Leveraging the Mendelian Disorders of the Epigenetic Machinery to Systematically Map Functional Epigenetic Variation

Teresa R. Luperchio[1]*, Leandros Boukas[1,2]*, Li Zhang[1], Genay O. Pilarowski[1**], Jenny Jiang[1], Allison Kalinousky[1], Kasper D. Hansen[1,2#], Hans T. Bjornsson[1,3,4,#].

[1]Department of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA;
[2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA;
[3]Faculty of Medicine, School of Health Sciences, University of Iceland, Reykjavik, Iceland;
[4]Landspitali University Hospital, Reykjavik, Iceland.

*Shared first-author contribution
**Present address: Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA
[#]Corresponding authors: 1) Hans Tomas Bjornsson, email: hbjorns1@jhmi.edu; 2) Kasper Hansen, email: khanse10@jhu.edu

Abstract

The Mendelian Disorders of the Epigenetic Machinery (MDEMs) have emerged as a class of Mendelian disorders caused by loss-of-function variants in epigenetic regulators. Although each MDEM has a different causative gene, they exhibit several overlapping disease manifestations. Here, we hypothesize that this phenotypic convergence is a consequence of common abnormalities at the epigenomic level, which directly or indirectly lead to downstream convergence at the transcriptomic level. This implies that identifying abnormalities shared across multiple MDEMs could pinpoint locations where epigenetic variation is causally related to disease phenotypes. To test our hypothesis, we perform a comprehensive interrogation of chromatin (ATAC-Seq) and expression (RNA-Seq) states in B cells from mouse models of three MDEMs (Kabuki types 1&2 and Rubinstein-Taybi syndromes). We build on recent work in covariate-powered multiple testing to develop a new approach for the overlap analysis, which enables us to find extensive overlap primarily localized in gene promoters. We show that disruption of chromatin accessibility at promoters often leads to disruption of downstream gene expression, and identify a total of 463 loci and 249 genes commonly disrupted across the three MDEMs. As an example of how widespread dysregulation leads to specific phenotypes, we show that subtle expression alterations of multiple, directly relevant genes, collectively contribute to IgA deficiency in KS1 and RT. In contrast, we predict that KS2 does not have IgA deficiency, and confirm this pattern in mice. We propose that the joint study of MDEMs offers a principled approach for systematically mapping functional epigenetic variation in mammals.

Keywords: histone machinery, immune dysfunction, chromatin, histone modification, epigenetics, Mendelian disorders

Introduction:

A long-standing and fundamental problem in epigenetics is the identification of specific epigenetic changes that *causally* mediate phenotypes through the alteration of transcriptional states. While statistical associations between many diseases/traits and epigenetic changes have been detected, it is typically extremely challenging to rule out the influence of confounders such as the environment, and to decide whether these associations are primary causes versus secondary consequences[1,2]. As a result, to date there are surprisingly few examples of causal relationships between epigenetic alterations and specific phenotypes; notable exceptions include

disorders of genomic imprinting[3], disorders caused by repeat-expansion-induced aberrant promoter hypermethylation[4,5], and metastable epialleles[6,7].

The recent advent and widespread clinical use of exome sequencing has led to the emergence of a novel class of Mendelian disorders, termed the Mendelian Disorders of the Epigenetic Machinery (MDEMs)[8]. MDEMs are caused by coding variants disrupting genes encoding for epigenetic regulators, which are generally very intolerant to loss-of-function variation[9]. This implies the following causal chain underlying MDEM pathogenesis: a coding variant disrupts an epigenetic regulator, leading to downstream epigenomic abnormalities, which in turn give rise to the phenotype, likely by perturbing the transcriptome (**Figure 1a**). As a result, MDEMs may provide a unique lens into the causal relationship between epigenetic/transcriptomic variation and disease. Indeed, studies of Kabuki syndrome type 1 - one of the most extensively studied MDEMs to date - have begun unraveling the underpinnings of the neural[10], growth[11], cardiac[12], and immune defects[13–15] seen in this disorder.

With studies of individual MDEMs, however, it remains challenging to distinguish the causal, disease-driving molecular alterations, from noise or passenger effects. As a consequence, it has been difficult to pinpoint specific target genes/loci that contribute to pathogenesis, and it is also unclear if such targets span a small portion of the genome, or whether the disease phenotypes arise through the combined effects of multiple perturbations widely distributed across the genome.

Results:

Joint analysis of multiple MDEMs to identify causally relevant epigenetic and transcriptomic variation

We set out to design a principled approach to overcome limitations in discovering functionally relevant epigenetic variation, such as confounding effects from the environment and reverse causality from the disease process. Our approach leverages a cardinal and thus far unexploited feature of MDEMs, namely their overlapping phenotypic features, despite the causative genetic variants disrupting distinct genes. Such common MDEM features include intellectual disability, growth defects, and immune dysfunction[8]. We hypothesized that these arise because the different primary genetic defects lead to common downstream epigenomic alterations, which themselves create common transcriptomic alterations (**Figure 1b**). The hypothesis of such a convergent pathogenesis motivates a joint analysis of more than one MDEM, and suggests a simple filter to identify the causal variation at the

epigenetic/transcriptomic level: true, disease-driving signals should be detectable in multiple disorders (**Figure 1c**).

With this approach, two central practical considerations arise. First, what biological samples are most appropriate to allow for non-ambiguous interpretation of the results? We propose the use of MDEM mouse models, which have been shown to closely recapitulate many aspects of the human phenotypes[13,16,17]. Importantly, using mice allows us to: a) eliminate multiple confounders such as the environment, the genetic background, age, and sex, b) maintain a consistent sampling of disease-relevant cell types between individuals.

Second, what statistical approach should be employed for detection of the true common alterations? The simplest way would be to perform differential accessibility and expression analyses separately for each disorder, and obtain a list of the overlapping differential hits. However, this suffers from the major shortcoming that a given locus is required to exceed an arbitrary significance threshold (or rank) to be labeled as differential. When multiple MDEMs are studied, this can lead to severe loss of power and erroneous underestimation of the size of the overlap. To avoid this, we recast the problem as testing whether knowledge that a set of loci/genes are differential in a given MDEM is informative about the state (null or differential) of the same loci/genes in another MDEM. We show (**Methods**) that with this formulation, we can use conditional p-value distributions to: a) estimate the size of the set of overlapping abnormalities and test if it is greater than expected by chance, b) identify a set of genes that belong to this overlap, and c) decouple a) from b), so that only the identification of specific genes is affected by the multiple testing burden.

Here, as proof-of-principle, we implement our proposed approach using mouse models of 3 MDEMs: two that were clinically indistinguishable prior to the discovery of the underlying genes (Kabuki syndrome types 1 and 2; KS1 and 2, caused by haploinsufficiency in histone methyltransferases *KMT2D* and *KDM6A*, respectively), and one that shares phenotypes but is clinically distinct (Rubinstein-Taybi type 1; RT1, caused by haploinsufficiency in histone acetyltransferase *CREBBP*). All three syndromes have been previously found to exhibit immune system dysfunction. In KS1, this includes combined variable immunodeficiency with low IgA, as well as abnormal cell maturation[8,18,19]. In RT1, it includes hypogammaglobulinemia with a reduction of mature B cells[20], while in KS2 the immune phenotype has been less extensively studied, with some evidence of mild hypogammaglobulinemia[18,21]. Given this potential overlap, we chose to profile positively selected B cells (CD19+; **Methods**) from the peripheral blood of mutant mice, and that of age- and sex-matched wild-type littermates (**Figure 1d**).

Genome-wide chromatin accessibility profiling reveals extensive overlap between the epigenetic aberrations of the three MDEMs

We first used ATAC-seq to profile genome-wide chromatin accessibility, employing a modified fastATAC protocol (**Methods**)[22,23]. Starting with a differential accessibility analysis of KS1 vs wild-type mice (**Methods**), we discovered 3,121 ATAC peaks differentially accessible at the 10% FDR level. Of these, 824 (26.4%) overlapped promoters (defined as +/- 2kb from the TSS), and 2,297 (74%) were in distal regulatory elements (defined as ATAC peaks outside of promoters).

We then compared KS1 to KS2, focusing on promoters first. Using a new approach to detect overlap between lists of differential features from different experiments (**Methods**), we found that 71.5% of promoter peaks differentially accessible in KS1 are also differential in KS2 (**Figure 2a;** $p < 2.2e-16$); at the 10% FDR level, we identified 645 such peaks (**Supplemental Table 1**). For 630 of the 645 (97.7%), accessibility is altered in the same direction in the two syndromes (**Figure 2b**) Out of the 630 promoter peaks shared between KS1 and KS2, we discovered that 62.2% are differential in RT1 as well (**Figure 2c;** $p = 0.0034$), again with highly concordant effect sizes (**Figure 2d**).

In total, we identified 313 promoter peaks showing disruption across the 3 disorders at the 10% FDR level (**Supplemental Table 2**). This is ~4 times more than what we obtain if we perform separate differential analyses and compute the intersection of the resulting differential hits (78 peaks), highlighting that our new approach provides substantial gain in empirical power. A principal component analysis shows that the accessibility signal of these commonly disrupted promoter peaks separates each of the three mutant genotypes from their wild-type littermates (**Figure 2e, f**). This is most surprising for the case of KS1 and KS2, which were similar enough to not be separated prior to discovery of the causative genes (**Figure 2e**).

