Genome Biology

RESEARCH

Open Access

# Prevalence of and gene regulatory constraints on transcriptional adaptation in single cells

Ian A. Mellis[1,2*], Madeline E. Melzer[3,4,5], Nicholas Bodkin[3,4,5] and Yogesh Goyal[3,4,5,6*]

Yogesh Goyal is the lead contact.

*Correspondence: im2613@cumc.columbia.edu; yogesh.goyal@northwestern.edu

[1] Department of Pathology and Cell Biology, Columbia University Vagelos College of Physicians and Surgeons, New York, NY, USA
[2] Aaron Diamond AIDS Research Center, Columbia University Vagelos College of Physicians and Surgeons, New York, NY, USA
[3] Department of Cell and Developmental Biology, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA
[4] Center for Synthetic Biology, Northwestern University, Chicago, IL, USA
[5] Robert H. Lurie Comprehensive Cancer Center, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA
[6] CZ Biohub Chicago, LLC, Chicago, IL, USA

## Abstract

**Background:** Cells and tissues have a remarkable ability to adapt to genetic perturbations via a variety of molecular mechanisms. Nonsense-induced transcriptional compensation, a form of transcriptional adaptation, has recently emerged as one such mechanism, in which nonsense mutations in a gene trigger upregulation of related genes, possibly conferring robustness at cellular and organismal levels. However, beyond a handful of developmental contexts and curated sets of genes, no comprehensive genome-wide investigation of this behavior has been undertaken for mammalian cell types and conditions. How the regulatory-level effects of inherently stochastic compensatory gene networks contribute to phenotypic penetrance in single cells remains unclear.

**Results:** We analyze existing bulk and single-cell transcriptomic datasets to uncover the prevalence of transcriptional adaptation in mammalian systems across diverse contexts and cell types. We perform regulon gene expression analyses of transcription factor target sets in both bulk and pooled single-cell genetic perturbation datasets. Our results reveal greater robustness in expression of regulons of transcription factors exhibiting transcriptional adaptation compared to those of transcription factors that do not. Stochastic mathematical modeling of minimal compensatory gene networks qualitatively recapitulates several aspects of transcriptional adaptation, including paralog upregulation and robustness to mutation. Combined with machine learning analysis of network features of interest, our framework offers potential explanations for which regulatory steps are most important for transcriptional adaptation.

**Conclusions:** Our integrative approach identifies several putative hits—genes demonstrating possible transcriptional adaptation—to follow-up on experimentally and provides a formal quantitative framework to test and refine models of transcriptional adaptation.

## Background

Cells can sense changing conditions and exhibit robustness in response to perturbations [1–9]. Several mechanisms underpinning this plasticity have been proposed and more continue to be reported. These mechanisms—operational at multiple levels of biological organization—include protein feedback loops, adaptive mutations, and single-cell molecular variabilities [10–16]. One recently reported robustness mechanism is a type of transcriptional adaptation, wherein nonsense mutations, i.e., mutations that result in premature termination of protein synthesis, can trigger the transcription of related genes, including paralogs [14, 15, 17–19]. This adaptation, known as nonsense-induced transcriptional compensation, can enable cells and tissues to escape otherwise fatal mutations and function normally [14, 15]. Besides revealing a new kind of transcriptional regulation, this discovery proposed a resolution of the long-standing discrepancy in molecular and morphological phenotypes between many knockdowns and knockouts of the same gene across several animal model systems [2, 20–23]. In particular, reducing the expression of a gene by using antisense oligos (including morpholinos) often results in more severe defects than for a gene knockout, contrary to the expectation that complete removal of the gene would, in principle, result in a stronger phenotype. Various existing explanations prior to these studies, including off-target effects associated with knockdowns [2], could not fully account for the paradox.

Mechanisms underlying nonsense-induced transcriptional compensation have been studied for a handful of curated genes in select developmental settings, particularly in vertebrates and *C. elegans* [18, 19, 24–30]. Recent studies identified components of the COMPASS complex, a histone methylase, and regulators of nonsense-mediated decay, including Upf genes [14, 15, 31], as important mediators of nonsense-induced transcriptional compensation. For example, premature termination codons in gene *egfl7* in zebrafish caused upregulation of the Emilin gene family via protein Upf1, resulting in a near-absence of vascular defects in zebrafish [26]. However, it is unclear if this compensatory behavior is pervasive in other genes, species, and contexts. While some studies have indicated the absence of this kind of compensation in organisms such as yeast [32, 33], to date, no genome-wide investigation of this behavior has been undertaken for different mammalian cell types and contexts. As a result, several questions remain unanswered. For example, is this adaptive behavior limited to certain genes associated with specific signaling pathways? Similarly, do such compensating gene families tend to be functionally similar, e.g., transcription factors, cytoskeleton molecules, or enzymes? Moreover, is this behavior intrinsic to a gene or dependent on its extrinsic environment (i.e., cell type or local regulation)? Furthermore, how prevalent is this behavior across mammalian systems and contexts, e.g., cancer or differentiation? One hypothesis is that nonsense-induced transcriptional compensation occurs only in very specific biological circumstances and organisms. Alternatively, it is possible that the transcriptional adaptation as a mode of cellular robustness is ubiquitous and occurs more commonly than hitherto appreciated. Both of these scenarios have different implications. For instance, the latter hypothesis implies that functional genetic screens for phenotypic outcomes will need to account for transcriptional adaptation. Computational analysis of existing datasets can lead the way in resolving these alternatives, which can subsequently be tested experimentally.

Another set of questions center around the regulatory constraints on upregulated paralogs and their downstream effector molecules [34, 35]. In particular, nonsense-induced transcriptional compensation can result in incomplete phenotypic penetrance, such that there are either attenuated defects or a subset of cells or organisms which continue to have strong defects despite compensation. In some cases, compensation can happen without necessarily rescuing a phenotypic defect induced by knockout mutations [28, 36–38]. These observations, coupled with the documented evidence that transcription is bursty [39], raise the possibility that inherent stochasticity underlying the compensatory gene regulatory networks may translate into single-cell differences. Single-cell differences, in turn, could result in incomplete penetrance, particularly for phenotypes resulting from variable downstream effects on relevant effector gene expression. However, precisely how compensated expression fluctuations manifest into downstream effects has not been formally investigated. For example, what is the ensemble of gene expression distributions of effector molecules post-compensation? Under what conditions can we expect the system to exhibit robustness of the distribution shape and mean? In a similar vein, does the answer depend on the nature of interactions or network size (negative or positive; one or multiple paralogs)? Resolving these single-cell possibilities with the analysis of single-cell sequencing datasets coupled with theoretical formulations can provide plausible mechanistic bases for the observed phenotypic penetrance, and aid in the design of predictive experiments, especially as single-cell technologies become more accessible [40, 41].

Here, we combine computational analysis of existing datasets with mathematical modeling of stochastic gene regulatory interactions to address the questions posed above. First, we argue that a systematic bioinformatic analysis of publicly available transcriptome-wide datasets that rely on CRISPR-Cas9-mediated mutagenesis can, in principle, suggest the presence of transcriptional adaptation, or lack thereof. Indeed, our unbiased computational pipeline surveying dozens of publicly available datasets, spanning both bulk and single-cell-resolved datasets, not only recovers known and validated gene targets that display nonsense-induced transcriptional compensation but also reveals the breadth of genes, cell types, and biological contexts across which nonsense-induced transcriptional compensation can be operational. Second, we extend the analysis of nonsense-induced compensatory effects to downstream regulatory targets of mutated genes, using annotated transcription factor regulons. We show that transcription factors that display potential transcriptional adaptation have more stable downstream regulatory targets after mutation. Lastly, we develop stochastic mathematical models of biallelic gene regulation and simulate over tens of millions of cells. We find that even a relatively parsimonious model of transcriptional adaptation can recapitulate paralog upregulation after mutation and diverse population-level gene expression distributions of downstream effectors qualitatively similar to those observed in real data. Our integrative framework is generalizable and lays the foundation for future work to test our findings experimentally and to refine models of transcriptional compensation.

## Results

### A generalizable framework for analyzing CRISPR-Cas9 knockouts paired with RNA-sequencing reveals upregulation of knockout-target paralogs

We wondered whether transcriptional adaptation to mutation—specifically nonsense-induced transcriptional compensation—is common in vertebrates, and if so, does it occur in specific genes belonging to specific signaling pathways or in broader gene sets across biological contexts. To address this question, we took advantage of a feature common in published experimental designs: CRISPR-Cas9-based knockout engineering. When paired with a guide RNA, Cas9 creates a double-stranded DNA break at a predefined site in a target gene, after which endogenous non-homologous end joining repair processes induce a random insertion-deletion (indel) mutation [42–45]. On average, two thirds of indels in coding regions will induce a frameshift by random chance. In turn, this frameshift will render the resulting open reading frame of the mutant different from the wild-type and cause a premature termination codon [46]. Since nonsense-induced transcriptional compensation is proposed to occur as a result of premature termination codons, we hypothesized that transcriptomic data from Cas9-based knockout experiments could reveal the presence—or absence—of potential nonsense-induced transcriptional compensation (Fig. 1A). Furthermore, even if a specific nonsense or frameshift allele is not isolated and expanded (i.e., the mutated population is polyclonal), at least two thirds of Cas9-affected alleles in Cas9-treated cells will be nonsense mutants.

Since nonsense-induced transcriptional compensation can depend on sequence homology [14, 15], we first developed a robust methodology for choosing genes that may compensate for a knockout target. There are several documented methods for choosing related, potentially compensatory, genes. These range from considering whole protein families to identifying more recently ancestrally related paralog genes to performing genome-wide local alignment searches [14, 26]. We decided to use paralog genes in our analysis as they are consistently annotated, and identifying Ensembl-annotated paralogs does not depend on individually optimized local alignment search parameters [47]. We then performed a comprehensive literature search for published, publicly available datasets for CRISPR-Cas9 knockout experiments paired with bulk RNA sequencing of both nontemplate controls and knockout target samples. We analyzed mouse and human samples. Furthermore, we prioritized published datasets that included multiple parallel knockout experiments. In total, we screened over 200 datasets in the NIH's Gene Expression Omnibus (GEO) and identified 36 GEO entries with a total of 220 initially analyzable CRISPR gene targets meeting our experimental design criteria, including 76 in mice and 144 in humans (Fig. 1B, Additional File 1: Table S1). After quality control, paralog lookup, and differential expression filters, we proceeded to analyze a total of 74 gene targets and their respective nontemplate controls (see the "Methods" section). The datasets analyzed include knockouts engineered for the study of a variety of biological phenomena, including organ development, reprogramming to pluripotency, and tumor responses to targeted therapies, among others (Additional File 1: Table S1).

With our collection of quality-controlled datasets, we examined whether nonsense-induced transcriptional compensation may exist more widely than previously reported. Specifically, we asked whether paralogs of knockout targets were upregulated after knockout more frequently than would be expected by random chance (see

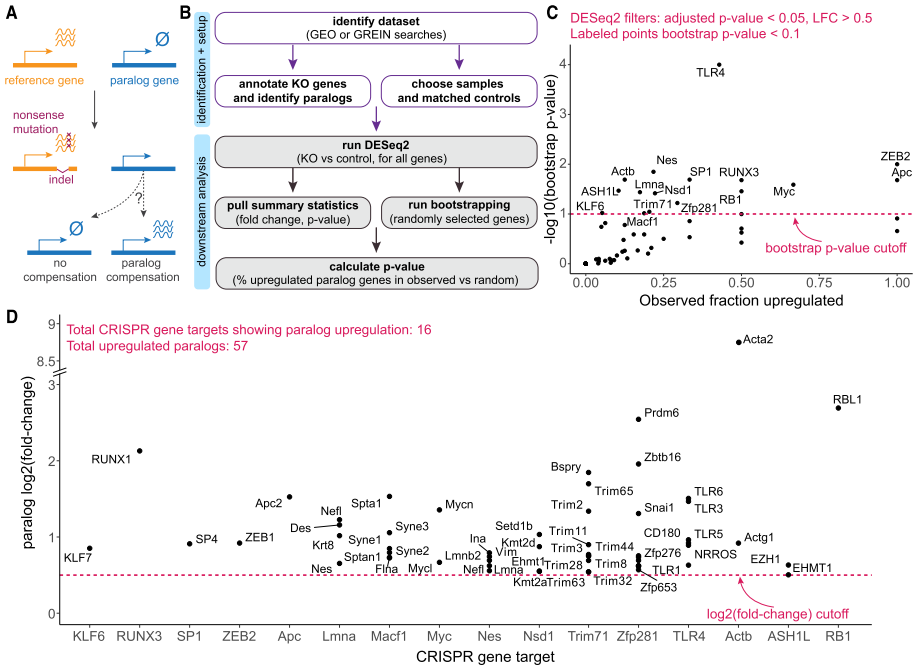


**Fig. 1** Inferring prevalence of transcriptional adaptation in transcriptomic datasets. **A** Schematic of transcriptional adaptation after Cas9-mediated mutagenesis. After indel mutations, frameshifts often occur, leading to premature termination codon formation and resultant nonsense-mediated decay. We hypothesized that paralogs may be upregulated in genes with nonsense-induced transcriptional compensation-type transcriptional adaptation. **B** Schematic of analytical workflow. We mined publicly available RNA-seq datasets for differential expression of paralogs of CRISPR/Cas9 knockout targets after mutation. Randomly selected bootstrap sampled genes are chosen to have similar average expression levels as respective paralogs of interest. **C** Per-knockout-target paralog differential expression results. The -log10 bootstrap *p*-value compared to the observed fraction of upregulated paralogs. Paralog differential expression counted if log2 fold-change > 0.5 and adjusted *p*-value < 0.05. **D** Per-upregulated-paralog differential expression magnitude. For significantly upregulated paralogs of knockout targets showing transcriptional adaptation in **C**, log2 fold-change relative to controls. Knockout targets on *x* axis in arbitrary order

the "Methods" section). We found that 16 out of 74 knockout targets had significant upregulation of their paralogs (Fig. 1C,D, Additional File 2: Fig. S1A, Additional File 1: Table S2). We confirmed this result was not specific to our thresholds for differential expression (adj. *p*-value < 0.05 and log2(fold change) > 0.5) by repeating the analysis using other differential expression or average expression paralog inclusion criteria (Additional File 2: Fig. S1B,C). Gene hits include *ASH1L*, *KLF6*, *RB1*, *RUNX3*, *SP1*, *TLR4*, and *ZEB2* in humans and *Actb*, *Apc*, *Lmna*, *Macf1*, *Myc*, *Nes*, *Nsd1*, *Trim71*, and *Zfp281* in mice. Our analysis is largely consistent with the published findings of related-gene upregulation after nonsense mutation of 3 target genes found to demonstrate nonsense-induced transcriptional compensation by El-Brolosy et al., 2019 (*Fermt2*, *Actg1*, *Actb*; Additional File 2: Fig. S1D). It is also important to note that our work recapitulated these earlier findings despite using different related-gene-inclusion criteria (local alignment in El-Brolosy et al., 2019, vs. paralog identity here). Furthermore, while we observed some degree of transcriptional upregulation in paralogs of all 3 target genes, only *Actb* was deemed significant by our bootstrap analysis pipeline. This result suggests that beyond the 16 significant hits reported in our study, our paralog-based analysis is relatively more stringent, perhaps leading to false-negative findings. Remarkably, there were 2 CRISPR

targets in our dataset that are paralogs of each other, *Lmna* and *Nes*, and we found that both were classified as hits. Moreover, for both *Lmna* and *Nes*, their mutual paralog gene *Nefl* was upregulated upon mutation of either *Lmna* or *Nes* (Fig. 1D). Therefore, despite conservative cutoffs, the shared upregulation of compensating, mutually paralogous genes across independent experiments illustrates the power of our approach and the reliability of our findings.

### Degree and frequency of paralog upregulation are similar across conditions for the same genes

We wanted to check whether paralog upregulation frequencies and paralog fold-changes were consistent across conditions for the same CRISPR target genes. Therefore, we compared paralog upregulation frequency for all 27 knockout targets that met our inclusion criteria for analysis across two conditions in a dataset published in Lackner et al., 2021 [48] (see the "Methods" section). Lackner and colleagues performed RNA-sequencing on different knockout mouse embryonic stem cell lines, both under standard naive stem cell culture conditions and after a day of differentiation (Fig. 2A). In our main analysis (Fig. 1C, D), we considered the results from Lackner et al., 2021, under standard culture

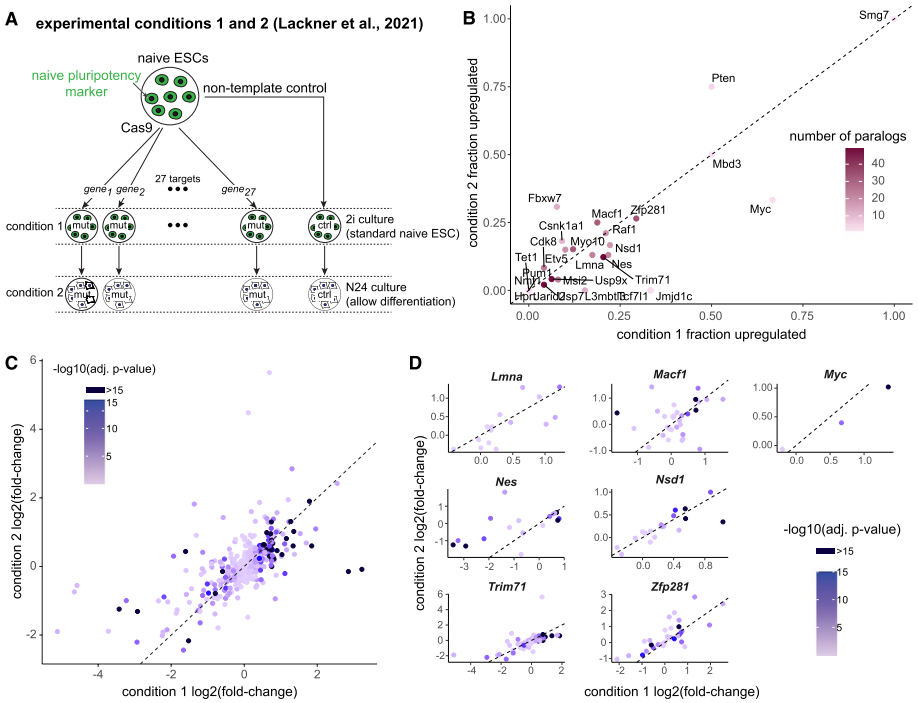


**Fig. 2** Repeatability of inferring prevalence of transcriptional adaptation across experimental conditions. **A** Schematic of experimental design for analysis of paralog upregulation frequency for 27 knockout targets in mouse naive embryonic stem cells (ESCs), adapted from Lackner et al., 2021. **B** Comparison of paralog upregulation frequency in 2i (standard naive ESC culture condition, condition 1, GSE145653-1) versus N24 (removal of 2i for 24 h, allowing exit from naive pluripotency, condition 2, GSE145653-2) for 27 targets. Color indicates the number of paralogs per knockout target. Dotted line indicates $y = x$. **C** Comparison of log2 fold-change of paralog expression across experimental conditions. Each point corresponds to a gene-paralog pair. Color indicates the -log10 adjusted *p*-value calculated from naive ESC dataset (condition 1), with any value greater than 15 being represented by the same dark blue. Dotted line indicates $y = x$. **D** Comparison of log2 fold-change of paralog expression across experimental conditions for individual knockout targets with bootstrap *p*-values < 0.1 (i.e., "hits") as in **C**

conditions to reduce potential confounding effects on expression changes associated with possible divergent differentiation outcomes in each knockout compared against differentiated non-template control lines. Nonetheless, we asked whether there is any agreement in paralog upregulation frequency across the standard undifferentiated and differentiated conditions. We found that for each knockout target, the fraction of paralogs upregulated in standard culture conditions was broadly correlated with the fraction of paralogs upregulated after a day of differentiation (Fig. 2B; $r = 0.832$), as did the degree of upregulation of each paralog (Fig. 2C, D, Spearman correlation coefficient = 0.677 for paralogs of all targets). Therefore, for the tested targets in mouse embryonic stem cells, paralog upregulation frequency is similar in two different conditions, further demonstrating the robustness of our approach.

