Title
Discovery of a Cellular Mechanism Regulating Transcriptional Noise

Author
Desai, Ravi

Publication Date
2020

Peer reviewed|Thesis/dissertation

Discovery of a Cellular Mechanism Regulating Transcriptional Noise

by
Ravi Desai

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Cell Biology

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Leor S. Weinberger
0C42BFB71565497...

Leor S. Weinberger
Chair

DocuSigned by:
Benoit G. Bruneau
[illegible]

Benoit G. Bruneau

DocuSigned by:
Jonathan S. Weissman, Ph.D.
A16A9953185F4C9...

Jonathan S. Weissman, Ph.D.

Committee Members

To my parents and wife for their help along this journey

Acknowledgements

This research was conducted with government funding under and awarded by the National Institute of Child Health and Human Development Grant F30HD095614-03 to Ravi Desai and the National Institute of General Medical Sciences Medical Scientist Training Program grant T32GM007618 to the UCSF Medical Scientist Training Program.

This work was made possible by my advisor, Leor S. Weinberger. Leor pushed me to think big and gave me the freedom to explore. His quantitative way of thinking has shaped this work and will surely impact my future as a physician-scientist. His persistent and indefatigable pursuit of an idea has helped me persevere through many failures. His focus on scientific communication has greatly improved my ability to convey the work that I have done.

I am grateful to my thesis committee members, Benoit Bruneau and Jonathan Weissman, for their wise advice which has helped push this work forward. I am grateful to the post-docs of the Weinberger Lab, especially Maike Hansen, Elizabeth Tanner and Sonali Chaturvedi, who have served as great role models for what a scientist should be.

I am grateful to my MSTP classmates, especially Hannah Joo and Tina Zheng, who have joined me on many fun adventures throughout my time at UCSF and have always been available to talk, cook and go on runs.

Lastly, I am grateful to my wife, Trena Mukherjee, who has always brought a smile to my face when I needed it most, even while on the other side of the country/planet at times. Her work in global public health and compassion for others has been inspiring and her love of exploring new cultures has broadened my horizons. She has challenged me to try new approaches and think creatively which I have found to be an essential skill in science.

Contributions

This work is largely adapted from a manuscript under review: Discovery of a Cellular Mechanism Regulating Transcriptional Noise. Desai RV, Hansen MMK, Martin B, Yu C, Ding S, Thompson M, Weinberger LS. *Unpublished* (June 2020)

R.V.D. and L.S.W. conceived and designed the study. R.V.D., C.U., S.D., and L.S.W conceived and designed the cellular reprogramming experiments. R.V.D. and C.U. performed the experiments. R.V.D., M.M.K.H., and B.M. analyzed data. R.V.D., M.M.K.H., B.M. and L.S.W. constructed and analyzed the mathematical models. R.V.D. and L.S.W. wrote the manuscript.

Discovery of a Cellular Mechanism Regulating Transcriptional Noise

Ravi Desai

Abstract

Stochastic fluctuations in gene expression ('noise') are often considered detrimental but, in other fields, fluctuations are harnessed for benefit (e.g., 'dither' or amplification of thermal fluctuations to accelerate chemical reactions). Here, we find that DNA base-excision repair amplifies transcriptional noise, generating increased cellular plasticity and facilitating reprogramming. The DNA-repair protein Apex1 recognizes modified nucleoside substrates to amplify expression noise—while homeostatically maintaining mean levels of expression—for virtually all genes across the transcriptome. This noise amplification occurs for both naturally occurring base modifications and unnatural base analogs. Single-molecule imaging shows amplified noise originates from shorter, but more intense, transcriptional bursts that occur via increased DNA supercoiling which first impedes and then accelerates transcription, thereby maintaining mean levels. Strikingly, homeostatic noise amplification potentiates fate-conversion signals during cellular reprogramming. These data suggest a functional role for the observed occurrence of modified bases within DNA in embryonic development and disease.

Contents

List of Figures

Chapter 1

Introduction

From Brownian motion to electrical 'shot' noise, fluctuations are fundamental to physical processes. Since the 1800s [1], fluctuations have been recognized to dynamically shape the distribution of microstates a system adopts, and modulation of fluctuations has been harnessed throughout engineering and the sciences. For example, in chemistry, thermal fluctuations—amplified via temperature increase (e.g., Bunsen Burners)—accelerate reactions [2]; in engineering, amplification of electrical, acoustic, or mechanical fluctuations (i.e., 'dither', from the Middle English "didderen" meaning to "tremble") is used for signal recovery [3], and in neuroscience, electrophysiological fluctuations—first reported in the 1950s [4]—are clinically amplified to improve sensorimotor function [5, 6, 7]. Such 'dither' approaches break Poisson dependency so that Δvariance $\neq \Delta$ mean.

Evolutionary theories dating to the 1960s [8, 9, 10] proposed that biological organisms maximized fitness by harnessing putative fluctuations to enable probabilistic 'bet-hedging' decisions. Subsequent studies showed that intrinsic molecular fluctuations in gene expression (i.e., stochastic 'noise'), modulated by gene-regulatory circuits, enabled probabilistic fate selection (Figure 1.1A) in diverse biological systems [11, 12, 13]. Open questions remain as to whether cellular noise control is limited to inherently locus-specific gene-regulatory circuits or if generalized noise-modulation mechanisms exist, if and how such mechanisms might orthogonally tune noise

independent of mean, and, given the detrimental effects of noise, if such putative mechanisms might be regulated 'on-demand' to potentiate cell-fate specification.

Non-genetic variability or noise in gene expression, often quantified by measurement of cell-to-cell variability in reporter expression, can arise from multiple sources, both intrinsic and extrinsic. Extrinsic noise refers to correlated variability across two or more promoters in the same cell. Cell cycle, cell size and abundances of communal cellular components like ribosomes contribute to extrinsic noise [14]. Intrinsic noise results from stochastic fluctuations in rates of biochemical reactions responsible for birth, maturation and death of mRNA and protein molecules [15, 16]. Although the relative contribution of intrinsic vs. extrinsic factors to expression noise differs for each gene, the small number of reactants involved in many biochemical processes renders stochasticity an inescapable aspect of gene expression that can either be amplified or suppressed.

Stochasticity or intrinsic noise first emerges from transcription, where the collision of small numbers of reactants undergoing Brownian motion culminates with discrete realizations: the birth of mRNAs [17, 18]. In the diffusion-limited case where noise is lowest, cell-to-cell divergence in transcript numbers follows a Poisson distribution where variance is equal to the mean and the Fano factor is equal to one [19]. In both prokaryotes and eukaryotes, however, many genes have transcript variances far greater than the mean (super-Poissonian, Fano factor >1). Elegant single-molecule imaging techniques have revealed a significant cause of this enhanced noise: many genes engage in short periods of high transcriptional activity that are interspersed with long periods of inactivity [17, 18, 20, 21, 22, 23, 24]. Toggling of a promoter between an inactive OFF state and a productive ON state, also known as transcriptional bursting, is a significant contributor to enhanced noise on both the mRNA and protein level [25].

The two-state random-telegraph model describes this bursting via two parameters: (i) the fraction of time a promoter is active $(K_{ON}/[K_{ON}+K_{OFF}])$, and (ii) the number of transcripts produced

during the ON state (burst size, K_{TX}/K_{OFF}) [26, 27, 28]. These bursting parameters are tuned by regulatory machinery [24] like histone acetyltransferases, which can increase burst frequency by facilitating nucleosome clearance from promoters thereby increasing mean transcriptional levels [29]. Increases in mean expression (μ) are typically accompanied by a stereotypical reduction in noise measured by coefficient of variation, CV, (σ/μ) (Figure 1.1B), whereas stressors that decrease mean are typically accompanied by an increase in noise [30, 31, 32]. This $1/\mu$ scaling of noise can be broken by gene-regulatory circuits such as feedback and feedforward loops [33], and some small-molecule pharmaceuticals can modulate transcriptional fluctuations/noise (σ/μ) independent of change in mean (μ) [34, 35]. Since some molecules can amplify expression noise of diverse unrelated promoters [34, 36], we asked if these molecules might be functioning via disruption or enhancement of a putative cellular noise-control mechanism.

In E. coli, DNA supercoiling and the cooperative recruitment of RNA polymerases have been proposed as general properties of transcription that together account for the bursting phenomenon and set mechanical bounds on transcriptional noise [37, 23, 38, 39]. According to the twin-supercoiled-domain model, translocation of an RNA polymerase leads to overwinding of downstream DNA (positive supercoils) and underwinding of upstream DNA (negative supercoils) [40, 41, 42]. This can create both positive and negative feedback loops where upstream negative supercoiling facilitates additional polymerase recruitment while buildup of downstream positive supercoils inhibits elongation [37, 39, 43]. For eukaryotic systems however, it is unclear whether the mechanical limitations imposed by supercoiling affect transcriptional bursting and gene expression noise more broadly? Furthermore, if supercoiling is a modifier of transcriptional bursting as shown in prokaryotes, could it serve as a lever to actively control gene expression noise in eukaryotes?

Here, we find that DNA base-excision repair amplifies transcriptional noise, generating increased cellular plasticity and facilitating reprogramming. Using mouse embryonic stem cells (mESCs) as our model eukaryotic system and a set of small molecules identified as transcriptional

noise enhancers, we uncovered how the DNA-repair protein Apex1 recognizes modified nucleoside substrates to amplify expression noise—while homeostatically maintaining mean levels of expression—for virtually all genes across the transcriptome. This noise amplification occurs for both naturally occurring base modifications and unnatural base analogs. Single-molecule imaging shows amplified noise originates from shorter, but more intense, transcriptional bursts that occur via increased DNA supercoiling which first impedes and then accelerates transcription, thereby maintaining mean levels. Stochastic modeling predicted a parameter space for tuning noise enhancement, which we experimentally validate using several modulators of DNA repair and topoisomerase activity. Strikingly, homeostatic noise amplification potentiates fate-conversion signals during cellular reprogramming. The results suggest that gene expression noise partly emerges from the biophysical constraints of DNA, which are tuned by topoisomerases. Additionally, these data suggest a functional role for modified bases within DNA as modulators of cellular plasticity in embryonic development and disease.

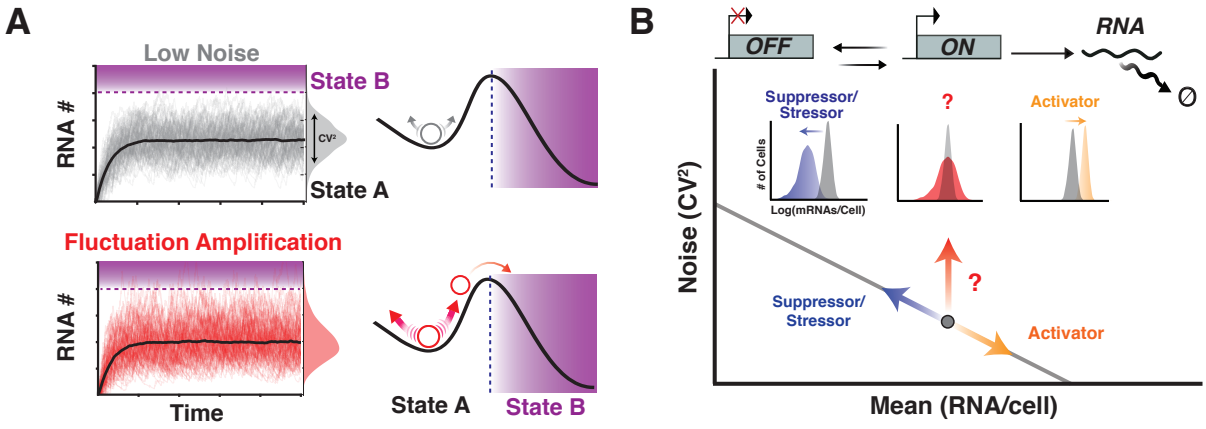


Figure 1.1: Search for a cellular noise-control mechanism.
(A) (Left) Monte-Carlo simulations of the two-state Random-Telegraph model of transcription showing low noise and higher noise trajectories with matched mean expression levels. Coefficient of Variation (σ^2/μ^2, CV^2) quantifies magnitude of fluctuations. (Right) The predicted facilitation of state transitions through 'dithering'. **(B)** (Top) Schematic of two-state Random-Telegraph model of transcription. (Bottom) Schematic of mean vs. CV^2 for mRNA abundance with solid gray line representing Poisson, inverse scaling of CV^2 as a function of mean. Question mark symbolizes unknown noise-control mechanisms that amplify fluctuations independently of mean. Histograms depict expected shift in mRNA copy number distributions.

Chapter 2

Identification of a global noise-enhancer

2.1 Introduction

To uncover what processes modulate transcriptional bursting and thus gene expression noise, we used the following logic: identify perturbations that alter bursting kinetics and then elucidate the mechanism of action. From a library of 1600 FDA-approved small molecules, the Weinberger lab recently identified 85 compounds that amplify expression noise of the HIV promoter without altering the promoter's mean-expression level [34]. From these results, we asked whether the noise-enhancing compounds alter a gene-specific property of the LTR promoter or perturb a more ubiquitous regulatory control on bursting kinetics which would manifest as generalizability to other cell-types and promoters.

2.2 Results

2.2.1 5'-iodo-2'-deoxyuridine (IdU) increases gene expression noise in Jurkat and K562 cells

I first tested a subset of noise-enhancer compounds on Jurkat cells harboring a d_2GFP reporter driven by a EF1αpromoter and K562 cells harboring a d_2GFP reporter driven by either a EF1αor

UBC promoter. I identified one compound, 5'-iodo-2'-deoxyuridine (IdU), which consistently increased expression noise (Figure 2.1).

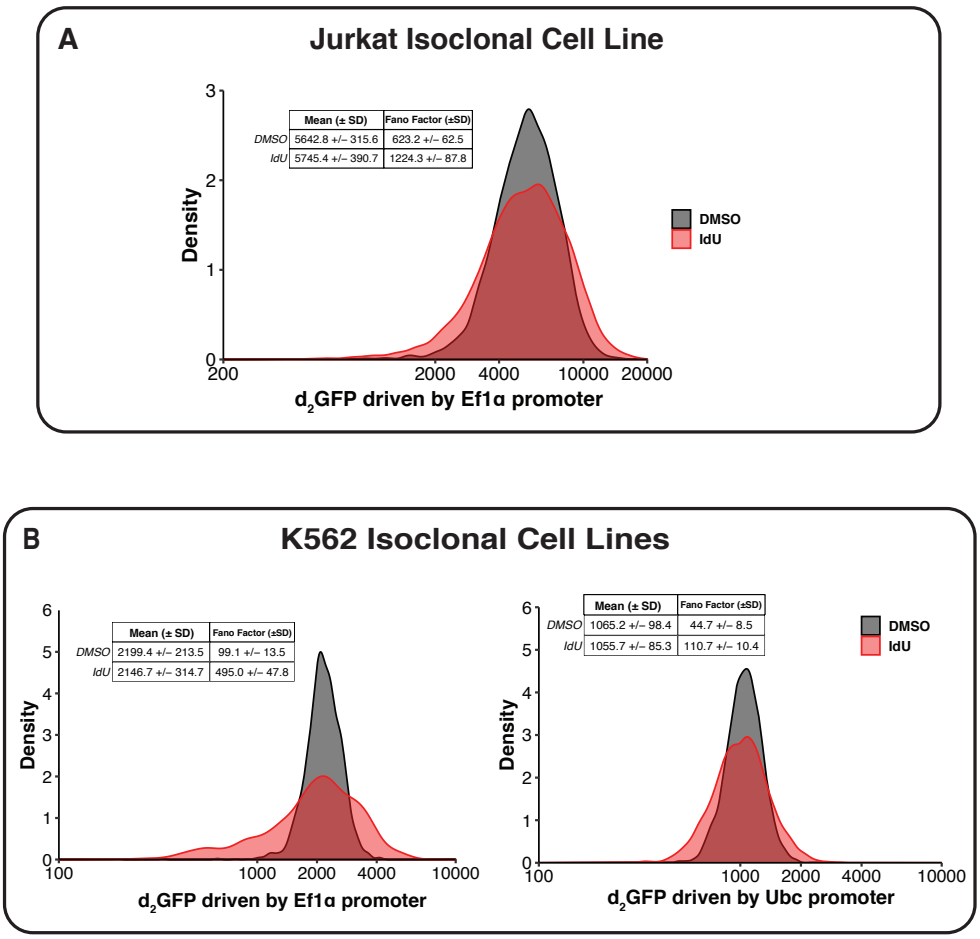


Figure 2.1: Nucleoside analog increases expression variability of housekeeping promoters in Jurkat and K562 cells.
(A) Representative flow cytomtery distributions of d_2GFP expression in an isoclonal population of Jurkat cells treated with either 20μM IdU or equivalent volume DMSO for 24 hours. Mean and SD are derived from 2 biological replicates. **(B)** Representative flow cytomtery distributions of d_2GFP expression in isoclonal populations of K562 cells treated with either 20μM IdU or equivalent volume DMSO for 24 hours. Mean and SD are derived from 2 biological replicates.

Methods

Noise Enhancer Testing on Isoclonal Jurkat and K562 Cells

Isoclonal Jurkat T Lymphocytes with lentivirally integrated EF1α-d_2GFP construct were previously described (36). Human immortalized myelogenous leukemia (K652, female) cells were cultured in RPMI-1640 medium (supplemented with L-glutamine, 10% fetal bovine serum, and 1% penicillin-streptomycin), at 37°C, 5% CO_2, in humidified conditions at 2×10^5 to 2×10^6 cells/mL. Isoclonal K562 cells with lentivirally integrated EF1α-d_2GFP and UBC- d_2GFP constructs were previously described (36).

Jurkat and K562 cells were seeded into 12-well plates at densities of 0.2×10^6 and 0.4×10^6 cells/mL respectively in media containing 20 μM IdU (Sigma, cat:I7125, dissolved in DMSO) or equivalent volume of DMSO for 24 hours. Flow cytometry was performed using a BD LSRII cytometer. Treated cells were run unfixed and live to avoid additional sources of variability from fixation. 50k live cells were collected from each sample for noise measurements. Conservative gating for a live subset of approximately 3k cells of similar size, volume, and state, was applied on the FSC vs. SSC to reduce extrinsic noise contributions as previously described (15,18).

2.2.2 IdU increases global gene expression noise in mouse embryonic stem cells

To test whether IdU has an effect on the global structure of mRNA noise, single-cell sequencing was performed on mouse embryonic stem cells (mESCs) treated with either 10μM IdU or DMSO for 24 hours. mESCs were chosen as the model system due to their extensive characterization in both developmental and noise biology. Strikingly, single-cell RNA sequencing (scRNA-seq) of mESCs maintained in 2i/LIF media—after filtering and normalization using Seurat [44]—showed

that IdU amplified cell-to-cell variability in transcript levels (i.e., transcript noise) for virtually all genes across the genome—4,578 genes analyzed—with little alteration in mean transcript levels for most genes, as analyzed by either CV^2 or variance versus mean (Figures 2.2A-B). To account for the Poisson scaling of variance on mean, transcript noise was also quantified using the Fano factor, which measures how noise deviates from Poisson scaling [45, 27, 46]. Despite mean-expression levels exhibiting minimal changes (Figure 2.2C), the Fano factor increased for $>90\%$ of genes (Figure 2.2D) with lowly expressed genes showing a slightly greater change in Fano (Figure 2.3). These results of a global increase in transcript noise with little change in mean levels are in stark contrast to the effects of transcriptional activators or cellular stressors that alter noise in a stereotypic manner together with changes in mean [31, 21].

To account for technical noise and quantify statistical significance of changes in noise and mean, we used an established Bayesian hierarchical model [47, 48] to create probabilistic, gene-specific estimates of both mean expression and cell-to-cell transcript variability. Of the 4,578 genes, the algorithm classified 945 genes ($\sim 20\%$) as highly variable, whereas 113 genes ($\sim 2\%$) showed a significant change in mean expression (Figure 2.2E-F). Bulk RNA-seq measurements of mean abundances—performed using ERCC spike-ins for normalization—confirmed the scRNA-seq findings (Figure 2.4). Thus, analyses from two methods (Seurat and BASiCS) show that IdU induces a significant increase in transcript variability (expression noise) but comparatively little change in mean expression in mESCs.

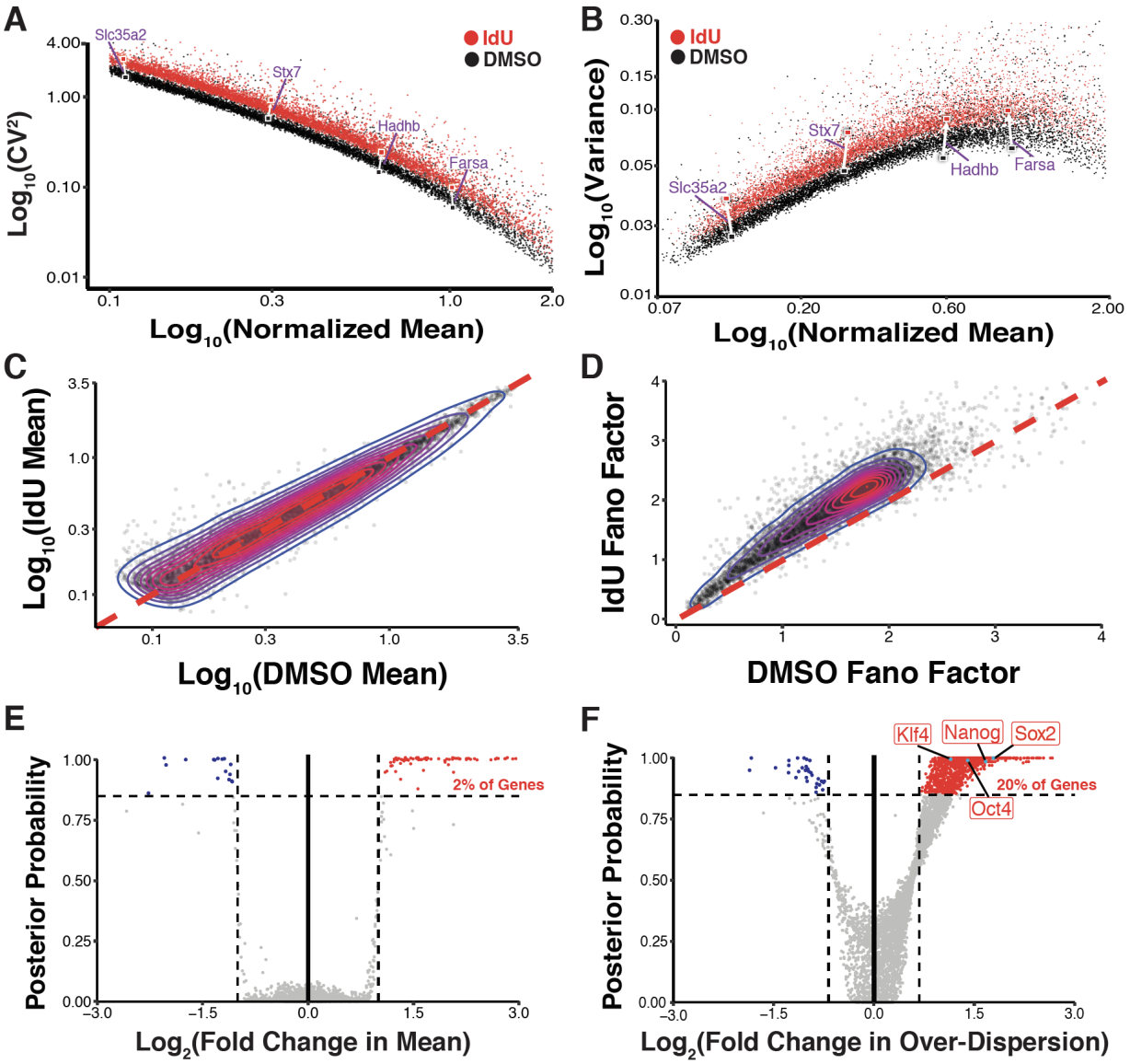


Figure 2.2: Genome-wide amplification of cell-to-cell mRNA variability (i.e., 'noise') independent of mean.

(A-D) scRNA-seq of mESCs treated with DMSO (black) or 10μM IdU (red) for 24h. 812 and 744 transcriptomes (filtered and normalized with Seurat) from DMSO and IdU treatments, respectively, were analyzed. (A) Mean expression vs. CV^2 and (B) mean vs. variance for 4,578 genes. Four examples of housekeeping genes (purple) demonstrate how IdU increases expression fluctuations with minimal change in mean (white arrows). (C) Mean expression and (D) Fano factor of 4,578 genes in DMSO vs. IdU treatments. Overlay of density contours reveals how center of mass lies on diagonal for mean values while lying above the diagonal for Fano factor measurements. **(E-F)** BASiCS analysis of scRNA-seq data for 4,578 genes. (E) Fold change in mean vs. certainty (posterior probability) that gene is up- or down-regulated. With IdU treatment, 113 genes (red) were classified as differentially expressed (>2-fold change in mean with >85% probability). (F) Fold change in over-dispersion vs. certainty (posterior probability) that gene is highly- or lowly-variable. 945 genes (red) were classified as highly variable (>1.5-fold change in over-dispersion with >85% probability).

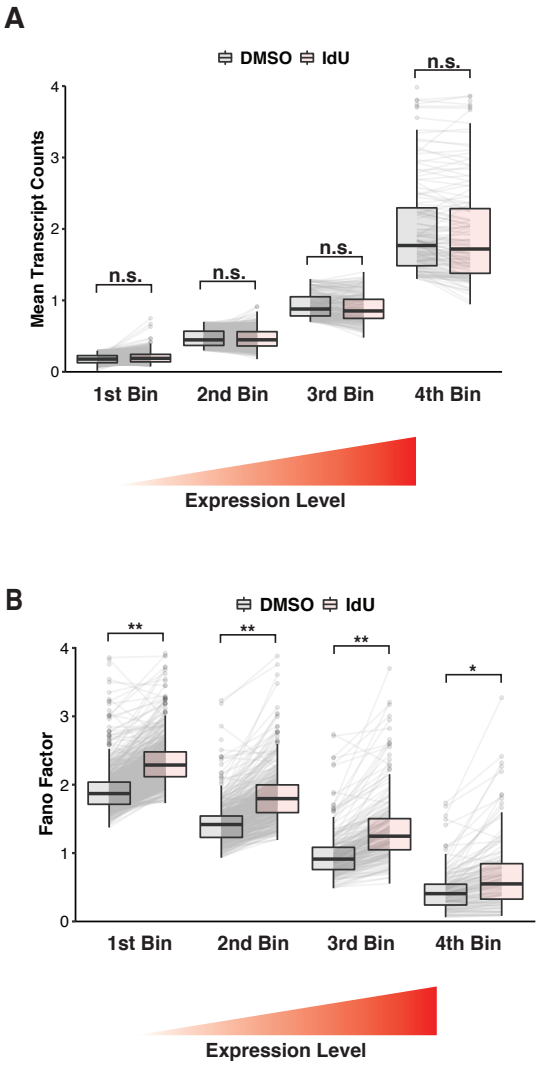


Figure 2.3: Noise enhancement occurs for genes across all expression levels.
4,578 genes from scRNA-seq dataset were binned into one of four groups (quartiles) based on mean expression level in DMSO condition. **(A)** Comparison of mean expression level for each gene in IdU and DMSO treatment groups. Boxplots show median ±interquartile range of mean values for genes within each bin. Solid lines connect the same gene in the DMSO and IdU boxplots. P values were calculated using a two-tailed, paired Student's t test. **(B)** Comparison of Fano factor for each gene in IdU and DMSO treatment groups. Boxplots show median ±interquartile range of Fano factors for genes within each bin. Solid lines connect the same gene in the DMSO and IdU boxplots. P values were calculated using a two-tailed, paired Student's t test. **p < 0.001, *p = 0.0016

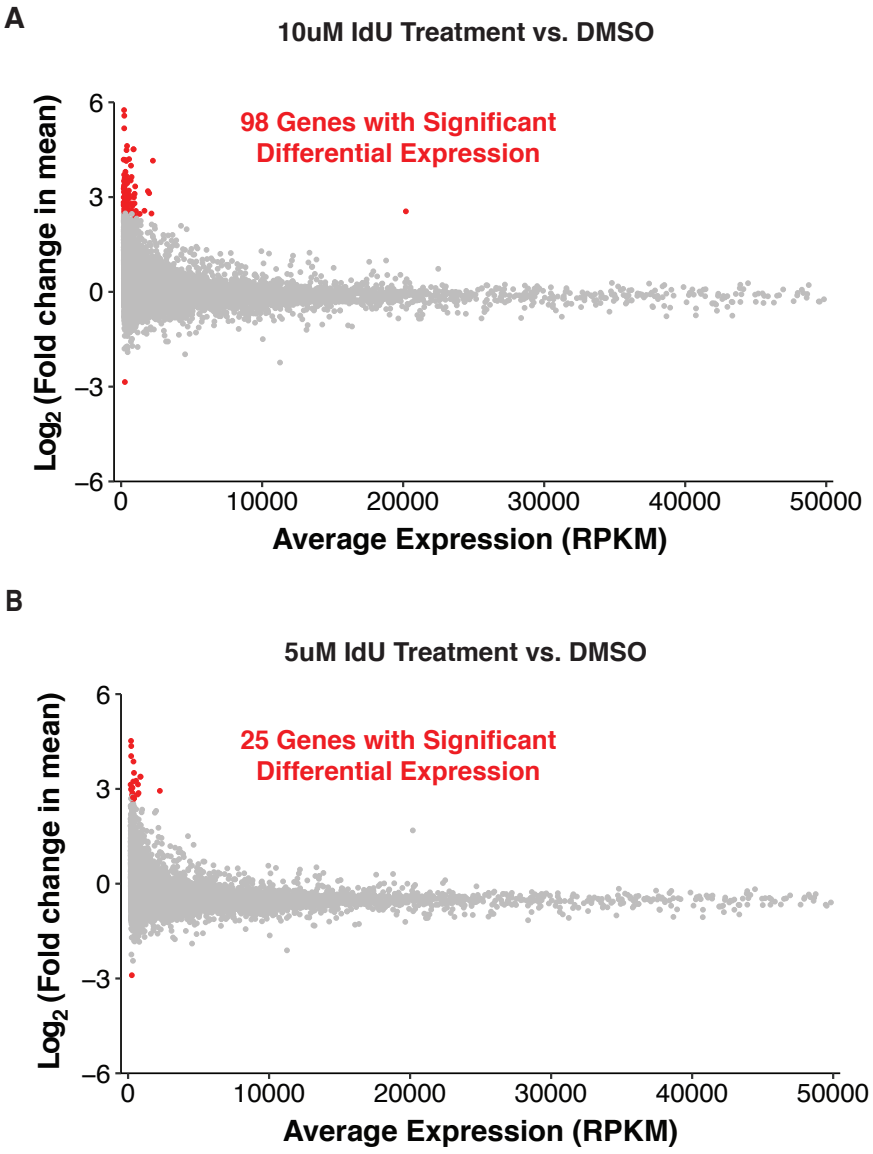


Figure 2.4: IdU causes minimal change in mean gene expression levels as measured by bulk RNA-seq.
Transcript abundances were normalized using ERCC spike-in counts. Differential mean testing was conducted with a threshold of fold change > 2 and an FDR cutoff of 0.05. Genes considered differentially expressed are highlighted in red. **(A)** Mean transcript abundance vs. fold change (Log_2) in mean for 12,502 genes. Comparison is between 10μM IdU and DMSO treatments. 98 genes were identifed as differentially expressed. **(B)** Mean transcript abundance vs. fold change (Log_2) in mean for 12,054 genes. Comparison is between 5μM IdU and DMSO treatments.

Methods

Single-Cell RNA Sequencing Preparation and Analysis

Mouse E14 embryonic stem cells (male) were routinely cultured in feeder-free conditions on gelatin-coated plates with ESGRO-2i medium (Millipore, cat:SF016-200) at 37°C, 5% CO_2, in humidified conditions [45].

1x10^6 mESCs were seeded in a gelatin-coated, 10cm dish in 2i/LIF media. 24 hours following seeding, cultures were replenished with 2i/LIF media containing 10μM IdU or an equivalent volume of DMSO for 24 hours. After treatment, cells were trypsinized with TrypLE and spun down for 5 minutes at 200 x g. Single-cell suspensions were prepared in DPBS at a concentration of 83,000 cells/ml. Approximately 3000 cells from each sample were loaded into a chip and processed with the Chromium Single Cell Controller (10x Genomics). To generate single-cell gel beads in emulsion (GEMs), DMSO- and IdU- treated samples were assigned unique indexes using Single Cell 3′ Library and Gel Bead Kit V2 (10x Genomics, cat:120237). Sequencing was performed on an Illumina HiSeq4000 with a paired-end setup specific for 10x libraries.

Data were aligned to mm10 reference genome using 10x Cell Ranger v2. Quality control, normalization and analysis were carried out using two packages: Seurat and BASiCS. For analysis in Seurat, gene-barcode matrices were filtered and normalized using the "LogNormalize" method, resulting in 812 and 744 transcriptomes from DMSO and IdU samples. Transcript variability was quantified using variance (σ^2), coefficient of variation ($\frac{\sigma^2}{\mu^2}$), and Fano factor ($\frac{\sigma^2}{\mu}$). During the normalization procedure in Seurat, counts for the i^{th} gene in the j^{th} cell (x_{ij}) are multiplied by the following scaling factor: $S = \frac{10000}{\sum_{i=1}^{n} x_{ij}}$, where n is the number of genes in the dataset. The scaling factor is therefore dependent on the number of UMIs detected per cell. The coefficient of variation is insensitive to this scaling factor as it is a dimensionless quantity (i.e, σ and μ are scaled by the same factor and thus cancel out when calculating coefficient of variation). However, the Fano factor,

which has units, must be re-scaled to account for the differential effect that this normalization procedure has on σ^2 vs. μ (i.e., σ^2 gets scaled by S^2 while μ gets scaled by S). To negate the carryover of this scaling factor, calculated Fano factors from the Seurat-normalized dataset were multiplied by $\frac{1}{S}$ where S is a unique value for the DMSO and IdU samples: $\frac{10000}{\text{avg. number of UMIs per cell in sample}}$. On average, 4151.3 and 4191.4 UMIs were detected per cell in DMSO and IdU samples respectively.

For analysis using BASiCS, quality control and filtering was performed using the BASiCS_Filter function resulting in an identical number of transcriptomes (812 and 744) as produced by Seurat. Posterior estimates of mean and over-dispersion for each gene were computed using a Markov Chain Monte Carlo (MCMC) simulation with 40,000 iterations and a log-normal prior. For differential mean testing, a threshold of fold change >2 with an FDR cutoff of 0.05 was used. Differential variability was tested with a threshold of fold change >1.5 with an FDR cutoff of 0.05. Only genes with no change in mean expression (4,458 of 4,578) were considered for interpreting changes in variability.