Turning our attention to distal regulatory elements, and using the same approach, we saw a similar picture, albeit with weaker shared signal (**Supplemental Figures 1a-d**). Specifically, 55.6% elements differential in KS1 were estimated as differential in KS2 ($p < 2.2e-16$), with 958 confidently labeled such elements (10% FDR; **Supplemental Table 3**). As with promoters, we observed agreement in directionality. Of these 958 elements, 35.4% are shared with RT1 ($p = 0.0053$), yielding a total of 150 commonly disrupted distal regulatory elements across the 3 MDEMs (10% FDR; **Supplemental Table 4**). We note that, collectively, the common hits show a 5.3-fold enrichment at promoters compared to distal elements (Fisher's test, $p < 2.2e-16$).

Finally, comparing the 3 MDEMs in a pairwise fashion, we observed that KS1 and KS2 share a greater proportion of their abnormalities compared to either KS1 with RT1 or KS2 with RT1 (**Supplemental Figure 1e**).

Commonly disrupted promoters, but not distal regulatory elements, show bias towards increased accessibility in KS1 and KS2

We explored the direction in which the accessibility of the disrupted peaks changes. We found that, at promoters, both the KS1 and KS2 mutants exhibit a substantial shift towards increased accessibility (84.6% and 92.7%, respectively, of significantly disrupted promoter peaks; **Figure 2g**). The same shift is observed in the promoter peaks commonly disrupted across the MDEMs (**Figure 2g**), even though in the RT1 mutants the majority (65.4%) of differentially accessible promoter peaks show the opposite pattern, with a shift towards decreased accessibility (**Figure 2g**). In contrast to promoters, we saw that the accessibility of distal regulatory elements shows a largely balanced disruption in all cases (**Supplemental Figure 1f**).

Transcriptome profiling reveals many expression alterations at genes downstream of promoters with disrupted accessibility

We next interrogated the transcriptome using RNA-seq (**Methods**) to: a) test whether the identified epigenetic aberrations in each disorder have direct transcriptional consequences and characterize the latter, and b) identify the common expression aberrations across the three disorders, and assess the extent to which these are direct rather than more downstream effects of the common epigenetic aberrations. As chromatin accessibility and gene expression are dynamic, we generated the RNA-seq samples in parallel with the ones used for ATAC-seq, from a subset of the same individual mice; this allowed us to capture both chromatin and transcriptional status at a single time point.

First, for each disorder, we obtained the top differential promoter peaks as ranked by p-value, and estimated the percent of genes downstream of these promoters that show differential expression; we repeated this by sliding the threshold for inclusion of the top ranked peaks from 1000 to 5000. When considering the top 1000 promoter peaks, the percentage of differentially expressed downstream genes is 36.2% in KS1, 41.3% in KS2, and in 49.5% in RT1 (**Methods; Supplemental Table 5** contains such genes detected at the 10% FDR level). In all three syndromes, this percentage gradually drops substantially as the promoter cutoff becomes less stringent (**Figure 3a),** indicating a clear relationship between abnormal promoter accessibility

and downstream gene expression dysregulation. Emphasizing this relationship, we discovered strong concordance between the direction of abnormal changes at the disrupted promoter-gene pairs: increased or decreased promoter accessibility correlates with increased or decreased gene expression, respectively (**Figure 3b, c, d;** Pearson correlation between promoter accessibility logFC and gene expression logFC = 0.67 for KS1, 0.79 for KS2, and 0.82 for RT1)

Finally, we compared the proportion of differentially expressed genes downstream of the commonly disrupted promoter peaks, to the same proportion for genes downstream of the top 1000 disrupted promoter peaks unique for each disorder. We invariably found the genes downstream of the commonly disrupted peaks to have a higher chance of dysregulated expression (**Figure 3e)**, supporting our hypothesis that the chromatin alterations at these peaks are more likely to be functional.

A substantial proportion of the common expression aberrations among the three MDEMs arise without concomitant disruption of promoter accessibility

To further dissect the relationship between the common expression and chromatin abnormalities in the three MDEMs, we sought to define a set of genes commonly differentially expressed, without utilizing prior information about the accessibility of their promoter peaks.

Utilizing our method, we discovered high overlap between KS1 and KS2, mirroring the findings at the chromatin level (**Figure 4a, b**). Specifically, we found 372 differentially expressed genes shared between them with concordant direction of effect (10% FDR; **Supplemental Table 6**). We then estimated 70% of these to be differential in RT1 (**Figure 4c, d**), resulting in 249 genes shared across the 3 disorders, with a preponderance of downregulated ones (**Figure 4e, f, g; Supplemental Table 7;** 171 downregulated vs 78 upregulated genes).

While these 249 genes show significant enrichment in the set of genes with shared disruption of promoter accessibility (p=3.06e-5), the magnitude of this enrichment was modest (24 genes in the intersection; odds ratio = 2.79, **Table 1**). The number of such genes increases to 81 when we also include those harboring commonly disrupted regulatory elements nearby (+/- 1Mb from their promoter peaks). Taken together, these results indicate that there is convergent dysregulation at the expression level in these three MDEMs, which is downstream of the common differential epigenetic alterations. Further underscoring this, we found that in KS1 and RT1 – but not in KS2 – the top differentially expressed genes are no more likely to have disrupted promoters than genes further down the differential list (**Figure 4h**).

Integration of transcription factor motifs with chromatin and expression abnormalities highlights some potentially disrupted regulatory connections

It is well-appreciated that chromatin accessibility is intimately linked to transcription factor binding. Accessibility patterns are often established subsequently to recruitment of epigenetic regulators by transcription factors at specific genomic locations[24], while other transcription factors can only bind their cognate motifs if these reside within pre-accessible sites[25,26]. We therefore investigated the transcription factor motifs encoded within the differentially accessible peaks in the three disorders, using a set of 233 non-redundant motifs (**Methods**). We focused on differentially accessible peaks within promoters of differentially expressed genes, reasoning that these are more likely to reveal potentially disrupted regulatory connections with functional relevance. In all three syndromes, however, no single motif reached significant enrichment at the 10% FDR level when compared to other peaks.

Regulatory wiring disruption can occur not only because of altered motif accessibility, but also theoretically because of abnormal expression of the cognate transcription factors themselves. We thus also performed a search for motif enrichment in promoter peaks corresponding to differentially expressed genes, regardless of whether these peaks are differentially accessible or not, and then intersected the resulting motifs with our differentially expressed genes. The significant hits (**Supplemental Table 8**) included *Runx1*, as well as motifs recognized by transcription factors of the NF-kB pathway.

The collective effect of individually subtle aberrations in multiple genes is likely responsible for perturbed IgA production and abnormal B-cell maturation

Our results establish the existence of widespread epigenetic and transcriptional aberrations that are largely shared across the three disorders, suggesting functionality. We therefore asked whether they can shed light on some specific aspects of the immune dysfunction. We first performed a pathway analysis of the commonly disrupted genes (either at the expression or promoter accessibility level; **Methods**). This yielded several potentially affected pathways (**Supplemental Tables 9 and 10**). However, most of these were of general relevance and did not pinpoint very specific pathologies.

We then reasoned that we might gain more insight by focusing on two of the specific phenotypes seen in KS1: abnormal B-cell maturation, and IgA deficiency[13,20]. We set out to test if these are attributable to the collective dysregulation of multiple genes, or to the abnormal expression of a select few. To define relevant gene sets, we first obtained the set of all transcription factors encoded in the mouse genome that are expressed in CD19+ B cells

(**Methods**); this choice was motivated by the fact that transcription factors are critical regulators of cellular differentiation and maturation. We then examined the ranks of these transcription factors in the KS1 p-value distribution, and observed a strong shift indicative of global dysregulation (**Figure 5a;** $p < 2.2e-16$). For IgA deficiency, we assembled a list of 75 genes known to lead to IgA deficiency when individually knocked out in mouse (**Methods**). Examination of the KS1 p-value ranks of these genes also highlighted a collective shift towards lower p-values (**Figure 5b;** $p=0.02$). Together, these results strongly suggest the collective (but often minor) dysregulation of many genes.

Turning our attention to KS2 and RT1, we observed similar results for transcription factors, with substantial contribution from a set of transcription factors commonly dysregulated across the three MDEMs (**Figure 5a, c**). However, when assessing the IgA deficiency genes, we only observed the signal in RT1, and not in KS2 (**Figure 5b, d**). This was surprising, given the high phenotypic similarity between KS1 and KS2. To verify these predictions at the phenotypic level, we measured serum IgA in the KS1/2 and wild-type mice. In agreement with the collective behavior of IgA-related genes, we found no difference in IgA levels between the KS2 and wild-type, while we recapitulated our previous result of IgA deficiency in KS1 mice[13] (**Figure 5e;** $p = 0.8$ for KS2 vs WT, $p = 0.0008$ for KS1 vs WT).

