Systematic analysis of the determinants of gene expression noise in embryonic stem cells

Andre J. Faure[1, 2], Jörn M. Schmiedel[1,2], and Ben Lehner[1,2,3]

[1] EMBL-CRG Systems Biology Unit, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, 08003 Barcelona, Spain.
[2] Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain.
[3] Institució Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain
Correspondence should be addressed to B.L. (ben.lehner@crg.eu) [1]

[1] Lead Contact

SUMMARY

Isogenic cells in a common environment show substantial cell-to-cell variation in gene expression, often referred to as “expression noise”. Here we use multiple single cell RNA sequencing datasets to identify features associated with high or low expression noise in mouse embryonic stem cells. These include the core promoter architecture of a gene, with CpG island promoters and a TATA box associated with low and high noise, respectively. High noise is also associated with ‘conflicting’ chromatin states – the absence of transcription-associated histone modifications or the presence of repressive ones in active genes. Genes regulated by pluripotency factors through super-enhancers show high and correlated expression variability, consistent with fluctuations in the pluripotent state. Together, our results provide an integrated view of how core promoters, chromatin, regulation and pluripotency fluctuations contribute to the variability of gene expression across individual stem cells.

INTRODUCTION

Understanding the origins of phenotypic variation is a fundamental goal of biology (Lehner, 2013). Genetic and environmental differences are two important sources of variation, but even isogenic organisms and cells growing in a controlled environment can show substantial phenotypic variation, including in the expression of individual genes (Burga and Lehner, 2013; Raj and van Oudenaarden, 2008). Cell-to-cell or individual-to-individual variation in gene expression can have substantial phenotypic consequences, including determining the outcome of inherited detrimental mutations (Burga et al., 2011; Eldar et al., 2009; Raj et al., 2010) and generating phenotypic diversity as a "bet-hedging" strategy to facilitate adaptation in an unpredictable environment (Casanueva et al., 2012; Kussell and Leibler, 2005; Thattai and van Oudenaarden, 2004; Wolf et al., 2005). Expression variation can also be important during multi-cellular development, influencing the differentiation potential of stem cells, lineage decisions and receptor choices (Balázsi et al., 2011; Raj and van Oudenaarden, 2008).

Cell-to-cell heterogeneity in gene expression (often generically referred to as "noise") arises from the stochastic nature of the molecular processes involved in transcription and post-transcriptional regulation (Elowitz et al., 2002; Ozbudak et al., 2002; Sanchez and Golding, 2013; Schmiedel et al., 2015), from variation in the internal states of each cell such as the concentration or activity of key cellular components (Neves et al., 2010; Pedraza and van Oudenaarden, 2005; Stewart-Ornstein et al., 2012), and from differences in the external micro-environment (Battich et al., 2015; Snijder et al., 2009).

Transcription typically occurs in "bursts" motivating a multi-step model where transcription is restricted to an "active" state reachable from one or more "inactive" states involving chromatin remodelling and/or recruitment of co-factors(Coulon et al., 2013; Peccoud and Ycart, 1995; Zoller et al., 2015). The causes of cell-to-cell variation in expression have been more extensively investigated in microbes than in mammalian cells (Balázsi et al., 2011). In budding yeast, the core promoter is an important influence on the expression noise of each gene. Yeast genes can be classified into two classes by their promoter nucleosome organisation where high nucleosome occupancy close to the transcription start site (TSS) is associated with high noise, whereas a depleted proximal-nucleosome state is characterised by low transcriptional variability (Tirosh and Barkai, 2008). These two classes also distinguish genes with strikingly different patterns of expression responsiveness/plasticity across conditions and during evolution (Choi and Y.-J. Kim, 2009; Field et al., 2008; Lehner, 2008; Tirosh and Barkai, 2008). Underlying promoter sequence properties partially explain these differences (Sanchez and Golding, 2013) with the presence of a well-defined TATA box increasing expression noise and plasticity (Hornung et al., 2012; Landry et al., 2007; Lehner, 2010; Segal et al., 2006; Tirosh et al., 2006). In mammalian cells, the presence of a TATA box is also associated with increased noise, which has been attributed to a relatively reduced number of inferred inactive promoter cycles or rate-limiting steps (Zoller et al., 2015). However, mammalian cells have more diverse core promoter architectures (Lenhard et al., 2012) and the association of specific promoter features with noise has not been systematically investigated.

In multicellular organisms, different classes of genes can vary quite substantially in their chromatin architectures and how their chromatin relates to expression(Rach et al., 2011; Vavouri and Lehner, 2012). Some of these differences in chromatin have been related to patterns of gene expression across conditions and cell types. For example, in flies,

worms and mammals, genes with tissue-specific expression tend to have low levels of active chromatin modifications even when they are highly expressed (Pérez-Lluch et al., 2015), consistent with what was previously reported comparing between genes with CpG island (CGI) and non-CGI promoters (Vavouri and Lehner, 2012) and broad and narrow distributions of transcription start site usage (Rach et al., 2011). In mammalian cells, broad domains of H3K4me3 around key cell identity genes have also been associated with more stable expression between individual cultured cells and biological replicates of cell populations (Benayoun et al., 2014). Moreover, in *C. elegans*, high levels of the transcription-elongation associated histone modification H3K36me3 have been associated with increased expression stability during aging (Pu et al., 2015).

To better understand influences on expression variation in mammalian cells, we have focussed on a single cell type for which multiple single cell RNA-sequencing datasets as well as many additional functional genomic datasets are available – mouse embryonic stem cells. In an integrative analysis, we show that both the core promoter and the chromatin architecture of a gene is predictive of altered levels of cell-to-cell expression noise. In particular, active chromatin states in a gene body are associated with reduced noise and conflicting chromatin states – the absence of activation-associated histone modifications or the presence of repressive ones – are associated with increased noise. Moreover, genes with super-enhancers have unusually high cell-to-cell expression variation, and this variation is correlated between genes. Additional targets of pluripotency transcription factors have correlated expression, consistent with it reflecting fluctuations in pluripotency. This pluripotency fluctuation is associated with anti-correlated expression of differentiation genes and, in serum conditions, of expression from bivalent promoters in general. Our results provide an integrated view of the influences on the variability of expression across individual stem cells.

RESULTS

To systematically investigate the determinants of cell-to-cell expression variability in mouse embryonic stem cells (mESCs) we collated existing single cell RNA sequencing (scRNA-seq) datasets. We used three datasets from two studies employing unique molecular identifiers (UMI)-based methods: Islam *et al.* (Islam et al., 2014) (41 cells in serum medium), Grün *et al.* (Grün et al., 2014) (59 cells in serum, 75 cells in 2i medium), as well as three non-UMI based datasets from Kolodziejczyk *et al.* (Kolodziejczyk et al., 2015) (262 cells in serum, 300 cells in 2i and 147 cells in alternative 2i medium).

To compare expression noise between sets of genes with varying properties we processed the raw single cell transcript counts in three steps.

First, We quantified expression noise as the coefficient of variation, CV, which, consistent with previous analyses (Bar-Even et al., 2006; Fan et al., 2016; Grün et al., 2014; Islam et al., 2014; Newman et al., 2006) is strongly anti-correlated with the mean expression level of each gene in all of these datasets (Figure 1A, Figure S1A): lowly expressed genes tend to exhibit higher levels of noise than highly expressed genes and *vice-versa*.

Second, we accounted for this global dependency using the SCDE/PAGODA method that normalises each gene's observed expression variance to its genome-wide expected value (Fan et al., 2016; Kharchenko et al., 2014). PAGODA also controls for various sources of technical variation inherent in single-cell sequencing protocols including differences in total read count, gene length and experimental batch(Fan et al., 2016) (Figure 1B, Figure S1B). Consistent with previous results (Buettner et al., 2015; Klein et al., 2015; Kolodziejczyk et al., 2015; Macosko et al., 2015), we found that the cell cycle represents a significant source of gene expression heterogeneity in these datasets (Figure S2A-F). We therefore factored out this component of variation in order to minimise cell cycle phase-related biases from the results of our downstream noise analyses (Figure S2G-J).

Finally, we further normalised each gene's adjusted variance by calculating its adjusted variance rank in a sliding window of the 100 genes with most similar mean expression . This additional sliding window rank normalisation step removes the dependency of noise variance on mean expression magnitude (Figure 1C, Figure S1C) and represents a methodological extension to the existing PAGODA framework to accommodate comparisons of total gene set noise levels. The resulting *relative noise rank* measure, which is corrected for the confounding effect of transcript abundance, provides a robust way to compare relative noise levels on a unified scale within and between datasets and experiments (Figure 1D,E). Our analyses are based on comparisons of these relative noise rank estimates between sets of genes using the Mann–Whitney U Test, a non-parametric equivalent of the t-test. As a measure of effect size we use the area under the ROC curve (AUC) statistic (Figure 1F), which can be directly derived from the Mann-Whitney U statistic (Mason and Graham, 2002). In this context, the AUC is interpreted as the probability that a randomly chosen gene from a given gene set will have higher noise than a randomly chosen gene not in the gene set. Given the technical difficulty of obtaining reliable and comparable estimates of (absolute) transcript variability from scRNA-seq data, we focussed our analyses on datasets that made use of UMIs (Grün et al., 2014; Islam et al., 2014).

We tested the feasibility of using currently available scRNA-seq datasets, with their limited capture efficiencies, in our analysis. We used simulations to create *in silico* scRNA-seq datasets to test the influence of cell number and transcript capture efficiency on the accuracy of gene set noise estimates (Figure S3). As expected, low cell numbers and low capture efficiency diminish the effect sizes of noise estimates. We find, however, that even datasets with as low as 50 cells and 3% capture efficiency only reduce effect size estimates by less than two-fold (Figure S4), showing that our analysis approach is appropriate for the used scRNA-seq datasets.

Core promoter architectures associated with high or low expression noise

Both global analyses (Bar-Even et al., 2006; Newman et al., 2006) and systematic mutagenesis have identified multiple features of yeast promoters that modulated noise through effects on transcriptional kinetics (Bai et al., 2010; Blake et al., 2006; Carey et al., 2013; Dadiani et al., 2013; Hornung et al., 2012; Murphy et al., 2010; To and Maheshri, 2010). We compiled diverse experimental and sequence-based information on gene regulation in mESCs and then tested whether sets of genes with shared promoter features, chromatin states or chromatin domains have unusually high or low expression noise at the mRNA level, both individually and in multivariate models.

First, we tested ten mammalian core promoter features for noise biases: five sequence motifs (TATA box, Initiator motif, GC motif, CCAAT motif, CpG island) (Dreos et al., 2015), broad and sharp transcription initiation defined by Cap Analysis of Gene Expression (CAGE) (Lizio et al., 2015) and an alternative three-class definition (single initiation site, multiple initiation sites, broad initiation region) (Dreos et al., 2015).

The presence of a TATA box is the core promoter feature most associated with elevated expression noise in all the tested datasets (Figure 1D and Figure 2A), consistent with the well-established influence of the TATA box on expression noise in yeast (Hornung et al., 2012; Landry et al., 2007; Tirosh et al., 2006) and previous analyses in mammalian cells (Zoller et al., 2015). In contrast to the TATA box, the presence of a CpG island (CGI) is associated with comparatively low noise, that is, transcriptional consistency, particularly when the CGI extends into the gene body (Figure 2A,C). We consistently observe that promoters characterised by a single or limited number of transcription start sites (TSSs) are associated with higher levels of noise compared to promoters with broad initiation (Figure 1D, Figure 2A,C,D, Figure S5). Notably, these trends are absent when using “pool-and-split” control samples generated by pooling thousands of cells and then splitting their mRNAs into single-cell equivalents (Grün et al., 2014) (Figure 2B, Figure S6B,C, Figure S5B,D,F,H). Moreover, neither mean expression level, GC content nor gene length can account for these relationships (Figure S6A). This suggests that our noise metric reflects biological and not technical variation in measured transcript levels between individual cells. Accordingly, our observations are insensitive to technical choices made during analysis (Figure S7, S8, and S9).

We next combined these partially overlapping core promoter features in a sparse logistic regression model (Figure 2E). Overall these results indicate that CGI and TATA status are the most important and consistent core promoter features associated with transcriptional noise. However, TSS width (a quantitative description of Sharp/Broad TSS) also had consistently negative coefficients in the regression models, suggesting

that (TATA/CGI) sequence-defined promoter type is not sufficient to explain its influence on noise (Figure 2E).

