Bias toward long gene misregulation in synaptic disorders can be an artefact of amplification-based methods

Ayush T. Raman[1,2,8], Amy E. Pohodich[2,3,8], Ying-Wooi Wan[2,4], Hari Krishna Yalamanchili[2,4], Bill Lowry[5], Huda Y. Zoghbi[2,3,4,6,*], Zhandong Liu[1,2,7,9,*]

[1]Graduate Program in Quantitative and Computational Biosciences, Baylor College of Medicine, Houston, TX 77030, USA; [2]Jan and Dan Duncan Neurological Research Institute at Texas Children's Hospital, Houston, Texas 77030, USA; [3]Department of Neuroscience, Baylor College of Medicine, Houston, Texas 77030, USA; [4]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA; [5]Department of Molecular, Cell and Developmental Biology, University of California, Los Angeles, Los Angeles, CA 90095, USA; [6]Howard Hughes Medical Institute, Baylor College of Medicine, Houston, Texas 77030, USA; [7]Department of Pediatrics, Section of Neurology, Baylor College of Medicine, Houston, TX, USA. [8]These authors contributed equally to this work. [9]Lead Contact. [*]Co-Corresponding Authors: Z.L. (zhandong.liu@bcm.edu) and H.Y.Z. (hzoghbi@bcm.edu).

Keywords: Long gene misregulation bias, *MECP2*, transcriptome profiling, microarray, RNA-sequencing, Nanostring

SUMMARY (149 words)

Several recent studies have suggested that genes that are longer than 100 kilobases are more likely to be misregulated in neurological diseases associated with synaptic dysfunction, such as autism and Rett syndrome. These length-dependent transcriptional changes are modest in *Mecp2*-mutant samples, but, given the low sensitivity of high-throughput transcriptome profiling technology, the statistical significance of these results needs to be re-evaluated. Here, we show that the apparent length-dependent trends previously observed in MeCP2 microarray and RNA-Sequencing datasets, particularly in genes with low fold-changes, disappeared after accounting for baseline variability estimated from randomized control samples. As we found no similar bias with NanoString technology, this long-gene bias seems to be particular to PCR amplification-based platforms. In contrast, authentic long gene effects, such as those caused by topoisomerase inhibition, can be detected even after adjustment for baseline variability. Accurate detection of length-dependent trends requires establishing a baseline from randomized control samples.

HIGHLIGHTS

- Length-dependent gene misregulation is not intrinsic to *Mecp2* disruption.
- Topoisomerase inhibition produces an authentic long gene bias.
- PCR amplification-based high-throughput datasets are biased toward long genes.

INTRODUCTION

The capacity for large-scale analysis of transcriptional changes in human disease has attracted considerable research attention, most recently in studies related to autism spectrum disorders, including Angelman syndrome, Rett syndrome (RTT), Fragile X syndrome, and autism itself (Zoghbi and Bear, 2012). Microarray and RNA-Seq studies have demonstrated that these disorders involve the dysregulation of thousands of neuronal genes. Several recent studies have also suggested that the genes dysregulated in these syndromes tend to be those that consist of more than 100 kilobases (Katz et al., 2016; Zylka et al., 2015). This intriguing length bias has been observed across both epigenetic and transcriptional datasets such as Angelman syndrome (Huang et al., 2011), Rett syndrome (Gabel et al., 2015; Kinde et al., 2016; Sugino et al., 2014), Fragile X syndrome (Gabel et al., 2015; Ouwenga and Dougherty, 2015) autism (King et al., 2013; Sullivan et al., 2015). The degree of bias tends to be fairly mild, however, long genes are themselves overrepresented in the brain compared to other tissues in the body (Zylka et al., 2015). It seems worthwhile to examine this apparent bias more closely in gene expression datasets.

The afore-mentioned gene expression studies (Gabel et al., 2015; King et al., 2013; Sugino et al., 2014; Sullivan et al., 2015) partitioned the entire genome into hundreds of overlapping bins (or windows), with each bin containing hundreds of genes. Within each bin, the average fold-change in wildtype or untreated brain tissue was compared to that observed in the knock-out or treatment groups, and a running average $\log_2$ fold-change was plotted against the average gene length. In these running average plots, long genes demonstrated a non-zero mean compared to short genes. These analyses did not, however, establish a baseline of inherent variation among samples within a given genotype, and they did not employ a statistical test to determine the significance of the length-dependent changes. It should be noted that variations in measured gene expression can arise because of RNA priming (Hansen et al., 2010; Li et al., 2010), GC-content (Risso et al., 2011), transcript length (Oshlack and Wakefield, 2009), or library preparation (Lahens et al., 2014), all of which must be accounted for in order to avoid unwarranted biological conclusions (Robert and Watson, 2015; Wan et al., 2014).

We, therefore, analysed a comprehensive list of large datasets derived from different transcriptome profiling technologies and set out to determine the best way to enhance the signal-to-noise ratio. To this end, we began by analysing technical replicates using benchmark datasets. Using these datasets, we developed an approach to reliably identify patterns with respect to gene regulation, and we then applied our approach to analyse datasets for which long gene trends have been reported.

RESULTS

Baseline length dependency should first be estimated from the control groups: the Topotecan study as a positive control

Preferential dysregulation of long genes is generally estimated by computing the average gene expression fold-changes between experimental groups and plotting this fold-change against the gene length (Gabel et al., 2015; King et al., 2013; Sugino et al., 2014), also known as running average plots (red curve in Fig 1A, Experimental Procedures). However, the statistical significance of running average plots has never been evaluated in the current literature. Here, we propose an approach to estimate statistical significance by constructing a null distribution of the running average plot from randomized control samples (Figure S1).