Finally, we found no evidence that these collective defects in the expression of transcription factors and IgA-deficiency associated genes are driven by similar shifts towards abnormal promoter accessibility ($p > 0.1$ in all cases).

Discussion:

Our study shows that three MDEMs caused by loss-of-function variants in three distinct epigenetic regulators, share common abnormalities at the chromatin and gene expression level. These abnormalities show evidence of functionality, as illustrated by the fact that: a) many chromatin changes at promoters are linked to downstream gene expression changes, and b) systematic expression changes affect genes known to contribute to specific, well-characterized phenotypic features (IgA deficiency, abnormal B cell maturation) seen in these MDEMs.

In terms of understanding the pathogenesis of MDEMs, our results clearly point towards a generalized, systems-level dysregulation, with a multitude of cellular processes/pathways affected. From our present study it is unclear how exactly these combine to ultimately give rise to the phenotypic manifestations; elucidating this will be an important challenge going forward. It is also worth noting that the emergent picture bears similarities to the molecular basis of complex diseases, with many widely distributed, small-effect perturbations, ultimately generating

the phenotype[27]. This is perhaps not unexpected, given that epigenetic regulators are typically trans-acting proteins that act at many locations. It also suggests that, even though MDEMs are single-gene Mendelian disorders, they might best be conceptualized as effectively complex disorders. This may also explain the broadness of the phenotype in MDEMs[28], and the decreased penetrance of many phenotypes in patients, that are fully penetrant in mouse models.

It is notable that we find greater molecular overlap between KS1 and KS2 than between either of them and RT1, in agreement with the greater similarity between the two KS types at the phenotypic level. It should be mentioned, however, that: a) this result may be partly driven by the fact that KS1/2 were compared against a common wild-type group, and b) specific sub-phenotypes provide exceptions to this rule, as evidenced by the multi-genic abnormalities in IgA deficiency genes, which are shared between KS1 and RT1 but are absent in KS2. Together, these results suggest that deep phenotyping of MDEMs combined with molecular characterization can yield new insights into the pattern of their shared features.

One unexpected finding was that, at promoters, almost all of the commonly disrupted peaks exhibit a shift towards a more open chromatin state, even though the causative mutations of all three disorders are thought to push towards a more closed chromatin state, based on the histone marks they affect[29]. One possible explanation is that these common hits represent indirect effects, arising downstream of the initial effects of the mutations. Alternatively, the hypothesis that loss of the epigenetic regulators disrupted in our three disorders would lead to closed chromatin may not hold. Finally, there is the possibility that the causative mutations lead to a non-specific cellular compensatory response, which causes increased chromatin openness at several genomic locations such as the adaptive stress response[30]. The latter is supported by the fact that many of the shared genes uncovered in this study are not known to be direct KMT2D, KDM6A or CREBBP target genes. Regardless of the exact reason, this observation warrants future exploration.

Finally, we note that our study differs from recent studies of DNA methylation in the peripheral blood of MDEM patients. In these studies, the goal is to derive “episignatures” with the capacity for robust phenotypic prediction. While the predictive power of these episignatures is evidenced by the ability to provide a diagnosis[31], the functional relevance episignatures is, however, uncertain, for two main reasons. First, peripheral blood is a mixture of several cell types, and changes in cell-type composition - which often occur in disease - can severely confound the differential DNAm analysis[32,33]. Second, the great majority of these MDEMs are caused by variants in histone modifiers or chromatin remodelers. As a result, even if cell-type heterogeneity were to be ignored, these DNAm aberrations are by definition secondary events.

In contrast, our strategy is specifically designed to yield a catalog of abnormalities with primary functional role in common MDEM pathogenesis.

In summary, we propose the study of the Mendelian Disorders of the Epigenetic Machinery as a principled approach for systematically mapping causally relevant epigenetic variation in mammals. The commonly discovered hits among the three disorders studied here almost exclusively demonstrate an increase in open chromatin at promoters, which is counterintuitive to the function of the individual causative genes and may either suggest a previously unexpected role for them, or an undescribed systemic compensatory response. Finally, we suggest that MDEMs are effectively complex disorders, arising from widely distributed epigenetic perturbations across the genome.

Methods:

Mice

We performed all mouse experiments in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals and all were approved by the Animal Care and Use Committee of The Johns Hopkins University. We genotyped mice using standard genotyping and PCR methods.

KS1. Kmt2d$^{+/\beta Geo}$ mice are fully backcrossed to C57BL/6J and this backcrossing is verified by SNP genotyping.[17] These mice are also known as *Mll2*$^{Gt(RRt024)Byg}$, and originally obtained from BayGenomics and backcrossed in the Bjornsson laboratory.

Primers:

βGeo F-CAAATGGCGATTACCGTTGA, R-TGCCCAGTCATAGCCGAATA;

Tcrd (control) F-CAAATGTTGCTTGTCTGGTG, R-GTCAGTCGAGTGCACAGTTT

KS2. Kdm6a$^{+/-}$ mice were acquired from European Mouse Mutant Archive (EMMA) but this model has also been called: *Kdm6a*$^{tm1a(EUCOMM)Wtsi}$. Mice were crossed with flippase expressing mice (B6.Cg-Tg(ACTFLPe)9205Dym/J) (Jackson Laboratories) to remove the third exon of Kdm6a, and then progeny were crossed with Cre expressing mice driven by CMV (B6.C-Tg(CMV-cre)1Cgn/J) (Jackson Laboratories), to generate the *Kdm6a*$^{tm1d(EUCOMM)Wtsi}$ allele. Mice were backcrossed on C57BL/6J to maintain the *Kdm6a*$^{tm1d(EUCOMM)Wtsi}$ allele.

Primers:

Kdm6aTm1c F-AAGGCGCATAACGATACCAC, Floxed LR-ACTGATGGCGAGCTCAGACC;

Tcrd (control) F-CAAATGTTGCTTGTCTGGTG, R-GTCAGTCGAGTGCACAGTTT

RT1. Crebbp$^{+/-}$ mice also known as *Crebbp*^{tm1Dli}, were acquired from Jackson laboratories but established by Kung et al.[16] These mice were maintained on a C57BL/6J background in the Bjornsson laboratory.

Primers:

Wild-Type. We used littermates for wildtype controls.

Sex disaggregation

We performed all experiments in female mice to enable comparison between all three disease models as *KDM6A* (KS2 model) is present on the X chromosome and is lethal in male mice when knocked out. Therefore, we are unable to present sex-disaggregated data.

Blood cell isolation

We obtained peripheral blood from 2.5-3.5 month old female mice by facial vein bleed. 150-250 ul blood was collected in K_2EDTA blood collection tubes (BD Microtainer 365974) and red blood cells were lysed for 7-15 minutes at room temperature in 2mL red blood cell lysis solution (15.5mM NH4Cl, 1mM KHCO3, 0.01mM EDTA). We diluted lysed blood with excess balanced salt solution (Gey's or 1xPBS), manually removed large clots using pipet tip, and spun at 500g for 10 minutes 4'. Second lysis at room temperature was performed for samples with large amounts of remaining red blood cells then spun. We isolated CD19+ B cells by positive selection using CD19+ microbeads for mouse (Miltenyi 130-052-201) following manufacturer protocols, then counted and aliquoted samples on ice to further process for ATAC-seq and RNA-seq.

ATAC-seq

We performed ATAC-seq using a modified FastATAC protocol[22,23]. Specifically, we resuspended 5k cells per reaction in 1xPBS and quickly spun to remove residual EDTA from isolation steps, and then resuspended in tagmentation reaction mix for 30 min (2.5uL TD1, 1X TD Buffer, Illumina Nextera DNA, FC-121-1030; .25uL 1% digitonin, Promega G9441; 1xPBS;) gently shaking (300rpm on Eppendorf thermomixer) at 37'. We purified reactions using Zymo DNA Clean and Concentrator-5 kit (Zymo D4013) following manufacturer protocols and eluted with 10.5uL water to recover 10uL. Each reaction was then amplified and indexed as described[22], total sample amplification cycles range from 6-10 cycles. After indexing and amplification, we purified samples using Select-A-Size purification columns (Zymo D4080) with a cutoff of 150bp to remove adapter dimers to allow for efficient sequencing on patterned flow cells, checked library size on BioAnalyzer using DNA High Sensitivity reagents (Agilent 5067-4626) and determined concentration using Qubit dsDNA HS Assay Kit (ThermoFisher Q32851). We pooled and sequenced on Illumina HiSeq4000 using PE flow cells with 100-8-8-100 read length using standard manufacturer protocols. Samples were clustered to aim for 60M reads per sample Samples were demultiplexed using Illumina pipeline bcl2fastq2 v2.20 with all defaults except --use-bases-mask Y100n, I8, I8, Y100n.