Effect of chromatin on expression noise

We next tested for associations between chromatin and gene expression variability using matched ChIP-seq data (Yue et al., 2014) (Figure 3A, Figure S10). Four different gene sub-regions were interrogated: the transcription start site (TSS), core promoter (TSS-200bp, TSS+100bp), promoter (TSS-500bp, TSS+2000bp) and the whole gene body (TSS to transcription termination site, TTS).

Given the well-established role of CTCF in mediating three-dimensional chromatin interactions (Merkenschlager and Odom, 2013), which show substantially variability between individual cells (Nagano et al., 2013), we were surprised to find that its presence or absence near a gene had no detectable association with expression noise (Figure 3A). Likewise, despite the link between the accumulation of RNA polymerase II (RNAP2) near the promoter and transcriptional pausing (Lenhard et al., 2012), the presence of an RNAP2 peak is only weakly associated with decreased noise (Figure 3A). Moreover, neither the presence of the co-factor EP300 nor its active enhancer-associated histone modification H3K27ac (Creyghton et al., 2010) are strongly associated with noise.

However, we discovered three promoter-proximal histone modifications consistently associated with increased noise (H3K27me3, H3K4me1, H3K9me3) and three histone modifications associated with low noise when occurring within the gene body (H3K36me3, H3K4me3, H3K9ac, Figure 3A). Like TATA boxes and CGIs, these trends are not evident in results obtained using “pool-and-split” control datasets implying that any residual variability of a technical origin does not underlie these relationships (Figure S6B,C).

We observed that the generally repressive H3K27me3 mark is associated with elevated noise when present on expressed genes, particularly when targeted to the TSS (Figure 3A). The vast majority (91%) of promoters with H3K27me3 in mESCs also possess peaks for H3K4me3, a mark generally present at active promoters. The co-occurrence of these two opposing histone modifications, predominantly at CGI promoters, is termed bivalency and has been linked to transcriptional “priming” of genes that are rapidly either up- or down-regulated at subsequent developmental time-points (Azuara et al., 2006; Bernstein et al., 2006; Boyer et al., 2006; Lee et al., 2006). This class of bivalent genes show unusually high levels of cell-to-cell transcriptional variability (Figure 3B), while promoters possessing only H3K4me3 or H3K27me3 are less noisy (Figure 3C, Figure S11A,B,G,H). Using co-ChIP data obtained from sequential ChIP experiments with two different antibodies (Weiner et al., 2016), we confirm that it is the simultaneous presence of these opposing marks at the same promoters in single cells which is associated with increased transcriptional noise (Figure S11E-H) (Bernstein et al., 2006; Pan et al., 2007).

Of the other histone modifications characterized, H3K4me1, which is associated with active or poised enhancers (Creyghton et al., 2010), is associated with increased variability when present at the TSS. Likewise, when H3K9me3 is present, it is associated with elevated noise (Figure. 3A). The paucity of detected genes with H3K9me3 at their

promoters testifies to this mark's role in repression (median odds ratio of the association between H3K9me3 presence and gene detection=0.5, Fisher's Exact Test P=4e-5). Promoters marked by H3K4me3 show little bias in transcriptional noise, whereas when this modification occurs in the gene body it is consistently associated with low cell-to-cell variability. The same is true for two other modifications associated with transcription – H3K9ac and H3K36me3 – whose presence in the gene body, but not around the TSS, is associated with more stable expression across individual cells (Figure 3A).

Consequently, the lack of H3K36me3 in the body of genes is strongly associated with higher than expected expression noise across single cells (Figure 4A, Figure S11C). We confirmed that genes lacking H3K36me3 in ESCs also tend to have lower than expected levels of H3K36me3 when expressed in other cell types (Figure S12A-C), which also holds when correcting for expression level (Figure S12D-F). Reproducibly noisy genes (above average relative noise rank in all UMI-based datasets) that lack H3K36me3 peaks within their bodies in ESCs are also enriched for significant variation in average expression level (FDR=5%, fold change>4) across a panel of eight different mouse cell types (odds ratio=2.1, P=8e-3). Genes that become active with unusually low levels of H3K36me3 in their gene bodies therefore vary in expression at two scales – across individual cells and also across tissues.

To test whether higher order features of the genome are predictive of noise, we defined sets of genes according to their occurrence within four classes of topologically associating domains (TADs) (Dixon et al., 2012; Nora et al., 2012) and two classes of Lamina-associated domains (LADs) (Guelen et al., 2008). None of the TAD sub-classes – defined based on their average chromatin state – showed any significant noise association (Figure S10A-C). However, genes with detectable expression in single cells despite their occurrence within the generally repressive environment of constitutive LADs (cLADs) – i.e. regions proximal to the nuclear periphery in all assayed cell types – tend to have slightly elevated noise levels (Figure S13A,C,D). This suggests that the location of an active gene in a repressed neighbourhood may result in high cell-to-cell variability in expression.

Integrative analysis of chromatin and promoter features

To further examine the relationship between chromatin states, core promoter architectures and transcriptional noise, we used logistic regression to address which features best distinguish genes in the upper noise tercile from those in the lower tercile (Figure 4B).

Overall, the logistic regression suggests that chromatin features (the quantitative ChIP enrichments) are more important than promoter sequence features when predicting expression noise. TATA box and CpG island status do not contribute to models in two out of three cases (zero coefficients in Figure 4B), suggesting that their impact on noise can be accounted for by their influence on chromatin. Also, although H3K4me3 in the gene body is individually associated with decreased noise (also see Figure 3A), other features in the combined model can account for its contribution. On the other hand, several histone modifications including H3K27me3, H3K4me1, H3K9me3, H3K9ac and H3K36me3 are consistently retained (non-zero coefficients in all models). This suggests that their effects on noise are (i) at least partially independent of each other and (ii) cannot be simply explained by core promoter architecture features.

These results also confirm the primacy of gene body H3K36me3 and promoter H3K27me3 in predicting expression noise levels. However, other features consistently contribute to the predictions including gene body H3K9me3 and core promoter H3K9ac, which are both associated with lower noise levels in the integrated model. In other words, better estimates of transcriptional noise are obtained by combining multiple aspects of the chromatin state.

Taken together, the individual and joint analyses of chromatin modifications support a model in which genes with "conflicting" chromatin modifications are associated with high noise, i.e. where a "mismatch" or "conflict" between a gene's chromatin state and its transcriptional state is associated with increased expression variability. To further test this idea we classified genes by the number of conflicting chromatin modifications that they carry. Here we defined "mismatched" chromatin states as H3K27me3 at the promoter, H3K9me3 or H3K4me1 at the TSS, the absence of H3K9ac, H3K36me3 or H3K4me3 within the gene body, or the absence of H3K27ac at the promoter. These choices were motivated by these marks' observed associations with mean expression level (Figure S14E-G). Consistent with the "conflict" model, genes with a greater number of mismatched chromatin states indeed have higher noise (Figure 4C, Figure S11D). This is particularly true for genes with high mean expression levels (Figure 4C, Figure S11D), and is not accounted for by covariation with promoter types (Figure S15).

Lastly, although mean expression level is individually a poor predictor of noise (Figure S16A), it ranks highly in terms of its importance when combined with chromatin features in the integrated model (Figure S16B,C). In other words, when controlling for chromatin state, increased expression level is positively associated with higher levels of noise. A similar effect is seen for gene length whose contribution is greater in the integrated model (Figure S16).

Correlated fluctuations in pluripotency underlie the high noise of genes with super-enhancers

Super-enhancers (SEs) are large genomic regions defined by high levels of chromatin marks and transcriptional cofactors associated with active enhancers (H3K4me1, H3K27ac, MED1) that are densely occupied by master transcription factors (OCT4/POU5F1, SOX2, NANOG) regulating the expression of nearby cell identity genes (Parker et al., 2013; Whyte et al., 2013). We found that genes with super-enhancers have particularly high levels of noise (Figure 5A, Figure S13A). This is true for ESCs grown in serum where fluctuations in the pluripotent state are well-established (Chambers et al., 2007; Kalmar et al., 2009), but also for cells grown in 2i conditions (Figure S13A,B).

We tested whether the high noise genes with super-enhancers could be due to coherent fluctuations in their expression levels within individual cells. For this we used the PAGODA method to determine the levels of coordinated variability within sets of genes from single cell RNA-seq data (Fan et al., 2016). This revealed that super-enhancer target gene expression is coordinated genome-wide for cells in serum medium, but not in 2i conditions (Figure 5B,C, Figure S17A). In serum we see that the expression correlation of super-enhancer target genes is higher than the vast majority of other gene sets tested (P=0.08; P=0.04 for UMI-based dataset; see Figure S10 for the complete list

of gene sets). As a negative control we included super-enhancer targets from other cell types (myotube, macrophage, T-helper and pro-B cells) in our analysis; these sets covary to a much lower degree than genes with super-enhancers in ES cells (Figure 5B, Figure S17A, black dashed lines).

We hypothesised that coherent super-enhancer target gene expression variability is due to shared regulation by the master regulators OCT4, SOX2, and NANOG. These factors bind cooperatively to thousands of other genes in the ESC genome apart from SE targets (Whyte et al., 2013). We therefore asked whether additional genes that covary with super–enhancer target genes are also enriched for previously validated *in vivo* binding events for all three master regulators at their promoters as determined by ChIP-chip/-seq in ESCs (J. Kim et al., 2008; Whyte et al., 2013). This is indeed the case, with OCT4, SOX2, and NANOG targets strongly enriched amongst the genes most correlated with the expression of genes with SEs (Figure 5D, Figure S17B). Similarly, OCT4, SOX2, and NANOG targets are more strongly correlated with SE targets than expected by chance – a finding that is not replicated using SE targets from other cell types (Figure 5E, Figure S17C). However, "conventional" OCT4, SOX2, and NANOG targets as defined by ChIP-chip/-seq have lower levels of covariation with each other than do super-enhancer OCT4, SOX2, and NANOG targets (Figure 5B, Figure S17A). Thus, some but not all OCT4, SOX2, and NANOG targets detectably co-vary in expression across single cells with the presence of a super-enhancer associated with stronger covariation with the pluripotency network fluctuations.

We investigated whether the correlated fluctuations in genes with super-enhancers and other OCT4, SOX2, and NANOG targets are associated with fluctuations in the expression of differentiation genes. The expression of genes annotated to Gene Ontology (GO) biological process terms including the words "differentiation" or "development" indeed tend to be anti-correlated with super-enhancer targets (Figure 5F). This result is consistent across multiple datasets and independent of the growth medium – it is observed for cells in both serum and 2i (Figure S14A). Thus, the coordinated fluctuations in super-enhancer targets across single cells are associated with a fluctuation in the pluripotent state.

Finally, in serum medium, we also detected an anti-correlation between the expression of super-enhancer targets and the expression of all genes with bivalent promoters (Figure 5H,I, Figure S18A,B, median AUC=0.42, P=3e-4; compare to Figure S14B-D). This might be expected given the established link between this particular chromatin state and genes with differentiation-type functions. On exit from pluripotency polycomb target genes are generally up-regulated (Lee et al., 2006). However, this trend persists when (i) ignoring genes with "differentiation" or "development" GO annotations and (ii) excluding OCT4, SOX2, and NANOG targets (Figure S18C,D). Interestingly, there is no consistent association between super-enhancer target correlation for gene sets defined based on promoter type (Figure 5G). For example, although TATA box genes have unusually high noise levels, their fluctuations in single cells are not correlated with those of super-enhancer targets. In summary, these results indicate that a bivalent chromatin state is associated with high noise in embryonic stem cells and that, additionally in serum conditions, fluctuations in the expression of bivalent genes tend to be aligned with fluctuations in pluripotency, i.e. anti-correlated with fluctuations in super-enhancer targets.

High mRNA stability masks the increased transcriptional noise of expression from a single X chromosome

The independent stochastic production of transcripts from multiple loci is expected to reduce transcriptional noise – even when controlling for mean levels (Cook et al., 1998). In male ESCs, genes on the X chromosome are expressed from a single locus in contrast to predominantly biallelic expression of autosomal genes. However, in all of the single cell RNA-seq datasets, X-linked genes have similar noise levels to genes expressed from the autosomes present in two copies in these cells (Figure S13A,E, Figure S10). These results are consistent with those obtained using replicate populations of bulk measurements from populations of cells showing unexpectedly low variation of genes on the X chromosome in contrast to other monoallelically expressed genes (Yin et al., 2009).