Gene features and sequences from the GRCm38 reference were used for analysis of gene characteristics that potentiate noise enhancement. TAD boundary locations in mESCs were taken from previously established Hi-C maps [49]. DAVID v6.8 was used to test for gene ontology (GO) enrichment among highly variable genes. All tested genes (4,458) from BASiCS were used as background. Bonferroni-corrected p-values (adjusted p-values) were used to visualize GO enrichment. Cell cycle determination was performed using *cyclone* as implemented in *scran* [50]. The default set of cell cycle marker genes for mESCs (mouse_cycle_markers.rds) was used. Cells were assigned to G1, S, and G2/M phases using their normalized genes counts produced by Seurat. Pseudotime analysis was conducted using *destiny* [51] with the Seurat-normalized cell-gene matrix as input. Gene-gene correlation matrices were assembled by first filtering out genes from the Seurat-normalized matrix whose mean abundance <1 in each treatment group to avoid spurious

correlations that may emerge from low expression. 961 genes remained for downstream analysis. Pearson correlation for each gene pair was calculated. Clustering of gene-pairs based on similarity in correlation patterns was performed using the hierarchical clustering method within the *seriation* package. Change in correlation strength was calculated by subtracting absolute value of gene-pair correlation in DMSO condition from IdU condition.

Bulk RNA Sequencing Preparation and Analysis

$2x10^5$ mESCs were seeded in each well of a gelatin-coated, 6-well plate in 2i/LIF media. 24 hours following seeding, cultures were replenished with 2i/LIF media containing 10μM IdU, 5μM IdU or an equivalent volume of DMSO in triplicate for 24 hours. After treatment, cells were trypsinized with TrypLE and RNA was extracted using a RNeasy minikit (Qiagen) according to manufacturer's instructions. ERCC spike-in RNA (2μl diluted at 1:100) was added to each RNA extraction (Ambion, cat:4456740). A total of 9 cDNA libraries were prepared with an NEBNext Ultra II RNA Library Prep kit (NEB, cat:E7770S) and sequenced with an Illumina HiSeq4000. Sequencing yielded a median of ≈40 million single-end reads per library. Read quality was checked via FASTQC. Reads were aligned to an edited version of the mm10 reference genome containing the ERCC spike-in sequences using TopHat with default parameters. Transcript level quantification was performed using Cufflinks with default parameters. The quantification matrix was then imported into R and analyzed via DESeq2. Samples were normalized using ERCC transcripts as controls for size factor estimation. Differential mean testing was conducted with a threshold of fold change >2 and an FDR cutoff of 0.05.

2.2.3 Noise-enhancement occurs through intrinsic mechanism: reciprocal modulation of burst duration and intensity

To examine if certain characteristics could explain a gene's potential for noise enhancement we examined (i) gene length, (ii) promoter and (iii) gene-body AT content, (iv) number of exons, (v) TATA-box inclusion and (vi) strand orientation. None of these characteristics exhibited predictive power or correlated with a gene's potential for noise enhancement (Figures 2.5A-F). However, genes susceptible to high noise enhancement were preferentially located within the interior of topologically associated domains (TADs), suggesting gene topology influences potency of noise enhancement (Figure 2.5G). Ontology analysis of highly variable genes showed enrichment of house-keeping pathways along with pluripotency maintenance factors, particularly Sox2, Oct4, Nanog and Klf4 (Figure 2.6). As these pluripotency maintenance factors are key influencers of cell-fate specification, we next focused on the molecular mechanisms driving their amplified transcript noise.

We first tested whether the enhanced variability arose from extrinsic factors, which include cell-cycle phase and cell-type identity. Cells within the scRNA-seq dataset were computationally assigned a cycle stage (G1, S, G2/M) [50] which showed that Nanog, Oct4, Sox2 and Klf4 were highly variable in each cell-cycle phase, indicating that their variability is not cell-cycle dependent (Figure 2.7). Moreover, pseudo-time analysis showed no bifurcations, indicating transcriptional variability was not due to a differentiation-induced mixture of cell-types (Figure 2.8).

Extrinsic variability may also arise from the coordinated propagation of noise through gene-regulatory networks [52, 53] and can be measured by gene-to-gene correlation matrices [54, 55]. If the increase in global transcript noise is extrinsic, expression correlation between network partners would increase or remain unchanged. Analysis of gene-to-gene correlation matrices showed that $\sim$80% of gene-gene pairs lost correlation strength following IdU treatment (Figure 2.9), indicating that enhanced expression noise is uncorrelated and not consistent with an extrinsic noise source.

Exclusion of these extrinsic noise sources suggested that IdU amplifies intrinsic noise arising from stochastic fluctuations in transcript birth (promoter toggling) or death (degradation).

To test whether a change in promoter toggling could account for IdU-enhanced noise, we used single-molecule RNA FISH (smRNA-FISH) to count both nascent and mature transcripts of Nanog, a master regulator of pluripotency. Spot counting was performed on a mESC line in which both endogenous alleles of Nanog are fused to eGFP. This fusion does not alter mRNA or protein half-life or impair differentiation potential [56]. To target mature transcripts, smRNA-FISH probes to eGFP were used, and to minimize extrinsic noise, analyses were limited to cells of similar sizes (Figure 2.10A). Consistent with scRNA-seq, smRNA-FISH showed a large increase in cell-to-cell variability of mature Nanog transcripts ($\sim$2-fold increase in Fano) with little change in mean Nanog levels (Figure 2.10B). Fewer IdU-treated cells exhibited active transcriptional centers (TCs), but the number of nascent mRNAs at each TC increased (Figures 2.10C-D). Fitting of the two-state random-telegraph model to smRNA-FISH data revealed that increased variability was due to a shortened burst duration (increased k_{OFF}) and amplified transcription rate (higher k_{TX}) (Figure 2.10E). These results represent direct validation of previous predictions [27, 34] that enhanced noise could arise from reciprocal changes in transcriptional burst duration ($1/k_{OFF}$) and intensity (k_{TX}).

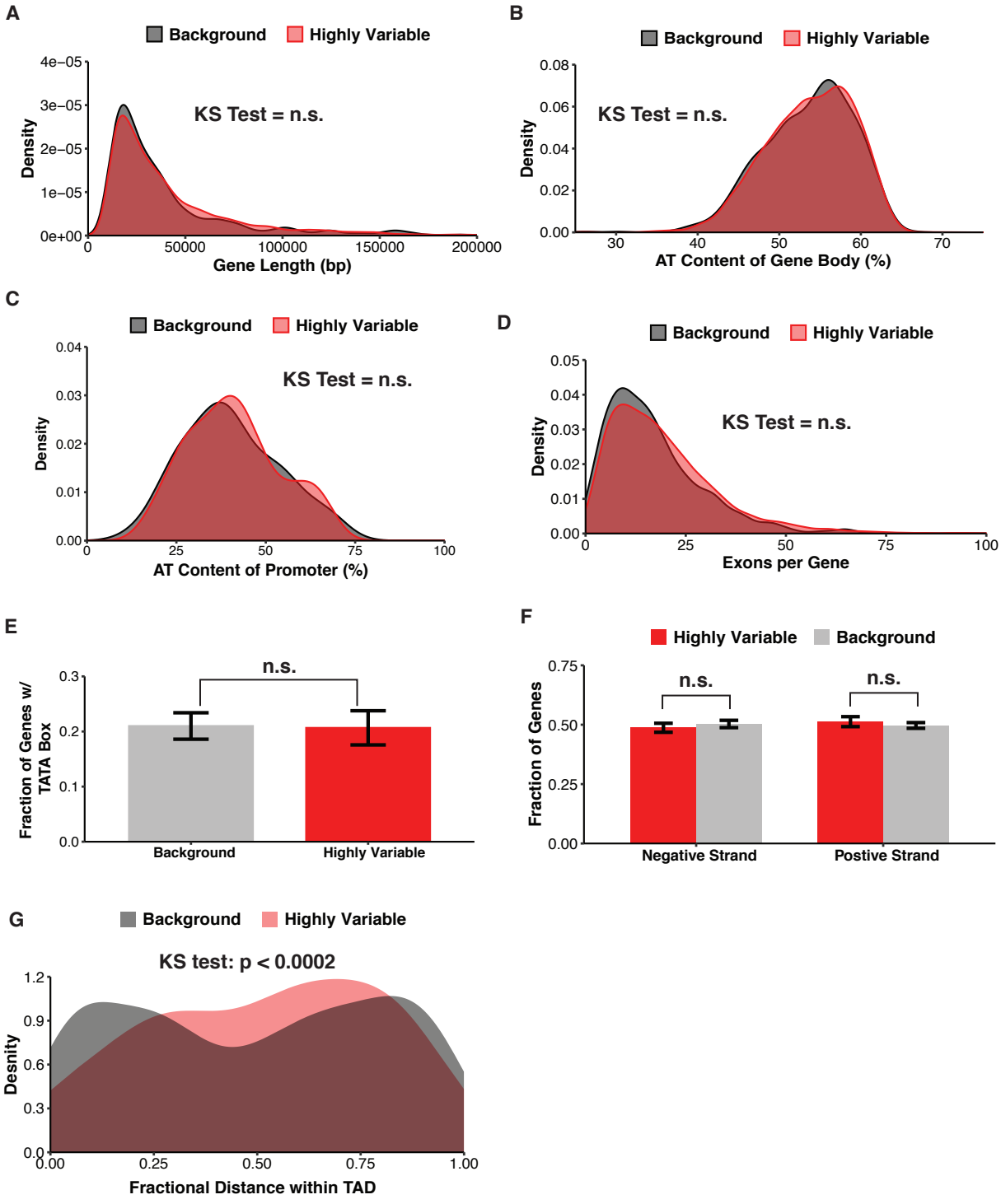

A
Background
Highly Variable
KS Test = n.s.
Density
Gene Length (bp)
B
KS Test = n.s.
AT Content of Gene Body (%)
C
KS Test = n.s.
AT Content of Promoter (%)
D
KS Test = n.s.
Exons per Gene
E
n.s.
Fraction of Genes w/ TATA Box
F
n.s.
n.s.
Fraction of Genes
Negative Strand
Postive Strand
G
KS test: p < 0.0002
Desnity
Fractional Distance within TAD

Figure 2.5: Noise-enhanced genes tend to be centrally located within topologically associating domains.
Comparisons are between 945 genes classified as highly variable and 3,513 genes classified as non-variable (background) according to BASiCS algorithm. Gene characteristics and sequences were taken from Ensembl GRCm38 reference genome. **(A)** Distributions of gene lengths for highly variable and background genes. Length was calculated as distance between Ensembl gene start and end coordinates which correspond to outermost transcript start and end coordinates. **(B)** Percentage of base-pairs in gene body (based on gene start and end coordinates) that are A:T. **(C)** Percentage of base-pairs in 200bp region upstream of gene start that are A:T. **(D)** The number of exons was averaged over all transcripts associated with a gene. Distributions of average exon quantity for genes in the highly variable and background group were then plotted. **(E)** Fraction of genes with TATA sequence in 200bp region upstream of gene start. Data represent mean and SD from bootstrapping procedure with 10,000 resamplings of 100 genes from each group with replacement. **(F)** Fraction of genes whose coding sequence is located on negative and positive strands. Data represent mean and SD from bootstrapping procedure with 10,000 resamplings of 100 genes from each group with replacement. **(G)** Fractional distance of gene within TAD was calculated as (gene start coordinate - TAD start coordinate)/(TAD end coordinate - TAD start coordinate).

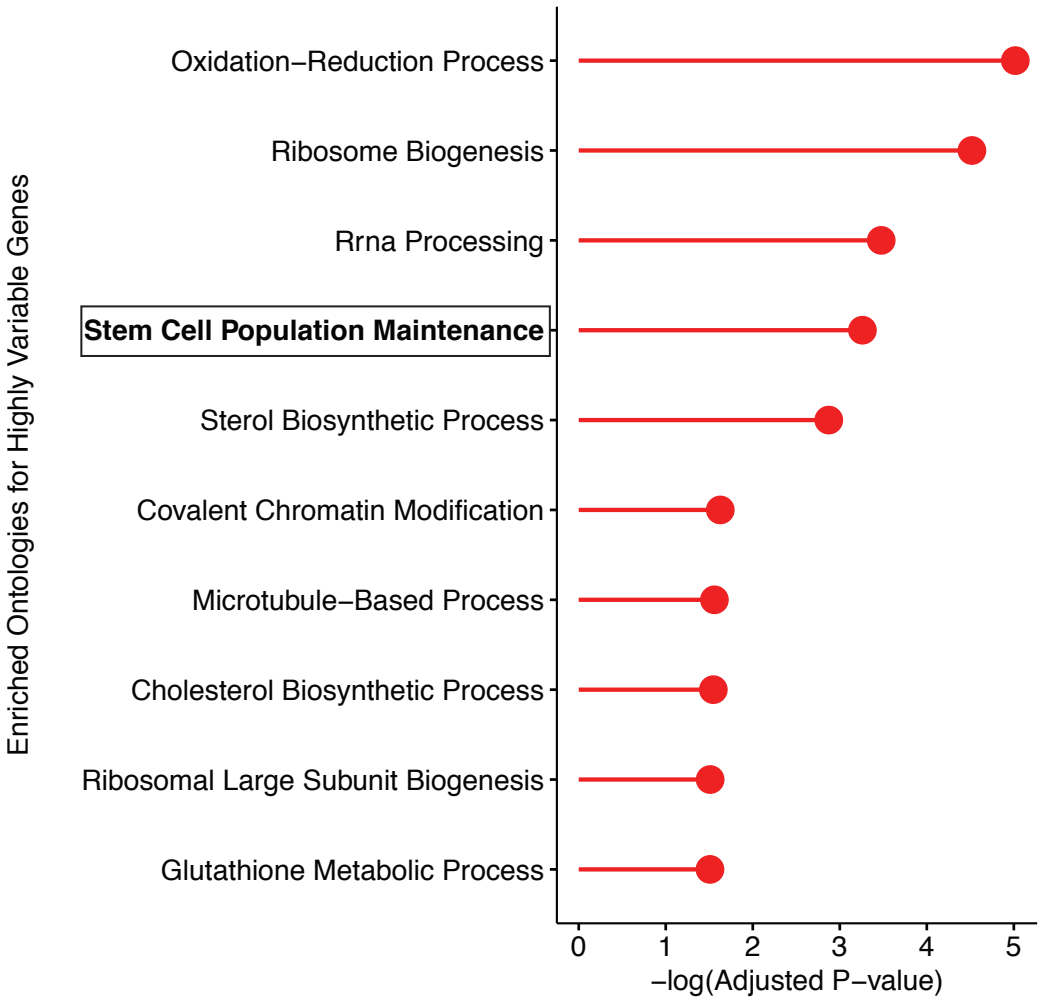


Figure 2.6: Ontology analysis of variably expressed genes shows enrichment for housekeeping and pluripotency maintenance pathways.
DAVID v6.8 was used to identify enriched ontologies among the 945 genes classified as highly variable according to BASiCS algorithm.

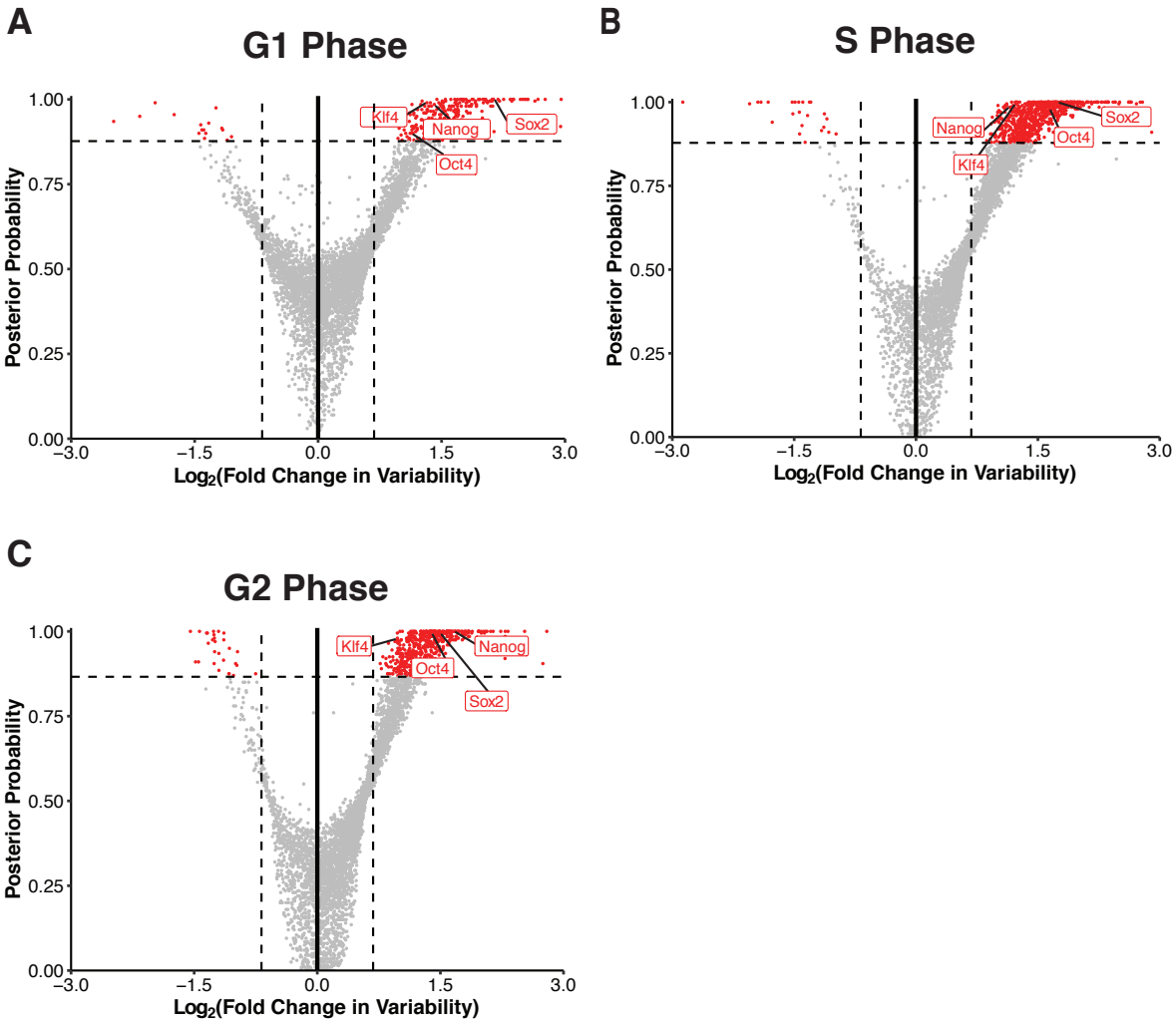


Figure 2.7: Noise-enhancement of pluripotency factors occurs in all three phases of the cell cycle.
A total of 1556 cells in the scRNA-seq dataset were classified into one of three cell-cycle phases. Differential variability testing was then conducted between cells in the DMSO and IdU treatment groups with the same cycle classification.

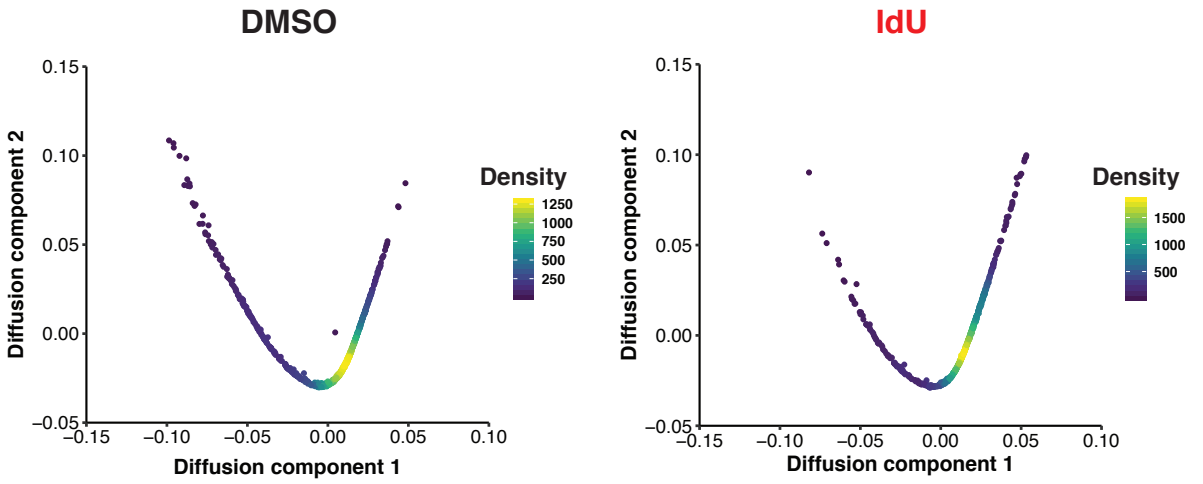


Figure 2.8: Transcript variability is not caused by bifurcation of mESCs into separate developmental lineages.
Pseudotime analysis of IdU-treated cells shows no differentiation of mESCs into separate developmental lineages.

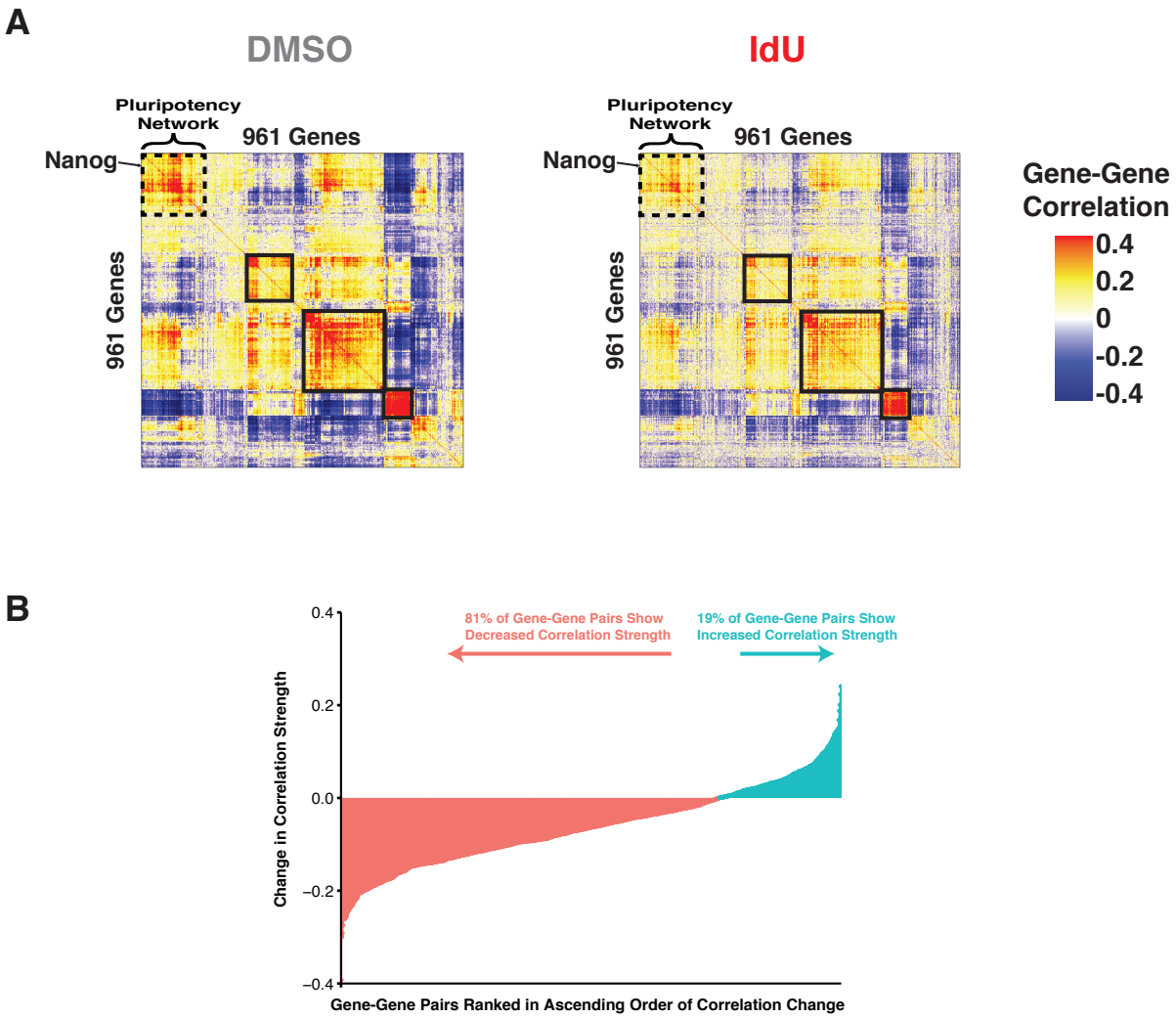


Figure 2.9: Majority of gene-gene pairs show a decrease in correlation strength.
(A) Pearson correlations of expression for gene pairs in scRNA-seq dataset. Hierarchical clustering reveals networks of genes (highlighted in black rectangles) sharing similar correlation patterns. Dashed rectangle highlights network enriched with pluripotency factors like Nanog. IdU treatment causes a fading of heatmap, indicating weakened expression correlations. **(B)** The Pearson correlation of expression for 923,521 (961 x 961) gene-gene pairs were compared between DMSO and IdU treatment groups. For each gene-gene pair, the absolute value of the correlation strength in DMSO was subtracted from the absolute value of the correlation strength in IdU. Negative values indicate loss of correlation in expression.

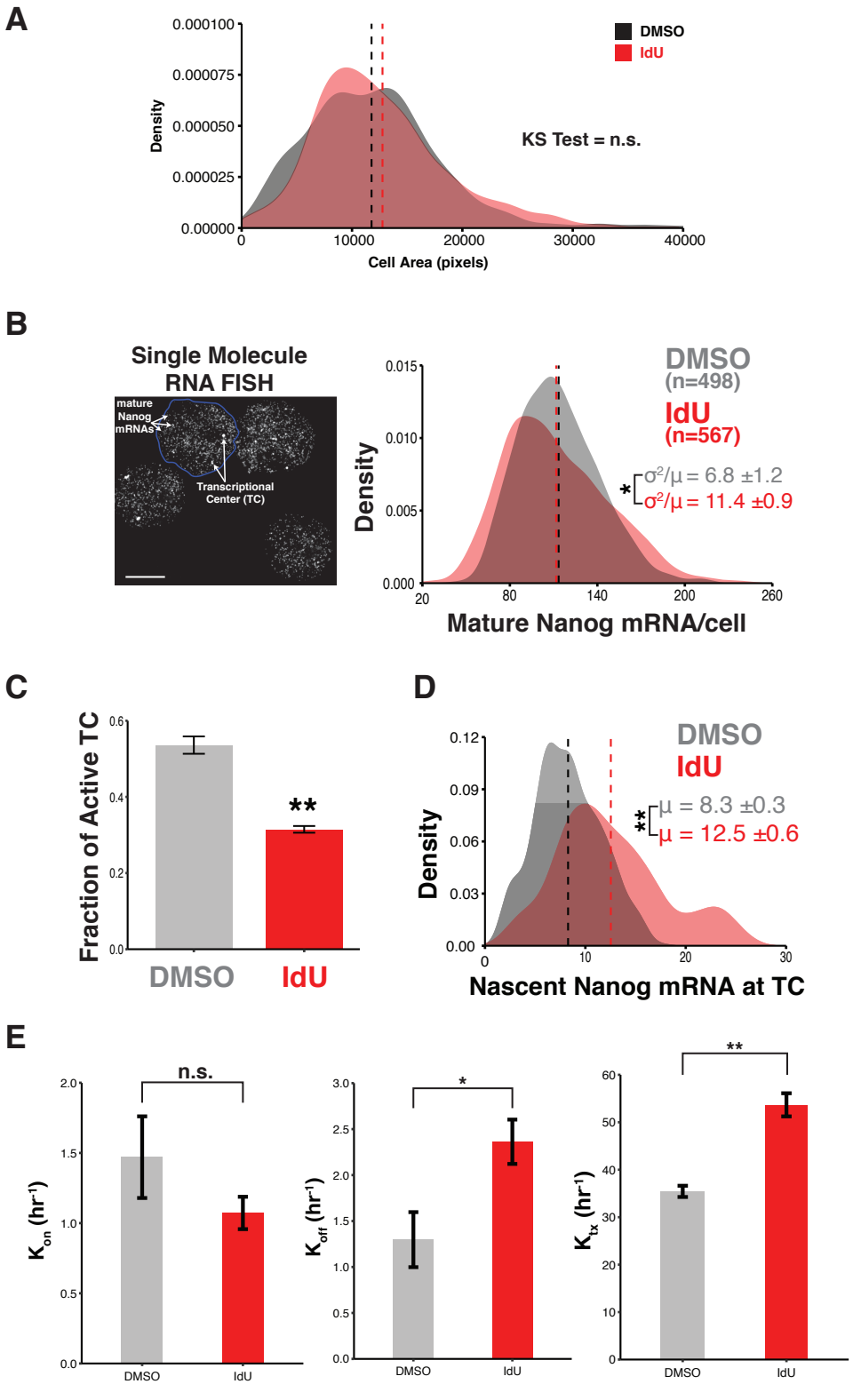

A
DMSO
IdU
KS Test = n.s.
Density
Cell Area (pixels)
B
Single Molecule RNA FISH
mature Nanog mRNAs
Transcriptional Center (TC)
DMSO (n=498)
IdU (n=567)
σ²/μ = 6.8 ±1.2
σ²/μ = 11.4 ±0.9
Density
Mature Nanog mRNA/cell
C
Fraction of Active TC
DMSO
IdU
D
DMSO
IdU
μ = 8.3 ±0.3
μ = 12.5 ±0.6
Density
Nascent Nanog mRNA at TC
E
n.s.
K_on (hr⁻¹)
K_off (hr⁻¹)
K_tx (hr⁻¹)
DMSO
IdU

Figure 2.10: Shortened burst duration and increased transcription rate causes enhanced cell-to-cell variability in Nanog mRNA counts.
(A) Distribution of cell sizes for analyzed cells in DMSO and IdU conditions. Cell size was calculated as number of pixels within segmented cell boundary. Dashed lines represent means of each distribution. Data represent pooling of cells from all four biological replicates for each condition. KS test shows no significant difference between cell size distributions. **(B-D)** Results of smRNA-FISH used to count nascent and mature Nanog mRNA in Nanog-GFP mESCs treated with DMSO or 10μM IdU for 24 hours in 2i/LIF media. Data are from four biological replicates. (B) (Left) Representative micrograph (maximum intensity projection) in which Nanog transcripts are labelled with probe-set for eGFP. Bright foci correspond to transcriptional centers as verified by intron probe set. Scale bar is 5um. (Right) Distributions of mature Nanog transcripts/cell. Dashed lines represent mean. IdU treatment increases cell-to-cell variability of transcript abundance as reported by averaged Fano factors (± SD), *p =0.0011 by a two-tailed, unpaired Student's t test. (C) Fraction of possible transcriptional centers that are active as detected by overlap of signal in exon and intron probe channels. Each cell is assumed to have 2 possible transcriptional centers (TCs). Data represent mean and SD. With IdU, the fraction of possible of TCs that are active decreases, **p =6.9 x 10^{-5} by a two-tailed, unpaired Student's t test. (D) Distributions of nascent Nanog mRNA per TC. With IdU, active TCs have more nascent mRNAs, **p =1.0 x 10^{-4} by a two-tailed, unpaired Student's t test. **(E)** Inference of parameters for 2-state model of transcription. P values were calculated using a two-tailed, unpaired Student's t test. *p = 0.0017, **p = 0.0001

Methods

Single Molecule RNA FISH

Probes for detection of nascent and mature Nanog transcripts were developed using the designer tool from Stellaris (LGC Biosearch Technologies). 30 probes (TAMRA conjugated) for mature Nanog mRNA were targeted towards the 3′ GFP segment of transcripts. 48 probes (Quasar 670 conjugated) for nascent Nanog mRNA were targeted towards the first intronic sequence as taken from the mm10 genome reference. Probes were designed using a masking level of 5, and at least 2 base pair spacing between single probes.

$1x10^5$ Nanog-GFP mESCs were seeded into each well of a gelatin-coated, 35mm Ibidi dish (quad-chambered, cat:80416) in 2i/LIF media. 24 hours following seeding, media was replaced with 2i/LIF containing 10μM IdU or equivalent volume DMSO. After 24 hours of treatment, cells were then fixed with DPBS in 4% paraformaldehyde for 10 minutes. Fixed cells were washed with DPBS and stored in 70% EtOH at 4°C for one hour to permeabilize the cell membranes. Probes were diluted 200-fold and allowed to hybridize at 37°C overnight. Wash steps and DAPI (Thermo) staining were performed as described (https://www.biosearchtech.com/support/resources/stellaris-protocols).

To minimize photo-bleaching, cells were imaged in a buffer containing 50% glycerol (Thermo), 75 μg/mL glucose oxidase (Sigma Aldrich), 520 μg/mL catalase (Sigma Aldrich), and 0.5 mg/mL Trolox (Sigma Aldrich). Images were taken on a Zeiss Axio Observer Z1 microscope equipped with a Yokogawa CSU-X1 spinning disk unit and 100x/1.4 oil objective. Approximately 20 xy locations were randomly selected for each condition. For each xy location, Nyquist sampling was performed by taking 30, 0.4μM steps along the z-plane.

Image analysis and spot counting was performed using FISH-quant [57]. Cells were manually

segmented and analysis was conducted on cells of a similar size to minimize extrinsic noise. Transcriptional centers (TCs) were identified by signal overlap in exon, intron and DAPI channels. The amount of nascent mRNA at TCs was quantified through a weighted superposition of point spread functions.

