Published online 12 July 2012

Nucleic Acids Research, 2012, Vol. 40, No. 19 ***e153***
doi:10.1093/nar/gks663

Recurrent transcriptional clusters in the genome of mouse pluripotent stem cells

Stavroula Skylaki and Simon R. Tomlinson*

MRC Centre for Regenerative Medicine, Institute for Stem Cell Research, School of Biological Sciences, The University of Edinburgh, 5 Little France Drive, Edinburgh EH16 4UU, UK

Received April 23, 2012; Revised June 13, 2012; Accepted June 14, 2012

ABSTRACT

A number of studies have shown that transcriptome analysis in terms of chromosomal location can reveal regions of non-random transcriptional activity within the genome. Genomic clusters of differentially expressed genes can identify genomic patterns of structural organization, underlying copy number variations or long-range epigenetic regulation such as X-chromosome inactivation. Here we apply an integrative bioinformatics analysis to a collection of 315 freely available mouse pluripotent stem cell samples to discover transcriptional clusters in the genome. We show that over half of the analysed samples (56.83%) carry whole or partial-chromosome spanning clusters which recur in genomic regions previously implicated in chromosomal imbalances. Strikingly, we found that the presence of such large-clusters is linked to the differential expression of a limited number of genes, common to all samples carrying clusters irrespectively of the chromosome where the cluster is found. We have used these genes to train and test classification models that can predict samples that carry large-scale clusters on any chromosome with over 90% accuracy. Our findings suggest that there is a common downstream activation in these cells that affects a limited number of nodes. We propose that this effect is linked to selective advantage and identify potential driver genes.

INTRODUCTION

Approaches that take into account the chromosomal mapping of transcriptional data have been used in the past for the identification of general structural genomic features such as the regional clustering of 'housekeeping' genes (1) as well as transgenic insertions within cell lines (2), gross aneuploidy (3,4) and subtle chromosomal patterns around translocation breakpoints (5). Non-random changes in the expression levels of specific genomic regions can also be linked to the perturbation of normal epigenetic regulation, such as X-chromosome inactivation, or long-range epigenetic silencing in cancer (6,7).

Especially in the field of cancer biology, where karyotypic abnormalities are prevalent, a number of studies have described the quantitative relationship between copy number (CN) and gene expression which affects a great percentage of the genes in the aberrant regions (3,4,8,9). The widespread genomic instability in various cancer types can be a challenge for the researcher as it is often not possible to decipher which aberrations contribute to cancer growth and which are the downstream effect of a compromised genomic stability. As a result, the combined analysis of large collections of transcriptional and genomic data from microarray platforms has been thus far a common approach for discovering new oncogenes or tumour suppressors and distinguishing them from the functionally unrelated bystanders (10).

However, for the majority of published pluripotent stem cell experiments, large-scale integrated analysis of combined genomic and transcriptional data from a single sample is unattainable due to lack of available datasets. This is especially the case for model organisms besides human. For mouse pluripotent stem cells, for example, there is not a single large-scale study to-date that performs comparative analysis between genomic and transcriptional data. Two recent studies in human pluripotent stem cells have used gene expression data to identify patterns of chromosomal aberrations in embryonic stem cells (ESCs), induced pluripotent stem cells (iPSCs) and other multipotent cell types (11,12). These studies used a limited percentage of available array comparative genomic hybridization (aCGH) and single-nucleotide polymorphism (SNP) arrays to validate the observed patterns and extended the analysis to samples with no corresponding genomic data. This approach shows that by departing from the paradigm of the combined analysis, the interrogation of the large collection of readily available transcriptional data becomes

*To whom correspondence should be addressed. Tel: +44 131 651 9554; Fax: +44 131 651 9501; Email: simon.tomlinson@ed.ac.uk

possible. In addition, positional transcriptome analysis simultaneously informs on three different layers of information: genomic content, epigenome and transcriptional regulation.

In mouse ESCs, small clusters of differentially expressed (DE) genes have been identified around the pluripotency marker *Nanog* locus as a result of complex epigenetic regulation during development (13) and at the imprinted *Dlk1-Dio3* gene cluster during reprogramming due to epigenetic silencing (14). Moreover, recurring chromosomal aberrations have been primarily mapped to chromosomes 8 and 11 in mouse ESCs (15–17) and chromosome 8 and 14 in mouse iPSCs (18). Interestingly, frequent genomic alterations have been also reported in human ESCs, mapping primarily to chromosomes 12, 17 and X (19–22). Recently, it has been shown that human iPSCs also demonstrate compromised genomic integrity which is especially evident during the process of reprogramming (11,23–25). It has been suggested that specific aneuploidies tend to recur because of their ability to confer growth advantage and/or resistance to apoptosis and differentiation (26). When such 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 populations. Consistent with this hypothesis, mouse ESCs with a trisomy 8 have been found to outgrow normal cells with a diploid karyotype in competitive cultures (15).

Given the above mentioned evidence for positional transcriptional patterns in mouse pluripotent stem cells, we sought to investigate the chromosomal mapping of recurrent clusters of DE genes by analysing a large collection of samples. We hypothesise that, regardless of their molecular origins, recurrent clusters in multiple pluripotent stem cell populations are likely to be the result of positive selection. We used an integrative bioinformatics approach to identify candidate genes that may be driving the selection that has been previously associated with the presence of such patterns. Our findings provide evidence for a recurring set of DE genes in samples that contain large-scale clusters, independently of the genomic location of the clusters, and suggest a common downstream mechanism which may be associated with selective growth advantage.