The time-scales of transcription are known to influence expression variability (Paulsson, 2004; Pedraza and Paulsson, 2008). Specificially, longer lived mRNAs are expected to exhibit lower expression variability, due to improved averaging over transcriptional bursts (Zoller et al., 2015). It was recently shown that genes expressed from the X chromosome have unusually high mRNA stability (Faucillion and Larsson, 2015). We therefore hypothesized that the increased mRNA stability of X-linked genes might be masking the increased noise resulting from mono-allelic expression. Combining data from four independent studies(Friedel et al., 2009; Schwanhäusser et al., 2011; Sharova et al., 2009; Tippmann et al., 2012) we could validate that X-chromosome transcripts have longer half-lives than those expressed from other chromosomes (Figure 6A, Figure S19A-D). The extreme stability of mRNAs transcribed from the X chromosome can be clearly be appreciated when comparing the average half-life of these transcripts to those of random sets of autosomal genes ($P<1e-4$, Figure 6B, Figure S19E-H).

We find that, in mESCs, mRNA half-life is indeed negatively associated with mRNA expression noise with an effect size similar to the strongest chromatin modifications (Figure S20A). We next tested whether, given their unusually high mRNA stability, X-linked genes do – or do not – have high noise. We constructed random gene sets of varying size and average mRNA half-life using a weighted sampling strategy. We then tested each of these random gene sets for biased expression noise. The strength of a gene set's mRNA stability bias is strongly anti-correlated with its average level of expression noise (median Spearman's rho=-0.62, Figure S19I-K). This analysis reveals that genes on the X chromosome do indeed have increased noise (median residual AUC=0.03, empirical P=0.03) compared to other gene sets with matched mRNA stabilities (Figure 6C, Figure S19I-K). Thus, the unusually stable transcripts of X-linked genes contribute to reducing their mRNA noise levels to match those of genes expressed from the autosomes. However, given the differential effects of mRNA stability on mRNA and protein noise (Ozbudak et al., 2002), it remains to be tested whether this noise buffering effect in X-linked genes also extends to the protein level.

By contrast, we observe that super-enhancer targets have noise levels that are much higher than would be expected given their average mRNA half-life, indicating that although they produce short-lived transcripts this alone is insufficient to explain their unusually high noise levels. Likewise, biases in mRNA stability are insufficient to explain the contribution of other promoter and chromatin features associated with high or low noise (Figure 6C, Figure S19I-K, for further analysis see Figure S20.).

DISCUSSION

We have conducted a genome-scale survey of genomic features associated with gene expression variation in single mouse embryonic stem cells. Our results provide an integrated view of the regulatory properties associated with stable or variable gene expression across individual ES cells (Figure 6D).

This composite portrait promotes the formulation of specific hypotheses about how gene regulation is coordinated. For example, we observe that chromatin is an important determinant of transcriptional stability across individual cells and that "conflicting" chromatin states in active genes are associated with more variable expression. When combined with analyses of the variability of expression across cell types (Benayoun et al., 2014; Pérez-Lluch et al., 2015) and during aging (Pu et al., 2015; Sen et al., 2015), this suggests that chromatin may be more important for modulating the robustness vs. responsiveness of gene expression in time and across conditions, rather than modulating expression levels per se.

We also found that genes with super-enhancers have unusually high noise in mESCs, and this noise is correlated across genes with super-enhancers and, to a lesser extent, also with other genes that are targets of OCT4, SOX2, and NANOG. This suggests fluctuations in the pluripotent state as an important upstream driving factor in the cell-to-cell variation in the expression of these genes. Accordingly, many genes with functions in differentiation and development had expression that detectably anti-correlates with the expression of super enhancer targets, and we detected an anti-correlation between pluripotency-driven fluctuations and the expression from bivalent promoters in general. This is consistent with results showing that polycomb target genes are de-repressed upon exit from pluripotency (Lee et al., 2006), and we suggest that bivalent genes both have high noise in general and fluctuations that are anti-correlated with the pluripotent state when grown in serum.

In total, this work demonstrates that core promoter architecture, chromatin state, the presence of super-enhancers, and mRNA half-life–all associated in definable ways to a gene's mRNA expression variability.

AUTHOR CONTRIBUTIONS

A.J.F. performed all analyses. J.M.S. performed simulations. A.J.F., J.M.S. and B.L. designed the study, interpreted the results, and wrote the manuscript.

ACKNOWLEDGEMENTS

This work was supported by a European Research Council (ERC) Consolidator grant (616434), the Spanish Ministry of Economy and Competitiveness (BFU2011-26206 and "Centro de Excelencia Severo Ochoa 2013-2017" SEV-2012-0208), the AXA Research Fund, the Bettencourt Schueller Foundation, Agencia de Gestio d'Ajuts Universitaris i de Recerca (AGAUR), FP7 project 4DCellFate (277899), and the EMBL-CRG Systems Biology Program. AF was supported by a MINECO post-doctoral grant (FPDI-2013-17783). We thank Mirko Francesconi and Fran Supek for their helpful suggestions.

REFERENCES

Azuara, V., Perry, P., Sauer, S., Spivakov, M., Jørgensen, H.F., John, R.M., Gouti, M., Casanova, M., Warnes, G., Merkenschlager, M., et al., (2006). Chromatin signatures of pluripotent cell lines. Nat Cell Biol 8, 532–538. doi:10.1038/ncb1403

Bai, L., Charvin, G., Siggia, E.D., Cross, F.R., (2010). Nucleosome-depleted regions in cell-cycle-regulated promoters ensure reliable gene expression in every cell cycle. Developmental Cell 18, 544–555. doi:10.1016/j.devcel.2010.02.007

Baker, S.C., Bauer, S.R., Beyer, R.P., Brenton, J.D., Bromley, B., Burrill, J., Causton, H., Conley, M.P., Elespuru, R., Fero, M., et al., (2005). The External RNA Controls Consortium: a progress report. Nat Meth 2, 731–734. doi:10.1038/nmeth1005-731

Balázsi, G., van Oudenaarden, A., Collins, J.J., (2011). Cellular decision making and biological noise: from microbes to mammals. Cell 144, 910–925. doi:10.1016/j.cell.2011.01.030

Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M., O'Shea, E., Pilpel, Y., Barkai, N., (2006). Noise in protein expression scales with natural protein abundance. Nat Genet 38, 636–643. doi:10.1038/ng1807

Battich, N., Stoeger, T., Pelkmans, L., (2015). Control of Transcript Variability in Single Mammalian Cells. Cell 163, 1596–1610. doi:10.1016/j.cell.2015.11.018

Benayoun, B.A., Pollina, E.A., Ucar, D., Mahmoudi, S., Karra, K., Wong, E.D., Devarajan, K., Daugherty, A.C., Kundaje, A.B., Mancini, E., et al., (2014). H3K4me3 Breadth Is Linked to Cell Identity and Transcriptional Consistency. Cell 158, 673–688. doi:10.1016/j.cell.2014.06.027

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al., (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125, 315–326. doi:10.1016/j.cell.2006.02.041

Blake, W.J., Balázsi, G., Kohanski, M.A., Isaacs, F.J., Murphy, K.F., Kuang, Y., Cantor, C.R., Walt, D.R., Collins, J.J., (2006). Phenotypic consequences of promoter-mediated transcriptional noise. Molecular Cell 24, 853–865. doi:10.1016/j.molcel.2006.11.003

Boyer, L.A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L.A., Lee, T.I., Levine, S.S., Wernig, M., Tajonar, A., Ray, M.K., et al., (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature 441, 349–353. doi:10.1038/nature04733

Bray, N.L., Pimentel, H., Melsted, P., Pachter, L., (2016). Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34, 525–527. doi:10.1038/nbt.3519

Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann, S.A., Marioni, J.C., Stegle, O., (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Biotechnol. doi:10.1038/nbt.3102

Burga, A., Casanueva, M.O., Lehner, B., (2011). Predicting mutation outcome from early stochastic variation in genetic interaction partners. Nature 480, 250–253. doi:10.1038/nature10665

Burga, A., Lehner, B., (2013). Predicting phenotypic variation from genotypes, phenotypes and a combination of the two. Current Opinion in Biotechnology 24, 803–809. doi:10.1016/j.copbio.2013.03.004

Carey, L.B., van Dijk, D., Sloot, P.M.A., Kaandorp, J.A., Segal, E., (2013). Promoter sequence determines the relationship between expression level and noise. Plos Biol 11, e1001528. doi:10.1371/journal.pbio.1001528

Casanueva, M.O., Burga, A., Lehner, B., (2012). Fitness trade-offs and environmentally

induced mutation buffering in isogenic C. elegans. Science 335, 82–85. doi:10.1126/science.1213491

Chambers, I., Silva, J., Colby, D., Nichols, J., Nijmeijer, B., Robertson, M., Vrana, J., Jones, K., Grotewold, L., Smith, A., (2007). Nanog safeguards pluripotency and mediates germline development. Nature 450, 1230–1234. doi:10.1038/nature06403

Choi, J.K., Kim, Y.-J., (2009). Intrinsic variability of gene expression encoded in nucleosome positioning sequences. Nat Genet 41, 498–503. doi:10.1038/ng.319

Cook, D.L., Gerber, A.N., Tapscott, S.J., (1998). Modeling stochastic gene expression: implications for haploinsufficiency. Proc. Natl. Acad. Sci. U.S.A. 95, 15641–15646.

Coulon, A., Chow, C.C., Singer, R.H., Larson, D.R., (2013). Eukaryotic transcriptional dynamics: from single molecules to cell populations. Nat Rev Genet 14, 572–584. doi:10.1038/nrg3484

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., et al., (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the national Academy of Sciences 107, 21931–21936. doi:10.1073/pnas.1016071107

Dadiani, M., van Dijk, D., Segal, B., Field, Y., Ben-Artzi, G., Raveh-Sadka, T., Levo, M., Kaplow, I., Weinberger, A., Segal, E., (2013). Two DNA-encoded strategies for increasing expression with opposing effects on promoter dynamics and transcriptional noise. Genome Research 23, 966–976. doi:10.1101/gr.149096.112

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., Ren, B., (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376–380. doi:10.1038/nature11082

Dreos, R., Ambrosini, G., Groux, R., Cavin Périer, R., Bucher, P., (2016). The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. Nucleic acids research. doi:10.1093/nar/gkw1069

Dreos, R., Ambrosini, G., Périer, R.C., Bucher, P., (2015). The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools. Nucleic acids research 43, D92–6. doi:10.1093/nar/gku1111

Eldar, A., Chary, V.K., Xenopoulos, P., Fontes, M.E., Losón, O.C., Dworkin, J., Piggot, P.J., Elowitz, M.B., (2009). Partial penetrance facilitates developmental evolution in bacteria. Nature 460, 510–514. doi:10.1038/nature08150

Elowitz, M.B., Levine, A.J., Siggia, E.D., Swain, P.S., (2002). Stochastic gene expression in a single cell. Science 297, 1183–1186. doi:10.1126/science.1070919

Ernst, J., Kellis, M., (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol 28, 817–825. doi:10.1038/nbt.1662

Fan, J., Salathia, N., Liu, R., Kaeser, G.E., Yung, Y.C., Herman, J.L., Kaper, F., Fan, J.-B., Zhang, K., Chun, J., et al., (2016). Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nat Meth. doi:10.1038/nmeth.3734

Faucillion, M.L., Larsson, J., (2015). Increased Expression of X-Linked Genes in Mammals Is Associated with a Higher Stability of Transcripts and an Increased Ribosome Density. Genome Biol Evol 7, 1039–1052. doi:10.1093/gbe/evv054

Field, Y., Kaplan, N., Fondufe-Mittendorf, Y., Moore, I.K., Sharon, E., Lubling, Y., Widom, J., Segal, E., (2008). Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. PLoS Comput Biol 4, e1000216. doi:10.1371/journal.pcbi.1000216