Rate calculations for random-telegraph model

From smRNA-FISH data for Nanog, the kinetic parameters of the random-telegraph model were inferred using the empirically derived values of mRNA mean (μ), mRNA Fano factor ($Fano$), transcriptional center frequency (f_{ON}) and transcriptional center size (TC_{mRNA}) [45]. The transcription rate (k_{tx}) is calculated as:

$$k_{tx} = TC_{mRNA}\frac{k_{elongation}}{L} \tag{2.1}$$

where $k_{elongation}$ is the elongation rate of RNAPII (1.9 kb/min) [58] and L is the length of the transcribed region of Nanog. The degradation rate (k_{decay}) is calculated as:

$$k_{decay} = \frac{f_{ON}\cdot k_{tx}}{\mu} \tag{2.2}$$

The rate of promoter activation (k_{ON}) is given by:

$$k_{ON} = k_{decay}\left(-\frac{\mu(f_{ON}-1)+f_{ON}(Fano-1)}{Fano-1}\right) \tag{2.3}$$

The rate of promoter inactivation (k_{OFF}) is given by:

$$k_{OFF} = -k_{decay}\left(-\frac{\mu(f_{ON}-1)+f_{ON}(Fano-1)}{Fano-1}\right)\cdot\left(\frac{1}{f_{ON}}-1\right) \tag{2.4}$$

2.2.4 Enhancement of transcriptional variability propagates to the protein level

To test if enhanced transcript variability transmitted to the protein level, we performed flow-cytometric analysis of Nanog-GFP reporter protein. In IdU-treated cells, the Nanog protein Fano factor increased by 3-fold, with little change in mean, indicating that mRNA variability from altered promoter toggling indeed resulted in changes to protein noise (Figure 2.11A). The increase in protein noise showed no dependency on cell-cycle (Figure 2.11D-E) despite G1-to-S cell-cycle progression being slightly slowed by IdU treatment (Figure 2.11B-C). Consistent with the extrinsic noise analysis above, there was no evidence of aneuploidy following IdU treatment (Figure 2.11B), precluding the possibility that increased noise results from a sub-population of cells with non-physiologic gene-copy numbers.

Given that Nanog noise was intrinsic and transmitted to the protein level, we next tested a previous theoretical prediction about Nanog. When cultured in 2i/LIF, mESCs exhibit Nanog protein expression that is unimodal and high, but when cultured in serum/LIF, mESCs exhibit bimodal Nanog expression with both a low Nanog state and a high Nanog state (Figure 2.12) [59]. Theories predicted that increased transcriptional noise would drive greater excursions from the high Nanog state into the low Nanog state [60]. We found that in mESCs cultured in serum/LIF, IdU-induced amplification of Nanog noise did indeed generate greater excursions into the low Nanog state (Figure 2.12), verifying theoretical predictions. This result demonstrates how promoter toggling can drive Nanog state-switching thereby altering differentiation potential.

To verify that enhanced noise is not a population-level phenomenon brought on by differential responses to IdU in distinct cellular subpopulations (i.e., verify 'ergodicity' and that individual cells exhibit increased fluctuations), we used live-cell time-lapse imaging to quantify both the magnitude (intrinsic-CV^2) and frequency content (1/half-autocorrelation time) of Nanog fluctuations. Single-cell tracking of individual cells showed that IdU induced a 2-fold increase in the magnitude (intrinsic-CV^2) of fluctuations (Figure 2.13A-B), and auto-correlation analysis of detrended trajectories showed a broadening of the frequency distribution to higher spectra, indicating reduced memory of protein state (Figure 2.13C). These higher frequency fluctuations are consistent with amplification of a non-genetic, intrinsic source of noise [61, 62] because genetic sources of cellular heterogeneity, such as promoter mutations, would lead to longer retention of protein states (increased memory) [63]. In silico sorting of cells based on starting Nanog expression verified that noise enhancement was not dependent on memory of initial state (Figure 2.14). Fluctuations in promoter toggling therefore drive individual cells to dynamically explore a larger state-space of Nanog expression. To further validate that IdU perturbs an intrinsic source of noise, we used a mESC line in which the two endogenous alleles of Sox2 are tagged with P2A-mClover and P2A-tdTomato, respectively, to enable quantification of the intrinsic and extrinsic components of noise. Treatment with IdU increased Sox2 intrinsic noise greater than 2-fold across all expression levels (Figure 2.15) further validating that IdU enhances intrinsic noise.

To test if a generalized stress response could explain the noise enhancement induced by IdU, we subjected Nanog-GFP mESCs to UV radiation for 15, 30, or 60 minutes. Both the mean and Fano factor of Nanog expression decreased for all timepoints, which markedly differs from IdU treatment (Figure 2.16). This result further indicates that IdU does not perturb an extrinsic or global noise source; rather, it perturbs an intrinsic source of noise (i.e., promoter toggling).

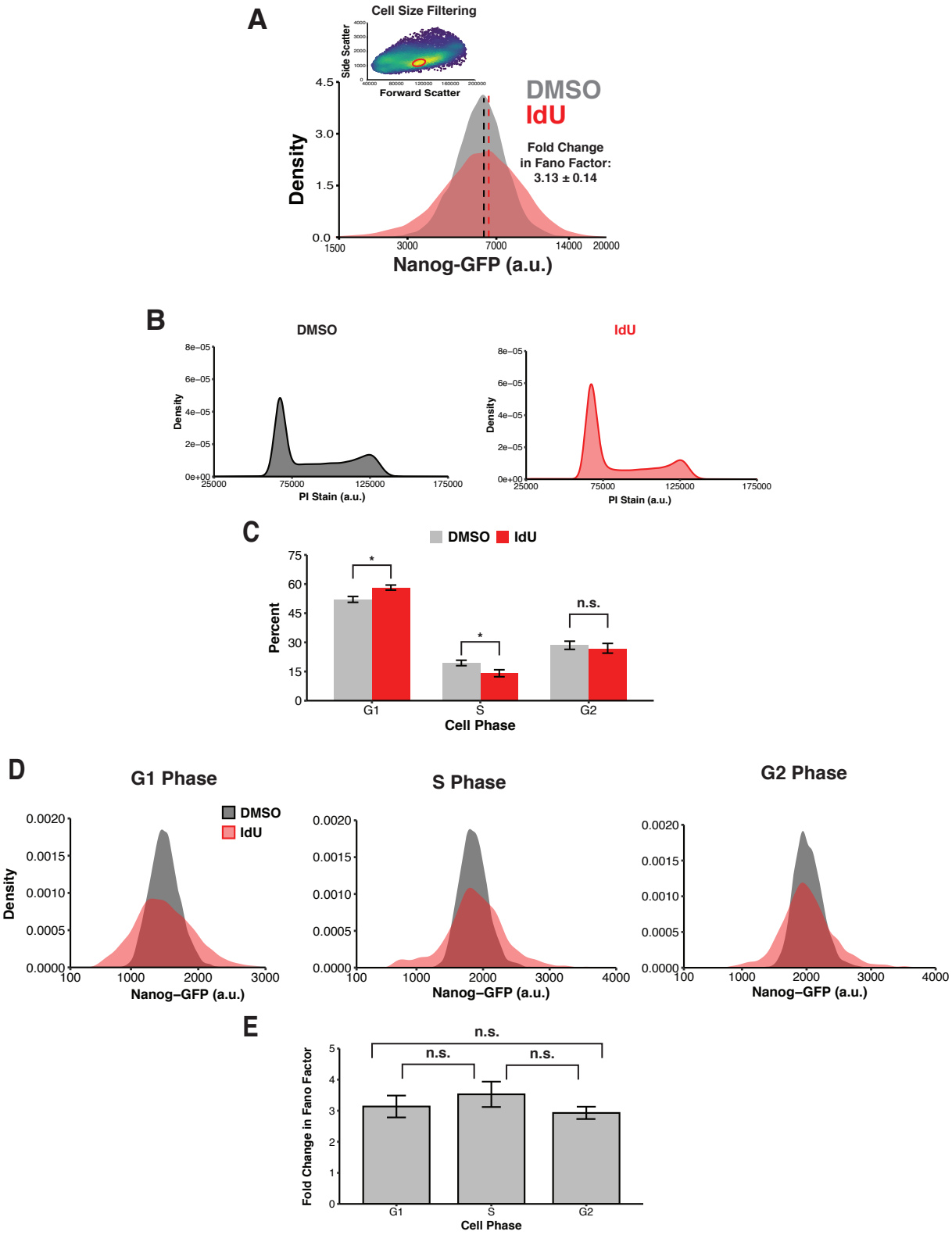

A
Cell Size Filtering
Side Scatter
Forward Scatter
DMSO
IdU
Fold Change
in Fano Factor:
3.13 ± 0.14
Density
Nanog-GFP (a.u.)
B
DMSO
IdU
Density
PI Stain (a.u.)
C
DMSO
IdU
Percent
n.s.
G1
S
G2
Cell Phase
D
G1 Phase
S Phase
G2 Phase
DMSO
IdU
Density
Nanog–GFP (a.u.)
E
n.s.
Fold Change in Fano Factor
Cell Phase

Figure 2.11: Noise-enhancement of Nanog protein expression is independent of cell-cycle state.
(A) Representative flow cytometry distribution of Nanog-GFP expression in mESCs treated with DMSO or 10μM IdU for 24h in 2i/LIF. Dashed lines represent mean. Fold change in Fano factor (± SD) obtained from three biological replicates. IdU increases cell-to-cell variability in Nanog protein expression. Inset: Representative flow cytometry dot-plot showing conservative gating on forward and side scatter to filter extrinsic noise arising from cell size heterogeneity. **(B)** Representative flow cytometry distributions of propidium iodide staining for Nanog-GFP mESCs treated with either DMSO or 10μM IdU for 24 hours. No signs of aneuploidy are visible, indicating transcriptional variability is not due to cell-to-cell variability in gene copy numbers. **(C)** Percent of cells in each phase of the cell cycle for DMSO and IdU treatments based on propidium iodide staining. IdU treatment slightly slows entry into S phase. Data represent mean and SD of three biological replicates. P values were calculated using a two-tailed, unpaired Student's t test. *p < 0.01 **(D)** Representative flow cytometry distributions of Nanog-GFP for mESCs within the G1, S and G2 phases of the cell cycle. mESCs were treated with 10μM IdU or equivalent volume DMSO for 24h followed by propidium iodide staining. **(E)** IdU-induced noise-enhancement of Nanog-GFP protein levels is unchanged across all three phases of the cell cycle. Nanog-GFP Fano factor with IdU treatment was normalized to DMSO control for calculation of fold change. Data represent mean and SD of three biological replicates. P values were calculated using a two-tailed, unpaired Student's t test.

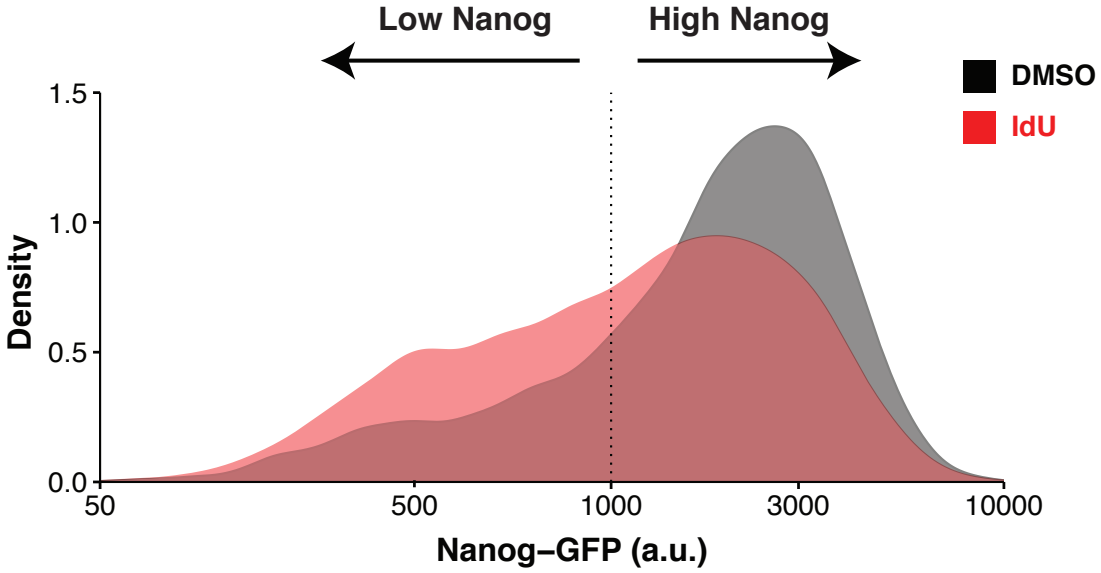


Figure 2.12: Increased transcriptional noise drives a greater number of mESCs into the low-Nanog state while cultured in serum/LIF.
Flow cytometry distribution of Nanog-GFP expression for mESCs cultured in serum/LIF and treated with 10μM IdU or equivalent volume DMSO for 24h. Data is pooled from three biological replicates.

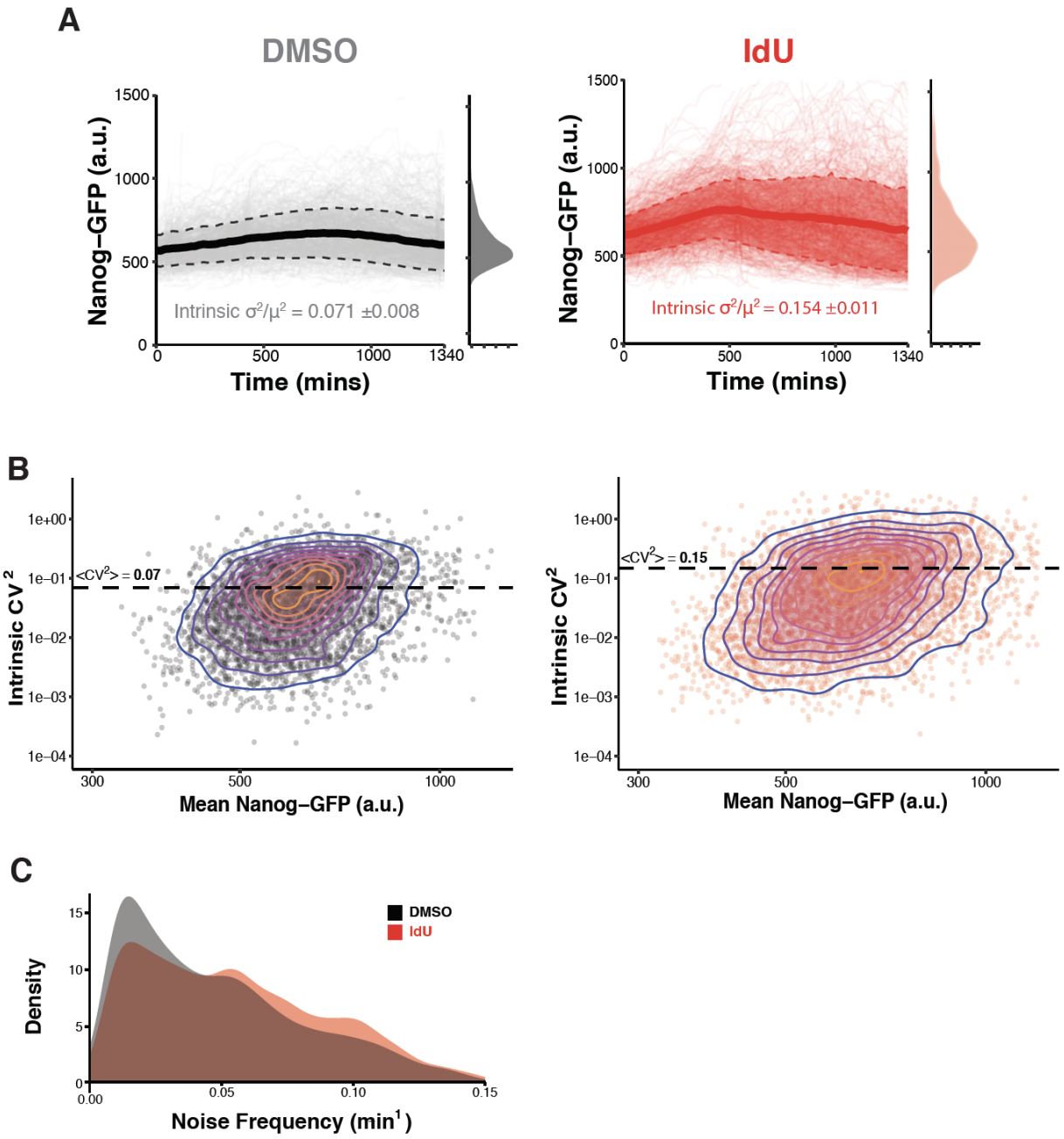

A
DMSO
IdU
Nanog–GFP (a.u.)
Time (mins)
Intrinsic σ²/μ² = 0.071 ±0.008
Intrinsic σ²/μ² = 0.154 ±0.011
B
Intrinsic CV²
Mean Nanog–GFP (a.u.)
<CV²> = 0.07
<CV²> = 0.15
C
Density
Noise Frequency (min¹)
DMSO
IdU

Figure 2.13: Time-lapse imaging demonstrates that altered kinetics of promoter toggling cause individual cells to experience larger fluctuations in Nanog protein expression.
(A) Time-lapse imaging of Nanog-GFP mESCs treated with either DMSO (n = 1513) or 10μM IdU (n = 1414) in 2i/LIF. Image acquisition began immediately after addition of compounds. Trajectories from two replicates of each condition are pooled, with solid and dashed lines representing mean and standard deviation of trajectories respectively. Distributions of Nanog-GFP represent expression at final time-point. Intrinsic-CV2 of each detrended trajectory was calculated, with the average (± SD) of all trajectories reported. **(B)** Each point represents a single-cell fluorescence trajectory (DMSO on left, n = 1513; IdU on right, n = 1414). Single-cell fluorescence trajectories were detrended by subtracting time-dependent population average for Nanog-GFP fluorescence. The mean Nanog-GFP fluorescence for each raw trajectory is then plotted versus the CV^2 of the detrended version of the trajectory to isolate intrinsic noise. The dashed lines represent the average intrinsic CV^2 of all trajectories for each treatment group. Time-lapse imaging shows that for individual cells the magnitude of Nanog protein fluctuations increases with IdU treatment. **(C)** Distributions of noise frequencies from autocorrelation functions of each detrended trajectory. Noise frequency is calculated as the inverse of the autocorrelation time ($\tau_{1/2}$). Shorter but more productive transcriptional bursts with IdU treatment pushes the frequency content of Nanog-GFP fluctuations to higher spectra.

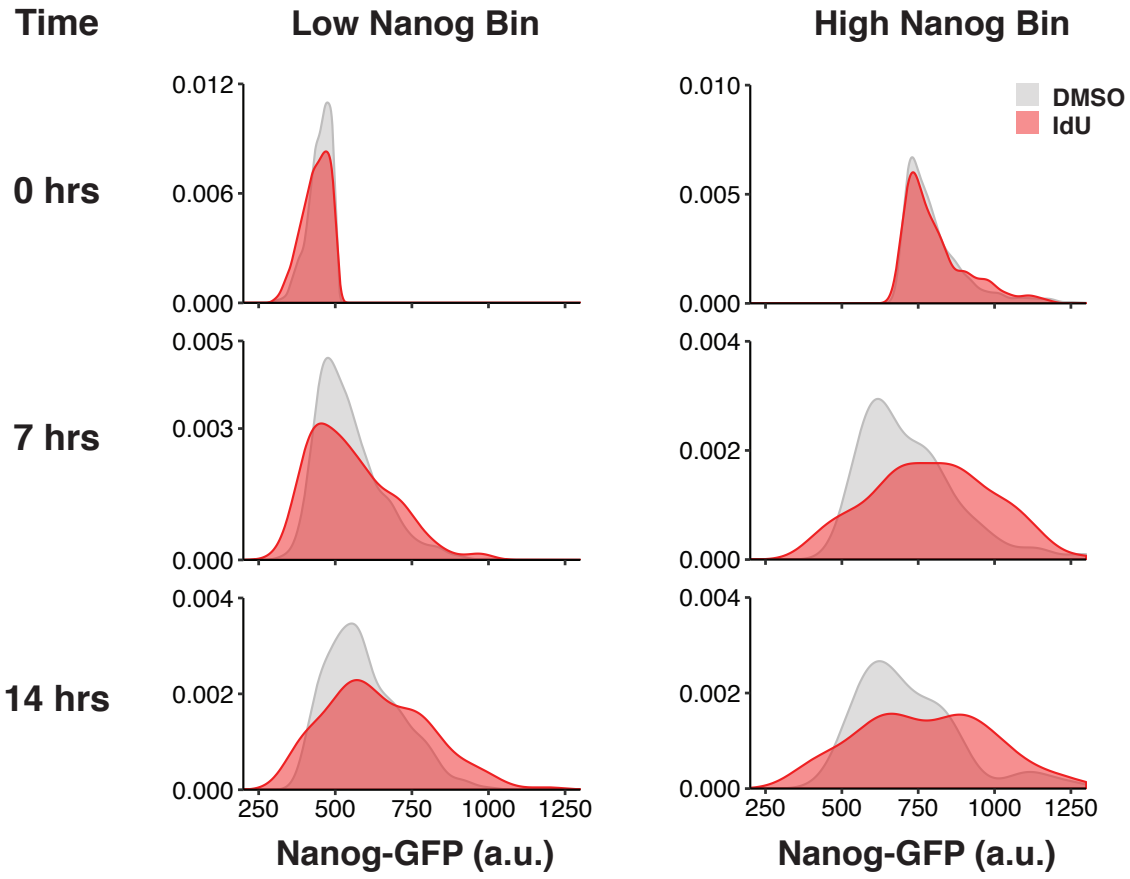


Figure 2.14: Amplification of expression fluctuations occurs independently of starting Nanog level.

Single-cell trajectories whose starting fluorescence value was below 500 a.u. or above 700 a.u. were binned into low and high groups respectively. Only trajectories whose starting point coincided with addition of DMSO or IdU at time zero were used. Distributions of trajectory fluorescence values at zero, seven, and 14 hours into treatment conditions are shown. By 14 hours into IdU treatment, there is visible interconversion of cells between the low and high Nanog states, indicating that memory of initial Nanog expression level is erased. This precludes the possibility that noise enhancement is due to promoter mutations that create sub-populations of cells with stable expression of Nanog at low and high levels.

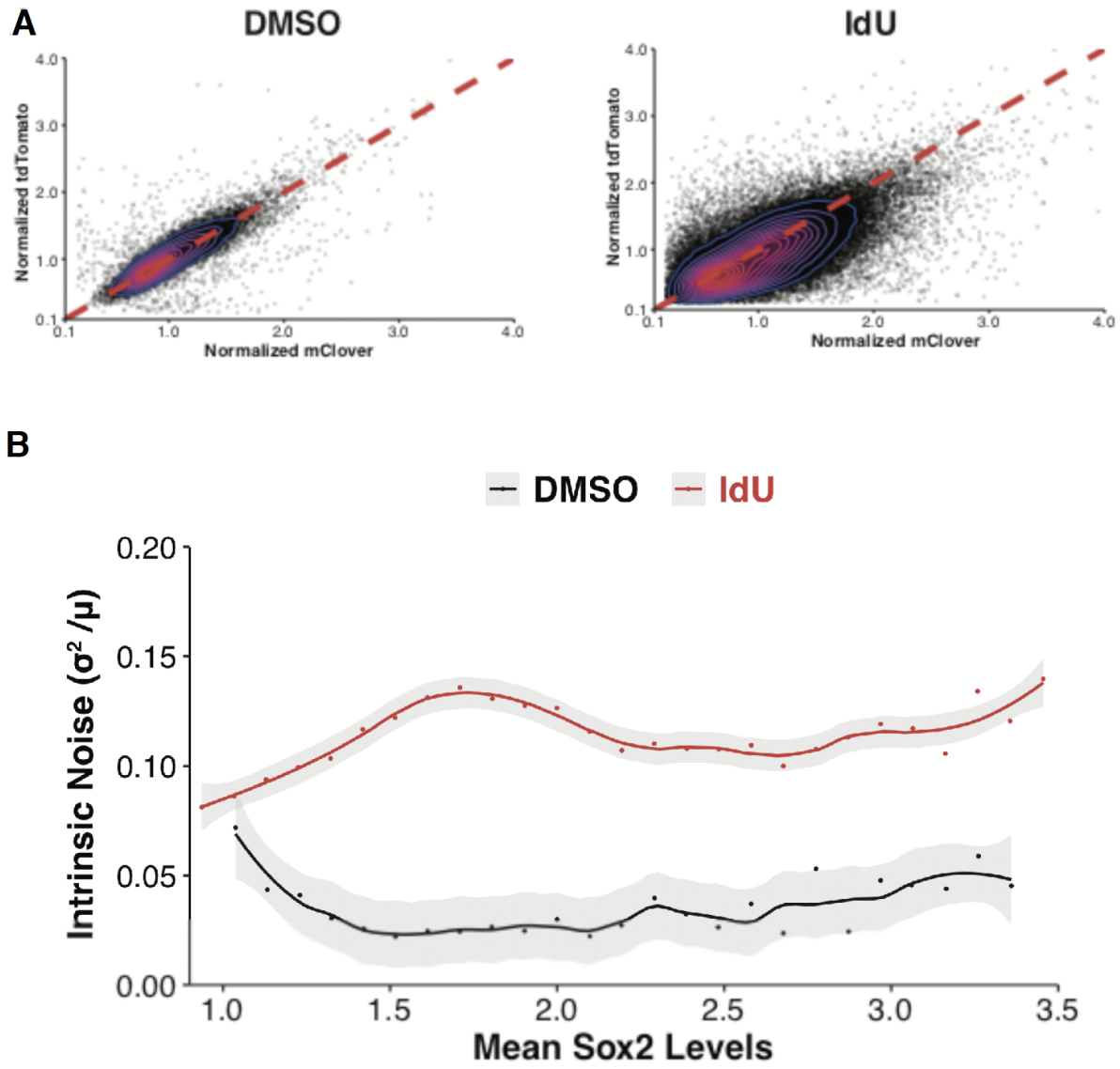
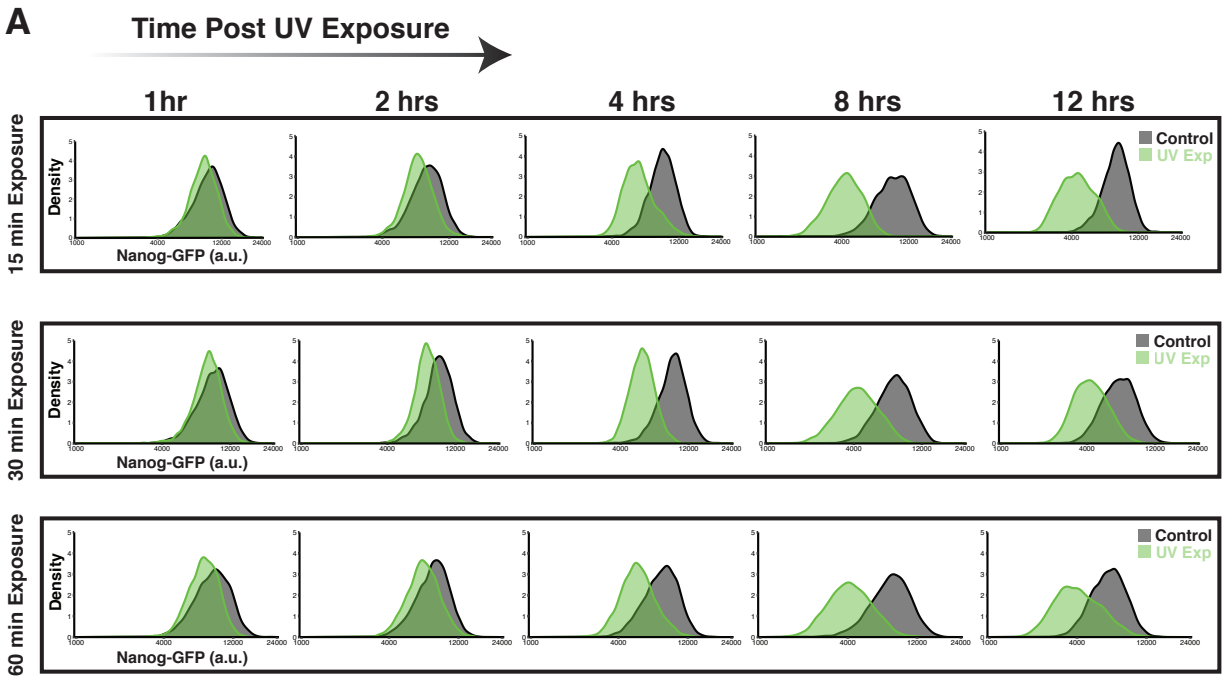


Figure 2.15: IdU treatment increases intrinsic noise of Sox2 expression.
(A) Flow cytometry dot-plot of mESCs with Sox2 dual color tags. Dashed red line has slope of one. mCLover and tdTomato fluorescence values were normalized to population average. Data shown is pooled from three biological replicates. **(B)** Cells were binned according to total Sox2 expression from both alleles. Each point represents the intrinsic noise (Fano factor) of Sox2 expression for cells within a particular bin. Grey shadings represent 95% confidence intervals as determined by bootstrapping. Smooth lines are produced from loess regression. IdU increases Sox2 intrinsic noise across all expression levels.


A
Time Post UV Exposure
1hr
2 hrs
4 hrs
8 hrs
12 hrs
15 min Exposure
30 min Exposure
60 min Exposure
Density
Nanog-GFP (a.u.)
Control
UV Exp

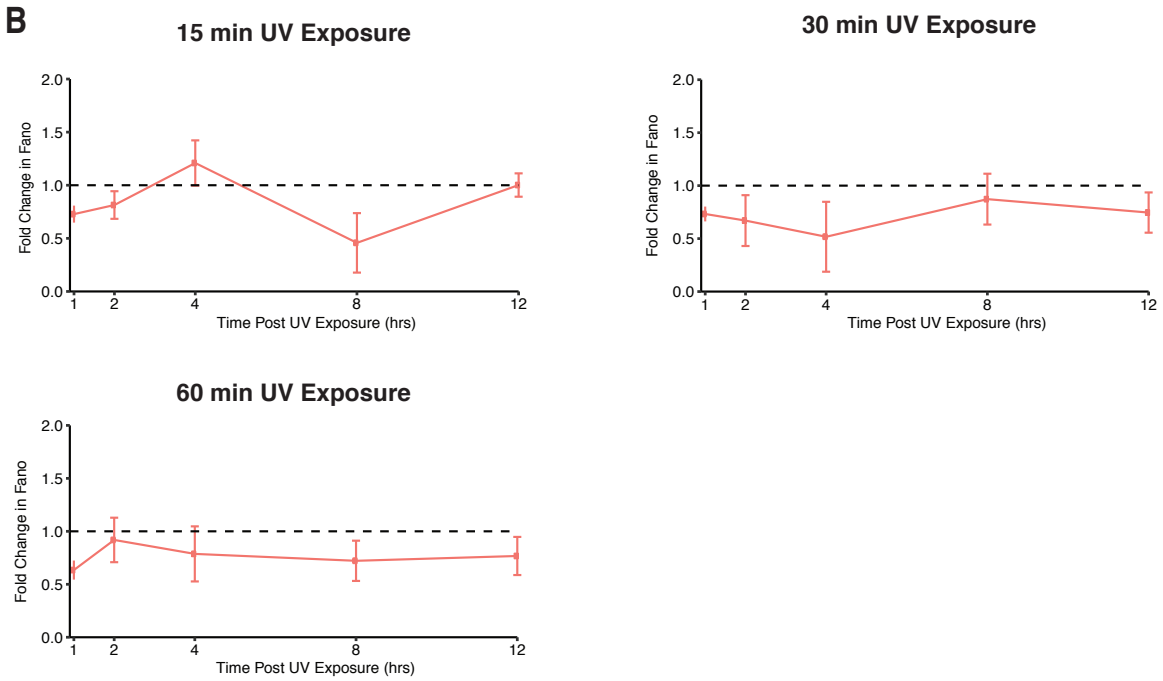

B
15 min UV Exposure
30 min UV Exposure
60 min UV Exposure
Fold Change in Fano
Time Post UV Exposure (hrs)

Figure 2.16: UV-stress reduces Nanog mean and Fano factor.
(A) Representative flow cytometry distributions of Nanog-GFP expression from UV-exposed (green) and control (grey) cell populations. Cells were analyzed one, two, four, eight, and 12 hours post exposure. **(B)** For each exposure group (15, 30, and 60 minutes), the fold change in Fano factor is calculated as the Fano factor for Nanog in the UV-exposed population normalized to the Fano factor of its respective control population. Data points represent mean and SD of two biological replicates. Across all time points (except 4 hour point in 15 minute exposure group) UV stress reduces the Fano factor of Nanog.

Methods

Extrinsic Noise Filtering on Flow Cytometry Data

All flow cytometry data were collected on BD FACSCalibur, LSRII or LSRFortessa X-20 with 488-nm laser used to detect GFP. For all measurements of Nanog-GFP mean and variability, >50k cells are collected per sample. Gating of cytometry data was performed with FlowJo. Prior to quantification of Nanog-GFP mean and variability, the smallest possible forward- and side-scatter region containing at least 3k cells was used to isolate cells of similar size and shape. This filters out gene expression variability arising from cell-size heterogeneity as previously established [30, 34, 45].

Cell-cycle Analysis by Propidium-Iodide Staining

$2x10^5$ Nanog-GFP mESCs were seeded in each well of a gelatin-coated, 6-well plate in 2i/LIF media. 24 hours following seeding, media was replaced with 2i/LIF media containing 10μM IdU or an equivalent volume of DMSO in triplicate for 24 hours. After treatment, cells were washed with DPBS, dissociated with TrypLE, pelleted, washed with DPBS, and resuspended in ice-cold 70% ethanol. Samples were stored overnight at -20°C and pelleted the following day at 200g for 5 minutes at 4°C. Cells were washed twice with DPBS supplemented with 0.5% BSA to prevent cell loss. Pellets were resuspended in 150μL of DPBS supplemented with 0.1mg/ml RNAse A (Thermo) and 30 μg/ml Propidium Iodide (Thermo). After overnight incubation at 4°C, cells were directly analyzed on a BD LSRII cytometer.

Noise Enhancer Testing in Serum/LIF culture

Serum/LIF media was prepared with 85% DMEM (supplemented with 2mM of L-glutamine), 15% FBS, 0.1mM 2-mercaptoethanol, and 1000U/ml of LIF (Sigma Aldrich). Nanog-GFP mESCs

grown feeder-free in 2i/LIF were passaged and seeded onto gelatin-coated 10cm dishes in serum/LIF media. Cells were passaged twice in serum/LIF media prior to noise enhancer testing. $4x10^5$ Nanog-GFP mESCs were seeded into each well of a gelatin-coated 6-well plate in serum/LIF media. 24 hours following seeding, media was replaced with serum/LIF supplemented with either 10μM IdU or equivalent volume DMSO in triplicate. After 24 hours of treatment, cells were run unfixed and live on BD LSRII flow cytometer.