METHODS

Data collection and processing

For the initial analysis phase, we have collected 481 public domain gene expression samples (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; http://www.ncbi.nlm.nih.gov/geo) and ArrayExpress (http://www.ebi.ac.uk/arrayexpress) public databases (see Figure 1A, Supplementary Table S1). The raw CEL files obtained were normalized using the Robust Multiple-Array Average (RMA) (27) and Present/Absent flags were extracted by the MAS5.0 algorithm (28), both methods from the 'affy' package of

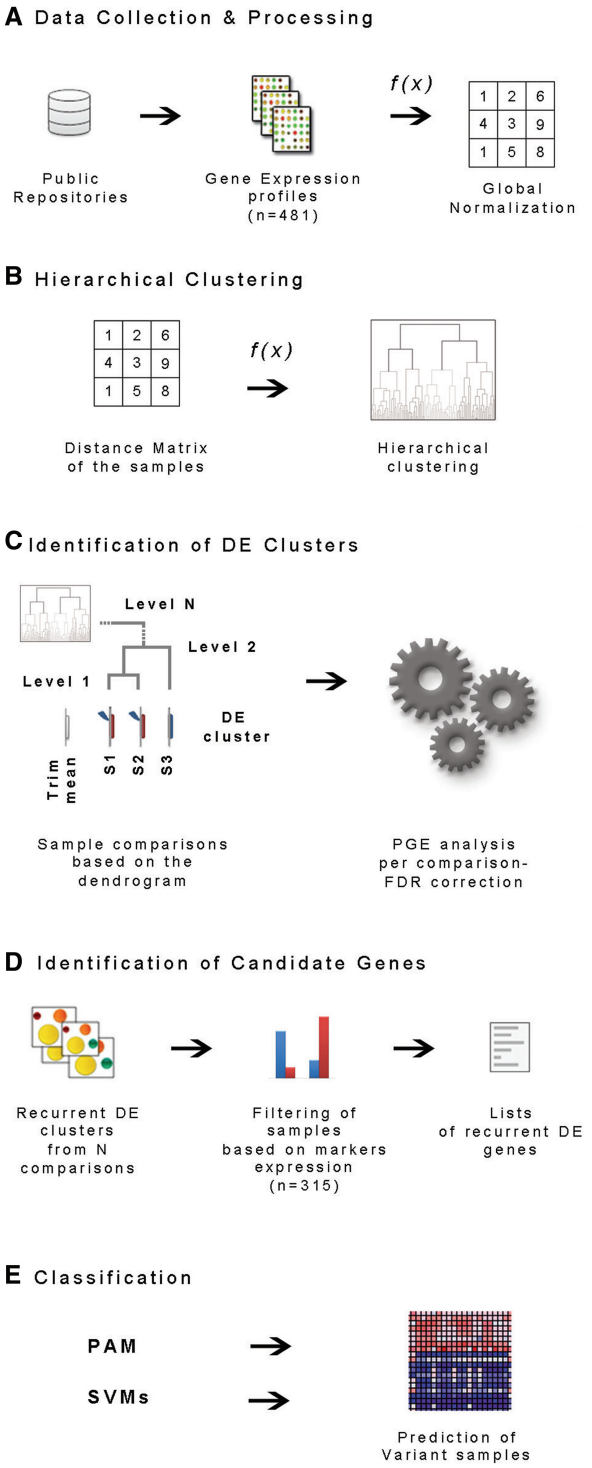

Figure 1. The integrative analysis workflow. (**A**) Collection and global normalization of 481 publicly available samples. (**B**) Pearson's correlation derived distance matrix and agglomerative hierarchical clustering with average linkage of the normalized data. (**C**) PGE analysis with MultiLevel Otsu thresholding for identification of recurrent aberrant localized expression across the dendrogram. (**D**) Catalogue of recurrent DE clusters. Filtering of samples according to the expression of pluripotency and lineage-specific markers resulting to the *Nanog*-high subgroup of 315 pluripotent samples. Identification of DE genes between the *Nanog*-high *Normal* versus *Variant* group, the *Normal-Chr8* versus *Variant-Chr8* groups and the *Normal-Chr11* versus *Variant-Chr11* groups. (**E**) Training and testing of classification models using PAM and SVMs for the prediction of *Variant* samples.

the Bioconductor suite (http://www.bioconductor.org/) in the R statistical environment (29).

Hierarchical clustering of samples

In order to obtain a measure of similarity between samples and subsequently groups of samples, we have performed agglomerative hierarchical clustering with average linkage, using a distance matrix based on the Pearson's correlation of the samples (Figure 1B). Large data collections, such as the one analysed in this manuscript, may present variations due to differences in RNA quality or hybridization processing, culture conditions or experimental treatments between different labs (30). In order to account for this complexity, we designed an iterative strategy where each sample (or group of samples after the leaf nodes of the dendrogram) is compared with the sample (or group of samples) with the most correlated transcriptome available in the matrix in a branch-wise manner, according to the dendrogram obtained from the hierarchical clustering (Figure 1C). This approach can reveal the unique subtle changes of each sample that differentiate it from its most similar neighbour. It therefore deviates from previous methodologies 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 (11,12).