Friedel, C.C., Dölken, L., Ruzsics, Z., Koszinowski, U.H., Zimmer, R., (2009). Conserved principles of mammalian transcriptional regulation revealed by RNA half-life. Nucleic

acids research 37, e115. doi:10.1093/nar/gkp542
Grün, D., Kester, L., van Oudenaarden, A., (2014). Validation of noise models for single-cell transcriptomics. Nat Meth 11, 637–640. doi:10.1038/nmeth.2930
Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W., et al., (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature 453, 948–951. doi:10.1038/nature06947
Haberle, V., Forrest, A.R.R., Hayashizaki, Y., Carninci, P., Lenhard, B., (2015). CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. Nucleic acids research 43, e51. doi:10.1093/nar/gkv054
Hiratani, I., Ryba, T., Itoh, M., Yokochi, T., Schwaiger, M., Chang, C.-W., Lyou, Y., Townes, T.M., Schübeler, D., Gilbert, D.M., (2008). Global reorganization of replication domains during embryonic stem cell differentiation. Plos Biol 6, e245. doi:10.1371/journal.pbio.0060245
Hornung, G., Bar-Ziv, R., Rosin, D., Tokuriki, N., Tawfik, D.S., Oren, M., Barkai, N., (2012). Noise-mean relationship in mutated promoters. Genome Research 22, 2409–2417. doi:10.1101/gr.139378.112
Hurt, J.A., Robertson, A.D., Burge, C.B., (2013). Global analyses of UPF1 binding and function reveal expanded scope of nonsense-mediated mRNA decay. Genome Research 23, 1636–1650. doi:10.1101/gr.157354.113
Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., Linnarsson, S., (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Meth 11, 163–166. doi:10.1038/nmeth.2772
Kalmar, T., Lim, C., Hayward, P., Muñoz-Descalzo, S., Nichols, J., Garcia-Ojalvo, J., Arias, A.M., (2009). Regulated Fluctuations in Nanog Expression Mediate Cell Fate Decisions in Embryonic Stem Cells. Plos Biol 7, e1000149. doi:10.1371/journal.pbio.1000149
Kharchenko, P.V., Silberstein, L., Scadden, D.T., (2014). Bayesian approach to single-cell differential expression analysis. Nat Meth 11, 740–742. doi:10.1038/nmeth.2967
Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S.L., (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14, R36. doi:10.1186/gb-2013-14-4-r36
Kim, J., Chu, J., Shen, X., Wang, J., Orkin, S.H., (2008). An Extended Transcriptional Network for Pluripotency of Embryonic Stem Cells. Cell 132, 1049–1061. doi:10.1016/j.cell.2008.02.039
Kim, J.K., Kolodziejczyk, A.A., Illicic, T., Teichmann, S.A., Marioni, J.C., (2015). Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. Nat Commun 6, 8687–. doi:10.1038/ncomms9687
Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., Kirschner, M.W., (2015). Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. Cell 161, 1187–1201. doi:10.1016/j.cell.2015.04.044
Kolodziejczyk, A.A., Kim, J.K., Tsang, J.C.H., Ilicic, T., Henriksson, J., Natarajan, K.N., Tuck, A.C., Gao, X., Bühler, M., Liu, P., Marioni, J.C., et al., (2015). Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. Cell Stem Cell 17, 471–485. doi:10.1016/j.stem.2015.09.011
Kussell, E., Leibler, S., (2005). Phenotypic diversity, population growth, and information in fluctuating environments. Science 309, 2075–2078. doi:10.1126/science.1114383
Landry, C.R., Lemos, B., Rifkin, S.A., Dickinson, W.J., Hartl, D.L., (2007). Genetic properties influencing the evolvability of gene expression. Science 317, 118–121.

doi:10.1126/science.1140247
Langmead, B., Salzberg, S.L., (2012). Fast gapped-read alignment with Bowtie 2. Nat Meth 9, 357–359. doi:10.1038/nmeth.1923
Lee, T.I., Jenner, R.G., Boyer, L.A., Guenther, M.G., Levine, S.S., (2006). Control of Developmental Regulators by Polycomb in Human Embryonic Stem Cells. Cell.
Lehner, B., (2013). Genotype to phenotype: lessons from model organisms for human genetics. Nat Rev Genet 14, 168–178. doi:10.1038/nrg3404
Lehner, B., (2010). Conflict between noise and plasticity in yeast. PLoS Genet 6, e1001185. doi:10.1371/journal.pgen.1001185
Lehner, B., (2008). Selection to minimise noise in living systems and its implications for the evolution of gene expression. Molecular Systems Biology 4, 170. doi:10.1038/msb.2008.11
Lenhard, B., Sandelin, A., Carninci, P., (2012). Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat Rev Genet 13, 233–245. doi:10.1038/nrg3163
Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., Sahin, S., Abugessaisa, I., Fukuda, S., Hori, F., Ishikawa-Kato, S., et al., (2015). Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol 16, 22. doi:10.1186/s13059-014-0560-6
Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al., (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161, 1202–1214. doi:10.1016/j.cell.2015.05.002
Merkenschlager, M., Odom, D.T., (2013). CTCF and Cohesin: Linking Gene Regulatory Elements with Their Targets. Cell 152, 1285–1297. doi:10.1016/j.cell.2013.02.029
Murphy, K.F., Adams, R.M., Wang, X., Balázsi, G., Collins, J.J., (2010). Tuning and controlling gene expression noise in synthetic gene networks. Nucleic acids research 38, 2712–2726. doi:10.1093/nar/gkq091
Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., Fraser, P., (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature. doi:10.1038/nature12593
Neves, das, R.P., Jones, N.S., Andreu, L., Gupta, R., Enver, T., Iborra, F.J., (2010). Connecting variability in global transcription rate to mitochondrial variability. Plos Biol 8, e1000560. doi:10.1371/journal.pbio.1000560
Newman, J.R.S., Ghaemmaghami, S., Ihmels, J., Breslow, D.K., Noble, M., DeRisi, J.L., Weissman, J.S., (2006). Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature 441, 840–846. doi:10.1038/nature04785
Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., et al., (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature 485, 381–385. doi:10.1038/nature11049
Ozbudak, E.M., Thattai, M., Kurtser, I., Grossman, A.D., van Oudenaarden, A., (2002). Regulation of noise in the expression of a single gene. Nat Genet 31, 69–73. doi:10.1038/ng869
Pan, G., Tian, S., Nie, J., Yang, C., Ruotti, V., Wei, H., Jonsdottir, G.A., Stewart, R., Thomson, J.A., (2007). Whole-Genome Analysis of Histone H3 Lysine 4 and Lysine 27 Methylation in Human Embryonic Stem Cells. Cell Stem Cell 1, 299–312. doi:10.1016/j.stem.2007.08.003
Parker, S.C.J., Stitzel, M.L., Taylor, D.L., Orozco, J.M., Erdos, M.R., Akiyama, J.A., van Bueren, K.L., Chines, P.S., Narisu, N., Black, B.L., et al., (2013). Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk

variants. Proceedings of the national Academy of Sciences.
Paulsson, J., (2004). Summing up the noise in gene networks. Nature 427, 415–418. doi:10.1038/nature02257
Peccoud, J., Ycart, B., (1995). Markovian Modeling of Gene-Product Synthesis. Theoretical population biology.
Pedraza, J.M., Paulsson, J., (2008). Effects of molecular memory and bursting on fluctuations in gene expression. Science 319, 339–343. doi:10.1126/science.1144331
Pedraza, J.M., van Oudenaarden, A., (2005). Noise Propagation in Gene Networks. Science 307, 1965–1969. doi:10.1126/science.1109090
Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S.W.M., Solovei, I., Brugman, W., Gräf, S., Flicek, P., Kerkhoven, R.M., van Lohuizen, M., et al., (2010). Molecular Maps of the Reorganization of Genome-Nuclear Lamina Interactions during Differentiation. Molecular Cell 38, 603–613. doi:10.1016/j.molcel.2010.03.016
Pérez-Lluch, S., Blanco, E., Tilgner, H., Curado, J., Ruiz-Romero, M., Corominas, M., Guigó, R., (2015). Absence of canonical marks of active chromatin in developmentally regulated genes. Nat Genet 47, 1158–1167. doi:10.1038/ng.3381
Pu, M., Ni, Z., Wang, M., Wang, X., Wood, J.G., Helfand, S.L., Yu, H., Lee, S.S., (2015). Trimethylation of Lys36 on H3 restricts gene expression change during aging and impacts life span. Genes & development 29, 718–731. doi:10.1101/gad.254144.114
Rach, E.A., Winter, D.R., Benjamin, A.M., Corcoran, D.L., Ni, T., Zhu, J., Ohler, U., (2011). Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. PLoS Genet 7, e1001274. doi:10.1371/journal.pgen.1001274
Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y., Tyagi, S., (2006). Stochastic mRNA Synthesis in Mammalian Cells. Plos Biol 4, e309. doi:10.1371/journal.pbio.0040309
Raj, A., Rifkin, S.A., Andersen, E., Oudenaarden, A.V., (2010). Variability in gene expression underlies incomplete penetrance. Nature 463, 913. doi:doi:10.1038/nature08781
Raj, A., van Oudenaarden, A., (2008). Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. Cell 135, 216–226. doi:10.1016/j.cell.2008.09.050
Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al., (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell. doi:10.1016/j.cell.2014.11.021
Sanchez, A., Golding, I., (2013). Genetic determinants and cellular constraints in noisy gene expression. Science 342, 1188–1193. doi:10.1126/science.1242975
Schmiedel, J.M., Klemm, S.L., Zheng, Y., Sahay, A., Bluthgen, N., Marks, D.S., van Oudenaarden, A., (2015). MicroRNA control of protein expression noise. Science 348, 128–132. doi:10.1126/science.aaa1738
Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., Selbach, M., (2011). Global quantification of mammalian gene expression control. Nature 473, 337–342. doi:10.1038/nature10098
Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I.K., Wang, J.-P.Z., Widom, J., (2006). A genomic code for nucleosome positioning. Nature 442, 772–778. doi:10.1038/nature04979
Sen, P., Dang, W., Donahue, G., Dai, J., Dorsey, J., Cao, X., Liu, W., Cao, K., Perry, R., Lee, J.Y., et al., (2015). H3K36 methylation promotes longevity by enhancing transcriptional fidelity. Genes & development 29, 1362–1376. doi:10.1101/gad.263707.115

Serra, F., Baù, D., Filion, G., Marti-Renom, M.A., (2016). Structural features of the fly chromatin colors revealed by automatic three-dimensional modeling., bioRxiv. doi:10.1101/036764

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., Cavalli, G., (2012). Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. Cell 148, 458–472. doi:10.1016/j.cell.2012.01.010

Sharova, L.V., Sharov, A.A., Nedorezov, T., Piao, Y., Shaik, N., Ko, M.S.H., (2009). Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells. DNA Res. 16, 45–58. doi:10.1093/dnares/dsn030

Smit, A.F.A., Hubley, R., Green, P., (1996). RepeatMasker Open-3.0 http://www.repeatmasker.org.

Snijder, B., Sacher, R., Rämö, P., Damm, E.-M., Liberali, P., Pelkmans, L., (2009). Population context determines cell-to-cell variability in endocytosis and virus infection. Nature 461, 520–523. doi:10.1038/nature08282

Soneson, C., Love, M.I., Robinson, M.D., (2016). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research 4. doi:10.12688/f1000research.7563.1

Stewart-Ornstein, J., Weissman, J.S., El-Samad, H., (2012). Cellular Noise Regulons Underlie Fluctuations in Saccharomyces cerevisiae. Molecular Cell 45, 483–493. doi:10.1016/j.molcel.2011.11.035

Thattai, M., van Oudenaarden, A., (2004). Stochastic gene expression in fluctuating environments. Genetics 167, 523–530.