Sox2 two-color reporter assay

The endogenous alleles of Sox2 are tagged with P2A-mClover and P2A-tdTomato. Both fluorophores have a PEST tag, thus shortening their half-lives to approximately 2.5 hours. $2x10^5$ Sox2-dual-tag mESCs were seeded in each well of a gelatin-coated, 6-well plate in 2i/LIF media. 24 hours following seeding, cultures were replenished with 2i/LIF media containing 10μM IdU or an equivalent volume of DMSO in triplicate for 24 hours. Cells were run unfixed and live on BD LSRII flow cytometer. Intrinsic noise was calculated as in Elowitz et. al [64]. Data from all three replicates were pooled together. No cell-size gating was performed as assay allows for separation of extrinsic noise. To align fluorescence values of mClover and tdTomato on the same scale, each cell's fluorescence intensity was normalized to the mean expression level of that fluorophore for the population. Since Sox2 expression spans several orders of magnitude, cells were binned according to their total Sox2 expression (mClover + tdTomato). Bins with fewer than 100 cells were discarded. Intrinsic noise (CV^2) of Sox2 expression for each bin was calculated using the following formula:

$$\eta^2_{intrinsic} = \frac{\langle (tdTomato_i - mClover_i)^2 \rangle}{2\langle tdTomato \rangle \langle mClover \rangle} \tag{2.5}$$

This value was then multiplied by the mean Sox2 expression for each bin to obtain the Fano

factor. Given that the number of cells in each bin differs and variance estimates are affected by sample size, we calculated 95% confidence intervals around the Fano factor for each bin through bootstrapping. Bin populations were resampled 10,000 times with replacement.

UV stress assay

$1x10^5$ Nanog-GFP mESCs were seeded in each well of gelatin-coated, 12-well plates in 2i/LIF media. 24 hours following seeding, cultures were exposed to 3kJ of 365nM light (Fotodyne UV Transilluminator 3-3000 with 15W bulbs) for 15, 30 or 60 minutes at room temperature in the dark. Control plates were left at room temperature in the dark for equivalent periods of time. Cells from UV-exposed and control plates were run unfixed and live on BD FACS Calibur cytometer 1,2,4,8, and 12 hours post-exposure in replicate. Extrinsic noise filtering via cell-size gating was performed prior to calculation of Nanog Fano factor.

Live-cell time-lapse microscopy

$1x10^5$ Nanog-GFP mESCs were seeded into each well of a gelatin-coated, 35mm Ibidi dish (quad-chambered, cat:80416) in 2i/LIF media. 24 hours following seeding, media was replenished with 2i/LIF containing 10μM IdU or equivalent volume DMSO in replicate. Time-lapse imaging commenced immediately after addition of compounds with IdU- and DMSO- treated cells imaged in the same experiment (neighboring wells). Imaging was performed on a Zeiss Axio Observer Z1 microscope equipped with Yokogawa CSU-X1 spinning disk unit and a Cool-SNAP HQ2 14-bit camera (PhotoMetrics). 488nM laser line (50% laser power, 500-ms excitation) was used for GFP imaging. Samples were kept in an enclosed stage that maintained humidified conditions at 37°C and 5% CO_2. Images were captured every 20 minutes for 24 hours. For each xy location, three z-planes were sampled at 4-μm intervals. The objective used was 40x oil, 1.3 N.A.

Cell segmentation, tracking and GFP quantification were carried out using CellProfiler [65]. Tracking of cells was manually verified. Segmented cells tracked for less than 4 hours were discarded. Cell division triggered the start of 2 new trajectories. After illumination correction and background subtraction, the mean GFP fluorescence intensity of a segmented cell was taken from each z-plane and averaged over the entire z-stack. For each trajectory, noise autocorrelation ($\tau_{1/2}$) and noise magnitude (intrinsic-CV^2) were calculated as previously described [61]. Fluorescence trajectories were first detrended (normalized) by subtracting the population time-dependent average fluorescence to isolate intrinsic noise. Distributions of noise frequency ranges (F_N) were extracted from normalized autocorrelation functions (ACFs) of individual trajectories, where $F_N = \frac{1}{\tau_{1/2}}.\tau_{1/2}$ is the value of τ (lag time) where the normalized ACF reaches a value of 0.5.

Chapter 3

DNA repair protein, Apex1, homeostatically enhances transcriptional noise via increased DNA supercoiling

3.1 Results

3.1.1 5'-bromo-2'-deoxyuridine (BrdU), 5-hydroxymethylcytosine (hmC), and 5-hydroxymethyluridine (hmU) increase Nanog expression noise

To pinpoint the molecular mechanism, 14 additional nucleoside analogs were screened for noise enhancement effects. 5'-bromo-2'-deoxyuridine (BrdU), 5-hydroxymethylcytosine (hmC), and 5-hydroxymethyluridine (hmU) also increased Nanog Fano factor to varying degrees (Figure 3.1A). Intriguingly, hmU and hmC are naturally produced by the Ten Eleven Translocation (Tet) family of enzymes during oxidation of thymine and methylated cytosine respectively [66, 67]. Given that these base modifications are removed via base-excision repair (BER), we surmised that their incorporation and removal from genomic DNA may be responsible for noise enhancement (Figure 3.1B) [68, 69].

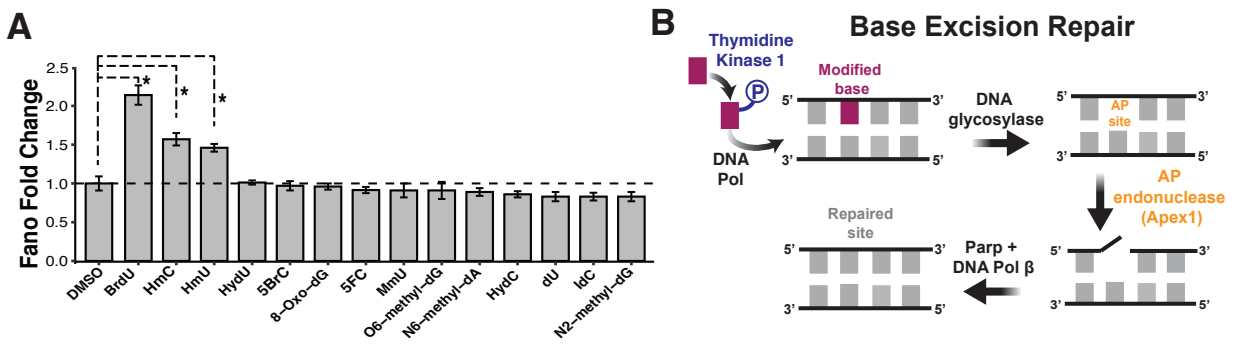


Figure 3.1: Screening of additional nucleoside analogs identifies naturally occurring base modifications that increase gene expression noise.

(A) Screening of 14 additional nucleoside analogs. Nanog-GFP mESCs grown in 2i/LIF were supplemented with 10μM of nucleoside analog for 24h. Fano factor for Nanog protein expression was normalized to DMSO. Data represent mean (± SD) of two biological replicates. BrdU, hmC and hmU increase Nanog expression variability as compared to DMSO, *p ¡ 0.01 by a Kruskal-Wallis test followed by Tukey's multiple comparison test. **(B)** Schematic of nucleoside analog incorporation into genomic DNA and removal via base excision repair pathway.

Methods

Nucleoside analog screening

14 nucleoside analogs (compound names and sources listed in Table S3) were resuspended in DMSO. 1x10^5 Nanog-GFP mESCs were seeded in gelatin-coated 12-well plates in 2i/LIF media. 24 hours after seeding, media was swapped with 2i/LIF containing 10μM of nucleoside analog or equivalent volume DMSO in replicate. After 24 hours of treatment, cells were run unfixed and live on BD LSRII cytometer. Extrinsic noise filtering via cell-size gating was performed prior to calculation of Nanog Fano factor. Fano factor for Nanog-GFP expression for each treatment was normalized to DMSO control.

3.1.2 CRISPRi knockdown of Apex1 and Tk1 ablates noise-enhancement from IdU

To test this, we knocked down 25 genes (3 gRNAs/gene) involved in nucleoside metabolism and DNA repair using CRISPRi, and quantified how these knockdowns affected IdU's noise enhancement. We identified 2 genes: AP Endonuclease 1 (Apex1) and thymidine kinase 1 (Tk1) whose knockdown abrogated noise enhancement (Figure 3.2A). Knockdown was confirmed via RT-qPCR (Figure 3.2B).

Within the base excision repair pathway, Uracil DNA glycosylase (Udg) and Thymine DNA glycosylase (Tdg) have overlapping roles in removal of modified and mismatched bases [70]. Although, knockdown of each enzyme individually failed to ablate Nanog noise-enhancement, it is possible that Udg and Tdg compensate for each other. To account for this possibility, dual gRNA expression cassettes were used to simultaneously express 9 combinations (3x3) of gRNAs for Udg and Tdg in CRISPRi Nanog-GFP mESCs. Simultaneous knockdown of Tdg and Udg in mESCs

failed to ablate the noise-enhancing effects of IdU (Figure 3.2C).

Tk1 adds a requisite gamma-phosphate group to diphosphate nucleotides prior to genomic incorporation (Figure 3.1B) [71]. Our knockdown results indicated that phosphorylation of IdU by Tk1 and subsequent incorporation of phosphorylated IdU into the genome may be necessary for noise enhancement. To validate this, we tested the effect of treatment with 10μM IdU combined with excess thymidine, a competitive substrate of Tk1. Competitive inclusion of thymidine returned Nanog noise to baseline levels (Figure 3.3), indicating that noise enhancement is dependent on IdU incorporation. The reduction in Nanog noise with addition of exogenous thymidine also suggests that IdU-induced noise amplification is not a generic effect of nucleotide imbalances within the cell.

Apex1 (a.k.a., Ref-1, Ape1) plays a pivotal role in the BER pathway as it incises DNA at apurinic/apyrimidinic sites via an endonuclease domain, allowing for subsequent removal of the sugar backbone and patching of the gap [70, 72]. To confirm the knockdown results, we attempted to knockout Apex1 in mESCs. However, the knockout was lethal, in agreement with previous reports [73]. As an alternative, we used a small-molecule inhibitor (CRT0044876) specific for the Apex1 endonuclease domain [74]. Unexpectedly, the combination of CRT0044876 with IdU synergistically increased Nanog protein variability, without significantly changing the mean (Figure 3.4).

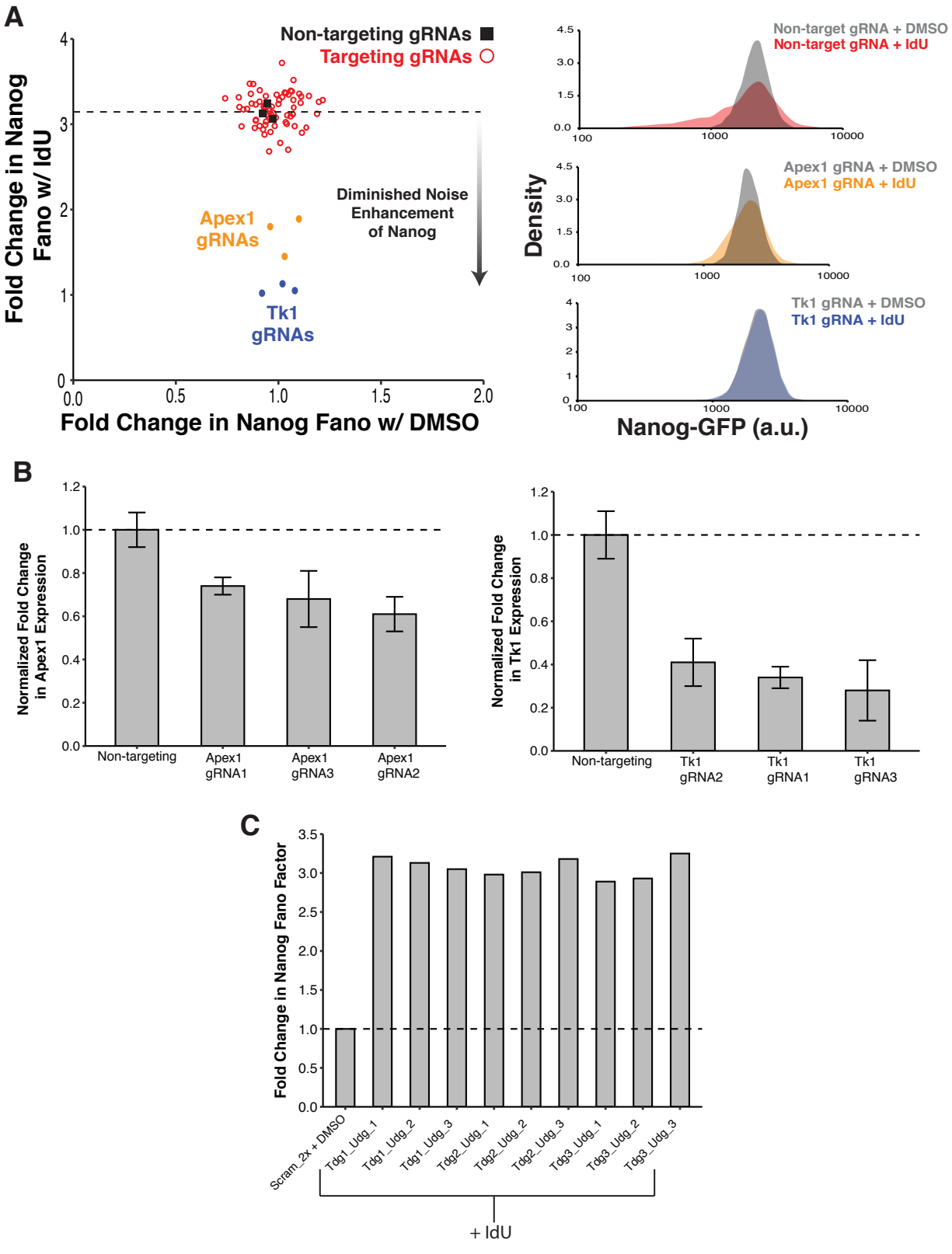

A
Non-targeting gRNAs
Targeting gRNAs
Fold Change in Nanog Fano w/ IdU
Fold Change in Nanog Fano w/ DMSO
Diminished Noise Enhancement of Nanog
Apex1 gRNAs
Tk1 gRNAs
Non-target gRNA + DMSO
Non-target gRNA + IdU
Apex1 gRNA + DMSO
Apex1 gRNA + IdU
Tk1 gRNA + DMSO
Tk1 gRNA + IdU
Density
Nanog-GFP (a.u.)
B
Normalized Fold Change in Apex1 Expression
Non-targeting
Apex1 gRNA1
Apex1 gRNA3
Apex1 gRNA2
Normalized Fold Change in Tk1 Expression
Non-targeting
Tk1 gRNA2
Tk1 gRNA1
Tk1 gRNA3
C
Fold Change in Nanog Fano Factor
Scram_2x + DMSO
Tdg1_Udg_1
Tdg1_Udg_2
Tdg1_Udg_3
Tdg2_Udg_1
Tdg2_Udg_2
Tdg2_Udg_3
Tdg3_Udg_1
Tdg3_Udg_2
Tdg3_Udg_3
+ IdU

Figure 3.2: Noise amplification independent of mean is dependent on Apex1 and Tk1.
(A) (Left) CRISPRi screening for genetic dependencies of IdU noise enhancement. Nanog-GFP mESCs stably expressing dCas9-KRAB-p2A-mCherry were transduced with a single gRNA expression vector with BFP reporter. 75 gRNAs (25 genes, 3 gRNAs/gene) were tested in addition to 3 non-targeting control gRNAs. Two days following transduction, each gRNA-expressing population of mESCs was treated with DMSO or 10μM IdU for 24h in 2i/LIF media. Nanog-GFP protein expression was measured for mCherry/BFP double positive cells. Nanog Fano factor for DMSO and IdU treatment of each gRNA population was normalized to Nanog Fano factor of non-targeting gRNA+DMSO population. Each point represents a gRNA. Dashed horizontal line represents average noise enhancement of Nanog from IdU in the background of non-targeting gRNA expression (black squares). Knockdown of Apex1 and Tk1 diminishes noise enhancement of Nanog from IdU. (Right) Representative flow cytometry distributions of Nanog expression for mESCs expressing non-targeting (top-right), Apex1 (middle-right), or Tk1 (bottom-right) gRNAs and treated with DMSO or 10μM IdU. **(B)** Validation of CRISPRi knockdown of Apex1 and Tk1 via qPCR measurements. $\Delta\Delta C_t$ method was used with the empty-vector cell population as the control. Levels of Apex1 and Tk1 repression are relative to the non-targeting (scrambled) population. Data represent mean and SD of two biological replicates. **(C)** Simultaneous knockdown of Udg and Tdg fails to ablate Nanog noise enhancement with IdU treatment. Two days following transduction with dual gRNA lentiviral vectors, each gRNA-expressing population of mESCs was treated with DMSO or 10μM IdU for 24h in 2i/LIF media. Nanog-GFP protein expression was measured for mCherry/BFP double positive cells. Nanog Fano factor for IdU treatment of each gRNA population was normalized to Nanog Fano factor of non-targeting gRNA (2x scram)+DMSO population.

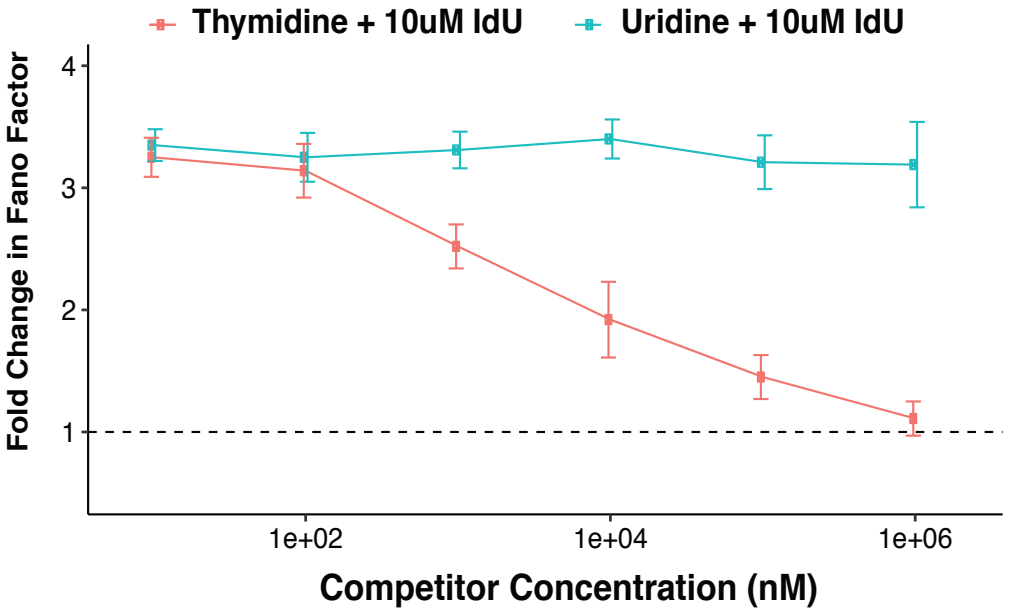


Figure 3.3: Thymidine competition ablates Nanog noise-enhancement from IdU.
The Fano factor of Nanog for each concentration combination is normalized to DMSO control. For all treatment combinations, IdU concentration is held constant at 10μM. Concentration of thymidine (red) and uridine (blue) is reported on the x-axis. Combination of 100μM thymidine and 10μM IdU returns Nanog Fano factor to baseline level (DMSO control). Uridine, which is not a substrate of Tk1, fails to ablate IdU-induced noise-enhancement. Data points represent mean and SD of three biological replicates.

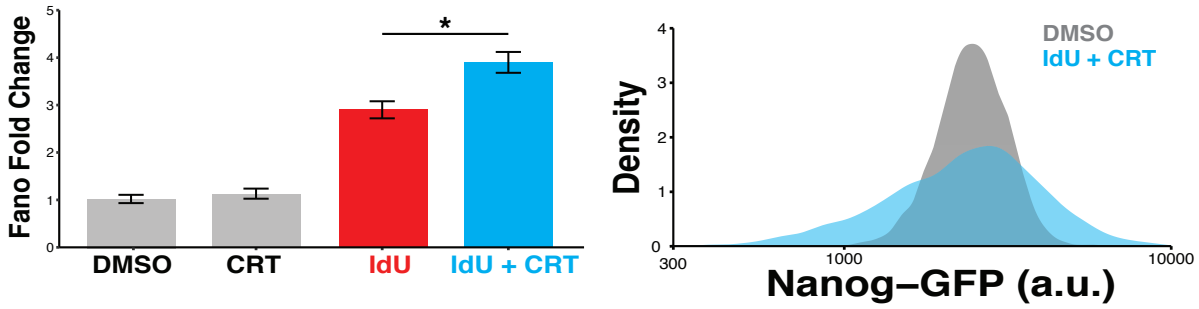


Figure 3.4: Small-molecule inhibition of Apex1 endonuclease domain synergistically increases cell-to-cell variability when combined with IdU.
(Left) mESCs were treated with DMSO, 100μM CRT0044876, 10μM IdU or 10μM IdU + 100μM CRT0044876 for 24h in 2i/LIF. Nanog Fano factor for each treatment was normalized to DMSO control. Data represent mean (± SD) of three biological replicates. Inhibition of Apex1 endonuclease domain in combination with IdU synergistically increases cell-to-cell variability of Nanog expression, *p = 0.0028 by a two-tailed, unpaired Student's t test. (Right) Representative flow cytometry distributions of Nanog expression for mESCs treated with DMSO or 10μM IdU + 100μM CRT0044876.

Methods

Generation of stable CRISPRi Nanog-GFP mESC line

To stably integrate the CRISPRi machinery into the ROSA26 locus of Nanog-GFP mESCs, AAVS1 homology arms of the CRISPRi knockin construct (krab-dCas9-p2a-mCherry, Addgene:73497) were swapped with ROSA26 homology arms. The dox-inducible promoter of this construct was replaced with a constitutive CAGGS promoter and the kanamycin resistance cassette was replaced with puromycin resistance. Two million Nanog-GFP mESCs were nucleofected with the CRISPRi knockin construct and left to recover for 48 hours. Puromycin (1µg/ml) selection was run until single colonies could be picked. Clonal CRISPRi Nanog-GFP mESC lines were assessed for mCherry expression and ability to knockdown Nanog. We selected the clone with the highest percentage of mCherry-positive cells.

CRISPRi gRNA design and cloning

gRNA sequences were were taken from the mCRISPRi-v2 library [75]. gRNA oligos were annealed and cloned into the pU6-sgRNA EF1Alpha-puro-T2A-BFP lentiviral vector (Addgene:60955) using the BstXI/BlpI ligation strategy [75].

CRISPRi screening for genetic dependencies of noise enhancer

25 genes involved in nucleotide metabolism, DNA repair, and chromatin remodeling were screened for their potential role in noise enhancement from IdU. Three gRNAs were designed per gene. Three non-targeting controls (scrambled gRNAs) were taken from the mCRISPRi-v2 library [75]. Each gRNA expression plasmid was separately packaged into lentivirus in HEK293T cells as previously described [75]. For each gRNA lentivirus, 1.5×10^5 CRISPRi Nanog-GFP mESCs were spinoculated with filtered viral supernatant for 90 minutes at 200 x g in replicate. Following

spinoculation, infected cells were seeded into gelatin-coated, 6-well plates in 2i/LIF media. 48 hours following seeding, media was swapped with 2i/LIF supplemented with either 10μM IdU or equivalent volume DMSO. Consequently, for every knockdown there is a DMSO and IdU treatment group. After 24 hours of treatment, cells were run unfixed and live on a BD LSRII flow cytometer. To minimize technical variability, analysis was restricted to cells with homogeneous levels of dCas9-KRAB and gRNA expression through stringent gating on mCherry/BFP double-positive cells. Extrinsic noise filtering through cell-size gating was then applied. For each gRNA, Nanog Fano factor for the DMSO and IdU treatments were normalized to the Nanog Fano factor of the non-targeting controls treated with DMSO.

qPCR verification of CRISPRi knockdown

To verify CRISPRi knockdown of Apex1 and Tk1, each of the six gRNA-expression plasmids targeting these two genes along with a non-targeting control and empty vector were packaged into lentivirus. $1.5\text{x}10^5$ CRISPRi Nanog-GFP mESCs were spinoculated with filtered viral supernatant for 90 minutes at 200 x g in replicate. Following spinoculation, infected cells were seeded into gelatin-coated, 6-well plates in 2i/LIF media. 72 hours following seeding, $1\text{x}10^6$ mCherry/BFP double-positive cells from each infected cell population were sorted on a FACSAria II. Total RNA was extracted using an RNeasy Mini Kit (QIAGEN cat:74104) and reverse-transcribed using a QuantiTect Reverse Transcription Kit (QIAGEN cat:205311). cDNA from each independent biological replicate was plated in triplicate and run on a 7900HT Fast Real-Time PCR System (Thermo) using designed primers and Fast SYBR Green Master Mix (Applied Biosystems, cat:4385612). Expression of GAPDH was used for normalization. Relative mRNA levels of Apex1 and Tk1 were calculated by the $\Delta\Delta C_t$ method using the empty-vector populations as the control. All reported levels of repression are relative to the non-targeting control.

Tk1 competition assay

$1x10^5$ Nanog-GFP mESCs were seeded in each well of gelatin-coated, 12-well plates in 2i/LIF media. 24 hours following seeding, media was replaced with 2i/LIF supplemented with 10μM IdU in combination with thymidine (Sigma cat:T1895) or uridine (Sigma cat:U3003) at concentrations ranging from 0 to 100μM. Concentration combinations were done in triplicate. After 24 hours of treatment, cells were run unfixed and live on BD FACS Calibur cytometer. Extrinsic noise filtering via cell-size gating was performed prior to calculation of Nanog Fano factor.

3.1.3 Homeostatic noise-amplification by Apex1 is mediated by DNA supercoiling

The contrasting effects of Apex1 knockdown and catalytic inhibition implied that a physical rather than enzymatic quality of the protein is responsible for modulation of transcriptional bursting. In support of this, Apex1 induces helical distortions and local supercoiling to identify mismatched bases [76]. Furthermore, catalytically inactive Apex1 binds DNA with higher affinity [77]. Therefore, CRT0044876 may lengthen Apex1 residence times on DNA, thus synergistically amplifying topological reformations. Taken together with evidence that supercoiling sets mechanical bounds on transcriptional bursting [23, 78, 38], we next asked whether Apex1 recruitment impacts supercoiling levels.

To assay supercoiling, we used a psoralen-crosslinking assay in which mESCs are incubated with biotinylated-trimethylpsoralen (bTMP), which preferentially intercalates into negatively supercoiled DNA [79, 80]. To eliminate DNA replication as a contributor of supercoiling, aphidicolin is added to inhibit DNA polymerases prior to bTMP incubation [81]. IdU treatment significantly increased genomic supercoiling as demonstrated by a ∼2-fold increase in bTMP intercalation (Figure 3.3A). The combination of IdU and CRT0044876 further increased intercalation, suggesting

that supercoiling levels are correlated with noise enhancement through increased Apex1-DNA interactions (Figure 3.5A). IdU treatment followed by a short incubation with bleomycin (which decreases supercoiling through double-stranded breaks) reduced bTMP intercalation below the DMSO control level, indicating IdU alone in uncoiled DNA does not increase intercalation (Figure 3.5B).

If DNA topology influences transcriptional bursting, additional modifiers of supercoiling should also affect Nanog noise. Topoisomerase 1 and 2a (Top1 and Top2a, respectively) relax coiled DNA through the introduction of single- and double-stranded breaks, respectively. Knockdown of Top1 and Top2a via CRISPRi, increased Nanog protein variability (Figure 3.6A). Furthermore, inhibition of topoisomerase activity with the small-molecule inhibitors topotecan and etoposide recapitulated these effects (Figure 3.6B).

To further test whether increased DNA supercoiling is responsible for noise-enhancement, we hypothesized that overexpression of topoisomerases should help alleviate Apex1-induced supercoiling and thus reduce noise-enhancement from IdU. To test this hypothesis, a Topoisomerase 1 expression vector under a weak CMV promoter was lentivirally integrated into Nanog-GFP mESCs. Overexpression of Topoisomerase 1 reduced IdU-mediated noise-enhancement of Nanog by 31% as compared to the wildtype population of Nanog-GFP mESCs treated with IdU (Figure 3.7). Taken together with psoralen-crosslinking and topoisomerase inhibition data, these results suggest that Apex1-induced supercoiling tunes gene-expression fluctuations without altering mean expression levels.

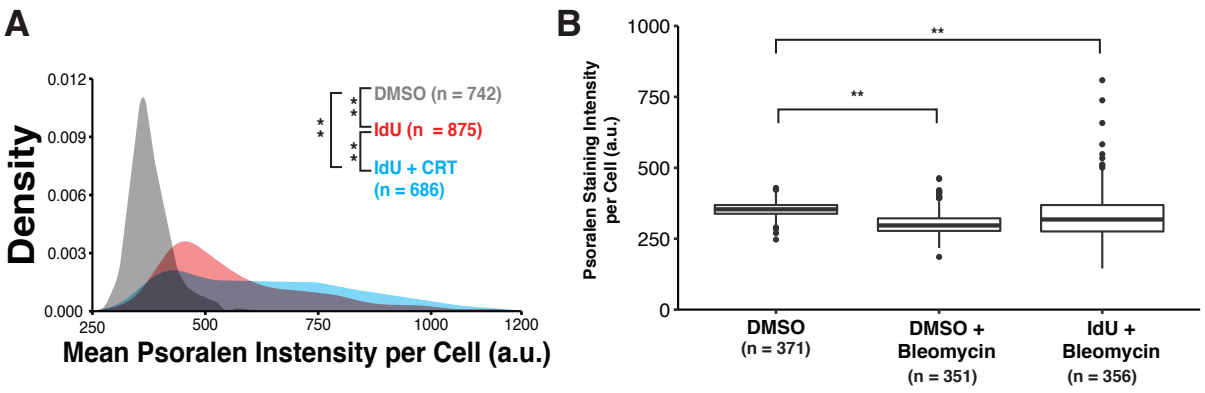


Figure 3.5: Apex1 recruitment to DNA increases negative-supercoiling levels.
(A) Single-cell quantification of negative supercoiling levels using psoralen-crosslinking assay. mESCs were treated with DMSO, 10μM IdU or 10μM + 100μM CRT0044876 for 24h in 2i/LIF. 1μM aphidicolin was added to cultures 2h prior to incubation with biotinylated-trimethylpsoralen (bTMP). Following UV-crosslinking, cells were stained with streptavidin-Alexa594 conjugate and DAPI. Distributions for nuclear intensities of bTMP staining are shown. Data are pooled from two biological replicates of each treatment. IdU treatment increases negative supercoiling as compared to DMSO control, $^{**}p < 0.0001$. IdU in combination with CRT0044876 further increases supercoiling levels as compared to DMSO ($^{**}p < 0.0001$) and IdU alone ($^{**}p < 0.0001$). P values were calculated using Kruskal-Wallis test followed by Tukey's multiple comparison test. **(B)** Bleomycin treatment reduces bTMP intercalation into DNA, validating assay sensitivity for negative supercoiling levels. Boxplots show median ±interquartile range of single-cell bTMP staining intensities. Treatment of mESCs with 100μM bleomycin was performed for 1 hour just prior to bTMP incubation. Bleomycin reduces the mean bTMP staining intensity for cells treated with DMSO or 10μM IdU as compared to DMSO control with no bleomycin treatment ($^{**}p < 0.0001$). The reduction in bTMP staining when IdU is coupled with bleomycin indicates that IdU alone in uncoiled DNA does not increase bTMP intercalation. Data shown are pooled from two biological replicates. P values were calculated using Kruskal-Wallis test followed by Tukey's multiple comparison test.

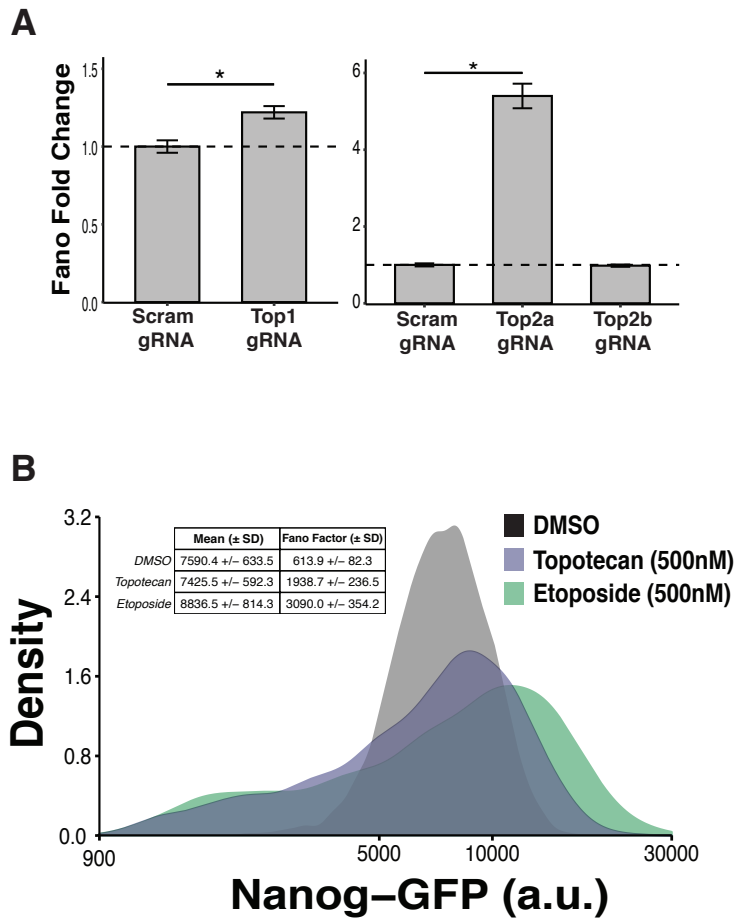


Figure 3.6: Loss of Topoisomerase activity increases Nanog expression variability.
(A) CRISPRi Knockdown of Topoisomerases involved in relaxation of DNA supercoiling. Nanog Fano factor was normalized to scrambled gRNA population. Data represent mean (± SD) of three biological replicates. Knockdown of Top1 (*p = 0.002) and Top2a (*p = 0.003) increases Nanog expression variability. P values were calculated by two-tailed, unpaired Student's t test. **(B)** Small-molecule inhibition of Topoisomerase I and II increases Nanog expression variability. Representative flow cytometry distributions of Nanog-GFP expression in mESCs treated with DMSO, 500nM topotecan or 500nM etoposide for 24 hours in 2i/Lif media. Extrinsic noise filtering via cell-size gating was performed prior to calculation of Nanog Fano factor. Table inset shows mean and Fano factor (±SD) of Nanog expression averaged over three biological replicates of each treatment.

