (2006). These studies were chosen because they provide the most comprehensive analysis to date. The percentage of abnormal karyotypes (any type of aberration) is 57% for the current study, 75.86% in Liu et al. (1997) and 39.77% in Sugawara et al. (2006). Trisomy 8 is the most frequent chromosomal change in all three studies and trisomy 11 is the second most frequently observed aneuploidy. Instances of chromosomal aberrations in chromosomes 6 and 14 were much more predominant in the present study but they constitute partial chromosomal gains or losses rather than trisomies or whole chromosome losses. The identification of frequent patterns on chromosomes 6 and 14 could reflect the ability of the proposed method to detect subtler changes. Nonetheless, high percentages of chromosome 6 and 14-specific aberrations have been also identified by multicolour FISH by Guo et al. (2005) (10/10 metaphases had a 14q duplication and 11/11 a loss of 6q in the E14.1 and W9.5 ES cell lines respectively).

In the case of the sex chromosomes, the instances of chromosome X specific aneuploidies identified by the present study could reflect different states of XCI between different ESC samples. Alternatively, it is possible that the detected up-regulation of chromosome X corresponds to differences between female cell lines with

two active X chromosomes and male ES cell lines. It is worth noting however that the majority of ES cell lines for included in the study have been annotated as male (~70% of lines for which annotation was available). It has been reported previously that the derivation of male ES cell lines is more efficient and they maintain better genomic stability than female ES cell lines (Robertson et al., 1983; Huynh and Lee, 2003). Information for the sex chromosome constitution of the iPSC samples examined here was not readily available from the annotation.

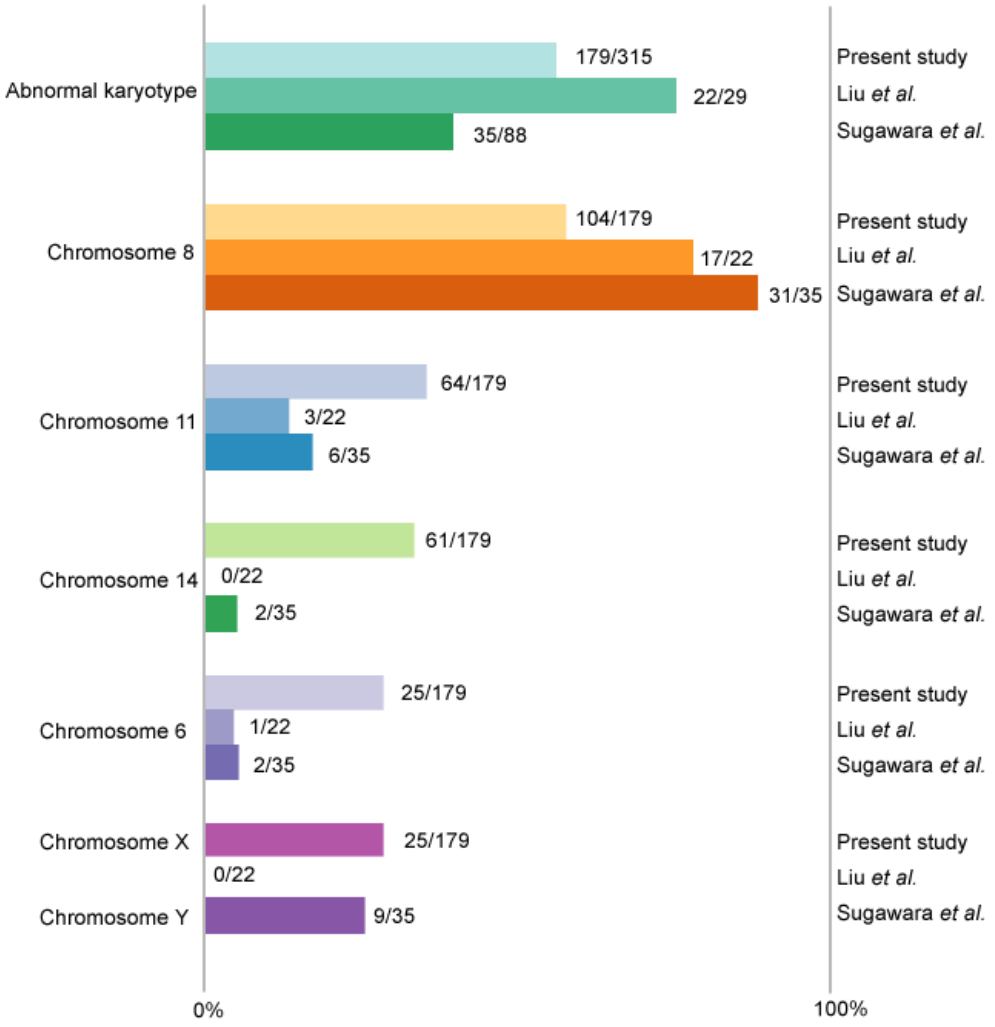

Figure 4.9 Comparison to published large-scale cytogenetic studies.
Comparison of the frequencies of predicted abnormalities per chromosome in the present study and two independent cytogenetic studies of mouse ESCs (Liu et al., 1997; Sugawara et al., 2006).

Growth advantage of ES cells in culture has also been linked with sex chromosome specific aneuploidies (Sugawara et al., 2006; Robertson et al., 1983). In the study of

Sugawara et al. (2006), 9 of the 35 aberrant cell lines were XO with 2 of the 9 lines containing a small sub-population of XY karyotypes, 1 of XX karyotypes and the rest were homogeneously XO. Y-chromosome loss has also been reported at an average frequency of 2% of subclones from 40,XY ES cell lines (Eggan et al., 2002). Cell lines are typically sexed by PCR analysis of the male-specific *Sry* gene on chromosome Y (Lambert et al., 2000). A similar analysis at the transcriptional level was not possible since *Sry* transcripts were only detected in very low levels in the ESCs of this study. The expression of four more genes transcribed from Y chromosome was analysed in order to investigate the possibility of a chromosome Y loss. These were the genes *Ddx3y* (*Dby*), *Eif2s3y*, *Ube1y1* and *Uty* located at the male-specific region of the gene-poor mouse Y chromosome. Their expression levels in the annotated as male only ES cell lines (221 samples) was analysed and compared to each gene's trim-mean expression (0.05% of values removed). Since the average percentage of XO male ES cell lines has been reported to be around 2% (Eggan et al., 2002), it can be hypothesised that the trim-mean across the male sub-group will be representative of the baseline expression of these genes in male samples. Samples were divided in three categories: up-regulated Y chromosome, down-regulated Y chromosome and no change in Y chromosome, based on the expression of the four genes. A sample was assigned to the one of the two aberrant categories if more than two of the four Y-specific genes were differentially expressed when compared to their trim-mean value (>1.5 FC).

The analysis of the occurrences of a deregulated Y chromosome (over- or under-expression of more than half the Y-linked genes examined) revealed that 28.96% of the male samples show a down-regulation of Y-chromosome while 12.67% show an up-regulation. Interestingly, an up-regulated Y-chromosome coincides in 76.92% of the samples with a trisomy 8 (p-value<0.01, hypergeometric). This could indicate a regulatory role of elements on chromosome 8 that can affect the expression of the chromosome Y-specific genes. In the study of Sugawara et al. (2006), 6 out of the 9 XO lines had also a trisomy 8. In the present study, 25% of male ES cell lines with predicted trisomy 8 showed a down-regulation of chromosome Y (p-value<0.01, hypergeometric). Figure 4.10 displays a heatmap representation of the expression levels of the four Y-linked genes examined in combination with the presence of a predicted chromosome 8 – specific aneuploidy. However, there is no way to conclusively infer whether the deregulation of the Y-chromosome linked genes is a result of the underlying loss or gain of the Y chromosome or a transcriptional

regulation event. These preliminary data could be further investigated only with the use of genomic techniques such as FISH chromosome labelling, PCR analysis or banding.

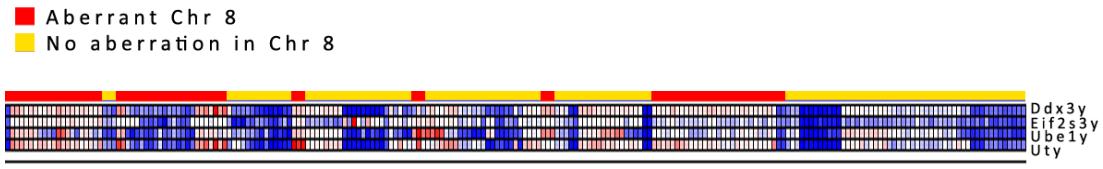

Figure 4.10 Comparative analysis of Chr8 aberrations and expression levels of four Y-linked genes. Heatmap representation of the expression levels of the four Y-linked genes *Ddx3y* (*Dby*), *Eif2s3y*, *Ube1y1* and *Uty* in male mouse ES cell lines in relation to the existence of a predicted chromosome 8 specific aneuploidy.

4.6.4. Chromosomal breakdown of patterns

The aberrant clusters from the PGE dendrogram-based analysis described in section 4.4 were visually inspected using DI.S.C.O. and statistically significant large whole- or partial-chromosome spanning clusters were further examined. It is reasonable to hypothesise that large expression domains are characteristic of underlying aneuploidy rather than transcriptional regulation. Large-region epigenetic silencing has, however, been reported in cancer (the largest region observed thus far has been reported by Frigola et al. (2004) which consisted of a 4 Mb methylated region in colorectal cancer). Figure 4.11 presents the chromosomal breakdown of the resulting clusters for the *Nanog*-high subgroup of 315 ESC and iPSC samples specifically. 179 samples (56.83%) of the *Nanog*-high subgroup contained whole or partial-chromosome spanning clusters while the percentage of aberrant samples in the *Nanog*-low subset is much lower (30%) as already discussed in section 4.6.2 (Figure 4.8).

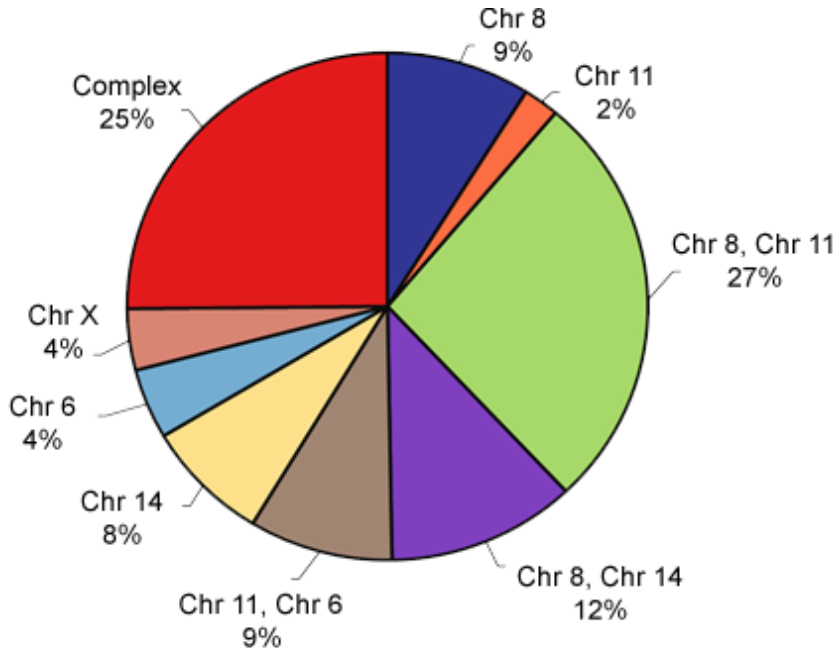

Figure 4.11 Breakdown of percentages for the aberrant chromosomes and the associated aberrant chromosome pairs for the *Nanog*-high subset of 315 pluripotent cell lines.

In Figure 4.11, chromosome 8 is most often affected either solely (9% of samples) or in combination with chromosomes 11 and 14 (27% and 12% of aberrant samples respectively). Chromosomes 8 and 11 are the chromosomes most frequently affected by aneuploidy and, in fact, 70% of the predicted aberrant samples carry whole or partial-chromosome aberrations on at least one of these two chromosomes.

Frequently recurring pairs of predicted chromosome anomalies which tend to recur across many different experiments were also discovered. These include chromosomes 8 and 11 (hypergeometric, p-value=0.001), chromosomes 8 and 14 (hypergeometric, p-value= 3.20E-06), chromosomes 11 and 6 (hypergeometric, p-value= 0.019) and chromosomes 14 and 17 (hypergeometric, p-value= 4.00E-11). A detailed breakdown of recurring aberrant chromosomal pairs for the *Nanog*-high subgroup is presented in Figure 4.12.

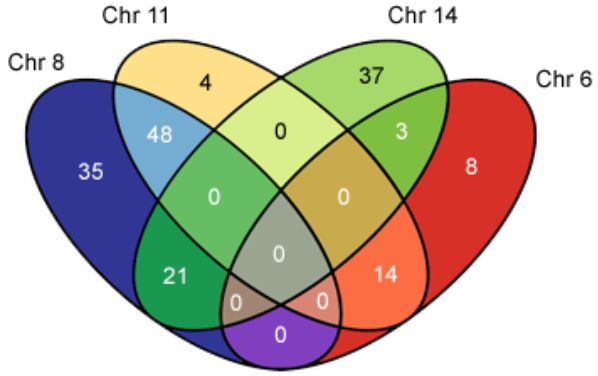

Figure 4.12 Venn-diagram representing the co-occurrence of aberrations between chromosomes 6, 8, 11 and 14.

Finally, a detailed comparison between ESC and iPSC-specific aberrations of the autosomes revealed that in both cases more than half of the samples are predicted to carry one or more chromosomal aberrations (Figure 4.13). However, chromosome 11 patterns are mostly present in ESCs. Chromosome 8 and 14, in contrast, account for a large number of aberrations in both populations. In iPSCs, the chromosome X changes, which are predicted gains, could reflect differences between male and female lines such as different states of XCI. Unfortunately, annotation for the sex of the line for the majority of iPSC samples studies was not readily available online and thus, sex chromosomes have been excluded from this comparison.

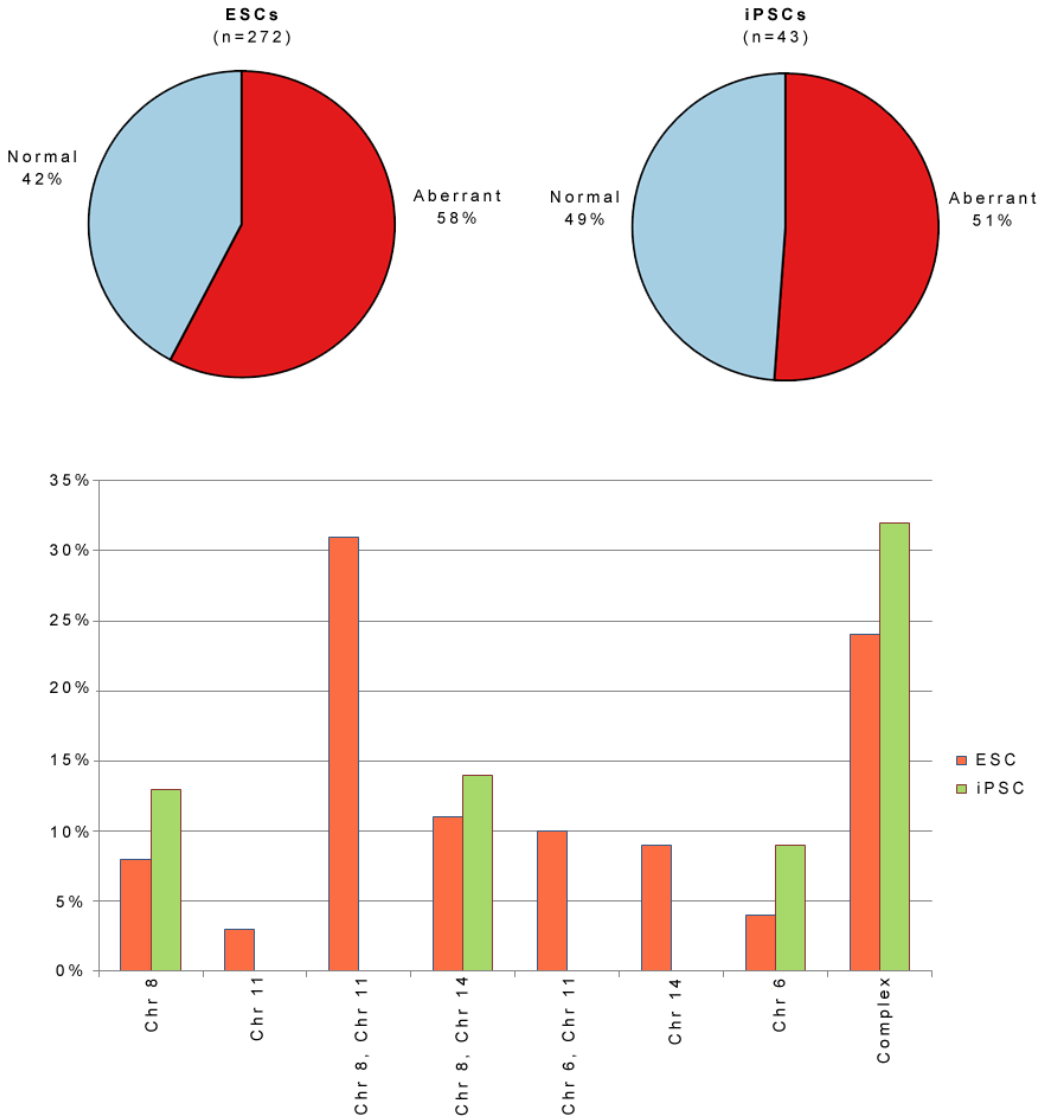

Figure 4.13 Genomic changes in ESCs versus iPSCs.
Detailed comparison between ESC and iPSC-specific patterns with overall percentages of predicted aberrant samples per group. The chromosomal breakdown of the aberrant samples per group is also presented.

4.7. Chromosome 14 & 17 complex recurring patterns

4.7.1. Identification of chromosome 14 and 17 recurring patterns

In one of the first aCGH studies performed in mouse ES cells, Hall described a complex pattern of aberration involving chromosome 14 and 17 (Hall, 2008). This pattern included a deletion on chromosome 14 of 36.9 Mb followed by an amplification of the distal part of chromosome 14 of around 28.1 Mb in the ZHBTc4.1 cell line. In addition, a small deletion in the chromosome 17qE1.3-qE3 locus was identified (11.3 Mb) in the same cell line. In this experiment, the ZHTBc4 cells were compared to the CGR8 male cell line (derived as described in (Nichols et al., 1990)). Hall reported one more chromosome 14 related amplification in the B6 PD1 cell line (wild-type C75BL/6 mouse derived ES cell line), namely an amplification of 14qD1-qE4 (aCGH analysis). Interestingly, a duplication of 14q has also been reported in W9.5 and trisomy 14 in E14.1 mouse ES cells after multicolour metaphase FISH analysis (Guo et al., 2005).

In Figure 4.14, the complex aberration of chromosomes 14 and 17 described by Hall is presented in DI.S.C.O (in-house gene expression data from the Affymetrix GeneChip Mouse Genome 430 2.0 Array). The relative gene expression changes between the ZHBTc4 and CGR8 cell lines is clearly depicted at the transcriptional level and the PGE algorithm identifies both the deletion followed by the amplification of the distal end of chromosome 14 as well as the deletion on chromosome 17 as indicated by the clustering output track (discussed in detail in section 3.6.2).

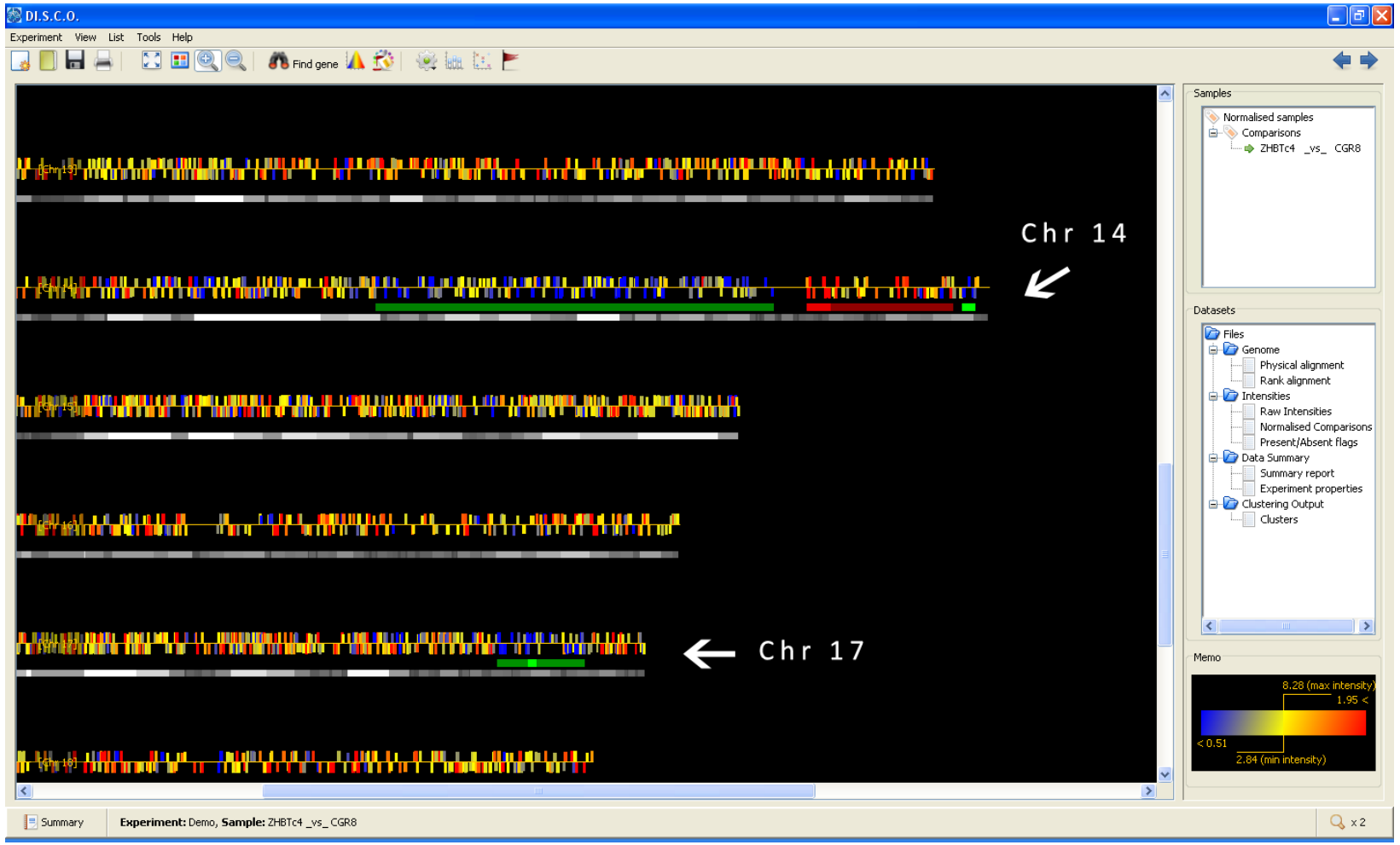

Figure 4.14 The ZHBTc4.1 cell line gene expression compared to the CGR8 mESCs in DI.S.C.O.
The complex aberration pattern in chromosomes 14 and 17 is highlighted by the output track of the PGE algorithm under the aberrant regions (green for down-regulation and red for up-regulation).

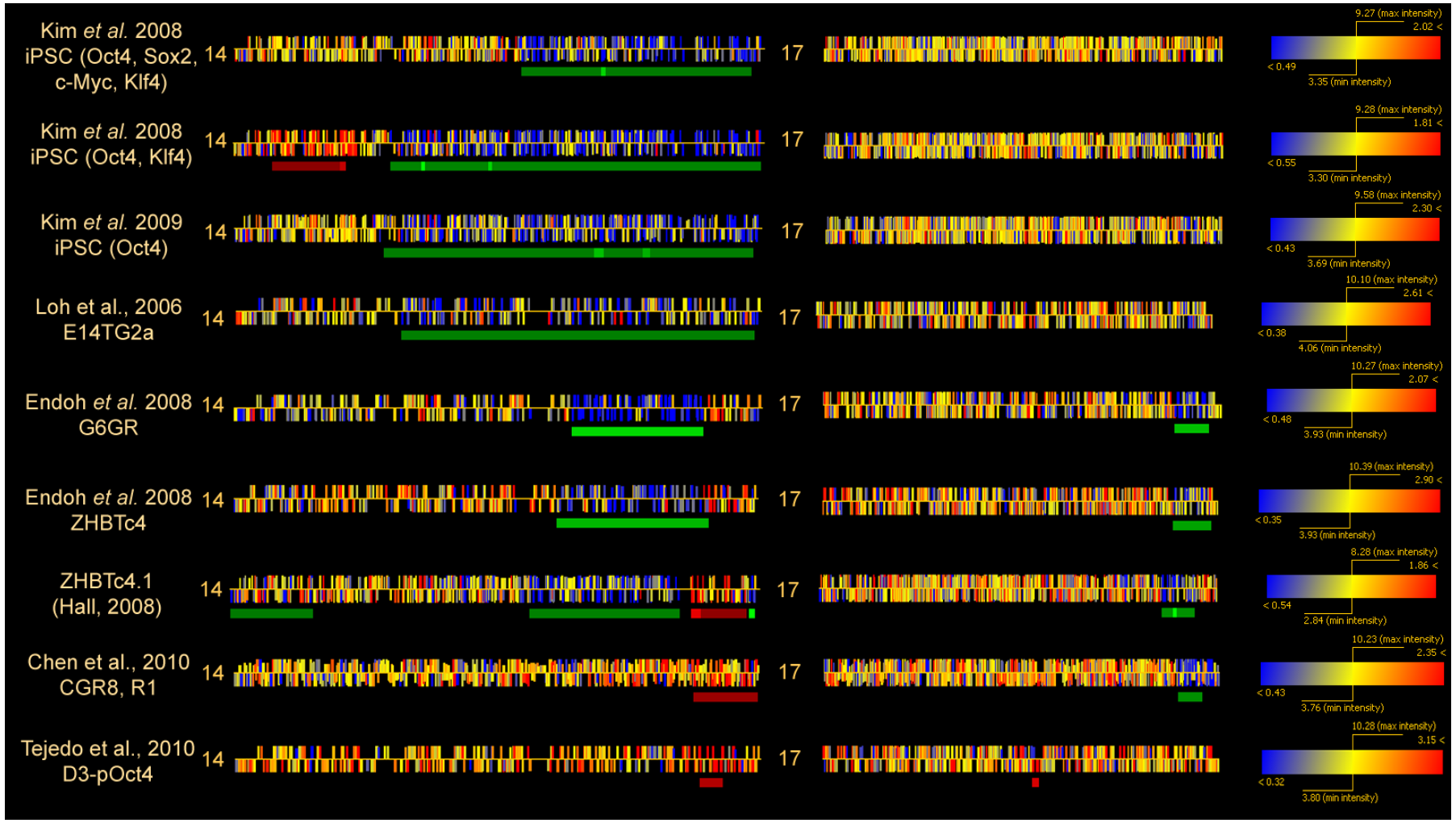

Figure 4.15 Recurring patterns of concordant changes in chromosome 14 and 17 uncovered by the-dendrogram-based PGE analysis.