#### Paralog upregulation is also observed in large-scale pooled single-cell CRISPR screens

Next, we wondered whether we could identify additional genes that may exhibit nonsense-induced transcriptional compensation in large pooled knockout experiments with single-cell resolution. Perturb-seq, CROP-seq, CRISP-seq, and related methods enable pooled parallel single-cell gene expression profiling of dozens or hundreds of knockout targets [49–51]. Here again, we reasoned that single-cell pooled datasets utilizing Cas9 or equivalent perturbation tools would cause indel mutations in the coding regions of the genes of interest. In addition to the benefits of single-cell resolution of gene expression and its high-throughput, Perturb-seq-style data offers consistency by using a common internal set of non-template-control-treated cells as a comparison for all knockout targets. We identified a large, quality-controlled, pre-processed Perturb-seq dataset with ~ 750 distinct guide RNAs using Cas9 as the knockout effector in patient-derived cancer cells. The dataset includes dozens of non-template-control guides and gene-targeting guides directed at > 200 target human genes with a wide variety of molecular functions, chosen due to their involvement with cell-intrinsic therapy resistance [52]. After quality controls, we considered cells representing 143 target genes with 429 total targeting guides as well as 37 non-template-control guides in the main analysis (Fig. 3A).

We then asked whether there was a trend toward increased expression at the single-cell level of any paralogs of each knockout target when compared against non-template-controls. Due to known drop-out events in single-cell RNA-sequencing, we initially focused our analysis on simply counting the fraction of cells with non-zero expression (i.e., "percent positive") of each paralog of a knockout target. We compared the percent positive cells treated with a targeting guide against cells treated with a control guide [53, 54]. For paralogs with a high baseline expression of at least 75% in control cells, we compared average expression levels instead of percent positive values. In general, the percent positives (and the means) in targeted cells and in controls were well correlated (Fig. 3B). Nonetheless, there were many paralog genes in which expression levels may have differed. We rank-ordered the differences between targeted and control cells for all 1792 paralog-target gene pairs and highlighted the paralogs with the 100 largest absolute increases in percent positive values after respective target knockout (or highest mean increases for those highly expressed at baseline). Eighty-five of the largest differences detected were observed when considering either change in percent-positivity or change in mean abundance, suggesting

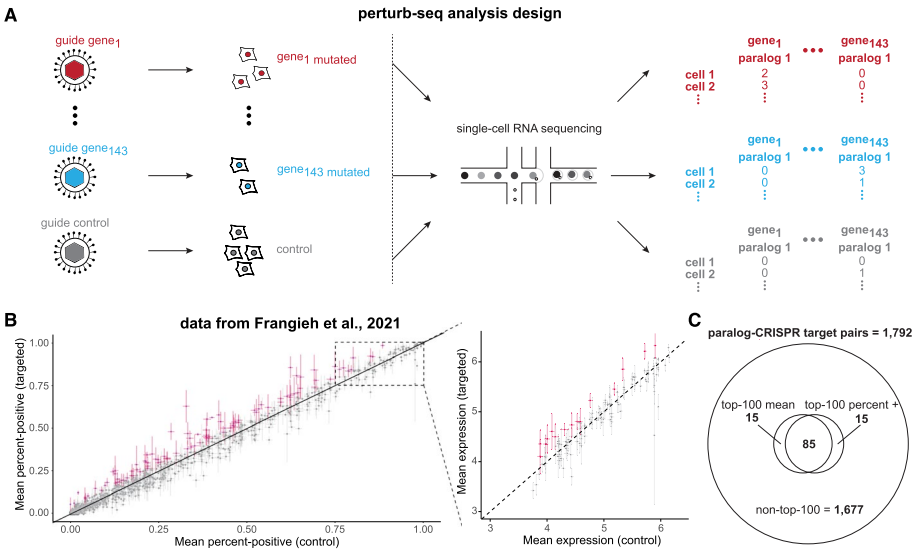


**Fig. 3** Inferring prevalence of transcriptional adaptation at single-cell resolution in Perturb-seq data. **A** Schematic of experimental design and paralog gene expression analysis, adapted from Frangieh et al., 2021. 143 gene targets, with a consistent batch of non-template control-treated cells, passing quality control filters (see the "Methods" section). **B** Perturb-seq-based single-cell paralog expression change after reference gene knockout. Per paralog, percentage of cells positive for that paralog's expression in non-template control guide-treated cells on *x*-axis, percentage of cells positive for that paralog's expression in cells treated with guides targeting a reference CRISPR target for which the gene is a paralog. For paralog genes with percent-positive > 75% in non-template controls, mean expression plotted in inset at right. Paralogs ranked in the top-100 of absolute increases per quantification method marked in magenta. All paralogs of all knockouts shown meeting minimum cell count and UMI count in gray, inclusion criteria listed in the "Methods" section. **C** Venn diagram summarizing the number of paralog-CRISPR target pairs with the top-100 largest increases in mean expression and/or percent-positive expression levels. 1792 total paralog-target pairs, of which 1677 are not in the top-100 largest increase lists. 85 paralog-target pairs in the top-100 largest difference list by both mean and percent-positive analysis

that the identification of a target-paralog pair was not entirely dependent on the choice of paralog-upregulation measure (Fig. 3C). There were 43 distinct knockout targets with top-100-increase paralogs, including transcription factors, RNA-binding proteins, signaling pathway components, cell cycle regulators, and cell surface proteins (Table 1, Additional File 1: Table S3). Additionally, we did a secondary analysis on the subset of gene targets that did not pass our conservative total cell filtering yet showed potential paralog upregulation (Additional File 2: Fig. S1E). Particularly, 6 of the 23 paralogs of *TUBB* (tubulin beta class I) may have been upregulated after *TUBB* mutation (Additional File 2: Fig. S1E). Intriguingly, others have experimentally shown that mutations of tubulins can lead to tubulin paralog upregulation in a mouse model of neurodegeneration [36].

We next asked whether any specific annotated biological pathways, molecular functions, or cellular components were enriched for genes that display transcriptional adaptation by paralogs, either in bulk or in single-cell datasets. Therefore, we conducted gene set enrichment analysis by comparing human knockout targets with significantly upregulated paralogs (in bulk data) or paralogs in the top-100-increase sets (in single-cell data) against the full set of human knockout targets tested (see the "Methods" section) [55]. No gene sets were overrepresented among the knockout targets with paralog

**Table 1** List of CRISPR targets with multiple paralogs demonstrating large increases in paralog expression. All CRISPR targets with more than 1 paralog in the top-100 list (see the "Methods" section) and with at least 10% of all annotated paralogs in the top-100 lists. A complete list of CRISPR targets with paralogs in the top-100 list is in Additional File 1: Table S3

| CRISPR target | # paralogs in top-100 | # annotated paralogs | Fraction of paralogs in top-100 |
|---|---|---|---|
| *TMED10* | 5 | 9 | 0.56 |
| *ILF2* | 5 | 13 | 0.39 |
| *CDK6* | 8 | 26 | 0.31 |
| *PABPC1* | 6 | 21 | 0.29 |
| *NONO* | 2 | 7 | 0.29 |
| *SMAD4* | 2 | 7 | 0.29 |
| *IRF4* | 2 | 8 | 0.25 |
| *FRZB* | 3 | 14 | 0.21 |
| *EIF4A1* | 7 | 36 | 0.19 |
| *PPIA* | 3 | 16 | 0.19 |
| *RAB27A* | 11 | 66 | 0.17 |
| *DDX39A* | 6 | 36 | 0.17 |
| *DDX17* | 5 | 36 | 0.14 |
| *CCND1* | 2 | 17 | 0.19 |
| *HLA-C* | 2 | 17 | 0.19 |
| *HLA-F* | 2 | 17 | 0.19 |
| *SERPINE2* | 3 | 27 | 0.11 |
| *FKBP4* | 2 | 18 | 0.11 |

upregulation. Therefore, transcriptional adaptation, at least for the targets tested, may not be limited to specific biological contexts, since we observed that it is not strongly correlated with functionally defined gene sets or particular signaling pathways.

We also wondered whether higher expression of any genes implicated in the proposed mechanisms of transcriptional adaptation were associated with targets exhibiting paralog upregulation. Therefore, from the two largest datasets, one human and one mouse, we extracted the expression levels of 12 genes associated with transcriptional adaptation (members of the COMPASS complex and Upf genes important for nonsense-mediated decay) [14, 15, 31, 48, 56–58]. We compared these 12 genes' expression levels for each knockout-control pairing in the dataset, grouped by whether the knockout target displayed paralog upregulation. We did not observe any major differences in the 12 genes' expression levels between targets with signs of transcriptional adaptation versus those without, in either humans or mice (Additional File 2: Fig. S2). Note, however, that the analysis was limited by low numbers of knockout targets, particularly in the groups displaying paralog upregulation. Moreover, it is also possible that protein-level, rather than transcript-level, regulation of COMPASS complex components, NMD pathway components, and other effectors of transcriptional adaptation drive paralog upregulation [17]. Future high-throughput and multimodal studies with more robust datasets will play a critical role in clarifying the role of COMPASS complex component or Upf gene expression levels in transcriptional adaptation.

**Several knockout target and paralog features are not associated with paralog upregulation**

Since we observed variability in upregulation of different paralogs for a target gene upon CRISPR mutation, we wondered whether paralog-intrinsic factors might be associated with whether a given paralog participates in transcriptional adaptation. Specifically, recent studies have indicated that genes that have some degree of local sequence homology with a nonsense-mutated gene are more likely to be upregulated after mutation, but this has not been systematically investigated [14]. We checked whether paralog upregulation has any association with the degree of sequence homology (i.e., transcript-wide percent homology) with the knockout target. We found no significant correlation between the degree of sequence homology and the expression change after mutation for bulk CRISPR targets demonstrating transcriptional adaptation (Additional File 2: Fig. S3A). Next, given the requirement for degraded transcripts from mutated genes in proposed mechanisms of transcriptional adaptation, we wondered whether longer knockout target genes might more often have paralog upregulation. We found no significant correlation between length of gene and the paralog upregulation frequency (Additional File 2: Fig. S3B).

Next, we asked whether shared genomic regulatory elements, including enhancers and promoters, between a paralog and its respective CRISPR target was associated with paralog upregulation after knockout. First, we mined the GeneHancer database of human enhancers and promoters annotated with their regulated genes [59]. We found only two paralogs that shared enhancers or promoters with their respective CRISPR target (out of 489 paralog-target pairs with both genes in the database). There was no difference in odds of paralog upregulation between the two paralogs with shared enhancers or promoters and those without (Fisher's exact test $p$-value = 0.13, see the "Methods" section). In *Caenorhabditis elegans*, studies have shown for at least one pair of genes that transcriptional adaptation requires partial homology between a promoter and a subsequence of the mutated transcript [29]. We hypothesized that even for genes not related to each other, sharing an upstream regulatory element with the CRISPR target may predispose another gene to upregulation. Our hypothesis, coupled with the fact that the number of paralogs with shared regulatory elements was low, led us to further extend the check for association between sharing regulatory elements and upregulation after knockout to any genes in the database. Therefore, we repeated a check of shared regulatory elements for all genes in a given experiment in the GeneHancer database with the CRISPR target. Here again, we found no clear association between a gene being upregulated and whether it shared an enhancer or promoter with the CRISPR target (Fisher's exact test $p$-value range (0.12, 1) across human CRISPR targets; see the "Methods" section).

We also checked whether 3D genome architecture may be associated with paralog upregulation after target knockout. We mined the TADKB database [60] of human topologically associated domains to check for knockout target-paralog pairs both located in shared annotated domains in any available human cell type (see the "Methods" section). We considered genes co-located in the same topologically-associated domain if any Ensembl-annotated transcription start sites for both genes occurred within the domain. We found that only 4 out of 562 target-paralog pairs tested were co-located in the same topologically-associated domains. There was no difference in odds of paralog upregulation between the few paralogs with shared domains and those without (Fisher's

exact test $p$-value = 0.23). There is evidence that transcriptional adaptation may lead to the mutated gene's own upregulation, such as at an unmutated allele in the case of heterozygosity or endogenous loci if mutated transcripts are injected into the cell [61]. Therefore, we hypothesized that if there were an association between 3D chromatin architecture and gene upregulation by transcriptional adaptation, it is possible that any neighboring gene with shared architectural regulatory features is more often upregulated. We repeated a check of shared domain co-location for CRISPR targets and any genes for which differential expression could be calculated. As with paralogs, we found minimal association between any gene sharing a topologically-associated domain with the CRISPR target and whether that gene was upregulated (unadjusted Fisher's exact test $p$-value range (0.039, 1); adjusted range (0.79, 1); see the "Methods" section). The only gene with an unadjusted $p$-value < 0.05 was *SMARCA4* ($p = 0.039$), which was not a hit in the main analysis of paralog upregulation after *SMARCA4* mutation and which with adjustment for multiple hypothesis testing did not meet a false-discovery rate threshold of < 0.05 (adjusted $p$-value = 0.79).

We then wondered whether paralog expression similarity to knockout target expression across contexts is predictive of paralog upregulation after target mutation. Therefore, we performed an analysis at a functional level of genomic regulation, agnostic of specific regulatory features. Specifically, we calculated paralog expression level correlation with knockout target expression level across 54 human tissue types in the GTEx dataset (see the "Methods" section). We then compared the correlation coefficient for the expression of each gene-paralog pair against that pair's paralog expression change after target mutation (Additional File 2: Fig. S1F). We found no positive association between gene expression correlation and paralog expression change after target knockout (instead, in fact, there was a slight anticorrelation (Spearman rho = −0.11, $p$-value = 0.01).) Of note, most paralogs have positive average expression correlation with the respective target genes across tissues, consistent with prior results [62].

Taken together, our results suggest that nonsense-induced transcriptional compensation may exist for several vertebrate genes, and the paralog(s) that get upregulated do not necessarily depend linearly on the degree of sequence homology or mutated gene length. Nor does paralog upregulation appear to be associated with sharing annotated enhancers or promoters or topologically associated domains with the mutated gene, granted the datasets were sparse and few. As new genomic datasets continue to be reported, future studies may systematically explore whether alternative paralog- or knockout-target-specific characteristics are associated with upregulation after mutation.

### Robustness of regulons at bulk and single-cell resolution for transcription factors exhibiting transcriptional adaptation

Previous studies have shown that paralog upregulation via transcriptional adaptation can enable the preservation or robustness of molecular and morphological phenotypes downstream of mutated regulators. Therefore, we wondered whether we would observe relative downstream buffering effects for targets exhibiting transcriptional adaptation versus those that do not. Specifically, we wanted to know if there was robustness of expression distribution shape and average level for downstream targets of transcription factors—which can activate or repress downstream genes—that showed signs

of transcriptional adaptation via paralog upregulation (Fig. 4A). To address this question, we isolated transcription factors from the bulk RNA-seq and Perturb-seq single-cell RNA-seq datasets previously analyzed for paralog upregulation (Additional File 1: Table S1 and [52]).

For the bulk RNA-seq datasets, we first checked whether any CRISPR targets with significant paralog upregulation were transcription factors. Next, we searched the DoRothEA regulon database for common downstream targets of CRISPR targets and their paralogs (i.e., common regulons) [63, 64]. We found common regulons with high-quality annotations for 4 target-paralog pairs for targets demonstrating possible transcriptional adaptation: *RUNX3-RUNX1*, *SP1-SP4*, *ZEB2-ZEB1*, and *Myc-Mycn*. In parallel, we repeated regulon searches for all target-paralog pairs for targets that did not appear to demonstrate transcriptional adaptation, as well. We hypothesized that transcriptional adaptation by transcription factor paralogs should, on average, buffer extreme changes in the expression of downstream genes after CRISPR target mutation. To test our hypothesis, we calculated differential expression of common regulon genes after CRISPR target knockout. We found that common regulon genes downstream of transcription factors with signs of transcriptional adaptation were significantly less often differentially expressed (10.2% of regulon genes) than those downstream of transcription factors

(See figure on next page.)