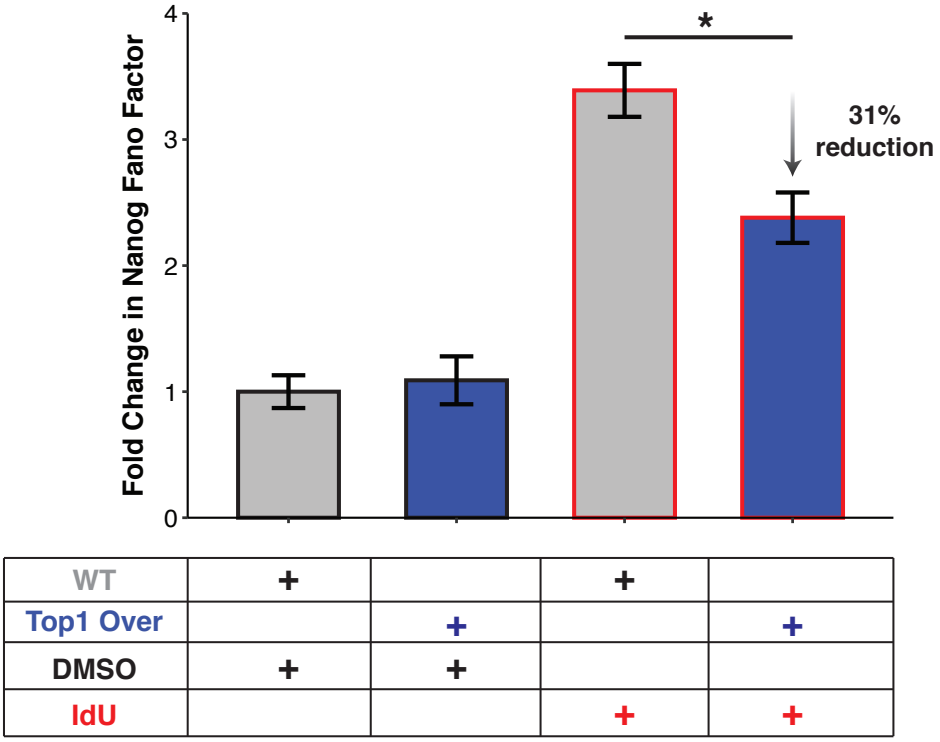


Figure 3.7: Overexpression of Topoisomerase 1 partially ablates noise-enhancement of Nanog expression.
Wildtype (WT) and topoisomerase 1 (Top1) overexpressiong Nanog-GFP mESCs were treated with DMSO or 10μM IdU for 24h in 2i/LIF media. Nanog Fano factor was normalized to wild-type population of Nanog-GFP mESCs treated with DMSO. Data represent mean (± SD) of three biological replicates. Overexpression of Top1 (*p = 0.0036) reduces IdU-mediated enhancement of Nanog variability by 31%. P values were calculated by two-tailed, unpaired Student's t test.

Methods

Biotinylated-trimethylpsoralen (bTMP) supercoiling assay

$1x10^5$ Nanog-GFP mESCs were seeded into each well of a gelatin-coated, 35mm Ibidi dish (quad-chambered, cat:80416) in 2i/LIF media. 24 hours following seeding, media was replaced with 2i/LIF supplemented with 10μM IdU , 10μM IdU & 100μM CRT0044876, or equivalent volume DMSO in replicate. After 24 hours of treatment, media was replaced with 2i/LIF supplemented with 1μM aphidicolin for two hours. For control experiments, Nanog-GFP mESCs were cultured with or without 10μM IdU for 24 hours followed by treatment with 100μM bleomycin for one hour. Cells were then washed 1xDPBS and then permeabilized with 0.1% Tween-20 in DPBS for 15 minutes. Cells were then incubated with 0.3mg/ml EZ-Link Psoarlen-PEG3-Biotin (Thermo cat:29986) for 15 minutes. Cultures were then exposed to 365nM light (AlphaImager HP with 15W bulbs, ProteinSimple) for 15 minutes at room temperature. Cells were then washed 2xDPBS, fixed with cold 70% ethanol for 30 minutes at 4°C, and then washed 2xDPBS. Cells were then incubated with Alexa Fluor 594 Streptavidin (Thermo cat:S32356) for one hour at room temperature in the dark, washed 2xDPBS, and stained with DAPI for 10 minutes at room temperature in the dark. Cells were imaged in a buffer containing 50% glycerol (Thermo), 75 μg/mL glucose oxidase (Sigma Aldrich), 520 μg/mL catalase (Sigma Aldrich), and 0.5 mg/mL Trolox (Sigma Aldrich). Images were taken on a Zeiss Axio Observer Z1 microscope equipped with a Yokogawa CSU-X1 spinning disk unit and 63x/1.4 oil objective. Approximately 20 xy locations were randomly selected for each condition. For each xy location, three z-planes were sampled at 4-μm intervals. Nuclear segmentation using DAPI signal and quantification of psoralen staining intensity were carried out using CellProfiler. After illumination correction and background subtraction, the mean psoralen fluorescence intensity of a segmented nucleus was taken from each z-plane and averaged over the entire z-stack.

Topoisomerase 1 overexpression assay

Topoisomerase 1 (NM 009408) expression vector (Origene, MR218547L3) with puromycin selection marker was packaged into lentivirus using HEK293T cells. Nanog-GFP mESCs were spinoculated with lentiviral supernatant. Three days following transduction, infected cells were subjected to seven days of puromycin (1ug/ml) selection to isolate cells stably overexpressing Topoisomerase 1.

$1x10^5$ wildtype and Topoisomerase 1 overexpressing Nanog-GFP mESCs were seeded in gelatin-coated 12-well plates in 2i/LIF media. 24 hours after seeding, media was swapped with 2i/LIF containing 10μM of IdU or equivalent volume DMSO in triplicate. After 24 hours of treatment, cells were run unfixed and live on BD LSRII cytometer. Extrinsic noise filtering via cell-size gating was performed prior to calculation of Nanog Fano factor. Fano factor for Nanog-GFP expression for each treatment was normalized to the DMSO control for wildtype Nanog-GFP mESCs.

3.1.4 Noise amplification is correlated with increased promoter nucleosome occupancy

In addition to examining DNA supercoiling levels, we wanted to investigate the effects that IdU treatment has on nucleosome positioning. In both yeast and human contexts, transcriptional variability is correlated with the level of nucleosome occupancy in the proximal promoter region (ppr, 200 base pairs upstream of TSS) [82]. High noise genes tend to have greater nucleosome occupancy within the ppr. It is believed that the additional chromatin-remodeling required for gene activation increases transcriptional noise.

To test whether IdU treatment is perturbing nucleosome positioning within promoter regions, I performed ATAC-seq on mESCs treated with DMSO or 10μM IdU for 24 hours in 2i/LIF media. Analysis revealed that IdU treatment increases nucleosome occupancy and decreases the peak-to-

trough ratio for the Nanog promoter region which is classified as a depleted proximal nucleosome gene in the DMSO condition (Figure 3.8B). Examination of the average nucleosome occupancy profile for all DPN genes shows a similar pattern: IdU treatment causes a decrease in the peak-to-trough ratio with increased nucleosome occupancy within the proximal promoter region (Figure 3.8C). Integration of the single-cell RNA-seq and ATAC-seq datasets revealed a strong positive correlation between the level of IdU-mediated noise enhancement for a gene and the increase in nucleosome occupancy for the promoter region (Figure 3.8D).

The correlation between strength of noise-enhancement and increased nucleosome occupancy may be related to chromatin remodeling that occurs during DNA repair or it is a direct consequence of Apex1-induced supercoiling which is known to cause shifting of nucleosomes [79]. If this is the case, the level of nucleosome remodeling within a promoter may be indicative of the supercoiling level within that region.

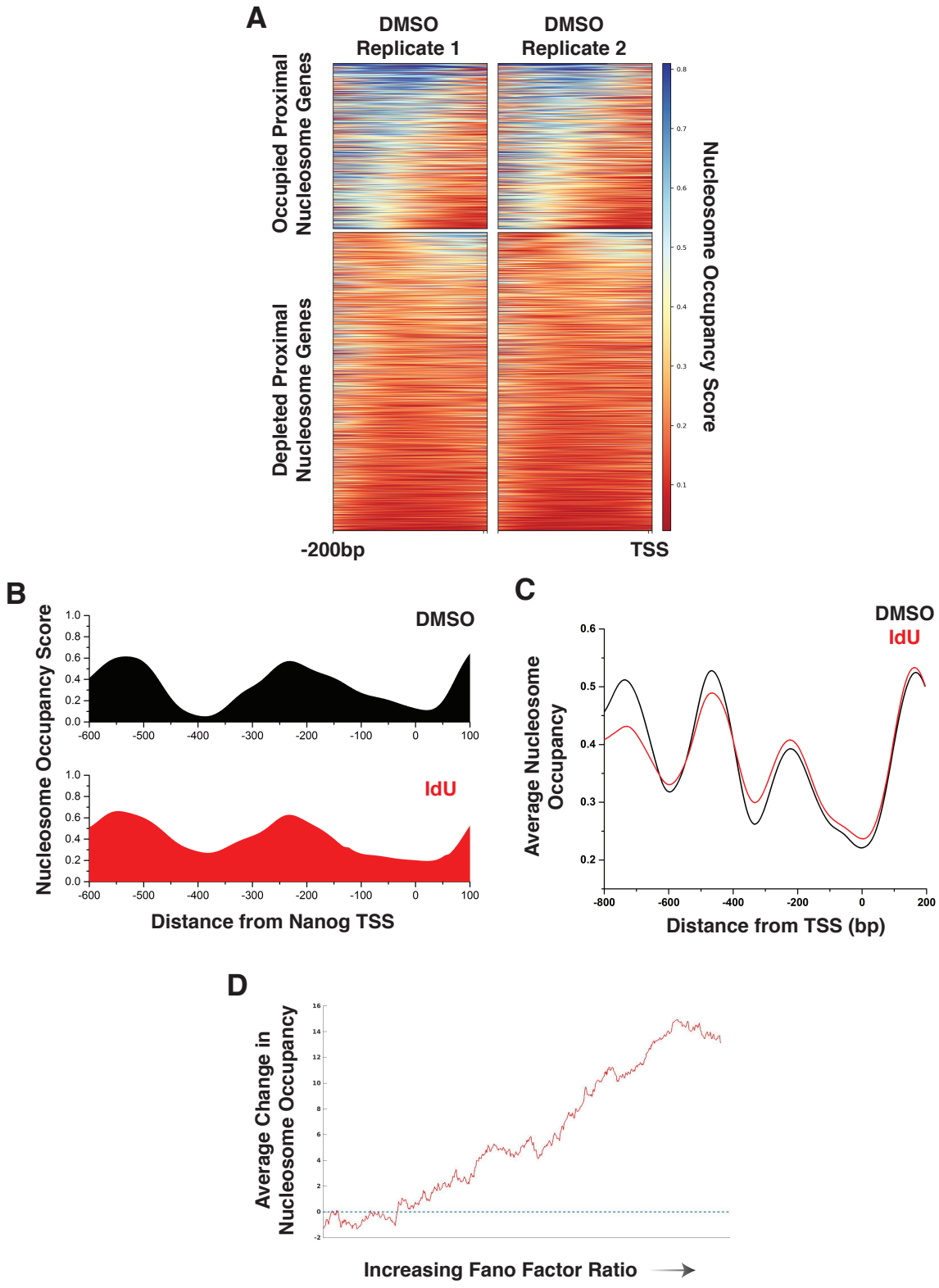

A
DMSO
Replicate 1
DMSO
Replicate 2
Occupied Proximal
Nucleosome Genes
Depleted Proximal
Nucleosome Genes
Nucleosome Occupancy Score
-200bp
TSS
B
DMSO
IdU
Nucleosome Occupancy Score
Distance from Nanog TSS
C
DMSO
IdU
Average Nucleosome
Occupancy
Distance from TSS (bp)
D
Average Change in
Nucleosome Occupancy
Increasing Fano Factor Ratio

Figure 3.8: Noise amplification is correlated with increased promoter nucleosome occupancy. **(A)** ATAC-seq results demonstrate presence of two gene classes based on promoter nucleosome occupancy. ATAC-seq analysis was performed on mESCs treated with DMSO or 10μM IdU for 24 hours. Heatmaps display nucleosome occupancy scores for 13,345 promoter regions. Data for 2 replicates are shown. Genes are classified as having an occupied (OPN) or depleted (DPN) proximal nucleosome based on the nucleosome occupancy score of the 200 base pair region directly upstream of the transcript TSS. **(B)** IdU treatment causes increased nucleosome occupancy for the promoter region of Nanog. Data represent average of two replicates. **(C)** The average promoter nucleosome occupancy profile for all DPN genes shows a decreased peak-to-trough ratio in the IdU condition. IdU treatment minimizes the distance between valley peaks and troughs indicating that nucleosome positioning has been destabilized. **(D)** Strength of IdU-mediated noise enhancement for a gene is positively correlated with change in nucleosome occupancy of the promoter region. Using single-cell RNA-seq data, genes are ranked in ascending order according to the ratio of their Fano factor in the IdU and DMSO conditions. The cumulative nucleosome occupancy score of the 400bp region directly upstream of the TSS is calculated for the IdU and DMSO conditions. The change in nucleosome occupancy for each gene is then plotted against its Fano factor ratio. A smoothing average of the data points was then performed. Genes with the greatest increase in Fano factor also experienced the greatest increase in nucleosome occupancy within their promoter region.

Methods

ATAC-seq analysis of nucleosome positioning.

$2x10^5$ mESCs were seeded in each well of a gelatin-coated, 6-well plate in 2i/LIF media. 24 hours following seeding, cultures were replenished with 2i/LIF media containing 10μM IdU or an equivalent volume of DMSO in replicate for 24 hours. After treatment, cells were trypsinized with TrypLE and ATAC-seq libraries were prepared according to Buenrostro et. al. [83].

Libraries were sequenced in a paired-end fashion on an Illumina HiSeq4000. Sequencing yielded a median of ∼50 million paired-end reads per library. Read quality was checked via FASTQC. Adaptor sequences were then trimmed from each read using atactk. Trimeed reads were then aligned to the mm9 reference genome using the burroughs-wheeler algorithm. Duplicate reads were removed using Picard. Reads aligning to the sex chromosomes and mitochondrial genome were also removed. MACS callpeak function was then used to identify euchromatic regions based on read density. Nucleoatac was then used to compute nucleosome occupancy scores for promoter regions of annotated transcripts in the mm9 genome.

Chapter 4

Mathematical modeling and simulations of Apex1 activity reveals a transcription-coupled repair mechanism

4.1 Introduction

To delineate how Apex1 alters transcriptional dynamics of Nanog gene expression in a way that increases noise without changing mean, we developed a series of models that allow for Apex1 interaction at different stages of the transcription process. Through stochastic simulation of these models and comparison to experimental data, the aim is to develop greater mechanistic insight into what stages of the transcription process Apex1 affects. The tested models listed in Figure 4.2 are adapted from the two-state random telegraph model. We assume in each of the models that IdU incorporation into the genome leads to recruitment of Apex1 ($k_{incorpo}$). The resulting interaction results in a transcriptionally non-productive state. Unbinding of Apex1 is triggered by completion of repair (k_{repair}).

4.2 Detailed mathematics and derivation of parameter constraints

We derive here some relations, at equilibrium, between the kinetic rates of the diverse models. These relationships are then used to constrain the parameter phase space for a given set of data.

4.2.1 Model 0

This is the *null* model and consists of only the canonical two-state random-telegraph. This model is used as a *null* hypothesis, in particular for *log-likelihood* and *AIC*- based model selection.

4.2.2 Model 1

We assume that IdU incorporation and subsequent interaction of Apex1 with the chromatin, occurs only in the OFF state of the promoter. Biologically, this may occur if control mechanisms inhibit DNA repair during active transcription.

The differential equations governing the gene fractions in the different states and the mRNA counts are as follows:

$$\begin{cases} \frac{dOFF}{dt} = k_{ON} \cdot ON + k_{repair} \cdot OFF^* - (k_{incorpo} + k_{ON}) \cdot OFF \\ \\ \frac{dON}{dt} = k_{ON} \cdot OFF - k_{OFF} \cdot ON \\ \\ \frac{dOFF^*}{dt} = k_{incorpo} \cdot OFF - k_{repair} \cdot OFF^* \\ \\ \frac{dRNA}{dt} = k_{RNA} \cdot ON - k_{decay} \cdot RNA \end{cases}$$

Using the fact that:

$$\frac{K_{ON}}{K_{ON}+K_{OFF}} \equiv \frac{ON}{ON+OFF+OFF^*} \tag{4.1}$$

we derive the following constraint between k_{repair} and $k_{incorpo}$, where K_{OFF} represents the transition rate to the macroscopic OFF state consisting of the OFF (k_{OFF}) and OFF*($k_{incorpo}$) states in Model 1:

$$k_{repair} = \frac{K_{ON}+K_{OFF}}{K_{ON}} \cdot \frac{k_{ON}}{k_{ON}+k_{OFF}+k_{incorpo}} - k_{incorpo} \tag{4.2}$$

4.2.3 Model 2

In this model, Apex1 can interact with chromatin in both the ON and OFF states of the promoter. If Apex1 interacts with the chromatin in the ON state, this leads to a turning off of the system. Molecularly this may be seen as a strong inhibitory effect mediated by Apex1: the interaction may recruit chromatin modifiers (e.g. histone deacetylases, histone methyltransferases) that silence gene expression. In the same way, stalled polymerases, at both the promoter proximal region and further in the gene, may unbind DNA.

This model can be described by the following set of ODEs:

$$\begin{cases} \frac{dOFF}{dt} = k_{OFF} \cdot ON + k_{repair} \cdot OFF^* - (k_{incorpo} + k_{ON}) \cdot OFF \\ \\ \frac{dON}{dt} = k_{ON} \cdot OFF - (k_{OFF} + k_{incorpo}) \cdot ON \\ \\ \frac{dOFF^*}{dt} = k_{incorpo} \cdot OFF + k_{incorpo} \cdot ON - k_{repair} \cdot OFF^* \\ \\ \frac{dRNA}{dt} = k_{RNA} \cdot ON - k_{decay} \cdot RNA \end{cases}$$

Using equation (4.1) we derive the following constraint between k_{repair} and $k_{incorpo}$:

$$\frac{k_{incorpo}}{k_{repair}} = \frac{K_{OFF}}{K_{ON}} \cdot \frac{k_{ON}}{k_{OFF}} - 1 \tag{4.3}$$

4.2.4 Model 3

Here we assume that Apex1 can still interact in the ON state but that does not alter the ”primed” characteristic of the gene expression system. The system is thus in a *transcriptionally non-productive* ON* state. In other words transcription can not be achieved when Apex1 interacts with the chromatin but the transcriptionally permissive chromatin and molecular context is not altered: primed polymerases remain, transcription enhancing epigenetic marks are not erased, etc.

This model can be described by the following set of ODEs:

$$\begin{cases} \frac{dOFF}{dt} = k_{OFF} \cdot ON + k_{repair} \cdot OFF^* - (k_{incorpo} + k_{ON}) \cdot OFF \\ \\ \frac{dON}{dt} = k_{ON} \cdot OFF + k_{repair} \cdot ON^* - (k_{OFF} + k_{incorpo}) \cdot ON \\ \\ \frac{dOFF^*}{dt} = k_{incorpo} \cdot OFF - k_{repair} \cdot OFF^* \\ \\ \frac{dON^*}{dt} = k_{incorpo} \cdot ON - k_{repair} \cdot ON^* \\ \\ \frac{dRNA}{dt} = k_{RNA} \cdot ON - k_{decay} \cdot RNA \end{cases}$$

We can thus derive the following constraint:

$$\frac{k_{incorpo}}{k_{repair}} = \frac{1}{k_{ON} + k_{OFF}} \cdot [k_{ON} \cdot \frac{K_{OFF}}{K_{ON}} - k_{OFF}] \tag{4.4}$$

The details of this derivation are described below for Model 4.

4.2.5 Model 4

Model 4 is based on Model 3, but an amplification step was added. When the system transitions from ON* to ON, the basal transcription rate (k^0_{RNA}) increased by a multiplicative factor: $k_{RNA2} = coop \cdot k^0_{RNA}$. Thus we assume that there is molecular memory of the repair event. This may be rooted in a modification of supercoiling, polymerase accumulation, or chromatin remodel-

ing.

This model can be described by the following set of ODEs :

$$\begin{cases} \frac{dOFF}{dt} = k_{OFF} \cdot ON + k_{repair} \cdot OFF^* - (k_{incorpo} + k_{ON}) \cdot OFF \\ \\ \frac{dON}{dt} = k_{ON} \cdot OFF + k_{repair} \cdot ON^* - (k_{incorpo} + k_{OFF}) \cdot ON \\ \\ \frac{dOFF^*}{dt} = k_{incorpo} \cdot OFF - k_{repair} \cdot OFF^* \\ \\ \frac{dON^*}{dt} = k_{incorpo} \cdot ON - k_{repair} \cdot ON^* \\ \\ \frac{dRNA}{dt} = \langle k_{RNA} \rangle \cdot ON - k_{decay} \cdot RNA \end{cases}$$

Where $k_{incorpo} = [IdU] \cdot k^0_{incorpo}$. We assume that IdU is in excess and thus [IdU] remains constant.

At equilibrium :

$$\frac{dOFF}{dt} = \frac{dON}{dt} = \frac{dOFF^*}{dt} = \frac{dON^*}{dt} = \frac{dRNA}{dt} = 0$$

$$\Longrightarrow \begin{cases} OFF_{eq} = \frac{k_{OFF}}{k_{ON}} \cdot ON_{eq} \\ ON_{eq} = \frac{k_{ON}}{k_{OFF}} \cdot OFF_{eq} \\ OFF^*_{eq} = \frac{k_{incorpo}}{k_{repair}} \cdot OFF_{eq} \\ ON^*_{eq} = \frac{k_{incorpo}}{k_{repair}} \cdot ON_{eq} \\ RNA_{eq} = \frac{\langle k_{RNA} \rangle}{k_{decay}} \cdot ON_{eq} \end{cases} \tag{4.5}$$

By definition:

$$\frac{K_{ON}}{K_{ON} + K_{OFF}} \equiv \frac{ON}{ON + OFF + ON^* + OFF^*} \tag{4.6}$$

where K_{OFF} represents the transition rate to the macroscopic OFF state (defined by no transcription) consisting of the OFF, OFF*, and ON* states. Using the set of equations in (4.5) and (4.6) we obtain:

$$\frac{K_{ON}}{K_{ON} + K_{OFF}} = \frac{k_{repair}}{k_{repair} + k_{incorpo}} \cdot \frac{k_{ON}}{k_{ON} + k_{OFF}} \tag{4.7}$$

Equation (4.7) can be seen as :

$$P(ON_{macro}) = P(\text{"}Repaired\ state\text{"}) \cdot P(ON_{micro})$$

Because $P(ON_{macro}) \equiv P(\text{"}Repaired\ state\text{"} \cap ON_{micro})$ we can deduce that "*Repaired state*" and ON_{micro} are independent probabilistic events.

Equation (4.7) can be rewritten as:

$$\frac{K_{ON}}{K_{OFF}} = \frac{k_{repair} \cdot k_{ON}}{k_{OFF} \cdot (k_{repair} + k_{incorpo}) + k_{ON} \cdot k_{incorpo}} \tag{4.8}$$

Which gives us:

$$\frac{k_{incorpo}}{k_{repair}} = \frac{1}{k_{ON} + k_{OFF}} \cdot [k_{ON} \cdot \frac{K_{OFF}}{K_{ON}} - k_{OFF}] \tag{4.9}$$

Let us define $\langle k_{RNA} \rangle$ and k^0_{RNA} as the mean transcription rate in the presence of IdU and the transcription rate in the control condition (DMSO), respectively. By construction of the model:

$$\langle k_{RNA} \rangle = P(ON \mid OFF) \cdot k^0_{RNA} + P(ON \mid ON^*) \cdot coop \cdot k^0_{RNA} \tag{4.10}$$

The $coop$ term represents the amplification of the transcription rate following completion of repair.

$P(ON \mid OFF)$ and $P(ON \mid ON^*)$ represent the probability that the gene transitioned to the ON state from the OFF and ON* states respectively.

Or :

$$\begin{cases} P(ON \mid OFF) = \frac{k_{ON} \cdot OFF}{k_{ON} \cdot OFF + k_{repair} \cdot ON^*} \\ \\ P(ON \mid ON^*) = \frac{k_{repair} \cdot ON^*}{k_{ON} \cdot OFF + k_{repair} \cdot ON^*} \end{cases} \tag{4.11}$$

Combining equations (4.9) and (4.11) we get:

$$\begin{cases} P(ON \mid OFF) = \frac{k_{OFF}}{k_{OFF} + k_{incorpo}} \\ \\ P(ON \mid ON^*) = \frac{k_{incorpo}}{k_{OFF} + k_{incorpo}} \end{cases} \tag{4.12}$$

Rewriting (4.10) using (4.12) we obtain :

$$\frac{\langle k_{RNA} \rangle}{k^0_{RNA}} = \frac{1}{k_{OFF} + k_{incorpo}} \cdot (k_{OFF} + coop \cdot k_{incorpo}) \tag{4.13}$$

$$\implies coop = \frac{k_{OFF} + k_{incorpo}}{k_{incorpo}} \cdot \frac{\langle k_{RNA} \rangle}{k^0_{RNA}} - \frac{k_{OFF}}{k_{incorpo}} \tag{4.14}$$

Equation (4.14) can be rewritten to explicitly take into account [IdU] :

$$coop = \frac{k_{OFF} + k^0_{incorpo} \cdot [IdU]}{k^0_{incorpo} \cdot [IdU]} \cdot \frac{\langle k_{RNA} \rangle}{k^0_{RNA}} - \frac{k_{OFF}}{k^0_{incorpo} \cdot [IdU]} \tag{4.15}$$

4.2.6 Model 5

In this last model we make the same assumptions as in model 4 except that Apex1 interaction with chromatin occurs *only* in the ON state. This does not imply that IdU incorporation and subsequent Apex1 interaction only occurs in the ON state. Our assumption supposes that the interaction of Apex1 in the OFF state is negligible quantitatively speaking compared to that in the ON state.

The ODEs describing this model are as follows:

$$\begin{cases} \frac{dOFF}{dt} = k_{OFF} \cdot ON - k_{ON} \cdot OFF \\ \\ \frac{dON}{dt} = k_{ON} \cdot OFF + k_{repair} \cdot ON^* - (k_{incorpo} + k_{OFF}) \cdot ON \\ \\ \frac{dON^*}{dt} = k_{incorpo} \cdot ON - k_{repair} \cdot ON^* \\ \\ \frac{dRNA}{dt} = \langle k_{RNA} \rangle \cdot ON - k_{decay} \cdot RNA \end{cases}$$

The *coop* expression remains unchanged from Model 4 but the ratio between k_{repair} and $k_{incorpo}$ changes as follows:

$$\frac{K_{ON}}{K_{OFF}} \equiv \frac{ON}{OFF + ON^*} \tag{4.16}$$

Thus we obtain:

$$\frac{K_{ON}}{K_{OFF}} = \frac{k_{repair} \cdot k_{ON}}{k_{OFF} \cdot k_{repair} + k_{incorpo} \cdot k_{ON}} \tag{4.17}$$

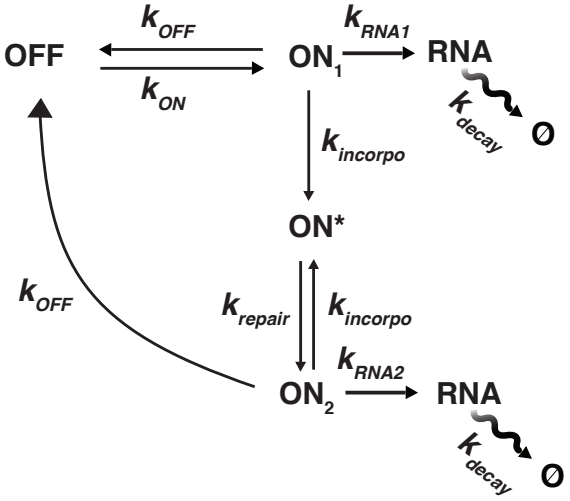
And :

$$\frac{k_{incorpo}}{k_{repair}} = \frac{K_{OFF}}{K_{ON}} - \frac{k_{OFF}}{k_{ON}} \tag{4.18}$$

4.3 Chemical Master Equation

For all models we constructed an associated stochastic scheme. As an example, Model 5 can be rewritten using the following scheme:

$$\begin{cases} OFF \xrightarrow{k_{ON}} ON_1 \\ ON_1 \xrightarrow{k_{OFF}} OFF \\ ON_1 \xrightarrow{k_{RNA1}} RNA \\ ON_1 \xrightarrow{k_{incorpo}} ON^* \\ ON^* \xrightarrow{k_{repair}} ON_2 \\ ON_2 \xrightarrow{k_{incorpo}} ON^* \\ ON_2 \xrightarrow{k_{OFF}} OFF \\ ON_2 \xrightarrow{k_{RNA2}} RNA \\ RNA \xrightarrow{k_{decay}} \varnothing \end{cases} \tag{4.19}$$



ON_2 represents the repaired state of the gene, which results in a higher transcription rate (k_{RNA2}). Using this scheme, we can construct the chemical master equation (CME) describing the time dependent distributions of mRNA copy number:

$$\frac{d\mathbf{P}(m,t)}{dt} = \mathbf{A}\cdot\mathbf{P}(m,t) + \delta(\mathbf{E}-\mathbf{I})[m\mathbf{P}(m,t)] + \Delta(\mathbf{E}^{-1}-\mathbf{I})[\mathbf{P}(m,t)] \tag{4.20}$$

Where $\mathbf{A}$, Δ, and δ are the transition, transcription, and degradation matrices respectively:

$$\mathbf{A} = \begin{bmatrix} -k_{ON} & 0 & -k_{OFF} & -k_{OFF} \\ 0 & -k_{repair} & k_{incorpo} & k_{incorpo} \\ k_{ON} & 0 & -(k_{OFF}+k_{incorpo}) & 0 \\ 0 & k_{repair} & 0 & -(k_{incorpo}+k_{OFF}) \end{bmatrix}$$

$$\Delta = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & k_{RNA1} & 0 \\ 0 & 0 & 0 & k_{RNA2} \end{bmatrix}$$

$$\delta = \begin{bmatrix} k_{decay} & 0 & 0 & 0 \\ 0 & k_{decay} & 0 & 0 \\ 0 & 0 & k_{decay} & 0 \\ 0 & 0 & 0 & k_{decay} \end{bmatrix}$$

$\mathbf{P}(m,t)$ is a four-element column vector consisting of the time-dependent mRNA probability distributions while in the $OFF, ON^*, ON_1, and ON_2$ states respectively. $\mathbf{E}$ and $\mathbf{E}^{-1}$ are the forward and backward shift operators while $\mathbf{I}$ is the identity matrix.

At steady-state, the mRNA probability distribution can be reconstructed as a sum of the binomial moments [84]

$$P(m) = \sum_{k \geq m} (-1)^{m-k} \binom{k}{m} b_k, m = 0,1,2,... \tag{4.21}$$

Where b_k is the k^{th} binomial moment of the distribution given by:

$$b_k = \frac{1}{\prod_{i=1}^{k} det(i\delta - \mathbf{A})} \cdot \prod_{i=k}^{1} [\mathbf{u}_N (i\delta - \mathbf{A})^* \Delta] \cdot \mathbf{b}_0, k = 1,2,... \tag{4.22}$$

where $\mathbf{u}_N$ = [1,1,1,1]. $(i\delta - \mathbf{A})^*$ and $det(i\delta - \mathbf{A})$ are the adjugate and the determinant of matrix $(i\delta - \mathbf{A})$ respectively. b_1 is equivalent to the mean mRNA abundance of the system at equilibrium. $\mathbf{b}_0$ is the corresponding eigenvector for the zero eigenvalue of $\mathbf{A}$. The 4 elements of $\mathbf{b}_0$ therefore

represent the fraction of time spent in the $OFF, ON^*, ON_1, and\ ON_2$ states at equilibrium. The n^{th} component of $\mathbf{b}_0$ is given by:

$$b_0^{(n)} = \prod_{i=1}^{3} \frac{\beta_i^{(n)}}{\alpha_i}, 1 \leq n \leq 4 \tag{4.23}$$

where $\alpha_1, \alpha_2, \alpha_3$ are the three non-zero eigenvalues of $\mathbf{A}$. $\beta_1^{(n)}, \beta_2^{(n)}, \beta_3^{(n)}$ are the three eigenvalues of the sub-matrix $\mathbf{M}_n$ which is constructed by removing the n^{th} row and n^{th} column of $\mathbf{A}$.

The Fano factor for mRNA counts at equilibrium is then given by:

$$FF = \frac{2b_2 + b_1 - b_1^2}{b_1} \tag{4.24}$$

An exact simulation of such a stochastic process is given by the Gillespie algorithm. We implemented the algorithm using a homemade script in Julia 1.1.1. For each model, 1500 simulations were run for a virtual duration of 200 hours. Since the time is not discrete we used a "parsing" algorithm, based on recursive binary search, to align all the traces on a common time scale of 2000 intervals. For each interval of the discretized time we computed the average and variance of the number of mRNAs, taking into account all traces. The analytic relationships described above are used to verify the inferred kinetic rates from stochastic simulations.

4.4 Estimation of model parameters from experimental data

From the smRNA FISH data, we can infer the *macroscopic* kinetic rates. The kinetic rates computed in the control condition are at the basis of the simulations and have to be considered as constant for all the models and associated results.

Use of these macroscopic rates along with the relationships developed in section 2 results in one remaining degree of freedom for our models: k_{repair} (or $k_{incorpo}$ for Model 1).

The confrontation between experimental data and the results of stochastic simulations for each of the models should allow us to define the consistency of our models and thus gain mechanistic insight into how Apex1 affects transcriptional dynamics.