Identification of DE clusters

In order to identify clusters of DE genes, we have considered all the probesets for which genome mapping annotation was available (43 109 probesets in total). Multiple probesets for a single gene were replaced with their median value resulting in 26 524 probesets. For each comparison in the dendrogram, we estimated suitable fold change (FC) cut-off values for differential expression by applying a novel approach, the MultiLevel Otsu method used in image processing (31). The average cut-offs used across the dendrogram were >1.56 FC for over-expressed genes and <0.66 FC for under-expressed. In addition, for each comparison, we filtered out probesets which were absent in more than 50% of the samples in the comparison. Next, we used the Positional Gene Enrichment (PGE) algorithm (32) to identify clusters of DE genes (Figure 1C). Briefly, PGE uses an adaptive genomic window approach to identify chromosomal regions that are over-represented in user provided gene lists. We have implemented the PGE algorithm in Java and run the method with the lists of all up-regulated and all down-regulated genes from the previous step. We used the rank position of each probeset on the chromosome instead of its physical coordinates in order to minimize regional biases due to gene-dense regions or gene deserts. For each comparison of samples, the PGE algorithm corrects the *P*-value of the discovered clusters for multiple testing using the False Discovery Rate (FDR) (33). We filtered out clusters with an adjusted *P*-value ≤ 0.01. To additionally assess the statistical significance of the predicted clusters across the whole dendrogram, we calculated an empirical FDR based on randomization by generating 1000 permutations of randomized genomic mappings of the FC values, keeping the dendrogram topography constant. Finally, once the specific chromosomal clusters were discovered, the global trimmed mean of each gene was used to predict the type of cluster, i.e. up- or down-regulation (the 0.05% of outlier expression values per gene was discarded). The final list of clusters was filtered for an adjusted *P*-value $<$ 1.0E-4 and cluster size of at least 10 DE genes (Figure 2).

Visual inspection and validation

We visually inspected the chromosomal clusters by plotting the rank position of each gene across the chromosomes using Di.S.C.O. (Discovery of Subtle Clustered Organization), a custom-developed software tool (Skylaki *et al.*, in preparation). Expression levels were presented by a colour gradient defined by the MultiLevel Otsu-derived thresholds, whereas each gene was represented by the median value of all its corresponding probesets (see Supplementary Figure S1).

Selection of pluripotent ESCs and iPSCs samples based on markers expression

To distinguish between mouse ESCs, iPSCs and their differentiating or partially reprogrammed counterparts, we examined the available sample annotation and the expression of hallmark pluripotency genes such as *Nanog* (34,35), as well as a range of differentiation markers (Figure 1D). It can be hypothesized that the high expression of pluripotency markers in combination with low expression of lineage-specific genes reflects cell populations rich in pluripotent stem cells. This filtering step was essential in order to focus on transcriptional changes that are specific in pluripotent stem cells and not the obvious result of cell mixtures in different stages of differentiation or reprogramming. The resulting subset of 315 homogeneous pluripotent populations (272 ESC and 43 iPSC samples), from here on referred to as *Nanog*-high samples, was used at the final stage of the analysis for the identification of recurring DE genes in samples that carry DE clusters as well as the training and testing of classification models as presented hereafter.

Differential gene expression analysis

As mentioned previously, we were specifically interested in analysing the positional transcriptional patterns of the *Nanog*-high subgroup which more closely represents the pluripotent state. In addition, we focused on whole- or partial-chromosome spanning clusters which are likely to reflect underlying aneuploidies since co-regulation of large genomic regions is not commonly observed as a result of transcriptional regulation. The 315 *Nanog*-high samples were divided in two groups: the group termed as *Normal* consists of samples where no large-scale DE clusters could be identified in the genome, whereas the group termed as *Variant* comprises of samples that bear large-scale chromosomal clusters of DE genes in at least one chromosome.

In order to determine whether there is a distinct transcriptional signature that can be associated with the