**Fig. 4** Robustness of regulon expression associated with transcription factor transcriptional adaptation. **A** Analysis schematic: is there differential regulon gene expression robustness after mutation of upstream transcription factors, associated with whether the transcription factors demonstrate transcriptional adaptation by paralogs? **B** Change in gene expression for each regulon gene after reference CRISPR transcription factor target mutation, compared against respective controls (log2(fold change) from DESeq2). Each point represents one downstream regulon member gene. Gray: regulon genes downstream of CRISPR target-paralog pairs not appearing to be involved in transcriptional adaptation. Red: regulon genes downstream of CRISPR target-paralog pairs with apparent transcriptional adaptation. Regulon genes called differentially expressed if DESeq2 adjusted $p$-value < 0.05 and abs(log2(fold change)) > 0.5. Fisher's exact test for difference of odds of differential expression between the two groups of CRISPR target regulons, $p = 1.82 \times 10^{-4}$. **C** Gene expression distributions of a representative transcription factor that demonstrates possible transcriptional adaptation: *SMAD4*, in non-template controls ("control", gray) and in *SMAD4*-guide-treated cells ("targeted", orange). Fisher's exact test $p$-values for difference in odds of cells being positive in control vs. targeted populations. Abundance values log(TPM + 1) reported by [52]. **D** Gene expression distributions of a representative top-100 paralog gene of a transcription factor that demonstrates possible transcriptional adaptation: *SMAD1*, a paralog of *SMAD4*, in non-template controls (gray) and in *SMAD4*-guide-treated cells (orange). Fisher's exact test $p$-values for difference in odds of cells being positive in control vs. targeted populations. Abundance values log(TPM + 1) reported by [52]. **E** Gene expression distributions of three representative regulon genes of a transcription factor that demonstrates possible transcriptional adaptation: *ID3*, *TNFRSF11B*, and *ID2*, regulon genes of both *SMAD4* and *SMAD1*, in non-template controls (gray) and in *SMAD4*-guide-treated cells (orange). Fisher's exact test $p$-values for difference in odds of cells being positive in control vs. targeted populations. Bimodality coefficient calculated for non-zero subpopulation of each distribution. Abundance values log(TPM + 1) reported by [52]. **F** Change in gene expression for each regulon gene after reference CRISPR transcription factor target mutation, compared against non-template controls (i.e., percent-positive in CRISPR targeted cells minus percent-positive in non-template control-treated cells, for genes 75% or less in non-template controls). Each point represents one downstream regulon member gene. Gray: regulon genes downstream of CRISPR target-paralog pairs not appearing to be involved in transcriptional adaptation. Orange: regulon genes downstream of CRISPR target-paralog pairs with apparent transcriptional adaptation. Asymptotic test of difference in coefficient of variation for groups of unequal size, $p = 0.0017$. **G** Empirical distribution of standard deviation of change in percent-positive for downsampled ($n = 49$, same as transcriptional-adaptation group size) no-transcriptional-adaptation group regulon members without replacement, 1000 downsamples, in gray. Observed standard deviation of change in percent-positive transcriptional-adaptation group standard deviation to this distribution, in orange

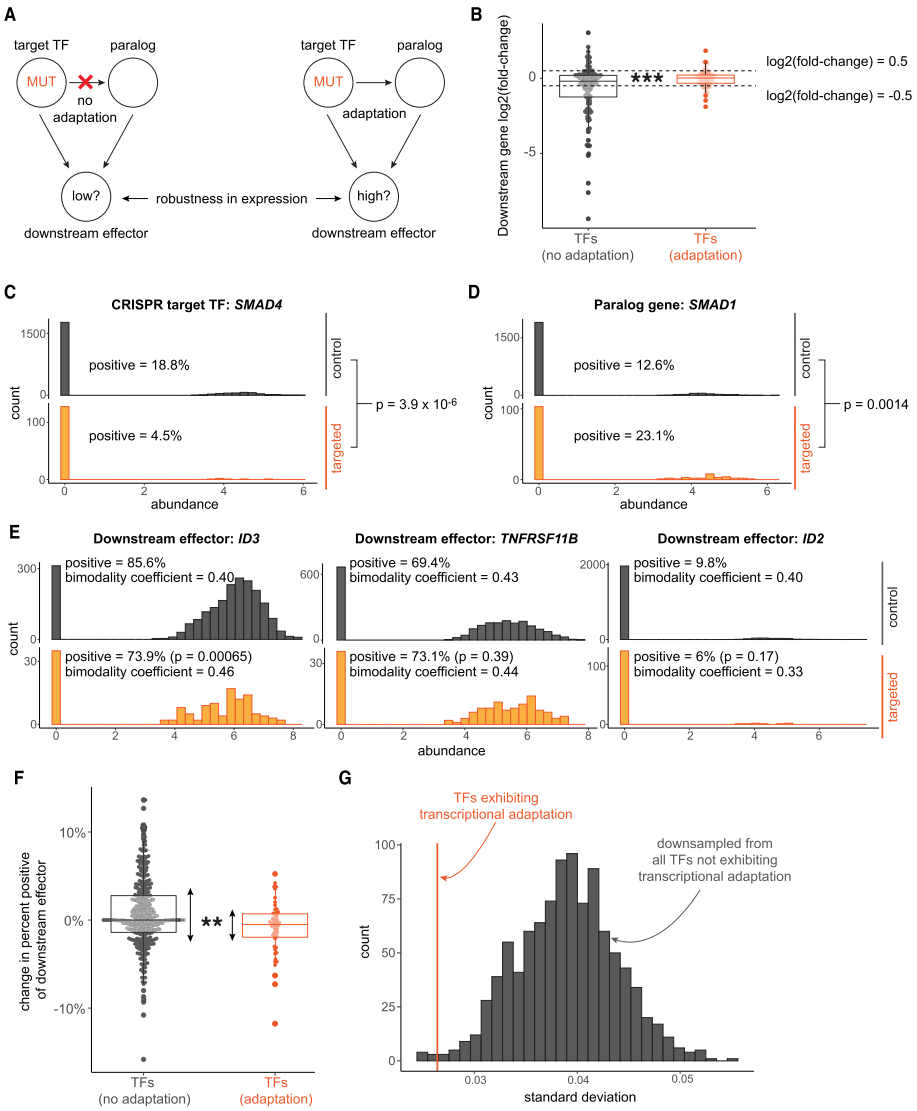


**Fig. 4** (See legend on previous page.)

without signs of transcriptional adaptation (34.1% of regulon genes; Fisher's exact test $p$-value $= 1.82 \times 10^{-4}$; Fig. 4B).

For the Perturb-seq single-cell dataset, we asked whether any of the top-100 upregulated paralogs were paralogs of known human transcription factors. Six pairs of top-100 CRISPR targets-paralog genes spanning 4 distinct CRISPR targets were transcription factors. We then searched the DoRothEA regulon database for common downstream targets of CRISPR targets and their paralogs (i.e., common regulons) [63, 64]. We found overlapping regulons with high-quality annotations for 3 of the 6 top-100 target-paralog pairs: *IRF4-IRF1*, *SMAD4-SMAD1*, and *TFAP2A-TFAP2C* (Fig. 4C–E). For each regulon, we then plotted single-cell transcript abundances for the CRISPR target (Fig. 4C), the paralog (Fig. 4D), and downstream genes (Fig. 4E) both in non-template control cells and in cells treated with guides specific to the CRISPR target. By inspection, we observed that some downstream genes demonstrated robust, consistent gene expression

distribution shapes, while others demonstrated partial expression robustness with possible new modes of gene expression (i.e., the possible presence of bimodality or multimodality, with some gene expression distribution differences showing modest increases in bimodality coefficient; see the "Methods" section) or decreased average expression (Fig. 4E).

As with the bulk RNA-seq datasets, we hypothesized that transcriptional adaptation by transcription factor paralogs would, on average, buffer extreme changes in downstream genes after CRISPR target mutation. Therefore, we repeated regulon searches and calculated the change in downstream gene expression for target-paralog pairs that did not appear to demonstrate transcriptional adaptation and for those exhibiting transcriptional adaptation. Compared to regulons of CRISPR target transcription factor-paralog pairs that do not appear to demonstrate transcriptional adaptation, the spread of gene expression changes downstream of top-100 target-paralog pairs was indeed narrower ($p = 0.0017$, see the "Methods" section) (Fig. 4F, G). In summary, transcriptional adaptation by paralogs of mutated transcription factors is associated with buffering of extreme expression changes in their mutual downstream regulon genes, in both bulk and single-cell datasets.

### Building a minimal network model of the effects of nonsense-induced transcriptional compensation

We demonstrated the existence of transcriptional adaptation in mice and humans across multiple contexts. Particularly, the results from Perturb-seq datasets suggest incomplete penetrance of transcriptional adaptation at a population level, with single-cell differences in the frequency and magnitude of related paralog upregulation (Fig. 3) as well as downstream effector molecules (Fig. 4). While such publicly-available datasets provide an important view of nonsense-induced transcriptional compensation, several questions related remain unanswered. For example, can simple gene regulatory networks recapitulate single-cell variability in compensating paralogs? Furthermore, under what conditions is transcriptional adaptation capable of inducing robustness across a population of cells, in that the compensating paralog expression precisely mimics wild-type expression at a single-cell level? Of note, robust paralog expression alone may not be sufficient, as paralog activity (e.g., that of a paralogous enzyme or a transcription factor) can differ substantially from the original gene. Similarly, gene regulatory network effects can result in distributions of effector molecules in single cells that are non-trivial to predict, yet they can have profound phenotypic implications. For example, mutations in regulators of *C. elegans* intestinal fate can result in downstream effector expression heterogeneity, further dependent on the continued function of other regulatory network components [65]. Therefore, it is important to identify the major control knobs that may confer robustness, or lack thereof, to (1) qualitatively recapitulate the computational findings from experimental datasets and (2) obtain a plausible ensemble of single-cell variabilities and their sources in networks exhibiting transcriptional adaptation.

To address this gap, we built a theoretical framework to model the ensemble of single-cell transcriptional-adaptation-containing network output possibilities with a minimal set of stochastic biochemical reactions. We chose to model cells in which a gene that exhibits nonsense-induced transcriptional compensation controls the expression of a

downstream effector molecule. Briefly, in our initial minimalistic model comprised of 13 parameters (see the "Methods" section), we simulate transcription of an upstream regulator, A, with a paralog, A', exhibiting nonsense-induced transcriptional compensation, and a downstream target, B, in a diploid genome (Fig. 5A, the "Methods" section). Gene product A in wild-type regulates the transcription of downstream pathway member, B. Mutation of A is compensated for by nonsense-mediated expression enhancement of A', which also regulates transcription of B when present (Fig. 5A and Additional File 2: Supplementary Note).

To model the effect of nonsense-mediated expression enhancement of A' on B, we used an expanded version of the telegraph model of transcription [66] as a building block in our model: each gene can reversibly switch between a transcriptionally inactive state (to which, $r_{off}$) and one or more active states (to which, $r_{on}$) (Additional File 2: Fig. S4A). When active, the gene product is transcribed in a Poisson process at a rate ($r_{prod}$). Degradation of each product also occurs as a Poisson process ($r_{deg}$). We specify the directed interaction between mutated gene A regulating the transcription of paralog gene A', which represents nonsense-induced transcriptional compensation, by adding a parameter ($r_{add}^{NITC}$) with dependency on the real-time abundance of mutated A gene product modified by a Hill function (Hill coefficient $n$), to account for the nonlinearity of gene regulatory interactions. We combine steps leading to transcription by making the quasi-equilibrium assumption, commonly used in models of gene regulatory networks due to differences in individual reaction timescales [67, 68]. We represent the differential regulation of B by A and A' by specifying two distinct active states for B: the active state directed by A and the active state directed by A'. The active states of B each have a respective production rate. In sum, our minimalistic model includes 12 varying parameters and 1 fixed parameter. We condensed the parameter search space to 8 independent and interpretable variables by focusing on parameter relationships in relation to a subset of critical network parameters (see the "Methods" section, Fig. S4A-D).

#### Gene expression distribution shape varies widely across the parameter search space

To understand how relationships between the parameters in the model relate to network output, we simulated the network model using eight independent variables [69] (see the "Methods" section, Supplementary Note, Fig. SN1). We chose parameter search ranges based on studies empirically documenting transcriptional burst kinetics in mammalian cells, when available [66, 70–72]. An autocorrelation analysis on simulations confirmed that they are ergodic, thus enabling us to condense long-timescale traces into "single-cell-like sub-simulations" (Additional File 2: Fig. S5, Supplementary Note).

Manual inspection of distributions of random samples showed that the distributions tended to fall into 5 general classes of distribution shape: low-expression, unimodal symmetric, left-skewed, right-skewed, and bimodal (or multimodal) (Fig. 5B, C, Additional File 2: Fig. S6A, B). We next sought to automate the process of describing the expression level distributions per gene per genotype in each simulation for several tens of thousands of simulation runs (see Additional File 2: Fig. S4E, Supplementary Note). To systematically classify distribution shapes, we developed a heuristic algorithm based on summary statistics and verified the accuracy with manual checks (see the "Methods" section, Supplementary Note). With our fairly accurate classifier (80–100% per class; Additional File

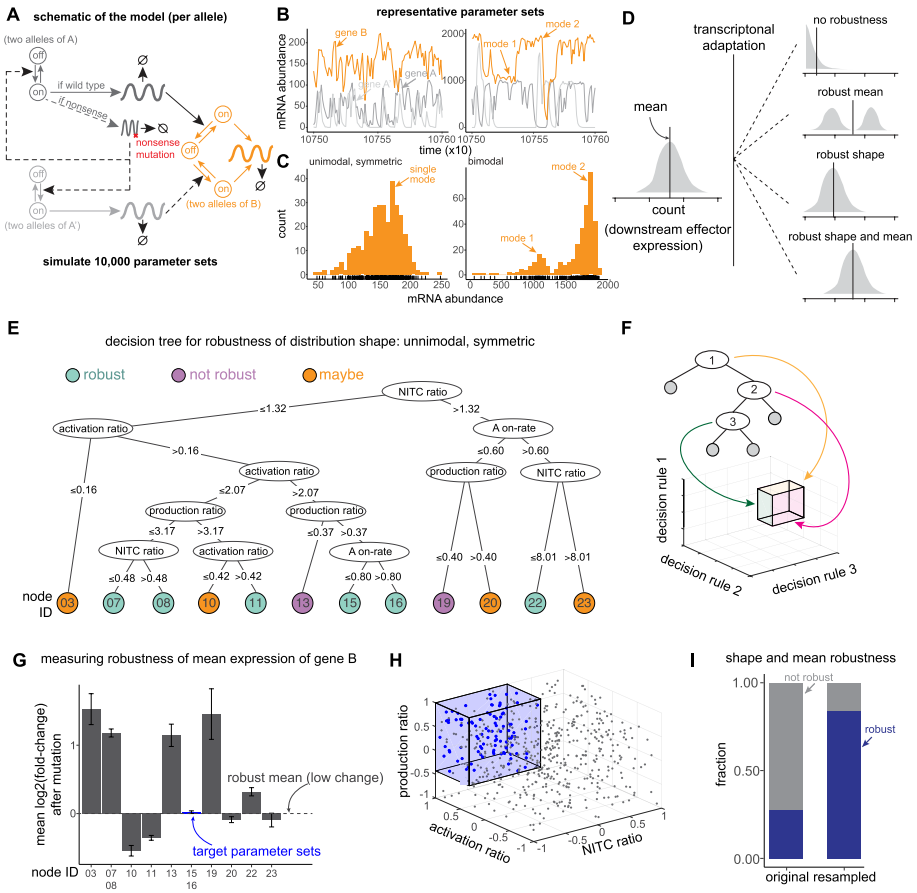


**Fig. 5** Outputs of simulated gene regulatory networks with transcriptional adaptation and robustness to mutation dependency on model parameters. **A** Schematic of a gene regulatory network with transcriptional adaptation to mutation. Two alleles of each gene, with bursty transcription of gene products at each allele. A mutated reference gene, A (dark gray), regulates downstream effector gene B (orange). When mutated, nonsense copies of A product upregulate a paralog of A, called A′ (light gray). A′ can also regulate B, albeit with different strengths. Hill functions are used in propensities for regulatory relationships between gene products and target alleles. See the "Methods" section for full model specification. Parameter descriptions in the table at right of the panel. **B** Example simulation output and inference of single-cell expression distributions from pseudo-single-cells taken every 300 time-steps. See the "Methods" section. **C** Example classification of gene expression distribution shapes. See the "Methods" section for classification algorithm. **D** Analysis schematic: under what network conditions does the network output (i.e., distribution of B expression levels) remain unchanged, either in shape or in average expression level, after mutation of gene A? **E** Decision tree trained on model parameters to classify parameter sets, restricted to those in which B is unimodal symmetric in the wildtype genotype, by whether B distribution remains unimodal symmetric or if it changes distribution shape class. Nodes marked "robust": > 70% of considered parameter sets robust, "maybe": 30–70% robust, "not robust": < 30% robust. **F** Analysis schematic: How parameter subspaces for terminal decision tree nodes relate to rules. Each rule limits the parameter subspace for the terminal node. Colored planes correspond to the decision rule-defined thresholds. When a subspace only encodes one decision rule for a given parameter, the subspace range is limited to the corresponding full parameter space boundary (e.g., if a decision rule sets only a new minimum parameter value, then the maximum parameter value for the subspace is the maximum parameter value of the full parameter space). **G** Mean B expression changes after mutation in each decision tree leaf for unimodal symmetric robust parameter sets. **H** Parameter subspace bounded by decision tree rules for nodes 15 + 16 in **E**, resampled. **I** Enrichment of robustness of both shape and mean for unimodal symmetric distributions of B after mutation, subsampled from the full parameter space and from the subspace marked in **D**. Robustness of mean here defined as absolute log2 fold-change after mutation < 0.35

2: Fig. S7A), we asked if we could capture the frequency of distribution shape change (or lack of shape change) after mutation as a proxy for robustness in distribution shape. For gene product B, we found that while many simulations demonstrated shape class changes after mutation, a large number instead did not (Additional File 2: Fig. S4E), qualitatively similar to our analysis on published datasets (Fig. 4). This robustness in distribution shape after mutation could translate into downstream phenotypic robustness.

In our initial set of simulations, we considered paralogs with no basal expression, i.e., paralogs regulated exclusively by transcriptional compensation. Since paralogs exhibit basal expression in many real world datasets, we ran new simulations incorporating basal paralog expression (adding an additional independent variable: basal paralog expression level). We observed qualitatively similar results in that the simulations produced all 5 distribution classes observed in the original model with no basal expression and that some parameter sets exhibited robustness of distribution shape to mutation (Additional File 2: Fig. S8 and Supplementary Note).

As many human and mouse genes have multiple annotated paralogs [47], we built a model incorporating multiple compensatory paralogs (see the "Methods" section and Additional File 2: Fig. S9). We observed qualitatively similar results on the diversity of distribution shapes and robustness of effector molecule B (Additional File 2: Fig. S9B, C, Supplementary Note, Fig. SN2-4). Similarly, since transcription factors can be repressors [73], we also simulated networks with repressive effects on downstream targets (see the "Methods" section; Additional File 2: Fig. S10), and found the resulting distributions to be broadly consistent with activating networks (Additional File 2: Fig. S10B, C; Supplementary Note, Fig. SN5, 6).

#### Gene expression distribution shape depends on model parameters

The shape and variability of a single cell's gene expression distribution can affect phenotypic penetrance and disease progression. We questioned how compensatory gene regulatory mechanisms might control the shape of emerging population-level distributions. Using summary statistics describing gene B in the heterozygous genotype, we checked whether there were any associations between independent model variables and gene expression distribution summary statistics. Bimodality coefficient features prominently in the distribution shape classification algorithm, so we first focused on parameter associations with bimodality coefficient. We found that the log ratio of B production rates in A- versus A'-directed on-states was more strongly correlated with bimodality coefficient ($r = 0.32$) than other relevant variables (Additional File 2: Fig. S7B).