4.5 Model selection: Comparison of simulation results to experimental data

4.5.1 Information theory-based approach: MLE and Akaike's criterion

Workflow

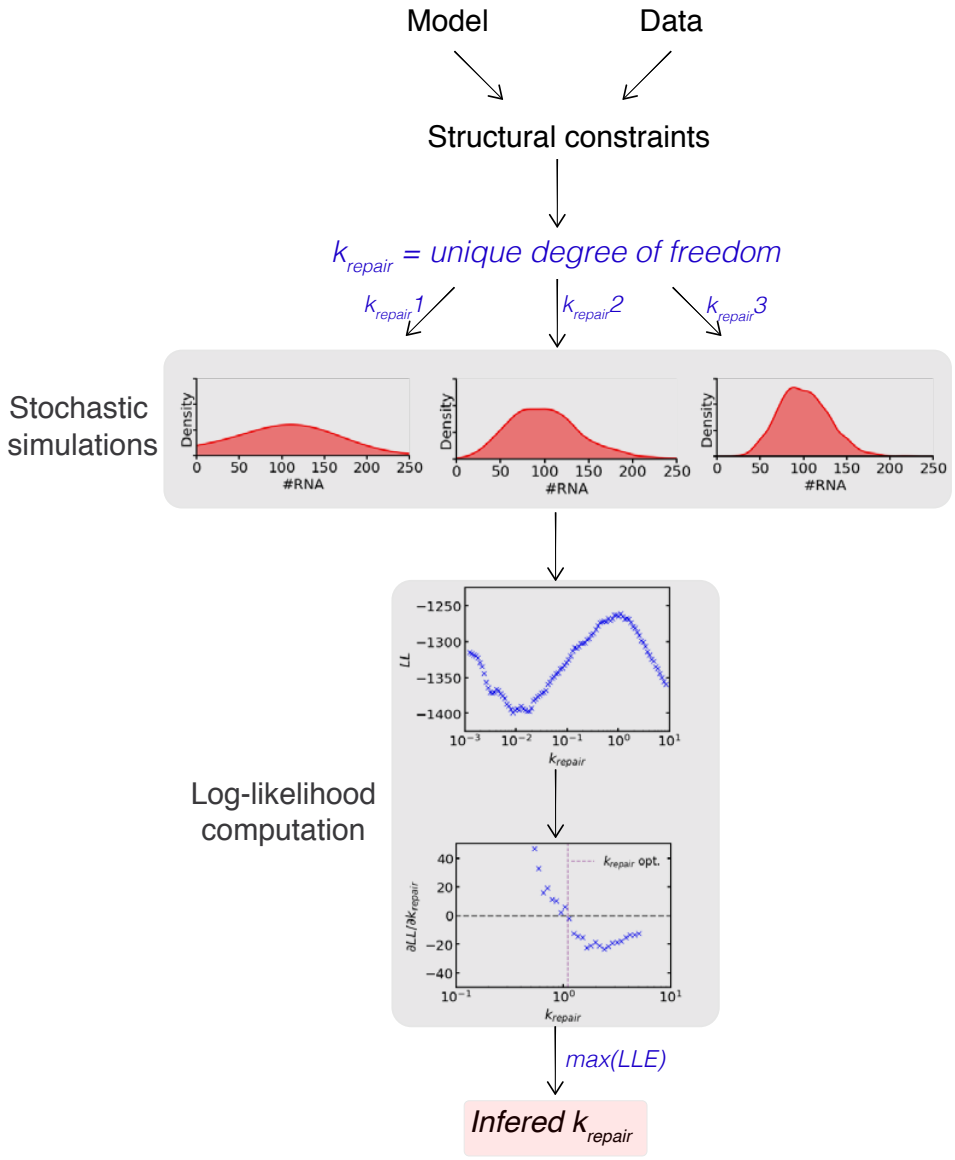


Figure 4.1: Workflow for information theory-based approach of model selection.

Method of Log-likelihood and AIC computation

For each of the 500 chosen values (logarithmically spaced) of our unique degree of freedom (k_{repair}), we ran 50,000 simulations for a total duration of 200h. Only the last point of each simulation was stored. Then the steady-state distribution of RNA was computed using the same binning as in the experimental data. The computation of the log-likelihood is as follows:

Let $\mathbf{X}$ be the vector of n empirical observations and let $\mathscr{X}_\Delta : \{x^\Delta, x^\Delta \in [x, x+\Delta]\}$, where Δ is the size of a bin and $\mathscr{X}_\Delta$ is the set of bins. We define $\hat{p}(x^\Delta)$ as the probability of observing an experimental value within a particular bin (x^Δ). $P(x^\Delta)$ is the probability of a given $x^\Delta \in [x, x+\Delta]$ for a given model. The likelihood function $\mathscr{L}$ is defined as :

$$\mathscr{L} \equiv \prod_{x_i \in \mathbf{X}} P(x_i, x_i \in [x, x+\Delta]) = \prod_{x^\Delta \in \mathscr{X}_\Delta} P(x^\Delta)^{n \cdot \hat{p}(x^\Delta)} \tag{4.25}$$

The log-likelihood is then given as:

$$\log(\mathscr{L}) = \sum_{x^\Delta \in \mathscr{X}_\Delta} n\hat{p}(x^\Delta) \cdot \log(P(x^\Delta)) \tag{4.26}$$

$\mathscr{L}$ is a function of k_{repair}. We then try to find the value of k_{repair} that maximizes $\log(\mathscr{L})$:

$$\hat{k}_{repair,LLE} = \text{argmax}_k(\log(\mathscr{L}(k))) \tag{4.27}$$

With $k_{repair} \in [10^{-4}, 10]$, by assumption. After computing $\log(\mathscr{L}^{(k)})$ for each of the 500 values of k_{repair}, we apply a smoothing (moving average) to the data, take the derivative, smooth the derivative, find the two points on each side of the abscissa, and then interpolate the point for which the derivative is equal to zero using a linear interpolation. The maximum $\log(\mathscr{L})$ is computed after

the first smoothing. Then, we compute the macroscopic behavior of the system using the inferred value of k_{repair}.

The model selection is based on the *AIC* and the resulting measures $\Delta_i AIC$ and w_i [85]. Because $\hat{k}_{repair,LLE}$ is dependent on the empirical distribution, we can assume that it is only an estimate of the *true* value, which we'll call k^0_{repair}. Thus we want to reduce as much as possible the distance between $\hat{k}_{repair,LLE}$ and k^0_{repair}. This optimization problem allows us to derive the so-called Akaike's information criterion (AIC) as a measure to compare models. *AIC* is an estimate of the expected relative distance between the fitted model and the unknown true mechanism that actually generated the observed data:

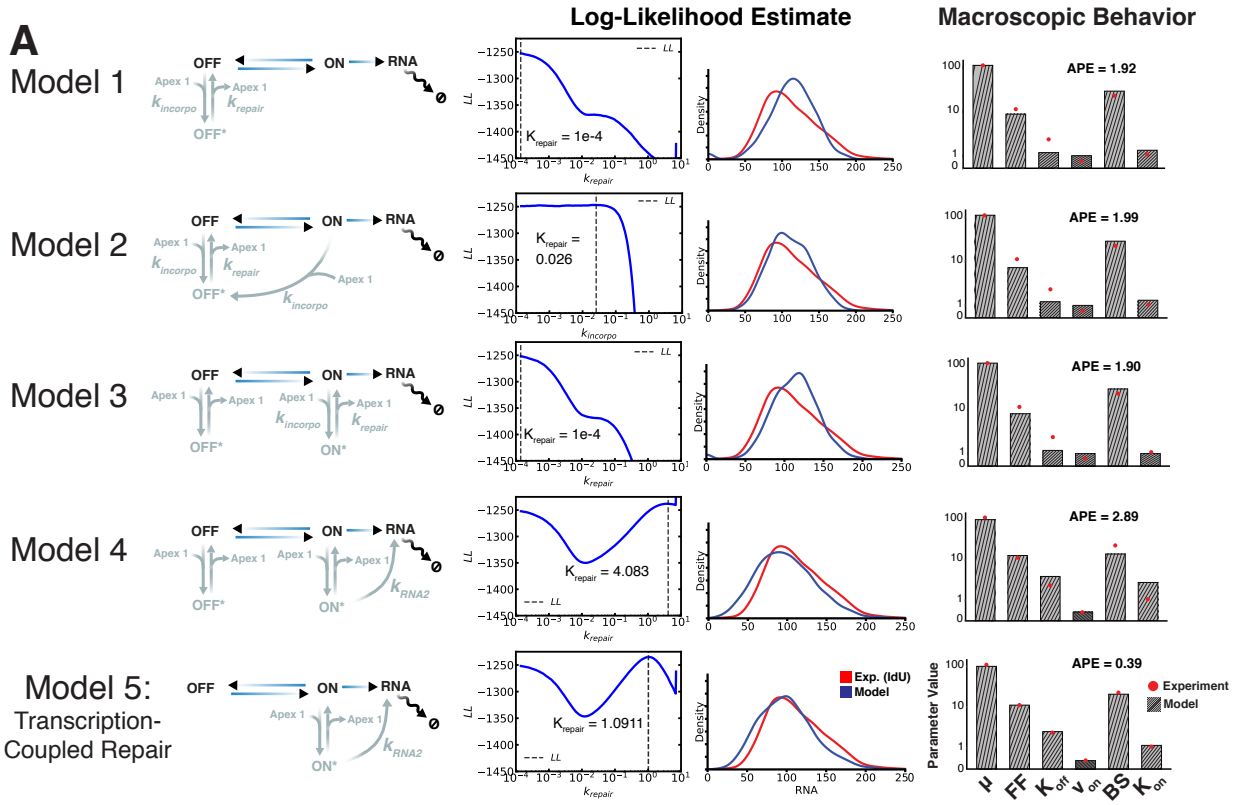
$$AIC = -2\log(\mathscr{L}(\hat{k}_{repair,LLE})) + 2K \tag{4.28}$$

with K the number of degrees of freedom of the system.

Results

As shown in Figures 4.2A-B, Model 5 is selected on the basis of AIC. Model 4 is second-best. Model 5 qualitatively and quantitatively matches experimental data with its inferred k_{repair}.

We used bootstrapping to assess the quality of the inference of k_{repair} for Model 5 using MLE estimator. This method allows the computation of the confidence/credible interval (CI) for k_{repair}, and, in the framework of Bayesianism, a posterior distribution $P(Data|k_{repair})$ using a non-informative prior [85]. The distribution of the MLE is peaked around a particular value, suggesting parameter identifiability (Figure 4.2C).



B

Model	max (LL)	Δ_iAIC	w_i
Control	-1247	24	0.000
1	-1253	36	0.000
2	-1247	24	0.000
3	-1252	34	0.000
4	-1238	6	0.05
5	-1235	0	0.95

C

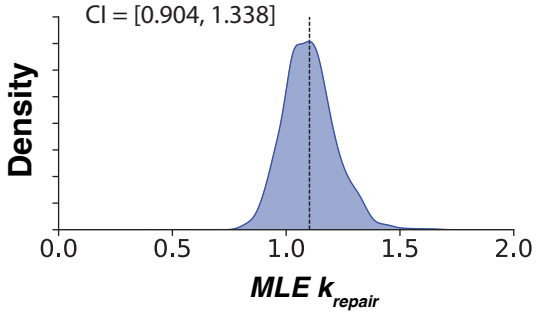


Figure 4.2: MLE-based approach for model selection reveals transcription-coupled repair mechanism best recapitulates experimental data.
(A) (First Column) Schematic of simulated models incorporating Apex1 into standard 2-state model of transcription. (Second Column) For each model, 500 logarithmically-spaced values of $k_{repair} \in [10^{-4}, 10]$ were simulated. For each simulated value of k_{repair}, log-likelihood is calculated as described in supplementary text 5.1.2 and plotted. Dashed vertical line in each plot denotes value of k_{repair} that maximizes log-likelihood estimate. (Third Column) Comparison of experimental Nanog mRNA distribution (red) to simulated distributions of Nanog mRNA (blue) for each model using value of k_{repair} that maximizes log-likelihood. (Fourth Column) Macroscopic behavior of simulation results (using value of k_{repair} that maximizes log-likelihood estimate) are compared to experimental data. Bars represent simulation values of Nanog gene expression system while red points with vertical line represent experimental data on Nanog expression from smRNA-FISH of mESCs treated with 10μM IdU. **(B)** For each tested model, the maximum log-likelihood value is listed along with the associated $\Delta_i AIC$. Model 5 (transcription-coupled repair) best describes experimental data based on these metrics. **(C)** Distribution and confidence interval (CI) of inferred k_{repair} values (based on MLE) for Model 5 using bootstrapping method in which the empirical distribution of Nanog mRNA counts from smRNA-FISH data was re-sampled 1000 times with replacement (supplementary text 5.1.4). Bootstrapping results show a well peaked distribution indicating practical parameter identifiability for k_{repair}.

4.5.2 APE-based approach

Workflow

The second approach for model selection that we employed involves comparison of the macroscopic behavior of simulation results to experimental data. After deriving constraints on the phase space spanned by k_{repair} we simulate each of the models over a range of k_{repair} values. From these stochastic simulations, we then extract the macroscopic behavior of the system for a given model. We computed the mean number of mRNA and the Fano factor at equilibrium, on the last 100 time points of the traces. k_{off} and k_{on} were computed using a non-linear curve fitting assuming exponentially distributed residence times (Poisson process). Both ν_{on} and the burst size were computed using their basic definitions. The density of mRNA population was computed using the last points of the simulations. A classic kernel density was used to represent the data.

For each model, we then infer the best value of k_{repair} based on the minimization of a loss function (absolute percentage error, a.k.a. APE). This quantitative approach is coupled with visual comparison of model behavior to experimental data. The model whose inferred k_{repair} value minimizes divergence between model and data behavior is thus chosen. A graphical representation of such a process is given in Figure 4.3.

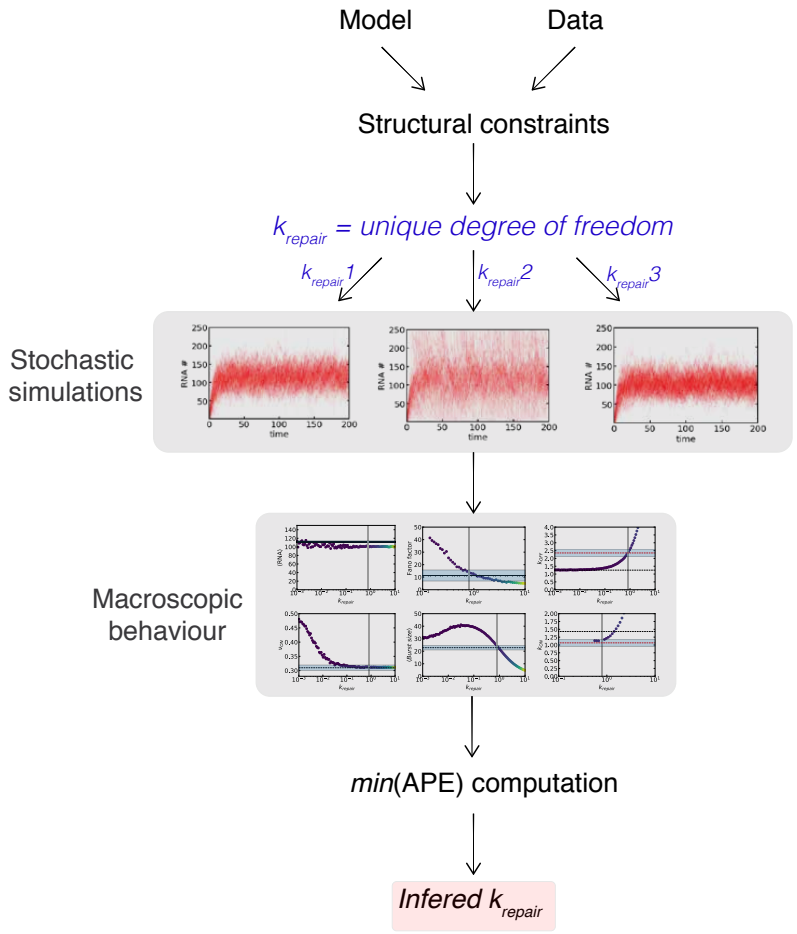


Figure 4.3: Workflow for APE-based approach of model selection.

Method of APE calculation

To discriminate between models based on their macroscopic behavior, we need a measure that quantifies the discrepancy between model-derived results and experimental data. Because the macroscopic observations (Fano factor, k_{ON}, etc.) are of different orders of magnitude, we need a relative measure to avoid the largest parameters carrying the highest weight on the error. We chose to use the absolute percentage error (APE). The procedure is as follows:

Consider the vector $\mathbf{M} \equiv (\langle RNA \rangle, \nu_{ON}, BS, FF, k_{ON}, k_{OFF})$. Thus, $\mathbf{M}_{model}$ and $\mathbf{M}_{exp}$ contain all

the macroscopic observations from modeling and experiment respectively. $\mathbf{M}_{model}$ is a function of k_{repair} or $k_{incorpo}$ equivalently. $\hat{k}_{repair,APE}$ is the best inferred value of the degree of freedom for a particular model. It is given by:

$$\hat{k}_{repair,APE} = \text{argmin}_k \left| \frac{\mathbf{M}_{exp} - \mathbf{M}_{model}(k)}{\mathbf{M}_{exp}} \right| \tag{4.29}$$

With $k \in [10^{-4}, 10]$, by assumption. This notation implies that we are minimizing the $\ell 1$ norm (sum of the vector components). In an operative manner, we simulated each model for 250 logarithmic distributed values of k_{repair} using the previously described Gillespie algorithm.

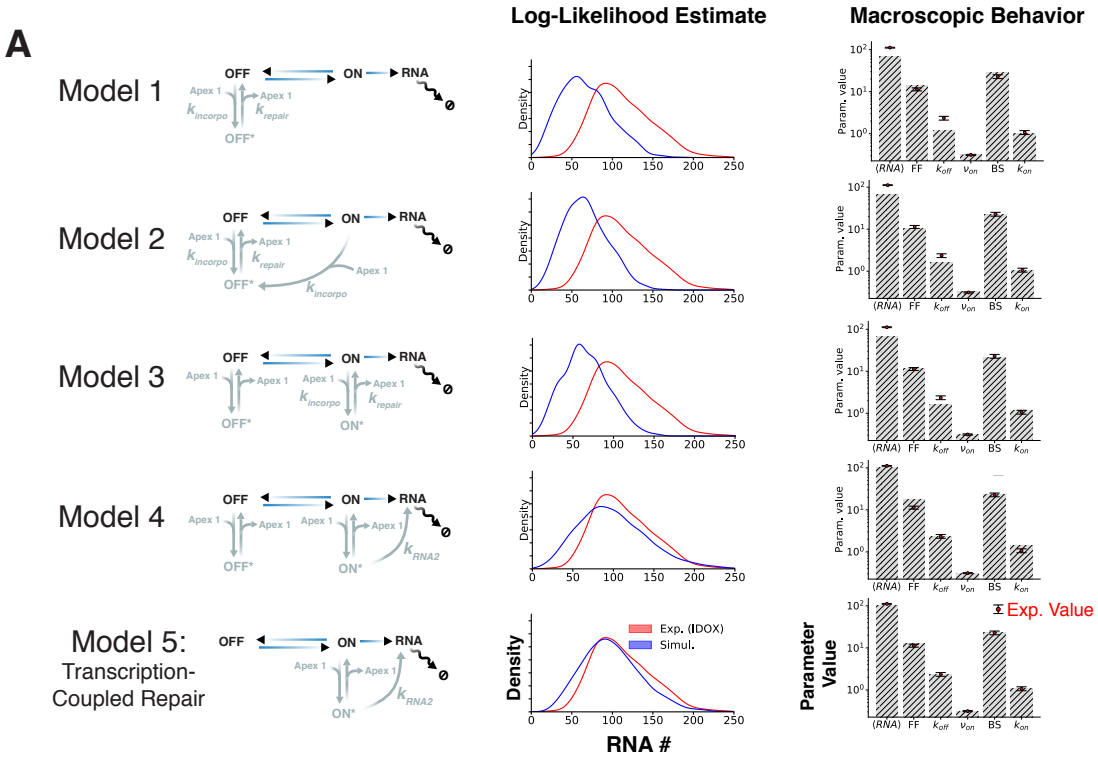
To validate our approach, we devised an alternative loss function where the $\ell 1$ norm is computed using a non-biased (i.e symmetric) measure of relative prediction accuracy: the absolute log accuracy (ALA). The procedure is as follows:

$$\hat{k}_{repair,ALA} = \text{argmin}_k \left| \log \left(\frac{\mathbf{M}_{model}(k)}{\mathbf{M}_{exp}} \right) \right| \tag{4.30}$$

With $k \in [10^{-4}, 10]$, by assumption.

Results

Using both the APE and ALA approaches, we obtained *exactly* the same results for parameter inference (except for model 1 for which $\hat{k}_{incorpo,ALA} = 0.39$) and model selection. According to the APE-based approach for parameter inference and model selection, model 5 best recapitulates the macroscopic behavior observed in the data (Figure 4.4).



B

Model	APE	$\hat{k}_{repair,APE}$	log(L)
1	1.43	3.27	-1650.7
2	0.89	0.43	-1766.7
3	0.92	0.74	-1752.5
4	0.99	1.71	-1315.5
5	0.27	0.89	-1256.3

Figure 4.4: APE-based approach for model selection concurs with MLE-based approach, identifying TCR model as best match to experimental data.
(A) (First Column) Schematic of simulated models incorporating Apex1 into standard 2-state model of transcription. (Second Column) Comparison of experimental Nanog mRNA distribution (red) to simulated distributions of Nanog mRNA (blue) for each model using value of k_{repair} that minimizes absolute percentage error. (Third Column) Macroscopic behavior of simulation results (using value of k_{repair} that minimizes absolute percentage error) are compared to experimental data (supplementary text 5.2). Bars represent simulation values of Nanog gene expression system while red points with vertical line represent experimental data on Nanog expression from smRNA-FISH of mESCs treated with 10μM IdU. **(B)** Values of k_{repair} that minimize the absolute percentage error for each model are listed. Model 5 (TCR model) yields the smallest APE and the largest log-likelihood.

4.5.3 Selection of Model 5

Based on both the MLE- and APE-based approaches, model 5 best matches experimental FISH data and scRNA-seq data for Nanog. Model 5 is most similar to a transcription-coupled repair (TCR) mechanism in which repair events only occur while the gene is transcriptionally permissive. The key adjustment we make is the introduction of a higher transcription rate upon completion of repair. This modification is justified by our model selection process, as we find that the second-best model (model 4) also has an amplified transcription rate following Apex1 interaction and subsequent DNA repair. Interestingly, for the same value of k_{repair}, transcription-coupled repair (model 5) leads to a less significant increase in noise and better maintenance of mean as compared to model 4 in which repair can also take place in the OFF state (model 4). This implies that coupling of repair to transcription is the most efficient method of DNA repair in terms of minimizing excess transcriptional variability.

Several molecular mechanisms can lead to the amplified transcription rate that we have incorporated into model 5. Apex1 binding may both momentarily silence RNA transcription while also inducing increased chromatin supercoiling and chromatin remodelling. This could lead to a subsequent increase in transcription efficiency through increased initiation and/or RNApol II processivity. This feedback mechanism may be perceived as a homeostatic process, allowing maintenance of the mean mRNA production despite a perturbation in template integrity. It is important to note that the dynamic binding and unbinding of Apex1 triggers noise enhancement and mean maintenance more than the repair per-se. One implication of this is that other protein-DNA dynamic interactions may lead to unavoidable noise modulation through structural constraints like supercoiling. The strength of such modulation will depend on the kinetic rates of interaction.

4.6 Sensitivity analysis of TCR model (Model 5)

We have seen that a model incorporating Apex1 interaction with chromatin and an associated transcriptional amplification, can recapitulate an increase in noise without alteration of the mean number of mRNA produced by the Nanog gene expression system. We next asked how this behavior - increase in noise without dramatic modification of mean expression - is related to the dynamics of Apex1 binding and unbinding with chromatin. k_{repair} and $k_{incorpo}$ are the kinetic rates describing such interaction.

We conducted a phase plane analysis of the system mean and Fano factor for both k_{repair} and $k_{incorpo}$ (Figure 4.5A-B). We assume that the cooperativity is fixed and equal to the deduced cooperativity from the previous analysis using Nanog FISH data for $10\mu M$ IdU. As expected, the Fano factor increases as $k_{incorpo}$ increases (for $k_{incorpo}$ lower than ≈ 1). This suggests a positive dose-dependent relationship between IdU and noise (Figure 4.5B).

We observe an inverse relation for k_{repair}, where noise increases as k_{repair} decreases. Experimentally, we use a small-molecule inhibitor of the Apex1 endonuclease domain (CRT0044876) to decrease k_{repair}. It is interesting to highlight that when $k_{incorpo}$ is higher than ≈ 1 the Fano factor starts to decrease. These observations can be understood looking at equation (4.31) for the Fano factor. For $k_{incorpo} > 1$, ν_{on} decreases slowly and the effective $\langle k_{RNA} \rangle$ starts to increase slowly as compared to when $k_{incorpo} \in [0.1, 1]$ (Figure 4.5C-D). These changes are counteracted by a larger increase in K_{OFF}. The behavior for the mean number of RNA produced with increasing $k_{incorpo} > 1$ can also be understood using the previous considerations: the decrease in mean corresponds to a decrease of the frequency in the ON state that is not counteracted by a strong enough cooperativity. All the results can be understood using the following formula for the Fano factor:

$$FF = 1 + \frac{(1-\nu_{on}) \cdot k_{RNA}}{K_{ON} + K_{OFF} + k_{decay}} \tag{4.31}$$

We next wanted to define the parameter regime for k_{ON} and k_{OFF} in which homeostatic maintenance of mean expression is possible with transcription-coupled repair (model 5). For this analysis, simulations were run with values of $k_{ON}, k_{OFF} \in [10^{-3}, 10]$ for both the null model (standard 2-state model, DMSO condition) and model 5. For the same values of k_{ON}, k_{OFF}, the fold change in mean of mRNA counts was calculated by comparing results of model 5 to the null model. This provides insight into how IdU treatment may impact expression of genes with different bursting kinetics. When $k_{OFF} >> k_{ON}$, the addition of IdU in Model 5 increases the average number of mRNA produced as compared to the null model (Figure 4.5E). This can be explained by a competition between the OFF and ON^* states and by the fact that in this portion of the phase space $k_{ON} < k_{OFF} < k_{incorpo}$ (and $k_{incorpo} < k_{repair}$). Therefore, the probability of presence in the ON_2 state (transcriptionally more productive state), increases. This implies that IdU treatment would increase the mean of very lowly expressed genes. This was seen experimentally in bulk RNA-seq measurements of transcript abundance in mESCs as the ≈ 100 genes that showed an increase in mean with IdU treatment were from the lowest expression regime (Figure 2.3). The exact inverse effect is observed in the upper left corner of the heatmap, where $k_{ON} > k_{OFF} > k_{incorpo}$. Thus for highly expressed genes, the effect of IdU on mean expression is minimal as seen experimentally.

When all the kinetics rates of the system are fixed, increasing the cooperativity, and thus the effective transcription rate, leads both to an increase in mean and Fano factor (Figure 4.5F).

Sensitivity analysis of the Apex1 TCR model revealed that orthogonal modulation of Nanog mean and noise is possible within a large portion of the parameter space (Figures 4.5A-B). As validation, we tested the effect of 96 concentration combinations of IdU and CRT0044876 to perturb the rates of Apex1 binding and unbinding respectively. The experimental results confirmed

model predictions, showing that Nanog noise could be tuned independently of the mean (Figure 4.6A). Testing of BrdU and hmU further validated that parameter regimes exist where noise can be regulated independent of mean (Figures 4.6B-C). The hmU data in particular showed that the BER pathway can amplify noise while maintaining mean expression when removing a naturally occurring base modification.

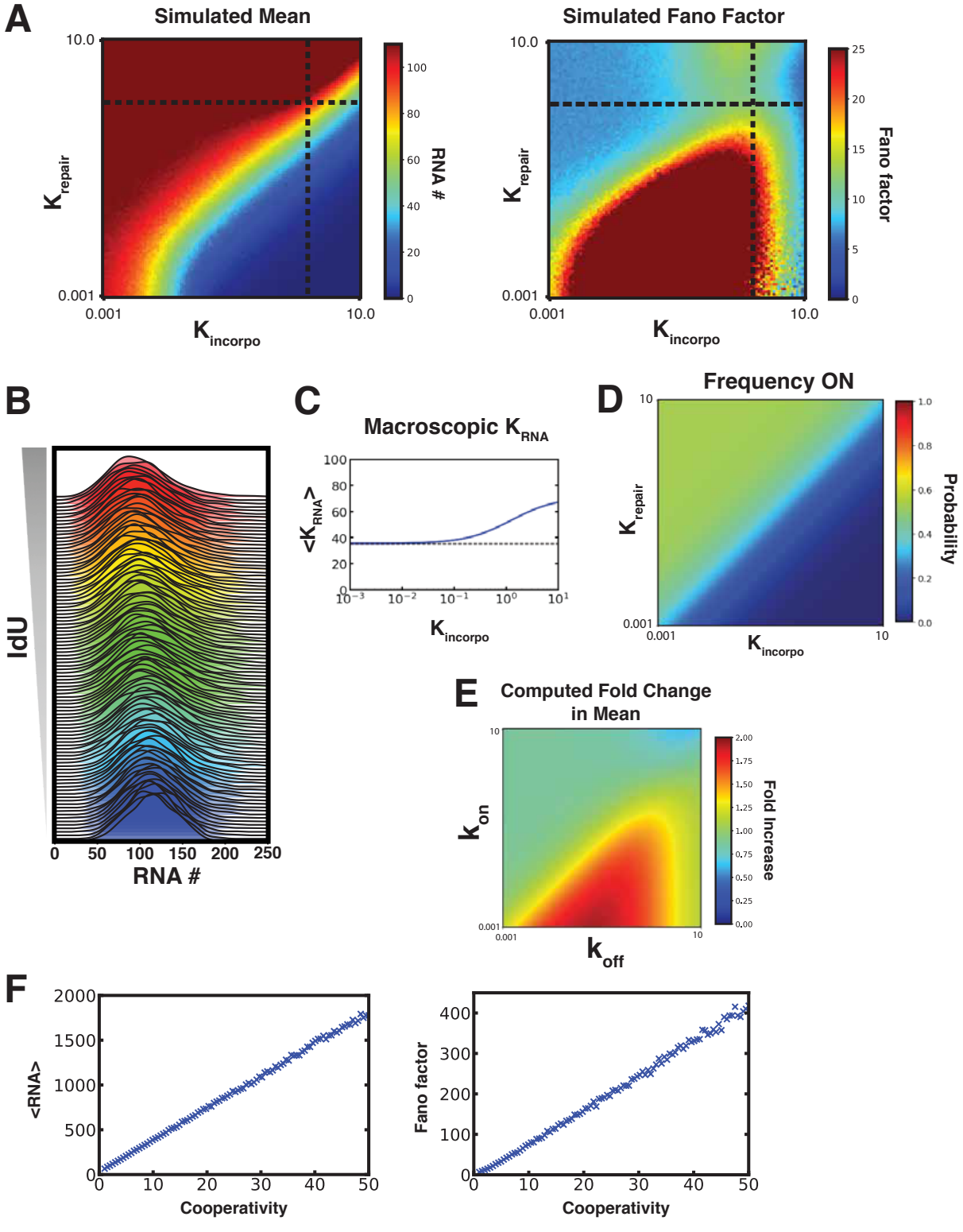
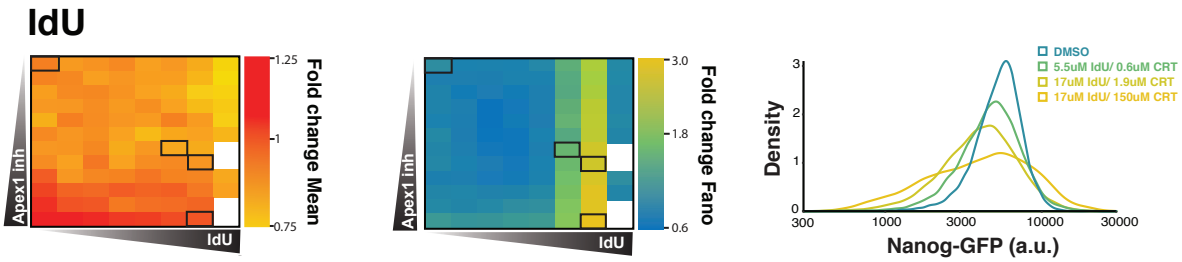
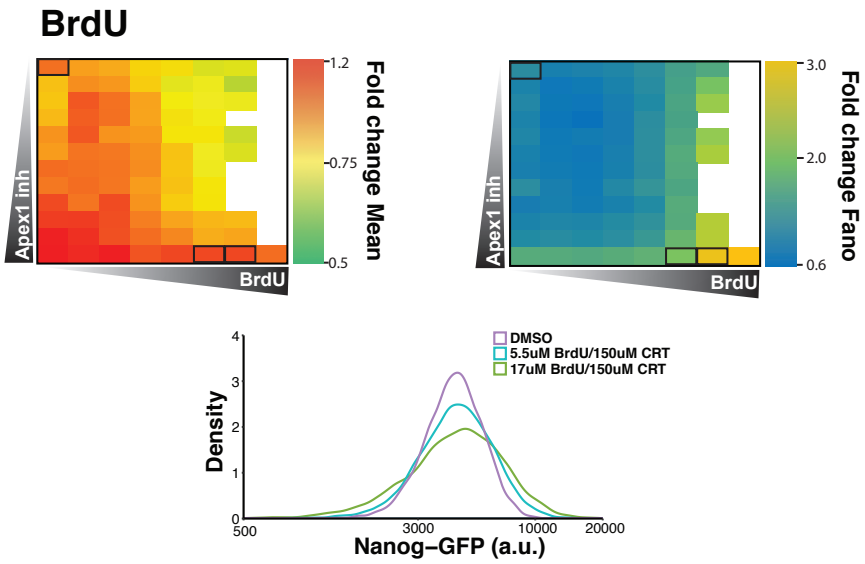

A
Simulated Mean
Simulated Fano Factor
K repair
K incorpo
RNA #
Fano factor
B
IdU
RNA #
C
Macroscopic K RNA
<K RNA>
K incorpo
D
Frequency ON
Probability
E
Computed Fold Change in Mean
k on
k off
Fold Increase
F
<RNA>
Cooperativity
Fano factor
Cooperativity

Figure 4.5: Sensitivity analysis of model parameters reveals phase-space for modulation of Nanog variability independently of mean.
(A) Heatmaps displaying mean (left) and Fano factor (right) of Nanog mRNA from simulation results of TCR model as a function of k_{repair} and $k_{incorpo}$ values spanning four orders of magnitude. Dashed horizontal and vertical lines represent inferred values of k_{repair} and $k_{incorpo}$ that match experimental Nanog gene expression system in the presence of 10μM IdU. Multiple regions of the parameter phase-space exhibit constant mean output with unique levels of variability (Fano factor) demonstrating how mean and variability are tuned independently. **(B)** Simulated distributions of Nanog mRNA with increasing concentration of IdU which increases $k_{incorpo}$. Simulation results demonstrate how TCR model allows for maintenance of mean output with increasing variability as concentration of IdU is increased. **(C)** Effective transcription rate of Nanog gene expression system as a function of $k_{incorpo}$. As IdU incorporation and subsequent Apex1 recruitment increases, the effective transcription rate increases as well. This represents the compensatory mechanism of model 5 allowing for maintenance of mean output with increasing incorporation of IdU. **(D)** Heatmap displaying fraction of time that the Nanog gene expression is in the macroscopic ON state as a function of k_{repair} and $k_{incorpo}$ values. **(E)** Heatmap displaying fold change in mean as a function of microscopic k_{off} and k_{on} values (supplementary text 6.2). Fold change is calculated as the output of Model 5 relative to model 0 (canonical 2-state model) for the same set of k_{off} and k_{on} values. For a gene whose $k_{off} >> k_{on}$, addition of IdU to the system increases the mean output. **(F)** Mean mRNA and Fano factor of Model 5 output as a function of the cooperativity term which describes how strongly the transcription rate is amplified following completion of repair.