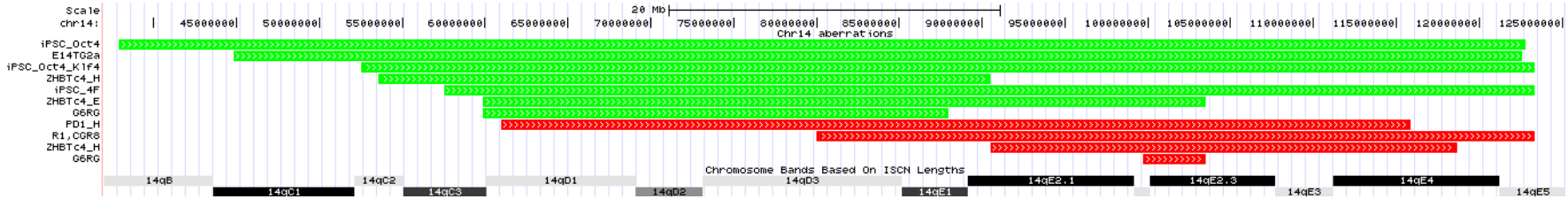

Figure 4.16 Mapping of the chromosome 14 specific aberrations across the specific genomic region of chromosome 14 (USCS browser).
Red tracks represent amplifications and green tracks represent deletions as indicated by the dendrogram-based PGE analysis. Samples included in the analysis: iPSC_Oct4 (Kim et al., 2009), iPSC_Oct4_Klf4 and iPSC_4Factors (Kim et al., 2008a), E14TG2a (Loh et al., 2006), R1,CGR8 (Chen et al., 2010), ZHBTc4_H and PD1_H(Hall, 2008), ZHBTc4_E and G6GR (Endoh et al., 2008).

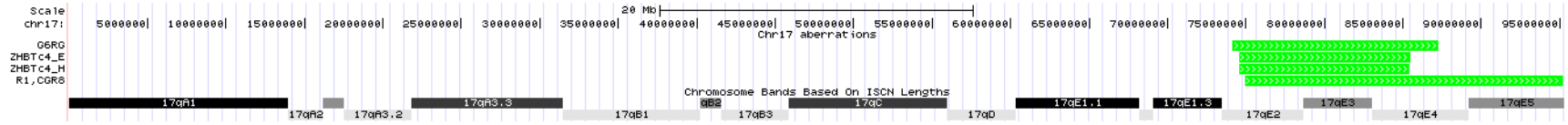

Figure 4.17 Mapping of the chromosome 17 specific aberrations across the specific genomic region of chromosome 17 (USCS browser).
Red tracks represent amplifications and green tracks represent deletions as indicated by the dendrogram-based PGE analysis. Samples included in the analysis: R1,CGR8 (Chen et al., 2010), ZHBTc4_H (Hall, 2008), ZHBTc4_E and G6GR (Endoh et al., 2008).

The recurrent chromosome 14 amplifications in two independent ES cell lines in the study of Hall (2008) and two more ES cell lines in the study of (Guo et al., 2005) were intriguing. Chromosome 14 specific aberrations were also frequent in the present study, commonly in combination with chromosome 17 aberrations. Therefore, it was intuitive to attempt to identify a common region that could possibly reveal the gene or genes under selection. In Figure 4.15, a summary of the samples with a predicted chromosome 14 and/or 17 aberration is presented. With the exception of the iPSC samples of (Kim et al., 2008a) and (Kim et al., 2009), the rest of the data are ES cell lines. In more detail, both Hall (2008) and (Endoh et al., 2008) used the ZHBTc4 cell line. In the ZHBTc4 cell line, *Oct4* expression is sustained from a tetracycline (Tc)–repressible transgene on an *Oct4* null background (Niwa et al., 2000). Upon Tc treatment, *Oct4* is rapidly silenced and cells start to differentiate. The G6GR cell line in (Endoh et al., 2008) is generated by random integration of the linearized Gata6-GR expression vector into E14TG2a-derived EB5 ES cells (Shimosato et al., 2007; Kawasaki et al., 2000; Niwa et al., 2000). Loh et al. (2006) also uses E14TG2a mouse embryonic stem cells transfected with *Nanog* RNAi or *Oct4* RNAi. Finally, Chen et al. (2010) uses two ES cell lines, the CGR8 and R1, but the specific origin of the sub-clones used is not provided.

Table 4.1 gives a detailed description of the ES cell lines with an identified chromosome 14 and/or chromosome 17 aberration and traces their parental cell lines. One initial observation is that the chromosome 14-17 specific pattern seems to be related to the CGR8 and E14TG2a cell lines and their genetically modified sub-clones, ZHBTc4 and G6GR. However, there is an instance of the R1 cell line carrying the exact same pattern in the study of Chen et al. (2010). No other chromosome 14/17 aberrations at the specific loci were discovered in the remaining R1 and CGR8 cell lines included in the present study. However, an overlapping amplification of chromosome 14qE1-qE2.3 (87816437-103476845) was also discovered in three subclones of the D3-pOct4 cell line (Tejedo et al., 2010).

Table 4.1 Description of ES cell lines carrying a chromosome 14 and/or a chromosome 17 aberration and their parental ES cell lines.

ES Cell Line	Description
CGR8	129/Ola-derived wild-type ES cells (Nichols et al., 1990).
D3	129/Sv blastocysts derived ES cells (Doetschman et al., 1985).
D3-pOct4	D3-derived ES cells by electroporation of the linearized p-Oct4-eGFP-pgk hygro plasmid (Tejedo et al., 2010).
E14TG2a	129/Ola-derived HPRT-negative ES cells (Hooper et al., 1987).
EB5	E14TG2a ES cells carrying IRES-BSD in one of *Oct4* locus, which allows selection of *Oct4*-positive undifferentiated stem cells (Niwa et al., 2000).
G6GR	EB5-derived ES cells by random integration of the linearized Gata6-GR expression vector (Shimosato et al., 2007).
R1	Hybrid of two 129 substrains (129X1/SvJ and 129S1/SV-+^{p}+$^{Tyr\text{-}c}$*Kitl*$^{Sl\text{-}J}$/+) (Nagy et al., 1993).
ZHBTc4	ZHTc6-derived ES cells maintained by tetracycline-regulatable *Oct4* transgene. Both of *Oct4* locus are disrupted by IRES-zeocin and IRES-BSD KO vectors (Niwa et al., 2000).
ZHTc6	CGR8-derived ES cells carrying tetracycline regulatable *Oct4* transgene and IRES-zeocin cassette in one of *Oct4* locus, which allow selection or elimination of *Oct4*-positive stem cells (Niwa et al., 2000).

Upon closer examination, the following observation can be made: the majority of ES cell lines that demonstrate the specific chromosome 14-17 aberrant pattern have at some point undergone a genetic manipulation of the *Oct4* locus by integration of an IRES-BSD or IRES-zeocin cassette (ZHBTc4 and G6GR), or the p-*Oct4*-eGFP-pgk hygro plasmid (D3-pOct4) with the exception of the CGR8 and R1 cell lines of Chen et al. (2010). In addition, the E14TG2a cell line used in the study of Loh et al. (2006) has been targeted with *Nanog* and *Oct4* RNAi. However, even the pSUPER-puro empty vector samples that have been used as a control in this study carry the deletion on chromosome 14, thus, it seems that the aberration had already been present before induction of RNAi.

Why the same complex pattern of chromosome 14-17 concordant aberration would appear in three different ES cell lines and their derivatives remains unclear. The presence of the pattern seems to coincide in some of the ES cell lines with genetic

manipulations of the *Oct4* locus which may have resulted in suboptimal levels of *Oct4* expression in a past time-point of the ES cell line (no other genetically engineered *Oct4* cell line was available in the present analysis). One hypothesis could be that this complex pattern of aberration confers an advantage to the cells that can potentially compensate for the low *Oct4* expression levels. This scenario might also explain the presence of the deletion of distal chromosome 14 observed in the iPSC cell lines. If such a hypothesis is true, it specifically implicates the region of deletion of chromosome 14 which is overlapping in all samples examined with the exception of the CGR8 and R1 cell lines of Chen et al. (2010). However, it does not explain the presence of the chromosome 14 amplification and/or the chromosome 17 deletion in parental cell lines such as CGR8, R1 and E14TG2a where no targeting of the *Oct4* locus has been performed.

Another possibility is that the deletion on chromosome 14 and the amplification on chromosome 14 coupled with the deletion on chromosome 17 are two independent events that happened to simultaneously occur in specific cell lines. The cell lines can be actually divided into three groups: the ones that carry only the chromosome 14 partial deletion (iPSC cell lines and E14TG2a), the ones that carry the chromosome 14 partial deletion followed by amplification at the chromosome's distal end and chromosome 17 partial deletion (ZHBTc4, G6GR, CGR8 and R1) and finally the ones that carry the chromosome 14 distal end amplification only (D3-pOct4).

It is also possible that some of the recurring pattern are mirroring a single event that occurred at the paternal cell line, which giving a selective advantage to the cells, propagated through serial passaging and outgrew the subsequent sub-clones or new cell line derivatives. This could be specifically the case for the ZHBTc4 cell line, which carries the exact same pattern in all clones examined in the course of this project, including an additional study by Sharov et al. (2008) using the NIA Mouse 44 K Microarray v2.1 and v2.2 (Agilent Technologies) (data not shown).

Finally, it should be noted that complex genomic events such as translocations would not be identified with aCGH analysis but they could potentially be identified at the transcriptional level if they result in concordant changes of the expression levels of the genes in the affected region. Therefore, the possibility of a recurrent complex translocation between chromosome 14 and 17 at their respective amplified and deleted

regions cannot be excluded. Even though the exact nature of the chromosome 14/17 pattern, i.e. translocation, sequential events in culture, remains unclear, it has been validated with aCGH analysis at the genomic level (Hall, 2008). The performed validation provides strong additional support for the effectiveness of the proposed methodology to infer patterns of abnormalities at the transcriptional level.

4.7.2. Determination of the minimal overlapping regions of recurring chromosome 14 and 17 aberrant patterns and candidate genes.

Given the high frequency of recurrence of the specific pattern of aberration in chromosome 14 and 17, it was interesting to identify the minimal overlapping region that is common to all affected pluripotent cell lines. Locating the common interval and the genes that are consistently differentially expressed in the region could potentially reveal the gene or genes that play a role in the selective advantage conferred to the cell by the presence of the specific genomic change(s). The obvious caveat with this approach is that the selection mechanism and the chromosomal aberration may not be the same. In addition, not all the genes that are differentially expressed in the aberrant region are necessarily linked to the selection mechanisms but they could rather be differentially expressed because of the underlying DNA CN change. In fact, it is possible that the key candidates whose deregulation has contributed to the selective advantage of the cells are no longer differentially expressed upon the domination of the abnormal cell population and the abolition of the competitive environment. However, the identification of consistently differentially expressed genes within a region that is recurrently aberrant may shed light to the composition and functionality of the aberration.

In order to narrow down to the minimal overlapping genomic region that is aberrant across all the samples with a chromosome 14-17 abnormality, the STAC (Significance Testing for Aberrant Copy number) method has been applied (Diskin, 2006). The STAC method has been originally developed in order to provide a statistical framework for the identification of non-random gain or losses across multiple aCGH experiments. For the purposes of this study, it can be easily generalised for the discovery of statistically significant recurring clusters of aberrant transcription, as it only accepts the genomic location of the aberrant intervals, independently of the method used to generate them.

Briefly, the STAC method uses the null model that a chromosomal aberration is equally likely to occur in any genomic interval of the chromosome. This model can be however erroneous in breakage-prone chromosomal regions (Bailey and Bedford, 2006). The null distribution is calculated by permutations of random rearrangements of the predicted aberrant intervals across the chromosome. Two metrics are calculated: the frequency and the footprint of an aberrant interval. The frequency is defined as the number of times the interval is aberrant across the samples examined, while the footprint takes into account the length of the different intervals and how precise their alignment is (Diskin, 2006).

Figure 4.18, Figure 4.19 and Figure 4.20 show the results of the STAC analysis for the three chromosomal regions, the partial loss of chromosome 14, the distal gain of chromosome 14 and the distal loss of chromosome 17. These regions have been examined independently since it is not safe to conclude that they are linked events (especially in the case of the chromosome 14 deletion followed by amplification). For this analysis, the chromosome was divided in 1Mb intervals and 500 permutations have been performed in order to obtain the null distribution. The minimal overlapping regions for the deletion on chromosome 14 was 59 Mb to 88 Mb, for the amplification on chromosome 14 was 99 Mb to 104 Mb and for the deletion on chromosome 17 was 74 Mb to 86 Mb (p-values<0.05). In addition, expression analysis was performed in order to identify transcripts that were differentially expressed in more than half of the profiles examined for each type of aberration. The analysis was focused on transcripts that were present (P flag) in the samples and they demonstrated a FC higher or equal to 2 FC in more than half of the samples. The resulting gene lists are presented in Table 4.2, Table 4.3 and Table 4.4.

In the chromosome 14 partial deletion, 18 transcripts were consistently down-regulated in more than four of the seven samples examined (<2FC). Among these, *Fgf17* and *Wdfy2* are implicated in signal transduction pathways and *Bin3* functions as a cancer suppressor involved in cell proliferation and cell motility pathways (Prendergast et al., 2009).

In the chromosome 14 distal end amplification, 9 genes were found consistently over-expressed in at least three out of five cell lines examined. *Lmo7* over-expression has been linked to breast cancer and it is implicated in the regulation of the actin

cytoskeleton, motility, migration, and adhesion (Hu et al., 2011). Interestingly, a member of the Zn-finger transcription factor of the Kruppel-like family, *Klf5,* can be found among them. *Klf5* is highly implicated in ESC self-renewal and it directly regulates the expression of *Oct4* and *Nanog* (Parisi et al., 2008; Hall et al., 2009). It has been also shown that *Klf5* regulates ESC proliferation by promoting phosphorylation of *Akt1* via induction of *Tcl1* (Ema et al., 2008). Furthermore, it is downstream of LIF signalling and its over-expression has been shown to significantly reduce ESC differentiation (Bourillot et al., 2009). It is possible, therefore, that the over-expression of *Klf5* is conferring selective advantage to the cells by enhancing self-renewal and reducing differentiation.

Finally, in the chromosome 17 distal deletion, only 5 transcripts were found consistently under-expressed (<2FC) in three out of the four ES cell lines examined. Among them, *Lrpprc* has been implicated with hexose metabolism, prostaglandin synthesis, and glycosphingolipid biology and it has been suggested that it could play an adaptive role in cell survival and apoptosis (Gohil et al., 2010; Michaud et al., 2011). The role of these genes in ES cell biology remains, however, unclear.

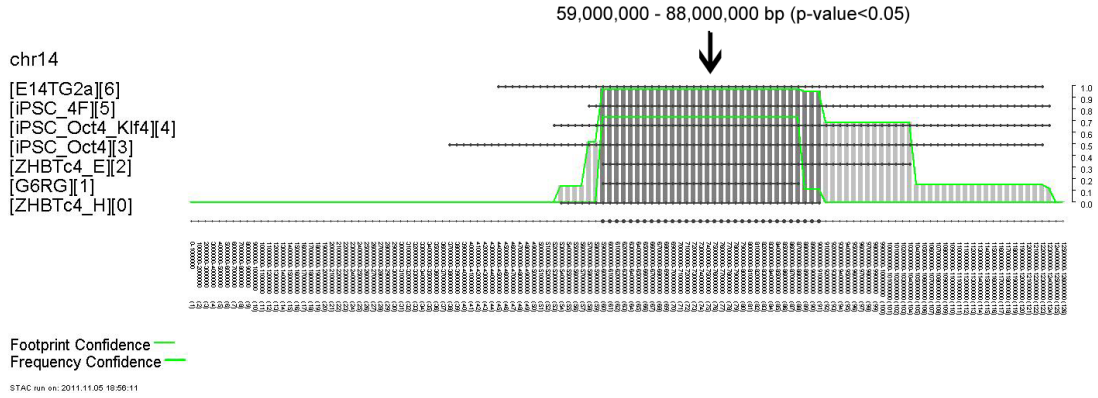

Figure 4.18 Identification of the minimal overlapping region of chromosome 14 deletion by STAC analysis (Diskin, 2006).

Table 4.2 Differentially expressed genes in the chr14 partial loss.
Genes in the minimal overlapping region of the chromosome 14 distal end amplification which demonstrate gene expression <0.5 FC in at least four of the seven ES and iPS cell lines examined (Hall,2008; Endoh et al., 2008; Kim et al., 2008a; Kim et al., 2009; Loh et al., 2006).

Affy ID	Gene Symbol	Average GE
1448499_a_at	Ephx2	0.15
1426255_at	Nefl	0.17
1456239_at	Fgf17	0.19
1451545_at	Tdrd3	0.27
1439411_a_at	Xpo7	0.30
1452436_at	Loxl2	0.31
1418399_at	Kctd9	0.34
1452254_at	Mtmr9	0.35
1418000_a_at	Itm2b	0.36
1425662_at	Cdadc1	0.37
1450385_at	Kpna3	0.38
1417691_at	Bin3	0.40
1426643_at	Elp3	0.41
1430291_at	Dock5	0.41
1456433_at	Rcbtb1	0.42
1424390_at	Nupl1	0.42
1426012_a_at	2610301G19Rik	0.44
1434517_at	Wdfy2	0.46

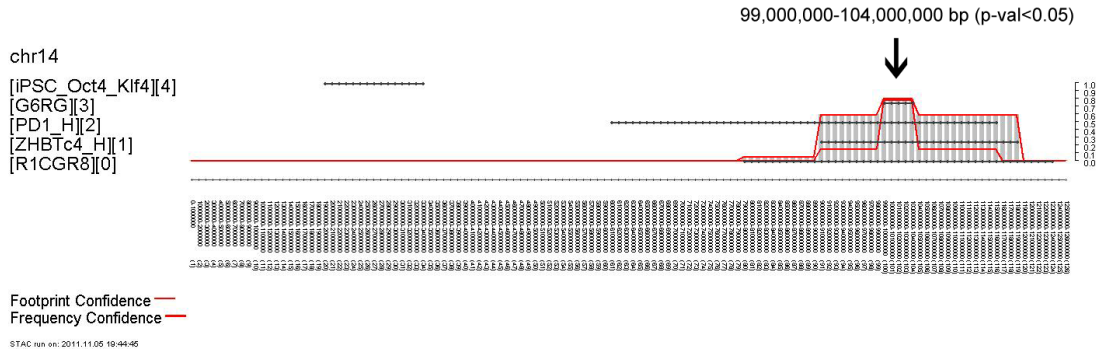

Figure 4.19 Identification of the minimal overlapping region of chromosome 14 amplification by STAC analysis (Diskin, 2006).

Table 4.3 Differentially expressed genes in the chr14 partial gain.
Genes in the minimal overlapping region of the chromosome 14 distal end amplification which demonstrate gene expression >2 FC in at least three of the five ES cell lines examined (Hall,2008; Tejedo et al., 2010; Chen et al., 2010; Endoh et al., 2008).

Affy ID	Gene Symbol	Average GE
1452402_at	Uchl3	15.58
1435863_at	Commd6	5.08
1419163_s_at	Dnajc3a	5.04
1426886_at	Cln5	4.82
1455056_at	Lmo7	4.57
1449855_s_at	Uchl3 /// Uchl4	3.64
1426609_at	Dis3	3.54
1454635_at	Fbxl3	2.62
1451021_a_at	Klf5	2.58

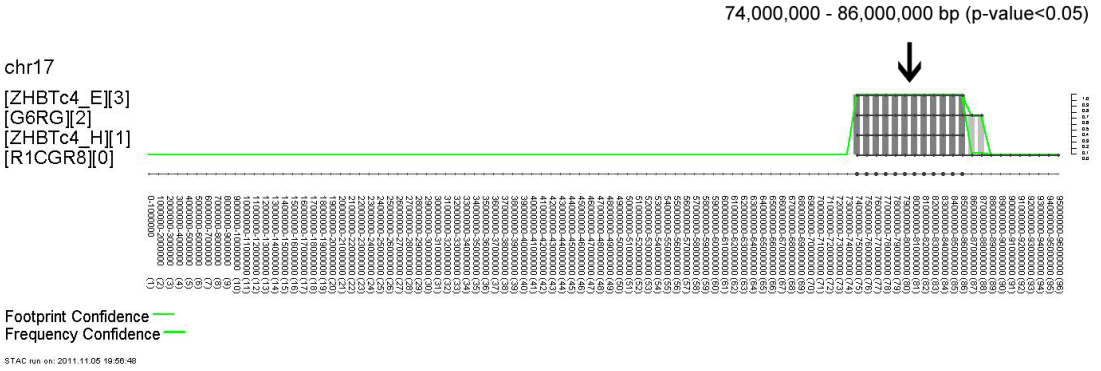

Figure 4.20 Identification of the minimal overlapping region of chromosome 17 deletion by STAC analysis (Diskin, 2006).

Table 4.4 Differentially expressed genes in the chr17 partial loss.
Genes in the minimal overlapping region of the chromosome 17 distal part deletion which demonstrate gene expression <0.5 FC in at least three of the four ES cell lines examined (Hall,2008) (Chen et al., 2010; Endoh et al., 2008).

Affy ID	Gene Symbol	Average GE
1454838_s_at	Pkdcc	0.16
1428229_at	Prkcn	0.32
1428230_at	Prkcn	0.36
1416445_at	2810405J04Rik	0.38
1424353_at	Lrpprc	0.46

4.8. Conclusions

This chapter has presented a sensitive, transcription-tailored integrated framework for the analysis of a large number of publicly available datasets for contiguous changes in gene expression levels along the chromosomes. Large chromosomal clusters of concordant changes in gene expression levels are diagnostic of underlying aneuploidies. The majority of the predicted aberrant genomic clusters map to chromosomes 8 and 11. These results are consistent with those from previous cytogenetic studies examining culture adaptation in murine ES cell lines (Liu et al., 1997; Sugawara et al., 2006) and highlight the necessity of rigorous karyotype validation of stem cell lines maintained in culture for prolonged time periods. The prediction power of the proposed method and the large scale of the data analysis revealed a complex pattern of genomic regions which are prone to be concordantly aberrant, such as the chromosome pairs 8 and 11, 6 and 11, 8 and 14 and 14 and 17. Importantly, many of the events identified here are likely to be of a functional significance, since they have been repeatedly selected for in culture.

The integrated analysis has revealed that in a set of 315 pluripotent samples selected for high *Nanog* expression, 56.83% carry large-scale aberrant transcriptional intervals. A recent study has linked the occurrence of trisomy 12 in hESCs to the *Nanog-Gdf3* presence in the amplified chromosome (Mayshar et al., 2010) and since over-expression of *Nanog* leads to enhanced self-renewal, proposed this as a likely mechanism for driving the aneuploidy. Although, such an effect cannot be excluded in specific cases, in the present study there are a great number of *Nanog*-high aberrant samples, irrespectively of specific chromosome change carried. It is likely that a chromosomal change that promotes cell growth and/ or blocks differentiation and apoptosis, would be selected in a self-renewing, *Nanog*-positive cell in culture in order to eventually dominate the entire cell population. As a result, the generated mixture of cells will show a bias towards self-renewing pluripotent state and therefore carry markers of such cells including *Nanog*.

In addition, the proposed methodology presents several advantages in comparison to the previously published method of Mayshar and colleagues. As opposed to applying a global averaged baseline (Mayshar et al., 2010), the analysis presented here uses

sample-specific baselines with sample-specific thresholds which can improve the ability to identify aberrant patterns (as demonstrated in Figure 4.3). Moreover, the PGE method is specifically designed for the task of inferring positional enrichment in transcriptional data. Mayshar and colleagues apply three different methods for this task (EASE, Expander and CGH-Explorer). EASE and Expander can only identify enrichment in chromosome cytobands or whole chromosomes while CGH-explorer, designed for aCGH analysis, can identify continuous changes on the chromosomes. The three methods were used in combination to cross-validate their predictions. It is not stated however why only the list of up-regulated genes was used for the EASE and Expander analysis.

Finally, the identification of key genes whose deregulation is linked with the presence of aneuploidy can give important insights into ES cell biology and culture condition requirements. This can be achieved by determining the minimal recurring regions across many ES cell lines in order to locate the gene or genes that demonstrate consistently aberrant transcriptional levels. This approach has been presented in the case of the complex chromosome 14-17 recurring pattern of aberrations.

5. Transcriptional signatures of aneuploidy

The analysis of recurring abnormalities in murine ESCs and iPSCs has revealed a transcriptional signature common to all the samples of this study carrying any type of predicted aneuploidy, and specific transcriptional signatures linked to the most commonly recurring chromosome 8 and 11-specific aneuploidies. This chapter describes the methodological steps taken for the identification of the distinct transcriptional signatures of aneuploidy and how they can be used in combination with machine learning techniques to predict the presence of aberrations in uncharacterised samples.

5.1. Introduction

As it has been discussed in chapter 4, the prediction of recurrent patterns of aneuploidy in a high percentage of mouse ES and iPS cell populations examined in the present study (over 50% of the samples) has highlighted a very widespread problem. It is evident that rigorous and continuous validation of pluripotent cell lines is necessary in order to verify their genomic integrity and avoid the misinterpretation of experimental data as a result of the presence of underlying chromosomal aberrations. Especially in the case of human pluripotent cell populations the accumulation of aneuploidies may jeopardise the clinical safety of potential therapeutic applications.