Beyond associations between individual parameters and summary statistics, we wondered whether parameter combinations were more likely to give rise to particular distribution shape classes upon mutation. While we could visualize the (complicated) densities of different distribution shapes in two dimensions at a time (Additional File 2: Fig. S11), we wanted to quantitatively assess the combined effects of all eight independent variables on the emergent network output. To address this question, we trained decision trees to classify all 5 distribution shapes at once, restricting the tree depth to varying degrees (see the "Methods" section, Additional File 2: Fig. S12). We found that multiple parameter combinations could lead to enrichment or depletion of different distribution shapes. To better understand

parameter combinations predictive of individual distribution shapes, we separately trained decision tree classifiers for each observed distribution class alone: unimodal symmetric, bimodal, right-skewed, left-skewed, and low-average (see Additional File 2: Supplementary Note, Fig. SN1, 7–10). Our analysis highlighted the variable complexity of parameter subspaces leading to distribution shapes of gene B in the heterozygous genotype (see Additional File 2: Supplementary Note, Fig. SN1). Additionally, there are multiple routes or parameter combinations that each can lead to a particular distribution shape. For example, the unimodal symmetric decision tree had a total of 31 significant decision rules up to 6 layers deep per combination (see Additional File 2: Supplementary Note, Fig. SN1). We next asked whether the subspaces defined by a tree's decision rules were indeed predictive of a particular distribution shape and, if so, how strongly so. Resampling of new parameter combinations from constrained subspaces identified by decision trees resulted in strong enrichment of the respective class, demonstrating the utility of our approach (Additional File 2: Fig. S13).

In further exploring the data, we identified an unexpected gene expression shape in one out of more than a hundred randomly inspected gene B expression distributions: trimodal with 3 non-zero modes (Additional File 2: Fig. S6C, D, and Supplementary Note). Examination of the gene expression traces suggested that the rate of B mRNA degradation was too slow to reach zero before an allele was activated by A or A' (Additional File 2: Fig. S6C, D). To verify whether this distribution was associated with specific parameter combinations, we resampled parameter combinations from four progressively larger subspaces centered around the originally observed parameter set, ran new simulations, and inspected the output to check for trimodality (Additional File 2: Fig. S6E, F). Indeed, we found a strong enrichment of trimodal distributions (as high as 94% in the smallest subspace centered around the original parameter set) compared to the full original parameter space (< 1%) (Additional File 2: Fig. S6G). As expected, the enrichment decreased monotonically with increasing size of the sampling hypercube (Additional File 2: Fig. S6G). In sum, this analysis further established that we can use parameter set combinations to predict distribution shapes.

#### Gene expression distribution robustness to mutation is dependent on model parameters

We next asked whether we could identify model predictors of robustness to mutation, insofar as robustness could exist in our model (Fig. 5D). Focusing on unimodal symmetric distributions in the wild-type state, we trained a classifier on whether or not a given parameter set resulted in gene B remaining unimodal symmetric in the heterozygous mutant-A genotype. We found a limited number of significant decision rules for unimodal symmetric robustness: a total of 11 decision rules, up to 5 layers deep, with 12 total groupings (terminal nodes) of parameter sets (Fig. 5E). This parameter space is bounded by the three condensed decision rules: (1) ratio of $r_{add}{}^{NITC}$ to $r_{on,basal,A} \leq 1.32$ (up to moderate strength NITC), (2) ratio of $r_{add}$ of A on B to $r_{add}$ of A' on B > 2.07 (relatively less frequent A'-directed bursting of B), and (3) ratio of $r_{prod}$ of B in the A-directed on state to the $r_{prod}$ of B in the A'-directed on state > 0.37 (A'-directed B production rate can be either weaker or stronger than A-directed, but not by too much) (Fig. 5E, F). When we further accounted for average expression of B as a measure of robustness, only

2 out of 5 terminal nodes (nodes 15 and 16) from shape robustness analysis remained (Fig. 5G). New parameter sets resampled from within the node 15 + 16 parameter subspace were strongly enriched for robustness of shape and average expression level as compared to all parameters (Fig. 5H, I). These results may aid in generating hypotheses for future experiments, for example, to test the constraints on differences between CRISPR target and paralog transcription factors' differential affinity for binding to target regulatory regions (since these constraints could, in theory, map to activation or production ratio differences, for appropriately chosen networks).

## Discussion

We developed a computational framework integrating bioinformatic analysis, mathematical modeling, and machine learning to uncover the genome-wide prevalence and gene regulatory constraints on a recently reported kind of transcriptional adaptation, nonsense-induced transcriptional compensation, in single mammalian cells. We found transcriptional upregulation of paralogs after reference gene mutation to be pervasive, but not necessarily ubiquitous, across cell types and contexts, including cancer, development, and cellular reprogramming. Furthermore, the genes identified as exhibiting possible transcriptional adaptation were neither associated with any single signaling pathway nor did they exhibit any observable molecular functional congruence between each other. Additionally, we did not observe correlation between expression levels of proposed mediators of transcriptional adaptation (e.g., COMPASS complex components) and whether a CRISPR target demonstrated paralog upregulation. Our relatively parsimonious model consisting of transcriptional bursting and stochastic interactions between genes in a biallelic compensatory network could produce a range of population-level distributions of downstream targets upon compensation, underscoring the complex ensemble of fate-space that compensatory networks can access. Finally, our regulon robustness results synthesize two separate earlier analyses: paralog upregulation bioinformatic analysis and single-cell network simulations, in that transcriptional adaptation was associated with downstream regulon gene robustness after transcription factor mutation. Collectively, our computational framework provides a basis for further mechanistic experimental and computational studies on the origins and manifestations of nonsense-mediated transcriptional adaptation.

The fact that transcriptional adaptation occurred across a wide range of processes and for gene sets not necessarily belonging to a single regulatory module or signaling pathway highlights the need to consider their implications when screening for any phenotypic outcomes. One way to address this concern is to perform screens with perturbation methods that avoid nonsense mutations. Techniques such as CRISPRi, already being used in pooled screens [74, 75], or other methods of engineering knockdowns, could be helpful. Alternatively, if knockout is a requirement of the experimental design, engineering whole-gene deletion alleles could help decouple effects of transcriptional adaptation from that of specific gene knockouts. Another opportunity is presented by recently reported combinatorial CRISPR screens (e.g., [24, 74]), which include paired knockout of two or more genes in the same cells, which could identify gene sets for which transcriptional adaptation confounds the outcomes. For example, combined knockout of a reference gene and its paralogs could help to overcome the effects of transcriptional

adaptation, while paired knockout of a reference gene and other interacting genes could help to disentangle reference- vs. paralog-specific functions. When such experimental designs are impractical or infeasible, the interpretation of nonsense-based knockout results could account for possible transcriptional adaptation by concurrent measurement of the expression of paralogs of the knockout target. In this way, caution must be taken in interpreting the phenotypic changes, or lack thereof, if a knockout target shows paralog upregulation.

Recent advances in sequencing and genome editing technologies have enabled perturbation and profiling of the molecular makeup of single cells at unprecedented throughput. For greater resolution of differences between genetic perturbation methods, high-throughput parallel treatments of the same target genes with RNA interference, CRISPRi, and Cas9-based knockouts could reveal specific effects of post-transcriptional, epigenetic, or mutation-based methods. Perhaps such parallel experiments could reveal transcriptional adaptation, or yet unknown mechanisms, by which cells retain robustness to genetic perturbations. Similarly, the adoption of recent single-cell CRISPR screening frameworks, such as CROP-seq and Perturb-seq [49–51], coupled with high-depth sequencing can lead the way in identifying single-cell manifestations of transcriptional adaptation. These experimental findings can, in principle and if at higher quantitative resolution, be projected onto the distributions from our theoretical formulations. Given the non-linearities associated with sequencing datasets, bona fide gene targets identified from sequencing studies can be tested in single cells with single-molecule fluorescent in situ techniques, which measure the absolute expression counts in individual cells for greater quantitative resolution [76]. Similarly, single-cell methods such as CROP-seq and Perturb-seq [49–51] only offer fixed snapshots in time, limiting our ability to discriminate between two conceptually distinct scenarios possible: mutation-induced relative increases of paralog expression in surviving cells irrespective of initial expression levels vs. selection for pre-existing stably relatively higher paralog-expressing (rare) cells in the baseline population. Coupling single-cell RNA sequencing datasets with cellular lineage information, leveraging recently reported barcoding technologies [8, 77, 78] that enable longitudinal tracking of individual cells before and after nonsense mutation for genes exhibiting transcriptional adaptation, can address such questions on the dynamics of transcriptional adaptation in single cells.

The mapping between simulation and wet-lab experiment can uncover plausible network and parameter constraints for individual compensating genes and could provide evidence for particular compensating gene regulatory steps affected by transcriptional adaptation. For example, one study used single-molecule approaches to study the effect of nonsense-mediated decay in U2OS cells with and without nonsense immunoglobulin-μ genes. They showed that UPF1 depletion increased the speed of transcriptional elongation in the wild-type but not in the nonsense immunoglobulin-μ gene [79]. Furthermore, regulatory network mappings at a single-cell level could also help explain incomplete phenotypic penetrance reported in association with transcriptional adaptation. Another set of questions center around whether gene length, number of introns and exons, chromosomal locations, and chromatin landscape play a role in which gene families exhibit nonsense-induced transcriptional compensation. Additionally, such mappings can help with the design and interpretation of functional genetic

screens by taking into account genes known to be exhibiting transcriptional adaptation and the extent of its impact. The breadth of genes that appear to have transcriptional compensation also invites study of potential negative consequences of nonsense-induced paralog—or other related gene—upregulation. Might some compensatory changes be deleterious and, if so, could such deleterious changes explain select negative phenotypes previously ascribed to haploinsufficiency or gene dosage effects [80]? In a similar vein, our framework could be extended to analyze cases where paralogs are downregulated upon Cas9-induced nonsense mutations, potentially revealing new biology.

During the process of mining published datasets from disparate studies, we found several cases where the phenomenology could reflect what the two landmark studies [14, 15] term as "transcriptional adaptation," but the results were not explicitly contextualized (nor definitively proven) as such. Localization of UPF proteins to compensating loci dependent on nonsense-mutated RNA in [79], tubulin family upregulation after *Tubb4a* mutation in [36], and others, such as the knockdown-knockout discrepancies reviewed in [2], could contribute to the field of transcriptional adaptation. Since our work and other recent studies point to the plausible presence of transcriptional adaptation across many contexts, e.g., even for coupling between maternal and zygotic gene regulation during early embryogenesis [81], we propose a push towards consistent and universal usage of the term "transcriptional adaptation" to describe upregulation of compensating related genes after a mutation dependent on the mutated transcript. A common consensus may facilitate faster discovery and reconciliation of paradoxical findings across contexts moving forward.

One limitation of our work is that a majority of the analysis was performed on datasets from bulk RNA sequencing studies, limiting a quantitative single-cell mapping with simulations. As single-cell sequencing datasets, and single-cell transcriptomics via other methods that enable absolute expression counts, such as SeqFISH, MERFISH, and optical pooled screens [82–84], become more accessible, bioinformatic analysis can inform model architecture and parameters and move towards more predictive models. Another limitation of our framework is that we focused primarily on mice and human datasets given the breadth of available datasets. In principle, our bioinformatic pipeline can be generalized to include other animal systems to reveal both species-specific and universal gene targets displaying transcriptional compensation [14, 15]. Lastly, simulations of gene regulatory networks are inherently simplifying, and while we specified reactions and assumptions that have been shown to model small numbers of interacting genes well [71, 72, 85], these models do not account for all regulatory interactions in a cell explicitly [39, 66, 86].

## Conclusions

In summary, our integrative analyses highlight the genome-wide prevalence of and gene regulatory constraints on transcriptional adaptation in mammalian cells. We show that upregulation of paralogs after reference gene mutation is common, but not necessarily ubiquitous, across cell types and contexts. This behavior is not restricted to genes in specific pathways or encoding products with specific molecular functions. Transcription factors that show evidence of transcriptional adaptation have downstream regulons that are more robust to the transcription factor's mutation compared to regulons

of transcription factors without mutation-induced paralog upregulation. Lastly, simulations of a gene regulatory network with transcriptional adaptation produce a variety of expression distributions of downstream targets upon compensation, recapitulating observed diverse regulon expression changes after transcription factor mutation. Altogether, our work provides a strong foundation for future mechanistic experimental and computational studies of transcriptional adaptation.

## Methods

### Selection of CRISPR-Cas9 transcriptomics datasets

We searched published literature, preprints, and the Gene Expression Omnibus (GEO: https://www.ncbi.nlm.nih.gov/geo/) for RNA sequencing datasets generated from experiments designed to measure differential gene expression after Cas9-mediated knockout of a gene of interest. We used the following search terms: "CRISPR/Cas9" or "CRISPR-Cas9" and "RNA-seq." We focused our search for publicly available data on the GEO RNA-seq Experiments Interactive Navigator (GREIN). GREIN contains processed data from thousands of GEO entries with human, mouse, and rat RNA-seq samples using a common pipeline for alignment, quality control, and transcript quantification [87]. In GREIN, we used search terms "CRISPR/Cas9" or "CRISPR-Cas9." Based on these search results, we manually checked the experimental designs of more than 200 publicly available RNA-seq experiments. We only considered experiments in which (1) there was CRISPR/Cas9-based knockout of the target, in which the stated strategy was not to target intergenic regulatory sequences, (2) annotated matched control samples treated with non-template control gRNA, and (3) multiple replicates of both targeted and control RNA-seq samples. We prioritized studies with multiple knockout targets, which enabled us to check for the presence or absence of paralog upregulation for as many targets as possible. For the few non-GREIN datasets included here, we only considered studies that provided mapped read counts per gene ID or, optimally, also provided processed differential gene expression calculations. Ultimately, we found 36 studies with datasets meeting inclusion criteria described above, described in Additional File 1: Table S1 [14, 48, 50, 58, 88–115].

### Identification of paralogs of knockout targets

For bulk RNA-seq datasets, we queried the Ensembl database version 110 for paralogs of knockout targets, using Ensembl REST API (Version 15.6) [47]. We searched by Gene Symbol, and extracted paralogs of all returned Ensembl IDs. We extracted both gene identifiers and Ensembl-annotated coding sequence overlap percentages between knockout targets and each of their paralogs. Ensembl gene IDs were then converted to standardized gene names using g:Profiler (Version e109_eg56_p17_1d3191d) [116]. For Perturb-seq data from [52], we used the BiomaRt package v2.40.5 in R v3.6.1 to search Ensembl version 105 for paralogs, searching by Gene Symbol and extracting all returned paralog Gene Symbols [117].

### Differential gene expression assessment

We wanted to identify differentially expressed genes across the dozens of knockout samples we reanalyzed. We used DESeq2 for differential expression analysis [118]. When

available, we used author-provided gene expression change calculations based on DESeq2 (for results from [58]). For all remaining datasets, for which DESeq2 results were not already available, the authors did provide mapped count data on GEO and/or they were available on GREIN. For these count-based results, we implemented DESeq2 ourselves, using the PyDESeq2 package, using default settings, comparing knockout samples against the matched controls from their respective studies [119].

For these studies, we implemented filters to consider knockout targets that we would a priori expect to have some detectable loss of gene dosage that would need to be compensated for by transcriptional adaptation. We first confirmed that the average library size of considered samples was at least approximately 1 million reads per sample. We then included genes only if they were expressed at a level of 10 raw counts or higher across all samples. We chose to classify paralogs as upregulated if DESeq2 reported an adjusted $p$-value $\leq 0.05$ and a log2 fold-change $\geq 0.5$. In supplementary analyses, we also show results when paralogs are classified as upregulated using either (1) only the adjusted $p$-value $\leq 0.05$ filter or (2) adjusted $p$-value $\leq 0.05$, log2 fold-change $\geq 0.5$, and basemean $\geq 10$ filters. The final analysis included all knockout target genes with any significant paralog differential expression, up or down, irrespective of log2 fold-change.

#### Checking for genomic regulatory feature association with paralog upregulation

We downloaded annotated associations between human enhancer or promoter elements and human genes from the GeneHancer v5.19 database (https://www.genecards.org/Guide/DatasetRequest, last accessed April 4, 2024). We converted gene identifiers to GeneCards IDs using gprofiler2 in R. We mined the GeneHancer database for associations between enhancers or promoters and genes. For assessment of paralogs only, we considered all knockout target-paralog pairs together. We checked for a difference in odds of paralog upregulation based on whether the paralog and CRISPR target shared any annotated enhancer or promoter using Fisher's exact test in R. For the analysis of all genes, for each CRISPR target, we repeated the check of shared enhancers or promoters, and for each CRISPR target checked for a difference in odds of upregulation using Fisher's exact test.

We downloaded annotated topologically-associated domains (TAD) coordinates in the hg19 human genome (http://dna.cs.miami.edu/TADKB/download/TAD_annotations.tar.gz, last accessed April 29, 2024). We considered annotated transcription start sites for any Ensembl transcript for each considered gene, identified using biomaRt, pulling from the February 2014 build of hg19. Consistent with the publicly accessible browser built by the authors of TADKB [60], we considered annotated TADs at 50 kb resolution called using the directionality index method, in any of the human cell types in the database. For each CRISPR target, paralog, or any other gene, we searched for annotated TADs overlapping any of their annotated transcription start sites. For the paralog analysis, we combined all paralog-target pairs together and checked for a difference in odds of paralog upregulation based on co-location in any TAD shared with the knockout target using Fisher's exact test in R. For the analysis of all genes, for each CRISPR target, we repeated the check of overlapping TADs, and for each CRISPR target, we checked

for a difference in odds of other-gene upregulation using Fisher's exact test. For *p*-value adjustment, we used the Benjamini–Hochberg method in p.adjust() in R.

#### Gene expression correlation analysis

For human genes, we downloaded GTEx Analysis V8 median expression levels (in TPM) in each tissue from the GTEx portal on April 15, 2024 (GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_median_tpm.gct.gz). For each pair of genes, one knockout target and one paralog, we plotted the median expression in each tissue of the respective species and calculated a Spearman correlation coefficient on log(TPM + 1). We then compared this correlation coefficient to the log2fold-change of the paralog as a check for association between cross-tissue average expression level correlation with the CRISPR target and whether a paralog displayed upregulation after CRISPR target knockout. We further calculated a Spearman correlation coefficient to check for association between the pairwise correlations and paralog log2(fold changes).