ATAC-seq mapping and peak calling

We mapped the ATAC-seq reads to the mm10 mouse assembly using bowtie2[34], with default parameters. We removed duplicate reads with the "MarkDuplicates" function from Picard (http://broadinstitute.github.io/picard/), and subsequently also removed mitochondrial reads using samtools[35]. We then created genotype-specific meta-samples, by merging all the individual bam files corresponding to samples from mice of a given genotype. This yielded one

meta-sample for KS1, one for KS2, and one for RT1. For wild-type mice, we created two such meta-samples, one from the wild-type littermates of the KS1 and KS2 cohorts (to which the KS1 and KS2 mutant mice were compared to), and one for the wild-type littermates of the RT1 cohort (to which the RT1 mutant mice were compared to). For each of the 5 resulting meta-samples, we then called peaks using MACS2[36], with the "keep-dup" parameter equal to "all".

ATAC-seq differential analysis

We first defined the set of features to be tested as differential, by unionizing the peaks from all meta-samples. This yielded 78,193 genomic intervals (median size = 690bp, 95th percentile = 1,774, range = 151 - 11363). To verify that these intervals are not likely to be false positives, we compared them to publicly available DNase Hypersensitivity Sites in B cells (CD19+) from the ENCODE project (https://www.encodeproject.org/experiments/ENCSR000CMM/). We converted the DHS coordinates from mm9 to mm10 using liftOver. We then unionized the intervals from the two DHS replicates to create a common set of 112,728 DHS's. We found that 78,101 of our 78,193 regions (99.88%) overlapped DHS's, providing strong orthogonal evidence that they represent true B cell regulatory regions.

We then counted the number of reads from each sample that map to each of the 78,193 features, using the featureCounts function from the Rsubread R package[37], with the following parameters: requireBothEndsMapped = TRUE, countChimericFragments = FALSE, countMultiMappingReads = FALSE, minOverlap = 3. This resulted in a count matrix with rows corresponding to features (the aforementioned genomic intervals), and columns to samples. This count matrix served as input for the differential analysis, which we performed using DESeq2[38], retaining only features with a median (across samples) count greater than 10. We used Surrogate Variable Analysis[39] to estimate unobserved confounding variables, and adjusted for those in the differential analysis (without explicitly including other covariates in the model).

To derive the list of features overlapping promoters, we first obtained promoter coordinates with the "promoters" function from the EnsDb.Mmusculus.v79 R package, with the parameters "upstream" and "downstream" both equal to 2000. We subsequently restricted to protein-coding transcripts, using the "tx_biotype" filter. The overlapping features were then obtained using the findOverlaps function from the GenomicRanges R package.

RNA-seq

We spun approximately 100k-500k cells at 300-500g for 5 min at 4', homogenized in Trizol (Invitrogen 15596018) and stored at -80 until extraction. We extracted and isolated RNA by

phase separation using standard protocols followed by purification using the Direct-zol RNA microprep kit (Zymo R2060) with an on-column DNAse step per manufacturer directions. Once purified, we quantified RNA using Quant-iT RiboGreen RNA Assay Kit (ThermoFisher R11490) or Qubit RNA HS Assay Kit (ThermoFisher Q32852), and checked quality by Bioanalyzer with RNA 6000 Pico Kit (Agilent 5067-1513). All samples show high quality RNA with RIN greater than 9. We used 20ng RNA per KS1 & KS2 & matched wild-type sample and 100ng per RT & matched wild-type sample as input to capture mRNA (NEBNext Poly(A) mRNA Magnetic Isolation Module; NEB #E7490) followed by library generation using NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (NEB E7760/E7765) per manufacturer protocols. We determined library size and quality using BioAnalyzer with DNA High Sensitivity reagents (Agilent 5067-4626), and determined concentration using Qubit dsDNA HS Assay Kit (ThermoFisher Q32851) and KAPA Library Quantification Kit for qPCR (KAPA KK4824). We pooled samples and sequenced on Illumina HiSeq4000 using PE flow cells with 100-8-8-100 read length using standard manufacturer protocols. Samples were clustered to aim for 60M reads per sample. Samples were demultiplexed using Illumina pipeline bcl2fastq2 v2.20.

RNA-seq mapping and differential analysis

We first obtained a fasta file (Mus_musculus.GRCm38.cdna.all.fa.gz) containing all mouse cDNA sequences from Ensembl (http://uswest.ensembl.org/Mus_musculus/Info/Index, version 91, downloaded January 2018). We used this file to build an index and pseudo-map the RNA-seq reads with Salmon (v0.10)[40]. We subsequently imported the resulting transcript quantifications into R to get gene-level counts, using the tximport R package[41]. The differential analysis was then performed with DESeq2, following the same steps as with ATAC-seq.

Principal Component Analysis

All PCA plots were generated as follows. We first applied a variance stabilizing transformation to the count matrices (either genes-by-samples or genomic-intervals-by-samples), as implemented in the vst function from DESeq2. We then used the resulting matrix to perform the PCA with the plotPCA function.

Testing for statistically significant overlap between two lists of differential features and identifying the common hits

Our problem is cast in the following setting. Assume we have performed two experiments, each of which involves measuring multiple features (e.g. genes or peaks) in two conditions and performing a differential analysis. The two experiments measure the same set of features. Because the two experiments investigate different biological systems, we don't expect the set of (true) differential features to be identical. But we are interested in the extent of the overlap between the two sets of features, specifically

a) Is there statistically significant overlap between the two sets of differential features and how big is it?
b) Which features are differential in both lists?

Our approach to these questions is a conditional approach: we ask, does information about the result in experiment 1 affect our interpretation of experiment 2?

We first (arbitrarily) designate one of the two experiments as experiment 1. We test m features, and for each feature i, we let X_i be a factor with values in $\{0,1\}$ expressing whether the feature was significant in experiment 1 ($X_i = 0$) or not ($X_i = 1$). We are interested in whether the variable $X = (X_1, \dots, X_m)$ is an informative covariate for experiment 2, using terminology from recent work in covariate-powered multiple hypothesis testing[42,43].

We now consider experiment 2. We split the features into two groups, conditional on the results in experiment 1. Group 0 consists of the features which were found to be differential in experiment 1 (the size of group 0 is n) and group 1 consists of the features which were not differential in experiment 1 (of which we have $m - n$). We let each group have its own proportion of differential features, that is we introduce parameters $\pi_{0\,|\,0}$ and $\pi_{0\,|\,1}$. Let $\mathrm{Y_i}$ be the indicator whether the i'th feature is differential in experiment 2 or not. Then

$$P(Y_i = 0) = P(Y_i = 0|X_i = 0)P(X_i = 0) + P(Y_i = 0|X_i = 1)P(X_i = 1)$$
$$= \pi_{0\,|\,0}\mathrm{P}(\mathrm{X_i} = 0) + \pi_{0\,|\,1}\mathrm{P}(\mathrm{X_i} = 1)$$

Our null hypothesis is that experiment 1 is not informative about experiment 2, or in other words

$$H_0: \pi_{0\,|\,0} = \pi_{0\,|\,1}$$

Under this null hypothesis $P(Y_i = 0) = \pi_{0\,|\,0}$. Furthermore, $\pi_{0\,|\,0}$ should have the same distribution as the proportion of significant features in a random sample of n features from experiment 2, which we term $\pi_0^{(n)}$.

This gives us the following method for assessing H_0:

1. Analyze experiment 1 and decide which features are differential or not.

2. Analyze experiment 2, but only the features which are differential in experiment 1. Estimate $\hat{\pi}_{0\,|\,0}$.
3. Repeatedly, draw n features and estimate $\hat{\pi}_0^{(n)}$ to get a null distribution.

In practice we can estimate $\hat{\pi}_{0\,|\,0}$ and $\hat{\pi}_0^{(n)}$ using a number of different methods. We use Storey's method as implemented in the qvalue package in R[44,45], with the “pi0.method” parameter set to "bootstrap". This tells us the size of the overlap and the extent to which it is significant. To estimate which features are in the overlap, we use the "qvalue" function on the features which are significant in the analysis of experiment 1, with the FDR level set to 10%.

Identification of differentially expressed genes with differentially accessible promoter peaks

For figure **3b**, we first selected the genes downstream of the top 1000 differentially accessible promoter peaks, the latter being ranked based on their p-values in each disorder. Out of these genes, we retained those differentially expressed using the “qvalue” function from the qvalue R package, with the gene p-values as input and the “fdr.level” parameter set to 0.1. In cases where there were more than one peaks in the same promoter, we calculated the median logFC across these peaks. For **Figure 3a,** we slid the threshold for the top differentially accessible promoter peaks in each disorder from 1000 to 5000, and estimated the proportion of differentially expressed downstream genes used the “pi0est” function from the qvalue R package with the “pi0.method” parameter set to “bootstrap”.

Finally, for **Figure 4h** we employed the analogous procedure to **Figure 3a** in order to estimate the proportion of differentially accessible peaks in promoters of differentially expressed genes for different thresholds.

Reactome pathway analysis

We used the goseq R package[46] to perform pathway analyses for the commonly disrupted genes, based on Reactome pathways[47]. As our assayed gene set, we used the set of all genes included in all three differential expression analyses, or the set of all genes that had at least one promoter peak included in all three differential accessibility analysis. As our differential gene set, we used the set of genes commonly differentially expressed across the three MDEMs, or the set of genes with at least one commonly differentially accessible promoter peak. The top 20 enriched pathways are provided in Supplemental Tables 9 and 10, respectively.