Tippmann, S.C., Ivanek, R., Gaidatzis, D., Schöler, A., Hoerner, L., van Nimwegen, E., Stadler, P.F., Stadler, M.B., Schübeler, D., (2012). Chromatin measurements reveal contributions of synthesis and decay to steady-state mRNA levels. Molecular Systems Biology 8, 593. doi:10.1038/msb.2012.23

Tirosh, I., Barkai, N., (2008). Two strategies for gene regulation by promoter nucleosomes. Genome Research 18, 1084–1091. doi:10.1101/gr.076059.108

Tirosh, I., Weinberger, A., Carmi, M., Barkai, N., (2006). A genetic signature of interspecies variations in gene expression. Nat Genet 38, 830–834. doi:10.1038/ng1819

To, T.L., Maheshri, N., (2010). Noise Can Induce Bimodality in Positive Transcriptional Feedback Loops Without Bistability. Science 327, 1142–1145. doi:10.1126/science.1178962

Vavouri, T., Lehner, B., (2012). Human genes with CpG island promoters have a distinct transcription-associated chromatin organization. Genome Biol 13, R110. doi:10.1186/gb-2012-13-11-r110

Weddington, N., Stuy, A., Hiratani, I., Ryba, T., Yokochi, T., Gilbert, D.M., (2008). ReplicationDomain: a visualization tool and comparative database for genome-wide replication timing data. BMC Bioinformatics 9, 530. doi:10.1186/1471-2105-9-530

Weiner, A., Lara-Astiaso, D., Krupalnik, V., Gafni, O., David, E., Winter, D.R., Hanna, J.H., Amit, I., (2016). Co-ChIP enables genome-wide mapping of histone mark co-occurrence at single-molecule resolution. Nat Biotechnol 34, 953–961. doi:10.1038/nbt.3652

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., Young, R.A., (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 153, 307–319. doi:10.1016/j.cell.2013.03.035

Wolf, D.M., Vazirani, V.V., Arkin, A.P., (2005). Diversity in times of adversity:

probabilistic strategies in microbial survival games. J. Theor. Biol. 234, 227–253. doi:10.1016/j.jtbi.2004.11.020

Yin, S., Wang, P., Deng, W., Zheng, H., Hu, L., Hurst, L.D., Kong, X., (2009). Dosage compensation on the active X chromosome minimizes transcriptional noise of X-linked genes in mammals. Genome Biol 10, R74. doi:10.1186/gb-2009-10-7-r74

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B.D., et al., (2014). A comparative encyclopedia of DNA elements in the mouse genome. Nature 515, 355–364. doi:10.1038/nature13992

Zoller, B., Nicolas, D., Molina, N., Naef, F., (2015). Structure of silent transcription intervals and noise characteristics of mammalian genes. Molecular Systems Biology 11, 823.

FIGURES

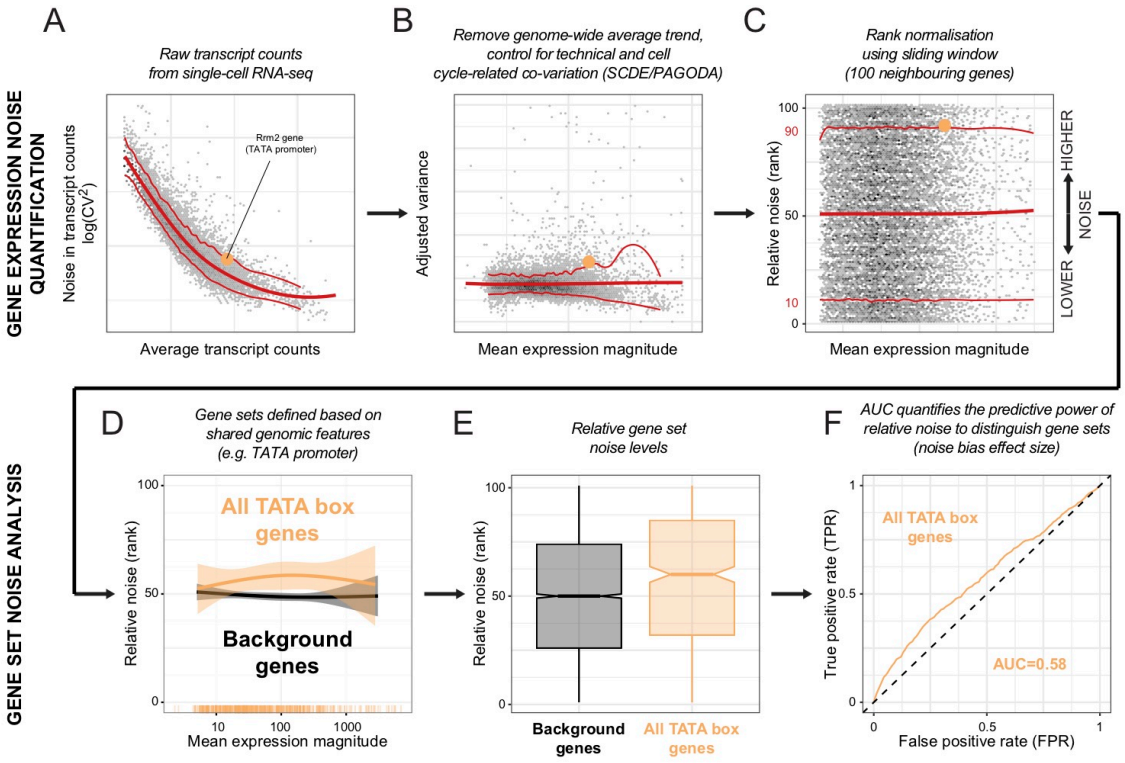

Figure 1. Gene expression noise quantification and gene set analysis procedure. A. Noise in transcript counts versus average transcript counts for all detected genes as determined using raw count data from a single-cell RNA-seq experiment: Grün *et al.* (serum)(Grün et al., 2014). **B.** We used SCDE/PAGODA to account for the global dependency of gene expression noise on the mean expression level, as well as to control for technical and cell cycle-related co-variation. **C.** An additional rank normalisation step removes the effect of mean expression on noise variance. A local smooth (*loess*) using all the data (and corresponding running upper and lower deciles) is shown in red in each panel. The Rrm2 gene, which possesses a TATA box at its promoter, is highlighted in orange in each panel. **D.** We defined sets of genes based on various shared genomic features, chromatin states and annotations. Indicated is the result of a Binomial smooth of relative noise (rank) versus mean expression magnitude shown separately for detected genes with a well-defined TATA box (orange) and the set of remaining background genes (black; detected genes for which TATA status was available). Shaded regions indicate 95% confidence intervals. Tick marks on the x-axis depict the distribution of individual gene mean expression magnitudes. See Figure S5C for similar results based on noise estimates from all three UMI-based datasets. Also see Figure S21A,B for comparisons to results where expression noise is represented by adjusted variance values (i.e. without rank normalisation). To quantify the relative noise bias of genes belonging to a given gene set with respect to background genes (**E**), we use the area under the ROC curve (AUC) statistic, which can be directly derived from the Mann-Whitney U statistic i.e. a non-parametric equivalent of the t-statistic (**F**).

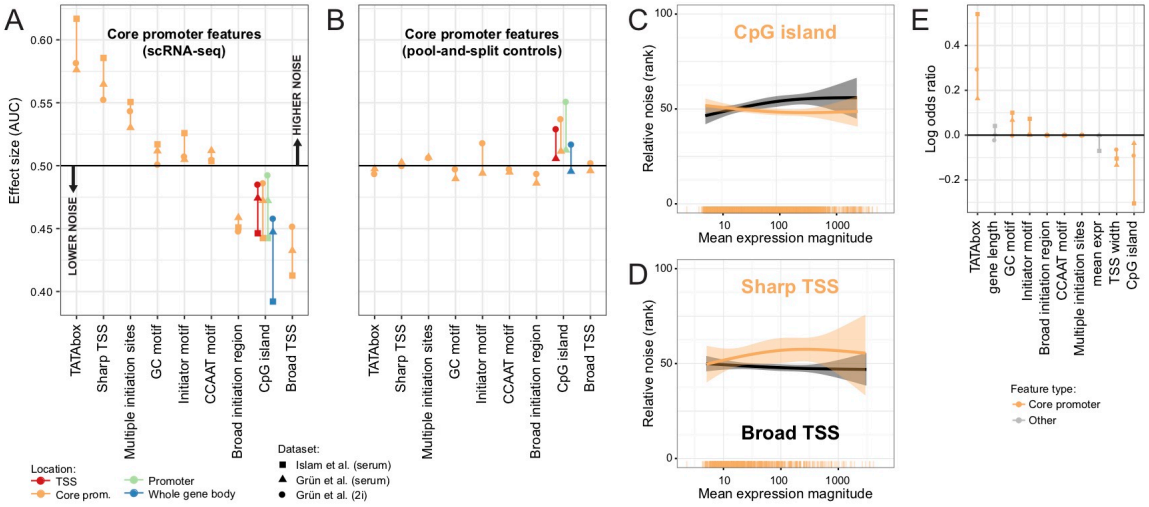

Figure 2. Gene set noise analysis reveals associations between core promoter architectures and cell-to-cell gene expression variability. A. Noise biases of gene sets constructed based on a selection of core promoter architecture features. Effect sizes are indicated as AUC, where a value of 0.5 corresponds to a random sample of genes. Each data point is shown in biological triplicate (three UMI-based single-cell RNA-seq datasets) and colours indicate the location of the corresponding feature (see legend). **B.** Results of an identical analysis to that shown in **A,** but instead using "pool-and-split" control datasets from Grün *et al.* where biological variation was eliminated by pooling thousands of cells and then splitting their mRNAs into single-cell equivalents. **C.** Binomial smooth results of relative noise (rank) versus mean expression magnitude shown separately for detected genes with a CpG island (CGI, orange) and the set of remaining background genes (black; detected genes for which CGI status was available). Shaded regions indicate 95% confidence intervals. Tick marks on the x-axis depict the distribution of individual gene mean expression magnitudes. **D.** A similar noise-mean curve to that shown in **C** constructed using genes with Sharp or Broad TSSs. Panels C-D show results obtained using noise estimates from Grün et al. (serum)(Grün et al., 2014). See Figure S5C,E,G for similar results based on noise estimates from all three UMI-based datasets. Also see Figure S21C-H for comparisons to results where expression noise is represented by adjusted variance values (i.e. without rank normalisation). **E.** Feature coefficients in a penalised logistic regression model (Lasso) distinguishing genes with noise levels in the upper tercile from those in the lower tercile (genes with intermediate noise levels in the mid tercile were excluded). The regularisation parameter *lambda* was chosen to yield the most regularized model such that deviance from 10-fold cross-validation was within one standard error of the minimum. Apart from previously considered (binary) core promoter sequence features, we also included TSS width (a quantitative description of Sharp/Broad TSS), mean expression level and gene length as continuous features in the model. See Figure S22 for comparisons to feature coefficients corresponding to single (univariate) and multi-feature logistic regression models (i.e. conventional as opposed to penalised logistic regression).