A


IdU
Apex1 inh
IdU
1.25
1
0.75
Fold change Mean
Apex1 inh
IdU
3.0
1.8
0.6
Fold change Fano
DMSO
5.5uM IdU/ 0.6uM CRT
17uM IdU/ 1.9uM CRT
17uM IdU/ 150uM CRT
Density
Nanog-GFP (a.u.)
300
1000
3000
10000
30000

B


BrdU
Apex1 inh
BrdU
1.2
0.75
0.5
Fold change Mean
Apex1 inh
BrdU
3.0
2.0
0.6
Fold change Fano
DMSO
5.5uM BrdU/150uM CRT
17uM BrdU/150uM CRT
Density
Nanog–GFP (a.u.)
500
3000
10000
20000

C

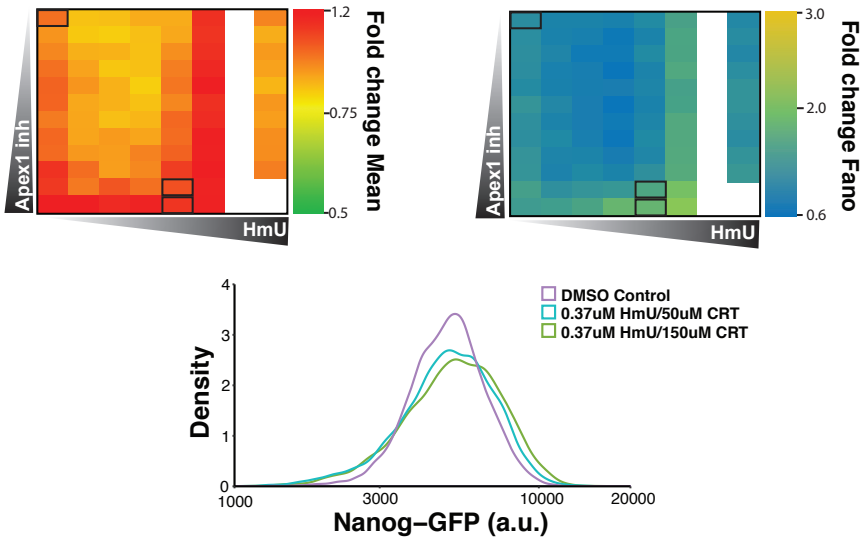

HmU
Apex1 inh
HmU
1.2
0.75
0.5
Fold change Mean
Apex1 inh
HmU
3.0
2.0
0.6
Fold change Fano
DMSO Control
0.37uM HmU/50uM CRT
0.37uM HmU/150uM CRT
Density
Nanog–GFP (a.u.)
1000
3000
10000
20000

Figure 4.6: Treatment with IdU, BrdU or HmU in combination with CRT0044876 allows for tuning of Nanog variability independently of the mean.
(A) Testing of 96 concentration combinations of IdU and CRT0044876 (apex1 endonuclease domain inhibitor) to validate tunability of Nanog variability. IdU and CRT0044876 were used to increase binding and decrease unbinding of Apex1 respectively. Nanog-GFP mESCs grown in 96-well plates were treated with 12 concentrations of CRT0044876 ranging from 0 to 150μM in combination with 8 concentrations of IdU ranging from 0 to 50μM. Data represent average of two biological replicates. (Leftmost and Center Panels) 96-well heatmaps displaying fold change in Nanog mean and Fano factor for each drug combination as compared to DMSO (top-leftmost well). Insufficient number of cells (¡50,000) for extrinsic noise filtering were recorded from white wells. (Rightmost Panel) Representative flow cytometry distributions from highlighted wells (black rectangles). Nanog variability increases independently of the mean. **(B)** Testing of 96 concentration combinations of BrdU and CRT0044876. Nanog-GFP mESCs grown in 96-well plates were treated with 12 concentrations of CRT0044876 ranging from 0 to 150uM in combination with 8 concentrations of BrdU ranging from 0 to 50uM. Data represent average of two biological replicates. (Top left and top right panels) 96-well heatmaps displaying fold change in Nanog mean and Fano factor for each drug combination as compared to DMSO (top-leftmost well). Insufficient number of cells (<50,000) for extrinsic noise filtering were recorded from white wells. (Bottom Panel) Representative flow cytometry distributions from highlighted wells (black rectangles). Nanog variability increases independently of the mean. **(C)** Testing of 96 concentration combinations of HmU and CRT0044876. HmU is a naturally found, Tet-induced oxidation product of thymine. Nanog-GFP mESCs grown in 96-well plates were treated with 12 concentrations of CRT0044876 ranging from 0 to 150uM in combination with 8 concentrations of HmU ranging from 0 to 10uM. Data represent average of two biological replicates. (Top left and top right panels) 96-well heatmaps displaying fold change in Nanog mean and Fano factor for each drug combination as compared to DMSO (top-leftmost well). Insufficient number of cells (¡50,000) for extrinsic noise filtering were recorded from white wells. (Bottom Panel) Representative flow cytometry distributions from highlighted wells (black rectangles). As with IdU and BrdU, Nanog variability increases independently of the mean.

Methods

96 dose combination for testing of noise phase space

Compound plates containing 96 concentration combinations of IdU, BrdU, or HmU with CRT0044876 were prepared by the Gladstone Assay Development and Drug Discovery Core using an Agilent Bravo liquid handling system. All wells contained equivalent volumes of DMSO. Compound mixtures were suspended in 200μL of 2i/LIF media. $1x10^4$ Nanog-GFP mESCs were seeded into each well of a gelatin-coated, 96-well dish in 200μL of 2i/LIF media. 24 hours after seeding, 100μL of media was removed from each well and 100μL of compound-containing 2i/LIF was added in replicate. IdU and BrdU concentrations ranged from 0 to 50μM while HmU concentrations ranged from 0 to 10μM. CRT0044876 ranged from 0 to 150μM. After 24 hours of treatment, cells were detached using TrypLE and plates were run on BD LSRFortessa high-throughput system. After extrinsic noise filtering via cell-size gating, Nanog mean and Fano factor for each treatment were normalized to DMSO control well. Reported fold changes in mean and Fano factor are the average of two replicates.

4.7 TCR model provides unifying mechanism for noise-enhancement of genes with different bursting kinetics.

In analyzing the Nanog gene expression we have found that IdU treatment leads to recruitment of DNA repair machinery while a gene is transcriptionally permissive (TCR model). This repair activity makes a second ON state accessible to the system. This state is characterized by an increased k_{RNA} which is sufficient to recapitulate the experimental observations of mean maintenance with increased noise strength (Fano factor). We next asked whether the inferred TCR model can explain the experimental data collected for other noise-enhanced genes within the scRNA-seq

dataset of mESCs treated with $10\mu M$ IdU for 24 hours.

To simulate the TCR model for additional genes, estimates for the following parameters are needed: $k_{ON}, k_{OFF}, k_{RNA1}$(basal transcription rate in DMSO), $k_{decay}, and \langle k_{RNA} \rangle$ (effective transcription rate in IdU). To derive estimates of $k_{ON}, k_{OFF}, k_{RNA1}, and \langle k_{RNA} \rangle$ we used a moments-matching technique described in [86, 87], where the first, second, and third exponential moments of the mRNA distributions in the DMSO and IdU conditions are used to calculate the parameters of a Poisson-beta distribution (describes 2-state model) that best fits experimental count data. The parameter estimates are derived in proportion to k_{decay}. Values of k_{decay} were retrieved from an existing dataset of mRNA degradation rates in mESCs [88].

Of the 945 genes classified as highly variable with IdU treatment, 314 genes remained for downstream analysis based on availability of mRNA degradation rates and high confidence parameter estimates (Figure 4.7A). A consistent pattern emerged for genes classified as highly variable: 80% exhibited increased rates of promoter inactivation (K_{OFF}) and 84% had increased transcription rates (K_{TX}) (Figures 4.7B-C).

With the above parameter estimates, we next computed the constraints on the cooperativity term and $k_{incorpo}$ as a function of k_{repair} using the relationships derived in section 4.2. There is again one remaining degree of freedom in our model system: k_{repair}. Using the MLE-based approach outlined in section 4.5, simulation results for a range of k_{repair} values were compared against scRNA-seq data to identify $\hat{k}_{repair,LLE}$ for each of the 314 genes.

Once $\hat{k}_{repair,LLE}$ was identified, the macroscopic values of K_{ON} and K_{OFF} from simulation results were compared to experimentally derived estimates of K_{ON} and K_{OFF} from scRNA-seq data for each gene (Figure 4.8). Overall, simulated values for macroscopic rates of promoter toggling in the IdU condition align with experimental results, indicating that the TCR model holds explana-

tory power for noise-enhanced genes beyond just Nanog. The use of a second, transcriptionally-enhanced ON state appears to be a unifying mechanism for maintenance of transcriptional homeostasis during DNA repair across a broad range of genes with different bursting kinetics. This suggests that the TCR mechanism for mean maintenance is robust to the initial bursting characteristics of a gene.

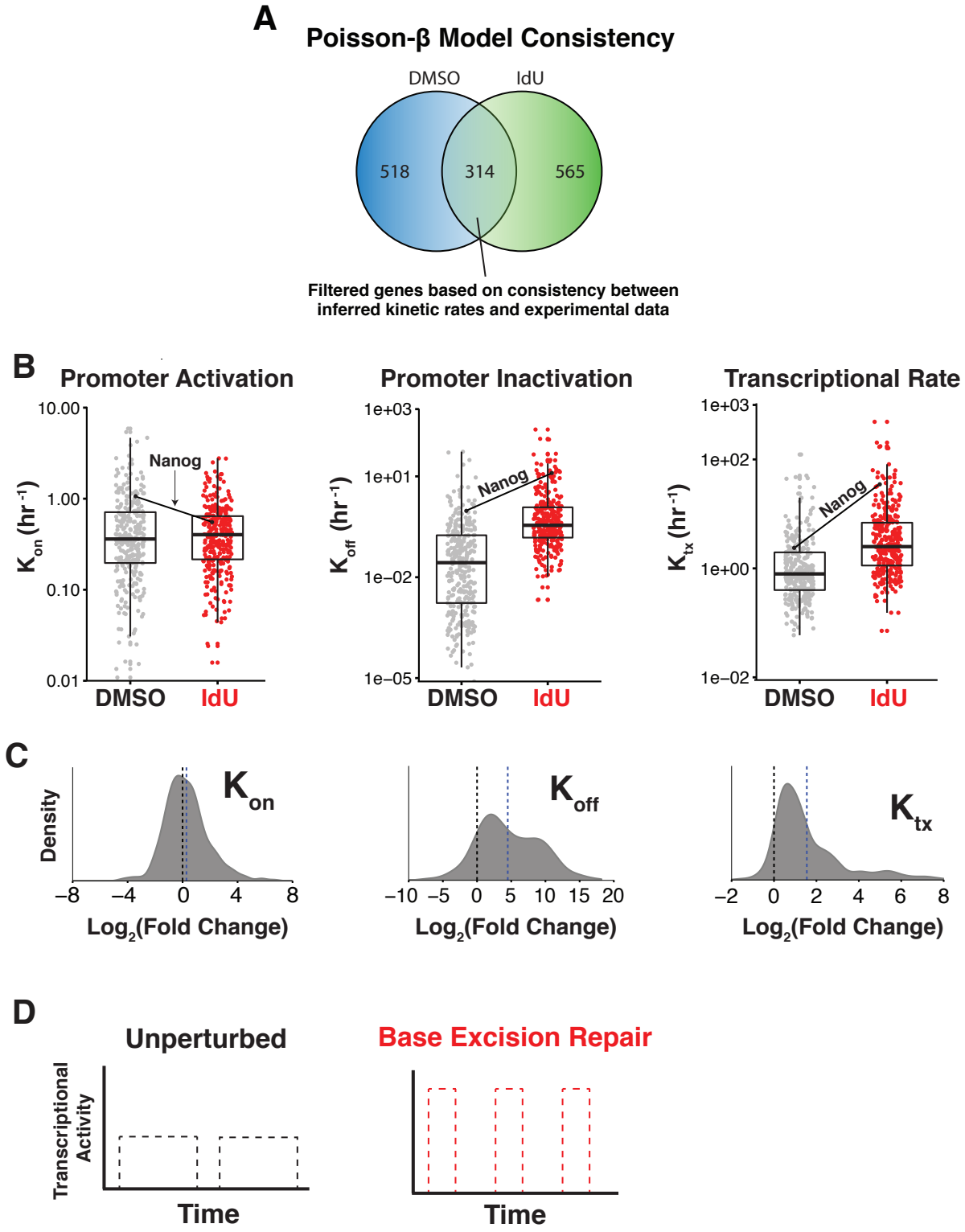

A
Poisson-β Model Consistency
DMSO
IdU
518
314
565
Filtered genes based on consistency between inferred kinetic rates and experimental data
B
Promoter Activation
Promoter Inactivation
Transcriptional Rate
Nanog
DMSO
IdU
C
Density
Log2(Fold Change)
D
Unperturbed
Base Excision Repair
Transcriptional Activity
Time

Figure 4.7: Highly variable genes exhibit shorter but more intense transcriptional bursts. **(A)** Of the 945 genes classified as highly variable with IdU treatment, we were able to estimate parameters of the 2-state model for 314 of these genes. **(B)** Boxplots show median ±interquartile range of parameter estimates with each point representing a gene. **(C)** Distributions of fold change in bursting kinetics between IdU and DMSO conditions for 314 highly variable genes. Dashed blue line signifies mean of distribution. Majority of highly variable genes exhibit increased K_{OFF} and K_{tx}, which is consistent with TCR model. **(D)** Base-excision repair orchestrates shorter but more intense transcriptional bursts to maintain mean expression for genes with diverse bursting kinetics.

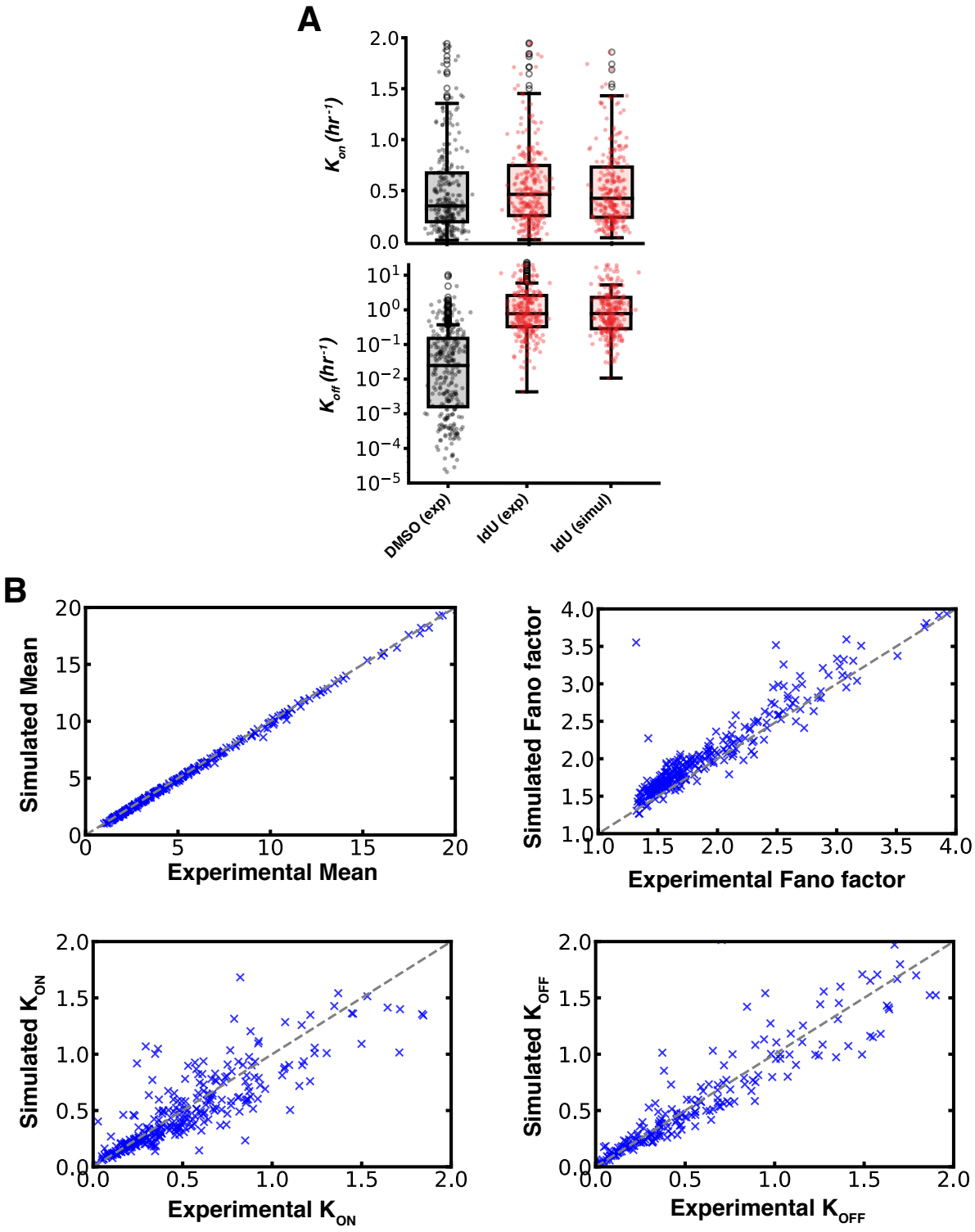

A
K_{on} (hr^{-1})
K_{off} (hr^{-1})
DMSO (exp)
IdU (exp)
IdU (simul)
B
Simulated Mean
Experimental Mean
Simulated Fano factor
Experimental Fano factor
Simulated K_{ON}
Experimental K_{ON}
Simulated K_{OFF}
Experimental K_{OFF}

Figure 4.8: TCR model provides unifying mechanism for noise-enhancement of genes with different bursting kinetics.
(A) Experimental values (exp) of macroscopic K_{ON} and K_{OFF}, as derived from a moments-matching technique applied to scRNA-seq data, are compared to predicted values (simul) derived from simulations of TCR model. Each point represents a gene. Boxplots show median $\pm$interquartile range of parameter values. **(B)** Experimental values of mean, Fano factor, K_{ON}, and K_{OFF} (based on scRNA-seq data) are compared to simulated values derived from TCR model.

Methods

Estimation of promoter toggling kinetics from scRNA-seq data

Gene expression data from the scRNA-seq dataset were fit to the 2-state model using the D3E algorithm, allowing for estimation of $k_{ON}, k_{OFF}, and k_{tx}$ in proportion to the rate of mRNA degradation which is the lone parameter that is not estimable from this dataset alone [87]. Parameter estimation was conducted using the methods of moments approach with the normalise and removeZeros options. Analysis was run for the 945 genes classified as highly variable according to the BASiCS algorithm. The Cramer-von Mises test was used for goodness-of-fit testing. Values of k_{decay} were then retrieved from an existing dataset of mRNA degradation rates in mESCs [88], with the assumption that degradation rates are unchanged between DMSO and IdU conditions. Parameter estimates were then verified against experimental values of mean mRNA counts using the following relationship: $\langle RNA \rangle = \frac{k_{ON}}{k_{ON}+k_{OFF}} \cdot \frac{k_{RNA}}{k_{decay}}$. Genes whose predicted mean was within 10% of experimental value were used for downstream analysis. 314 genes passed this filtering process based on availability of mRNA degradation rates and alignment of parameter estimates with expected mean mRNA counts.

Chapter 5

Homeostatic noise-amplification potentiates responsiveness to cell-fate signals

5.1 Results

5.1.1 Amplification of transcriptional fluctuations destabilizes cellular identity resulting in greater cellular plasticity.

We next asked if this amplification of transcriptional variability acted to enhance cellular plasticity as previously suggested [89]. Using a neural-network approach, we reconstructed the Waddington landscape based on a predictive model of gene-gene interactions inferred from the scRNA-seq data [90]. In this approach, each cell has a characteristic energy determined by its proximity to an attractor state, with lower energy values corresponding to greater stability. The analysis indicated that cells exposed to IdU lie at a higher altitude on this landscape, indicating destabilization of cellular identity and greater developmental plasticity (Figure 5.1).

Numerical simulations of the TCR model then verified that IdU-mediated amplification of transcriptional noise has the potential to increase responsiveness to activation stimuli (Figure 5.2). The

complementary abilities of IdU-mediated noise amplification to destabilize cellular identity and potentiate responsiveness to fate signals suggested that IdU might facilitate cellular reprogramming.

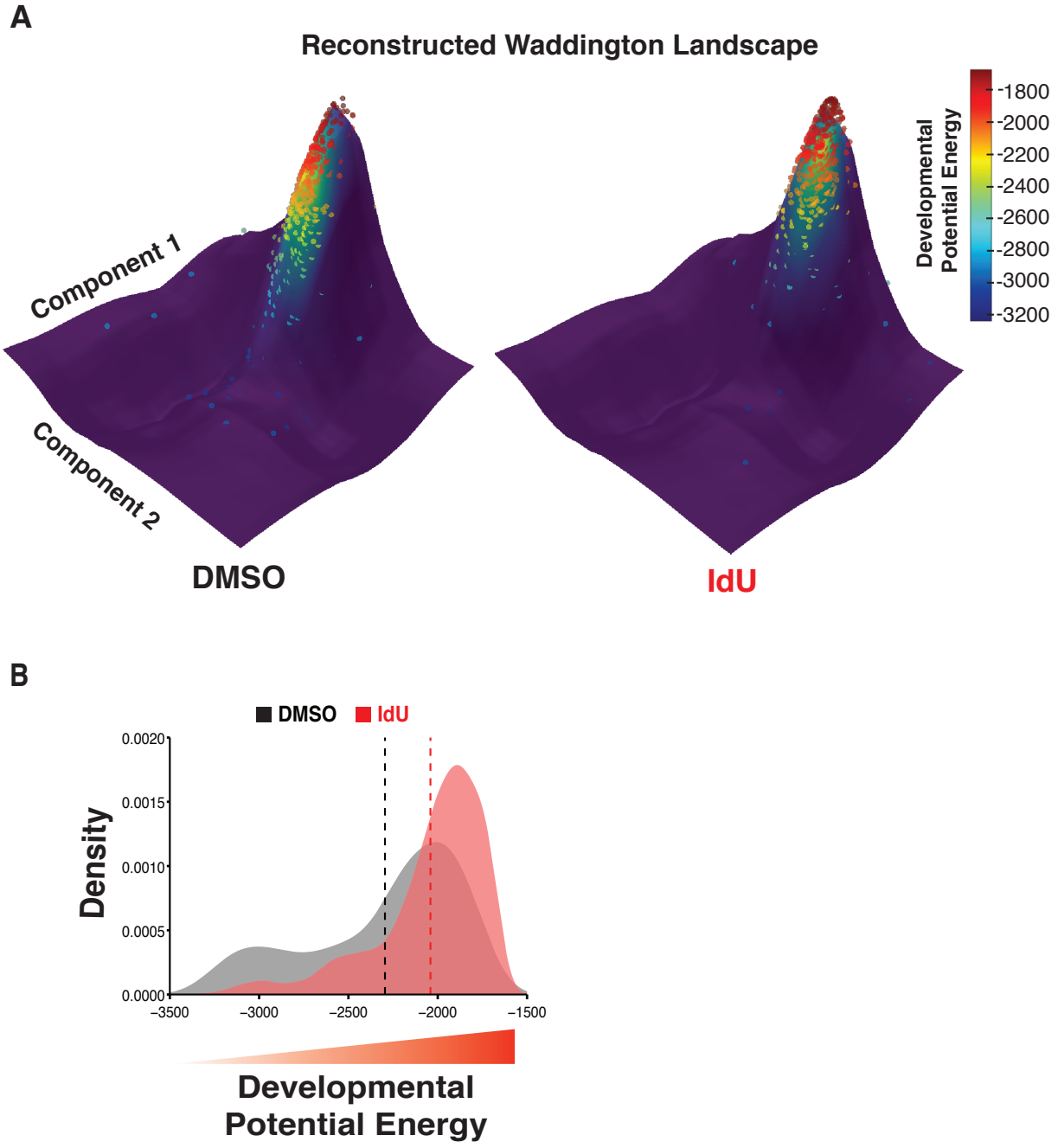

A
Reconstructed Waddington Landscape
Component 1
Component 2
DMSO
IdU
Developmental Potential Energy
-1800
-2000
-2200
-2400
-2600
-2800
-3000
-3200
B
DMSO
IdU
Density
0.0020
0.0015
0.0010
0.0005
0.0000
−3500
−3000
−2500
−2000
−1500
Developmental Potential Energy

Figure 5.1: Amplification of transcriptional fluctuations destabilizes cellular identity resulting in greater cellular plasticity.
(A) Reconstruction of Waddington's landscape using scRNA-seq data of mESCs treated with DMSO (left) or IdU (right). Each point represents a cell. A Gaussian process latent variable model (GP-LVM) was used for dimensionality reduction to create a 2-D map of cell clustering, represented by component 1 (y-axis) and component 2 (x-axis). The z-axis represents the calculated potential energy (distance from an attractor) of a cell's gene expression state with lower values indicating greater proximity to an attractor and thus lower developmental potential. Cells are colored according to their height (developmental potential energy, z-axis value) on the landscape as denoted by the associated color bar. Underlying shading of landscape represents density of points with purple being the least dense and yellow being the most dense. **(B)** Distributions of developmental potential energy (z-axis values from Waddington landscape in panel A) for mESCs treated with DMSO or 10μM IdU for 24 hours. Dashed vertical lines signify the mean of each distribution, with IdU-treated cells demonstrating greater potential energies.

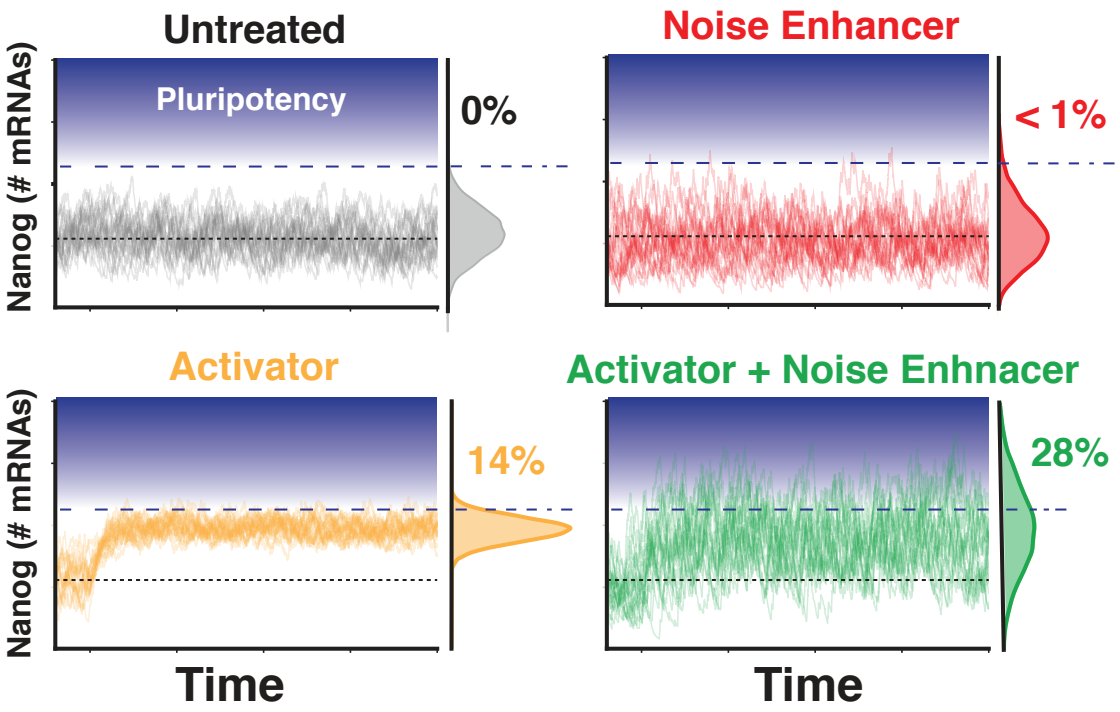


Figure 5.2: Homeostatic noise-amplification synergizes with canonical activators of gene expression to increase threshold crossing.
Simulations of the TCR model for Nanog gene expression in the presence of DMSO (top left), IdU (top right), an activator (increased KON, decreased KOFF) of promoter activity (bottom left) and an activator combined with IdU (bottom right). Homeostatic noise amplification potentiates responsiveness to an activator of gene expression as demonstrated by increased threshold crossing (14% to 28%).

Methods

Estimation of cellular developmental potential and Waddington Landscape reconstruction

The HopLand algorithm (continuous Hopfield network) was used to create a predictive model of gene-gene interactions from scRNA-seq data [90]. As input, raw count data for a total of 800 randomly chosen cells (400 from DMSO and 400 from IdU treatment groups) were used. Each neuron of the network corresponds to a gene. Genes whose variance fell within in the top 10% were used for construction of a neural network resulting in 512 nodes. The weight matrix describing pair-wise interactions between nodes was initialized using the gene-gene Pearson correlation matrix. A Gaussian process latent variable model (GP-LVM) was used for dimensionality reduction to create a 2-D map of cell clustering (x and y coordinates on Waddington Landscape). Energy values (z coordinate on landscape) were calculated using the Lyapunov function which is a measure of stability. Lower energy values indicate greater proximity to an equilibrium point (attractor state) and thus less developmental potential.

5.1.2 IdU treatment potentiates reprogramming of mouse embryonic fibroblasts into pluripotent stem cells

To experimentally verify these predictions, we tested if IdU could potentiate conversion of differentiated cells into pluripotent stem cells using two cellular reprogramming systems. The first assay utilized mouse embryonic fibroblasts (MEFs) that express GFP from the endogenous Nanog locus and harbor stably integrated, doxycycline-inducible cassettes for three of the Yamanaka factors: Oct4, Sox2, and Klf4 (OSK). As confirmation that IdU acts as a noise-enhancer in this system, treatment of secondary MEFs with IdU for 48 hours in standard MEF media caused increased variability in Nanog protein expression (Figure 5.3A) with no changes in cell-cycle progression (Figure 5.3B). Strikingly, IdU supplementation for the first 48 hours of a 10-day reprogramming course

enhanced the formation of pluripotent colonies as measured by alkaline phosphatase staining (Figure 5.4A). Bulk RNA-seq at days 2 and 5 of reprogramming (Figure 5.4B) and flow-cytometric analysis at day 10 (Figure 5.4C) demonstrate that early-stage noise-enhancement accelerates activation of the pluripotency program. To confirm the results in an orthogonal reprogramming assay, Oct4-GFP primary MEFs were transduced with retroviral vectors expressing Oct4, Sox2, Klf4, and c-Myc. IdU supplementation for the 48 hours immediately following transduction caused a 2.4-fold increase in the number of Oct4-GFP(+) colonies (Figure 5.4D), further demonstrating how amplification of intrinsic gene expression fluctuations can potentiate cell-fate conversion.

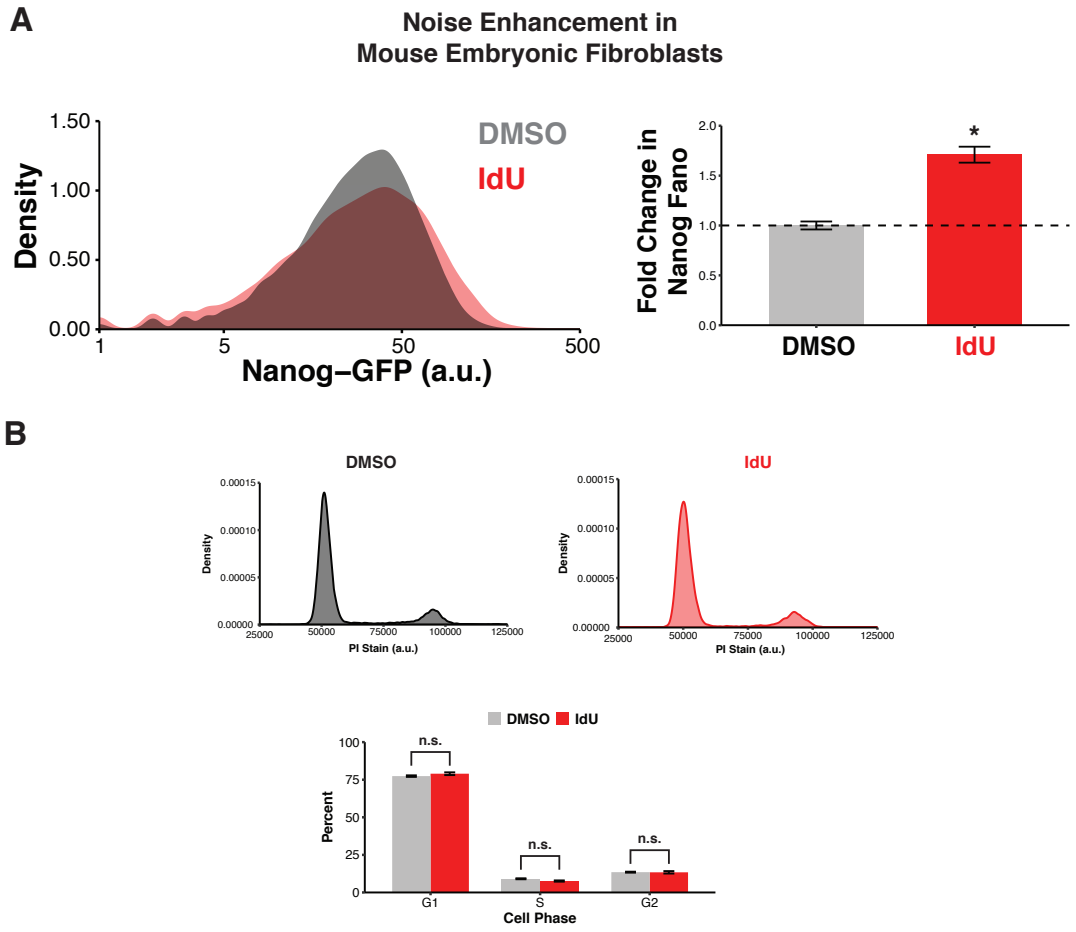


Figure 5.3: IdU treatment enhances Nanog expression noise in mouse embryonic fibroblasts (MEFs).