#### Estimation of expected frequency of paralog upregulation per knockout target

The null hypothesis for analysis of paralog upregulation is that for any group of paralogs of a knockout target, the number of paralog genes upregulated after knockout is simply reflective of randomly selecting any similarly expressed genes in the dataset and checking whether they were upregulated. In order to check whether the paralog upregulation pattern observed for a particular knockout target was reflective of randomly selecting similarly expressed genes in the dataset (instead of the paralogs), we developed an algorithm for bootstrapping the null distribution of paralog upregulation frequency for each knockout target. For each knockout target, the algorithm is implemented as follows:

1. Rank order all genes in the dataset by basemean across all samples
2. For each paralog, randomly select a gene within 51 ranks
3. For each randomly selected similarly expressed gene, check the fold change after knockout and whether the gene qualifies as upregulated based on dataset-specific thresholds (in figure legends)
4. Count the number of upregulated randomly selected genes and divide by the total number of paralogs for a bootstrap sample of the paralog upregulation frequency
5. Repeat steps 2–4 10,000 times to build an empirical null distribution of the paralog upregulation frequency
6. To calculate a *p*-value, calculate the fraction of the empirical null distribution that is at least as large as the observed fraction of paralogs that are upregulated

#### Perturb-seq-based single-cell gene expression reanalysis

We wanted to identify possible changes in single-cell gene expression distributions after knockout of a library of CRISPR targets. Therefore, we reanalyzed Cas9-based pooled knockout single-cell RNA-seq, Perturb-seq, data from [52]. We downloaded published processed log-transformed UMI-based transcript quantification tables (in log(TPM + 1)) from https://singlecell.broadinstitute.org/, accession SCP1064, "Control" condition, last

accessed June 15, 2023. For the main analysis, we only considered knockout targets with sufficient cells for a minimal quantitative analysis: a minimum of 2000 UMI per cell, at least 30 cells total, with no fewer than 5 cells annotated in any one of the three included targeting guide RNAs. In a supplementary analysis to identify lower-confidence targets with possible transcriptional adaptation, we considered removing the 30-cell and 5-cell filters. For transcript quantification, within each guide, we either counted the number of cells with non-zero expression and divided by total cells for that guide (for percent-positive), or we averaged expression levels over all cells for that guide (for mean). Within each gene within a given condition (nontemplate controls or targeted for a given gene), we averaged over all appropriate guides.

#### Regulon robustness analysis

For bulk RNA-seq data-derived regulon gene expression analyses, we focused on human and mouse transcription factors, as defined by the most recent version of AnimalTFDB3, last accessed November 21, 2023 [120]. We searched for overlapping regulons between a knockout target and the paralog gene of interest in DoRothEA, only considering downstream genes with annotation confidence level A, B, or C (out of a possible range of A-E, see original source for evidence level descriptions) [63, 64]. We compared regulon genes for transcription factor CRISPR target-paralog pairs in the bulk RNA-seq dataset that demonstrated possible transcriptional adaptation as defined by significant paralog upregulation frequency ($p < 0.1$, see the "Estimation of expected frequency" section, above). There were 68 annotated regulon genes across the four target-paralog pairs with transcriptional adaptation versus 138 annotated regulon genes for all target-paralog pairs without transcriptional adaptation. Regulon genes were considered differentially expressed if DESeq2 adjusted $p$-value $< 0.05$ and abs(log2FoldChange) $> 0.5$. We calculated a $p$-value between the groups of regulon genes using Fisher's exact test, testing whether the odds of regulon gene differential expression were different between the transcriptional-adaptation and no-transcriptional-adaptation groups.

For Perturb-seq data-derived gene expression distribution analyses, we chose to focus on human transcription factor genes, as defined by the most recent version of AnimalTFDB3, last accessed July 28, 2023 [120]. We searched for overlapping regulons between a knockout target and the paralog gene of interest in DoRothEA, only considering downstream genes with annotation confidence level A, B, or C (out of a possible range of A-E, see original source for evidence level descriptions) [63, 64]. We compared regulon genes for transcription factor CRISPR target-paralog pairs in the Perturb-seq dataset that demonstrated possible transcriptional adaptation as defined by having a top-100 paralog, versus all pairs that that did not demonstrate transcriptional adaptation, as defined by both not having any top-100 paralogs and being a pair in the interquartile range of changes in paralog expression after knockout. There were 55 annotated regulon genes for the three target-paralog pairs with transcriptional adaptation versus 439 annotated regulon genes for all target-paralog pairs without transcriptional adaptation.

The average difference in expression of regulon genes in both groups was approximately zero, with a spread of values about that mean (Fig. 4B). In order to test for transcriptional adaptation-associated buffering of gene expression changes of regulon genes after CRISPR target mutation, we performed two checks for significance of a difference

in the spread size between the two groups; a larger spread would indicate a larger number of extreme expression changes. The first test was an asymptotic test of difference in coefficient of variation for groups of unequal size, using the cvequality v0.1.3 package in R [121, 122]. The second test was a check of the plausibility of the null hypothesis that equal size samples from the transcriptional-adaptation and no-transcriptional-adaptation groups have the same standard deviation. We empirically downsampled ($n = 49$ with percent-positive < 75% in controls, of 55 total genes, same as transcriptional-adaptation group size) the no-transcriptional-adaptation group without replacement, 1000 times, to generate an empirical null distribution of sample standard deviations from the no-transcriptional-adaptation group. We then compared the observed transcriptional-adaptation group standard deviation to this distribution and observed that it was among the lowest downsampled values (Fig. 4G), suggesting that the transcriptional-adaptation group is unlikely to have a similar sample standard deviation to the no-transcriptional-adaptation group.

#### Gene set enrichment analysis

We wondered whether genes involved in any specific biological processes or contexts were overrepresented in the set of genes whose paralogs were significantly upregulated. Therefore, we performed gene set enrichment analysis to check for over-enrichment of any Gene Ontology—Biological Process terms, comparing the following sets of hits against their respective background sets of tested CRISPR targets.

1. Bulk RNA-seq CRISPR targets with bootstrap $p$-value < 0.1, against a background set of all fully analyzed, human genes only.
2. Single-cell RNA-seq CRISPR targets from Frangieh et al., 2021, among those meeting minimum cell count thresholds above, with any paralog in the top-100 largest increase in percent positive list, or if control percent positive > 0.75 with any paralog in top-100 largest increase in mean list, against background of all targets in library
3. Combined (1) and (2) hits against their combined respective backgrounds

We used the clusterProfiler R package v3.12.0 for gene ontology over-representation testing [55].

#### Networks

Gene regulatory networks are represented as directed graphs. Genes are nodes, and regulatory relationships are edges (e.g., A stimulates B leads to an edge from node A to node B; Fig. 5). The biological mechanism presented in recent studies on nonsense-induced transcriptional compensation implies a minimum set of regulatory relationships between an ancestral regulator, its paralog genes, and a downstream target gene [14, 15]. We model gene regulatory networks with, for each gene, two alleles with transcriptional burst activity independent of each other, consistent with observations of transcriptional burst regulation [123]. The edges between a given regulator gene product and the target gene alleles are set at equal weight, reflecting no regulatory differences at the allele level.

For an upstream regulator gene A that has nonsense-induced transcriptional compensation, there is at least one compensating gene, A' (referred to as a paralog here). Gene

A' encodes product $A'$. The downstream regulatory target of A and A' is gene B. Gene B encodes product $B$. Upon mutation of A, the mutant allele of gene A produces product $A_{nonsense}$ instead of product $A_{wt}$. During nonsense-induced transcriptional compensation, $A_{nonsense}$ can regulate alleles of gene A (mutated or not), as well as gene A', but no longer regulates gene B.

#### Core transcriptional bursting model

Our network is built of component genes whose alleles are each modeled with an expanded version of the classic telegraph model, similar to prior work (Fig. 5; [66]). Each allele can reversibly enter an active ("on"; transcribing) or inactive ("off"; quiescent) state, with high or low (by default 0) production rates of that gene's product, respectively. We assume that any gene product is effectively immediately translated or processed to the relevant functional form capable of regulating a downstream target. In the case of $\{A_{wt}, A', B\}$, this assumption applies to post-transcriptional regulation, translation, and post-translational processing. In the case of $A_{nonsense}$, this assumption applies to the hypothesized but unknown mechanisms of nonsense-induced transcriptional compensation. Therefore, for alleles of genes A and A', and their respective products, there are five consistent reactions:

$$off \rightarrow on$$

$$on \rightarrow off$$

$$off \rightarrow off + mRNA$$

$$on \rightarrow on + mRNA$$

$$mRNA \rightarrow \varnothing$$

Since for all genes we assume there is no leaky expression in the off state, in the following sections we ignore the possible $off \rightarrow off + mRNA$ reaction. Alleles of gene B are described by a related but larger set of reactions to reflect the different consequences of regulation by $\{A, A'\}$, and are described below.

As previously described, we make use of reaction propensities under the assumption of the law of mass action, where each propensity function $p_i(x)dt$ gives the probability of reaction $R_i$ occurring in the time step $dt$, for a small $dt$. In the models presented here, gene regulation affects either the reaction rate of an allele entering the active state or entering the inactive state.

##### *Activation (positive regulation) model*

In a model of activating interacting genes in a gene regulatory network, all regulation affects the reaction rate of activation of target alleles. In order to simulate differential effects of A-stimulated B alleles versus A'-stimulated B alleles, we used an expanded gene model for B. The expanded model allows for multiple on-states, corresponding to

the different upstream regulators; the different on-states are allowed to have different production rates. These different production rates can reflect biological differences in gene regulation at B loci, including, but not limited to, chromatin changes, differences in transcription factor recruitment, and effects on transcriptional machinery. Therefore, for gene B the full set of reactions is:

$$off \rightarrow on^{(A)}$$

$$off \rightarrow on^{(A')}$$

$$on^{(A)} \rightarrow off$$

$$on^{(A')} \rightarrow off$$

$$on^{(A)} \rightarrow on^{(A)} + B$$

$$on^{(A')} \rightarrow on^{(A')} + B$$

$$B \rightarrow \varnothing$$

The rates in the reactions described in Fig. 5, S4 are:

| Parameter | Description |
|---|---|
| $r_{on}$ | Activation rate of an allele |
| $r_{off}$ | Inactivation rate of an allele |
| $r_{prod}$ | Production rate of mRNA from an active allele |
| $r_{deg}$ | Degradation rate of mRNA |
| $r_{add}^{NITC}$ | Additional activation of an allele (A or A') upon nonsense-induced transcriptional compensation caused by product $A_{nonsense}$ |
| $r_{add}^{A,B}$, $r_{add}^{A',B}$ | Activation of B by A or A' to one of two respective B active states specified in the model |
| $d$ | Factor by which the mRNA production rate of B is lower in the A'-directed B active state than in A-directed B active state |
| $n$ | Hill coefficient |
| $k$ | Dissociation constant of the Hill function |

The full model is therefore described as follows. The basal on-rates of A' and B alleles are assumed to be 0. The mRNA production rate in the off state of all alleles is fixed at 0.

| Reaction | Reaction rate | Reaction propensity |
|---|---|---|
| **Gene A alleles, wildtype** | | |
| $off \rightarrow on$ | $r_{on}^{A} + r_{add}^{NITC} \frac{A_{nonsense}{}^{n}}{k^{n}+A_{nonsense}{}^{n}}$ | $(r_{on}^{A} + r_{add}^{NITC} \frac{A_{nonsense}{}^{n}}{k^{n}+A_{nonsense}{}^{n}}) * off$ |
| $on \rightarrow off$ | $r_{off}$ | $r_{off} * on$ |
| $on \rightarrow on + A_{wt}$ | $r_{prod}$ | $r_{prod} * on$ |
| $A_{wt} \rightarrow \varnothing$ | $r_{deg}$ | $r_{deg} * A_{wt}$ |
| **Gene A alleles, nonsense mutant** | | |
| $off \rightarrow on$ | $r_{on}^{A} + r_{add}^{NITC} \frac{A_{nonsense}{}^{n}}{k^{n}+A_{nonsense}{}^{n}}$ | $(r_{on}^{A} + r_{add}^{NITC} \frac{A_{nonsense}{}^{n}}{k^{n}+A_{nonsense}{}^{n}}) * off$ |
| $on \rightarrow off$ | $r_{off}$ | $r_{off} * on$ |

| Reaction | Reaction rate | Reaction propensity |
|---|---|---|
| $on \rightarrow on + A_{nonsense}$ | $r_{prod}$ | $r_{prod} * on$ |
| $A_{nonsense} \rightarrow \varnothing$ | $r_{deg}$ | $r_{deg} * A_{nonsense}$ |
| **Gene A′ alleles** | | |
| $off \rightarrow on$ | $r_{on}^{A'} + r_{add}^{NITC} \frac{{A_{nonsense}}^n}{k^n + {A_{nonsense}}^n}$ | $(r_{on}^{A'} + r_{add}^{NITC} \frac{{A_{nonsense}}^n}{k^n + {A_{nonsense}}^n}) * off$ |
| $on \rightarrow off$ | $r_{off}$ | $r_{off} * on$ |
| $on \rightarrow on + A'$ | $r_{prod}$ | $r_{prod} * on$ |
| $A' \rightarrow \varnothing$ | $r_{deg}$ | $r_{deg} * A\prime$ |
| **Gene B alleles** | | |
| $off \rightarrow on^{(A)}$ | $r_{add}^{A,B} \frac{{A_{wt}}^n}{k^n + {A_{wt}}^n}$ | $(r_{add}^{A,B} \frac{{A_{wt}}^n}{k^n + {A_{wt}}^n}) * off$ |
| $off \rightarrow on^{(A')}$ | $r_{add}^{A\prime,B} \frac{{A\prime}^n}{k^n + {A\prime}^n}$ | $(r_{add}^{A',B} \frac{{A'}^n}{k^n + {A'}^n}) * off$ |
| $on^{(A)} \rightarrow off$ | $r_{off}$ | $r_{off} * on^{(A)}$ |
| $on^{(A')} \rightarrow off$ | $r_{off}$ | $r_{off} * on^{(A')}$ |
| $on^{(A)} \rightarrow on^{(A)} + B$ | $r_{prod}$ | $r_{prod} * on^{(A)}$ |
| $on^{(A')} \rightarrow on^{(A')} + B$ | $\frac{r_{prod}}{d}$ | $\frac{r_{prod}}{d} * on^{(A')}$ |
| $B \rightarrow \varnothing$ | $r_{deg}$ | $r_{deg} * B$ |

where $on, on^{(A')}, on^{(A')}, off \in \{0, 1\}$, for A and A' $on + off = 1$, for B $on^{(A)} + on^{(A\prime)} + off = 1$. Products $A_{wt}, A_{nonsense}, A', B$ represent the count of their respective gene products at a given time. Parameter values in reaction rates are explained in the previous table, above.

### *Multiple-paralog model*

Several real genes have multiple annotated paralogs, which in principle could all potentially compensate for mutations in the CRISPR-target gene [14, 15, 47]. Therefore, we also developed a gene regulatory network model in which gene A can be compensated for by two paralogs, genes A'1 and A'2 with products $A'_1, A'_2$ respectively. Similar to the model in the previous section, each allele of A'1 and A'2 is regulated by nonsense-induced transcriptional compensation for A. An expanded model accounting for differences in regulation of B alleles by gene products $A'_1, A'_2$ leads to 3 possible on states: those directed by $A_{wt}, A'_1, A'_2$ respectively.

| **Gene B alleles** | | |
|---|---|---|
| $off \rightarrow on^{(A)}$ | $r_{add}^{A,B} \frac{{A_{wt}}^n}{k^n + {A_{wt}}^n}$ | $(r_{add}^{A,B} \frac{{A_{wt}}^n}{k^n + {A_{wt}}^n}) * off$ |
| $off \rightarrow on^{(A'_1)}$ | $r_{add}^{A'_1,B} \frac{{A'_1}^n}{k^n + {A'_1}^n}$ | $(r_{add}^{A'_1,B} \frac{{A'_1}^n}{k^n + {A'_1}^n}) * off$ |
| $off \rightarrow on^{(A'_2)}$ | $r_{add}^{A'_2,B} \frac{{A'_2}^n}{k^n + {A'_2}^n}$ | $(r_{add}^{A'_2,B} \frac{{A'_2}^n}{k^n + {A'_2}^n}) * off$ |
| $on^{(A)} \rightarrow off$ | $r_{off}$ | $r_{off} * on^{(A)}$ |
| $on^{(A'_1)} \rightarrow off$ | $r_{off}$ | $r_{off} * on^{(A'_1)}$ |
| $on^{(A'_2)} \rightarrow off$ | $r_{off}$ | $r_{off} * on^{(A'_2)}$ |
| $on^{(A)} \rightarrow on^{(A)} + B$ | $r_{prod}$ | $r_{prod} * on^{(A)}$ |
| $on^{(A\prime)} \rightarrow on^{(A'_1)} + B$ | $\frac{r_{prod}}{d}$ | $\frac{r_{prod}}{d} * on^{(A'_1)}$ |
| $on^{(A\prime)} \rightarrow on^{(A'_2)} + B$ | $\frac{r_{prod}}{d*s}$ | $\frac{r_{prod}}{d*s} * on^{(A'_2)}$ |
| $B \rightarrow \varnothing$ | $r_{deg}$ | $r_{deg} * B$ |

where $on^{(A'1)}, on^{(A'2)} \in \{0,1\}$, for A and A' $on + off = 1$, for B $on^{(A)} + on^{(A'1)} + on^{(A'2)} + off = 1$. Parameter values in reaction rates are explained in the previous section, above, with the addition of a new ratio:

| Parameter | Description |
|---|---|
| $s$ | Factor by which the mRNA production rate of B is lower in the A1′-directed B active state than in A′2-directed B active state |

#### *Repression (negative regulation) model*

We also explored gene regulatory network models in which the regulator products $A, A'$ are inhibitors of gene B rather than activators. In this model, although nonsense-induced transcriptional compensation continues to activate expression of A and A', $A, A'$ products increase the rate of B allele inactivation instead of activation. To simulate differential effects on B alleles, inspired by differential gene regulatory effects such as chromatin modifications and repressive transcription factor recruitment, we changed the gene model of B to include two off states (one directed by $A$ and one by$A\prime$), with one on state. The two off states could have different rates of reversion to the active state, thereby simulating more or less repressed loci. Therefore, while the model confined to the regulation of genes A and A' remain the same as above, the model for gene B becomes:

| Gene B alleles | | |
|---|---|---|
| $off^{(A)} \rightarrow on$ | $r_{on}$ | $r_{on} * off^{(A)}$ |
| $off^{(A')} \rightarrow on$ | $\frac{r_{on}}{d_{inh}}$ | $\frac{r_{on}}{d_{inh}} * off^{(A')}$ |
| $on \rightarrow off^{(A)}$ | $r_{add}^{A,B} \frac{{A_{wt}}^n}{k^n + {A_{wt}}^n}$ | $r_{add}^{A,B} \frac{{A_{wt}}^n}{k^n + {A_{wt}}^n} * on$ |
| $on \rightarrow off^{(A')}$ | $r_{add}^{A',B} \frac{{A'}^n}{k^n + {A'}^n}$ | $r_{add}^{A',B} \frac{{A'}^n}{k^n + {A'}^n} * on$ |
| $on \rightarrow on + B$ | $r_{prod}$ | $r_{prod} * on$ |
| $B \rightarrow \varnothing$ | $r_{deg}$ | $r_{deg} * B$ |

#### *Parameters*

We sought to characterize the breadth of possible gene regulatory network outputs given the presence of a transcriptional adaptation regulatory interaction under biologically plausible conditions [70, 72]. Extensive prior research has established transcriptional bursting as a core model of gene expression in eukaryotes. In the bursting model with zero leaky expression as presented here, upon activation of an allele the steady state expected mRNA abundance derived from that allele rises from 0 to $r_{prod}/r_{deg}$. When mRNA abundance is high enough to exceed a threshold specified by the dissociation constant of the Hill function ($k$) above, there is a higher probability of activation of alleles of downstream targets, further modulated by $r_{add}$ parameters, above. The dissociation constant is defined as:

$$k = x * \frac{r_{prod}}{r_{deg}}$$

where $x$ is the fraction of the active-steady-state expression level required to be exceeded for regulation of the target allele. In our simulations, we use $x = 0.5$ to represent regulation that can occur with some degree of expression of an upstream regulator, so that regulation is not presumed to require prolonged active high-expression states, consistent with published burst duration measurements [70]. Other studies have also simulated transcriptional bursts in gene regulatory networks based on these parameter ranges [72].