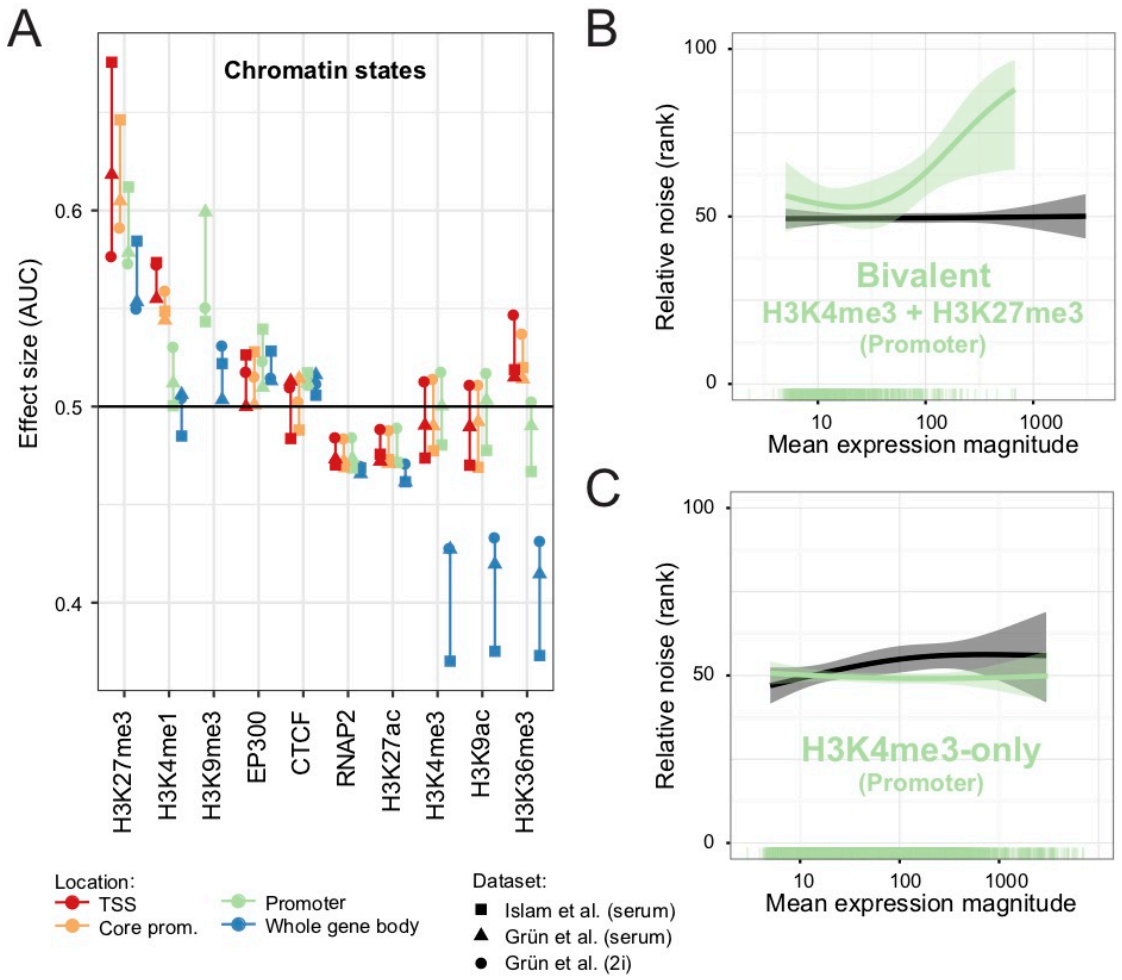

Figure 3. Chromatin states and domains associated with high or low expression noise. A. Noise biases of gene sets constructed based on shared ChIP-seq peaks at the indicated gene regions (see Figure 2A). **B-C.** Binomial smooth results of expression noise versus mean expression magnitude for genes with H3K4me3 and H3K27me3 peaks at their promoters (bivalent) or H3K4me3 peaks only (green). Black curves show background gene set trends, shaded regions indicate 95% confidence intervals and tick marks show individual gene mean expression magnitudes (as in Figure 2C-D). Panels B and C show results obtained using noise estimates from Grün et al. (serum)(Grün et al., 2014). See Figure S11A,B for similar results based on noise estimates from all three UMI-based datasets. Also see Figure S23A-D for comparisons to results where expression noise is represented by adjusted variance values (i.e. without rank normalisation).

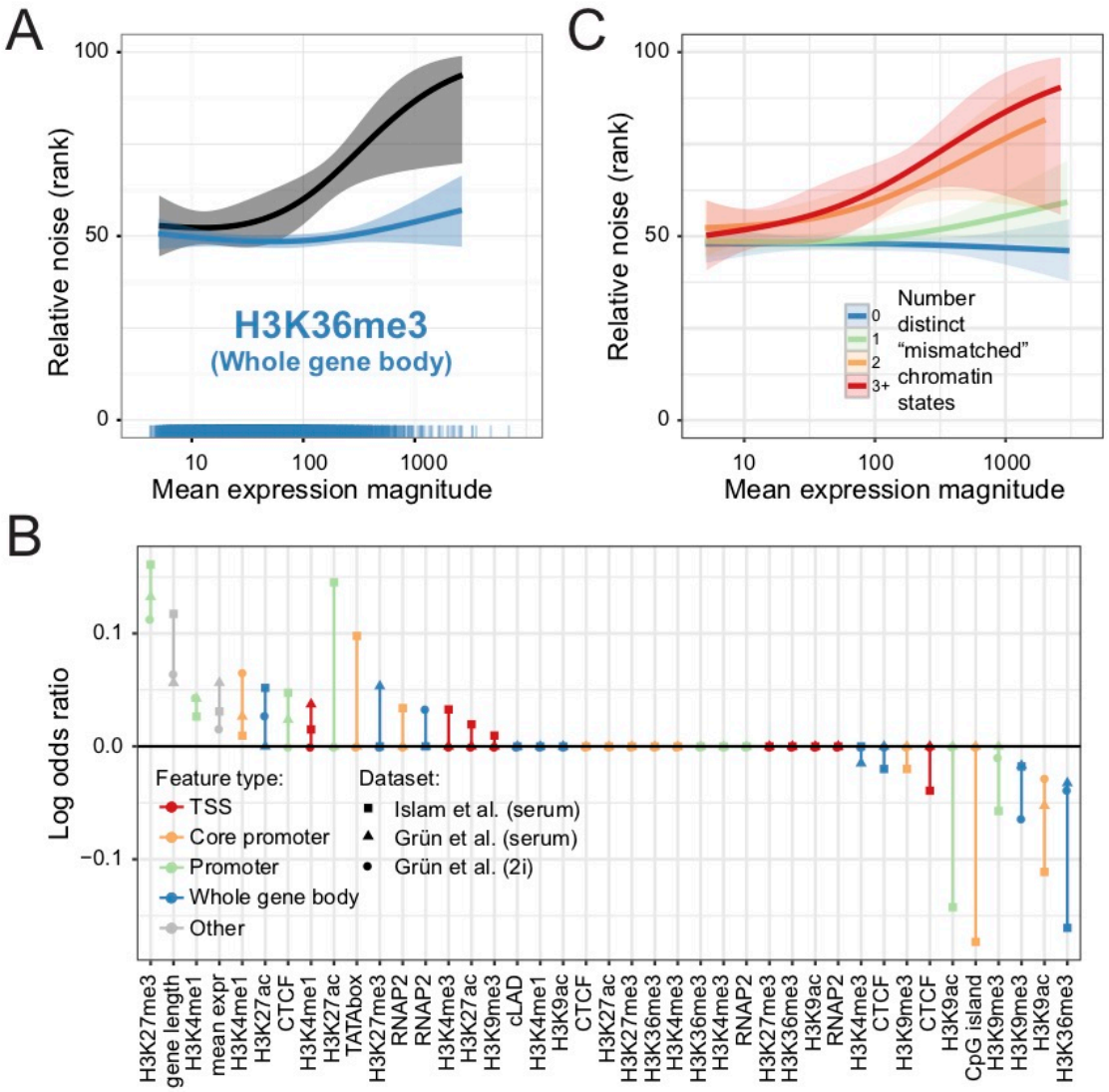

Figure 4. Integrative analysis of chromatin and promoter features. A. Binomial smooth results of expression noise versus mean expression magnitude for genes with H3K36me3 peaks overlapping their bodies (blue). Black curves show background gene set trends, shaded regions indicate 95% confidence intervals and tick marks show individual gene mean expression magnitudes (as in Figure 2C-D). **B.** Feature coefficients in a penalised logistic regression model (Lasso) distinguishing genes with noise levels in the upper tercile from those in the lower tercile (see Figure 2E). The regularisation parameter *lambda* was chosen to yield the most regularized model such that deviance from 10-fold cross-validation was within one standard error of the minimum. We included chromatin state datasets (quantitative ChIP enrichment), gene body constitutive LAD membership, the presence of a TATA box or CpG island at the core promoter (binary features), as well as mean expression level and gene length (continuous features) in the model. See Figure S16A,B for feature coefficients corresponding to single (univariate) and multi-feature logistic regression models (i.e. conventional as opposed to penalised logistic regression). **C.** Binomial smooth results of

expression noise versus mean expression magnitude for genes with increasing numbers of “mismatched” chromatin states defined based on average chromatin state feature associations with mean expression level (see Figure S14E-G). Panels A and C show results obtained using noise estimates from Grün et al. (serum)(Grün et al., 2014). See Figure S11C,D for similar results based on noise estimates from all three UMI-based datasets. Also see Figure S23E-H for comparisons to results where expression noise is represented by adjusted variance values (i.e. without rank normalisation).

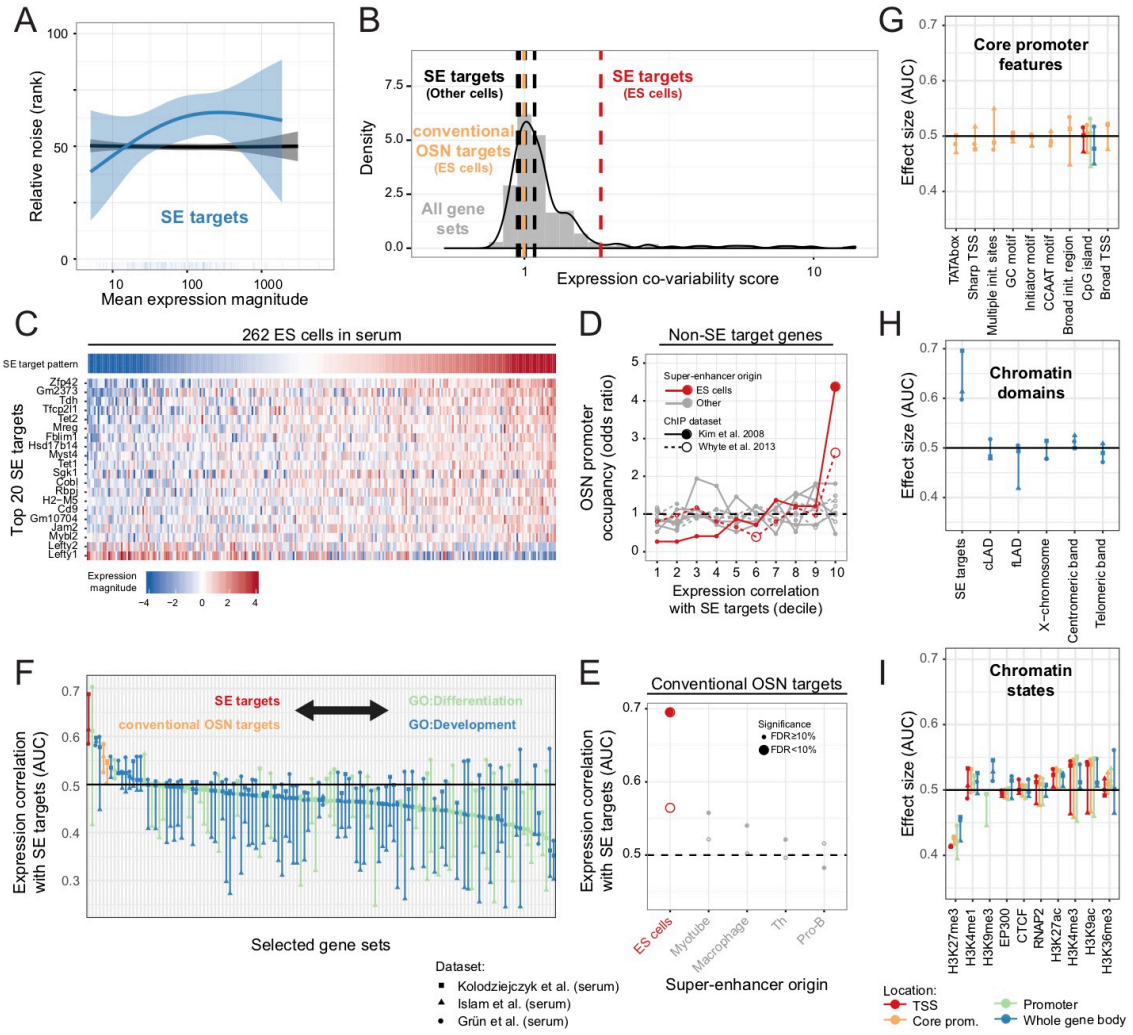

Figure 5. Super-enhancers are associated with high noise due to fluctuations in pluripotency. A. Binomial smooth results of expression noise versus mean expression magnitude for genes with SEs (blue). Black curves show background gene set trends, shaded regions indicate 95% confidence intervals and tick marks show individual gene mean expression magnitudes. Shown are results obtained using noise estimates from Grün et al. (serum)(Grün et al., 2014). See Figure S13B for results based on noise estimates from all three UMI-based datasets. **B.** Histogram (grey bars) and kernel density plot (black line) of weighted PCA-based expression covariability scores for gene sets corresponding to shared sequence, promoter architecture, chromatin state and domain membership features (grey bars). See Figure S10 for the complete list of gene sets. Vertical dashed lines indicate covariability scores of the selected gene sets