TF motifs

We obtained a bed file (mm10.archetype_motifs.v1.0.bed) containing the genomic positions of 233 non-redundant TF motifs, from https://www.vierstra.org/resources/motif_clustering [48]. We then restricted to motifs that had at least 1 base overlapping our set of unionized B cell peaks (see sections **ATAC-seq mapping and peak calling** and **ATAC-seq differential analysis**). Subsequently, we tested each motif for enrichment as described in the results section using the fisher.test function in R.

Gene catalogs

Transcription Factors:

We obtained a list of 1,254 genes encoding for human TFs from Barrera et al., 2016.[49] We the used the biomaRt R package to obtain the mouse orthologs of these TF genes, with the ENSEMBL ids as our filter. We only retained high-confidence orthologs ("mmusculus_homolog_orthology_confidence" equal to 1). Finally, we restricted to TFs included in all three differential analyses (KS1 vs WT, KS2 vs WT, and RT1 vs WT).

IgA deficiency genes:

We used the Mammalian Phenotype Browser on the Mouse Genome Informatics database[50] to obtain a catalog of genes known to lead to IgA deficiency when individually knocked-out. Specifically, we used "decreased IgA level" as the phenotype term and then obtained all the resulting genes, regardless of the genetic background. In cases of double knockouts, we included both genes.

ELISA for serum IgA levels

We performed ELISAs on serum IgA from peripheral blood samples as previously described.[13]

Code availability

All code for the analyses in this manuscript is available at https://github.com/hansenlab/mdem_overlap

Acknowledgements: H.T.B. and T.R.L. are supported by a grant from the Louma G. Foundation. H.T.B. is also supported by grants from the Icelandic Research Fund (#195835-051, #206806-051) and the Icelandic Technology Development Fund (#2010588-0611). K.D.H, H.T.B and LB were partly supported by the 2016 Discovery Award from Johns Hopkins University. L.B. was partly supported by the Maryland Genetics, Epidemiology and Medicine (MD-GEM) training program, funded by the Burroughs-Wellcome Fund.

References

1. Rakyan, V. K., Down, T. A., Balding, D. J. & Beck, S. Epigenome-wide association studies for common human diseases. *Nature Reviews Genetics* **12**, 529–541 (2011).
2. Lappalainen, T. & Greally, J. M. Associating cellular epigenetic models with human phenotypes. *Nature Reviews Genetics* **18**, 441–451 (2017).
3. Barlow, D. P. & Bartolomei, M. S. Genomic Imprinting in Mammals. *Cold Spring Harb Perspect Biol* **6**, a018382 (2014).
4. Sutcliffe, J. S. *et al.* DNA methylation represses FMR-1 transcription in fragile X syndrome. *Hum Mol Genet* **1**, 397–400 (1992).
5. LaCroix, A. J. *et al.* GGC Repeat Expansion and Exon 1 Methylation of XYLT1 Is a Common Pathogenic Variant in Baratela-Scott Syndrome. *The American Journal of Human Genetics* **104**, 35–44 (2019).
6. Bertozzi, T. M. & Ferguson-Smith, A. C. Metastable epialleles and their contribution to epigenetic inheritance in mammals. *Semin Cell Dev Biol* **97**, 93–105 (2020).
7. Rakyan, V. K., Blewitt, M. E., Druker, R., Preis, J. I. & Whitelaw, E. Metastable epialleles in mammals. *Trends in Genetics* **18**, 348–351 (2002).
8. Fahrner, J. A. & Bjornsson, H. T. Mendelian disorders of the epigenetic machinery: postnatal malleability and therapeutic prospects. *Hum. Mol. Genet.* **28**, R254–R264 (2019).
9. Boukas, L. *et al.* Coexpression patterns define epigenetic regulators associated with neurological dysfunction. *Genome Res.* (2019) doi:10.1101/gr.239442.118.

10. Carosso, G. A. *et al.* Precocious neuronal differentiation and disrupted oxygen responses in Kabuki syndrome. *JCI Insight* **4**, (2019).
11. Fahrner, J. A. *et al.* Precocious chondrocyte differentiation disrupts skeletal growth in Kabuki syndrome mice. *JCI Insight* **4**, (2019).
12. Ang, S.-Y. *et al.* KMT2D regulates specific programs in heart development via histone H3 lysine 4 di-methylation. *Development* **143**, 810–821 (2016).
13. Pilarowski, G. O. *et al.* Abnormal Peyer patch development and B-cell gut homing drive IgA deficiency in Kabuki syndrome. *Journal of Allergy and Clinical Immunology* **145**, 982–992 (2020).
14. Zhang, J. *et al.* Disruption of KMT2D perturbs germinal center B cell development and promotes lymphomagenesis. *Nature Medicine* **21**, 1190–1198 (2015).
15. Ortega-Molina, A. *et al.* The histone lysine methyltransferase KMT2D sustains a gene expression program that represses B cell lymphoma development. *Nature Medicine* **21**, 1199–1208 (2015).
16. Kung, A. L. *et al.* Gene dose-dependent control of hematopoiesis and hematologic tumor suppression by CBP. *Genes Dev* **14**, 272–277 (2000).
17. Bjornsson, H. T. *et al.* Histone deacetylase inhibition rescues structural and functional brain deficits in a mouse model of Kabuki syndrome. *Sci Transl Med* **6**, 256ra135 (2014).
18. Margot, H. *et al.* Immunopathological manifestations in Kabuki syndrome: a registry study of 177 individuals. *Genetics in Medicine* **22**, 181–188 (2020).
19. Lindsley, A. W. *et al.* Defects of B-cell terminal differentiation in patients with type-1 Kabuki syndrome. *Journal of Allergy and Clinical Immunology* **137**, 179-187.e10 (2016).
20. Saettini, F. *et al.* Prevalence of Immunological Defects in a Cohort of 97 Rubinstein–Taybi Syndrome Patients. *J Clin Immunol* **40**, 851–860 (2020).
21. Frans, G. *et al.* Mild humoral immunodeficiency in a patient with X-linked Kabuki syndrome. *Am J Med Genet A* **170**, 801–803 (2016).

22. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature Genetics* **48**, 1193–1203 (2016).
23. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* **10**, 1213–1218 (2013).
24. Voss, T. C. & Hager, G. L. Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nature Reviews Genetics* **15**, 69–81 (2014).
25. John, S. *et al.* Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature Genetics* **43**, 264–268 (2011).
26. Guertin, M. J. & Lis, J. T. Chromatin Landscape Dictates HSF Binding to Target DNA Elements. *PLOS Genetics* **6**, e1001114 (2010).
27. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).
28. Bjornsson, H. T. The Mendelian disorders of the epigenetic machinery. *Genome Res.* **25**, 1473–1481 (2015).
29. Fahrner, J. A. & Bjornsson, H. T. Mendelian disorders of the epigenetic machinery: tipping the balance of chromatin states. *Annu Rev Genomics Hum Genet* **15**, 269–293 (2014).
30. Brose, R. D. *et al.* Activation of the stress proteome as a mechanism for small molecule therapeutics. *Hum Mol Genet* **21**, 4237–4252 (2012).
31. Aref-Eshghi, E. *et al.* Evaluation of DNA Methylation Episignatures for Diagnosis and Phenotype Correlations in 42 Mendelian Neurodevelopmental Disorders. *The American Journal of Human Genetics* **106**, 356–370 (2020).
32. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).
33. Jaffe, A. E. & Irizarry, R. A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology* **15**, R31 (2014).

34. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).

35. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

36. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137 (2008).

37. Liao, Y., Smyth, G. K. & Shi, W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res* **47**, e47–e47 (2019).

38. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).

39. Leek, J. T. & Storey, J. D. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLOS Genetics* **3**, e161 (2007).

40. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14**, 417–419 (2017).

41. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res* **4**, 1521 (2016).

42. Ignatiadis, N., Klaus, B., Zaugg, J. B. & Huber, W. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods* **13**, 577–580 (2016).

43. Chen, X., Robinson, D. G. & Storey, J. D. The functional false discovery rate with applications to genomics. *Biostatistics* doi:10.1093/biostatistics/kxz010.

44. Storey, J. D. The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Statist.* **31**, 2013–2035 (2003).

45. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *PNAS* **100**, 9440–9445 (2003).

46. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology* **11**, R14 (2010).

47. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res* **48**, D498–D503 (2020).

48. Vierstra, J. *et al.* Global reference mapping of human transcription factor footprints. *Nature* **583**, 729–736 (2020).

49. Barrera, L. A. *et al.* Survey of variation in human transcription factors reveals prevalent DNA binding changes. *Science* **351**, 1450–1454 (2016).

50. Eppig, J. T. Mouse Genome Informatics (MGI) Resource: Genetic, Genomic, and Biological Knowledgebase for the Laboratory Mouse. *ILAR J* **58**, 17–41 (2017).