It is still not clear, however, why specific genomic abnormalities tend to recur in pluripotent cell populations. Therefore, it is equally important to discover the genes in the affected regions whose deregulation plays a functional role in the selective advantage that is believed to be conferred by aneuploidy. This could reveal key regulators involved in essential molecular processes of stem cell biology such as self-renewal, cell cycle control, differentiation and apoptosis.

The majority of studies that have assessed chromosomal abnormalities in human ESCs and iPSCs do not converge to a single mechanism. In a study assessing the role of the surface marker CD30, an established marker of ECs which confers resistance to

apoptosis (Pera et al., 1997), it was shown that adapted human ES cell lines express this receptor that could be used to identify subpopulations of abnormal cells in ES cell cultures (Herszfeld et al., 2006). However, in an attempt to verify this result in a different set of hES cell lines, Harrison et al. (2009) reported that CD30 expression is unable to segregate between karyotypically normal and abnormal cells, karyotypically abnormal cells are not protected from apoptosis and suggested that adaptation in culture can occur through different routes. For example, Enver et al. (2005) has proposed that culture adaptation represents a shift of the cell population towards self-renewal instead of differentiation. This study also identified the Notch ligand *DLK1* as the most differentially expressed gene between karyotypically normal and abnormal-culture adapted cells and indicated a potential role of the Notch pathway, which is implicated in hESC proliferation and self-renewal (Enver et al., 2005; Fox et al., 2008).

Chromosomal abnormalities of mouse or human ESCs and iPSCs can also be linked to oncogenesis and tumour progression. A recent study has suggested that during reprogramming, selection can be linked to deleted regions that contain tumour-suppressors and maintenance of cell lines in culture can be facilitated by the presence of amplifications at oncogene-containing loci (Laurent et al., 2011). Other studies have also implicated the over-expression of oncogenes in karyotypically unstable hES cell lines (Narva et al., 2010; Draper et al., 2004; Blum and Benvenisty, 2009; Enver et al., 2005).

Finally, amplified or deleted genes that take part in the core circuit of pluripotency also represent prime candidates for the understanding of the selective nature of specific aneuploidies. Mayshar et al. (2010) have reported the up-regulation of the cluster of hallmark pluripotency genes *NANOG* and *GDF3* in human iPS cell lines bearing the frequent 12p amplification. In addition, Laurent et al. (2011) have found CNVs in pseudogenes of *NANOG* and *OCT4* although their potential role needs still to be investigated. Furthermore, Hall (2008) first identified the up-regulation of *Klf5* in a mouse ES cell line within the amplified region of chromosome 14, part of the complex chromosome 14-17 aberrant pattern described in chapter 4. As discussed in section 4.7, the *Klf5* over-expression was also identified in multiple murine ES cell lines with the same chromosome 14 amplification by the present study.

While there is a plethora of studies examining human ESC and iPSC genomic integrity

and the putative genes under selection, no comprehensive study in mouse has been thus far performed. The next sections present the results of the current analysis for the identification of differentially expressed genes between samples predicted to be karyotypically normal compared to aberrant samples.

5.2. Differential Expression Analysis between Normal and Aberrant Samples

5.2.1. Using SAM to identify differentially expressed genes

In order to determine whether there is a distinct transcriptional signature that can be associated with specific types of aneuploidy, a two-class Significance Analysis of Microarrays (SAM) was performed (Tusher et al., 2001). SAM was chosen as it is one of the most popular methods for the identification of differentially expressed genes as indicated by the number of publications that employ the test (1884 citations of the original publication to present). In summary, SAM measures the relative difference in expression levels by analysing the gene expression change between conditions and the variation of the measurements within condition. It then defines the "expected relative difference" by performing a set of random data permutations. The user defines a cut-off for genes that significantly deviate from the expected relative difference and the FDR for the specified cut-off is calculated. In reality, the SAM test is a modified t-test and the SAM statistic is defined as following:

$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) - s_0}$$

where $d(i)$ is the relative difference of the expression levels of gene i between the conditions I and U, $\bar{x}_I$ and $\bar{x}_U$ are the average expression levels for gene i in conditions I and U, $s(i)$ is the "gene-specific scatter" and s_0 is a small positive constant that minimises the coefficient of variation. More specifically, the "gene-specific scatter" $s(i)$ is defined as:

$$s(i) = \sqrt{a\{\sum_m [x_m(i) - \bar{x}_I(i)]^2 + \sum_n [x_n(i) - \bar{x}_U(i)]^2\}}$$

with

$$a = \frac{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{(n_1 + n_2 - 2)}$$

where n_1 and n_2 the number of measurements in conditions I and U respectively.

SAM analysis was employed in order to identify the differentially expressed genes

between (i) normal versus abnormal samples (any type of chromosomal abnormality), (ii) samples with chromosome 8-specific patterns versus all other samples and (iii) samples with chromosome 11-specific patterns versus all other samples (See Table S3, Table S4 and Table S5 for complete lists of differentially expressed genes with FC≥1.5 and q-val≤0.05 in the attached CD). Chromosome 8 and chromosome 11 were specifically chosen for this analysis because they are the chromosomes most frequently affected by aneuploidy and, in fact, 70% of the predicted aberrant samples carry whole or partial-chromosome aberrations on at least one of these two chromosomes. The results for each type of comparison are presented in the following sections (5.2.2-5.2.4).

It should be noted that this analysis has been performed for the *Nanog*-high group of samples described in section 4.6.2. Samples with low-levels of *Nanog* expression are likely to represent cell populations highly heterogeneous in differentiating or partially reprogrammed cells and therefore not helpful to the study of the stability of self-renewing pluripotent cells. As it has been already discussed, the percentage of aberrant samples in the *Nanog*-low group is distinctly lower. As a result, an attempt to identify differentially expressed genes between karyotypically normal and aberrant samples when including the *Nanog*-low subgroup would only result to the identification of pluripotency-associated genes and would obscure the discovery of aneuploidy-associated genes. In order to avoid this problem, the analysis has focused on the relatively homogeneous pluripotent samples of the *Nanog*-high subgroup.

5.2.2. Global Dataset

In the global dataset, samples were divided in two classes: normal samples where no large-scale cluster of differentially expressed genes has been identified and predicted aberrant samples carrying any type of aberration in any chromosome. In total, 128 genes were found over-expressed and 543 genes were found under-expressed in the aberrant samples group (the complete list is available as part of the supplementary files provided with the attached cd, Table S5).

Figure 5.1 presents a heatmap representation of the top 50 up- and down-regulated genes between normal and aberrant samples according to SAM analysis. The lower

separate panel of the heatmap displays the expression of the three core pluripotency genes *Nanog*, *Pou5f1* (*Oct4*) and *Sox2* and demonstrates that there is no correlation between the aneuploidy signature and the pluripotency signature in the two groups of samples.

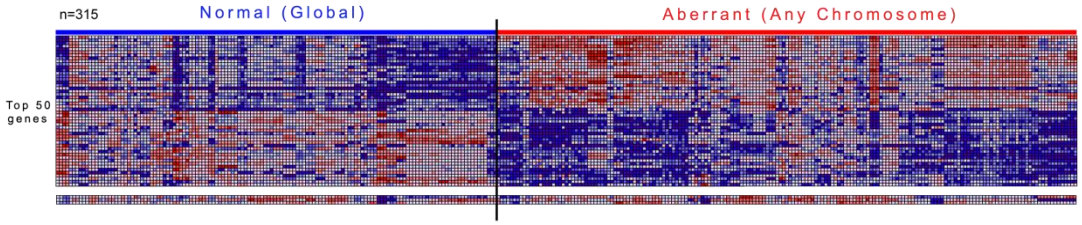

Figure 5.1 Heatmap of Normal versus Aberrant samples.
Heatmap representation of the top 50 differentially expressed genes between normal and aberrant samples carrying any type of chromosomal aberration (generated from SAM analysis). The panel of the three core pluripotency genes (*Nanog*, *Pou5f1* (*Oct4*) and *Sox2*) at the bottom of each heatmap demonstrates the independency of the aberrant chromosomal patterns from the core pluripotency program in the stem cell populations.

In addition, Table 5.1 presents the top 50 differentially expressed genes between the normal and aberrant group (ranked by FC) assigned to functional categories that could be linked to the selective advantage conferred by aneuploidy to the aberrant cells. Importantly, the presence of a deregulated gene expression signature identified across all predicted aneuploid samples suggests that there is a common secondary effect in these cells. It is possible that these cells operate under a positive selection mechanism the downstream consequences of which can be identified at the transcriptional level despite their different types of aneuploidy. Attention is drawn to the top up-regulated list (Table 5.1), typified by genes linked to pluripotency, genomic integrity and cell cycle. An example of this type of gene is *Pramel7* that has been recently reported to promote self-renewal in the absence of exogenous LIF in mouse ESCs (Casanova et al., 2011). Some other interesting examples of differentially over-expressed genes are *Crxos1*, a homeoprotein that has been shown to play a dual role in self-renewal and differentiation (Saito et al., 2010), the non-homologous end-joining repair gene *Lig4* (Frank et al., 2000), the genome maintenance regulator *Zscan4* (Zalzman et al., 2010) as well as the cell-growth modulator *Lin28* (Xu et al., 2009). The function of these genes is consistent with the properties of genes expected to drive aneuploidy.

Finally, the protein kinase, *Pkdcc* (Imuta et al., 2009), that has been found consistently down-regulated in the chromosome 17 deletion described in section 4.7, has been also identified in the global comparison. *Pkdcc,* also known as *Vlk* (vertebrate lonesome

kinase) is a novel protein kinase found to be induced upon differentiation of mouse embryonic stem cells to mesendoderm (Kinoshita et al., 2009). *Pkdcc* could be an example of a gene whose deletion prevents differentiation and, thus, indirectly enhances self-renewal in ESC cultures.

Table 5.1 Functional categories of the top 50 over and under-expressed genes in the *Global* set.

Functional Category	**Up-regulated genes (*Global set*)**	**Down-regulated genes (*Global set*)**
Cell Cycle / Growth	***Lin28***,*Ccnb1ip1,Dnajc2, Anapc10,Syce1*	*Grb10*
Survival	*Pou4f2*, ***Mras***	-
Protein metabolic process	*St8sia1, Anapc10, Dub1, Eif1a, Hck, Map2k6, Rpl39l, Eif2s2*	*Rps9*
Genomic integrity	***Lig4***, ***Zscan4***	-
Cell death	*Plagl1, Map2k6, Xaf1*	*Serpinh1 (Hsp47), Cdh11, Cyr61 (Ccn1)*
Stem cells	***Lin28, Mras, Pramel7, Crxos1***, ***Zfp42*** *(Rex1)*	-
Cancer	*Ceacam1, St8sia1*	*Malat1, Fus*
ECM	-	*Bgn, Col1a1 , Col1a2, Col3a1, Col5a2, Lox, Tnc, App*
Other/ Unknown function	*Calcoco2, Xlr3, Xlr4, 100043292, Pramel6, AU015836, LOC639910, LOC100038935, Spesp1, Hck , H19,Gsta3, Glod5, Snrpn /// Snurf, 2200001I15Rik,Snhg3, 2410004A20Rik, Glrx, Cox7a1, Sec23ip, Zfp560, Sdc4, 666185, Gprc5b*	*Acta2, Thbs1, Mid1, Tagln, Fstl1, Atrx, Prss23, Ptprf, Cd44, Cdk7, Hs6st2, Prtg, Pkdcc, LOC72520, F630007L15Rik, Axl, Lpp, Meg3, Prtg, Sox11, Ptgs2, A130040M12Rik*

The top 50 up- and down-regulated genes (ranked by FC) in the *Global* feature set (which in total includes 128 over-expressed and 543 under-expressed genes). In bold: candidates with literature evidence that supports functional significance in ESC self-renewal.

Table 5.2 GO enrichment analysis for the list of down-regulated genes in the global signature of aneuploidy (Benjamini corrected p-val<0.05)

Category	Term	P-Value	Benjamini
GOTERM_BP_5	blood vessel development	6.90E-11	6.90E-08
GOTERM_BP_5	vasculature development	1.20E-10	5.80E-08
GOTERM_BP_5	tissue development	2.40E-08	7.90E-06
GOTERM_BP_5	organ morphogenesis	1.30E-07	3.20E-05
GOTERM_BP_5	blood vessel morphogenesis	2.20E-07	4.40E-05
GOTERM_BP_5	regulation of cell-substrate adhesion	1.60E-06	2.70E-04
GOTERM_BP_5	muscle organ development	5.20E-06	7.30E-04
GOTERM_BP_5	enzyme linked receptor protein signalling pathway	5.90E-06	7.30E-04
GOTERM_BP_5	head morphogenesis	7.20E-06	7.90E-04
GOTERM_BP_5	chordate embryonic development	3.30E-05	3.30E-03
GOTERM_BP_5	chromatin modification	4.40E-05	4.00E-03
GOTERM_BP_5	in utero embryonic development	5.70E-05	4.70E-03
GOTERM_BP_5	face morphogenesis	8.70E-05	6.60E-03
GOTERM_BP_5	positive regulation of cell migration	1.20E-04	8.30E-03
GOTERM_BP_5	actin filament organization	1.70E-04	1.10E-02
GOTERM_BP_5	tissue morphogenesis	1.70E-04	1.10E-02
GOTERM_BP_5	positive regulation of cell motion	2.10E-04	1.20E-02
GOTERM_BP_5	positive regulation of cell-substrate adhesion	2.30E-04	1.30E-02
GOTERM_BP_5	gland development	3.00E-04	1.60E-02
GOTERM_BP_5	skeletal system morphogenesis	4.50E-04	2.20E-02
GOTERM_BP_5	angiogenesis	5.40E-04	2.50E-02
GOTERM_BP_5	embryonic organ development	6.50E-04	2.90E-02
GOTERM_BP_5	muscle tissue development	6.50E-04	2.80E-02
GOTERM_BP_5	regulation of cell migration	7.50E-04	3.10E-02
GOTERM_BP_5	heart development	9.70E-04	3.80E-02
GOTERM_BP_5	positive regulation of cell proliferation	1.10E-03	4.20E-02

Table 5.3 KEGG pathway enrichment analysis for the list of down-regulated genes in the global signature of aneuploidy (Benjamini corrected p-val<0.05)

Category	Term	P-Value	Benjamini
KEGG_PATHWAY	Focal adhesion	1.10E-11	1.40E-09
KEGG_PATHWAY	ECM-receptor interaction	1.30E-09	8.00E-08
KEGG_PATHWAY	TGF-beta signalling pathway	3.00E-04	1.20E-02
KEGG_PATHWAY	Pathways in cancer	1.10E-03	3.20E-02
KEGG_PATHWAY	Prostate cancer	1.80E-03	4.20E-02

GO biological process and KEGG pathway analyses were performed using DAVID (Huang et al., 2009b; Huang et al., 2009a) in order to investigate the functional enrichment of the identified up- and down-regulated gene lists (Table 5.2 and Table 5.3). The shift towards self-renewal at the expense of differentiation in the aberrant group of samples is also supported by the significantly enriched GO_BP categories that are related to organ morphogenesis and embryonic development and appear down-regulated in aberrant samples, as shown in Table 5.2. It is reasonable to hypothesise that GO_BP categories related to development and organ morphogenesis would appear enriched in cell populations with a high degree of differentiating cells where genes involved in the formation of the three germ layers would be up-regulated. A similar signature would not be present in cell populations highly homogeneous in rapidly self-renewing pluripotent cells as the ones we expect to see in cultures where the presence of aneuploidy has prevailed through selection. In addition, the down-regulated gene list shows significant enrichment in pathways that are highly connected with changes in the extracellular matrix (ECM), a common mechanism for tissue remodelling in development. No significant enrichment was observed for the list of up-regulated genes.

5.2.3. Chromosome 8

In the chromosome 8 dataset, samples were divided in two classes: samples that carry no chromosome 8-specific aberration (even though these samples may carry an aberration on another chromosome) and samples that bear a large-scale cluster of differentially expressed genes spanning any genomic region of chromosome 8. In total, 214 genes were found over-expressed and 611 genes were found under-expressed in

the chromosome 8 aberrant samples group (the complete list is available as part of the supplementary files provided with the attached cd, Table S3).

Figure 5.2 presents a heatmap representation of the top 50 up- and down-regulated genes between normal (in terms of chromosome 8 – specific aberrations) and samples carrying a chromosome 8-specific pattern according to SAM analysis. Again, the lower separate panel of the heatmap displays the expression of the three core pluripotency genes *Nanog*, *Pou5f1* (*Oct4*) and *Sox2*.

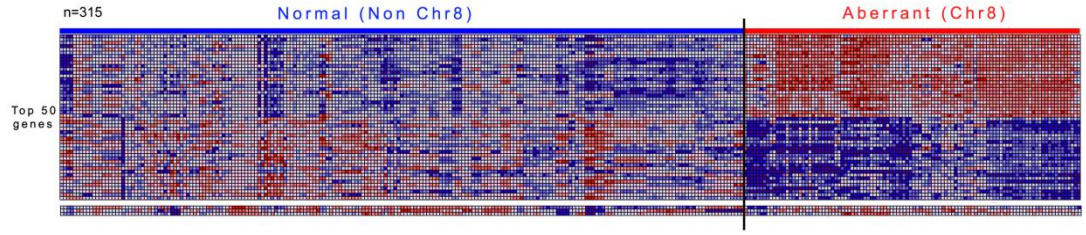

Figure 5.2 Heatmap of Normal-Chr8 versus Aberrant-Chr8 ESCs.
Heatmap representation of the top 50 differentially expressed genes between normal (in terms of chromosome 8 aberrations) and aberrant samples carrying a chromosome 8-specific aberration (generated from SAM analysis). The panel of the three core pluripotency genes (*Nanog*, *Pou5f1* (*Oct4*) and *Sox2*) at the bottom of each heatmap demonstrates the independency of the aberrant chromosomal patterns from the core pluripotency program in the stem cell populations.

In the case of chromosome 8, among the most highly up-regulated genes is *Bag4* (BCL2-associated athanogene 4), also known as the silencer of death domain (SODD), which confers resistance to TNF-induced apoptosis (Miki and Eddy, 2002). *Nob1, Mak16* (*Rbm13*) and *Nip7* are all involved in ribosome biogenesis in yeast (Lamanna and Karbstein, 2009; Pellett and Tracy, 2006) and in human (Morello et al., 2010). *Fnta* is a farnesyltransferase connected to the *Raf1* oncogene, critical for cell-cycle progression in human (Long et al., 2002). *Upf3a* takes part in the nonsense-mediated decay (NMD) and exon-junction (EJC) pathways and it is highly conserved in eukaryotes (Kim et al., 2005). In the down-regulated list, *Clic4* is a downstream target of p53 that induces apoptosis upon DNA damage and down-regulates *Bcl-2* (Fernández-Salas et al., 2002; Suh et al., 2004). *Rab10* (member RAS oncogene family) is responsible for insulin-stimulated GLUT4 translocation (Sano et al., 2008). The list of the top 10 up- and down-regulated genes in samples with chromosome 8-specific aberrations is presented in Table 5.4. The GO biological process and KEGG pathway enrichment performed in DAVID for the full lists of the over- and under-expressed genes is also presented in Table 5.5 and Table 5.6.

Table 5.4 Top 10 up- and down-regulated genes in samples with a chromosome 8-specific aberration (ranked by SAM score)

Affy ID	Gene Name	Fold Change	Genomic Location
Up-regulated			
1454957_at	Nob1	1.55	chr8:109936386-109948938 (-) // qD3
1426426_at	Mak16	1.65	chr8:32269908-32279217 (-) // qA3
1449186_at	Bag4	1.75	chr8:26877238-26895678 (-) // qA2
1452971_at	Upf3a	1.78	chr8:13785636-13798744 (+) // qA1.1
1423873_at	Lsm1	1.87	chr8:26895784-26913009 (+) // qA2
1454659_at	Dctd	1.86	chr8:49226001-49227020 (+) // qB1.2
1417465_at	Fnta	1.54	chr8:27109193-27126080 (-) // qA2
1449724_s_at	D8Ertd738e	1.61	chr8:86770135-86770593 (-) // qC3
1448801_a_at	Timm44	1.52	chr8:4259730-4274233 (-) // qA1.1
1448480_at	Nip7	1.71	chr8:109580776-109583149 (+)// qD3
Down-regulated			
1416648_at	Dync1h1	0.55	chr12:111839661-111905134(+)//qF1
1450784_at	Reck	0.50	chr4:43888401-43957677 (+) // qB1
1435930_at	Scaper	0.64	chr9:55397686-55434123 (-) // qB
1451148_at	Pink1	0.61	chr4:137869323-137882183 (-) // qD3
1415687_a_at	Psap	0.59	chr10:59740404-59765345 (+) // qB4
1423392_at	Clic4	0.56	chr4:134769821-134828686 (-) // qD3
1459275_at	Rnf17	0.52	chr14:57128119-57143869 (+) // qC3
1434206_s_at	Ppp2r5c	0.57	chr12:111818922-111821286(+)//qF1
1416183_a_at	Ldhb	0.52	chr19:22011804-22013056 (+) // qB
1429296_at	Rab10	0.53	chr12:3247427-3249567 (-) //qA1.1

It is particularly interesting that in the list of the top up-regulated genes there are three genes localised on the 8qA3 region, namely *Bag4*, *Lsm1* and *Fnta*. The genes *BAG4* and *LSM1* have been both described as breast cancer oncogenes in the syntenic 8p11-p12 recurrent amplicon in human. *BAG4* and *LSM1*, in combination with *C8ORF4*, influence growth factor independence and anchorage-independent growth of MCF10A breast cancer cells (Yang et al., 2006). In addition, *FNTA* has been also significantly associated with poor breast cancer outcome and it is considered as a possible therapeutic target

(Chin et al., 2006). The 8qA2 region could, therefore, be a potential selective locus on chromosome 8.

Table 5.5 GO and KEGG pathway enrichment analysis for the list of up-regulated genes in the chromosome 8 signature of aneuploidy (Benjamini corrected p-val<0.05)

Category	Term	P-Value	Benjamini
GOTERM_BP_5	ncRNA metabolic process	2.50E-06	1.50E-03
GOTERM_BP_5	ncRNA processing	1.50E-05	4.40E-03
GOTERM_BP_5	RNA metabolic process	1.10E-04	2.20E-02
GOTERM_BP_5	RNA processing	1.90E-04	2.80E-02
KEGG_PATHWAY	RNA degradation	5.30E-06	3.20E-04

Table 5.6 KEGG pathway enrichment analysis for the list of down-regulated genes in the chromosome 8 signature of aneuploidy (Benjamini corrected p-val<0.05)

Category	Term	P-Value	Benjamini
KEGG_PATHWAY	ECM-receptor interaction	7.40E-07	8.30E-05
KEGG_PATHWAY	Focal adhesion	1.70E-04	9.70E-03

5.2.4. Chromosome 11

In a similar way, in the chromosome 11 dataset samples were divided again in two classes: samples that carry no chromosome 11-specific aberration (even though these samples may carry an aberration on another chromosome) and samples that bear a large-scale cluster of differentially expressed genes spanning any genomic region of chromosome 11. 500 genes were found over-expressed and 189 genes were found under-expressed in chromosome 11 aberrant samples (the complete list is available as part of the supplementary files provided with the attached cd, Table S4).

Figure 5.3 presents a heatmap representation of the top 50 up- and down-regulated genes between normal (no chromosome 11 – specific aberrations but samples with abnormalities in other chromosomes included) and samples carrying a chromosome 11-specific pattern according to SAM analysis. Again, the lower separate panel of the heatmap displays the expression of the three core pluripotency genes *Nanog*, *Pou5f1* (*Oct4*) and *Sox2*.

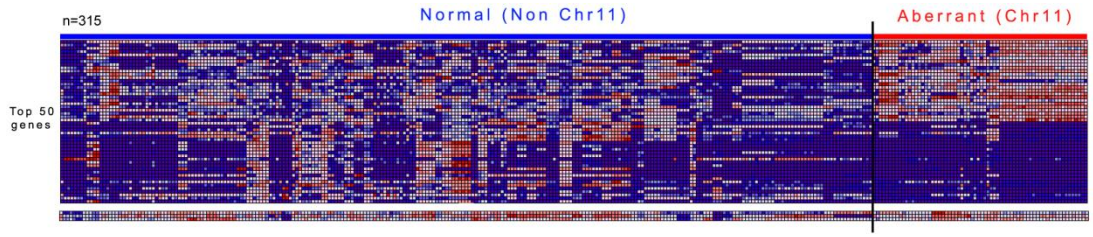

Figure 5.3 Heatmap of Normal-Chr11 versus Aberrant-Chr11 ESCs.
Heatmap representation of the top 50 differentially expressed genes between normal (in terms of chromosome 11 aberrations) and aberrant samples carrying a chromosome 11-specific aberration (generated from SAM analysis). The panel of the three core pluripotency genes (*Nanog*, *Pou5f1* (*Oct4*) and *Sox2*) at the bottom of each heatmap demonstrates the independency of the aberrant chromosomal patterns from the core pluripotency program in the stem cell populations.

For chromosome 11, the list of top 10 up- and down-regulated genes is presented in Table 5.7. Among them, *Pthr2* (else known as Bcl-2 inhibitor of transcription, *Bit*) is one of the up-regulated candidates in the feature set. *Pthr2* is an anoiksis effector that negatively regulates Erk activity (Kairouz-Wahbe et al., 2008) and is up-regulated upon endoplasmic reticulum (ER) stress (Yi et al., 2010). *Prpsap2* (phosphoribosyl pyrophosphate synthetase-associated protein 2) is responsible for de novo synthesis of purine and pyrimidine nucleotides, histidine and tryptophan, and NAD (Katashima et al., 1998). *Utp6* (UTP6, small subunit (SSU) processome component, homolog) which is again integrated in the mitochondrial caspase activation pathway, through interaction with pro-apoptotic *Apaf-1* (Piddubnyak et al., 2007). It is also worth noting that both the hedgehog (Hh) pathway receptor *Ptch1* and activator *Gli1* are in the list of the up-regulated genes (see full list in attached CD – Table S4). This pathway is required in normal development but its abnormal activation has been also associated with tumorigenesis (Goodrich and Scott, 1998).