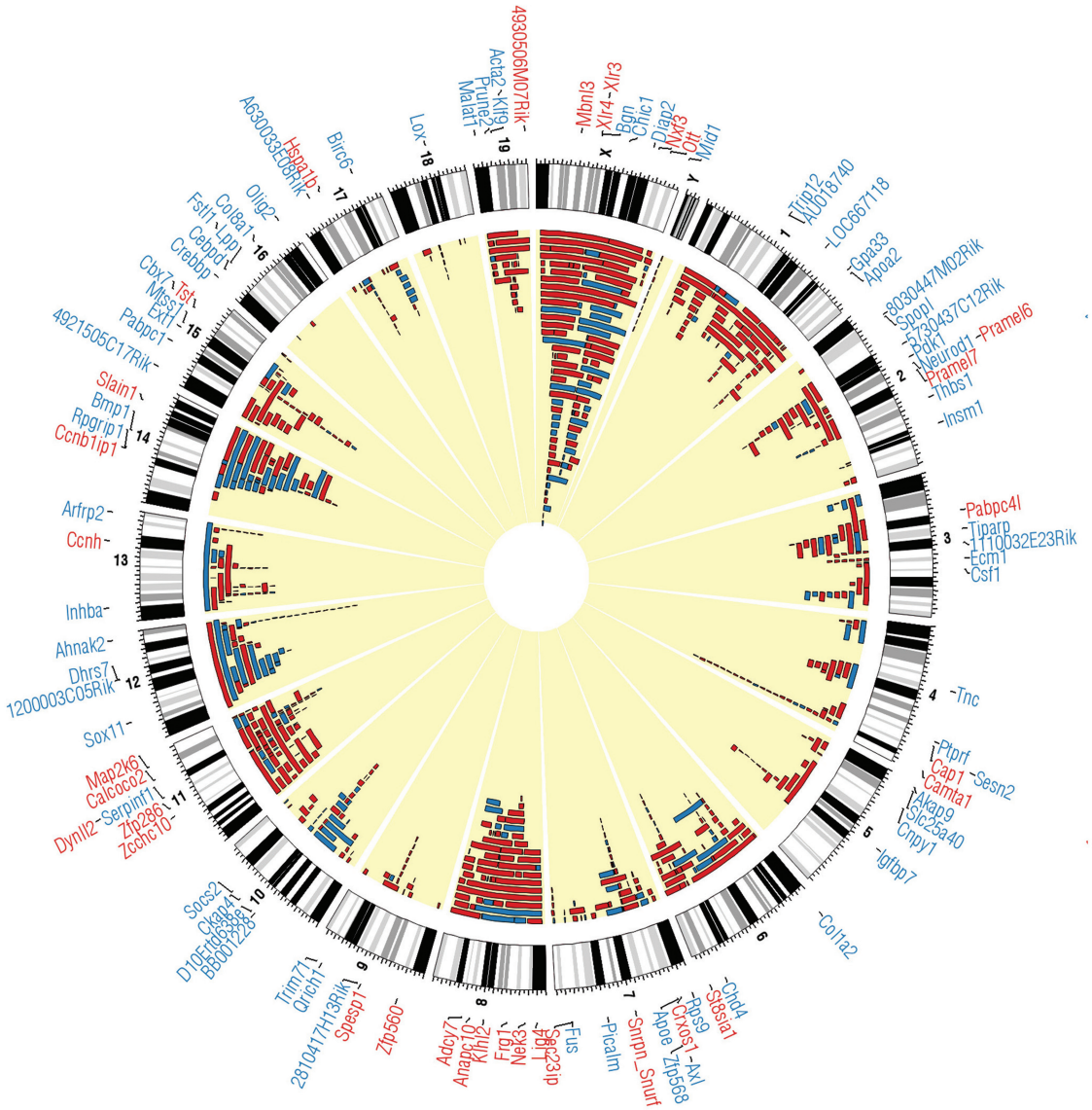

Figure 2. The circular karyotype of all predicted significantly over-expressed (red) and under-expressed (blue) DE clusters in the matrix and the genes that are DE between predicted *Normal* and *Variant* samples (red for up-regulated genes and blue for down-regulated). Larger effects observed in chromosomes 8, 11, 14 and X. For an example of the enhanced detection power of the approach, see also Supplementary Figure S1. For detailed description of the samples included in the analysis, see Supplementary Table S1.

presence of such large DE clusters in the *Nanog*-high *Variant* group, we performed differential expression analysis using a two-class Significance Analysis of Microarrays (SAMs) (36) (Figure 1D). The analysis was performed using the 'samr' package in R (500 permutations, FDR = 0.05). From this stage onwards, all probesets were considered and no replacement was performed, in order to account for the unique behaviour of each probeset which may represent alternative splicing or polyadenylation events. We compared (i) *Normal* versus *Variant* samples; (ii) samples with chromosome-8 specific patterns (*Variant-Chr8*) versus all other samples (*Normal-Chr8*); and (iii) samples with chromosome-11 specific patterns (*Variant-Chr11*) versus all other samples (*Normal-Chr11*). The *Normal-Chr8* and *–Chr11* groups also contained the rest of the samples that had DE clusters in any other chromosomes, besides 8 and 11 respectively. The lists of DE genes per comparison are presented in Supplementary Tables S2–S4 (with FC $\geq$ 1.5 and adjusted *P*-value $<$ 0.05). Chromosomes 8 and 11 were specifically chosen for this analysis because they are the chromosomes most frequently affected by aneuploidy and, in fact, 70% of the predicted *Variant* samples carry whole or partial-chromosome spanning clusters on at least one of these two chromosomes (see 'Results' section).

Classification

To investigate whether the set of DE genes common in samples that carry large DE clusters on any chromosome and in samples that carry chromosome-8 and -11 specific clusters can be predictive of the presence of such clusters,

we employed two well-established classification techniques: Prediction Analysis of Microarrays (PAMs) (37) and Support Vector Machines (SVMs) (38,39) (Figure 1E). PAM uses a nearest shrunken centroid approach to identify the genes that best separate between classes. We used the 'pamr' package in R (40). For the linear SVMs we used the 'e1071' package in R (41). 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 classes. For the construction of the SVM classifiers, we used a subset of 187 samples and 37 samples for training and validation, respectively. After selecting the best scoring classifier, we merged the training and validation subsets 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). Our decision was based on accuracy and F1 score, defined as follows:

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN),$$
$$\text{Precision} = TP/(TP+FP),$$
$$\text{Recall} = TP/(TP+FN),$$
$$\text{F1score} = 2 \times (\text{Precision} \times \text{Recall})/(\text{Precision}+\text{Recall})$$

where TP = True Positive, TN = True Negative, FP = False Positive and FN = False Negative.

For the chromosome-specific classifiers, we additionally accounted for differences in the number of input samples per class by adjusting the weight parameters of the SVM to be proportional to the number of samples in each class.

The Recursive Feature Elimination (RFE) method (42) was applied to linear SVMs to obtain small subsets of predictive genes.

GO analysis

Gene ontology (GO) enrichment (GOTERM_BP_5) was calculated using the DAVID functional annotation bioinformatics tool (43,44). For the GO analysis only probesets with $FC > 1.5$ and Q-value < 0.05 (from SAM) were considered. Enrichment significance was limited to a very stringent Benjamini–Hochberg adjusted P-value < 0.01.

RESULTS

A catalogue of DE clusters in mouse ESCs and iPSCs

The PGE analysis performed across the dendrogram generated a large set of DE clusters (Figure 2). The most prevalent recurring intervals that we have observed map to chromosomes 6, 8, 11, 14 and most commonly in chromosome X (Figure 2). It is plausible that a percentage of the observed clusters on chromosome X correspond to varying states of X chromosome inactivation (XCI), whereas others to DNA CN alterations. However, it should be noted that in mouse ESCs, all lines for which sample annotation was available (~70%) were annotated as male. Interestingly, 75.43% of the identified clusters are up-regulated, 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 have been reported to correspond to amplifications (72%) (45).