(A) Secondary MEFs with the endogenous Nanog locus tagged with GFP were treated with 4μM IdU or equivalent volume DMSO for 48 hours in MEF media. (Right) Representative flow cytometry distributions of Nanog-GFP expression in secondary MEFs after 48 hour treatment with IdU or DMSO. (Left) Quantification of Nanog Fano factor demonstrates that IdU treatment increases expression variability as compared to DMSO control (*p = 0.003, by a two-tailed, unpaired Student's t test). Data represent mean and SD of three biological replicates. **(B)** (Top) Representative flow cytometry distributions of propidium iodide staining for Nanog-GFP secondary MEFs treated with either DMSO or 4μM IdU for 48 hours in MEF media. No signs of aneuploidy are visible, indicating Nanog expression variability is not due to cell-to-cell variability in gene copy numbers. (Bottom) Percent of cells in each phase of the cell cycle for DMSO and IdU treatments based on propidium iodide staining. IdU treatment does not alter cell-cycle progression, indicating enhanced reprogramming is not due to accelerated cellular division. Data represent mean and SD of three biological replicates. P values were calculated using a two-tailed, unpaired Student's t test.

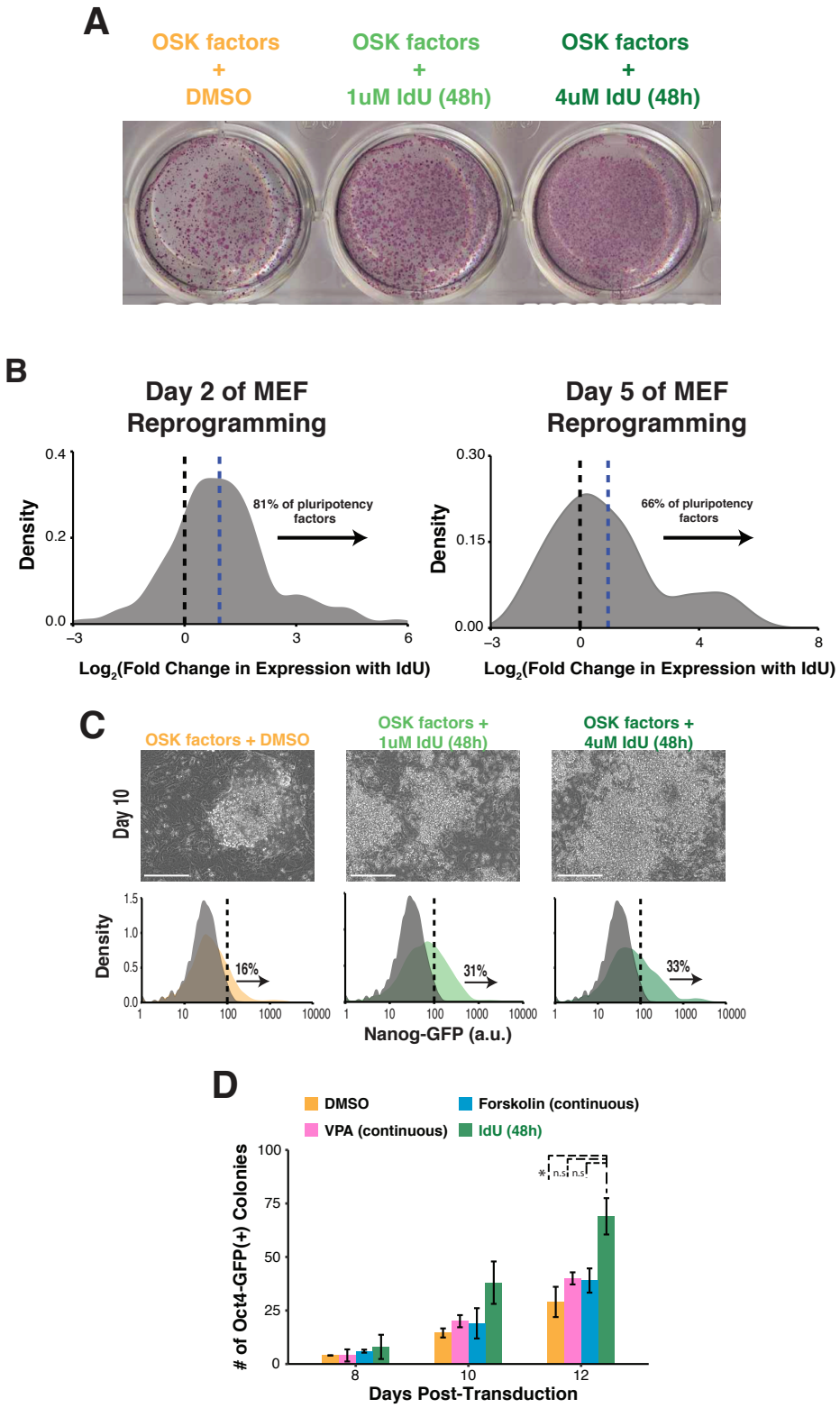

A
OSK factors + DMSO
OSK factors + 1uM IdU (48h)
OSK factors + 4uM IdU (48h)
B
Day 2 of MEF Reprogramming
Day 5 of MEF Reprogramming
Density
81% of pluripotency factors
66% of pluripotency factors
Log2(Fold Change in Expression with IdU)
C
OSK factors + DMSO
OSK factors + 1uM IdU (48h)
OSK factors + 4uM IdU (48h)
Day 10
Density
16%
31%
33%
Nanog-GFP (a.u.)
D
DMSO
Forskolin (continuous)
VPA (continuous)
IdU (48h)
# of Oct4-GFP(+) Colonies
n.s.
Days Post-Transduction

Figure 5.4: IdU treatment enhances conversion of MEFs into induced pluripotent stem cells (iPSCs).
(A) Nanog-GFP secondary MEFs (seeded at 10,000 cells/cm^2) harboring stably-integrated, doxycycline-inducible cassettes for Oct4, Sox2, and Klf4 (OSK) were subjected to 10 days of doxycycline treatment in combination with DMSO (first well), 1μM IdU (second well), or 4μM IdU (third well) for the first 48 hours of reprogramming. Alkaline phosphatase staining for pluripotent colonies of cells demonstrates how IdU treatment potentiates pluripotency induction. **(B)** Bulk RNA-seq was conducted on days 2 and 5 of doxycycline-induced reprogramming of secondary MEFs supplemented with 4μM IdU or equivalent volume DMSO for the first 48 hours. Distributions of fold change in expression for 129 pluripotency genes (taken from Mouse Genome Informatics, gene ontology term: 0019827) in the IdU condition as compared to the DMSO control are shown. Dashed blue line represents mean of distribution. 81% and 66% of the pluripotency factors show increased expression with the addition of IdU as compared to DMSO control at days 2 and 5 of reprogramming, respectively. Noise amplification during early stages of reprogramming accelerates activation of pluripotency network. **(C)** (Top) Micrographs of Nanog-GFP secondary MEFs at day 10 of doxycycline-induced reprogramming (scale bar = 100 μm). (Bottom) Flow cytometric analysis of Nanog-GFP activation at day 10 of reprogramming. Data are pooled from two replicates. **(D)** Oct4-GFP primary MEFs (seeded at 10,000 cells/cm^2) were retrovirally transduced with cDNAs encoding Oct4, Sox2, Klf4, and c-Myc. 24 hours after transduction, infected cells were treated with DMSO (continuously), 1mM valproic acid (VPA, continuously), 10μM forskolin (continuously), or 4μM IdU (first 48 hours). VPA and forskolin are established enhancers of cellular reprogramming. The number of Oct4-GFP(+) stem cell colonies were counted 8, 10, and 12 days from the start of drug treatment. Data represent mean and SD of 2 biological replicates. Treatment of transduced MEFs with IdU during early stages of reprogramming increases the number of Oct4-GFP(+) colonies that form as compared to DMSO control, *p =0.039 by one-way ANOVA with Bonferroni post hoc test.

Methods

Cellular reprogramming assays

Two cellular reprogramming systems were tested in this study: (1) Nanog-GFP secondary mouse embryonic fibroblasts (MEFs) harboring stably integrated, doxycycline-inducible cassettes for Oct4, Sox2, and Klf4. GFP is expressed from the endogenous Nanog locus. (2) Oct4-GFP primary MEFs that express GFP from the endogenous Oct4 locus.

Secondary MEFs were seeded onto gelatin-coated, 12-well plates at a density of 10,000 cells/cm^2 in MEF medium (DMEM supplemented with 10% FBS and 0.1mM non-essential amino acid, and 2mM Glutamax). 24 hours after seeding, wells were washed with DPBS and media was switched to ESC media (knockout DMEM, 10% FBS, 10% KSR, 2mM Glutamax, 0.1mM non-essential amino acid, 0.1mM 2-mercaptoethanol, 10^3 units/ml leukemia inhibitory factor) supplemented with 1μg/ml doxycycline. Additionally, IdU (1uM or 4uM) or equivalent volume DMSO (Day 0) were added to media. 48 hours after the start of IdU treatment, wells were washed with DPBS and media was replaced with ESC media supplemented with 1μg/ml doxycycline alone. Media was refreshed every other day until day 10 of reprogramming. Alkaline phosphatase staining was performed according to manufacturer's instructions using the Alkaline Phosphatase Detected Kit (Millipore). For flow cytometric analysis of Nanog-GFP expression, cells were dissociated with TrypLE and run unfixed on BD FACS Calibur cytometer.

Oct4-GFP primary MEFs were transduced with lentiviral vectors encoding Oct4, Sox2, Klf4, and c-Myc. Lentiviruses encoding these factors were individually packaged in PLAT-E cells (ATCC) using pMX-based vectors. 48 hours after transfection of lentiviral vectors, viral supernatant was collected and filtered. For infection, Oct4-GFP primary MEFs were seeded on gelatin-coated, 6-well plates at a density of 10,000 cells/cm^2 in MEF medium 24 hours prior to transduction (Day -2). Oct4, Sox2, Klf4, and c-Myc viruses were mixed in equal volume along with 5μg/ml

polybrene and incubated with primary MEFs for 24 hours in MEF medium (Day -1). Following infection, wells were washed with ESC media and cells were incubated with ESC media supplemented with 10μM Forskolin, 1mM Valproic Acid, 4μM IdU or equivalent volume DMSO (Day 0). ESC media was refreshed every other day. IdU supplementation was discontinued after 48 hours while Forskolin and Valproic Acid were kept in media continuously. Oct4-GFP(+) colonies were counted on days 8, 10 and 12.

Bulk RNA-seq of secondary MEFs undergoing reprogramming

Secondary MEFs were seeded onto gelatin-coated, 6-well plates at a density of 10,000 cells/cm^2 in MEF medium. For each timepoint (2- and 5-day), 4 wells were seeded (2 replicates for standard reprogramming and 2 replicates for IdU-assisted reprogramming). 24 hours after seeding, wells were washed with DPBS and media was switched to ESC media supplemented with 1μg/ml doxycycline. Additionally, 4μM IdU or equivalent volume DMSO (Day 0) were added to media. 48 hours after the start of reprogramming, cells for the 2-day timepoint in DMSO and IdU conditions were dissociated with TrypLE, pelleted, and snap frozen with liquid nitrogen. Media in the wells for the 5-day timepoint was refreshed with ESC media supplemented with 1μg/ml doxycycline alone. This was repeated on day 4. On day 5, remaining cells were dissociated and frozen identically to that of the 2-day timepoint.

RNA was extracted from each cell pellet using a RNeasy minikit (Qiagen) according to manufacturer's instructions. A total of 8 cDNA libraries were prepared with an NEBNext Ultra II RNA Library Prep kit (NEB, cat:E7770S) and sequenced with an Illumina HiSeq4000. Sequencing yielded a median of $\sim$50 million single-end reads per library. Read quality was checked via FASTQC. Reads were aligned to the mm10 reference genome using TopHat with default parameters. Transcript level quantification was performed using Cufflinks with default parameters.

5.2 Conclusions

Overall, these data reveal that a DNA-surveillance pathway exploits the biomechanical link between supercoiling and transcription to homeostatically enhance noise without altering mean-expression levels. This homeostatic noise-without-mean amplification appears to increase cellular plasticity, thus facilitating reprogramming of cellular identity. This raises intriguing implications for the role of naturally occurring oxidized nucleobases (e.g., hmU) in cell-fate determination, particularly since these base modifications are found at higher frequencies in embryonic stem-cell DNA [66]. Mechanistic insight from modeling and experimental perturbation of Apex1 suggest that homeostatic (i.e., orthogonal) noise amplification may also apply to other DNA-processing activities that interrupt transcription. It is important to note that homeostatic noise amplification cannot occur for all promoters (i.e., promoters with $K_{OFF} >> K_{ON}$ are precluded as they will exhibit increased mean) and propagation of transcriptional variability to the protein level likely depends on protein half-lives and thus may not occur for a large swath of proteins. The proteins monitored in this study either have naturally short half-lives (Nanog) or PEST tags (e.g. d_2GFP) which minimizes the buffering of transcriptional bursts conferred by longer protein half-lives [91]. The ability to independently control the mean and variance of gene expression may indicate that cells have the ability to amplify transcriptional noise for fate exploration and specification.

Bibliography

[1] Ludwig Boltzmann. Weitere studien uber das wärmegleichgewicht unter gasmolekulen, sitzungsber. kais. akad. wiss. wien math. *Naturwiss*, 66:275–370, 1872.

[2] Svante Arrhenius. Über die reaktionsgeschwindigkeit bei der inversion von rohrzucker durch säuren. *Zeitschrift für Physikalische Chemie*, 4:226, 1889.

[3] L. Roberts. Picture coding using pseudo-random noise. *IEEE Transactions on Information Theory*, 8(2):145–154, 1962.

[4] P. Fatt and B. Katz. Some observations on biological noise. *Nature*, 166(4223):597–598, 1950.

[5] Attila A. Priplata, James B. Niemi, Jason D. Harry, Lewis A. Lipsitz, and James J. Collins. Vibrating insoles and balance control in elderly people. *The Lancet*, 362(9390):1123–1124, 2003.

[6] Attila A. Priplata, Benjamin L. Patritti, James B. Niemi, Richard Hughes, Denise C. Gravelle, Lewis A. Lipsitz, Aristidis Veves, Joel Stein, Paolo Bonato, and James J. Collins. Noise-enhanced balance control in patients with diabetes and patients with stroke. *Annals of Neurology*, 59(1):4–12, 2006.

[7] Atefeh Aboutorabi, Mokhtar Arazpour, Mahmood Bahramizadeh, Farzam Farahmand, and Reza Fadayevatan. Effect of vibration on postural control and gait of elderly subjects: a systematic review. *Aging Clinical and Experimental Research*, 30(7):713–726, 2018.

[8] Dan Cohen. Optimizing reproduction in a randomly varying environment. *Journal of Theoretical Biology*, 12(1):119–129, 1966.

[9] Sankar Adhya and Max Gottesman. Promoter occlusion: Transcription through a promoter may inhibit its activity. *Cell*, 29(3):939–944, 1982.

[10] John L. Spudich and D. E. Koshland. Non-genetic individuality: chance in the single cell. *Nature*, 262(5568):467–471, 1976.

[11] R. Amini, A. A. Labudina, and C. Norden. Stochastic single cell migration leads to robust horizontal cell layer formation in the vertebrate retina. *Development*, 146(12), 2019.

[12] J. P. Bergman, M. J. Bovyn, F. F. Doval, A. Sharma, M. V. Gudheti, S. P. Gross, J. F. Allard, and M. D. Vershinin. Cargo navigation across 3d microtubule intersections. *Proc Natl Acad Sci U S A*, 115(3):537–542, 2018.

[13] J. Wang, P. Jenjaroenpun, A. Bhinge, V. E. Angarica, A. Del Sol, I. Nookaew, V. A. Kuznetsov, and L. W. Stanton. Single-cell gene expression analysis reveals regulators of distinct cell subpopulations among developing human neurons. *Genome Res*, 27(11):1783–1794, 2017.

[14] Peter S. Swain, Michael B. Elowitz, and Eric D. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12795–12800, 2002.

[15] M. Thattai and A. Van Oudenaarden. Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences*, 98(15):8614–8619, 2001.

[16] Long Cai, Nir Friedman, and X. Sunney Xie. Stochastic protein expression in individual cells at the single molecule level. *Nature*, 440(7082):358–362, 2006.

[17] Ido Golding, Johan Paulsson, Scott M. Zawilski, and Edward C. Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–1036, 2005.

[18] Arjun Raj, Charles S. Peskin, Daniel Tranchina, Diana Y. Vargas, and Sanjay Tyagi. Stochastic mrna synthesis in mammalian cells. *PLoS Biology*, 4(10):e309, 2006.

[19] Mads Kærn, Timothy C. Elston, William J. Blake, and James J. Collins. Stochasticity in gene expression: From theories to phenotypes. *Nature Reviews Genetics*, 6(6):451–464, 2005.

[20] Roy D. Dar, Brandon S. Razooky, Abhyudai Singh, V. Thomas Trimeloni, James M. McCollum, Chris D. Cox, Michael L. Simpson, and Leor S. Weinberger. Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, 109(43):17454–17459, 2012.

[21] David M. Suter, Nacho Molina, David Gatfield, Kim Schneider, Ueli Schibler, and Felix Naef. Mammalian genes are transcribed with widely different bursting kinetics. *Science*, 332(6028):472–474, 2011.

[22] Daniel Zenklusen, Daniel R. Larson, and Robert H. Singer. Single-rna counting reveals alternative modes of gene expression in yeast. *Nature Structural and Molecular Biology*, 15(12):1263–1271, 2008.

[23] Shasha Chong, Chongyi Chen, Hao Ge, and X. Sunney Xie. Mechanism of transcriptional bursting in bacteria. *Cell*, 158(2):314–326, 2014.

[24] Joseph Rodriguez and Daniel R. Larson. Transcription in living cells: Molecular mechanisms of bursting. *Annual Review of Biochemistry*, 89(1), 2020.

[25] William J. Blake, Mads Kærn, Charles R. Cantor, and J. J. Collins. Noise in eukaryotic gene expression. *Nature*, 422(6932):633–637, 2003.

[26] Jonathan R. Chubb, Tatjana Trcek, Shailesh M. Shenoy, and Robert H. Singer. Transcriptional pulsing of a developmental gene. *Current Biology*, 16(10):1018–1025, 2006.

[27] B. Munsky, G. Neuert, and A. Van Oudenaarden. Using gene expression noise to understand gene regulation. *Science*, 336(6078):183–187, 2012.

[28] Jean Peccoud and Bernard Ycart. Markovian modeling of gene-product synthesis. *Theoretical Population Biology*, 48(2):222–234, 1995.

[29] Damien Nicolas, Nick E. Phillips, and Felix Naef. What shapes eukaryotic transcriptional bursting? *Molecular BioSystems*, 13(7):1280–1290, 2017.

[30] John R. S. Newman, Sina Ghaemmaghami, Jan Ihmels, David K. Breslow, Matthew Noble, Joseph L. DeRisi, and Jonathan S. Weissman. Single-cell proteomic analysis of s. cerevisiae reveals the architecture of biological noise. *Nature*, 441(7095):840–846, 2006.

[31] Arren Bar-Even, Johan Paulsson, Narendra Maheshri, Miri Carmi, Erin O'Shea, Yitzhak Pilpel, and Naama Barkai. Noise in protein expression scales with natural protein abundance. *Nature Genetics*, 38(6):636–643, 2006.

[32] Roy D. Dar, Sydney M. Shaffer, Abhyudai Singh, Brandon S. Razooky, Michael L. Simpson, Arjun Raj, and Leor S. Weinberger. Transcriptional bursting explains the noise–versus–mean relationship in mrna and protein levels. *PLOS ONE*, 11(7):e0158298, 2016.

[33] Uri Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461, 2007.

[34] Roy D. Dar, Nina N. Hosmane, Michelle R. Arkin, Robert F. Siliciano, and Leor S. Weinberger. Screening for noise in gene expression identifies drug synergies. *Science*, 344(6190):1392–1396, 2014.

[35] M. M. K. Hansen, W. Y. Wen, E. Ingerman, B. S. Razooky, C. E. Thompson, R. D. Dar, C. W. Chin, M. L. Simpson, and L. S. Weinberger. A post-transcriptional feedback mechanism for noise suppression and fate stabilization. *Cell*, 173(7):1609–1621 e15, 2018.

[36] Yan Li, Yongli Shan, Ravi V. Desai, Kimberly H. Cox, Leor S. Weinberger, and Joseph S. Takahashi. Noise-driven cellular heterogeneity in circadian periodicity. *Proceedings of the National Academy of Sciences*, 117(19):10350–10356, 2020.

[37] Purnananda Guptasarma. Cooperative relaxation of supercoils and periodic transcriptional initiation within polymerase batteries. *BioEssays*, 18(4):325–332, 1996.

[38] Stuart A. Sevier, David A. Kessler, and Herbert Levine. Mechanical bounds to transcriptional noise. *Proceedings of the National Academy of Sciences of the United States of America*, 113(49):13983–13988, 2016.

[39] Sangjin Kim, Bruno Beltran, Irnov Irnov, and Christine Jacobs-Wagner. Long-distance cooperative and antagonistic rna polymerase dynamics via dna supercoiling. *Cell*, 179(1):106–119.e16, 2019.

[40] L. F. Liu and J. C. Wang. Supercoiling of the dna template during transcription. *Proceedings of the National Academy of Sciences*, 84(20):7024–7027, 1987.

[41] H. S. Koo, L. Claassen, L. Grossman, and L. F. Liu. Atp-dependent partitioning of the dna template into supercoiled domains by escherichia coli uvrab. *Proceedings of the National Academy of Sciences*, 88(4):1212–1216, 1991.

[42] Jie Ma, Lu Bai, and Michelle D. Wang. Transcription under torsion. *Science*, 340(6140):1580–1583, 2013.

[43] Yoshiko Hirota and Takashi Ohyama. Adjacent upstream superhelical writhe influences an escherichia coli promoter as measured by in vivo strength and in vitro open complex formation. *Journal of Molecular Biology*, 254(4):566–578, 1995.

[44] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420, 2018.

[45] Maike M. K. Hansen, V. Ravi Desai, Michael L. Simpson, and Leor S. Weinberger. Cytoplasmic amplification of transcriptional noise generates substantial cell-to-cell variability. *Cell Systems*, 7(4):384–397.e6, 2018.

[46] Alvaro Sanchez and Ido Golding. Genetic determinants and cellular constraints in noisy gene expression. *Science*, 342(6163):1188–1193, 2013.

[47] Catalina A. Vallejos, John C. Marioni, and Sylvia Richardson. Basics: Bayesian analysis of single-cell sequencing data. *PLOS Computational Biology*, 11(6):e1004333, 2015.

[48] Nils Eling, Arianne C. Richard, Sylvia Richardson, John C. Marioni, and Catalina A. Vallejos. Correcting the mean-variance dependency for differential variability testing using single-cell rna sequencing data. *Cell Systems*, 7(3):284–294.e12, 2018.

[49] Elphège P. Nora, Anton Goloborodko, Anne Laure Valton, Johan H. Gibcus, Alec Uebersohn, Nezar Abdennur, Job Dekker, Leonid A. Mirny, and Benoit G. Bruneau. Targeted degradation of ctcf decouples local insulation of chromosome domains from genomic compartmentalization. *Cell*, 2017.

[50] Antonio Scialdone, Kedar N. Natarajan, Luis R. Saraiva, Valentina Proserpio, Sarah A. Teichmann, Oliver Stegle, John C. Marioni, and Florian Buettner. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods*, 85:54–61, 2015.

[51] Philipp Angerer, Laleh Haghverdi, Maren Büttner, Fabian J. Theis, Carsten Marr, and Florian Buettner. Destiny: Diffusion maps for large-scale single-cell data in r. *Bioinformatics*, 32(8):1241–1243, 2016.

[52] Sara Hooshangi, Stephan Thiberge, and Ron Weiss. Ultrasensitivity and noise propagation in a synthetic transcriptional cascade. *Proceedings of the National Academy of Sciences of the United States of America*, 102(10):3581–3586, 2005.

[53] Juan H. Pedraza and Alexander Van Oudenaarden. Noise propagations in gene networks. *Science*, 307(5717):1965–1969, 2005.

[54] Mitra Mojtahedi, Alexander Skupin, Joseph Zhou, Ivan G. Castaño, Rebecca Y. Y. Leong-Quong, Hannah Chang, Kalliopi Trachana, Alessandro Giuliani, and Sui Huang. Cell fate decision as high-dimensional critical state transition. *PLoS Biology*, 14(12), 2016.

[55] Rhishikesh Bargaje, Kalliopi Trachana, Martin N. Shelton, Christopher S. McGinnis, Joseph X. Zhou, Cora Chadick, Savannah Cook, Christopher Cavanaugh, Sui Huang, and Leroy Hood. Cell population structure prior to bifurcation predicts efficiency of directed differentiation in human induced pluripotent cells. *Proceedings of the National Academy of Sciences of the United States of America*, 114(9):2271–2276, 2017.

[56] Cameron Sokolik, Yanxia Liu, David Bauer, Jade McPherson, Michael Broeker, Graham Heimberg, Lei S. Qi, David A. Sivak, and Matt Thomson. Transcription factor competition allows embryonic stem cells to distinguish authentic signals from noise. *Cell Systems*, 1(2):117–129, 2015.

[57] Florian Mueller, Adrien Senecal, Katjana Tantale, Hervé Marie-Nelly, Nathalie Ly, Olivier Collin, Eugenia Basyuk, Edouard Bertrand, Xavier Darzacq, and Christophe Zimmer. Fish-quant: Automatic counting of transcripts in 3d fish images. *Nature Methods*, 10(4):277–278, 2013.

[58] Hiroshi Ochiai, Takeshi Sugawara, Tetsushi Sakuma, and Takashi Yamamoto. Stochastic promoter activation affects nanog expression variability in mouse embryonic stem cells. *Scientific Reports*, 4(1):7125, 2015.

[59] E. Abranches, A. M. Guedes, M. Moravec, H. Maamar, P. Svoboda, A. Raj, and D. Henrique. Stochastic nanog fluctuations allow mouse embryonic stem cells to explore pluripotency. *Development*, 141(14):2770–9, 2014.

[60] Tibor Kalmar, Chea Lim, Penelope Hayward, Silvia Muñoz-Descalzo, Jennifer Nichols, Jordi Garcia-Ojalvo, and Alfonso Martinez Arias. Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS Biology*, 7(7), 2009.

[61] D. W. Austin, M. S. Allen, J. M. McCollum, R. D. Dar, J. R. Wilgus, G. S. Sayler, N. F. Samatova, C. D. Cox, and M. L. Simpson. Gene network shaping of inherent noise spectra. *Nature*, 439(7076):608–611, 2006.

[62] N. Rosenfeld. Gene regulation at the single-cell level. *Science*, 307(5717):1962–1965, 2005.

[63] Alex Sigal, Ron Milo, Ariel Cohen, Naama Geva-Zatorsky, Yael Klein, Yuvalal Liron, Nitzan Rosenfeld, Tamar Danon, Natalie Perzov, and Uri Alon. Variability and memory of protein levels in human cells. *Nature*, 444(7119):643–646, 2006.

[64] Michael B. Elowitz, Arnold J. Levine, Eric D. Siggia, and Peter S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.

[65] Claire McQuin, Allen Goodman, Vasiliy Chernyshev, Lee Kamentsky, Beth A. Cimini, Kyle W. Karhohs, Minh Doan, Liya Ding, Susanne M. Rafelski, Derek Thirstrup, Winfried Wiegraebe, Shantanu Singh, Tim Becker, Juan C. Caicedo, and Anne E. Carpenter. Cellprofiler 3.0: Next-generation image processing for biology. *PLOS Biology*, 16(7):e2005970, 2018.

[66] Toni Pfaffeneder, Fabio Spada, Mirko Wagner, Caterina Brandmayr, Silvia K. Laube, David Eisen, Matthias Truss, Jessica Steinbacher, Benjamin Hackner, Olga Kotljarova, David Schuermann, Stylianos Michalakis, Olesea Kosmatchev, Stefan Schiesser, Barbara Steigenberger, Nada Raddaoui, Gengo Kashiwazaki, Udo Müller, Cornelia G. Spruijt, Michiel Vermeulen, Heinrich Leonhardt, Primo Schär, Markus Müller, and Thomas Carell. Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell dna. *Nature Chemical Biology*, 10(7):574–581, 2014.

[67] Mamta Tahiliani, Kian Peng Koh, Yinghua Shen, William A. Pastor, Hozefa Bandukwala, Yevgeny Brudno, Suneet Agarwal, Lakshminarayan M. Iyer, David R. Liu, L. Aravind, and Anjana Rao. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian dna by mll partner tet1. *Science*, 324(5929):930–935, 2009.

[68] Yu Fei He, Bin Zhong Li, Zheng Li, Peng Liu, Yang Wang, Qingyu Tang, Jianping Ding, Yingying Jia, Zhangcheng Chen, N. Li, Yan Sun, Xiuxue Li, Qing Dai, Chun Xiao Song, Kangling Zhang, Chuan He, and Guo Liang Xu. Tet-mediated formation of 5-carboxylcytosine and its excision by tdg in mammalian dna. *Science*, 333(6047):1303–1307, 2011.

[69] Li Shen, Hao Wu, Dinh Diep, Shinpei Yamaguchi, Ana C. D'Alessio, Ho Lim Fung, Kun Zhang, and Yi Zhang. Genome-wide analysis reveals tet- and tdg-dependent 5-methylcytosine oxidation dynamics. *Cell*, 153(3):692–706, 2013.

[70] Tomas Lindahl and Richard D. Wood. Quality control by dna repair. *Science*, 286(5446):1897–1905, 1999.

[71] Elias S. J. Arnér and Staffan Eriksson. Mammalian deoxyribonucleoside kinases. *Pharmacology and Therapeutics*, 67(2):155–186, 1995.

[72] Bruce Demple, Tory Herman, and Davis S. Chen. Cloning and expression of ape, the cdna encoding the major human apurinic endonuclease: Definition of a family of dna repair enzymes. *Proceedings of the National Academy of Sciences of the United States of America*, 88(24):11450–11454, 1991.

[73] Steven Xanthoudakis, Richard J. Smeyne, James D. Wallace, and Tom Curran. The redox/dna repair protein, ref-1, is essential for early embryonic development in mice. *Proceedings of the National Academy of Sciences of the United States of America*, 93(17):8919–8923, 1996.

[74] Srinivasan Madhusudan, Fiona Smart, Paul Shrimpton, Jason L. Parsons, Laurence Gardiner, Sue Houlbrook, Denis C. Talbot, Timothy Hammonds, Paul A. Freemont, Michael J. E. Stern-

berg, Grigory L. Dianov, and Ian D. Hickson. Isolation of a small molecule inhibitor of dna base excision repair. *Nucleic Acids Research*, 33(15):4711–4724, 2005.

[75] Max A. Horlbeck, Luke A. Gilbert, Jacqueline E. Villalta, Britt Adamson, Ryan A. Pak, Yuwen Chen, Alexander P. Fields, Chong Yon Park, Jacob E. Corn, Martin Kampmann, and Jonathan S. Weissman. Compact and highly active next-generation libraries for crispr-mediated gene repression and activation. *eLife*, 5, 2016.

[76] Clifford D. Mol, Tadahide Izumi, Sankar Mitra, and John A. Talner. Dna-bound structures and mutants reveal abasic dna binding by ape1 dna repair and coordination. *Nature*, 403(6768):451–456, 2000.

[77] Daniel R. McNeill and David M. Wilson. A dominant-negative form of the major human abasic endonuclease enhances cellular sensitivity to laboratory and clinical dna-damaging agents. *Molecular Cancer Research*, 5(1):61–70, 2007.

[78] Namiko Mitarai, Ian B. Dodd, Michael T. Crooks, and Kim Sneppen. The generation of promoter-mediated transcriptional noise in bacteria. *PLoS Computational Biology*, 4(7):e1000109, 2008.

[79] Samuel Corless and Nick Gilbert. Effects of dna supercoiling on chromatin architecture. *Biophysics Reviews*, 8(3):245–258, 2016.

[80] Fedor Kouzine, Laura Baranello, and David Levens. *The use of psoralen photobinding to study transcription-induced supercoiling*, volume 1703, pages 95–108. 2018.

[81] Kimberley N. Babos, Kate E. Galloway, Kassandra Kisler, Madison Zitting, Yichen Li, Yingxiao Shi, Brooke Quintino, Robert H. Chow, V. Berislav Zlokovic, and Justin K. Ichida. Mitigating antagonism between transcription and proliferation allows near-deterministic cellular reprogramming. *Cell Stem Cell*, 25(4):486–500.e9, 2019.

[82] I. Tirosh and N. Barkai. Two strategies for gene regulation by promoter nucleosomes. *Genome Research*, 18(7):1084–1091, 2008.

[83] Jason D. Buenrostro, Beijing Wu, Howard Y. Chang, and William J. Greenleaf. Atac-seq: A method for assaying chromatin accessibility genome-wide. *Current Protocols in Molecular Biology*, 109(1), 2015.

[84] Jiajun Zhang and Tianshou Zhou. Promoter-mediated transcriptional dynamics. *Biophysical Journal*, 106(2):479–488, 2014.

[85] Kenneth P. Burnham, David Raymond Anderson, and Kenneth P. Burnham. *Model selection and multimodel inference : a practical information-theoretic approach*. Springer, New York, 2nd edition, 2002.

[86] Jong Kyoung Kim and John C. Marioni. Inferring the kinetics of stochastic gene expression from single-cell rna-sequencing data. *Genome Biology*, 14(1):1–12, 2013.

[87] Mihails Delmans and Martin Hemberg. Discrete distributional differential expression (d3e) - a tool for gene expression analysis of single-cell rna-seq data. *BMC Bioinformatics*, 17(1):110, 2016.

[88] L. V. Sharova, A. A. Sharov, T. Nedorezov, Y. Piao, N. Shaik, and M. S. H. Ko. Database for mrna half-life of 19 977 genes obtained by dna microarray analysis of pluripotent and differentiating mouse embryonic stem cells. *DNA Research*, 16(1):45–58, 2009.

[89] William J. Blake, Gábor Balázsi, Michael A. Kohanski, Farren J. Isaacs, Kevin F. Murphy, Yina Kuang, Charles R. Cantor, David R. Walt, and James J. Collins. Phenotypic consequences of promoter-mediated transcriptional noise. *Molecular Cell*, 24(6):853–865, 2006.

[90] Jing Guo and Jie Zheng. Hopland: Single-cell pseudotime recovery using continuous hopfield network-based modeling of waddington's epigenetic landscape. *Bioinformatics*, 33(14):i102–i109, 2017.

[91] Abhyudai Singh. Transient changes in intercellular protein variability identify sources of noise in gene expression. *Biophysical Journal*, 107(9):2214–2220, 2014.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
Ravi Desai
2DDA42B314E440C...
Author Signature

6/11/2020
Date