In the down-regulated list, *Zimp10* has been found to regulate the TGF-beta/Smad signalling pathway and depletion of the gene results in impaired Smad3/4-mediated transcription after TGF-beta induction (Li et al., 2006). *Vegfa* is part of the JAK/STAT signalling pathway and plays a key role in angiogenesis (Suganami et al., 2004). *Ltbp3* is also involved in the TGF-beta signalling pathway (Chen et al., 2002) and *Kdel3* is an ER sorting receptor (Aicha et al., 2007).

The enriched GO biological process categories and the KEGG pathway enrichment found for the list of down-regulated genes in the chromosome 11 aberrant samples is presented in Table 5.8 (DAVID analysis). No significant enrichment was identified in

the list of up-regulated genes.

Table 5.7 Top 10 up- and down-regulated genes in samples with a chromosome 11-specific aberration (ranked by SAM score)

Affy ID	Gene Name	Fold Change	Genomic Location
Up-regulated			
1427492_at	Pof1b	3.28	chrX:109752035-109758938 (-) // qE1
1451845_a_at	Ptrh2	1.65	chr11:86497484-86503689 (+) // qC
1418752_at	Aldh3a1	2.28	chr11:61022246-61031916 (+) // qB2
1427277_at	Six1	2.50	chr12:74142813-74147684 (-) // qC3
1424501_at	Utp6	1.59	chr11:79747632-79775901 (-) // qB5
1452062_at	Prpsap2	1.54	chr11:61543135-61575564 (-) // qB2
1423285_at	Coch	2.49	chr12:52694360-52706762 (+) // qC1
1435995_at	Mrpl22	1.64	chr11:57985191-57993068(+) //qB1.3
1424731_at	Nle1	1.66	chr11:82714250-82721729 (-) // qC
1457882_at	Etaa1	1.60	chr11:17838756-17839369(-) // qA3.1
Down-regulated			
1426936_at	F630007L15Rik	0.22	---
1436841_at	B230380D07Rik	0.57	chr9:70450772-70478922 (-) // qD
1426451_at	Spg11	0.59	chr18:43632469-43635374 (+) // qB3

1448715_x_at	Ccrn4l, Cog6, Sgip1	0.61	chr3:51028369-51055576 (-) // qC
			chr3:52786045-52821145 (-) // qC
			chr4: 102432968-102643782 // qC6

1437479_x_at	Tbx3	0.40	chr5:120134109-120134729 (+) //qF
1420909_at	Vegfa	0.53	chr17:46153941-46169322 (-) // qC
1455316_x_at	BC094435	0.64	chr1: 145208786-145214839 (-) //
1438312_s_at	Ltbp3	0.59	chr19:5758232-5758823 (+) // qA
1418538_at	Kdelr3	0.49	chr15:79346878-79372701 (+) // qE1
1435740_at	Zimp10	0.63	chr14:26477708-26486229 (+) // qA3

Table 5.8 GO biological process and KEGG pathway enrichment analysis for the list of down-regulated genes in the chromosome 11 signature of aneuploidy (Benjamini corrected p-val<0.05)

Category	Term	P-Value	Benjamini
GOTERM_BP_5	blood vessel development	6.60E-05	3.90E-02
GOTERM_BP_5	vasculature development	8.10E-05	2.40E-02
KEGG_PATHWAY	ECM-receptor interaction	4.80E-04	3.40E-02

5.3. Development of Classification Models for the Presence of Aneuploidy

5.3.1. Introduction

The identification of distinct transcriptional signatures associated with the presence of aneuploidy in any chromosome or chromosome-specific aneuploidies (such as the case of chromosome 8 and 11), as discussed in the previous sections, hinted towards the ability to train classification models that can predict these signatures. Two well-established classification techniques have been applied in order to investigate the possibility of identifying the aneuploidy signatures in uncharacterised samples: Prediction Analysis of Microarray (PAM) (Tibshirani et al., 2002) and Support Vector Machines (SVMs) (Vapnik, 1979).

In the present study, three different binary classification tasks have been investigated for the sub-group of 315 *Nanog*-high pluripotent samples. Firstly, in the *Global* set, all aberrant samples, regardless of the chromosomal mapping of the identified change, are classified against all normal samples. The *Chromosome 8* and *11* sets test the predictive ability of the classifiers to detect samples bearing a chromosome 8 or 11 specific aberration. The prediction power of the classifiers was validated by using the accuracy and F1-score (as discussed in section 2.4.4).

This is the first study that attempts to build classification models which will allow detection of samples of any type of aneuploidy as well as chromosome 8 and 11-specific aneuploidies (which account for 70% of the identified aneuploidies in ES and iPS cell lines), opening the way for a more rigorous screening of data throughout the whole ES cell community.

5.3.2. Prediction Analysis of Microarray (PAM)

5.3.2.1. Introduction

Prediction Analysis of Microarray (PAM) is a classification method developed specifically for the task of classification of microarray samples (Tibshirani et al., 2002). PAM is based on the shrunken nearest centroid methodology. Briefly, the method calculates a standardised centroid for each class and compares the gene expression profile of each new sample to each class centroids. The centroid of each class is defined as the average gene expression for each gene in the class divided by the within-class standard deviation for that gene. The new sample is then assigned to the class with the closest centroid in terms of squared distance. The term "shrunken" refers to the modification that can be applied to the calculated centroids by moving the centroids towards zero by adding or subtracting a fixed threshold. The threshold value is defined by the user guided by the results of K-fold cross validation performed by PAM for a range of threshold values. The threshold that produces the lowest misclassification rates from cross-validation is commonly selected. The shrinkage threshold can be also used as an indirect means of feature selection since the genes whose centroid is zero do not contribute to the classification model.

5.3.2.2. Methodology

The detailed methodology of the PAM method can be found in the original publication of Tibshirani et al. (2002) as well as the available *User guide and Manual* (Hastie et al., 2001). In the present study, PAM analysis has been performed for the *Global* set. Two classes have been used: the Normal and Aberrant. A sample is assigned to the class Normal if no large-scale aberration has been identified by the PGE-dendrogram based analysis described in section 4.4 (and after visual inspection using DI.S.C.O.). The class Aberrant consists of samples that have a predicted large-scale aberration in any chromosome. As stated previously, small clusters have been excluded at this stage of the analysis, since they are equally likely to reflect transcriptional regulation or underlying aneuploidies. Even though small clusters of co-regulated genes can be very informative or even reflecting small scale CN variations, this distinction cannot be made at the transcriptional level and, therefore, these samples have not been categorised as aberrant.

This section briefly describes the formulation of the method: let $x_{i,j}$ be the expression of a gene i in a sample j where $i = 1,2,\dots,p$ and $j = 1,2,\dots,n$. In addition, let C_k be the indices of the n_k samples that belong to class k, with $k = 1,2,\dots,K$ (in our case $K = 2$). The contribution to the centroid of the class k from the i^{th} gene can be defined as:

$$\bar{x}_{i,k} = \sum_{j \in C_k} \frac{x_{ij}}{n_k}$$

in other words the average of all the values that the i^{th} gene takes within the class k. Similarly, the i^{th} gene's contribution to the global centroid is:

$$\bar{x}_i = \sum_{j=1}^{n} \frac{x_{ij}}{n}$$

A t-test-like statistic is used to compare the expression of the i^{th} gene in the k^{th} class with the rest of the classes after normalisation by the within-class standard deviation of the gene:

$$d_{ik} = \frac{\bar{x}_{i,k} - \bar{x}_i}{m_k(s_i + s_0)}$$

where $m_k = (\frac{1}{n_k} - \frac{1}{n})^{\frac{1}{2}}$ is used to scale the standard error of the denominator and s_i is the within-class standard deviation:

$$s_i{}^2 = \frac{1}{n-K} \sum_k \sum_{i \in C_k} (x_{ij} - \bar{x}_{i,k})^2$$

In order to shrink the class centroids towards the global centroid, PAM employs *soft-thresholding.* A threshold Δ is chosen by K-fold cross validation and subtracted from each d_{ik} bringing each new d'_{ik} nearer to zero or equal to zero if negative:

$$d'_{ik} = \begin{cases} sign(d_{ik})(|d_{ik}| - \Delta), if\ |(d_{ik}| - \Delta) > 0 \\ 0, if\ |(d_{ik}| - \Delta) \leq 0 \end{cases}$$

The new centroid of the i^{th} gene in the k^{th} class is then calculated as:

$$\bar{x}_{i,k}{}' = \bar{x}_i + m_k s_i d'_{ik}$$

Depending on the chosen threshold Δ, genes that have been shrunken to zero will not contribute in the prediction rule. These genes are more likely to have a highly variable expression in the samples of the k^{th} class or insubstantial deviations from the class centroid.

Under a specific threshold Δ, the test samples are classified depending on the nearest class shrunken centroid. Let $x^* = (x_1^*, x_2^*, .., x_p^*)$ be the expression vector of the new test sample. Then, *discriminant score* of the test sample for the k^{th} class can be defined as:

$$d_k(x^*) = \sum_{i=1}^{p} \frac{(x_i^* - \bar{x}_{i,k}{}')^2}{(s_i + s_0)^2} - 2\log \pi_k$$

where the first term of the above equation is the distance from the class centroid while the second term is a correction term based on the prior probability π_k, where π_k gives the overall proportion of the k^{th} class in the sample population. Class assignment is finally performed based on the minimum $d_k(x^*)$ and the test sample x^* is assigned to the $C(x^*)$ class as from:

$$C(x^*) = \min_{\mathrm{k}} (d_k(x^*))$$

5.3.2.3. PAM – No feature selection

PAM analysis has been performed for the *Global* set with the "pamr" package in the R statistical environment (Tibshirani et al., 2011). Firstly, all genes (features) were included in the analysis. The classifier was trained with a set of 224 samples and 91 independent samples (the remaining dataset) were used for testing. In order to estimate an appropriate threshold for shrinkage a 10-fold cross validation was performed for the training dataset. A threshold value of 6 was selected in order to keep both the number of genes used in the prediction rule and the misclassification error to a minimum (Figure 5.4). Thus, even if the highest accuracy was achieved with a threshold value of 4.3 and more than 250 genes included in the prediction rule, high accuracy levels (82%) could be achieved with a threshold value of 6 and only 26 genes included in the prediction rule (Figure 5.4).

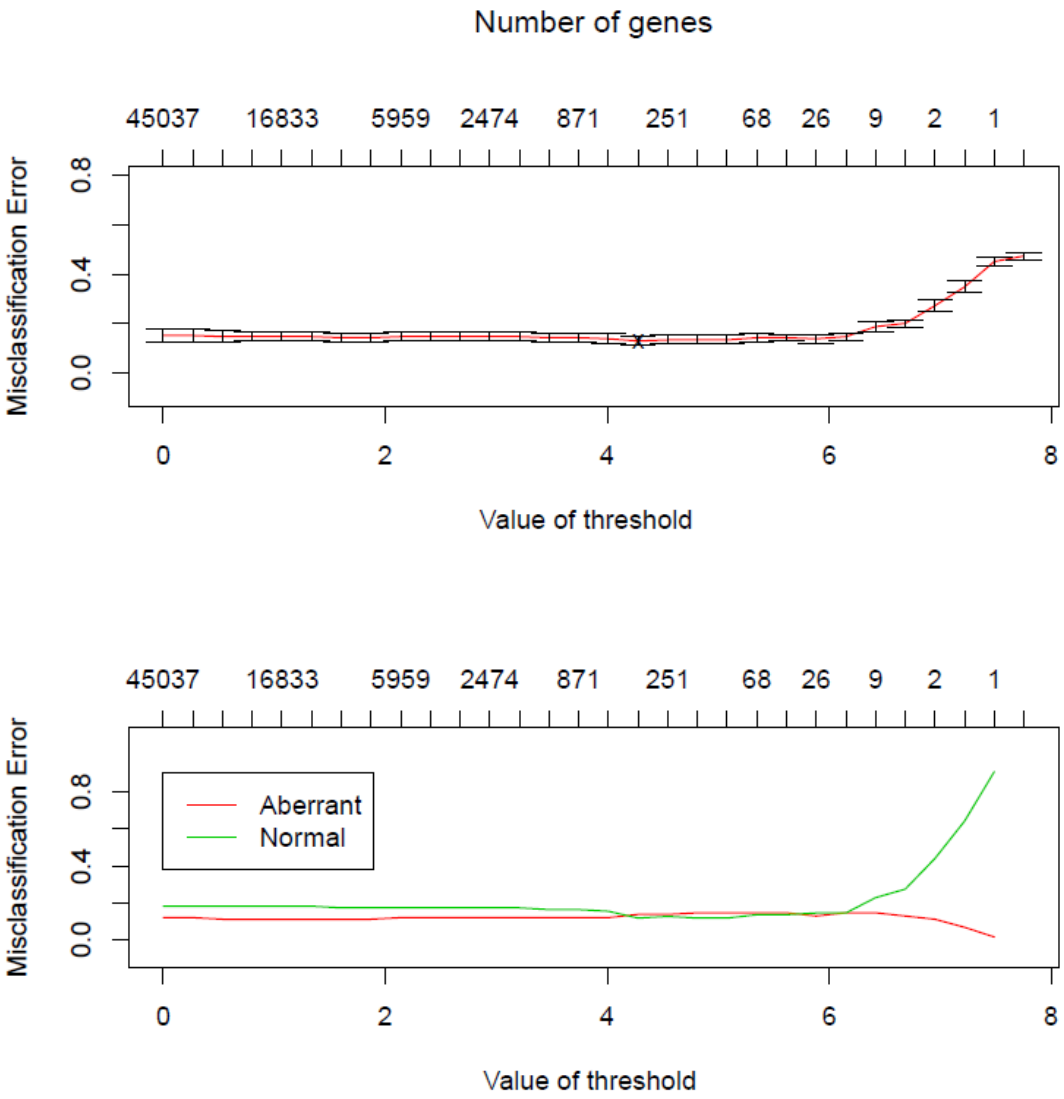

Figure 5.4 10-fold cross-validation curves for different threshold values and number of features (genes) used in the prediction rule.

Figure 5.5 presents the centroids of the significant genes for the Aberrant and Normal classes. The list of the 26 significant genes at the threshold of 6 is also presented in Table 5.9. These are the genes that PAM is using to apply the prediction rule.

Figure 5.5 Aberrant and Normal class centroids by PAM analysis (no feature selection).

Table 5.9 List of the significant genes used in the prediction rule.

Affy Id	Gene Name	Aberrant score	Normal score
1429483_at	Calcoco2	0.0915	-0.0966
1455831_at	Fus	-0.0300	0.0317
1448259_at	Fstl1	-0.1218	0.1285
1429376_s_at	Anapc10	0.0617	-0.0651
1426208_x_at	Plagl1	0.0054	-0.0056
1453599_at	Trim71	-0.0609	0.0643
1440261_at	Ap4e1	0.0227	-0.0240
1450843_a_at	Serpinh1	-0.0063	0.0066
1429051_s_at	Sox11	-0.0409	0.0432
1425458_a_at	Grb10	-0.0427	0.0450
1420841_at	Ptprf	-0.0140	0.0148
1418820_s_at	Zcchc10	0.0036	-0.0038
1426269_at	Vamp7	0.0221	-0.0233
1431353_at	Pabpc4l	0.0421	-0.0444
1439487_at	Lig4	0.0266	-0.0281
1452880_at	Znhit3	0.0167	-0.0177
1427488_a_at	Birc6	-0.0111	0.0117
1418585_at	Ccnh	0.0285	-0.0300
1452378_at	Malat1	-0.0193	0.0203
1459804_at	Crebbp	-0.0034	0.0036
1441959_s_at	1200003C05Rik	-0.0062	0.0066
1443924_at	Wnk3	0.0413	-0.0436
1420935_a_at	Srrm1	-0.0299	0.0315
1429375_at	Anapc10	0.0488	-0.0515
1451238_at	1200003C05Rik	-0.0016	0.0017
1456270_s_at	Pramel6	0.0075	-0.0079

5.3.2.4. PAM – Feature selection with SAM

Gene selection was performed by SAM analysis on the training dataset (with FC>1.5 and q-val<0.05). The lowest misclassification error was achieved with a threshold of ~3.7 and 63 genes included in the prediction rule (Figure 5.6). From the 26 genes identified by PAM without feature selection (section 5.3.2.3) 23 were also included in the 63 genes identified after feature selection by SAM. This high overlap can be also explained by the fact that the PAM method indirectly performs feature selection through the shrinkage threshold in a very similar way as the SAM method and both methods are based on the same principles. The feature selection, possibly in combination with a higher number of genes included in the prediction rule, has improved the accuracy of the method from 82% (no feature selection) to 87% (with feature selection by SAM) (see section 5.3.4 for a breakdown of the prediction accuracy of all models).

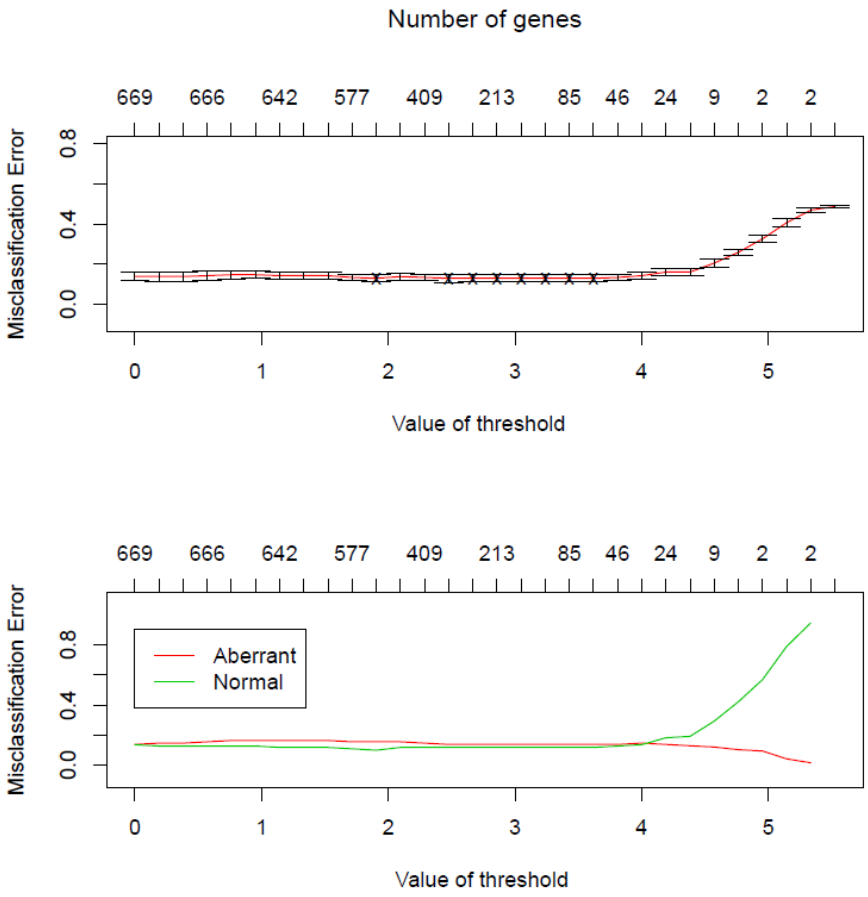

Figure 5.6 10-fold cross-validation curves for different threshold values and number of features (genes) used in the prediction rule (feature selection by SAM analysis).

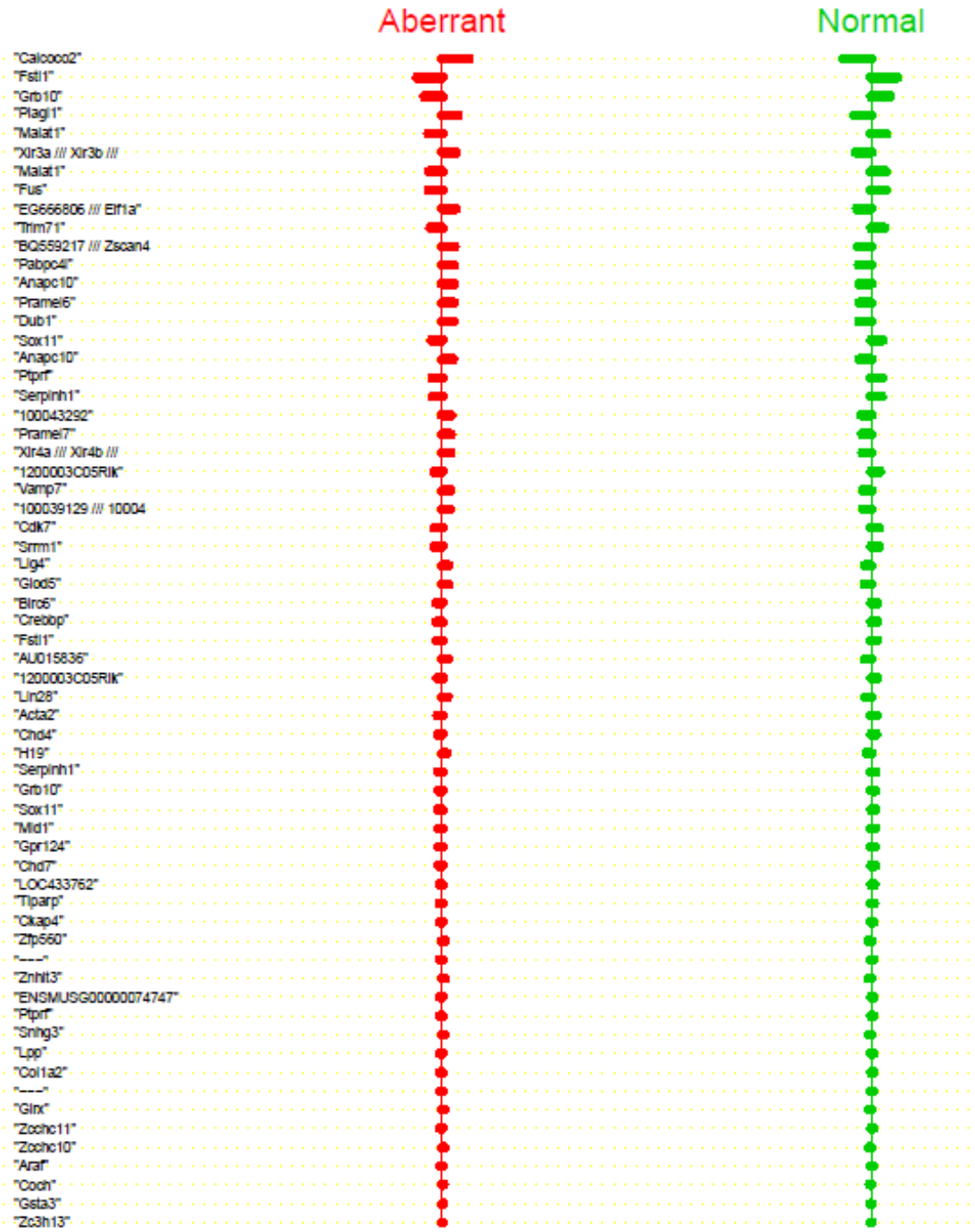

Figure 5.7 Aberrant and Normal class centroids by PAM analysis (feature selection by SAM analysis).

Table 5.10 List of the significant genes used in the prediction rule (after gene selection by SAM).

Affy ID	Gene Name	Aberrant-score	Normal-score
1429483_at	Calcoco2	0.12	-0.13
1448259_at	Fstl1	-0.12	0.12
1425458_a_at	Grb10	-0.08	0.09
1426208_x_at	Plagl1	0.07	-0.08
1452378_at	Malat1	-0.06	0.07
1455831_at	Fus	-0.06	0.07
1453599_at	Trim71	-0.06	0.06
1431353_at	Pabpc4l	0.05	-0.06
1429376_s_at	Anapc10	0.05	-0.06
1456270_s_at	Pramel6	0.05	-0.06
1429051_s_at	Sox11	-0.05	0.05
1429375_at	Anapc10	0.05	-0.05
1420841_at	Ptprf	-0.05	0.05
1450843_a_at	Serpinh1	-0.05	0.05
1441959_s_at	1200003C05Rik	-0.04	0.04
1426269_at	Vamp7	0.04	-0.04
1420935_a_at	Srrm1	-0.03	0.04
1439487_at	Lig4	0.03	-0.03
1427488_a_at	Birc6	-0.03	0.03
1459804_at	Crebbp	-0.03	0.03
1451238_at	1200003C05Rik	-0.02	0.03
1452880_at	Znhit3	0.01	-0.01
1418820_s_at	Zcchc10	0.01	-0.01
1418189_s_at	Malat1	-0.07	0.07
1420357_s_at	Xlr3	0.07	-0.07
1444529_at	Eif1a	0.06	-0.07
1457033_at	Zscan4	0.06	-0.06
1420773_at	Dub1	0.05	-0.06
1443961_at	100043292	0.04	-0.04
1439810_s_at	Pramel7	0.04	-0.04
1449347_a_at	Xlr4	0.04	-0.04
1427479_at	Eif1a	0.04	-0.04
1439511_at	Cdk7	-0.04	0.04

1460454_at	Glod5	0.03	-0.03
1416221_at	Fstl1	-0.03	0.03
1444038_at	AU015836	0.03	-0.03
1421749_at	Lin28	0.02	-0.03
1416454_s_at	Acta2	-0.02	0.02
1436343_at	Chd4	-0.02	0.02
1448194_a_at	H19	0.02	-0.02
1456733_x_at	Serpinh1	-0.02	0.02
1425457_a_at	Grb10	-0.02	0.02
1429372_at	Sox11	-0.02	0.02
1438239_at	Mid1	-0.02	0.02
1418379_s_at	Gpr124	-0.02	0.02
1437745_at	Chd7	-0.02	0.02
1431213_a_at	LOC433762	-0.01	0.02
1452160_at	Tiparp	-0.01	0.01
1426755_at	Ckap4	-0.01	0.01
1438026_at	Zfp560	0.01	-0.01
1457746_at	---	-0.01	0.01
1428127_at	ENSMUSG00000074747	-0.01	0.01
1420843_at	Ptprf	-0.01	0.01
1433789_at	Snhg3	0.01	-0.01
1438271_at	Lpp	-0.01	0.01
1450857_a_at	Col1a2	-0.01	0.01
1438345_at	---	-0.01	0.01
1416593_at	Glrx	0.01	-0.01
1439631_at	Zcchc11	-0.01	0.01
1440764_at	Araf	-0.01	0.01
1423285_at	Coch	0.01	-0.01
1423436_at	Gsta3	0.00	0.00
1430568_at	Zc3h13	0.00	0.00

The complete analysis results are presented for all methods in section 5.3.4 where the predictive power of all the different configurations tested here is discussed.