indicated. **C.** Heatmap showing normalised expression levels of the top 20 super-enhancer target genes in 262 single mouse ESCs in serum medium. Cells are sorted according to the aggregate SE target pattern. **D.** Enrichment for promoter OSN occupancy (odds ratio) calculated separately for ten equally sized sets of genes constructed based on positive expression correlation level with SE targets (increasing from 1 to 10). Significance was determined based on Fisher's Exact Test. Super-enhancer target and bystander genes themselves were excluded from the analysis. The analysis was repeated using SEs from mouse ES cells (red) as well as those obtained from four other cell types (grey; macrophages, C2C12 mouse myoblast, pro-B and T helper cells). Two independent definitions of OSN promoter occupancy were used based on either ChIP-chip or ChIP-seq enrichment for all three transcription factors (see inset legend). Data point size indicates significance level (see panel E). **E.** Bias in the positive expression correlation of conventional OSN target gene sets with SE targets from mouse ES cells (red) or other cell types (grey). Effect sizes are indicated as area under the receiver operating characteristic curve (AUC), where a value of 0.5 corresponds to a random sample of genes. Symbol size corresponds to significance level. Symbol type indicates ChIP dataset (see panel D inset legend) **F.** Bias in the expression correlation with mouse ESC SE targets for selected gene sets including SE target genes themselves (red), conventional OSN targets (orange) and gene sets based on gene ontology terms containing the words "differentiation" (green) or "development" (blue). Each data point is shown in biological triplicate (three serum medium single-cell RNA-seq datasets) and ranked according to the median effect size (AUC). Results from a similar analysis using 2i/a2i medium single-cell RNA-seq datasets is shown in Figure S14B-D. **G-I.** Bias in the expression correlation with mouse ESC SE targets for core promoter, chromatin domain and chromatin state gene sets. As in panel F, we restricted our analyses to serum medium single-cell RNA-seq datasets (where SE target induced effects are most pronounced). We also controlled for the effect of mean expression level on expression correlation using a sliding window approach (see Figure 1C). Panels B-E show results obtained using single-cell RNA-seq data from Kolodziejczyk *et al.* (serum)(Kolodziejczyk et al., 2015). See Figure S17 for results obtained using 2i/a2i medium UMI- and non-UMI-based datasets.

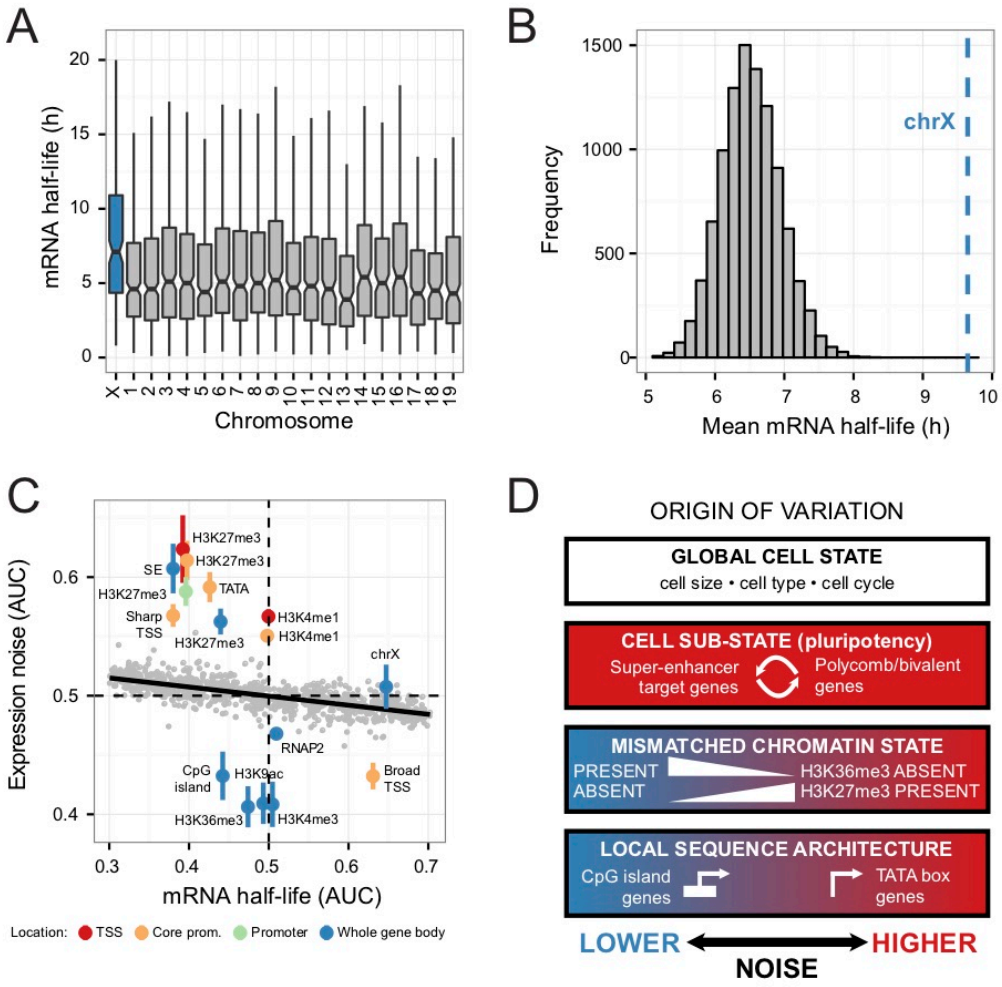

Figure 6. High mRNA stability compensates for the increased transcriptional noise of expression from a single X chromosome. A. Boxplots showing the distribution of mRNA half-lives for genes according to chromosome. **B.** Histogram of mean mRNA half-lives corresponding to random sets of autosomal genes (n=1e4) matching the number of assayed genes on the X chromosome (blue dashed line). **C.** Scatterplot of expression noise bias versus mRNA half-life bias for selected promoter/chromatin gene sets and random gene sets of varying size (200-2000) and mRNA half-life constructed using weighted sampling (grey; n=1e3). Effect sizes are indicated as area under the receiver operating characteristic curve (AUC), where a value of 0.5 corresponds to a random sample of genes. For chromatin promoter/chromatin gene sets, expression noise values are shown as the mean and standard error of three biological replicates (three UMI-based single-cell RNA-seq datasets). For random gene sets (grey), each data point represents the mean expression noise across the same three single-cell RNA-seq datasets (see Figure S19I-K for separate results for each dataset). Panels A-C show results obtained using mRNA decay data from Friedel *et al.* (murine NIH-3T3 fibroblasts)(Friedel et al., 2009). See Figure S19A-H for results obtained using three additional mRNA decay datasets. **D.** Summary scheme showing proposed hierarchical relationship of features associated with gene expression noise. Global differences in cell state (e.g. cell cycle, cell type, cell size) involve coordinated changes in many or all

genes (top panel). Local sequence properties (e.g. core promoter architecture) give rise to gene-specific fluctuations (bottom panel). Additionally, results from our gene set analysis reveal specific chromatin state features associated with expression noise (middle panels). Promoter H3K27me3 / bivalency, genes with super-enhancers and other “mismatched” repressive chromatin states in active genes are associated with high noise (e.g. the absence of H3K36me3). In serum, expression from bivalent H3K27me3-marked promoters anti-correlates with expression from super-enhancer targets indicating fluctuations in Polycomb activity across single cells.

STAR METHODS

QUANTIFICATION AND STATISTICAL ANALYSIS

UMI-based single-cell RNA-seq data pre-processing

Well-identifying barcode demultiplexed raw sequencing reads from Islam *et al.* were processed and filtered as described by the authors (Islam et al., 2014). For the raw pair-end sequencing data from Grün *et al.* (Grün et al., 2014), cell barcode sequences from the left mate were appended to right mate read names to enable post-alignment demultiplexing. Transcript sequences from all UMI-based datasets were mapped to the Ensembl v71 mouse transcriptome (GRCm38) using TopHat (v2.0.8) (D. Kim et al., 2013) with Bowtie2 (v2.1.0) (Langmead and Salzberg, 2012) and the following command-line options, which allow up to 5 mismatches and exclude reads with more than 24 alternative mappings to genomic sequence: *-M -g 24 -N 5 --read-edit-dist 5*. Transcript models were extended by 100bp upstream of the TSS to account for incomplete cap site knowledge. We also included an artificial chromosome consisting of concatenated ERCC spike-in sequences (Baker et al., 2005) each separated by 100bp of ambiguous (N) spacer sequence. After alignment, reads were filtered to remove those with greater than 3 mismatches to unambiguous bases. However, up to 5 matches to ambiguous bases were allowed in the case of ERCC spike-in sequences as they were found to often include a prepended 5' sequence of unknown origin (AATTC). In the case of reads with multiple mappings, we retained only the mapping where the read 5'/3' was closest to an orientation matching TSS/TTS (depending on the end of each fragment sequenced). Unreliable UMIs not supported by sufficient reads (fewer than 1/100 of the average of the nonzero UMIs) were discarded. Gene expression molecule counts were then determine by counting the number of overlapping unique read start:UMI combinations, excluding molecules only supported by a single read or those that could not be assigned unambiguously to a single gene. Only genes on canonical chromosomes and not overlapping mouse ENCODE "black list" regions (downloaded from the UCSC genome browser: https://genome.ucsc.edu/) were considered in our analysis. Molecule counts were also corrected for collision probability, but at the gene rather than the transcript level, as previously described (Grün et al., 2014). We filtered out the same cells as the authors in Islam *et al.* (Islam et al., 2014). For the data from Grün *et al.*, we retained only cells with at least 500 detected spike-in molecules and 10,000 detected endogenous mRNAs.

non-UMI-based single-cell RNA-seq data pre-processing

For the data from Kolodziejczyk *et al.*(Kolodziejczyk et al., 2015), estimated counts for all Ensembl v71 mouse (GRCm38) transcripts (GRCm38) were obtained using Kallisto v0.42.4 (Bray et al., 2016). Estimated counts at the gene level were summarized from transcript-level counts using the *tximport* R package (Soneson et al., 2016). As above, only genes on canonical chromosomes and not overlapping mouse ENCODE "black list" regions (downloaded from the UCSC genome browser: https://genome.ucsc.edu/) were considered in our analysis. Similarly to the authors, we retained only cells with less than 10% of reads mapping to the mitochondrial genome and at least 1 million reads in total (Kolodziejczyk et al., 2015).

Variance normalisation, gene set over-dispersion analysis and cell cycle correction with PAGODA

We used the PAGODA routines of the SCDE R package (Fan et al., 2016) to analyse transcriptional heterogeneity in all single-cell RNA-seq datasets. Molecule counts were first filtered with the *clean.counts* function requiring a gene to be detected in at least two cells for any given dataset (*min.detected=2*). Cell-specific error models were built using the *knn.error.models* function with the number of nearest neighbor cells to use during fitting (*k*) set to one quarter of the total number of cells in each dataset. For non-UMI-based datasets, which tend to suffer from higher levels of noise, a minimum of two reads was required for a given gene to be initially classified as a non-failed measurement (*min.count.threshold=2*). We then normalised gene expression variances relative to transcriptome-wide expectations using the *pagoda.varnorm* function, capping the adjusted variance at a value of five (*max.adj.var=5*) for all datasets, controlling for *batch* in the UMI-based datasets and *gene.length* in the non-UMI-based datasets. We also controlled for the total number of detected genes in each cell (influenced by library and cell size) using the *pagoda.subtract.aspect* function.

Chromatin/promoter type gene set or pathway covariability was determined for each valid set of genes (*min.pathway.size=10, max.pathway.size=2000*) using the *pagoda.pathway.wPCA* function. Briefly, the method runs a weighted PCA analysis on the cell-wise expression levels of the genes in each gene set, as well as analogous analyses of random gene sets of matched size (*n.randomizations=50*). It then tests whether the amount of shared variance captured by the first principal component (PC1) is significantly greater than expected by chance. We used gene sets based on GO terms as well as promoter and chromatin gene set definitions as described below. To control for cell cycle-related variation, we used the set of genes annotated to the GO biological process term “cell cycle”, followed by the *pagoda.subtract.aspect* function with input *aspect* a vector of cell scores corresponding to the first principle component. This procedure was repeated iteratively on each dataset separately until “cell cycle” gene covariability was indistinguishable from that expected by chance (FDR=5%).