Focusing on the subgroup of the 315 Nanog-high samples (Figure 3A and B), we could identify whole or partial-chromosome spanning clusters in 179 samples, 56.83% of the group. We further validated these clusters by plotting the gene expression levels on the chromosomes (see 'Methods' section and Supplementary Figure S1). Large expression domains are good predictors of underlying aneuploidy. The percentage of samples that carry such large-scale clusters of DE genes in the Nanog-low subset is much lower (30%, Figure 3A) than the one in the Nanog-high subgroup (56.83%). This difference may reflect differences in the frequency of pluripotent cells in cultures or the inability to detect these subtle signatures in mixtures of differentiating cells such as the ones in the *Nanog*-low group. These findings are consistent with previous cytogenetic studies in mouse pluripotent stem cells which also highlight recurrent changes of chromosome 8, 11 and 14 (Figure 3C) (15,17,18). Our method additionally identified a high number of clusters in chromosomes 6 and X and frequently recurring pairs of large chromosomal clusters which tend to appear across many different experiments. The latter include clusters on 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 the specific percentages of predicted clusters per chromosome for the *Nanog*-high subgroup is presented in Figure 3E.

Finally, a comparison between ESC and iPSC-specific clusters on the autosomes revealed that in both cases more than half of the samples carry at least one large-scale chromosomal cluster (58% of samples for ESCs and 51% for iPSCs) (Supplementary Figure S2). Interestingly, chromosome 11 patterns are mostly present in ESCs. In iPSCs, the chromosome X changes, which are predicted gains or up-regulations, could reflect differences between male and female lines such as different states of XCI. Unfortunately, we were unable to obtain the annotation for the sex of the line for the majority of iPSC samples studies and thus, sex chromosomes have been excluded from further analysis.

Recurring DE genes in samples carrying large DE clusters

The SAM analysis performed between *Nanog*-high *Normal* and *Variant* groups, *Normal-Chr8* and *Variant-Chr8* groups and *Normal-Chr11* and *Variant-Chr11* groups revealed sets of DE genes for each comparison (see Supplementary Tables S2–S4). A heatmap representation of the top 50 DE genes from each comparison is presented in Figure 4A–C.

The presence of a recurring set of DE genes across all *Variant* samples suggests that there is a common downstream effect in these samples independent of the genomic

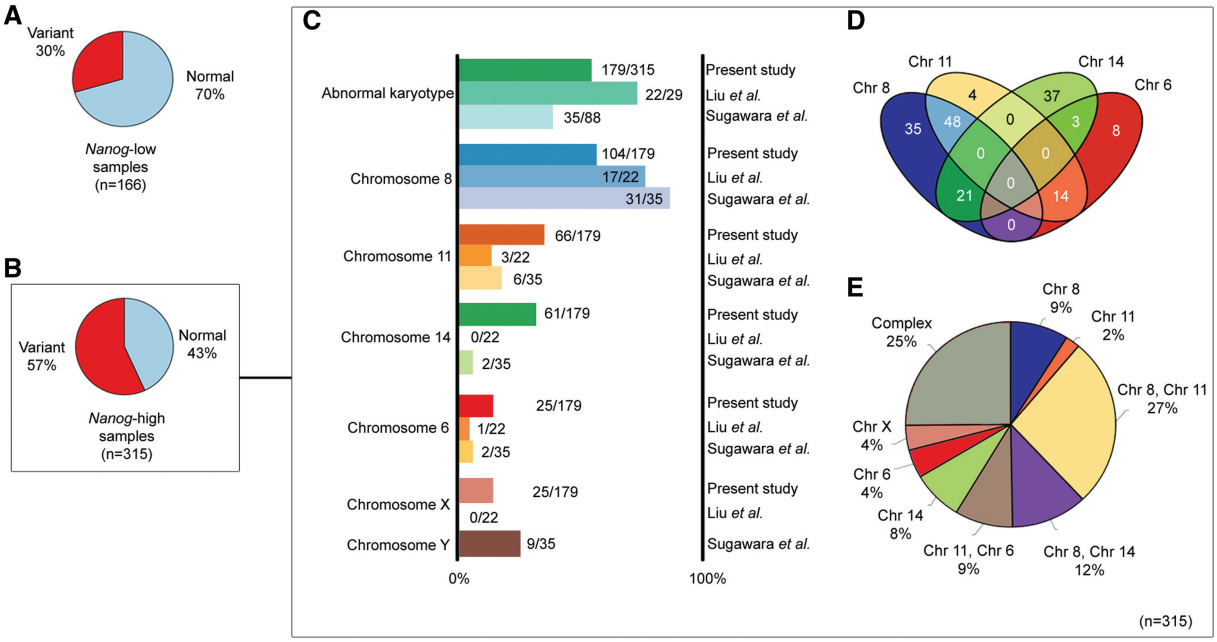

Figure 3. Description of the large-scale chromosome spanning DE clusters in the *Nanog*-high subgroup. (**A**) Percentages of *Variant* and *Normal* samples in the differentiating or partially reprogrammed *Nanog*-low group of samples ($n = 166$). (**B**) Percentages of *Variant* and *Normal* samples in the *Nanog*-high pluripotent group of samples ($n = 315$). The downstream analysis was focused on this subgroup of 315 samples. (**C**) Comparison of the frequencies of predicted abnormalities per chromosome in the present study and two independent cytogenetic studies of mouse ESCs (15,17). (**D**) Venn-diagram representing the co-occurrence of large DE clusters between chromosomes 6, 8, 11 and 14. Figure constructed in Venny (46). (**E**) Breakdown of percentages for the aberrant chromosomes and the associated aberrant chromosome pairs. For a detailed comparison between mouse ESCs and iPSCs, see also Supplementary Figure S2.

location of the DE cluster they carry. Importantly, the identified DE genes are not necessarily members of the identified clusters. We hypothesize that these cells operate under a positive selection mechanism the downstream consequences of which manifest at the transcriptional level despite their different types of DE clusters. The top up-regulated list (Table 1) is 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 (48). 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 (49), the non-homologous end-joining repair gene *Lig4* (50), the genome maintenance regulator *Zscan4* (51) as well as the cell-growth modulator *Lin28* (52). The function of these genes is consistent with the properties of genes expected to drive positive selection in competitive cultures.