5.3.3. Support Vector Machines (SVMs)

5.3.3.1. Introduction

The Support Vector Machines (SVMs) is a powerful supervised machine learning technique that has been applied successfully in many biological areas (Statnikov et al., 2004; Brown et al., 2000; Xu et al., 2010; Zhu et al., 2009; Guyon et al., 2002; Golub et al., 1999; Furey et al., 2000; Yeoh et al., 2002). SVMs have been shown to outperform other classification methods for gene expression microarray classification (Brown et al., 2000; Li et al., 2004; Statnikov et al., 2004; Lee et al., 2005; Statnikov and Aliferis, 2007; Xu et al., 2010). Briefly, SVMs map the input data onto a high-dimensional space, where classification can be achieved by defining a hyperplane that separates the data points of the two or more classes. The theoretical concepts of the methodology have been firstly proposed by Vapnik (1979). The current standard implementation of *soft-margin* SVM has been developed by Cortes and Vapnik (1995). Introductory texts and tutorials have been made also made available by (Stitson et al., 1996; Weston et al., 1996; Burges, 1998; Gunn, 1998) among others, while a comprehensive book focused on SVM classification is the one of Cristianini and Shawe-Taylor (2000). A brief description of the fundamental methodology in SVM classification follows in the next section.

5.3.3.2. Methodology

The concept of the *decision boundary* of a classification problem has been first introduced in section 2.4.2. SVMs try to identify the decision boundary that achieves the maximum margin between the classes (Figure 5.8). The margin represents the maximum distance between two hyperplanes parallel to the two sides of the decision boundary that do not contain any samples between them. In the case of non-separable data, the standard *soft-margin* implementation of (Cortes and Vapnik, 1995) allows the existence of mislabelled samples while still maximising the distance between the hyperplanes that are defined by rightly labelled samples. The maximum margin identification is based on the principles of the structural risk minimization (SRM) theory (that is the test error minimisation or minimisation of over-fitting) from (Vapnik, 1979) which shows that the generalisation ability of the classifier is improved as the margin that separates the two classes gets larger.

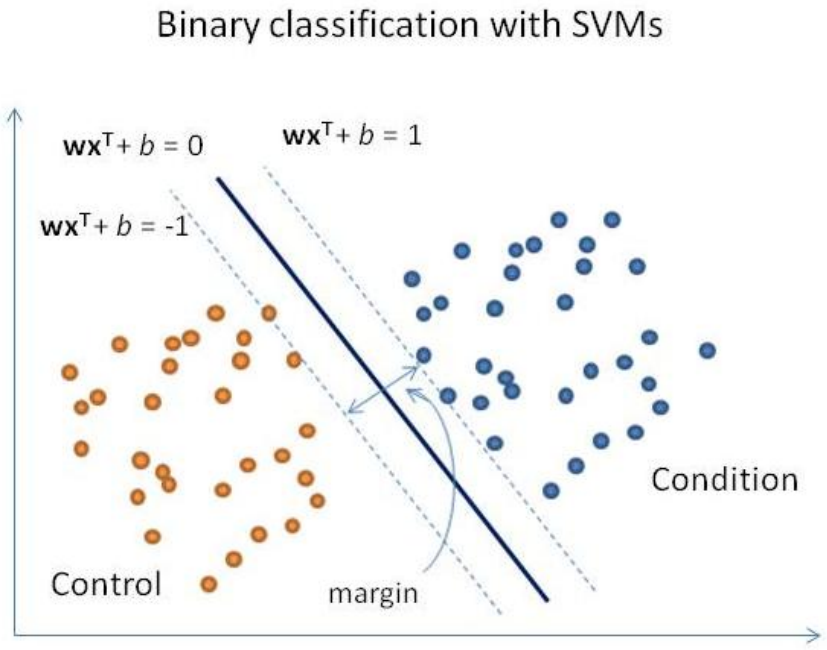

Figure 5.8 An SVM classifier attempts to find the decision boundary that maximises the margin between the two classes (blue and orange points). The samples on the dashed lines are called support vectors (SVs).

Following the notation used in section 2.4.2 and 5.3.2.2, let $\boldsymbol{x}_j$ be the vector of expression of a training sample j where $j = 1,2,\dots,n$. In addition, let C_k be the indices of the n_k samples that belong to class k, with $k = 1,2,\dots,K$ (in our case $K = 2$). SVMs identify the separating hyperplane with the maximum possible margin:

$$\boldsymbol{w}\boldsymbol{x}^T + b = 0$$

where $\boldsymbol{w}$ is the n-dimensional vector perpendicular to the hyperplane and b is the bias. The two margin boundary hyperplanes are formed by the samples of each class that are closest to the maximum-margin hyperplane (Figure 5.8). These samples are called the *support vectors* and the margin boundary hyperplanes are defined:

$$\boldsymbol{w}\boldsymbol{x}^T + b = \begin{cases} +1 \; for \; C_k = +1 \\ -1 \; for \; C_k = -1 \end{cases}$$

From the above equation, it can be derived that the margin is:

$$margin = \frac{1}{\|\boldsymbol{w}\|^2}$$

where $\|w\|$ is the norm of $\boldsymbol{w}$. The identification of the maximum margin is, therefore,

equivalent of minimising the norm of **w**. This allows us to formulate the SVM learning problem by introducing the Lagrange formalism to obtain the objective function L_p as following:

$$L_p = \frac{1}{2}\boldsymbol{w}\boldsymbol{w}^T - \sum_{j=1}^{n} a_j((\boldsymbol{w}^T\boldsymbol{x}_j + b)k_j - 1)$$

where the non-negativity constrains are multiplied with the Lagrange multipliers $a_j \geq 0$ for each $j = 1,2,\dots,n$. The minimisation problem is now equivalent to the identification of the $\boldsymbol{w}$, b and a_j that minimise L_p. This can be calculated by differentiating L_p with respect to w and b and equating to zero:

$$\begin{cases} \dfrac{\partial L_p}{\partial \boldsymbol{w}} = 0 \rightarrow \boldsymbol{w} = \displaystyle\sum_{j=1}^{n} a_j k_j \boldsymbol{x}_j \\ \dfrac{\partial L_p}{\partial b} = 0 \rightarrow \displaystyle\sum_{j=1}^{n} a_j k_j = 0 \end{cases}$$

with resulting classifier:

$$y(x) = sign(\sum_{j=1}^{n} a_j k_j \, \boldsymbol{x}_j{}^T\boldsymbol{x} + b)$$

The solution to the problem can be given follow quandratic programming as the dual problem in the Lagrange multipliers a_j. After the identification of $\boldsymbol{w}$ and b the classification of a new sample $\boldsymbol{x}^*$ can be simply performed by identifying the sign of the $\boldsymbol{w}^T\boldsymbol{x}^* + b$.

In the case of linearly non-separable classification, Cortes and Vapnik (1995) introduced the non-negative *slack variables* ξ_j with $j = 1,2,\dots,n$ and a penalty function for classification errors. The original inequalities are now defined as:

$$(\boldsymbol{w}^T\boldsymbol{x}_j + b)k_j \geq 1 - \xi_j$$

and the penalty function as:

$$\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{j=1}^{n} \xi_j$$

where C is the *regularisation parameter*, a positive real constant. The minimisation problem is now formulated as:

$$min_{\boldsymbol{w},b,\xi}(\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{j=1}^{n}\xi_j)$$

Such that

$$\big(\boldsymbol{w}^T\boldsymbol{x}_j + b\big)k_j \geq 1 - \xi_j, j = 1, \dots, n$$

$$\xi_j \geq 0, j = 1, \dots, n$$

In the case where no separation is possible in the original space, a kernel transformation can be applied that transforms the original input space into a high-dimensional *feature space* (Figure 5.9). Each sample $\boldsymbol{x}_j$ is transformed into a point $\varphi(\boldsymbol{x}_j)$ in the new feature space and the new linear discriminating function can be then defined as:

$$\boldsymbol{w}^T\varphi(\boldsymbol{x}_j) + b$$

since the decision boundaries at the transformed feature space are again linear.

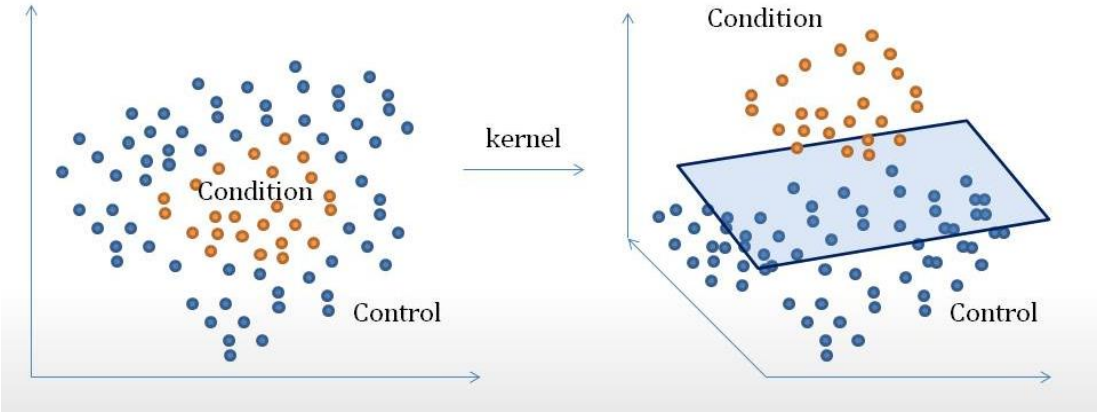

Figure 5.9 With the kernel trick a non-linear separable classification can become linear separable in high-dimensional feature space.

Some of the most common kernel functions used are the polynomial SVM (of degree d) $(\tau + \boldsymbol{x}^T\boldsymbol{z})^d$ and radial kernel (RBF) exp $(-\llbracket(\boldsymbol{x} - \boldsymbol{z})\rrbracket$ ^2/σ^2).

It should also be noted that only one parameter must be tuned in the case of the linear

SVM, that is the regularisation parameter C. In the case of more complex kernel functions more parameters must be tuned, usually by performing an exhaustive search of the parameter space and selection via cross-validation. A larger number of free parameters could turn, however, complex kernel methods more susceptible to over-fitting.

5.3.3.3. Model Validation

The classification accuracy of SVMs was tested with linear, radial, polynomial and sigmoid kernels with default parameters ("e1071" package in R) (Dimitriadou et al., 2005). SVM classifiers were trained for the *Global* set and for the chromosome-specific sets of *Chromosome 8* and *Chromosome 11*. For the chromosome-specific classifiers, the differences in the number of input samples in both classes was accounted for by adjusting the weight parameters of the SVM proportional to the number of samples in each class. This is typically performed in the case of unbalanced classes. For the training of the SVM the *three-way partitioning* method has been chosen as described in the section 2.4.4. A subset of 187 samples and 37 samples was used for training and validation, respectively. After selection of the best scoring classifier, the training and validation subsets were merged in order to train the classifiers again and obtain the final accuracy score on a test dataset of 91 entirely independent samples (the remainder of the complete data collection). Samples from the same studies were kept together in the training, validation or test datasets in order to avoid classification bias as a result of very similar transcriptional profiles used both in training and in testing the prediction power of the classifier (see also section 2.4.5).

In addition, in order to examine what is the minimum number of features required to achieve a low classification error and which genes are the most valuable for the creation of the prediction rule, the Recursive Feature Elimination algorithm for SVM (SVM-RFE) has been applied (Guyon et al., 2002; implementation in R provided by Ruifang and Visser (2010)). SVM-RFE performs feature selection by eliminating one feature at the time in an iterative manner. Features are ranked based on the **w** vector of the decision hyperplane (section 5.3.3.2) and the feature with the lowest ranking value is discarded. A small subset of the features can be then used to examine the classification performance. Guyon et al. (2002) implemented the algorithm for SVMs with linear kernels.

5.3.4. Predictive Power of the Classification Models

Table 5.11 presents the classification performance of the different models described thus far. Remarkably, by reducing the number of genes used for training by applying the SVM-RFE algorithm, small subsets of candidate genes that demonstrate a high class prediction power have been identified. For the *Global* set, the top 50 genes are sufficient to predict abnormalities with a high accuracy (91%). In the case of chromosome-specific SVMs it was possible to narrow the selection down to the top 10 ranked genes while still maintaining a high accuracy (over 80%). Interestingly, a selection of solely non-chromosome 8 mapped genes could still be used to train the classifier for chromosome 8 aberrations with an up to 71% accuracy, suggesting that there is a non-chromosome 8-specific program that is affected by the presence of the predicted transcriptional aberration on chromosome 8. It is also evident that for the *Global* set, where both methods have been applied, the SVM-RFE method outperforms the PAM algorithm.

Table 5.11 Performance of classifiers

Classifier	Set	Kernel	Feature Selection	Accuracy	F1 score
PAM	Global	-	None	0.82	0.88
PAM	**Global**	**-**	**SAM All**	**0.87**	**0.90**
SVM	Global	Linear	None	0.86	0.89
SVM	Global	Linear	SAM All	0.92	0.94
SVM	Global	Linear	SVM-RFE Top 100	0.89	0.92
SVM	Global	**Linear**	**SVM-RFE Top 50**	**0.91**	**0.94**
SVM	Global	Linear	SVM-RFE Top 10	0.55	0.59
SVM	Chr8	Linear	None	0.73	0.68
SVM	Chr8	Linear	SAM All	0.80	0.78
SVM	**Chr8**	**Linear**	**SVM-RFE Top 50**	**0.81**	**0.78**
SVM	Chr8	Linear	SVM-RFE Top 10	0.80	0.79
SVM	Chr8	Linear	*SVM-RFE - No Chr8*	0.71	0.63
SVM	Chr11	Linear	None	0.73	0.29
SVM	Chr11	Linear	SAM All	0.93	0.79
SVM	**Chr11**	**Linear**	**SVM-RFE Top 50**	**0.95**	**0.81**
SVM	Chr11	Linear	SVM-RFE Top 10	0.90	0.61

Best performing classifiers from the PAM (Tibshirani et al., 2002) and SVM (Vapnik, 1998) classification (with bold we highlight the classifier trained with the top 50 features in each set). Feature selection was performed from the SAM (Tusher et al., 2001) output list by the Recursive Feature Elimination (RFE) algorithm (Guyon et al., 2002). In the *SVM-RFE - No Chr8 feature* set, genes mapped to chromosome 8 were excluded from the up-regulated list.

5.4. Conclusions

This chapter has investigated the transcriptional signatures linked to *any* chromosomal aneuploidy identified in this study as well as chromosome 8- and 11-specific aneuploidies. For the *Global* set, which includes samples with *any* chromosomal aberration, it has been shown that a secondary transcriptional effect, common to all samples, can be identified. This is a very important observation that paves the way for the identification and subsequent validation of the deregulated genes/pathways which may be contributing to the selective advantage conferred by aneuploidy. Similar signatures have been identified for samples that carry a chromosome 8 or 11 aneuploidy.

The comparison between normal and aberrant profiles has revealed a set of differentially expressed genes highly connected to pluripotency, cell cycle and apoptosis. It has been proposed before that culture adaptation can occur through multiple mechanisms, in particular via cell cycle progression and deregulation of the p53 pathway or activation of anti-apoptotic pathways (Harrison et al., 2009). Prominent delegates of these processes are present in the selected features (*Lin28, Mras, Pramel7, Crxos1, Rex1, Lig4* and *Zscan4* among others). In addition, it is interesting to note that in some aneuploid cells there is compensation for the adverse effects of higher DNA copy numbers by modulating pathways involved in balancing protein stoichiometry such as ribosome biogenesis and protein degradation (Torres et al., 2008). A similar effect is observed in the case of chromosome 8 clusters which demonstrate enrichment in the Gene Ontology (GO) categories related to RNA processing (Table 5.5).

In this chapter, it has been also shown that it is in fact possible to use an unbiased pool of profiles from publicly available pluripotent samples, carrying complex or chromosome-specific aberrations, to train highly accurate classifiers using only a small set of diagnostic genes. Since genomic aberration are indeed quite often in pluripotent stem cell populations (chapter 4), the use of accurate classification models can provide a cheap, quick and effective validation tool for the scientific ESC community.

6. Discussion

6.1. General Discussion

This thesis was focused on two main goals: (i) the development of computational methods that can be used for the analysis of transcriptional data with a main interest in the prediction of underlying aneuploidy and (ii) the application of these methods in large collections of data in an attempt to better understand the biological mechanisms linked to the presence of genomic changes in pluripotent stem cell populations.

Each of the remaining sections of this chapter presents a recap of the methods developed and a summary of the results obtained from their application. More precisely, section 6.1.1 discusses the software application DI.S.C.O. that can be used for a single-experiment type of analysis. Section 6.1.2 is focused on the large-scale integrated analysis methodology and the findings resulting from its application in a large number of published pluripotent stem cell datasets. Finally, section 6.1.3 presents the analysis for the identification of the transcriptional signatures linked to aneuploidy and the development of classification models for their prediction in uncharacterised new datasets.

6.1.1. DI.S.C.O.

Chapter 3 presented DI.S.C.O., the software application that put the fundaments of this work. DI.S.C.O. is a stand-alone web-based application for the analysis of gene expression profiling data from microarray or RNA-seq technologies in the context of genomic position. DI.S.C.O. offers enhanced visualisation for the transcriptional view of the genome, coupled with three different computational methods for the identification of clusters of differentially expressed genes on the chromosomes. Genomic regions that show a positional enrichment in differentially expressed genes may be the direct consequence of underlying genomic abnormalities such as chromosomal gains or losses or even complex structural chromosomal aberrations. Gene clusters with concordantly altered levels of gene expression can also be due to epigenetic effects including DNA

methylation or histone modifications or the intrinsic structural organisation of the genome (such as the clustering of housekeeping genes (Lercher et al., 2002)).

Several tools exist to date that can perform this type of analysis (a detailed presentation has been given in Table 3.1). These tools demonstrate different limitations that can restrict their usability for the non-expert user. Such potential limiting factors are the need for previous familiarisation with specific statistical or programming packages, the complexity of the model and the number of user-adjusted parameters, the type of organism/platform/technology that the method can process and the effectiveness and the resolution of the method to identify significant patterns. DI.S.C.O. addressed these limitations by providing a powerful and user-friendly analysis platform that accepts gene expression profiling data of any type of organism (as long as there is a genome annotation available) supporting all major microarray technologies and RNA-seq. There is no prerequisite programming knowledge required from the user in order to run a data analysis with DI.S.C.O. In addition, great care has been taken so all the major functionality of the software can be run in an automated way with no additional configuration by the user (or semi-automated when it cannot be avoided). This includes automated discovery of the appropriate FC cut-offs for over- and under-expressed genes and initialisation of the computational methods with default parameters. The computational methods are based on different fundamental principles and they require different number of user-tuned parameters. In Chapter 3, the prediction power of the methods and the effect of the required parameters have been presented through extensive validation with artificial and biological data. It has been shown that the PGE method is the most appropriate for the automated analysis pipeline since it can effectively identify most patterns in biological datasets with the default parameters and the automatically defined FC thresholds. The NN method gives a good alternative although the gap parameter of the method might be dataset specific. The TV method poses the greatest challenge for the user and requires a previous knowledge of the type of pattern that the user seeks to discover in order to optimise the tuning of the method's parameters. Nonetheless, the combination of the Multi-Level Otsu thresholding and the PGE methods can identify subtle patterns in the dataset with no additional configuration from the user.

It could be suggested that the use of the fold change approach for the identification of differentially expressed genes poses a limitation to DI.S.C.O. Even though FC is a

popular metric, mostly because of its simplicity and straight-forward interpretability, it has been often criticised as an inadequate test static primarily since (i) it does not take into account the variance of the expression measurements of a gene across the samples and (ii) it does not offer any p-value for the confidence of an observation (Allison et al., 2006). The choice of FC as the test statistic of use in DI.S.C.O. was, however, a conscious choice. DI.S.C.O. is not focused on the identification of differentially expressed genes per se (there is a wealth of available tools and methods for this purpose) but rather on the identification of concordant changes at the expression levels of genomic regions. For this reason it was desirable to choose a metric that can be easily adjusted by the use of different thresholds so as to facilitate both the visualisation and the computational discovery of the patterns. As it has been applied in the dendrogram-based enrichment analysis for sample-to-sample comparisons, this metric should be also effective in the absence of replicates (which renders metrics that are based on the population standard deviation, such as t-test and SAM, inappropriate). In addition, in the case of positional enrichment, the mere existence of successive genes on the chromosome that share similar changes in their expression levels gives a measure of confidence for the observed pattern and renders the use of a more sophisticated test statistic unnecessary. Finally, by offering the possibility of Absent flags removal, the user can choose to remove transcripts at the lower end of expression which are more likely to produce false positives due to technical noise.

To conclude, the development of DI.S.C.O. was greatly motivated by the observation that a very limited number of publications, specifically in the field of stem cell biology, perform positional enrichment analysis in the context of gene expression data. In fact, at the onset of this project (2008) there was only a single publication that performed a similar type of analysis in human ESCs (Enver et al., 2005) and there are only two additional publications (Mayshar et al., 2010; Ben-David et al., 2011) at present and no relevant publication in mouse ESCs. Given the frequent occurrences of genomic aberrations in pluripotent stem cells (as discussed in section 2.2) and the broad popularity of the gene expression microarray analysis approach, it is essential to provide the scientific community with tools that can be easily, effectively and inexpensively used to perform a first-level validation of the genomic integrity of pluripotent stem cell lines at the transcriptional level.

6.1.2. Large-scale integrated search of mESCs and miPSCs datasets

In Chapter 4, some of the fundamental tools developed in DI.S.C.O., that is the PGE and the Multi-Level Otsu thresholding methods, were integrated in a framework for the analysis of a large-collection (largest to date) of 481 published mouse pluripotent stem cell samples from the Affymetrix GeneChip Mouse Genome 430 2.0 Array. The dendrogram-based positional enrichment methodology (described in section 4.4) is a fully automated analysis pipeline that takes into consideration the specific characteristics of each sample (or cluster of samples) in terms of expression values distribution and compares it in an iterative way with the most similar sample (or cluster of samples) in the matrix. In this way it is possible to reveal transcriptionally aberrant intervals that could potentially otherwise pass unnoticed because of great differences in the transcriptional profiles of distant cell populations or by comparing to a global averaged profile (as in the case of the work by Mayshar and colleagues (2010)). In the proposed approach, both the FC values thresholds and the normalisation scheme is defined automatically after examination of the intrinsic properties of each dataset. In addition, the data matrix is globally normalised using RMA in order to assure that the distribution of gene expression measurements is comparable between samples, Absent flags are removed in order to stabilise the data from noisy variations at the low level of the expression range and each gene is represented by a single value after replacement of replicate probe sets with their median value so as to remove bias in genomic loci where multiple probe sets of the same gene map.

Although it could have been possible to average replicate samples in an attempt to reduce potential biases in recurring aberrations present in the same cell line, it was chosen not to, mainly for three reasons:

(i) replicate samples from the same cell lines may reflect different degrees of heterogeneity in the cell population, i.e. different percentages of aneuploid and diploid cells in the sample.
(ii) greater number of data confers higher predictive ability, and finally,
(iii) by retaining individual replicates it is possible to identify patterns that are present in multiple instances of the same cell line and infer the type and frequency of aneuploidy in the specific cell line i.e. as it has been shown for the case of the ZHBTc4

cell line where all examined instances carry a complex chromosome 14/17 pattern hinting towards an early event that has been propagated in subsequent subclones (also see section 4.7).

The dendrogram-based positional enrichment analysis has been applied in an initial collection of 481 samples from mouse pluripotent stem cell studies and has revealed a catalogue of aberrant intervals. Samples which demonstrate low expression of the pluripotency markers *Oct4, Sox2* and *Nanog* were further filtered out from the dataset (referred to as *Nanog*-low samples). Samples with low expression of pluripotency markers are more likely to consist of highly heterogeneous cell populations of differentiating or partially reprogrammed cells and pluripotent cells. As it has been shown in chapter 4, the percentage of predicted aneuploidy in these samples is much lower than in the *Nanog*-high subgroup which may reflect a decreased ability of the method to identify patterns of aneuploidy in heterogeneous cell populations. It could also reflect that aneuploid cells are prone to undergo apoptosis upon differentiation (Tichy, 2011). The subsequent analysis was focused on the *Nanog*-high subgroup and particularly in samples that carry whole or large partial chromosome spanning aberrant intervals which can be diagnostic of underlying aneuploidies. From the 315 pluripotent samples selected for high *Nanog* expression, 56.83% carry large-scale aberrant transcriptional intervals. This surprising percentage stresses the need for rigorous and continuous validation of mouse pluripotent stem cell lines that can be initially achieved at the transcriptional level with tools such as DI.S.C.O.

The majority of the predicted aberrant genomic intervals mapped primarily to chromosomes 8 and 11 (70% of the samples carried an aberration in either one or both of these chromosomes). In addition, high frequencies of aberrant intervals were discovered in chromosomes 6, 14 and X. These results are consistent with the findings of several previously published cytogenetic studies in mESCs (Liu et al., 1997; Sugawara et al., 2006; Guo et al., 2005) indicating that the transcriptionally aberrant intervals are indeed most probably the result of aneuploidy rather than transcriptional or epigenetic regulation. In addition, complex patterns of recurring simultaneously aberrant chromosomes were unveiled including the pairs of chromosomes 8 and 11, 6 and 11, 8 and 14 and 14 and 17. It is not clear if genomic changes in the above pairs act synergistically by activating members of the same pathways or independently by influencing different biological processes that can lead to the prevalence of aneuploidy

in the cell population through advantageous selection. The hypothesis that these paired patterns actually represent structural subchromosomal changes (especially in the case of the chromosome 14/17 aberration that has been extensively discussed in section 4.7) cannot be excluded. For the chromosome 14/17 combined pattern, it has been demonstrated that the identification of the minimal overlapping genomic region which appears aberrant in many samples can reveal potential candidate genes that may be driving the selection. As such, one of the affected genes in the chromosome 14 distal duplication was *Klf5* that has been shown to play an important role in maintenance of ESC self-renewal (Parisi et al., 2008) and *Pkdcc*, a novel protein kinase found to be induced upon differentiation of mouse embryonic stem cells to mesendoderm (Kinoshita et al., 2009), was identified in the chromosome 17 deletion. Genes, which play a role in the maintenance or enhancement of self-renewal and the induction of differentiation, represent prime candidates that can be linked to the selective advantage conferred by their respective duplication or deletion. Unfortunately, for the most frequent aberrations involving chromosomes 8 and 11 such an analysis could not be performed since in the majority of the samples the aberrant pattern spanned the whole chromosome and most likely represented trisomy 8 and/or 11 as it has been also previously reported (Liu et al., 1997; Sugawara et al., 2006).