Published telegraph model inferred parameter ranges based on allele-resolved single-cell RNA sequencing data provide a useful guide to ranges for the core allele-level parameters [70]. We reanalyzed the data in Larsson et al., 2019, to summarize, for each reported gene expressed in mouse fibroblasts, relative to the degradation rate ($r_{deg}$ arbitrarily fixed at 1), what are the inferred values of $r_{on}, r_{off}, r_{prod}$. In order to preserve overall burst frequency and duration ranges, we focused on the calculated values of: $r_{on}$, $\frac{r_{on}}{r_{off}}$, $r_{prod}$. Based on the observed results, we conducted simulations over parameter ranges spanning approximately 2–3 orders of magnitude:

| Parameter | Meaning | Lower limit | Upper limit |
|---|---|---|---|
| $r_{deg}$ | Degradation rate of mRNA | 1 | 1 |
| $r_{prod}$ | Production rate of mRNA in on state | 1 | 1000 |
| $r_{on}$ | Activation rate of an allele (by default, an A allele) | 0.1 | 10 |
| On/off ratio | $\frac{r_{on}}{r_{off}}$; the factor by which the rate of inactivation is lower than activation | 0.01 | 2 |

Consistent with prior studies, we also explored a range of values of Hill coefficient $n$, to ensure that we adequately sampled over different steepness levels of the Hill function representing regulatory relationships [66].

| Parameter | Meaning | Lower limit | Upper limit |
|---|---|---|---|
| Hill coefficient | $n$; the Hill coefficient. How "switch-like" regulatory effects are | 0.1 | 5 |

Extending the basic allele-level telegraph model to our nonsense-induced transcriptional compensation gene regulatory network model, we needed to pick reasonable ranges for other parameters. For the primary simulations discussed in Fig. 5, we assumed zero basal activation of paralog A' alleles. We focused on characterizing the relative strengths of interactions, as that directly reflects the differences between a regulator, its paralog, and their downstream target rather than absolute simulation parameter values. Therefore, in order to explore network outputs over similarly large ranges of other interaction strengths (roughly two orders of magnitude), and to ensure that activation rates were at least partially overlapping with the observed rates from Larsson et al., 2019, we further considered rate ratios as follows:

| Parameter | Meaning | Lower limit | Upper limit |
|---|---|---|---|
| NITC ratio ($\delta_N$) | $\frac{r_{on}^{(A)}}{r_{add}^{NITC}}$; the factor by which the NITC-mediated contribution to activation rate compares to the reference gene (A) activation rate | 0.1 | 10 |

For each of the different models, we sampled over several other rate and ratio model parameters.

In the base case, a model with one paralog and stimulating regulation of gene B, we also considered:

| Parameter | Meaning | Lower limit | Upper limit |
|---|---|---|---|
| $r_{add}^{A,B}$ | Contribution to activation of B by A to A-directed B active state | 0.1 | 10 |
| A/A′ activation ratio ($\delta_A$) | $\frac{r_{add}^{A,B}}{r_{add}^{A',B}}$; the factor by which A is more effective at activating B alleles to A-directed active state than A′ to A′-directed active state | 0.1 | 10 |
| A/A′ production ratio ($\delta_P$) | $d$; the factor by which the mRNA production rate of B is lower in the A′-directed B active state than in A-directed B active state | 0.1 | 10 |

In a set of simulations of the stimulatory model including non-zero paralog A' expression at baseline, we also sampled over:

| Parameter | Meaning | Lower limit | Upper limit |
|---|---|---|---|
| Basal A′ ratio | $\frac{r_{on}^{(A)}}{r_{on}^{(A')}}$; the factor by which A allele activation rates are higher than A′ alleles at baseline | 1 | 100 |

In the model expanded to consider multiple paralogs, we conducted simulations with and without basal paralog expression. In one set of simulations, we fixed the effects of both paralogs, A'1 and A'2, on B, to be equal. In another set of simulations, we sampled over values of a new ratio describing how expression varies between A'1-directed and A'2-directed B-active states.

| Parameter | Meaning | Lower limit | Upper limit |
|---|---|---|---|
| A'1/A'2 prod ratio ($\delta_R$) | $s$; Factor by which the mRNA production rate of B is lower in the A'1-directed B active state than in A'2-directed B active state | 1 | 100 |

In a model of inhibitory regulation of gene B, instead of A/A' add-on ratio and A/A' prod ratio, we sampled over parameter ratios: A/A' add-off ratio and B,A/B,A' on ratio, to reflect the regulator contributions to inactivation rates to their respective B off states and their respective off states' rates of activation (to a common single active state; see Additional File 2: Fig. S10), which could differ.

| Parameter | Meaning | Lower limit | Upper limit |
|---|---|---|---|
| $r_{add,off}^{A,B}$ | Contribution to inactivation of B by A to A-directed B inactive state | 0.1 | 10 |
| A/A′ add-off ratio | $\frac{r_{add,off}^{A,B}}{r_{add,off}^{A',B}}$; the factor by which A is more effective at inactivating B alleles to A-directed inactive state than A′ to A′-directed inactive state | 0.1 | 10 |
| A/A′ B-activation ratio | $d_{inh}$; factor by which the activation rate of B is lower in the A′-directed B inactive state than in A-directed B inactive state | 0.1 | 10 |

The parameter sets (spanning 8 or 9 parameters or parameter ratios, depending on the model and assumptions) for each set of simulations are described in Additional File 1: Table S4.

For each primary set of simulations, we used Latin hypercube sampling to homogeneously sample over the multidimensional parameter spaces. In each case, we drew 10,000 parameter sets from within the upper and lower boundaries of each log10-transformed parameter range (https://www.mathworks.com/matlabcentral/fileexchange/45793-latin-hypercube). Log transformation was used to more evenly sample over orders of magnitude.

### Simulations

We simulated each of the 3 network models under at least 2 different assumed conditions related to differential effects or paralog expression (see Figures), for approximately 10,000 parameter sets, resulting in a total of approximately 60,000 simulations across 3 network models, each containing consecutive time periods simulating 3 genotypes: wildtype (neither A allele mutated), heterozygous (one A allele mutated), and homozygous-mutant (two A alleles mutated). We also conducted parameter subspace resampling for smaller numbers of parameter sets. We used Gillespie's next reaction method, as previously described [39, 66, 69]. We computed for a total of 300,000 timesteps per simulation, i.e., 100,000 in each genotype. In each simulation, at time $t=0$ all alleles were in the inactive state and the mRNA count was fixed at 1 for $A$ and 0 for all other products. We implemented the simulations in MATLAB R2017a, R2021b, and R2024a [66]. Each simulation took between 30 s and 12 min to run, depending on the parameter values, leading to a total simulation time of approximately two weeks using 8 cores running in parallel.

#### *Pseudo-single-cell analysis and autocorrelation*

In order to simulate snapshot single-cell population measurements of gene expression from these simulations, we split the simulation traces into 300-timestep-unit segments and used the first set of values (of DNA activation states and product abundances) as a sampled "pseudo-single-cell" measurement, as discussed in prior work [66]. We needed to confirm that our sub-simulation samples were not susceptible to unexpected autocorrelation that might interfere with using these samples as independent pseudo-random samples of single cells drawn from the underlying distributions emergent from the network simulations. Therefore, we used the stats::acf function in R to show that 100-step lags were within the 95% confidence intervals for random autocorrelation for a random sample of parameter sets, and set the lag at an even higher value, 300, out of an abundance of caution given the range of possible parameter combinations.

### Steady state analysis

We also wanted to confirm that the output of our numerically simulated stochastic network models fitted with other ways of estimating the outputs of these same networks, both as a quality control and to highlight the added benefits related to simulating variability with stochastic simulations in studies of complex networks. Therefore, we used the ode45 solver in MATLAB R2017a to deterministically estimate the steady state

outputs of all gene product levels in each genotype for 100 parameter sets. The systems of differential equations for each model included 18, 25, and 18 equations each, for the single-paralog stimulation, two-paralog stimulation, and single-paralog repression models, respectively. Initial conditions included all alleles set to the off state and all mRNA levels set to 0. A timespan of 500 units was used, and a random sample of results were inspected to ensure that mRNA level estimates had reached steady state. We then compared these estimated steady state outputs to the pseudo-single-cell population means from our simulations and observed very high concordance at the absolute abundance level (Additional File 2: Fig. S14). The simulations in which a gene product did not have a pseudo-single-cell mean similar to the ode45 steady state solution were most often for parameter sets with high Hill coefficients ($n$), reflecting high non-linearity in regulatory interactions.

#### Distribution shape statistics

We sought to describe the variability in gene expression emerging from gene regulatory networks with transcriptional adaptation and to quantify differences in aspects of variability between network outputs given different parameter values. Therefore, we calculated several summary statistics related to distribution shape to highlight important features of gene expression distributions.

For each gene product in each genotype, we calculated the first four empirical moments of gene expression distributions which describe different aspects of distribution shape. Briefly, the first moment is mean ($\mu$), which would parallel steady state output for a symmetric unimodal distribution. The second moment is variance, for an overall estimate of spread in the distribution. Instead of variance ($\sigma^2$) specifically, we focus our analyses instead on the coefficient of variation ($CV = \frac{\sigma}{\mu}$), which more directly parallels the percentage of spread relative to the mean. The third moment is skewness ($\gamma_1$), which will be positive for right-skewed distributions and negative for left-skewed distributions. The fourth moment is kurtosis, specifically here the excess kurtosis, ($\gamma_2$), which, among other uses, is positively correlated with the heaviness of a distribution's tails.

We also calculated several additional statistics related to distribution shape. Particularly for follow-up analyses, we computed the bimodality coefficient ($BC = \frac{\gamma_1^2+1}{\gamma_2+3*\frac{(n-1)^2}{(n-2)(n-3)}}$) [124, 125]. The bimodality coefficient has a number of useful features. It is a statistic with value constrained to [0, 1]. Uniform distributions will have $BC = \frac{5}{9}$, while unimodal distributions will have values closer to 0 and bimodal (or multimodal) distributions will have values closer to 1. Two additional statistics associated with distribution shape are Gini coefficient and entropy. Gini coefficient is constrained to [0, 1], where 1 corresponds to a distribution in which one cell has non-zero expression and all others have zero expression, and 0 corresponds to a distribution in which all cells express the same amount. We calculated entropy over the binned expression axis, considering 30 bins spread evenly across the range of expression values. As described in the "Results" section, we used these distribution shape statistics both directly in analyses of model parameter effects as well as indirectly in a shape classifier algorithm, below, which we then also used for analyses of model parameter effects on gene expression.

#### *Normalized distribution shape statistics*

Initial exploration of the distribution statistics revealed that several systematically correlated, often nonlinearly, with the overall sample mean. Therefore, to partially correct for differences in mean explaining differences in other statistics, we performed local regression (LOESS) of each statistic against mean, using default settings in the loess function in R with span = 0.1. The residual of the observed minus LOESS fitted statistic value at an observation's same mean can be considered as a mean-corrected version of the observed statistic. For $BC$, the LOESS residual is called $BC^{res}$. One such LOESS-corrected statistic (bimodality coefficient) was used in the distribution shape classifier below, in conjunction with the uncorrected statistic value.

### Distribution shape classification

We present several analyses centered on the question of when an expression distribution can remain robust to the mutation of an upstream regulator. Therefore, we built an algorithm for classifying distribution shapes to reflect plausibly important differences. We were particularly interested in a robust method for identifying whether a distribution was unimodal and symmetric, suggesting a degree of homogeneity in expression. For distributions that were bimodal (or multimodal), one could imagine different emergent properties in a population of cells, e.g., with bistability or other kinds of functional diversity. For distributions that were unimodal but not symmetric, i.e., skewed, one could imagine a bias toward low-frequency diversity in behavior, either being very high expressors or very low expressors. Lastly, we also needed to identify when expression levels were very low in general, reflecting overall minimal transcriptional activity.

The algorithm sorts distributions into 1 of 5 classes:

1. Low-expression
2. Unimodal symmetric
3. Right-skewed unimodal
4. Left-skewed unimodal
5. Bimodal (or multimodal)

It starts by considering whether $\mu < 10$. If so, the distribution is called low-expression. Next, if $BC > 0.555$ and if $BC^{res} > 0.1$, the distribution is called bimodal (or multimodal). After that, if $\gamma_1 > 1$, then the distribution is called right-skewed unimodal, and if $\gamma_1 < (-1)$, then the distribution is called left-skewed unimodal. All remaining distributions have relatively high expression, low absolute skewness, and low bimodality coefficient, and they are called unimodal symmetric. This algorithm was the result of numerous modified iterations during algorithm development, paired with manual inspections of classifier results on random selections of hundreds of simulation results. Classifier result accuracy (i.e., whether a distribution classification by the algorithm is in agreement with manual assignment) was high (greater than 80%, often greater than 90%) for all distribution shape classes. We focused our analyses on whether or not a distribution was called unimodal symmetric and whether the unimodal symmetric class could be made robust to mutation of an upstream regulator.

### Decision tree analysis

We wanted to check whether there were subspaces of parameter space that are enriched for gene regulatory outputs that display behavior of interest, e.g., robustness of shape to mutation of an upstream regulator, for unimodal symmetric distributions. Therefore, as previously described, we trained decision tree classifiers on simulation results paired with algorithm-assigned distribution shape classes, particularly for gene product *B*. According to the binary classifications described in the respective sections of Results, we performed decision tree optimization in R using partykit v1.2–16 and its associated dependencies, with alpha = 0.01, minbucket = 100 for variable selection. For the five-way shape classifier decision tree in Additional File 2: Fig. S12, we also added a max_depth = 3 constraint to aid in visualization. As discussed in the respective sections of the "Results" section, we validated decision tree results by resampling parameter sets from the parameter subspaces bounded by the decision rules enriching for the specific behavior of interest. In each case, we used Latin hypercube sampling, as described above, to sample 100 parameter sets from each subspace and conducted simulations, also as described above.

### Decision tree subspace-based resampling analysis

We extracted the subspace bounds of decision tree terminal nodes (i.e., leaves in ctree objects) using partykit::.list.rules.party() in R, for all 5 decision trees constructed for identifying enrichment of different distribution shapes of gene B in the heterozygous genotype (see Supplementary Note) using the algorithm in the "Distribution shape classification" section. When a parameter was only bounded by one decision rule in a subspace, we used the full parameter subspace boundary as the corresponding upper or lower limit of the subspace. For each parameter, we measured the empirical distribution of sub-full-parameter-range subspace range sizes on a log10 scale.

For trimodal shape resampling analysis (Fig. S6), we first specified the original parameter set in which we observed the trimodal distribution shape for gene B in the heterozygous genotype. We then picked subspaces in which each parameter range was centered on the original parameter set's respective value, bounded by a range of size (custom-smaller than minimum, minimum of the empirical distribution, 10th percentile of the empirical distribution, or 50th percentile of the empirical distribution). The sampled parameter ranges are listed in Additional File 1: Table S4.

### Statistical analysis

Unless noted otherwise in figure legends, error bars represent standard error of the mean. RNA-seq data analysis, including bootstrap resampling, was performed in Python v3.9.15 using gprofiler-official v1.0.0, matplotlib v3.6.2, numpy v1.23.5, pandas v1.5.2, pydeseq2 v0.3.5, requests v2.28.1, scipy v1.9.3, and seaborn v0.12.1. Latin hypercube sampling and simulations were run in MATLAB R2017a, R2021b, and R2024a. All remaining statistical analysis and graph generation was performed in R v3.6.1 and v4.3.2 using packages readxl v1.4.0, partykit v1.2–16, mvtnorm v1.1–3, libcoin v1.0–9,

ggalluvial v0.12.3, entropy v1.3.1, svglite v2.1.0, corrplot v0.92, ggrepel v0.9.1, e1071 v1.7–11, diptest v0.76–0, gridExtra v2.3, Hmisc v4.7–0, Formula v1.2–4, survival v3.3–1, lattice v0.20–45, ineq v0.2–13, magrittr v2.0.3, forcats v0.5.1, stringr v1.4.0, dplyr v1.0.9, purrr v0.3.4, readr v2.1.2, tidyr v1.2.0, tibble v3.1.7, ggplot2 v3.3.6, tidyverse v1.3.1, and gprofiler2 v0.2.3.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03351-2.

Additional file 1: Supplementary Tables. S1: Bulk RNA-seq datasets reanalyzed in this study. GEO entries, experimental designs, and descriptions of knockout targets are provided. S2: Bulk RNA-seq-based knockout targets analyzed for paralog upregulation. S3: Perturb-seq-based knockout target-paralog pairs with signs of potential paralog upregulation. Target-paralog pairs included if the paralog was among either in the top-100 increases in percent-positive cellsor in the top-100 increases in mean expression level. S4: Latin hypercube sampled parameter sets and trimodal analysis sampling ranges for simulations presented in this study.

Additional file 2: Supplementary Figures and Supplementary Note. Supplementary Note: Detailed description of Gillespie simulations and additional analyses of simulation results.

Additional file 3: Review history.

**Acknowledgements**
We thank members of the Goyal lab, including Rohan Sohini, for insightful discussions related to this work. We also thank Lea Schuh (Helmholtz Munich), Karun Kiani (University of Pennsylvania), and Granton Jindal (UCSD) for discussions and manuscript comments. We thank Aviv Regev (Genentech) for pointing us to relevant datasets, and Timothee Lionnet (New York University) for pointing us to relevant literature. We thank David Ho (Columbia University) for support and advice. We thank all authors who published accessible transcriptomics datasets.