Tables

Table 1

Gene	Gene expression	Chromatin accessibility	Function
Pard3b	↑	↑	Cell division and cell polarization processes
Nr3c2	↑	↑	Mineralocorticoid receptor
Epm2a	↑	↑	Serine/Threonine/Tyrosine phosphatase
Dip2c	↑	↑	Protein that interacts with transcription factors
Col4a5	↑	↑	Major collagen of basement membrane
Kif13a	↑	↑	Microtubule-based motor proteins
Reln	↑	↑	Secreted extracellular matrix protein
Zfp365	↑	↑	Transcription factor
Dmd	↑	↑	Connects cytoskeleton and the extracellular matrix
Maged1	↑	↑	p75 neurotrophin receptor mediated program
Chic1	↑	↑	Cysteine-rich hydrophobic (CHIC) domain-containing protein,
Pbx1	↑	↑	Transcription factor
Arhgef12	↑	↑	Guanine exchange factor
Slc36a4	↑	↑	Amino acid transporter
Golim4	↑	↑	Golgi protein
Bicd1	↑	↑	Role in intracellular cargo transport
Slc29a4	↑	↑	Monoamine transporter
Cask	↑	↑	Calcium/calmodulin-dependent serine protein kinase
Spry2	↑	↑	Inhibitory activity on receptor tyrosine kinase signaling proteins
Ndrg1	↑	↑	N-myc downregulated gene family member
Zfp827	↑	↑	Transcription factor
Slc25a4	↑	↑	Member of the mitochondrial carrier subfamily
Amz1	↓	↓	Zinc metalloprotease
Ppdpf	↓	↓	Regulator of exocrine pancreas development

Figures

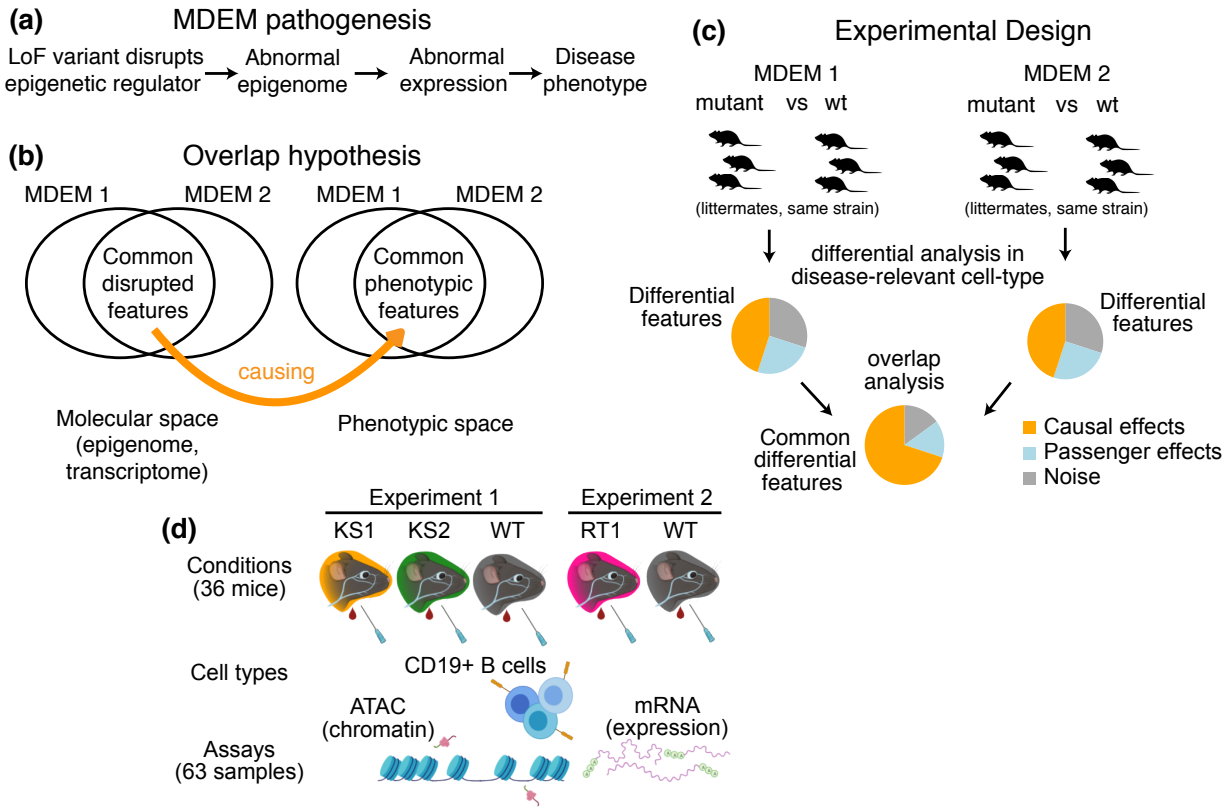


Figure 1. The theoretical framework for the present study.

- **a.** The causal chain of MDEM pathogenesis: the genetic disruption of an epigenetic regulator leads to epigenetic and transcriptomic consequences, which ultimately generate the phenotype.
- **b.** We hypothesize that the shared phenotypic features between MDEMs occur because of shared epigenetic and transcriptomic abnormalities.
- **c.** Our approach is designed to derive a set of abnormalities with high probability of causal relevance, by jointly comparing multiple MDEMs.
- **d.** Experimental design and workflow for sample generation in our present study. Created with BioRender.com

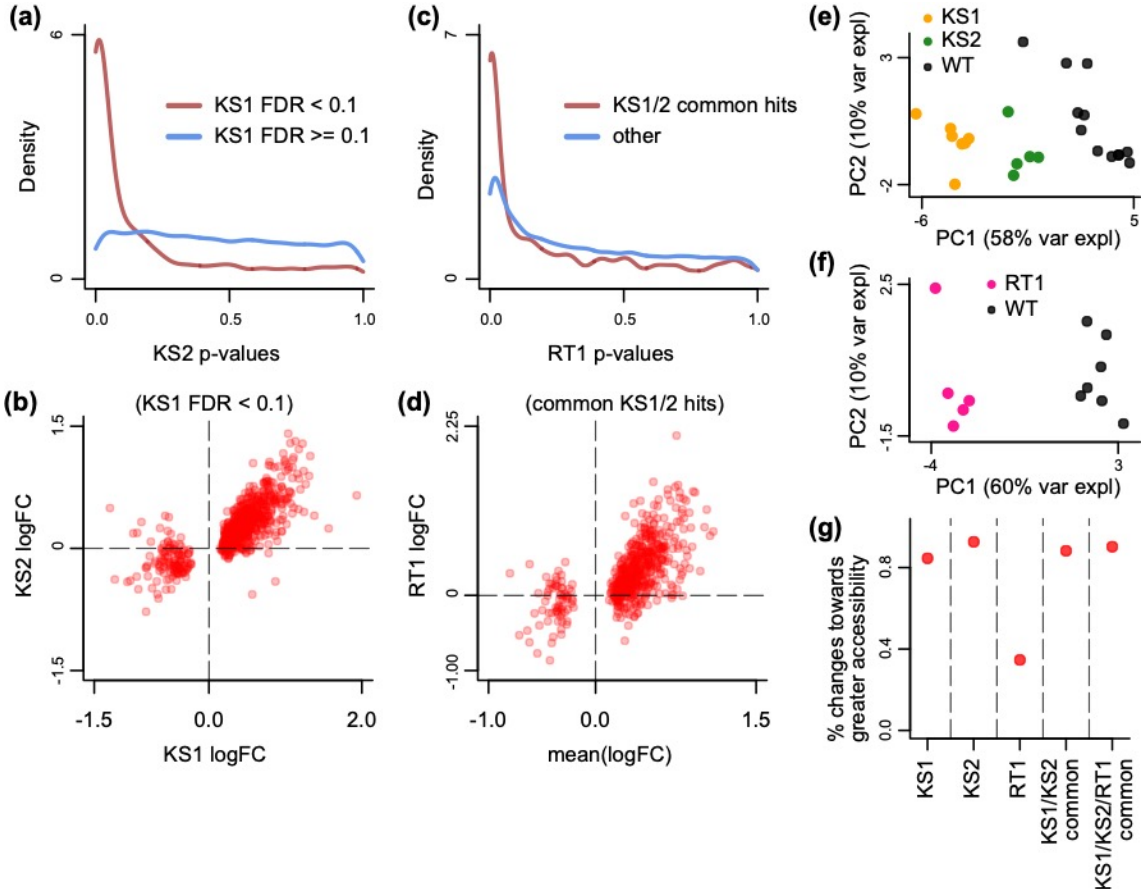


Figure 2. Evaluating the overlap between the differentially accessible promoter peaks in B cells in Kabuki type 1, Kabuki type 2, and Rubinstein-Taybi syndromes.

a. The distribution of p-values from the KS2 vs wild-type differential accessibility analysis for promoter peaks, stratified according to whether the same promoter peaks are significantly differentially accessible in KS1 (FDR < 0.1; red curve), or not (FDR >= 0.1; blue curve).

b. Scatterplot of log2(fold changes) from the KS1 vs WT (x-axis) promoter peak differential accessibility analysis against the corresponding log2(fold changes) in KS2 (y-axis). Each point corresponds to a peak. Shown are only peaks that are differentially accessible in KS1 (FDR < 0.1).

c. The distribution of p-values from the RT1 vs wild-type differential accessibility analysis for promoter peaks, stratified according to whether the same promoter peaks are commonly differentially accessible in KS1 and KS2 (FDR < 0.1, see Methods; red curve), or not (blue curve).

d. Scatterplot of log2(fold changes) from the RT1 vs WT (x-axis) differential accessibility analysis, against the mean log2(fold change) in KS1 and KS2. Each point corresponds to a peak. Shown are only peaks that are commonly differentially accessible in KS1 and KS2 (FDR < 0.1).

e. and

f. Principal component analysis plots using only the 313 promoter peaks identified as commonly differentially accessible between the three MDEMs. Each point corresponds to a mouse.

g. The proportion of differentially accessible promoter peaks that show increased accessibility in the mutant vs the wild-type mice.