Furthermore, Sugawara et al. (2006) and Eggan et al. (2002) have found that frequent aberrations also occur in the sex chromosomes. In the case of chromosome X, 25 samples (8% of samples) carried a large-scale aberration on chromosome X. It is not clear, however, if these patterns represent different degrees of X chromosome inactivation rather than underlying genomic changes. Of note, 70% of the ESC cell lines of this study for which annotation was available were sexed as male. The sex of the iPSC lines is not however often reported. It is believed that the activation of both X chromosomes in female pluripotent cells is closely linked with the establishment of the naïve state of pluripotency (Kim et al., 2011). It is therefore possible that a percentage of the aberrant patterns identified on chromosome X are the result of comparisons between male and female lines with two active chromosomes. An attempt was made to identify instances of chromosome Y deletions or gains by examining the expression of four Y-linked genes that are expressed in mESCs and miPSCs. In this case, it has not been possible to conclude whether the observed up- or down-regulation of these genes in a number of samples could reasonably be linked to a transcriptional regulation event rather than gain or loss of the chromosome. Further experiments will be needed to

clarify the frequency and role of sex chromosomes related genomic changes.

To conclude, this analysis shows that a surprising 56.83% of the 315 high *Nanog* expressing samples carry large-scale aberrant transcriptional intervals. The recent study of Mayshar and colleagues (2010) has attributed the high occurrence of trisomy 12 in hESCs to the *Nanog-Gdf3* presence in the amplified chromosome. Although, over-expression of *Nanog* confers enhanced self-renewal (Chambers et al., 2003) and this effect cannot be excluded, in the examined data of this study the great number of *Nanog* over-expressing aberrant samples, irrespectively of specific chromosome change, hints towards a different conclusion. In fact, it is possible that a single random event that promotes cell growth and/ or blocks differentiation or apoptosis, occurring in a pluripotent cell in culture leads to a rapid domination of the entire cell population. The resulting culture is more homogeneous in pluripotent, albeit aberrant, cells and thus confers both signatures (Figure 6.1).

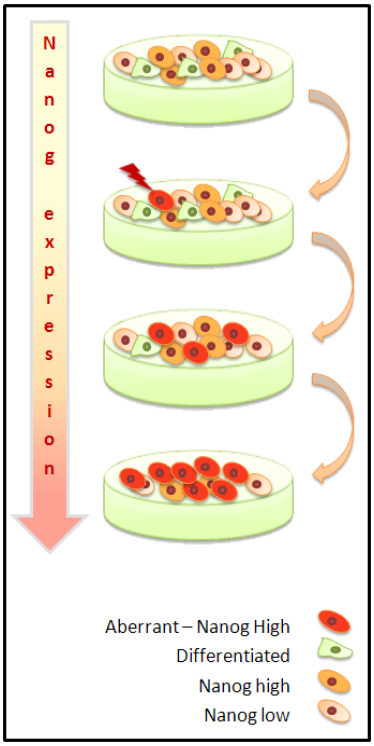

Figure 6.1 A schematic model of the overtaking of the cell population in culture by the Nanog over-expressing aberrant cells.

6.1.3. Transcriptional signatures of aneuploidy and classification of aneuploid samples

In chapter 6, the transcriptional profiles of aberrant and normal samples were compared in order to identify genes whose deregulation could be linked to the selective advantage conferred by aneuploidy. This comparison was effective mainly because samples with low expression of pluripotency markers have been filtered out. Given the higher percentages of aneuploid samples in the *Nanog*-high subgroup versus the *Nanog*-low subgroup, using the initial 481 samples dataset for this comparison would only reveal a high enrichment of the pluripotency signature in the aberrant samples and a differentiation signature in the normal samples. This finding might initially seem intriguing, it is however the result of the difference in percentages of aneuploid samples between the two subgroups.

The comparison between all normal versus all aberrant samples has revealed a number of putative candidates involved in pluripotency, cell cycle control and apoptosis. Among them, *Lin28, Mras, Pramel7, Crxos1, Rex1, Lig4* and *Zscan4* are prominent candidates. The functions of these genes are consistent with those that would be expected to drive selection under a competitive culture environment. Moreover, normal samples show a significant enrichment in processes involved in organ morphogenesis and embryonic development (as shown by GO biological process analysis), further supporting the shift towards self-renewal at the expense of differentiation in aberrant samples (as it has been also suggested by Enver et al. (2005)). Differential expression analysis was also performed for the case of the frequent chromosome 8 and 11 abnormalities. In the case of chromosome 8, an interesting candidate region has been identified at the 8qA2 amplicon, containing the *Bag4*, *Lsm1* and *Fnta* genes that have been described as breast cancer oncogenes in human (Yang et al., 2006; Chin et al., 2006).

In chapter 6, it has been also shown that it is in fact possible to use an unbiased pool of profiles from publicly available pluripotent samples, carrying complex or chromosome-specific aberrations, to train highly accurate classifiers using only a small set of diagnostic genes. In addition, it is shown for the first time that many different predicted aneuploidies in pluripotent cells are associated with a common transcriptional signature that can be used to assess the integrity of pluripotent cultures. The presence

of an aneuploidy-related transcriptional signature in pluripotent stem cells can be used for the identification of core pathways that can be subsequently targeted to develop aneuploidy anti-selective culture conditions. Such an approach has been effectively applied in trisomic MEFs and human cancer cell lines with compounds that are anti-selective for karyotypically abnormal cells (Tang et al., 2011). Finally, these findings pave the way towards the creation of low-cost assays for the validation of ESCs and iPSCs, a tool much needed given the high percentages of predicted aberrant samples in public data sets.

6.2. Open Questions and Future Directions

During the completion of this thesis, a primary goal was to move from a purely descriptive approach of the genomic integrity of pluripotent cells to an attempt to understand the biological mechanisms involved in the selective growth advantage of recurring abnormalities. The majority of published studies thus far mostly catalogue the observed abnormal karyotypes and can only hypothesise about the event(s) that contributed and promoted their presence. Here, a comprehensive list is provided of candidate genes whose deregulation is linked to specific types of aneuploidy and, most importantly, for any type of aneuploidy. The discovered global signature of aneuploidy, which can be also used to predict aberrations in novel uncharacterised samples, hints towards a common secondary effect(s) present in all cells that undergo culture adaptation. Further analysis of the putative candidates is required that can be only performed using specialised wet lab techniques.

In addition, it would be desirable to also apply the dendrogram-based positional enrichment methodology to a large-collection of human pluripotent stem cells. For example, the frequently aberrant human chromosome 17 shares orthology with a single mouse chromosome, the frequently aberrant chromosome 11 (specifically the distal part of mouse chromosome 11) (DeBry and Seldin, 1996). Furthermore, the cluster of predictive genes on the mouse 8qA2 amplicon is syntenic to a cluster of breast cancer oncogenes at the 8p11-p12 recurrent amplicon in human. Therefore, a comparative analysis between human and mouse pluripotent cell lines and their observed patterns of aneuploidy could aid narrow down the list of implicated genes by excluding inconsequential bystanders in the aberrant genomic loci.

Moreover, it has been shown here that classification models can predict the presence of aneuploidy with over 90% accuracy. However, mainly due to time limitations, an exhaustive search in the parameter space has not been performed. Such a search could identify the optimal set of parameters for different SVM classifiers and it is not unlikely that it could even increase the prediction power of the models.

Finally, it would be interesting to move from a 2D view of the genome towards a three dimensional representation of the transcription on the chromosomes. This approach could possibly reveal a whole new level of interactions between chromosome domains and chromosome territories. The study of the 3D spatial organisation of the chromosomes in the nucleus may shed additional light to our understanding of gene regulation and genomic stability (Meaburn and Misteli, 2007).

6.3. Final Conclusions

To sum up, this study:

i) Has provided a software application that can be used to identify clusters of differentially expressed genes on the chromosomes providing a powerful tool for the validation of the genomic integrity of cells at the transcriptional level,

ii) Has proposed an effective methodology for the analysis of the genomic integrity of large-collections of samples and the identification of recurring patterns of aneuploidy. The developed screening method can identify even small aneuploidies in mouse ES and iPS cells, using only existing genome-wide expression data,

iii) Has found that a majority (56.83%) of ES cell lines used for global expression studies are predicted to carry aneuploidies. These anomalies are likely to have led to systematic errors in the interpretation of expression data from many laboratories,

iv) Has identified a global signature of aneuploidy, diagnosed by a small gene set, and built classification models which will allow detection of samples of any type of aneuploidy as well as chromosome 8 and 11-specific aneuploidies (which account for

70% of the identified aneuploidies in ES and iPS cell lines), opening the way for a more rigorous screening of data throughout the whole ES cell community.

7. References

Aicha, S. B., Lessard, J., Pelletier, M., Fournier, A., Calvo, E., and Labrie, C. (2007). Transcriptional profiling of genes that are regulated by the endoplasmic reticulum-bound transcription factor AIbZIP/CREB3L4 in prostate cells. Physiological Genomics *31*, 295–305.

Allison, D. B., Cui, X., Page, G. P., and Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. Nat. Rev. Genet. *7*, 55–65.

Amaldi, E. (1998). On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. Theoretical Computer Science *209*, 237–260.

Ambroise, C., and McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. Proceedings of the National Academy of Sciences *99*, 6562–6566.

Ambrosetti, D. C., Basilico, C., and Dailey, L. (1997). 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.

Ambrosetti, D. C., Schöler, H. R., Dailey, L., and Basilico, C. (2000). 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.

Amit, M., Carpenter, M. K., Inokuma, M. S., Chiu, C. P., Harris, C. P., Waknitz, M. A., Itskovitz-Eldor, J., and Thomson, J. A. (2000). Clonally derived human embryonic stem cell lines maintain pluripotency and proliferative potential for prolonged periods of culture. Dev. Biol. *227*, 271–278.

Amon, A. (1999). The spindle checkpoint. Curr. Opin. Genet. Dev. *9*, 69–75.

Amps, K., Andrews, P. W., Anyfantis, G., Armstrong, L., Avery, S., Baharvand, H., Baker, J., Baker, D., Munoz, M. B., Beil, S., et al. (2011). Screening ethnically diverse human embryonic stem cells identifies a chromosome 20 minimal amplicon conferring growth advantage. Nature Biotechnology. Available at: http://www.nature.com/doifinder/10.1038/nbt.2051 [Accessed December 1, 2011].

Aoi, T., Yae, K., Nakagawa, M., Ichisaka, T., Okita, K., Takahashi, K., Chiba, T., and Yamanaka, S. (2008). Generation of Pluripotent Stem Cells from Adult Mouse Liver and Stomach Cells. Science *321*, 699–702.

Atkin, N. B., and Baker, M. C. (1982). Specific chromosome change, i(12p), in testicular tumours? Lancet *2*, 1349.

Atwood, C. (2011). Embryonic stem cells recent advances in pluripotent stem cell-based regenerative medicine (Rijeka, Croatia :: Intech,).

Avilion, A. A., Nicolis, S. K., Pevny, L. H., Perez, L., Vivian, N., and Lovell-Badge, R. (2003). Multipotent cell lineages in early mouse development depend on SOX2 function. Genes Dev. *17*, 126–140.

Awad, I. A. B., Rees, C. A., Hernandez-Boussard, T., Ball, C. A., and Sherlock, G. (2004). Caryoscope: an Open Source Java application for viewing microarray data in a genomic context. BMC Bioinformatics *5*, 151.

Bailey, S. M., and Bedford, J. S. (2006). Studies on chromosome aberration induction: what can they tell us about DNA repair? DNA Repair (Amst.) *5*, 1171–1181.

Baker, D. E. C., Harrison, N. J., Maltby, E., Smith, K., Moore, H. D., Shaw, P. J., Heath, P. R., Holden, H., and Andrews, P. W. (2007). Adaptation to culture of human embryonic stem cells and oncogenesis in vivo. Nat Biotech *25*, 207–215.

Barone, M. V., Pepperkok, R., Peverali, F. A., and Philipson, L. (1994). Id proteins control growth induction in mammalian cells. Proc. Natl. Acad. Sci. U.S.A. *91*, 4985–4988.

Barriot, R., Sherman, D. J., and Dutour, I. (2007). How to decide which are the most pertinent overly-represented features during gene set enrichment analysis. BMC Bioinformatics *8*, 332.

Bayani, J., and Squire, J. A. (2004). Traditional banding of chromosomes for cytogenetic analysis. Curr Protoc Cell Biol *Chapter 22*, Unit 22.3.

Beddington, R. S., and Robertson, E. J. (1989). An assessment of the developmental potential of embryonic stem cells in the midgestation mouse embryo. Development *105*, 733–737.

Ben-David, U., and Benvenisty, N. (2011). The tumorigenicity of human embryonic and induced pluripotent stem cells. Nature Reviews Cancer *11*, 268–277.

Ben-David, U., Mayshar, Y., and Benvenisty, N. (2011). Large-scale analysis reveals acquisition of lineage-specific chromosomal aberrations in human adult stem cells. Cell Stem Cell *9*, 97–102.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) *57*, 289–300.

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J. S., Dobson, J., Urashima, M., et al. (2010). The landscape of somatic copy-number alteration across human cancers. Nature *463*, 899–905.

Blake, J., Schwager, C., Kapushesky, M., and Brazma, A. (2006). ChroCoLoc: an application for calculating the probability of co-localization of microarray gene expression. Bioinformatics *22*, 765–767.

Blum, B., and Benvenisty, N. (2009). The tumorigenicity of diploid and aneuploid

human pluripotent stem cells. Cell Cycle *8*, 3822–3830.

Boland, M. J., Hazen, J. L., Nazor, K. L., Rodriguez, A. R., Gifford, W., Martin, G., Kupriyanov, S., and Baldwin, K. K. (2009). Adult mice generated from induced pluripotent stem cells. Nature *461*, 91–94.

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics *19*, 185–193.

Boué, S., Paramonov, I., Barrero, M. J., and Izpisúa Belmonte, J. C. (2010). Analysis of Human and Mouse Reprogramming of Somatic Cells to Induced Pluripotent Stem Cells. What Is in the Plate? PLoS ONE *5*, e12664.

Bourillot, P.-Y., Aksoy, I., Schreiber, V., Wianny, F., Schulz, H., Hummel, O., Hubner, N., and Savatier, P. (2009). Novel STAT3 target genes exert distinct roles in the inhibition of mesoderm and endoderm differentiation in cooperation with Nanog. Stem Cells *27*, 1760–1771.

Bradley, A., Evans, M., Kaufman, M. H., and Robertson, E. (1984). Formation of germ-line chimaeras from embryo-derived teratocarcinoma cell lines. Nature *309*, 255–256.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G. G., et al. (2003). ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic Acids Res. *31*, 68–71.

Brons, I. G. M., Smithers, L. E., Trotter, M. W. B., Rugg-Gunn, P., Sun, B., Chuva de Sousa Lopes, S. M., Howlett, S. K., Clarkson, A., Ahrlund-Richter, L., Pedersen, R. A., et al. (2007). Derivation of pluripotent epiblast stem cells from mammalian embryos. Nature *448*, 191–195.

Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., and Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci USA *97*, 262–267.

Buness, A., Kuner, R., Ruschhaupt, M., Poustka, A., Sültmann, H., and Tresch, A. (2007). Identification of aberrant chromosomal regions from gene expression microarray studies applied to human breast cancer. Bioinformatics *23*, 2273–2280.

Burdon, T. (1999). Suppression of SHP-2 and ERK Signalling Promotes Self-Renewal of Mouse Embryonic Stem Cells. Developmental Biology *210*, 30–43.

Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery *2*, 121–167.

Buzzard, J. J., Gough, N. M., Crook, J. M., and Colman, A. (2004). Karyotype of human ES cells during extended culture. Nature Biotechnology *22*, 381–382.

Callegaro, A., Basso, D., and Bicciato, S. (2006). A locally adaptive statistical procedure (LAP) to identify differentially expressed chromosomal regions. Bioinformatics *22*, 2658–2666.

Casanova, E. A., Shakhova, O., Patel, S. S., Asner, I. N., Pelczar, P., Weber, F. A., Graf, U., Sommer, L., Bürki, K., and Cinelli, P. (2011). Pramel7 Mediates LIF/STAT3-Dependent Self-Renewal in embryoniC Stem Cells. STEM CELLS *29*, 474–485.

Catalina, P., Cobo, F., Cortés, J. L., Nieto, A. I., Cabrera, C., Montes, R., Concha, A., and Menendez, P. (2007). Conventional and molecular cytogenetic diagnostic methods in stem cell research: a concise review. Cell Biol. Int. *31*, 861–869.

Catalina, P., Montes, R., Ligero, G., Sanchez, L., de la Cueva, T., Bueno, C., Leone, P. E., and Menendez, P. (2008). Human ESCs predisposition to karyotypic instability: Is a matter of culture adaptation or differential vulnerability among hESC lines due to inherent properties? Mol. Cancer *7*, 76.

Cervantes, R. B., Stringer, J. R., Shao, C., Tischfield, J. A., and Stambrook, P. J. (2002). Embryonic stem cells and somatic cells differ in mutation frequency and type. Proc. Natl. Acad. Sci. U.S.A. *99*, 3586–3590.

Chambers, I., and Smith, A. (2004). Self-renewal of teratocarcinoma and embryonic stem cells. Oncogene *23*, 7150–7160.

Chambers, I., and Tomlinson, S. R. (2009). The transcriptional foundation of pluripotency. Development *136*, 2311–2322.

Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S., and Smith, A. (2003). Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. Cell *113*, 643–655.

Chambers, I., Silva, J., Colby, D., Nichols, J., Nijmeijer, B., Robertson, M., Vrana, J., Jones, K., Grotewold, L., and Smith, A. (2007). Nanog safeguards pluripotency and mediates germline development. Nature *450*, 1230–1234.

Chari, R., Lockwood, W. W., and Lam, W. L. (2006). Computational methods for the analysis of array comparative genomic hybridization. Cancer Inform *2*, 48–58.

Chen, J., Liu, J., Han, Q., Qin, D., Xu, J., Chen, Y., Yang, J., Song, H., Yang, D., Peng, M., et al. (2010). Towards an Optimized Culture Medium for the Generation of Mouse Induced Pluripotent Stem Cells. Journal of Biological Chemistry *285*, 31066–31072.

Chen, Y., Dabovic, B., Annes, J. P., and Rifkin, D. B. (2002). Latent TGF-beta binding protein-3 (LTBP-3) requires binding to TGF-beta for secretion. FEBS Lett. *517*, 277–280.

Cheng, Y.-J., Tsai, J.-W., Hsieh, K.-C., Yang, Y.-C., Chen, Y.-J., Huang, M.-S., and Yuan, S.-S. (2011). Id1 promotes lung cancer cell proliferation and tumor growth through Akt-related pathway. Cancer Lett. *307*, 191–199.

Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W.-L., Lapuk, A.,

Neve, R. M., Qian, Z., Ryder, T., et al. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer Cell *10*, 529–541.

Chung, Y.-J., Jonkers, J., Kitson, H., Fiegler, H., Humphray, S., Scott, C., Hunt, S., Yu, Y., Nishijima, I., Velds, A., et al. (2004). A whole-genome mouse BAC microarray with 1-Mb resolution for analysis of DNA copy number changes by array comparative genomic hybridization. Genome Res. *14*, 188–196.

Cifola, I., Spinelli, R., Beltrame, L., Peano, C., Fasoli, E., Ferrero, S., Bosari, S., Signorini, S., Rocco, F., Perego, R., et al. (2008). Genome-wide screening of copy number alterations and LOH events in renal cell carcinomas and integration with gene expression profile. Mol. Cancer *7*, 6.

Coppe, A., Danieli, G. A., and Bortoluzzi, S. (2006). REEF: searching REgionally Enriched Features in genomes. BMC Bioinformatics *7*, 453.

Cortes, C., and Vapnik, V. (1995). Support-vector networks. Machine Learning *20*, 273–297.

Cory, S., and Adams, J. M. (2002). The bcl2 family: regulators of the cellular life-or-death switch. Nature Reviews Cancer *2*, 647–656.

Cowan, C. A., Atienza, J., Melton, D. A., and Eggan, K. (2005). Nuclear Reprogramming of Somatic Cells After Fusion with Human Embryonic Stem Cells. Science *309*, 1369–1373.

Cristianini, N., and Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods (Cambridge, New York: Cambridge University Press).

Damelin, M., Sun, Y. E., Sodja, V. B., and Bestor, T. H. (2005). Decatenation checkpoint deficiency in stem and progenitor cells. Cancer Cell *8*, 479–484.

DeBry, R. W., and Seldin, M. F. (1996). Human/mouse homology relationships. Genomics *33*, 337–351.

Deng, W., Tsao, S. W., Mak, G. W. Y., Tsang, C. M., Ching, Y. P., Guan, X.-Y., Huen, M. S. Y., and Cheung, A. L. M. (2010). Impact of G2 checkpoint defect on centromeric instability. Oncogene *30*, 1281–1289.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., and Weingessel, A. (2005). Misc Functions of the Department of Statistics (e1071), TU Wien.

Diskin, S. J. (2006). STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. Genome Research *16*, 1149–1158.

Doetschman, T. C., Eistetter, H., Katz, M., Schmidt, W., and Kemler, R. (1985). The in vitro development of blastocyst-derived embryonic stem cell lines: formation of visceral yolk sac, blood islands and myocardium. J Embryol Exp Morphol *87*, 27–45.

Draper, J. S., Pigott, C., Thomson, J. A., and Andrews, P. W. (2002). Surface antigens of

human embryonic stem cells: changes upon differentiation in culture*. Journal of Anatomy *200*, 249–258.

Draper, J. S., Smith, K., Gokhale, P., Moore, H. D., Maltby, E., Johnson, J., Meisner, L., Zwaka, T. P., Thomson, J. A., and Andrews, P. W. (2004). Recurrent gain of chromosomes 17q and 12 in cultured human embryonic stem cells. Nat Biotech *22*, 53–54.

Dyce, J., George, M., Goodall, H., and Fleming, T. P. (1987). Do trophectoderm and inner cell mass cells in the mouse blastocyst maintain discrete lineages? Development *100*, 685–698.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research *30*, 207–210.

Eggan, K., Rode, A., Jentsch, I., Samuel, C., Hennek, T., Tintrup, H., Zevnik, B., Erwin, J., Loring, J., Jackson-Grusby, L., et al. (2002). Male and female mice derived from the same embryonic stem cell clone by tetraploid embryo complementation. Nat. Biotechnol. *20*, 455–459.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. U.S.A. *95*, 14863–14868.

Ema, M., Mori, D., Niwa, H., Hasegawa, Y., Yamanaka, Y., Hitoshi, S., Mimura, J., Kawabe, Y.-ichi, Hosoya, T., Morita, M., et al. (2008). Krüppel-like factor 5 is essential for blastocyst development and the normal self-renewal of mouse ESCs. Cell Stem Cell *3*, 555–567.

Endoh, M., Endo, T. A., Endoh, T., Fujimura, Y.-ichi, Ohara, O., Toyoda, T., Otte, A. P., Okano, M., Brockdorff, N., Vidal, M., et al. (2008). Polycomb group proteins Ring1A/B are functionally linked to the core transcriptional regulatory circuitry to maintain ES cell identity. Development *135*, 1513–1524.

Enver, T., Soneji, S., Joshi, C., Brown, J., Iborra, F., Orntoft, T., Thykjaer, T., Maltby, E., Smith, K., Abu Dawud, R., et al. (2005). Cellular differentiation hierarchies in normal and culture-adapted human embryonic stem cells. Hum. Mol. Genet *14*, 3129–3140.

Evans, M. J., and Kaufman, M. H. (1981). Establishment in culture of pluripotential cells from mouse embryos. Nature *292*, 154–156.

Fernández-Salas, E., Suh, K. S., Speransky, V. V., Bowers, W. L., Levy, J. M., Adams, T., Pathak, K. R., Edwards, L. E., Hayes, D. D., Cheng, C., et al. (2002). mtCLIC/CLIC4, an organellular chloride channel protein, is increased by DNA damage and participates in the apoptotic response to p53. Mol. Cell. Biol *22*, 3610–3620.

Ferrari, F., Solari, A., Battaglia, C., and Bicciato, S. (2011). PREDA: an R-package to identify regional variations in genomic data. Bioinformatics *27*, 2446–2447.

Forman, G., and Scholz, M. (2010). Apples-to-apples in cross-validation studies: pitfalls

in classifier performance measurement. SIGKDD Explor. Newsl. *12*, 49–57.

Fox, V., Gokhale, P. J., Walsh, J. R., Matin, M., Jones, M., and Andrews, P. W. (2008). Cell-cell signaling through NOTCH regulates human embryonic stem cell proliferation. Stem Cells *26*, 715–723.

Frank, K. M., Sharpless, N. E., Gao, Y., Sekiguchi, J. M., Ferguson, D. O., Zhu, C., Manis, J. P., Horner, J., DePinho, R. A., and Alt, F. W. (2000). DNA ligase IV deficiency in mice leads to defective neurogenesis and embryonic lethality via the p53 pathway. Mol. Cell *5*, 993–1002.

Frigola, J., Song, J., Stirzaker, C., Hinshelwood, R. A., Peinado, M. A., and Clark, S. J. (2006). Epigenetic remodeling in colorectal cancer results in coordinate gene suppression across an entire chromosome band. Nat. Genet. *38*, 540–549.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics *16*, 906–914.