Total noise level bias analysis

We rank normalised the cell cycle corrected adjusted variance measures obtained from SCDE/PAGODA using a sliding window approach. For each gene, we ranked its adjusted variance measure compared to its 100 nearest mean expression level neighbours (50 higher and 50 lower; Figure S2C). The 49 most highly and lowly expressed genes in each dataset were discarded due to insufficient neighbours for this analysis. To determine the noise bias of individual gene sets, we tested the deviation of these ranks compared to “background” genes using the Mann–Whitney U Test. For gene body interval (peak overlap) feature associations we excluded short genes from the analysis (<1e4 bp). To verify that our results were robust to the chosen noise metric, we repeated our noise level bias analyses using (raw) adjusted variance values (i.e. without rank normalisation). We also used a more naïve approach to obtain noise estimates based on the distance to the running median (DM) (Newman et al., 2006) of squared CV values in log space (window width, *k=101*). Binomial smooth plots in noise-mean space for selected gene sets were generated using a second-order spline (*formula = y ~ splines::ns(x, 2)*) as argument to the *geom_smooth* function in *ggplot2* (*method=glm, family=binomial*). Shaded regions indicate 95% confidence intervals (*se=T*) constructed on the link scale (*link=logit*), and then back-transformed to the response scale. Smooth

plots using adjusted variance values (i.e. without rank normalisation) were similarly generated (*method=gam, family=gaussian*).

Simulation of single-cell RNA-seq data

An *in silico* dataset was generated based on the two Grün *et al.* (Grün et al., 2014) and one Islam *et al.* (Islam et al., 2014) UMI-based datasets. That is, average expression of genes was calculated from aggregated averages from all three datasets and multiplied by ten (assuming on average 10% capturing efficiency) to arrive at "biological" cell-wise transcript counts. Genes that were not detected in all three datasets were excluded to avoid a long tail of transcript counts towards very low expression levels (~below one transcript mean expression). This resulted in a set of 12,998 genes and a range of average biological transcript counts spanning four orders of magnitude.

Each gene was further assigned a gene expression noise value, assuming a general expression noise structure of the form *gene_expression_noise* = (*promoter_noise* + *1) / average_transcript_counts* + *extrinsic_noise*. Here, the first term in the sum corresponds to *intrinsic_noise* and *promoter_noise* η_p describes the deviation of intrinsic noise from the Poissonian limit due to transcriptional bursting (Pedraza and Paulsson, 2008). For each gene, *promoter_noise* was sampled from a log-normal distribution $\eta_{p,i} = \mathcal{LN}(0,2)$. For each gene, *extrinsic_noise* was sampled from a log-normal distribution $\eta_{ext,i} = \mathcal{LN}\left(\log\left(0.2/\sqrt{2}\right), 0.5\right)$. Transcript counts of each gene *i* across 1,000 cells *j* were drawn from a gamma distribution $X_{i,j} \sim \Gamma(\alpha_i, \beta_i)$, with gene-specific shape parameter $\alpha_i = {m_i^2}/{\left(\eta_{p,i}+1\right)} \cdot \mathcal{LN}(0, \eta_{ext,i})$ and gene-specific rate parameter $\beta_i = {m_i}/{\left(\eta_{p,i}+1\right)} \cdot \mathcal{LN}(0, \eta_{ext,i})$. Here, following Raj *et al.* (Raj et al., 2006), under the assumption of short, instantaneous transcriptional bursts, the shape parameter describes the burst frequency relative to the rate of mRNA degradation and the rate parameter describes the burst size. Multiplying both parameters with a random number drawn from a log-normal distribution with width proportional to the gene-specific extrinsic noise adds cell-specific variation in the parameters, resembling extrinsic influences.

Altogether, this leads to an *in silico* dataset of transcript counts of 12,998 genes in 1,000 cells. Justifying the *ad hoc* parameter choices, the Fano factor distribution resembles a previous estimate (Grün et al., 2014), with ~10% of genes with a Fano factor above 10, and an extrinsic noise floor of ~20% CV. From the initial *in silico* cell population, 16 separate *in silico* scRNA-seq experiments were simulated with varying cell numbers and varying capture efficiencies by randomly sampling cells (1000, 500, 100, 50) and in these cells randomly sampling transcripts (100%, 50%, 15%, 3% capture efficiency c_k). Following Kim & Marioni (J. K. Kim et al., 2015), the single-cell RNA sequencing experiments are modelled assuming incomplete capture of transcripts in the reverse transcription step, but omitting shot noise from sequencing itself and instead assuming saturated sequencing depth. The observed transcript counts in experiment *k* are thus drawn from a binomial distribution $Y_{i,j,k} \sim \text{Binomial}\left(X_{i,j}, c_{j,k}\right)$. Here, capture efficiency is further varied in a cell-specific fashion by drawing the cell-specific capture efficiency from a log-normal distribution $c_{j,k} \sim \mathcal{LN}\left(\log(c_k), 0.1\right)$.

Effect of cell number and transcript capture efficiency on gene set noise estimates

Pre-processing and variance normalisation of all *in silico* scRNA-seq experiments was performed as described above for real UMI-based datasets. Restricting our analysis to genes present in all 16 datasets, we generated 10 independent random sets of 1000 genes with either low (AUC=0.3), high (AUC=0.7) or unbiased (AUC=0.5) noise using adjusted variances from the *in silico* "true" dataset (population of 1000 cells; optimal transcript capture efficiency). Briefly, starting from a random set of 1000 genes (unbiased variance), low or high noise gene sets were constructed by iteratively substituting individual genes (of lower or higher variance respectively) until achieving the desired gene set noise bias. We then determined the individual and combined effects of reduced cell number and capture efficiency on these "true" gene set noise estimates using adjusted variances from all derived datasets. For comparison, we repeated this analysis using relative noise (rank) values in both gene set construction and noise estimation.

Logistic regression models to predict expression noise

Logistic regression models were built using various features to distinguish genes with noise levels in the upper versus lower tercile (genes with intermediate noise levels in the mid tercile were excluded). Model coefficients were used to assess feature importance in (i) single-feature models (including only the feature of interest), (ii) multi-feature or integrative models (including the feature of interest as well as other selected features), and (iii) penalised logistic regression models with lasso regularisation. Single- and multi-feature logistic regression models were fit with the *glm* function in R (*family=binomial, link=logit*) using iteratively reweighted least squares (*method=glm.fit*). Penalised logistic regression models were fit with the *glmnet* package in R (*family=binomial*) using lasso regularisation (*alpha=1*). In all cases, the regularisation parameter *lambda* was chosen to yield the most regularized model such that deviance from 10-fold cross-validation was within one standard error of the minimum (*type.measure=deviance*).

Continuous features (ChIP-seq enrichment, gene expression level, gene length, mRNA half-life, TSS width) were log transformed and standardized to have zero mean and unit variance before model fitting. Binary features (chromosome X, constitutive LAD, super-enhancer target, TATA box and CpG island status) were not standardized. Only genes with non-missing values for all corresponding features were used during model fitting. Core promoter sequence features from EPD were only available for a subset of 14,208 genes. To expand TATA box status to all genes for integrated models including chromatin features, we scanned the EPD TATA box position weight matrix (PWM) over all core promoter DNA sequences. We defined TATA box presence using a maximum score threshold corresponding to an FDR of 10%.

Promoter type and chromatin gene set definitions

Where available, we used mouse ESCs promoter type and chromatin state feature coordinates corresponding to GRCm38/mm10. Otherwise, we converted NCBI37/mm9 coordinates to GRCm38/mm10 using the UCSC genome browser batch coordinate conversion tool (liftOver).

Coordinates of mouse core promoter type and sequence elements (NCBI37/mm9) were downloaded from the Eukaryotic Promoter Database (EPD) (Dreos et al., 2016) using the "EPDnew selection tool" (http://epd.vital-it.ch/EPDnew_select.php) (Dreos et al., 2015). Coordinates of CpG islands (GRCm38/mm10) were downloaded from the UCSC

genome browser (https://genome.ucsc.edu/). Whole embryo (E11-E18) FANTOM5 Cap Analysis of Gene Expression (CAGE) data (Lizio et al., 2015) was downloaded and normalised using the CAGEr R/Bioconductor package (Haberle et al., 2015), followed by individual TSS clustering and aggregation to generate a set of consensus promoters. TSSs were defined as either “Sharp” or “Broad” based on a promoter width threshold of 25bp.

Coordinates of mouse ENCODE (Yue et al., 2014) ChIP-seq peaks (NCBI37/mm9) were downloaded from the UCSC genome browser (https://genome.ucsc.edu/ENCODE/). We used a Hidden Markov model (HMM)-based method (Ernst and Kellis, 2010) to obtain a 15-state chromatin classification based on peaks for seven histone modifications (H3K4me1, H3K4me3, H3K9ac, H3K9me3, H3K27ac, H3K27me3, H3K36me3) and CTCF. To define constitutive and facultative LADs we compared previously computed HMM state calls (Peric-Hupkes et al., 2010) (NCBI37/mm9) for ESCs to those from three other mouse cell types (neural progenitor cells, astrocytes and embryonic fibroblasts). Regions were defined as facultative LADs (fLADs) if they were present in mouse ESCs but absent in any of the other three cell types. Otherwise they were defined as constitutive (cLADs).

Raw Hi-C reads from mouse ESCs were mapped to the mouse genome (GRCm38/mm10) and filtered to obtain normalised contact matrices that were then used to call TADs on the autosomes using TADbit (Serra et al., 2016). In view of results showing that the borders of TADs are enriched for specific regulatory elements and genes with specific biological functions (Dixon et al., 2012), we defined a separate list of genomic domains of fixed size (100kb) centered on these TAD boundaries, which we termed inter-TADs. We used previously defined loops and contact domains from mouse lymphoblasts (Rao et al., 2014) (NCBI37/mm9) in the absence of sufficiently high resolution Hi-C data to call these regions in mouse ESCs. Normalised probe-level mouse ESC replication-timing data (Hiratani et al., 2008) (GRCm38/mm10) was downloaded from the ReplicationDomain database(Weddington et al., 2008) and early replication timing domains defined as contiguous genomic regions of positive signal. The remainder of the genome was defined as late replicating. We used previously defined super-enhancer (SE) regions (Whyte et al., 2013) from mouse ESCs (NCBI37/mm9) as well as four additional cell types (macrophages, C2C12 mouse myoblast, pro-B and T helper cells). We downloaded coordinates of RepeatMasker (Smit et al., 1996) repeat families and chromosomal bands from the UCSC genome browser, defining the first band of each chromosome as “Centromeric” and the last band of each chromosome as “Telomeric”.

K-means clustering of average ChIP-seq signals within TADs recapitulated previous (Sexton et al., 2012) results showing the existence of two major classes: those with either broadly active or repressive chromatin marks. The former are enriched for SEs, whereas the latter are enriched for cLADs. We therefore defined two further TAD classes based on their enrichment for these two features: “TAD (super)” and “TAD (LAD)”. The remaining TADs were classified as either “TAD (active)” or “TAD (repressed)” based on the aforementioned K-means clustering results (K=2). We likewise classified inter-TADs, loops and contact domains according to their occurrence within these four TAD classes. If a domain overlapped two or more TADs of a different class, we termed these “transition” domains.

We then determined the overlap of each of the above genomic features with four distinct gene sub-regions: transcription start site (TSS), core promoter (TSS-200bp, TSS+100bp), promoter (TSS-500bp, TSS+2000bp) and whole gene body (TSS to transcription termination site, TTS). However, the following features could only be sensibly defined at the whole gene level: short genes (<10kb), genes on the X-chromosome (chrX) or mitochondrial chromosome (chrM), genes targeted by nonsense mediated decay (NMD) (Hurt et al., 2013), genes with high/low mRNA stability (Sharova et al., 2009) (upper/lower quartile mRNA half-life). Similarly, promoter type feature occurrence was only interrogated within each gene's core promoter region.

We used a previously described strategy to determine SE target genes (Whyte et al., 2013): the closest active genes i.e. those with H3K4me3 or RNAP2 peaks within 2.5kb of the canonical TSS. SE bystander genes were defined as those overlapping a TAD shared with a SE ("TAD (super)") or within 50kb of a SE if not overlapping a TAD. Individual OSN targets (OCT4/POU5F1, SOX2, NANOG) and combination OSN targets (1/3, 2/3, 3/3) were defined based on predefined gene lists obtained using ChIP-chip data (J. Kim et al., 2008). Additionally, we used a definition based on the promoter overlap of enhancers defined by shared OSN peaks as determined by ChIP-seq (Whyte et al., 2013).

DATA AND SOFTWARE AVAILABILITY

All statistical analyses were performed using custom R scripts that are available upon request.