**Peer review information**
Kevin Pang was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

**Review history**
The review history is available as Additional file 3.

**Authors' contributions**
YG and IAM conceived and designed the project. IAM designed, performed, and analyzed the simulations in close consultation with YG. IAM conceptualized the analysis of publicly available datasets, and NB identified datasets and performed analysis under the guidance of IAM and YG. MEM performed a subset of analysis, original and revision draft review, and figure design with input from IAM and YG. IAM and YG made the figures. YG and IAM wrote the paper with input largely from MEM and some from NB.

**Funding**
YG acknowledges support from Northwestern University's startup funds and the Burroughs Wellcome Fund Career Awards at the Scientific Interface. NB acknowledges support from NIH T32 GM144295 and T32 GM142604 and funding to YG. MEM and YG acknowledge support from the National Institute for Theory and Mathematics in Biology through the National Science Foundation (DMS-2235451) and the Simons Foundation (MPTMPS-00005320). YG is a CZ Biohub Investigator. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

**Availability of data and materials**
GEO accession numbers for publicly accessible, previously published RNA-seq datasets analyzed here are noted in Additional File 1: Table S1 [126–161]. Processed Perturb-seq data tables are available at https://singlecell.broadinstitute.org/, accession SCP1064, "Control" condition and raw data are available at the Broad Data Use and Oversight System (https://duos.broadinstitute.org), accession DUOS-000124 [162]. Simulation outputs analyzed in this paper are available for download on Dryad (https://doi.org/10.5061/dryad.nk98sf82j) [163]. All code required to reproduce the figures in this paper is available on GitHub and Zenodo (https://doi.org/10.5281/zenodo.12775159) [164].

## Declarations

**Ethics approval and consent to participate**
Ethical approval was not required for this study.

**Competing interests**
The authors declare no competing interests.

Received: 13 December 2023 Accepted: 25 July 2024
Published online: 12 August 2024

**References**


1. Mellis IA, Edelstein HI, Truitt R, Goyal Y, Beck LE, Symmons O, et al. Responsiveness to perturbations is a hallmark of transcription factors that maintain cell identity in vitro. Cell Syst. 2021;12:885-99.e8.
2. El-Brolosy MA, Stainier DYR. Genetic compensation: a phenomenon in search of mechanisms. PLoS Genet. 2017;13:e1006780.
3. Shin J, MacCarthy T. Antagonistic coevolution drives whack-a-mole sensitivity in gene regulatory networks. PLoS Comput Biol. 2015;11:e1004432.
4. Habib N, Wapinski I, Margalit H, Regev A, Friedman N. A functional selection model explains evolutionary robustness despite plasticity in regulatory networks. Mol Syst Biol. 2012;8:619.
5. Karin O, Swisa A, Glaser B, Dor Y, Alon U. Dynamical compensation in physiological circuits. Mol Syst Biol. 2016;12:886.
6. Stelling J, Sauer U, Szallasi Z, Doyle FJ 3rd, Doyle J. Robustness of cellular functions. Cell. 2004;118:675–85.
7. Ma W, Trusina A, El-Samad H, Lim WA, Tang C. Defining network topologies that can achieve biochemical adaptation. Cell. 2009;138:760–73.
8. Goyal Y, Busch GT, Pillai M, Li J, Boe RH, Grody EI, et al. Diverse clonal fates emerge upon drug treatment of homogeneous cancer cells. Nature. 2023. https://doi.org/10.1038/s41586-023-06342-8
9. Braun E. The unforeseen challenge: from genotype-to-phenotype in cell populations. Rep Prog Phys. 2015;78:036602.
10. Hebbar A, Moger A, Hari K, Jolly MK. Robustness in phenotypic plasticity and heterogeneity patterns enabled by EMT networks. Biophys J. 2022;0. [cited 2022 Jul 20]. Available from: http://www.cell.com/article/S0006349522005902/abstract
11. McFaline-Figueroa JL, Srivatsan S, Hill AJ, Gasperini M, Jackson DL, Saunders L, et al. Multiplex single-cell chemical genomics reveals the kinase dependence of the response to targeted therapy [Internet]. bioRxiv. 2023. p. 2023.03.10.531983. [cited 2023 Mar 13]. Available from: https://www.biorxiv.org/content/10.1101/2023.03.10.531983v1
12. Filteau M, Hamel V, Pouliot M-C, Gagnon-Arsenault I, Dubé AK, Landry CR. Evolutionary rescue by compensatory mutations is constrained by genomic and environmental backgrounds. Mol Syst Biol. 2015;11:832.
13. Maamar H, Raj A, Dubnau D. Noise in gene expression determines cell fate in Bacillus subtilis. Science. 2007;317:526–9.
14. El-Brolosy MA, et al. Genetic compensation triggered by mutant mRNA degradation. Nature. 2019;568:193–7. https://doi.org/10.1038/s41586-019-1064-z.
15. Ma Z, Zhu P, Shi H, Guo L, Zhang Q, Chen Y, et al. PTC-bearing mRNA elicits a genetic compensation response via Upf3a and COMPASS components. Nature. 2019;568:259–63.
16. Aoki SK, Lillacci G, Gupta A, Baumschlager A, Schweingruber D, Khammash M. A universal biomolecular integral feedback controller for robust perfect adaptation. Nature. 2019. https://doi.org/10.1038/s41586-019-1321-1
17. Wilkinson MF. Genetic paradox explained by nonsense. Nature Publishing Group UK. 2019 [cited 2023 Jul 12]. https://doi.org/10.1038/d41586-019-00823-5
18. Diofano F, Weinmann K, Schneider I, Thiessen KD, Rottbauer W, Just S. Genetic compensation prevents myopathy and heart failure in an in vivo model of Bag3 deficiency. PLoS Genet. 2020;16:e1009088.
19. Serobyan V, Kontarakis Z, El-Brolosy MA, Welker JM, Tolstenkov O, Saadeldein AM, et al. Transcriptional adaptation in Caenorhabditis elegans. Elife. 2020;9. https://doi.org/10.7554/eLife.50014
20. Gao Y, Zhang Y, Zhang D, Dai X, Estelle M, Zhao Y. Auxin binding protein 1 (ABP1) is not required for either auxin signaling or *Arabidopsis* development. Proc Natl Acad Sci U S A. 2015;112:2275–80.
21. De Souza AT, Dai X, Spencer AG, Reppen T, Menzie A, Roesch PL, et al. Transcriptional and phenotypic comparisons of Ppara knockout and siRNA knockdown mice. Nucleic Acids Res. 2006;34:4486–94.
22. Yamamoto S, Jaiswal M, Charng W-L, Gambin T, Karaca E, Mirzaa G, et al. A drosophila genetic resource of mutants to study mechanisms underlying human genetic diseases. Cell. 2014;159:200–14.
23. Kok FO, Shin M, Ni C-W, Gupta A, Grosse AS, van Impel A, et al. Reverse genetic screening reveals poor correlation between morpholino-induced and mutant phenotypes in zebrafish. Dev Cell. 2015;32:97–108.
24. Ito T, Young MJ, Li R, Jain S, Wernitznig A, Krill-Burger JM, et al. Paralog knockout profiling identifies DUSP4 and DUSP6 as a digenic dependence in MAPK pathway-driven cancers. Nat Genet. 2021;53:1664–72.
25. Zhu P, Ma Z, Guo L, Zhang W, Zhang Q, Zhao T, et al. Short body length phenotype is compensated by the upregulation of nidogen family members in a deleterious nid1a mutation of zebrafish. J Genet Genomics. 2017;44:553–6.
26. Rossi A, Kontarakis Z, Gerri C, Nolte H, Hölper S, Krüger M, et al. Genetic compensation induced by deleterious mutations but not gene knockdowns. Nature. 2015;524:230–3.
27. Konjikusic MJ, Gray RS, Wallingford JB. The developmental biology of kinesins. Dev Biol. 2021;469:26–36.
28. Buglo E, Sarmiento E, Martuscelli NB, Sant DW, Danzi MC, Abrams AJ, et al. Genetic compensation in a stable slc25a46 mutant zebrafish: A case for using F0 CRISPR mutagenesis to study phenotypes caused by inherited disease. PLoS ONE. 2020;15:e0230566.
29. Welker JM, Serobyan V, Zaker Esfahani E, Stainier DYR. Partial sequence identity in a 25-nucleotide long element is sufficient for transcriptional adaptation in the Caenorhabditis elegans act-5/act-3 model. PLoS Genet. 2023;19:e1010806.
30. Fernandez-Abascal J, Wang L, Graziano B, Johnson CK, Bianchi L. Exon-dependent transcriptional adaptation by exon-junction complex proteins Y14/RNP-4 and MAGOH/MAG-1 in Caenorhabditis elegans. PLoS Genet. 2022;18:e1010488.
31. Xie A, Ma Z, Wang J, Zhang Y, Chen Y, Yang C, et al. Upf3a but not Upf1 mediates the genetic compensation response induced by leg1 deleterious mutations in an H3K4me3-independent manner. Cell Discov. 2023;9:63.
32. Kovács K, Farkas Z, Bajić D, Kalapis D, Daraba A, Almási K, et al. Suboptimal global transcriptional response increases the harmful effects of loss-of-function mutations. Mol Biol Evol. 2021;38:1137–50.

33. García-Martínez J, Medina DA, Bellvís P, Sun M, Cramer P, Chávez S, et al. The total mRNA concentration buffering system in yeast is global rather than gene-specific [Internet]. bioRxiv. 2021 [cited 2023 Jul 13]. p. 2021.01.14.426689. Available from: https://www.biorxiv.org/content/10.1101/2021.01.14.426689v2
34. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. Science. 2002;298:824–7.
35. Sorrells TR, Booth LN, Tuch BB, Johnson AD. Intersecting transcription networks constrain gene regulatory evolution. Nature. 2015;523:361–5.
36 Fertuzinhos S, Legué E, Li D, Liem KF Jr. A dominant tubulin mutation causes cerebellar neurodegeneration in a genetic model of tubulinopathy. Sci Adv. 2022;8:eabf7262.
37. Vincentz JW, Firulli BA, Toolan KP, Osterwalder M, Pennacchio LA, Firulli AB. HAND transcription factors cooperatively specify the aorta and pulmonary trunk. Dev Biol. 2021; Available from: https://www.sciencedirect.com/science/article/pii/S0012160621000737
38. Dickinson ME, Flenniken AM, Ji X, Teboul L, Wong MD, White JK, et al. High-throughput discovery of novel developmental phenotypes. Nature. 2016;537:508–14.
39. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S. Stochastic mRNA synthesis in mammalian cells. PLoS Biol. 2006;4:e309.
40. Mellis IA, Raj A. Half dozen of one, six billion of the other: what can small- and large-scale molecular systems biology learn from one another? Genome Res. 2015;25:1466–72.
41. Pillai M, Hojel E, Jolly MK, Goyal Y. Unraveling non-genetic heterogeneity in cancer with dynamical models and computational tools. Nature Computational Science. 2023;3:301–13.
42. Gasiunas G, Barrangou R, Horvath P, Siksnys V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc Natl Acad Sci U S A. 2012;109:E2579–86.
43. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012;337:816–21.
44. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, et al. Multiplex genome engineering using CRISPR/Cas systems. Science. 2013;339:819–23.
45. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, et al. RNA-guided human genome engineering via Cas9. Science. 2013;339:823–6.
46. You KT, Li LS, Kim N-G, Kang HJ, Koh KH, Chwae Y-J, et al. Selective translational repression of truncated proteins from frameshift mutation-derived mRNAs in tumors. PLoS Biol. 2007;5: e109.
47. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al. Ensembl 2022. Nucleic Acids Res. 2022;50:D988–95.
48. Lackner A, Sehlke R, Garmhausen M, Giuseppe Stirparo G, Huth M, Titz-Teixeira F, et al. Cooperative genetic networks drive embryonic stem cell transition from naïve to formative pluripotency. EMBO J. 2021;40:e105776.
49. Adamson B, Norman TM, Jost M, Cho MY, Nuñez JK, Chen Y, et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. Cell. 2016;167:1867-82.e21.
50. Datlinger P, Rendeiro AF, Schmidl C, Krausgruber T, Traxler P, Klughammer J, et al. Pooled CRISPR screening with single-cell transcriptome readout. Nat Methods. 2017;14:297–301.
51. Jaitin DA, Weiner A, Yofe I, Lara-Astiaso D, Keren-Shaul H, David E, et al. Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-Seq. Cell. 2016;167:1883-96.e15.
52. Frangieh CJ, Melms JC, Thakore PI, Geiger-Schuller KR, Ho P, Luoma AM, et al. Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion. Nat Genet. 2021;53:332–41.
53. Holland CH, Tanevski J, Perales-Patón J, Gleixner J, Kumar MP, Mereu E, et al. Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. Genome Biol. 2020;21:36.
54. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014;11:740–2.
55. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16:284–7.
56. Shilatifard A. The COMPASS family of histone H3K4 methylases: mechanisms of regulation in development and disease pathogenesis. Annu Rev Biochem. 2012;81:65–95.
57. Greulich F, Wierer M, Mechtidou A, Gonzalez-Garcia O, Uhlenhaut NH. The glucocorticoid receptor recruits the COMPASS complex to regulate inflammatory transcription at macrophage enhancers. Cell Rep. 2021;34:108742.
58. Torre EA, Arai E, Bayatpour S, Jiang CL, Beck LE, Emert BL, et al. Genetic screening for single-cell variability modulators driving therapy resistance. Nat Genet. 2021;53:76–85.
59. Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database. 2017;2017. https://doi.org/10.1093/database/bax028
60. Liu T, Porter J, Zhao C, Zhu H, Wang N, Sun Z, et al. TADKB: family classification and a knowledge base of topologically associating domains. BMC Genomics. 2019;20:217.
61 Jiang Z, El-Brolosy MA, Serobyan V, Welker JM, Retzer N, Dooley CM, et al. Parental mutations influence wild-type offspring via transcriptional adaptation. Sci Adv. 2022;8:eabj2029.
62. Ibn-Salem J, Muro EM, Andrade-Navarro MA. Co-regulation of paralog genes in the three-dimensional chromatin architecture. Nucleic Acids Res. 2017;45:81–91.
63. Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J. Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res. 2019;29:1363–75.
64. Mueller-Dott S, Tsirvouli E, Vazquez M, Flores ROR, Badia-i-Mompel P, Fallegger R, et al. Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities [Internet]. bioRxiv. 2023 [cited 2023 Apr 3]. p. 2023.03.30.534849. Available from: https://www.biorxiv.org/content/10.1101/2023.03.30.534849v1
65. Raj A, Rifkin SA, Andersen E, van Oudenaarden A. Variability in gene expression underlies incomplete penetrance. Nature. 2010;463:913–8.

66. Schuh L, Saint-Antoine M, Sanford EM, Emert BL, Singh A, Marr C, et al. Gene networks with transcriptional bursting recapitulate rare transient coordinated high expression states in cancer. Cell Syst. 2020;10:363-78.e12.
67. Phillips R, Belliveau NM, Chure G, Garcia HG, Razo-Mejia M, Scholes C. Figure 1 theory meets figure 2 experiments in the study of gene expression. Annu Rev Biophys. 2019;48:121–63.
68. Czuppon P, Pfaffelhuber P. Limits of noise for autoregulated gene expression. J Math Biol. 2018;77:1153–91.
69. Gillespie DT. Exact stochastic simulation of coupled chemical reactions. J Phys Chem. 1977;81:2340–61.
70. Larsson AJM, Johnsson P, Hagemann-Jensen M, Hartmanis L, Faridani OR, Reinius B, et al. Genomic encoding of transcriptional burst kinetics. Nature. 2019;565:251–4.
71. Suter DM, Molina N, Gatfield D, Schneider K, Schibler U, Naef F. Mammalian genes are transcribed with widely different bursting kinetics. Science. 2011;332:472–4.
72. Gupta A, Martin-Rufino JD, Jones TR, Subramanian V, Qiu X, Grody EI, et al. Inferring gene regulation from stochastic transcriptional variation across single cells at steady state. Proc Natl Acad Sci U S A. 2022;119: e2207392119.
73. DelRosso N, Tycko J, Suzuki P, Andrews C, Aradhana, Mukund A, et al. Large-scale mapping and mutagenesis of human transcriptional effector domains. Nature. 2023 https://doi.org/10.1038/s41586-023-05906-y
74. Replogle JM, Norman TM, Xu A, Hussmann JA, Chen J, Cogan JZ, et al. Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. Nat Biotechnol. 2020;38:954–61.
75. Replogle JM, Saunders RA, Pogson AN, Hussmann JA, Lenail A, Guna A, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. Cell. 2022;185:2559-75.e28.
76. Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S. Imaging individual mRNA molecules using multiple singly labeled probes. Nat Methods. 2008;5:877–9.
77 Grody EI, Abraham A, Shukla V, Goyal Y. Toward a systems-level probing of tumor clonality. iScience. 2023;26:106574.
78. Zhang Z, Melzer ME, Arun KM, Sun H, Eriksson C-J, Fabian I, et al. Synthetic DNA barcodes identify singlets in scRNA-seq datasets and evaluate doublet algorithms. Cell Genom. 2024;100592.
79. de Turris V, Nicholson P, Orozco RZ, Singer RH, Mühlemann O. Cotranscriptional effect of a premature termination codon revealed by live-cell imaging. RNA. 2011;17:2094–107.
80. Seidman JG, Seidman C. Transcription factor haploinsufficiency: when half a loaf is not enough. J Clin Invest. 2002;109:451–5.
81. Fishman L, Modak A, Nechooshtan G, Razin T, Erhard F, Regev A, et al. Cell-type-specific mRNA transcription and degradation kinetics in zebrafish embryogenesis from metabolically labeled single-cell RNA-seq. Nat Commun. 2024;15:3104.
82. Moffitt JR, Hao J, Wang G, Chen KH, Babcock HP, Zhuang X. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. Proc Natl Acad Sci U S A [Internet]. 2016. https://doi.org/10.1073/pnas.1612826113
83. Eng C-HL, Lawson M, Zhu Q, Dries R, Koulena N, Takei Y, et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. Nature. 2019;1.
84. Feldman D, Singh A, Schmid-Burgk JL, Carlson RJ, Mezger A, Garrity AJ, et al. Optical pooled screens in human cells. Cell. 2019;179:787-99.e17.
85. Rodriguez J, Ren G, Day CR, Zhao K, Chow CC, Larson DR. Intrinsic dynamics of a human gene reveal the basis of expression heterogeneity. Cell. 2019;176:213-26.e18.
86. Boe RH, Ayyappan V, Schuh L, Raj A. Allelic correlation is a marker of trade-offs between barriers to transmission of expression variability and signal responsiveness in genetic networks. Cell Syst. 2022;13:1016-32.e6.
87. Mahi NA, Najafabadi MF, Pilarczyk M, Kouril M, Medvedovic M. GREIN: an interactive web platform for re-analyzing GEO RNA-seq data. Sci Rep. 2019;9:7580.
88. Grevet JD, Lan X, Hamagami N, Edwards CR, Sankaranarayanan L, Ji X, et al. Domain-focused CRISPR screen identifies HRI as a fetal hemoglobin regulator in human erythroid cells. Science. 2018;361:285–90.
89. Wang X, Wang S, Troisi EC, Howard TP, Haswell JR, Wolf BK, et al. BRD9 defines a SWI/SNF sub-complex and constitutes a specific vulnerability in malignant rhabdoid tumors. Nat Commun. 2019;10:1881.
90. Huang B, Chen Z, Geng L, Wang J, Liang H, Cao Y, et al. Mucosal profiling of pediatric-onset colitis and IBD reveals common pathogenics and therapeutic pathways. Cell. 2019;179:1160-76.e24.
91. Yu X, Azzo A, Bilinovich SM, Li X, Dozmorov M, Kurita R, et al. Disruption of the MBD2-NuRD complex but not MBD3-NuRD induces high level HbF expression in human adult erythroid cells. Haematologica. 2019;104:2361–71.
92. Beyret E, Liao H-K, Yamamoto M, Hernandez-Benitez R, Fu Y, Erikson G, et al. Single-dose CRISPR-Cas9 therapy extends lifespan of mice with Hutchinson-Gilford progeria syndrome. Nat Med. 2019;25:419–22.
93. Xu W, Liu C, Deng B, Lin P, Sun Z, Liu A, et al. TP53-inducible putative long noncoding RNAs encode functional polypeptides that suppress cell proliferation. Genome Res. 2022;32:1026–41.
94. Guièze R, Liu VM, Rosebrock D, Jourdain AA, Hernández-Sánchez M, Martinez Zurita A, et al. Mitochondrial Reprogramming Underlies Resistance to BCL-2 Inhibition in Lymphoid Malignancies. Cancer Cell. 2019;36:369-84.e13.
95. Schade AE, Fischer M, DeCaprio JA. RB, p130 and p107 differentially repress G1/S and G2/M genes after p53 activation. Nucleic Acids Res. 2019;47:11197–208.
96. He Z, Thorrez L, Siegfried G, Meulemans S, Evrard S, Tejpar S, et al. The proprotein convertase furin is a pro-oncogenic driver in KRAS and BRAF driven colorectal cancer. Oncogene. 2020;39:3571–87.
97. Castellani CA, Longchamps RJ, Sumpter JA, Newcomb CE, Lane JA, Grove ML, et al. Mitochondrial DNA copy number can influence mortality and cardiovascular disease via methylation of nuclear DNA CpGs. Genome Med. 2020;12:84.
98. Perenthaler E, Nikoncuk A, Yousefi S, Berdowski WM, Alsagob M, Capo I, et al. Loss of UGP2 in brain leads to a severe epileptic encephalopathy, emphasizing that bi-allelic isoform-specific start-loss mutations of essential genes can cause genetic diseases. Acta Neuropathol. 2020;139:415–42.