Gagos, S., and Irminger-Finger, I. (2005). Chromosome instability in neoplasia: chaotic roots to continuous growth. Int. J. Biochem. Cell Biol. *37*, 1014–1033.

Garber, M., Grabherr, M. G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. Nature Methods *8*, 469–477.

Garge, N. R., Page, G. P., Sprague, A. P., Gorman, B. S., and Allison, D. B. (2005). Reproducible clusters from microarray research: whither? BMC Bioinformatics *6 Suppl 2*, S10.

Gearing, D. P., Gough, N. M., King, J. A., Hilton, D. J., Nicola, N. A., Simpson, R. J., Nice, E. C., Kelso, A., and Metcalf, D. (1987). Molecular cloning and expression of cDNA encoding a murine myeloid leukaemia inhibitory factor (LIF). EMBO J. *6*, 3995–4002.

Geigl, J. B., Obenauf, A. C., Schwarzbraun, T., and Speicher, M. R. (2008). Defining "chromosomal instability." Trends Genet. *24*, 64–69.

Gijsbers, A. C. J., and Ruivenkamp, C. A. L. (2011). Molecular karyotyping: from microscope to SNP arrays. Horm Res Paediatr *76*, 208–213.

Gilbert, S. (2000). Developmental biology 6th ed. (Sunderland Mass.: Sinauer).

Gohil, V. M., Nilsson, R., Belcher-Timme, C. A., Luo, B., Root, D. E., and Mootha, V. K. (2010). Mitochondrial and nuclear genomic responses to loss of LRPPRC expression. J. Biol. Chem. *285*, 13742–13747.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science *286*, 531–537.

Goodrich, L. V., and Scott, M. P. (1998). Hedgehog and patched in neural development

and disease. Neuron *21*, 1243–1257.

Gore, A., Li, Z., Fung, H.-L., Young, J. E., Agarwal, S., Antosiewicz-Bourget, J., Canto, I., Giorgetti, A., Israel, M. A., Kiskinis, E., et al. (2011). Somatic coding mutations in human induced pluripotent stem cells. Nature *471*, 63–67.

Gossler, A., Doetschman, T., Korn, R., Serfling, E., and Kemler, R. (1986). Transgenesis by means of blastocyst-derived embryonic stem cell lines. Proc. Natl. Acad. Sci. U.S.A. *83*, 9065–9069.

Gunn, S. R. (1998). Support Vector Machines for Classification and Regression (Southampton: Faculty of Engineering, Science and Mathematics, University of Southampton) Available at: http://users.ecs.soton.ac.uk/srg/publications/pdf/SVM.pdf.

Guo, J., Jauch, A., Heidi, H.-G., Schoell, B., Erz, D., Schrank, M., and Janssen, J. W. G. (2005). Multicolor karyotype analyses of mouse embryonic stem cells. In Vitro Cell. Dev. Biol. Anim. *41*, 278–283.

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. J. Mach. Learn. Res. *3*, 1157–1182.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning *46*, 389–422.

Hackstadt, A. J., and Hess, A. M. (2009). Filtering for increased power for microarray data analysis. BMC Bioinformatics *10*, 11.

Halbritter, F., Vaidya, H., and Tomlinson, S. R. (2012). GeneProf: Integrated Analysis of High-Throughput Sequencing Data. Nature Methods (In press).

Hall, J. (2008). Identification and investigation of the transcriptional targets of Oct4 and the LIF/Stat3 Signalling Pathway, within the context of the mouse embryonic stem cell genome.

Hall, J., Guo, G., Wray, J., Eyres, I., Nichols, J., Grotewold, L., Morfopoulou, S., Humphreys, P., Mansfield, W., Walker, R., et al. (2009). Oct4 and LIF/Stat3 Additively Induce Krüppel Factors to Sustain Embryonic Stem Cell Self-Renewal. Cell Stem Cell *5*, 597–609.

Hand, D., Manilla, H., and Smyth, P. (2001). Principles of Data Mining (The MIT Press).

Hanna, J., Cheng, A. W., Saha, K., Kim, J., Lengner, C. J., Soldner, F., Cassady, J. P., Muffat, J., Carey, B. W., and Jaenisch, R. (2010). Human embryonic stem cells with biological and epigenetic characteristics similar to those of mouse ESCs. Proceedings of the National Academy of Sciences *107*, 9222–9227.

Hanna, J., Markoulaki, S., Schorderet, P., Carey, B. W., Beard, C., Wernig, M., Creyghton, M. P., Steine, E. J., Cassady, J. P., Foreman, R., et al. (2008). Direct Reprogramming of Terminally Differentiated Mature B Lymphocytes to Pluripotency. Cell *133*, 250–264.

Hardiman, G. (2004). Microarray platforms--comparisons and contrasts.

Pharmacogenomics *5*, 487–502.

Harrison, N. J., Baker, D., and Andrews, P. W. (2007). Culture adaptation of embryonic stem cells echoes germ cell malignancy. International Journal of Andrology *30*, 275–281.

Harrison, N. J., Barnes, J., Jones, M., Baker, D., Gokhale, P. J., and Andrews, P. W. (2009). CD30 expression reveals that culture adaptation of human embryonic stem cells can occur through differing routes. Stem Cells *27*, 1057–1065.

Hartemink, A. J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. (2001). Maximum likelihood estimation of optimal scaling factors for expression array normalization. Proceedings of SPIE *4266*.

Hastie, T. J., Narasimhan, B., Tibshirani, R. J., and Chu, G. (2001). PAM "Prediction Analysis of Microarrays" User guide and manual (Stanford: Department of Statistics and Department of Health Research & Policy) Available at: http://www-stat.stanford.edu/~tibs/PAM/pam.pdf.

Herszfeld, D., Wolvetang, E., Langton-Bunker, E., Chung, T.-L., Filipczyk, A. A., Houssami, S., Jamshidi, P., Koh, K., Laslett, A. L., Michalska, A., et al. (2006). CD30 is a survival factor and a biomarker for transformed human pluripotent stem cells. Nat. Biotechnol. *24*, 351–357.

Hertzberg, L., Betts, D. R., Raimondi, S. C., Schäfer, B. W., Notterman, D. A., Domany, E., and Izraeli, S. (2007). Prediction of chromosomal aneuploidy from gene expression data. Genes Chromosomes Cancer *46*, 75–86.

Hochedlinger, K., and Jaenisch, R. (2006). Nuclear reprogramming and pluripotency. Nature *441*, 1061–1067.

Hong, Y., and Stambrook, P. J. (2004). Restoration of an absent G1 arrest and protection from apoptosis in embryonic stem cells after ionizing radiation. Proc. Natl. Acad. Sci. U.S.A. *101*, 14443–14448.

Hooper, M., Hardy, K., Handyside, A., Hunter, S., and Monk, M. (1987). HPRT-deficient (Lesch–Nyhan) mouse embryos derived from germline colonization by cultured cells. Nature *326*, 292–295.

Hosack, D. A., Dennis, G., Jr, Sherman, B. T., Lane, H. C., and Lempicki, R. A. (2003). Identifying biological themes within lists of genes with EASE. Genome Biol. *4*, R70.

Hovatta, O., Jaconi, M., Töhönen, V., Béna, F., Gimelli, S., Bosman, A., Holm, F., Wyder, S., Zdobnov, E. M., Irion, O., et al. (2010). A teratocarcinoma-like human embryonic stem cell (hESC) line and four hESC lines reveal potentially oncogenic genomic changes. PLoS ONE *5*, e10263.

Hu, Q., Guo, C., Li, Y., Aronow, B. J., and Zhang, J. (2011). LMO7 Mediates Cell-Specific Activation of the Rho-Myocardin-Related Transcription Factor-Serum Response Factor Pathway and Plays an Important Role in Breast Cancer Cell Migration. Molecular and Cellular Biology *31*, 3223–3240.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res *37*, 1–13.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc *4*, 44–57.

Hubbell, E., Liu, W.-M., and Mei, R. (2002). Robust estimators for expression analysis. Bioinformatics *18*, 1585–1592.

Hughes, T. R., Mao, M., Jones, A. R., Burchard, J., Marton, M. J., Shannon, K. W., Lefkowitz, S. M., Ziman, M., Schelter, J. M., Meyer, M. R., et al. (2001). Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat. Biotechnol. *19*, 342–347.

Hussein, S. M., Batada, N. N., Vuoristo, S., Ching, R. W., Autio, R., Närvä, E., Ng, S., Sourour, M., Hämäläinen, R., Olsson, C., et al. (2011). Copy number variation and selection during reprogramming to pluripotency. Nature *471*, 58–62.

Huynh, K. D., and Lee, J. T. (2003). Inheritance of a pre-inactivated paternal X chromosome in early mouse embryos. Nature *426*, 857–862.

Hyman, E., Kauraniemi, P., Hautaniemi, S., Wolf, M., Mousses, S., Rozenblum, E., Ringnér, M., Sauter, G., Monni, O., Elkahloun, A., et al. (2002). Impact of DNA Amplification on Gene Expression Patterns in Breast Cancer. Cancer Research *62*, 6240–6245.

Ihaka, R., and Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics *5*, 299–314.

van den IJssel, P. (2005). Human and mouse oligonucleotide-based array CGH. Nucleic Acids Research *33*, e192–e192.

Imreh, M. P., Gertow, K., Cedervall, J., Unger, C., Holmberg, K., Szöke, K., Csöregh, L., Fried, G., Dilber, S., Blennow, E., et al. (2006). In vitro culture conditions favoring selection of chromosomal abnormalities in human ES cells. J. Cell. Biochem. *99*, 508–516.

Imuta, Y., Nishioka, N., Kiyonari, H., and Sasaki, H. (2009). Short limbs, cleft palate, and delayed formation of flat proliferative chondrocytes in mice with targeted disruption of a putative protein kinase gene, Pkdcc (AW548124). Developmental Dynamics *238*, 210–222.

Inzunza, J., Sahlén, S., Holmberg, K., Strömberg, A.-M., Teerijoki, H., Blennow, E., Hovatta, O., and Malmgren, H. (2004). Comparative genomic hybridization and karyotyping of human embryonic stem cells reveals the occurrence of an isodicentric X chromosome after long-term cultivation. Mol. Hum. Reprod *10*, 461–466.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003a). Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. *31*, e15.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003b). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics *4*, 249–264.

Ishkanian, A. S., Malloff, C. A., Watson, S. K., deLeeuw, R. J., Chi, B., Coe, B. P., Snijders, A., Albertson, D. G., Pinkel, D., Marra, M. A., et al. (2004). A tiling resolution DNA microarray with complete coverage of the human genome. Nature Genetics *36*, 299–303.

Jordan, I. K., Mariño-Ramírez, L., and Koonin, E. V. (2005). Evolutionary significance of gene expression divergence. Gene *345*, 119–126.

Kahlem, P., Sultan, M., Herwig, R., Steinfath, M., Balzereit, D., Eppens, B., Saran, N. G., Pletcher, M. T., South, S. T., Stetten, G., et al. (2004). Transcript level alterations reflect gene dosage effects across multiple tissues in a mouse model of down syndrome. Genome Res *14*, 1258–1267.

Kairouz-Wahbe, R., Biliran, H., Luo, X., Khor, I., Wankell, M., Besch-Williford, C., Pascual, J., Oshima, R., and Ruoslahti, E. (2008). Anoikis effector Bit1 negatively regulates Erk activity. Proceedings of the National Academy of Sciences *105*, 1528–1532.

Karpievitch, Y. V., Hill, E. G., Leclerc, A. P., Dabney, A. R., and Almeida, J. S. (2009). An Introspective Comparison of Random Forest-Based Classifiers for the Analysis of Cluster-Correlated Data by Way of RF++. PLoS ONE *4*, e7087.

Katashima, R., Iwahana, H., Fujimura, M., Yamaoka, T., and Itakura, M. (1998). Assignment of the human phosphoribosylpyrophosphate synthetase-associated protein 41 gene (PRPSAP2) to 17p11.2-p12. Genomics *54*, 180–181.

Kawasaki, H., Mizuseki, K., Nishikawa, S., Kaneko, S., Kuwana, Y., Nakanishi, S., Nishikawa, S.-I., and Sasai, Y. (2000). Induction of Midbrain Dopaminergic Neurons from ES Cells by Stromal Cell–Derived Inducing Activity. Neuron *28*, 31–40.

Kim, D. H., Jeon, Y., Anguera, M. C., and Lee, J. T. (2011). X-chromosome epigenetic reprogramming in pluripotent stem cells via noncoding genes. Seminars in Cell & Developmental Biology *22*, 336–342.

Kim, J. B., Sebastiano, V., Wu, G., Araúzo-Bravo, M. J., Sasse, P., Gentile, L., Ko, K., Ruau, D., Ehrich, M., van den Boom, D., et al. (2009). Oct4-induced pluripotency in adult neural stem cells. Cell *136*, 411–419.

Kim, J. B., Zaehres, H., Wu, G., Gentile, L., Ko, K., Sebastiano, V., Arauzo-Bravo, M. J., Ruau, D., Han, D. W., Zenke, M., et al. (2008a). Pluripotent stem cells induced from adult neural stem cells by reprogramming with two factors. Nature *454*, 646–650.

Kim, J. B., Zaehres, H., Wu, G., Gentile, L., Ko, K., Sebastiano, V., Araúzo-Bravo, M. J., Ruau, D., Han, D. W., Zenke, M., et al. (2008b). Pluripotent stem cells induced from adult neural stem cells by reprogramming with two factors. Nature *454*, 646–650.

Kim, Y. K., Furic, L., DesGroseillers, L., and Maquat, L. E. (2005). Mammalian Staufen1 Recruits Upf1 to Specific mRNA 3'UTRs so as to Elicit mRNA Decay. Cell *120*, 195–208.

Kinoshita, M., Era, T., Jakt, L. M., and Nishikawa, S.-I. (2009). The novel protein kinase Vlk is essential for stromal function of mesenchymal cells. Development *136*, 2069–2079.

Kohavi, R., and John, G. H. (1997). Wrappers for feature subset selection. Artificial Intelligence *97*, 273–324.

Kohonen, T. (2001). Self-Organizing Maps 3rd ed. (Berlin [u.a.]: Springer).

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res *19*, 1639–1645.

Kunisato, A., Wakatsuki, M., Kodama, Y., Shinba, H., Ishida, I., and Nagao, K. (2010). Generation of induced pluripotent stem cells by efficient reprogramming of adult bone marrow cells. Stem Cells Dev. *19*, 229–238.

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. Nucleic Acids Research *37*, 4181–4193.

Lai, W. R., Johnson, M. D., Kucherlapati, R., and Park, P. J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. Bioinformatics *21*, 3763–3770.

Lamanna, A. C., and Karbstein, K. (2009). Nob1 binds the single-stranded cleavage site D at the 3'-end of 18S rRNA with its PIN domain. Proceedings of the National Academy of Sciences *106*, 14259–14264.

Lambert, J. F., Benoit, B. O., Colvin, G. A., Carlson, J., Delville, Y., and Quesenberry, P. J. (2000). Quick sex determination of mouse fetuses. J. Neurosci. Methods *95*, 127–132.

Laurent, L. C., Ulitsky, I., Slavin, I., Tran, H., Schork, A., Morey, R., Lynch, C., Harness, J. V., Lee, S., Barrero, M. J., et al. (2011). Dynamic changes in the copy number of pluripotency and cell proliferation genes in human ESCs and iPSCs during reprogramming and time in culture. Cell Stem Cell *8*, 106–118.

Lee, J. M., and Sonnhammer, E. L. L. (2003). Genomic gene clustering analysis of pathways in eukaryotes. Genome Res *13*, 875–882.

Lee, J. W., Lee, J. B., Park, M., and Song, S. H. (2005). An extensive comparison of recent classification tools applied to microarray data. Computational Statistics & Data Analysis *48*, 869–885.

Lefebvre, L., Dionne, N., Karaskova, J., Squire, J. A., and Nagy, A. (2001). Selection for transgene homozygosity in embryonic stem cells results in extensive loss of heterozygosity. Nat. Genet. *27*, 257–258.

Lefort, N., Feyeux, M., Bas, C., Féraud, O., Bennaceur-Griscelli, A., Tachdjian, G., Peschanski, M., and Perrier, A. L. (2008). Human embryonic stem cells reveal recurrent genomic instability at 20q11.21. Nat. Biotechnol *26*, 1364–1366.

Lengauer, C., Kinzler, K. W., and Vogelstein, B. (1997). Genetic instability in colorectal cancers. Nature *386*, 623–627.

Lenzi, L., Facchin, F., Piva, F., Giulietti, M., Pelleri, M. C., Frabetti, F., Vitale, L., Casadei, R., Canaider, S., Bortoluzzi, S., et al. (2011). TRAM (Transcriptome Mapper): database-driven creation and analysis of transcriptome maps from multiple sources. BMC Genomics *12*, 121.

Lercher, M., Urrutia, A. O., and Hurst, L. D. (2002). Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nature Genetics, 180–183.

Levin, A. M., Ghosh, D., Cho, K. R., and Kardia, S. L. R. (2005). A model-based scan statistic for identifying extreme chromosomal regions of gene expression in human tumors. Bioinformatics *21*, 2867–2874.

Li, H., Zhu, D., and Cook, M. (2008). A statistical framework for consolidating “sibling” probe sets for Affymetrix GeneChip data. BMC Genomics *9*, 188.

Li, T., Zhang, C., and Ogihara, M. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics *20*, 2429–2437.

Li, W., Wei, W., Zhu, S., Zhu, J., Shi, Y., Lin, T., Hao, E., Hayek, A., Deng, H., and Ding, S. (2009). Generation of Rat and Human Induced Pluripotent Stem Cells by Combining Genetic Reprogramming and Chemical Inhibitors. Cell Stem Cell *4*, 16–19.

Li, X., Thyssen, G., Beliakoff, J., and Sun, Z. (2006). The novel PIAS-like protein hZimp10 enhances Smad transcriptional activity. J. Biol. Chem. *281*, 23748–23756.

Liang, Q., Conte, N., Skarnes, W. C., and Bradley, A. (2008). Extensive genomic copy number variation in embryonic stem cells. Proc. Natl. Acad. Sci. U.S.A *105*, 17453–17456.

Liao, B.-Y., and Zhang, J. (2006). Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol. Biol. Evol. *23*, 530–540.

Liao, P.-sung, Chen, T.-sheng, and Chung, P.-choo (2001). A fast algorithm for multilevel thresholding. Journal of Information Science and Engineering *17*, 713–727.

Lingjaerde, O. C., Baumbusch, L. O., Liestol, K., Glad, I. K., and Borresen-Dale, A.-L. (2004). CGH-Explorer: a program for analysis of array-CGH data. Bioinformatics *21*, 821–822.

Lingjaerde, O. C., Baumbusch, L. O., Liestøl, K., Glad, I. K., and Børresen-Dale, A.-L. (2005). CGH-Explorer: a program for analysis of array-CGH data. Bioinformatics *21*, 821–822.

Lipshutz, R. J., Fodor, S. P., Gingeras, T. R., and Lockhart, D. J. (1999). High density synthetic oligonucleotide arrays. Nat. Genet. *21*, 20–24.

Liu, H., Zhu, F., Yong, J., Zhang, P., Hou, P., Li, H., Jiang, W., Cai, J., Liu, M., Cui, K., et al. (2008). Generation of Induced Pluripotent Stem Cells from Adult Rhesus Monkey Fibroblasts. Cell Stem Cell *3*, 587–590.

Liu, W.-m, Mei, R., Di, X., Ryder, T. B., Hubbell, E., Dee, S., Webster, T. A., Harrington, C. A., Ho, M.-h, Baid, J., et al. (2002). Analysis of high density expression microarrays with signed-rank call algorithms. Bioinformatics *18*, 1593–1599.

Liu, X., Wu, H., Loring, J., Hormuzdi, S., Disteche, C. M., Bornstein, P., and Jaenisch, R. (1997). Trisomy eight in ES cells is a common potential problem in gene targeting and interferes with germ line transmission. Dev. Dyn *209*, 85–91.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., et al. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat. Biotechnol. *14*, 1675–1680.

Loh, Y.-H., Wu, Q., Chew, J.-L., Vega, V. B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., et al. (2006). The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nat. Genet *38*, 431–440.

Long, S. B., Casey, P. J., and Beese, L. S. (2002). Reaction path of protein farnesyltransferase at atomic resolution. Nature *419*, 645–650.

Longo, L., Bygrave, A., Grosveld, F. G., and Pandolfi, P. P. (1997). The chromosome make-up of mouse embryonic stem cells is predictive of somatic and germ cell chimaerism. Transgenic Res *6*, 321–328.

Lyle, R., Gehrig, C., Neergaard-Henrichsen, C., Deutsch, S., and Antonarakis, S. E. (2004). Gene expression from the aneuploid chromosome in a trisomy mouse model of down syndrome. Genome Res *14*, 1268–1274.

MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability *1*, 281–297.

Maherali, N., Sridharan, R., Xie, W., Utikal, J., Eminli, S., Arnold, K., Stadtfeld, M., Yachechko, R., Tchieu, J., Jaenisch, R., et al. (2007). Directly Reprogrammed Fibroblasts Show Global Epigenetic Remodeling and Widespread Tissue Contribution. Cell Stem Cell *1*, 55–70.

Maitra, A., Arking, D. E., Shivapurkar, N., Ikeda, M., Stastny, V., Kassauei, K., Sui, G., Cutler, D. J., Liu, Y., Brimble, S. N., et al. (2005). Genomic alterations in cultured human embryonic stem cells. Nat. Genet. *37*, 1099–1103.

Martin, G. R. (1981). 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.

Martin, G. R. (1975). Teratocarcinomas as a model system for the study of embryogenesis and neoplasia. Cell *5*, 229–243.

Martin, G. R., and Evans, M. J. (1975). Differentiation of clonal lines of teratocarcinoma cells: formation of embryoid bodies in vitro. Proc. Natl. Acad. Sci. U.S.A. *72*, 1441–1445.

Martin, G. R., and Evans, M. J. (1974). The morphology and growth of a pluripotent teratocarcinoma cell line and its derivatives in tissue culture. Cell *2*, 163–172.

Masayesva, B. G., Ha, P., Garrett-Mayer, E., Pilkington, T., Mao, R., Pevsner, J., Speed, T., Benoit, N., Moon, C.-S., Sidransky, D., et al. (2004). Gene expression alterations over large chromosomal regions in cancers include multiple genes unrelated to malignant progression. Proc. Natl. Acad. Sci. U.S.A *101*, 8715–8720.

Masui, S., Nakatake, Y., Toyooka, Y., Shimosato, D., Yagi, R., Takahashi, K., Okochi, H., Okuda, A., Matoba, R., Sharov, A. A., et al. (2007). Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells. Nat. Cell Biol. *9*, 625–635.

Matsuda, T. (1999). STAT3 activation is sufficient to maintain an undifferentiated state of mouse embryonic stem cells. The EMBO Journal *18*, 4261–4269.

Matsumoto, M., and Nishimura, T. (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Transactions on Modeling and Computer Simulation *8*, 3–30.

Matzavinos, P. (2011). Development of a Workflow Componentfor Genome Analysis.

Mayshar, Y., Ben-David, U., Lavon, N., Biancotti, J.-C., Yakir, B., Clark, A. T., Plath, K., Lowry, W. E., and Benvenisty, N. (2010). Identification and classification of chromosomal aberrations in human induced pluripotent stem cells. Cell Stem Cell *7*, 521–531.

McClintick, J. N., and Edenberg, H. J. (2006). Effects of filtering by Present call on analysis of microarray experiments. BMC Bioinformatics *7*, 49.

Meaburn, K. J., and Misteli, T. (2007). Cell biology: Chromosome territories. Nature *445*, 379–781.

Michalak, P. (2008). Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomics *91*, 243–248.

Michaud, M., Barakat, S., Magnard, S., Rigal, D., and Baggetto, L. G. (2011). Leucine-rich protein 130 contributes to apoptosis resistance of human hepatocarcinoma cells. Int. J. Oncol. *38*, 169–178.

De Miguel, M. P., Fuentes-Julián, S., and Alcaina, Y. (2010). Pluripotent stem cells: origin, maintenance and induction. Stem Cell Rev *6*, 633–649.

Miki, K., and Eddy, E. M. (2002). Tumor necrosis factor receptor 1 is an ATPase regulated by silencer of death domain. Mol. Cell. Biol *22*, 2536–2543.

Mitalipova, M. M., Rao, R. R., Hoyer, D. M., Johnson, J. A., Meisner, L. F., Jones, K. L., Dalton, S., and Stice, S. L. (2005). Preserving the genetic integrity of human embryonic stem cells. Nat. Biotechnol *23*, 19–20.

Mitsui, K., Tokuzawa, Y., Itoh, H., Segawa, K., Murakami, M., Takahashi, K., Maruyama, M., Maeda, M., and Yamanaka, S. (2003). The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. Cell *113*, 631–642.

Morello, L. G., Hesling, C., Coltri, P. P., Castilho, B. A., Rimokh, R., and Zanchin, N. I. T. (2010). The NIP7 protein is required for accurate pre-rRNA processing in human cells. Nucleic Acids Research. Available at: http://nar.oxfordjournals.org/content/early/2010/08/26/nar.gkq758.abstract .

Morgan, W. F., Day, J. P., Kaplan, M. I., McGhee, E. M., and Limoli, C. L. (1996). Genomic Instability Induced by Ionizing Radiation. Radiation Research *146*, 247–258.

Myers, C. L., Dunham, M. J., Kung, S. Y., and Troyanskaya, O. G. (2004). Accurate detection of aneuploidies in array CGH and gene expression microarray data. Bioinformatics *20*, 3533–3543.

Na, J., Plews, J., Li, J., Wongtrakoongate, P., Tuuri, T., Feki, A., Andrews, P. W., and Unger, C. (2010). Molecular mechanisms of pluripotency and reprogramming. Stem Cell Research & Therapy *1*, 33.

Nagy, A., Rossant, J., Nagy, R., Abramow-Newerly, W., and Roder, J. C. (1993). Derivation of completely cell culture-derived mice from early-passage embryonic stem cells. Proc. Natl. Acad. Sci. U.S.A. *90*, 8424–8428.