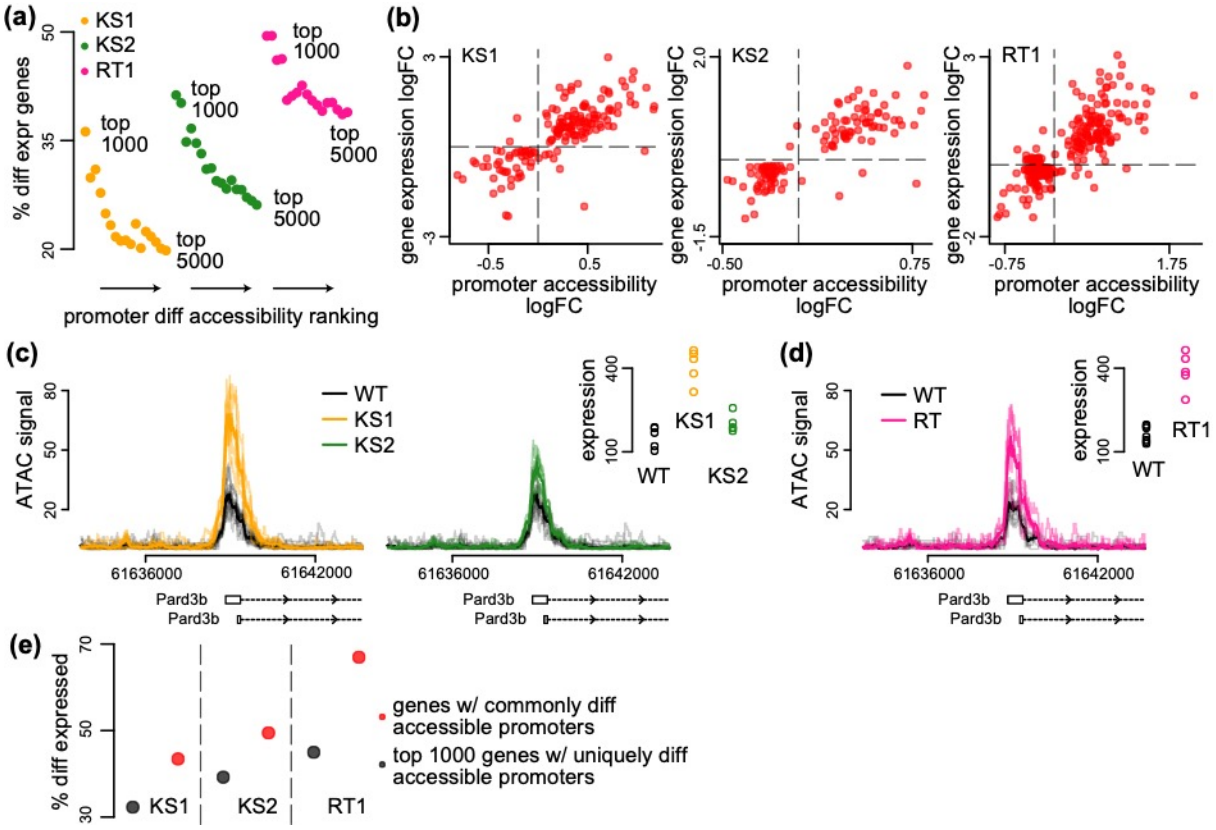


Figure 3. The relationship between differential accessibility of promoter peaks and differential expression of downstream genes in the three MDEMs.

a. The proportion of promoters with differentially expressed downstream genes in KS1, KS2, and RT, estimated for the top ranked differentially accessible promoter peaks. The estimation was repeated for different thresholds for inclusion into the top ranked list. For each MDEM, each point corresponds to a different threshold. Thresholds were slid from 1000 to 5000, in steps of 250.

b. Scatterplot of the accessibility log2(fold changes) of differentially accessible promoter peaks, against the expression log2(fold changes) of differentially expressed downstream genes, for each of the three MDEMs. Shown are only pairs where the promoter peak was within the top 1000 differentially accessible promoter peaks (ranked based on p-value), and the downstream gene was differentially expressed (10% FDR; methods). Each point corresponds to a gene-promoter pair. In cases where more than 1 peak in the same promoter was within the top 1000 differentially accessible peaks, the median(log2(fold change)) across all such peaks was calculated.

c. and

d. An example locus (*Pard3b*) with concordant changes in promoter peak accessibility and downstream gene expression in all three MDEMs

e. The proportion of promoters with differentially expressed downstream genes in KS1, KS2, and RT, estimated separately for the top 1000 uniquely differentially accessible promoters in each MDEM, versus the same proportion estimated for the genes downstream of the 313 commonly differentially accessible promoter peaks.

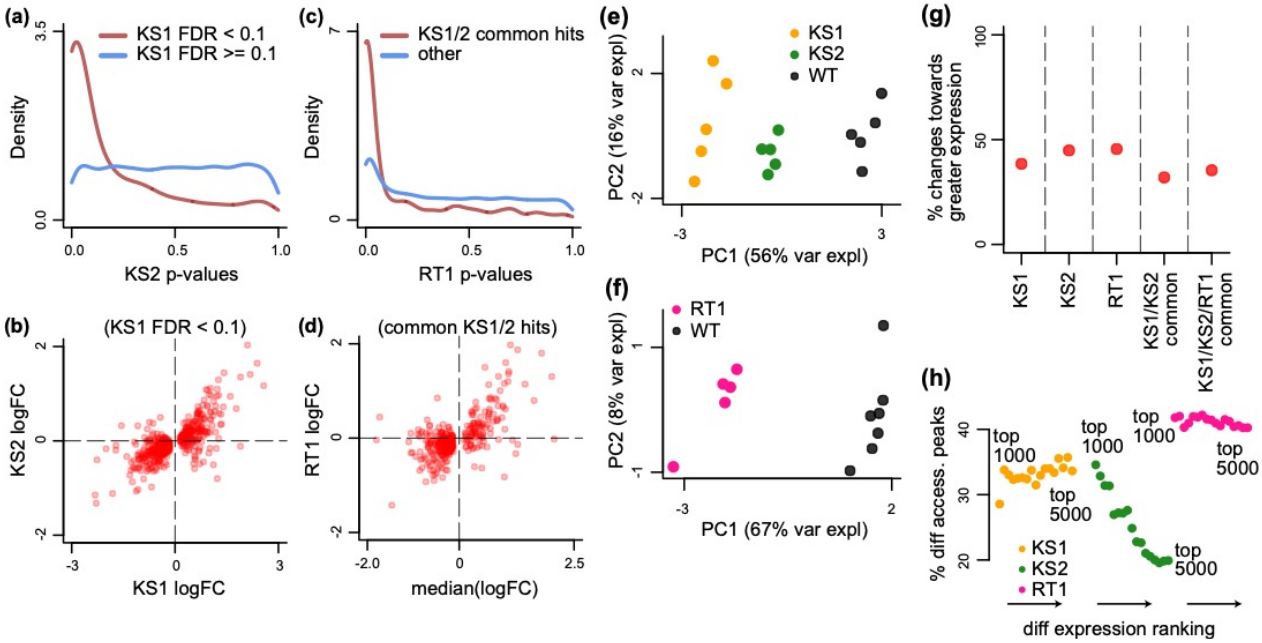


Figure 4. Evaluating the overlap between the differentially expressed genes in B cells in Kabuki type 1, Kabuki type 2, and Rubinstein-Taybi syndromes.

a. The distribution of p-values from the KS2 vs wild-type differential expression analysis, stratified according to whether the same genes are significantly differentially expressed in KS1 (FDR < 0.1; red curve), or not (FDR >= 0.1; blue curve).

b. Scatterplot of log2(fold changes) from the KS1 vs WT differential expression analysis (x-axis), against the corresponding log2(fold changes) in KS2 (y-axis). Each point corresponds to a gene. Shown are only genes that are differentially expressed in KS1 (FDR < 0.1).

c. The distribution of p-values from the RT1 vs wild-type differential expression analysis, stratified according to whether the same genes are commonly differentially expressed in KS1 and KS2 (FDR < 0.1, see Methods; red curve), or not (blue curve).

d. Scatterplot of log2(fold changes) from the RT1 vs WT (x-axis) differential expression analysis, against the mean log2(fold change) in KS1 and KS2. Each point corresponds to a gene. Shown are only genes that are commonly differentially expressed in KS1 and KS2 (FDR < 0.1).

e. and

f. Principal component analysis plots using only the 249 genes identified as commonly differentially expressed between the three MDEMs. Each point corresponds to a mouse.

g. The proportion of differentially expressed genes that show increased expression in the mutant vs the wild-type mice.

h. The proportion of genes with differentially accessible promoter peaks in KS1, KS2, and RT, estimated for the top ranked differentially expressed genes. The estimation was repeated for different thresholds for inclusion into the top ranked list. For each MDEM, each point corresponds to a different threshold. Thresholds were varied from 1000 to 5000, in steps of 250.