Classification of samples carrying large DE clusters

Given the high percentage of samples in our analysis that carries large DE clusters and the presence of distinct set of DE genes in these samples, we investigated the prediction power of these sets by training classification models using PAM and SVMs. The results are presented in Table 2. In all three case studies, that is *Variant* (any type of cluster), *Variant-Chr8* and *Variant-Chr11*, we achieved predictive accuracy higher than 80% using linear SVM classifiers with just a limited number of DE genes.

Remarkably, by applying the RFE method, we could identify small subsets of candidate genes that demonstrate a high class prediction power. For the *Variant* set, the top 50 genes are sufficient to predict the presence of DE clusters with an accuracy of 91%. In the case of chromosome-specific SVMs it was possible to narrow our selection down to the top 10 ranked genes while still maintaining a high accuracy (over 80%). The top 10 up-regulated genes in the *Variant-Chr8* group include the anti-apoptotic *Bag4* as well as *Lsm1*, both described as breast cancer oncogenes in the 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 (53). Interestingly, a recent study has implicated another anti-apoptotic gene, *BCL2L1*, in conferring growth advantage to human pluripotent cells carrying the 20q11.21 amplicon (54). The Bag4 and Bcl-2 anti-apoptotic protein families interact to regulate cell survival (55). The up-regulation of different members of the anti-apoptotic pathways in both mouse (present study) and human (54) may indicate the existence of a common reserved path towards selective growth in both organisms.

Finally, a selection of solely non-chromosome 8 mapped genes could still be used to train the classifier for chromosome 8 clusters with up to 71% accuracy (Table 2). This result suggests that there is a non-chromosome 8-specific program that is affected by the presence of the DE cluster on chromosome 8, further supporting the evidence for a secondary mechanism independent of the chromosomal location of the clusters.