Narva, E., Autio, R., Rahkonen, N., Kong, L., Harrison, N., Kitsberg, D., Borghese, L., Itskovitz-Eldor, J., Rasool, O., Dvorak, P., et al. (2010). High-resolution DNA analysis of human embryonic stem cell lines reveals culture-induced copy number changes and loss of heterozygosity. Nat Biotech *28*, 371–377.

Neri, T., Monti, M., Rebuzzini, P., Merico, V., Garagna, S., Redi, C. A., and Zuccotti, M. (2007). Mouse fibroblasts are reprogrammed to Oct-4 and Rex-1 gene expression and alkaline phosphatase activity by embryonic stem cell extracts. Cloning Stem Cells *9*, 394–406.

Nesbitt, M. N., and Francke, U. (1973). A system of nomenclature for band patterns of mouse chromosomes. Chromosoma *41*, 145–158.

Ng, H.-H., and Surani, M. A. (2011). The transcriptional and signalling networks of pluripotency. Nature Cell Biology *13*, 490–496.

Nichols, J., Evans, E. P., and Smith, A. G. (1990). Establishment of germ-line-competent embryonic stem (ES) cells using differentiation inhibiting activity. Development *110*, 1341–1348.

Nichols, J., Zevnik, B., Anastassiadis, K., Niwa, H., Klewe-Nebenius, D., Chambers, I., Schöler, H., and Smith, A. (1998). Formation of pluripotent stem cells in the

mammalian embryo depends on the POU transcription factor Oct4. Cell *95*, 379–391.

Nilsson, B., Johansson, M., Heyden, A., Nelander, S., and Fioretos, T. (2008). An improved method for detecting and delineating genomic regions with altered gene expression in cancer. Genome Biol *9*, R13.

Niwa, H., Burdon, T., Chambers, I., and Smith, A. (1998). Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3. Genes Dev. *12*, 2048–2060.

Niwa, H., Miyazaki, J., and Smith, A. G. (2000). Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. Nat. Genet. *24*, 372–376.

O'Neill, L. P., VerMilyea, M. D., and Turner, B. M. (2006). Epigenetic characterization of the early embryo with a chromatin immunoprecipitation protocol applicable to small cell populations. Nature Genetics *38*, 835–841.

Okita, K., Ichisaka, T., and Yamanaka, S. (2007). Generation of germline-competent induced pluripotent stem cells. Nature *448*, 313–317.

Okita, K., Nakagawa, M., Hyenjong, H., Ichisaka, T., and Yamanaka, S. (2008). Generation of mouse induced pluripotent stem cells without viral vectors. Science *322*, 949–953.

Otsu, N. (1979). A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man, and Cybernetics *9*, 62–66.

Pal, N. R., and Pal, S. K. (1993). A review on image segmentation techniques. Pattern Recognition *26*, 1277–1294.

Parisi, S., Passaro, F., Aloia, L., Manabe, I., Nagai, R., Pastore, L., and Russo, T. (2008). Klf5 is involved in self-renewal of mouse embryonic stem cells. J. Cell. Sci. *121*, 2629–2634.

Park, I.-H., Lerou, P. H., Zhao, R., Huo, H., and Daley, G. Q. (2008). Generation of human-induced pluripotent stem cells. Nature Protocols *3*, 1180–1186.

Peiffer, D. A., Le, J. M., Steemers, F. J., Chang, W., Jenniges, T., Garcia, F., Haden, K., Li, J., Shaw, C. A., Belmont, J., et al. (2006). High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res. *16*, 1136–1148.

Pellett, S., and Tracy, J. W. (2006). Mak16p is required for the maturation of 25S and 5.8S rRNAs in the yeast Saccharomyces cerevisiae. Yeast *23*, 495–506.

Pera, M. F., Bennett, W., and Cerretti, D. P. (1997). Expression of CD30 and CD30 ligand in cultured cell lines from human germ-cell tumors. Lab. Invest. *76*, 497–504.

Piddubnyak, V., Rigou, P., Michel, L., Rain, J.-C., Geneste, O., Wolkenstein, P., Vidaud, D., Hickman, J. A., Mauviel, A., and Poyet, J.-L. (2007). Positive regulation of apoptosis by HCA66, a new Apaf-1 interacting protein, and its putative role in

the physiopathology of NF1 microdeletion syndrome patients. Cell Death Differ *14*, 1222–1233.

Pignatelli, M., Serras, F., Moya, A., Guigó, R., and Corominas, M. (2009). CROC: finding chromosomal clusters in eukaryotic genomes. Bioinformatics *25*, 1552–1553.

Pillai, S., Rizwani, W., Li, X., Rawal, B., Nair, S., Schell, M. J., Bepler, G., Haura, E., Coppola, D., and Chellappan, S. (2011). ID1 facilitates the growth and metastasis of non-small cell lung cancer in response to nicotinic acetylcholine receptor and epidermal growth factor receptor signaling. Mol. Cell. Biol. *31*, 3052–3067.

Pollack, J. R., Sørlie, T., Perou, C. M., Rees, C. A., Jeffrey, S. S., Lonning, P. E., Tibshirani, R., Botstein, D., Børresen-Dale, A.-L., and Brown, P. O. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Proceedings of the National Academy of Sciences of the United States of America *99*, 12963–12968.

Prabhu, S., Ignatova, A., Park, S. T., and Sun, X. H. (1997). Regulation of the expression of cyclin-dependent kinase inhibitor p21 by E2A and Id proteins. Mol. Cell. Biol. *17*, 5888–5896.

Prendergast, G. C., Muller, A. J., Ramalingam, A., and Chang, M. Y. (2009). BAR the door: cancer suppression by amphiphysin-like genes. Biochim. Biophys. Acta *1795*, 25–36.

De Preter, K., Barriot, R., Speleman, F., Vandesompele, J., and Moreau, Y. (2008). Positional gene enrichment analysis of gene sets for high-resolution identification of overrepresented chromosomal regions. Nucleic Acids Res *36*, e43.

Van Prooijen-Knegt, A. C., Van Hoek, J. F., Bauman, J. G., Van Duijn, P., Wool, I. G., and Van der Ploeg, M. (1982). In situ hybridization of DNA sequences in human metaphase chromosomes visualized by an indirect fluorescent immunocytochemical procedure. Exp. Cell Res. *141*, 397–407.

Quinlan, A. R., Boland, M. J., Leibowitz, M. L., Shumilina, S., Pehrson, S. M., Baldwin, K. K., and Hall, I. M. (2011). Genome Sequencing of Mouse Induced Pluripotent Stem Cells Reveals Retroelement Stability and Infrequent DNA Rearrangement during Reprogramming. Cell Stem Cell *9*, 366–373.

Rebuzzini, P., Neri, T., Mazzini, G., Zuccotti, M., Redi, C. A., and Garagna, S. (2008a). Karyotype analysis of the euploid cell population of a mouse embryonic stem cell line revealed a high incidence of chromosome abnormalities that varied during culture. Cytogenet. Genome Res *121*, 18–24.

Rebuzzini, P., Neri, T., Zuccotti, M., Redi, C. A., and Garagna, S. (2008b). Chromosome number variation in three mouse embryonic stem cell lines during culture. Cytotechnology *58*, 17–23.

Rebuzzini, P., Pignalosa, D., Mazzini, G., Di Liberto, R., Coppola, A., Terranova, N., Magni, P., Redi, C. A., Zuccotti, M., and Garagna, S. (2011). Mouse embryonic stem cells that survive γ-rays exposure maintain pluripotent differentiation potential and

genome stability. Journal of Cellular Physiology. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21732352 [Accessed November 29, 2011].

Reubinoff, B. E., Pera, M. F., Fong, C. Y., Trounson, A., and Bongso, A. (2000). Embryonic stem cell lines from human blastocysts: somatic differentiation in vitro. Nat. Biotechnol. *18*, 399–404.

Robertson, E. J., Evans, M. J., and Kaufman, M. H. (1983). X-chromosome instability in pluripotential stem cell lines derived from parthenogenetic embryos. J Embryol Exp Morphol *74*, 297–309.

Robertson, E., Bradley, A., Kuehn, M., and Evans, M. (1986). Germ-line transmission of genes introduced into cultured pluripotential cells by retroviral vector. Nature *323*, 445–448.

Rodriguez, E., Houldsworth, J., Reuter, V. E., Meltzer, P., Zhang, J., Trent, J. M., Bosl, G. J., and Chaganti, R. S. (1993). Molecular cytogenetic analysis of i(12p)-negative human male germ cell tumors. Genes Chromosomes Cancer *8*, 230–236.

Ruifang, L., and Visser, H. M. (2010). Gene Expression Profiling to Predict Clinical Outcome of Breast Cancer - reproducing, analyzing and extending the Nature publication by Van't Veer et al. (Eindhoven: Philips Research Europe) Available at: http://repository.tudelft.nl/assets/uuid:1a8c7b1c-3152-4249-9994-bf2085e333a6/TN-2010-00626.pdf.

Saito, K., Abe, H., Nakazawa, M., Irokawa, E., Watanabe, M., Hosoi, Y., Soma, M., Kasuga, K., Kojima, I., and Kobayashi, M. (2010). Cloning of complementary DNAs encoding structurally related homeoproteins from preimplantation mouse embryos: their involvement in the differentiation of embryonic stem cells. Biol. Reprod *82*, 687–697.

Sano, H., Roach, W. G., Peck, G. R., Fukuda, M., and Lienhard, G. E. (2008). Rab10 in insulin-stimulated GLUT4 translocation. Biochem J *411*, 89–95.

Sarantidis, I. (2008). Algorithms to Explore the Chromosomal Clustering of Genes.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. Science *270*, 467–470.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O., and Davis, R. W. (1996). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc. Natl. Acad. Sci. U.S.A. *93*, 10614–10619.

Schnedl, W., Dann, O., and Schweizer, D. (1980). Effects of counterstaining with DNA binding drugs on fluorescent banding patterns of human and mammalian chromosomes. Eur. J. Cell Biol. *20*, 290–296.

Schoch, C., Kohlmann, A., Dugas, M., Kern, W., Schnittger, S., and Haferlach, T. (2006). Impact of trisomy 8 on expression of genes located on chromosome 8 in different AML subgroups. Genes Chromosomes Cancer *45*, 1164–1168.

Schöler, H. R., Ruppert, S., Suzuki, N., Chowdhury, K., and Gruss, P. (1990). New type of POU domain in germ line-specific protein Oct-4. Nature *344*, 435–439.

Schröck, E., du Manoir, S., Veldman, T., Schoell, B., Wienberg, J., Ferguson-Smith, M. A., Ning, Y., Ledbetter, D. H., Bar-Am, I., Soenksen, D., et al. (1996). Multicolor spectral karyotyping of human chromosomes. Science *273*, 494–497.

Sharan, R., Maron-Katz, A., and Shamir, R. (2003). CLICK and EXPANDER: a system for clustering and visualizing gene expression data. Bioinformatics *19*, 1787–1799.

Sharov, A. A., Masui, S., Sharova, L. V., Piao, Y., Aiba, K., Matoba, R., Xin, L., Niwa, H., and Ko, M. S. H. (2008). 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.

Shimosato, D., Shiki, M., and Niwa, H. (2007). Extra-embryonic endoderm cells derived from ES cells induced by GATA Factors acquire the character of XEN cells. BMC Developmental Biology *7*, 80.

Silva, J., Nichols, J., Theunissen, T. W., Guo, G., van Oosten, A. L., Barrandon, O., Wray, J., Yamanaka, S., Chambers, I., and Smith, A. (2009). Nanog is the gateway to the pluripotent ground state. Cell *138*, 722–737.

Simon, R., Radmacher, M. D., Dobbin, K., and McShane, L. M. (2003). Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J. Natl. Cancer Inst *95*, 14–18.

Skotheim, R. I., Monni, O., Mousses, S., Fosså, S. D., Kallioniemi, O.-P., Lothe, R. A., and Kallioniemi, A. (2002). New insights into testicular germ cell tumorigenesis from gene expression profiling. Cancer Res. *62*, 2359–2364.

Skylaki, S. (2007). Transcriptional Karyotyping of Stem Cell Gene Expression Data.

Smith, A. G. (2001). Stem cell biology (Cold Spring Harbor NY: Cold Spring Harbor Laboratory Press).

Smith, A. G., and Hooper, M. L. (1987). Buffalo rat liver cells produce a diffusible activity which inhibits the differentiation of murine embryonal carcinoma and embryonic stem cells. Dev. Biol. *121*, 1–9.

Smith, A. G., Heath, J. K., Donaldson, D. D., Wong, G. G., Moreau, J., Stahl, M., and Rogers, D. (1988). Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides. Nature *336*, 688–690.

Snijders, A. M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A. K., Huey, B., Kimura, K., et al. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. Nat. Genet. *29*, 263–264.

Sokal, R. R., and Michener, C. D. (1958). A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin, 1409–1438.

Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., Döhner, H.,

Cremer, T., and Lichter, P. (1997). Matrix-based comparative genomic hybridization: Biochips to screen for genomic imbalances. Genes Chromosom. Cancer *20*, 399–407.

Speicher, M. R., Gwyn Ballard, S., and Ward, D. C. (1996). Karyotyping human chromosomes by combinatorial multi-fluor FISH. Nat. Genet. *12*, 368–375.

Spits, C., Mateizel, I., Geens, M., Mertzanidou, A., Staessen, C., Vandeskelde, Y., Van der Elst, J., Liebaers, I., and Sermon, K. (2008). Recurrent chromosomal abnormalities in human embryonic stem cells. Nat Biotech *26*, 1361–1363.

Stadtfeld, M., Brennand, K., and Hochedlinger, K. (2008a). Reprogramming of Pancreatic β Cells into Induced Pluripotent Stem Cells. Current Biology *18*, 890–894.

Stadtfeld, M., Nagaya, M., Utikal, J., Weir, G., and Hochedlinger, K. (2008b). Induced pluripotent stem cells generated without viral integration. Science *322*, 945–949.

Statnikov, A., Aliferis, C. F., Tsamardinos, I., Hardin, D., and Levy, S. (2004). A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics *21*, 631–643.

Statnikov, A., and Aliferis, C. F. (2007). Are random forests better than support vector machines for microarray-based cancer classification? AMIA Annu Symp Proc, 686–690.

Stitson, M. O., Weston, J. A. E., Gammerman, A., Vovk, V., and Vapnik, V. (1996). Theory of Support Vector Machines (Surrey, England: Department of Computer Science, Royal Holloway University of London) Available at: http://ynucc.yu.ac.kr/~shkwon/lectures/ic/svm/svm_1.pdf.

Stransky, N., Vallot, C., Reyal, F., Bernard-Pierrot, I., de Medina, S. G. D., Segraves, R., de Rycke, Y., Elvin, P., Cassidy, A., Spraggon, C., et al. (2006). Regional copy number-independent deregulation of transcription in cancer. Nat. Genet *38*, 1386–1396.

Suda, Y., Suzuki, M., Ikawa, Y., and Aizawa, S. (1987). Mouse embryonic stem cells exhibit indefinite proliferative potential. J. Cell. Physiol. *133*, 197–201.

Suganami, E., Takagi, H., Ohashi, H., Suzuma, K., Suzuma, I., Oh, H., Watanabe, D., Ojima, T., Suganami, T., Fujio, Y., et al. (2004). Leptin stimulates ischemia-induced retinal neovascularization: possible role of vascular endothelial growth factor expressed in retinal endothelial cells. Diabetes *53*, 2443–2448.

Sugawara, A., Goto, K., Sotomaru, Y., Sofuni, T., and Ito, T. (2006). Current status of chromosomal abnormalities in mouse embryonic stem cell lines used in Japan. Comp. Med *56*, 31–34.

Suh, K. S., Mutoh, M., Nagashima, K., Fernandez-Salas, E., Edwards, L. E., Hayes, D. D., Crutchley, J. M., Marin, K. G., Dumont, R. A., Levy, J. M., et al. (2004). The Organellular Chloride Channel Protein CLIC4/mtCLIC Translocates to the Nucleus in Response to Cellular Stress and Accelerates Apoptosis. Journal of

Biological Chemistry *279*, 4632–4641.

Sumner, A. T. (1972). A simple technique for demonstrating centromeric heterochromatin. Exp. Cell Res. *75*, 304–306.

Sveinbojornsson, J. I. (2010). Genomic Clustering of Gene Expression Data.

Tada, M., Takahama, Y., Abe, K., Nakatsuji, N., and Tada, T. (2001). Nuclear reprogramming of somatic cells by in vitro hybridization with ES cells. Current Biology *11*, 1553–1558.

Takahashi, K., and Yamanaka, S. (2006). Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. Cell *126*, 663–676.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. Cell *131*, 861–872.

Tang, Y.-C., Williams, B. R., Siegel, J. J., and Amon, A. (2011). Identification of aneuploidy-selective antiproliferation compounds. Cell *144*, 499–512.

Tarca, A. L., Carey, V. J., Chen, X.-wen, Romero, R., and Drăghici, S. (2007). Machine Learning and Its Applications to Biology. PLoS Computational Biology *3*, e116.

Tejedo, J. R., Tapia-Limonchi, R., Mora-Castilla, S., Cahuana, G. M., Hmadcha, A., Martin, F., Bedoya, F. J., and Soria, B. (2010). Low concentrations of nitric oxide delay the differentiation of embryonic stem cells and promote their survival. Cell Death and Disease *1*, e80.

Tesar, P. J., Chenoweth, J. G., Brook, F. A., Davies, T. J., Evans, E. P., Mack, D. L., Gardner, R. L., and McKay, R. D. G. (2007). New cell lines from mouse epiblast share defining features with human embryonic stem cells. Nature *448*, 196–199.

Thomson, J. A., Itskovitz-Eldor, J., Shapiro, S. S., Waknitz, M. A., Swiergiel, J. J., Marshall, V. S., and Jones, J. M. (1998). Embryonic stem cell lines derived from human blastocysts. Science *282*, 1145–1147.

Tibshirani, R. J., Hastie, T., Narasimhan, B., and Chu, G. (2011). Prediction Analysis of Microarrays for R Available at: http://cran.r-project.org/web/packages/pamr/index.html.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl. Acad. Sci. U.S.A *99*, 6567–6572.

Tichy, E. D. (2011). Mechanisms maintaining genomic integrity in embryonic stem cells and induced pluripotent stem cells. Experimental Biology and Medicine *236*, 987–996.

Toedling, J., Schmeier, S., Heinig, M., Georgi, B., and Roepcke, S. (2005). MACAT--microarray chromosome analysis tool. Bioinformatics *21*, 2112–2113.

Torres, E. M., Sokolsky, T., Tucker, C. M., Chan, L. Y., Boselli, M., Dunham, M. J., and

Amon, A. (2007). Effects of aneuploidy on cellular physiology and cell division in haploid yeast. Science *317*, 916–924.

Torres, E. M., Williams, B. R., and Amon, A. (2008). Aneuploidy: cells losing their balance. Genetics *179*, 737–746.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. U.S.A *98*, 5116–5121.

Valor, L. M., and Grant, S. G. N. (2007). Clustered Gene Expression Changes Flank Targeted Gene Loci in Knockout Mice. PLoS ONE *2*, e1303.

Vapnik, V. (1979). Estimation of Dependences Based on Empirical Data: Empirical Inference Science (Information Science and Statistics) (Moscow: Nauka).

Vapnik, V. (1998). Statistical learning theory (New York [u.a.]: Wiley).

Veltman, J. A., Jonkers, Y., Nuijten, I., Janssen, I., van der Vliet, W., Huys, E., Vermeesch, J., Van Buggenhout, G., Fryns, J.-P., Admiraal, R., et al. (2003). Definition of a Critical Region on Chromosome 18 for Congenital Aural Atresia by ArrayCGH. The American Journal of Human Genetics *72*, 1578–1584.

Vrieling, H., Wijnhoven, S., van Sloun, P., Kool, H., Giphart-Gassler, M., and van Zeeland, A. (1999). Heterozygous Aprt mouse model: detection and study of a broad range of autosomal somatic mutations in vivo. Environ. Mol. Mutagen. *34*, 84–89.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10*, 57–63.

Werbowetski-Ogilvie, T. E., Bossé, M., Stewart, M., Schnerch, A., Ramos-Mejia, V., Rouleau, A., Wynder, T., Smith, M.-J., Dingwall, S., Carter, T., et al. (2009). Characterization of human embryonic stem cells with features of neoplastic progression. Nature Biotechnology *27*, 91–97.

Wernig, M., Meissner, A., Foreman, R., Brambrink, T., Ku, M., Hochedlinger, K., Bernstein, B. E., and Jaenisch, R. (2007). In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. Nature *448*, 318–324.

Weston, J. A. E., Stitson, M. O., Gammerman, A., Vovk, V., and Vapnik, V. (1996). Experiments with Support Vector Machines (London: Royal Holloway University of London).

Wineinger, N. E., Kennedy, R. E., Erickson, S. W., Wojczynski, M. K., Bruder, C. E., and Tiwari, H. K. (2008). Statistical issues in the analysis of DNA Copy Number Variations. Int J Comput Biol Drug Des *1*, 368–395.

Wu, H., Kim, K. J., Mehta, K., Paxia, S., Sundstrom, A., Anantharaman, T., Kuraishy, A. I., Doan, T., Ghosh, J., Pyle, A. D., et al. (2008). Copy number variant analysis of human embryonic stem cells. Stem Cells *26*, 1484–1489.

Xu, B., Zhang, K., and Huang, Y. (2009). Lin28 modulates cell growth and associates with

a subset of cell cycle regulator mRNAs in mouse embryonic stem cells. RNA *15*, 357–361.

Xu, H., Lemischka, I., and Ma'ayan, A. (2010). SVM classifier to predict genes important for self-renewal and pluripotency of mouse embryonic stem cells. BMC Systems Biology *4*, 173.

Yamashita, T., Honda, M., Takatori, H., Nishino, R., Hoshino, N., and Kaneko, S. (2004). Genome-wide transcriptome mapping analysis identifies organ-specific gene expression patterns along human chromosomes. Genomics *84*, 867–875.

Yang, S., Lin, G., Tan, Y.-Q., Zhou, D., Deng, L.-Y., Cheng, D.-H., Luo, S.-W., Liu, T.-C., Zhou, X.-Y., Sun, Z., et al. (2008). Tumor progression of culture-adapted human embryonic stem cells during long-term culture. Genes Chromosomes Cancer *47*, 665–679.

Yang, Z. Q., Streicher, K. L., Ray, M. E., Abrams, J., and Ethier, S. P. (2006). Multiple interacting oncogenes on the 8p11-p12 amplicon in human breast cancer. Cancer Res *66*, 11632–11643.

Yeoh, E.-J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R., Behm, F. G., Raimondi, S. C., Relling, M. V., Patel, A., et al. (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell *1*, 133–143.

Yi, P., Nguyên, D. T., Higa-Nishiyama, A., Auguste, P., Bouchecareilh, M., Dominguez, M., Bielmann, R., Palcy, S., Liu, J. F., and Chevet, E. (2010). MAPK scaffolding by BIT1 in the Golgi complex modulates stress resistance. J. Cell. Sci *123*, 1060–1072.

Ying, Q. L., Nichols, J., Chambers, I., and Smith, A. (2003). BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3. Cell *115*, 281–292.

Ying, Q.-L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., and Smith, A. (2008). The ground state of embryonic stem cell self-renewal. Nature *453*, 519–523.

Yoshida, K., Chambers, I., Nichols, J., Smith, A., Saito, M., Yasukawa, K., Shoyab, M., Taga, T., and Kishimoto, T. (1994). Maintenance of the pluripotential phenotype of embryonic stem cells through direct activation of gp130 signalling pathways. Mechanisms of Development *45*, 163–171.

Yu, J., Vodyanik, M. A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J. L., Tian, S., Nie, J., Jonsdottir, G. A., Ruotti, V., Stewart, R., et al. (2007). Induced Pluripotent Stem Cell Lines Derived from Human Somatic Cells. Science *318*, 1917–1920.

Zalzman, M., Falco, G., Sharova, L. V., Nishiyama, A., Thomas, M., Lee, S.-L., Stagg, C. A., Hoang, H. G., Yang, H.-T., Indig, F. E., et al. (2010). Zscan4 regulates telomere elongation and genomic stability in ES cells. Nature *464*, 858–863.

Zappone, M. V., Galli, R., Catena, R., Meani, N., De Biasi, S., Mattei, E., Tiveron, C., Vescovi, A. L., Lovell-Badge, R., Ottolenghi, S., et al. (2000). Sox2 regulatory sequences

direct expression of a (beta)-geo transgene to telencephalic neural stem cells and precursors of the mouse embryo, revealing regionalization of gene expression in CNS stem cells. Development *127*, 2367–2382.

Zhao, Z., Morstatter, F., Sharma, S., Alelyani, S., Anand, A., and Liu, H. (2010). Advancing Feature Selection Research - ASU Feature Selection Research (Tempe: School of Computing, Informatics, and Decision Systems Engineering, Arizona State University) Available at: http://featureselection.asu.edu/featureselection_techreport.pdf.

Zhu, Y., Shen, X., and Pan, W. (2009). Network-based support vector machine for classification of microarray samples. BMC Bioinformatics *10*, S21.

8. Appendix

In the interest of brevity, the supplemental material is provided in the attached CD. The inventory of supplemental files is as following:

1) The DI.S.C.O. application jar executable file,
2) The DI.S.C.O. application user manual,
3) Example dataset to run DI.S.C.O. The dataset includes:
    i) The Affymetrix GeneChip Mouse Genome 430 2.0 Array genome annotation,
    ii) RMA normalised gene expression values from the study of Kim et al. (2008) with GEO assession number GSE10806,
    iii) MAS5.0 calculated P/A flags for the GSE10806 dataset,
    iv) Mouse genome cytoband file,
    v) The normalisation scheme file for the GSE10806 dataset.
4) Tables:

Table S1: Detailed list of the mouse ESC and iPSC lines used in the present analysis with available annotation for each sample.

Table S2: List of identified chromosomal clusters of differentially expressed genes from the dendrogram-based PGE analysis. Information about the samples included in every comparison is also provided.

Table S3: List of differentially expressed genes identified by SAM analysis when comparing every sample with a chromosome 8-specific large-scale aberrations against every other sample.

Table S4: List of differentially expressed genes identified by SAM analysis when comparing every sample with a chromosome 11-specific large-scale aberrations against every other sample.

Table S5: List of differentially expressed genes identified by SAM analysis when comparing every sample with any type of large-scale aberration against every normal sample.