99 Guthridge JM, Lu R, Tran LT-H, Arriens C, Aberle T, Kamp S, et al. Adults with systemic lupus exhibit distinct molecular phenotypes in a cross-sectional study. EClinicalMed. 2020;20:100291.
100. Li B, Clohisey SM, Chia BS, Wang B, Cui A, Eisenhaure T, et al. Genome-wide CRISPR screen identifies host dependency factors for influenza A virus infection. Nat Commun. 2020;11:164.
101. Sorial AK, Hofer IMJ, Tselepi M, Cheung K, Parker E, Deehan DJ, et al. Multi-tissue epigenetic analysis of the osteoarthritis susceptibility locus mapping to the plectin gene PLEC. Osteoarthr Cartil. 2020;28:1448–58.
102. Temprine K, Campbell NR, Huang R, Langdon EM, Simon-Vermot T, Mehta K, et al. Regulation of the error-prone DNA polymerase Polκ by oncogenic signaling and its contribution to drug resistance. Sci Signal. 2020;13. https://doi.org/10.1126/scisignal.aau1453
103. Xu B, Qin T, Yu J, Giordano TJ, Sartor MA, Koenig RJ. Novel role of ASH1L histone methyltransferase in anaplastic thyroid carcinoma. J Biol Chem. 2020;295:8834–45.
104. Oksa L, Mäkinen A, Nikkilä A, Hyvärinen N, Laukkanen S, Rokka A, et al. Arginine methyltransferase PRMT7 deregulates expression of RUNX1 target genes in T-Cell acute lymphoblastic leukemia. Cancers. 2022;14. https://doi.org/10.3390/cancers14092169
105. Chin CV, Antony J, Ketharnathan S, Labudina A, Gimenez G, Parsons KM, et al. Cohesin mutations are synthetic lethal with stimulation of WNT signaling. Elife. 2020;9. https://doi.org/10.7554/eLife.61405
106. Ma W, Wang Y, Zhang R, Yang F, Zhang D, Huang M, et al. Targeting PAK4 to reprogram the vascular microenvironment and improve CAR-T immunotherapy for glioblastoma. Nat Cancer. 2021;2:83–97.
107. Wan C, Mahara S, Sun C, Doan A, Chua HK, Xu D, et al. Genome-scale CRISPR-Cas9 screen of Wnt/β-catenin signaling identifies therapeutic targets for colorectal cancer. Sci Adv. 2021;7. https://doi.org/10.1126/sciadv.abf2567
108. Krassovsky K, Ghosh RP, Meyer BJ. Genome-wide profiling reveals functional interplay of DNA sequence composition, transcriptional activity, and nucleosome positioning in driving DNA supercoiling and helix destabilization in C. elegans. Genome Res. 2021;31:1187–202.
109. Uğurlu-Çimen D, Odluyurt D, Sevinç K, Özkan-Küçük NE, Özçimen B, Demirtaş D, et al. AF10 (MLLT10) prevents somatic cell reprogramming through regulation of DOT1L-mediated H3K79 methylation. Epigenet Chromatin. 2021;14:32.
110. Jost M, Jacobson AN, Hussmann JA, Cirolia G, Fischbach MA, Weissman JS. CRISPR-based functional genomics in human dendritic cells. Elife. 2021;10. https://doi.org/10.7554/eLife.65856
111. Haring NL, van Bree EJ, Jordaan WS, Roels JRE, Sotomayor GC, Hey TM, et al. ZNF91 deletion in human embryonic stem cells leads to ectopic activation of SVA retrotransposons and up-regulation of KRAB zinc finger gene clusters. Genome Res. 2021;31:551–63.
112. Abraham HG, Ulintz PJ, Goo L, Yates JA, Little AC, Bao L, et al. RhoC modulates cell junctions and type I interferon response in aggressive breast cancers. Front Oncol. 2021;11:712041.
113. Dubey R, Lebensohn AM, Bahrami-Nejad Z, Marceau C, Champion M, Gevaert O, et al. Chromatin-remodeling complex SWI/SNF controls multidrug resistance by transcriptionally regulating the drug efflux pump ABCB1. Cancer Res. 2016;76:5810–21.
114. Avior Y, Lezmi E, Yanuka D, Benvenisty N. Modeling developmental and tumorigenic aspects of trilateral retinoblastoma via human embryonic stem cells. Stem Cell Reports. 2017;8:1354–65.
115. Tchasovnikarova IA, Timms RT, Douse CH, Roberts RC, Dougan G, Kingston RE, et al. Hyperactivation of HUSH complex function by Charcot-Marie-Tooth disease mutation in MORC2. Nat Genet. 2017;49:1035–44.
116. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2019;47:W191–8.
117. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005;21:3439–40.
118. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
119. Muzellec B, Teleńczuk M, Cabeli V, Andreux M. PyDESeq2: a python package for bulk RNA-seq differential expression analysis [Internet]. bioRxiv. 2022 [cited 2023 May 5]. p. 2022.12.14.520412. Available from: https://doi.org/10.1101/2022.12.14.520412v1
120 Zhang H-M, Liu T, Liu C-J, Song S, Zhang X, Liu W, et al. AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. Nucleic Acids Res. 2015;43:D76-81.
121. Feltz CJ, Miller GE. An asymptotic test for the equality of coefficients of variation from k populations. Stat Med. 1996;15:646–58.
122. Marwick B, Krishnamoorthy K. cvequality: tests for the equality of coefficients of variation from multiple groups. R software package version 01 [Internet]. 2019; Available from: https://scholar.google.ca/scholar?cluster=16493263674800530914&hl=en&as_sdt=0,5&sciodt=0,5
123. Symmons O, Chang M, Mellis IA, Kalish JM, Park J, Suszták K, et al. Allele-specific RNA imaging shows that allelic imbalances can arise in tissues through transcriptional bursting. PLoS Genet. 2019;15:e1007874.
124. Hartigan JA, Hartigan PM. The dip test of unimodality. aos. 1985;13:70–84.
125. Pfister R, Schwarz KA, Janczyk M, Dale R, Freeman JB. Good things peak in pairs: a note on the bimodality coefficient. Front Psychol. 2013;4:700.
126. El-Brolosy MA, Rossi A, Kontarakis Z, Kuenne C, Günther S, Fukuda N, Takacs C, Lai S, Fukuda R, Gerri C, Kikhi K, Giraldez AJ, Stainier DY. Genetic compensation is triggered by mutant mRNA degradation. Datasets. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE114212 (2019).
127. Grevet JD, Lan X, Hamagami N, Edwards CR, Sankaranarayanan L, Ji X, Brardwaj SK, Face CJ, Posocco DF, Abdulmalik O, Keller CA, Giardine BM, Sidoli S, Garcia BA, Chou ST, Liebhaber SA, Hardison RC, Shi J, Blobel GA. Domain-focused CRISPR-screen identifies HRI as a fetal hemoglobin regulator in human erythroid cells. Datasets. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE115687 (2018).

128. Wang X, Wang S, Park P, Roberts CW. BRD9 defines a novel SWI/SNF sub-complex and constitutes a specific vulnerability in malignant rhabdoid tumors [RNA-seq]. Datasets. Gene Expression Omnibus 2019. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE120233.
129. Frausto RF, Swamy VS, Morselli M, Pellegrini M, Aldave AJ. ZEB1 insufficiency causes corneal endothelial cell state transition and altered cellular processing. Datasets. Gene Expression Omnibus 2019. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE121680.
130. Azzo A, Yu X, Dozmorov M, Ginder GD. Disruption of the MBD2-NuRD complex but not MBD3-NuRD induces high level HbF expression in human adult erythroid cells. Datasets. Gene Expression Omnibus 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE121992.
131. Beyret E, Liao H, Fu Y, Yamamoto M, Hernandez-Benitez R, Erikson G, Reddy P, Izpisua Belmonte JC. Single-dose CRISPR/Cas9 therapy extends lifespan of mice with Hutchinson-Gilford progeria syndrome. Datasets. Gene Expression Omnibus 2019. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE122865.
132. Xu W, Deng B, Lin P, Liu C, Li B, Zhou K, Zhou H, Qu L, Yang J. p53-inducible long non-coding RNAs encode functional peptides in hepatocellular carcinoma cells [RNA-seq]. Datasets. Gene Expression Omnibus 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE125756.
133. Wang X, Wang S, Park P, Roberts CW. BRD9 defines a novel SWI/SNF sub-complex and constitutes a specific vulnerability in malignant rhabdoid tumors [RNA-seq 2]. Datasets. Gene Expression Omnibus 2019. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE125775.
134. Guieze R, Wu CJ, Lawlor M, Ott C. Genetic determinants of venetoclax resistance. Datasets. Gene Expression Omnibus 2019. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE128563.
135. Schade AE, Fischer M, DeCaprio JA. RNA-seq of human foreskin fibroblast cells lacking RB and/or p130 after doxorubicin treatment. Datasets. Gene Expression Omnibus 2019. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE128711.
136. He Z. Differential Gene expression of furin knockout (KO) DLD1, HCA7 and HT29 colorectal cancer cells. Datasets. Gene Expression Omnibus 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE130969.
137. Antony J, Giminez G, Horsfield JA. Expression profiling in STAG2 mutant K562 cells. Datasets. Gene Expression Omnibus 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE131448.
138. Castellani CA, Sumpter JA, Newcomb CE, Arking DE. HEK293 TFAM knockout expression study. Datasets. Gene Expression Omnibus 2019. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE134048.
139. Barakat TS. RNA-seq of UGP2 mutant human embryonic stem cells and in vitro differentiated neural stem cells. Datasets. Gene Expression Omnibus 2019. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE137129.
140. Qin Z, Li Q, Xiao Z. Bioinformatic Analysis of mRNA and miRNA Expression Patterns In p53 Knock-out C666-1 cells [mRNA]. Datasets. Gene Expression Omnibus 2021. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE138258.
141. Li B, Cui A, Hacohen N. Genome-wide CRISPR screen Identifies host dependency factors for influenza A virus infection. Datasets. Gene Expression Omnibus 2019. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE141171.
142. Loughlin J, Sorial AK, Cheung K. Multi-tissue epigenetic analysis of the osteoarthritis susceptibility locus mapping to the plectin gene PLEC. Datasets. Gene Expression Omnibus 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE143725.
143. Campbell NR, White RM. Regulation of the error-prone DNA polymerase polκ by oncogenic signaling and its contribution to drug resistance. Datasets. Gene Expression Omnibus 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE145313.
144. Lackner A, Sehlke R, Garmhausen M, Stirparo G, Huth M, Titz-Teixeira F, Ramesmayer J, van der Lelij P, Thomas HF, Ralser M, Santini L, Galimberti E, Sarov M, Stewart A, Smith A, Beyer A, Leeb M. Cooperative genetic networks drive embryonic stem cell transition from naïve to formative pluripotency. Datasets. Gene Expression Omnibus 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE145653.
145. Xu B, Qin T, Yu J, Giordano TJ, Sartor MA, Koenig RJ. RNA-seq BHT-101 cells and BHT-101 ASH1L KO cell lines. Datasets. Gene Expression Omnibus 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE147076.
146. Laukkanen S, Oksa L, Lohi O. SIX6 knockdown in Jurkat-Cas9 T-ALL cell line. Datasets. Gene Expression Omnibus 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE148658.
147. Torre EA, Arai E, Bayatpour S, Jiang C, Beck LE, Emert BL, Shaffer SM, Mellis IA, Budinich KA, Almeida F, Fane M, Weeraratna A, Shi J, Raj A. Genetic screening for single-cell variability modulators driving therapy resistance [WM989-A6-G3-Cas9 5a3]. Datasets. Gene Expression Omnibus 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE151825.
148. Kumar D, Narang V, Singhal A. RNA sequencing of NF-kB knockout (KO) U937 cells. Datasets. Gene Expression Omnibus 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE153158.
149. Chin CV, Antony J, Gimenez G, Horsfield JA. Expression profiling in cohesin mutant MCF10A epithelial and CMK leukaemia cells. Datasets. Gene Expression Omnibus 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE154086.
150. Ma W, Zhang D, Fan Y. Effects of PAK4 knockdown on gene expression in glioblastoma-associated endothelial cells. Datasets. Gene Expression Omnibus 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE154133.
151. Firestein R, Mahara S, Wan C, Sun CX. Genome-scale CRISPR-Cas9 screen of Wnt/β-catenin signalling identifies therapeutic targets for Colorectal Cancer (RNA-seq). Datasets. Gene Expression Omnibus 2021. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE156082.
152. Barbazuk WB, Shailesh L. CRISPR-Cas9 knockdown of RBM48 in K562 cells. Datasets. Gene Expression Omnibus 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE156471.
153. Uğurlu-Çimen D, Sevinç K, Önder T. DOT1L-interacting protein AF10 (MLLT10) is a barrier to somatic cell reprogramming. Datasets. Gene Expression Omnibus 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE161043.

154. Jost M, Jacobson AN, Fischbach MA, Weissman JS. CRISPR genome editing of human dendritic cells (treatments of knockout dendritic cells). Datasets. Gene Expression Omnibus 2021. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE161466.
155. Haring NL, van Bree EJ, Jordaan WS, Roels JR, Congrains Sotomayor G, Hey TM, White FT, Galland MD, Smidt MP, Jacobs FM. Genetic deletion of ZNF91 in human embryonic stem cells leads to ectopic activation of SVAs and collective upregulation of KRAB zinc finger gene clusters. Datasets. Gene Expression Omnibus 2021. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE162571.
156. Abraham H, Merajver S, Ulintz PJ. RhoC modulates cell junctions and type I interferon response in aggressive breast cancers. Datasets. Gene Expression Omnibus 2021. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE175787.
157. Dubey R, Lebensohn A, Bahrami-Nejad Z, Marceau C, Sikic BI, Carette J, Rohatgi R. Gene expression analysis of human haploid cells (HAP1) depleted of SMARCB1 and SMARCA4. Datasets. Gene Expression Omnibus 2016. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75515.
158. Avior Y, Benvenisty N. Modeling trilateral retinoblastoma using human embryonic stem cells. Datasets. Gene Expression Omnibus 2017. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84504.
159. Dai X, Li L, Li J, You C, Gonzalez G, Miao W, Hu J, Fu L, Xu Y, Gu W, Wang Y. The impact of YTHDF2 knockout on the distribution of 5-methylcytosine in RNA. Datasets. Gene Expression Omnibus 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE85887.
160. Timms RT, Tchasovnikarova IA, Lehner PJ. Assessing the impact of loss of ATF7IP and SETDB1 on the transcriptome. Datasets. Gene Expression Omnibus 2016. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE86813.
161. Datlinger P, Rendeiro AF, Schmidl C, Krausgruber T, Traxler P, Klughammer J, Schuster LC, Kuchler A, Alpar D, Bock C. Pooled CRISPR screening with single-cell transcriptome read-out. Datasets. Gene Expression Omnibus 2017. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92872.
162. Frangieh CJ, Melms JC, Thakore PI, Geiger-Schuller KR, Regev A, Izar B. DUOS-000124. Datasets. Broad data use and oversight system 2021. https://duos.broadinstitute.org.
163. Mellis IA, Melzer ME, Bodkin N, Goyal Y. Simulations of gene regulatory networks with transcriptional adaptation. 2024. Dryad. https://doi.org/10.5061/dryad.nk98sf82j.
164. Mellis IA, Melzer ME, Bodkin N, Goyal Y. GoyalLab/TA_prevalence_constraints_public: Publication Version. Zenodo 2024. https://zenodo.org/doi/10.5281/zenodo.12775158.

## Publisher's Note