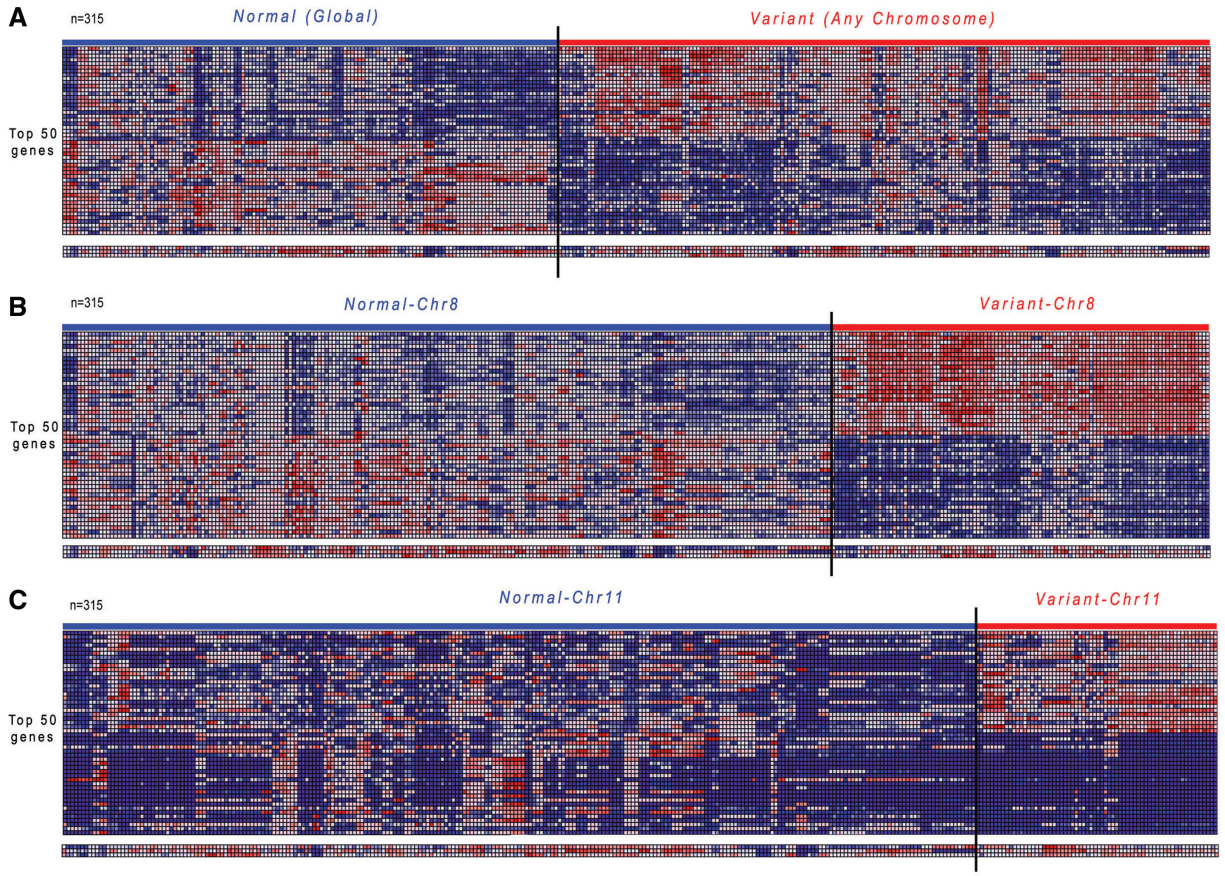

Figure 4. Heatmap representation of the top 50 genes generated from SAM analysis. The panel of the three core pluripotency genes (*Nanog*, *Pou5f1* (*Oct4*) and *Sox2*) at the bottom of each heatmap demonstrates the independency of the large DE clusters from the core pluripotency program in the stem cell populations. Figure constructed in GenePattern (47). (**A**) Heatmap of the global set where the *Variant* group consists of samples with any type of large-scale DE cluster. (**B**) Heatmap of the chromosome 8-specific set where the *Variant-Chr8* group consists of any sample with a chromosome 8-specific DE cluster. (**C**) Heatmap of the chromosome 11-specific set where the *Variant-Chr11* group consists of any sample with a chromosome 11-specific DE cluster. For the SAM-derived lists of DE genes for each comparison, refer to Supplementary Tables S2–S4.

Table 1. Functional categories of the top 50 over- and under-expressed genes in the *Variant* feature set

Functional category	Up-regulated genes (*Variant*)	Down-regulated genes (*Variant*)
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, St8sia1, Sec23ip, Zfp560, Sdc4, 666185, Glrx, Gprc5b*	*Acta2, Thbs1, Mid1, Tagln, Fstl1, Atrx, Prss23, Ptprf, Cd44, Cdk7, Hs6st2, Prtg, Pkdcc, LOC72520, F630007L15Rik, Axl, Fstl1, 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.

DISCUSSION

In summary, we have used a sensitive integrative method to analyse the transcriptome of the largest collection to-date of mouse ESCs and iPSCs samples. We were able to quantify the number of samples that carry a large-scale cluster of genes with concordant changes in expression levels and assign the greatest percentage of these intervals to chromosomes 8 and 11. These findings

Table 2. Performance of classifiers

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

Best performing classifiers (with bold we highlight the classifier trained with the top 50 features in each set). Feature selection was performed from the SAM output list by RFE. In the *RFE SVM—No Chr8 feature* set, genes mapped to chromosome 8 were excluded from the up-regulated list. Global: *Normal* and *Variant,* Chr8: *Normal-Chr8* and *Variant-Chr8*, Chr11: *Normal-Chr11* and *Variant-Chr11*.

are consistent with cytogenetic studies reporting recurrent aberrations on chromosomes 8 and 11 in murine ESC populations (15,17). A subset of the smaller recurrent intervals may be due to co-regulated functional gene clusters as has been previously observed for the *Nanog* locus in mouse ESCs (13), whose up-regulation was also detected in our analysis. The prediction power of the method and the large scale of the data analysis revealed a complex pattern of genomic regions which are prone to be concordantly DE, 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.

Our analysis shows that in a set of 315 pluripotent samples selected for high *Nanog* expression, 56.83% carry large-scale clusters of DE genes. As large-scale clusters of DE genes can be indicative of underlying aneuploidies, we hypothesise that the majority of these clusters, which overlap with previously reported hotspots of aneuploidy, can in fact be the effect of acquired chromosomal aberrations. The presence of such clusters is not a universal characteristic of normal pluripotent stem cells as the remaining 43.17% of the pluripotent samples carry no such large-scale changes and still demonstrate high expression of pluripotency markers. Therefore, these clusters are not essential for the survival of pluripotent stem cells under normal conditions but they rather may contribute towards the dominance of the affected cells in a selective culture environment, possibly through the deregulation of a small set of driver genes. It should be also noted that the majority of these clusters do not span the *Nanog* locus. A recent study has indicated that the occurrence of trisomy 12 in human iPSCs is a result of the up-regulation of the *NANOG-GDF3* cluster on chromosome 12 (11). The authors proposed that this is a likely mechanism for the driving the aneuploidy, since over-expression of *NANOG* leads to enhanced self-renewal. Such an effect may be possible in the presence of *NANOG*-spanning clusters, however, in our data, there is a great number of *Nanog*-high *Variant* samples, irrespectively of the genomic position of the DE cluster they carry. It is likely that a 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*.

The comparison between *Normal* and *Variant* profiles has revealed a set of DE genes highly connected to pluripotency, cell cycle and apoptosis. It has been proposed before that positive selection in culture can occur through multiple mechanisms, in particular via cell cycle progression and deregulation of the p53 pathway or activation of anti-apoptotic pathways (26). 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 CNs by modulating pathways involved in balancing protein stoichiometry such as ribosome biogenesis and protein degradation (56). A similar effect is observed in the case of chromosome 8 clusters which demonstrate enrichment in the GO categories related to RNA processing (Benjamini adjusted *P*-value = 2.23E-03).

Importantly, there is a recurring set of DE genes present in *Variant* samples, irrespectively of the genomic mapping of the cluster they carry. It is in fact possible to use this limited number of genes to train highly accurate classifiers in order to assess the transcriptional integrity of pluripotent cultures. We speculatively suggest that the presence of a recurring transcriptional signature indicates a downstream response mechanism that confers selective advantage to the affected cells and can be detected by the expression of a limited number of nodes. It could be additionally used for the identification of core pathways that can be subsequently targeted to develop anti-selective culture conditions for aneuploidy. Such an approach has been effectively applied in trisomic mouse embryonic fibroblasts (MEFs) and human cancer cell lines with compounds that are anti-selective for karyotypically abnormal cells (57).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–4 and Supplementary Figures 1 and 2.