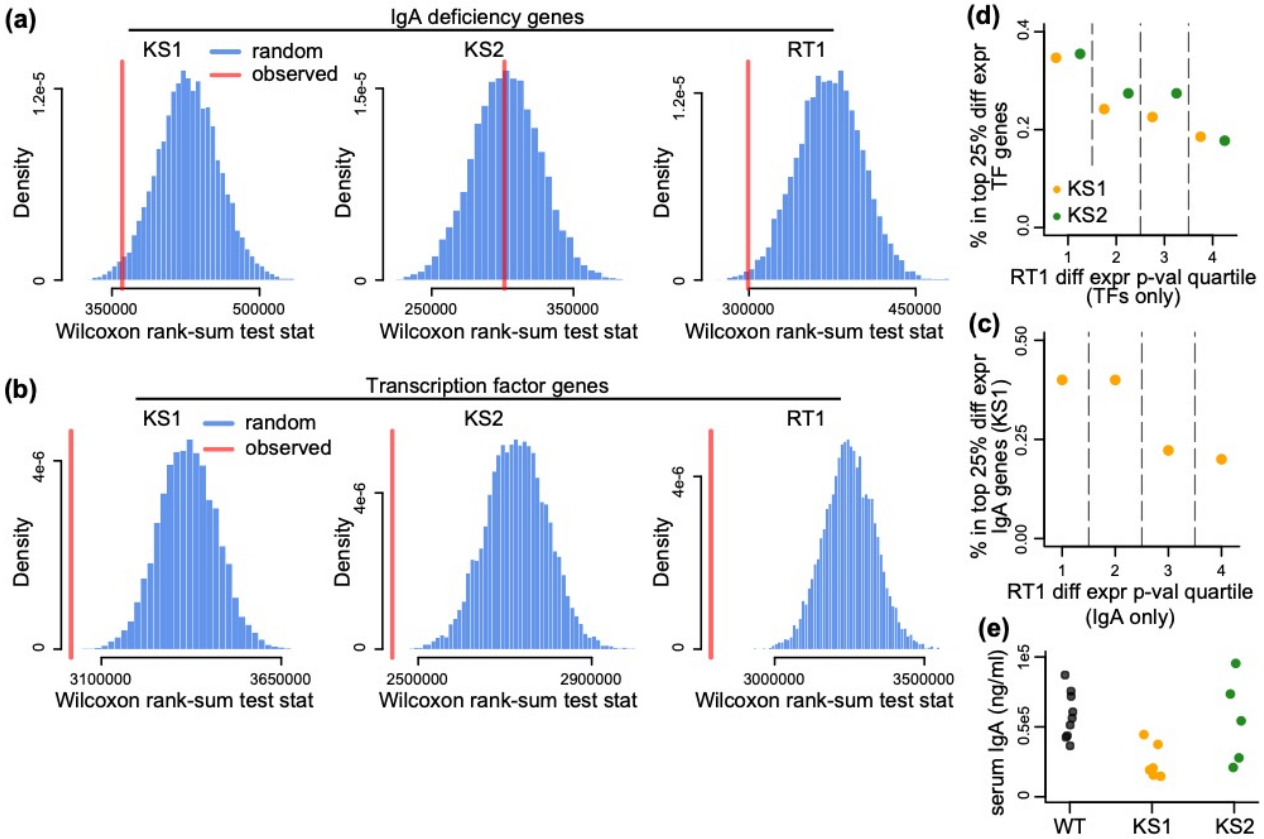


Figure 5. Evaluating genes known to encode for transcription factors, or individually contribute to IgA deficiency, for collective expression dysregulation.

a. The Wilcoxon rank-sum test statistic (red vertical line) computed after assembling a list of genes encoding transcription factors expressed in B cells (Methods), and comparing the distribution of their differential expression p-values to the p-value distribution of the rest of the genes included in the differential expression analysis. The blue distribution corresponds to the same statistic computed after randomly sampling gene sets of the same size as the IgA deficiency set, and comparing their p-value distribution to the p-values for the rest of the genes. The resampling was performed 10,000 times.

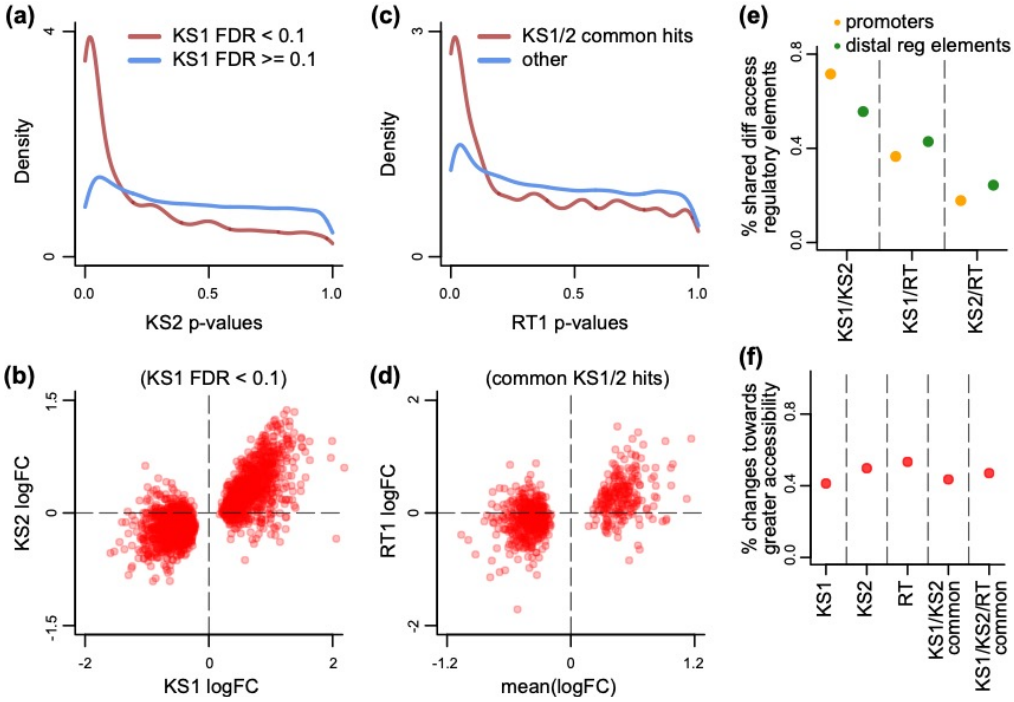
b. Same as a), but for genes known to individually contribute to IgA deficiency (Methods).

c. The percentage of TF genes that belong to the top 25% differentially expressed TFs in KS1 (orange dots), and KS2 (green dots), stratified according to their p-value quartile in RT.

d. Same as c), but for IgA deficiency genes compared in KS1 and RT.

e. Serum IgA levels in KS1, KS2 and wild-type mice.

Supplemental Figures



Supplemental Figure 1. Evaluating the overlap between the differentially accessible distal regulatory elements in B cells in Kabuki type 1, Kabuki type 2, and Rubinstein-Taybi syndromes.

a. The distribution of p-values from the KS2 vs wild-type differential accessibility analysis for promoter peaks, stratified according to whether the same promoter peaks are significantly differentially accessible in KS1 (FDR < 0.1; red curve), or not (FDR >= 0.1; blue curve).

b. Scatterplot of log2(fold changes) from the KS1 vs WT (x-axis) promoter peak differential accessibility analysis against the corresponding log2(fold changes) in KS2 (y-axis). Each point corresponds to a peak. Shown are only peaks that are differentially accessible in KS1 (FDR < 0.1).

c. The distribution of p-values from the RT1 vs wild-type differential accessibility analysis for promoter peaks, stratified according to whether the same promoter peaks are commonly differentially accessible in KS1 and KS2 (FDR < 0.1, see Methods; red curve), or not (blue curve).

d. Scatterplot of log2(fold changes) from the RT1 vs WT (x-axis) differential accessibility analysis, against the mean log2(fold change) in KS1 and KS2. Each point corresponds to a peak. Shown are only peaks that are commonly differentially accessible in KS1 and KS2 (FDR < 0.1).

e. The pairwise overlap between the differentially accessible peaks (promoters or distal regulatory elements) in the three MDEMs.

f. The proportion of differentially accessible distal regulatory elements that show increased accessibility in the mutant vs the wild-type mice.