ACKNOWLEDGEMENTS

The authors thank Dr V. Wilson, Prof. I. Chambers, Dr K. Kaji, F. Halbritter, F. Wymeersch and A. Spiliopoulou for discussions and revisions of the manuscript.

FUNDING

Biotechnology and Biological Sciences Research Council (BBSRC) and the EU seventh framework program EuroSyStem. Funding for open access charge: EU FP7 program EuroSyStem.

Conflict of interest statement. None declared.

REFERENCES

1. 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. *Nat. Genet.*, **31**, 180–183.
2. Valor,L.M. and Grant,S.G.N. (2007) Clustered gene expression changes flank targeted gene loci in knockout mice. *PLoS One*, **2**, e1303.
3. 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. *Proc. Natl Acad. Sci. USA*, **99**, 12963–12968.
4. 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 Res.*, **62**, 6240–6245.
5. 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.
6. 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.
7. 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.
8. 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. USA*, **101**, 8715–8720.
9. 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. *Gene. Chromosome. Canc.*, **46**, 75–86.
10. Huang,N., Shah,P.K. and Li,C. (2011) Lessons from a decade of integrating cancer copy number alterations with gene expression profiles. *Brief. Bioinformatics*, **13**, 305–316.
11. 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.
12. 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.
13. Levasseur,D.N., Wang,J., Dorschner,M.O., Stamatoyannopoulos,J.A. and Orkin,S.H. (2008) Oct4 dependence of chromatin structure within the extended Nanog locus in ES cells. *Genes Dev.*, **22**, 575–580.
14. Stadtfeld,M., Apostolou,E., Akutsu,H., Fukuda,A., Follett,P., Natesan,S., Kono,T., Shioda,T. and Hochedlinger,K. (2010) Aberrant silencing of imprinted genes on chromosome 12qF1 in mouse induced pluripotent stem cells. *Nature*, **465**, 175–181.
15. 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.
16. 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.
17. 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.
18. Chen,Q., Shi,X., Rudolph,C., Yu,Y., Zhang,D., Zhao,X., Mai,S., Wang,G., Schlegelberger,B. and Shi,Q. (2011) Recurrent trisomy and Robertsonian translocation of chromosome 14 in murine iPS cell lines. *Chromosome Res.*, **19**, 857–868.
19. 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.
20. 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.
21. 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.
22. 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.
23. 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.
24. 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.
25. 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.
26. 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.
27. Irizarry,R.A., Bolstad,B.M., Collin,F., Cope,L.M., Hobbs,B. and Speed,T.P. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, **31**, e15.
28. Hubbell,E., Liu,W.-M. and Mei,R. (2002) Robust estimators for expression analysis. *Bioinformatics*, **18**, 1585–1592.
29. Ihaka,R. and Gentleman,R. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Stat.*, **5**, 299–314.
30. Hegde,P., Qi,R., Abernathy,K., Gay,C., Dharap,S., Gaspard,R., Hughes,J.E., Snesrud,E., Lee,N. and Quackenbush,J. (2000) A concise guide to cDNA microarray analysis. *BioTechniques*, **29**, 548–550, 552–554, 556 passim.
31. Liao,P., Chen,T. and Chung,P. (2001) A fast algorithm for multilevel thresholding. *J. Inf. Sci. Eng.*, **17**, 713–727.
32. 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.
33. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B*, **57**, 289–300.
34. 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.
35. 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.

36. Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.
37. 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. USA*, **99**, 6567–6572.
38. Vapnik,V. (1979) *Estimation of Dependences Based on Empirical Data: Empirical Inference Science (Information Science and Statistics)*. Nauka, Moscow.
39. Cortes,C. and Vapnik,V. (1995) Support-vector networks. *Mach. Learn.*, **20**, 273–297.
40. Tibshirani,R.J., Hastie,T., Narasimhan,B. and Chu,G. (2011) Prediction Analysis of Microarrays for R.
41. Dimitriadou,E., Hornik,K., Leisch,F., Meyer,D. and Weingessel,A. (2005) *Misc Functions of the Department of Statistics (e1071)*. TU Wien.
42. Guyon,I., Weston,J., Barnhill,S. and Vapnik,V. (2002) Gene selection for cancer classification using support vector machines. *Mach. Learn.*, **46**, 389–422.
43. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
44. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
45. 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.
46. 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.
47. 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.
48. 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.
49. 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.
50. 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.
51. 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.
52. 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. *Nat. Biotechnol.*, **29**, 1132–1144.
53. Annunziata,C.M., Kleinberg,L., Davidson,B., Berner,A., Gius,D., Tchabo,N., Steinberg,S.M. and Kohn,E.C. (2007) BAG-4/SODD and associated antiapoptotic proteins are linked to aggressiveness of epithelial ovarian cancer. *Clin. Cancer Res.*, **13**, 6585–6592.
54. Torres,E.M., Williams,B.R. and Amon,A. (2008) Aneuploidy: cells losing their balance. *Genetics*, **179**, 737–746.
55. Tang,Y.-C., Williams,B.R., Siegel,J.J. and Amon,A. (2011) Identification of aneuploidy-selective antiproliferation compounds. *Cell*, **144**, 499–512.
56. Oliveros,J. VENNY. An interactive tool for comparing lists with Venn diagrams. *BioinfoGP, CNB-CSIC*.
57. Reich,M., Liefeld,T., Gould,J., Lerner,J., Tamayo,P. and Mesirov,J.P. (2006) GenePattern 2.0. *Nat. Genet.*, **38**, 500–501.