The first data that we analyzed was obtained from a study of the topoisomerase 1 inhibitor topotecan (King et al., 2013), which looked into the transcriptional effects of topotecan in autism. When we constructed a running average plot comparing the gene expression changes between topotecan drug-treated neurons (drug or D) and vehicle-treated cortical neurons (vehicle or V), we observed a preferential downregulation of long genes (the running average plot comparing drug vs. vehicle is indicated by the red curve in Figure 1A; Figure S1). To estimate the baseline variation among control samples, two random sets of vehicle-treated cultured cortical neurons were compared to each other (blue curve in Fig 1A, Experimental Procedures). Given that these untreated samples were obtained from littermates, we did not expect to observe any differences in gene expression and predicted that a running average plot comparing gene expression between vehicle-treated control samples would yield a horizontal line through y=0. However, we found that genes over 100kb in length tended to be down-regulated on average (blue curve in Figure 1A) when comparing gene expression levels between control samples. This effect was found for both RNA-Seq and microarray datasets (Figure 1A) and indicates that a portion of the length-dependent trend observed in the topotecan datasets is due to a length-dependent bias (i.e. noise) that can be observed even in the control samples.

To determine the significance of average fold-change trends, we applied a Student's t-test to each of the matching data bins from the drug vs. vehicle (D/V) and vehicle vs. vehicle (V/V) comparisons, followed by an adjustment for multiple hypothesis testing. For consistency, these plots are referred to as overlap plots (Experimental Procedures, Figure S1). At a false discovery rate of 0.05, only the long gene bins were statistically significant and showed preferential downregulation following topotecan treatment in both RNA-Seq and microarray datasets (lower panel in Fig 1A, red dots indicate statistically significant bins; Figure S1). In other words, although the control samples showed that long genes are downregulated at baseline (i.e. when comparing controls), topotecan treatment produced an even stronger

123 downregulation of long genes, providing sufficient signal to overcome the noise (or intra sample variation
124 between WT samples) observed in long genes at baseline. These datasets enabled us to establish a
125 statistical procedure as well as provided positive control for further analyses of long gene trends (King et
126 al., 2013; Mabb et al., 2016) in other studies.
127
128 **Gene length trends do not hold up in datasets for MeCP2 mouse models**
129 Studies of MeCP2-related disorders—both Rett syndrome (caused by loss-of-function mutations
130 in *MECP2*) and *MECP2* duplication syndrome (caused by duplication or even triplication of the locus)—
131 have provided a wealth of transcriptome data. Experiments in mouse models of both syndromes have
132 suggested that loss of MeCP2 function causes preferential upregulation of long genes and, conversely,
133 that gain of MeCP2 function leads to preferential downregulation of long genes (Gabel et al., 2015). We
134 chose to delve deeper into these datasets to explore the extent of the contribution of long genes to RTT
135 pathology. We applied our method to eleven MeCP2 datasets (Table 2) across seventeen different tissue
136 types (Baker et al., 2013; Ben-Shachar et al., 2009; Chahrour et al., 2008; Chen et al., 2015; Gabel et al.,
137 2015; Kishi et al., 2016; Samaco et al., 2012; Sugino et al., 2014; Zhao et al., 2013). We first computed
138 the running average plots and were able to reproduce the same results as reported previously (Gabel et al.,
139 2015; Sugino et al., 2014). However, when the baseline variation between wild-type (WT) samples is
140 plotted (blue curves in Fig. 1B), they extensively overlap with the running average plots from the *Mecp2*-
141 null (KO) samples (red curves in Fig. 1B; see also Figures S2A-K). This overlap between the curves for
142 the WT vs. WT comparisons and the KO vs. WT comparisons indicates that the signal originally reported
143 for the KO vs. WT comparison can be largely explained by noise (or inherent biological variation
144 between WT samples) in the dataset, as there is not a clear separation between the WT vs. WT curves and
145 the KO vs. WT curves in most brain regions surveyed. Further, we found that only a few long gene bins
146 showed significant preferential upregulation in *Mecp2*-null mice (FDR < 0.05) in these datasets.
147 However, we observed sporadic significant results for genes less than 100k (Figures 1B-1C; see also
148 Figures S2B-S2C, S2F-S2G and S2J). Likewise, we observed no preferential repression of long genes for
149 datasets from *Mecp2*-overexpression models (Tg) (Figure 1C; see also Figure S2L). Rather, we observed
150 more short genes to be preferentially dysregulated in the *Mecp2*-overexpression models (Figure 1C).
151 Thus, when assessing the bins of genes with significantly different expression between WT and KO mice,
152 we found that genes with a variety of lengths were altered in KO mice. Additionally, while there are
153 certainly some long genes with significantly altered expression in both KO and Tg mice, there is no
154 consistent and preferential long gene trend observed in the *Mecp2* datasets.
155
156 **Long gene trend is not present in Nuclear RNA profiles of MeCP2 mouse models**

157 A recent study reported that transcripts of long genes were downregulated in nuclear and nascent
158 RNA samples (Johnson et al., 2017) in contrast to previous studies (Gabel et al., 2015; Sugino et al.,
159 2014). The dataset was generated by combining an *in vivo* biotinylation system with Cre-loxP technology
160 that circumvented cellular heterogeneity of the brain and helped examine transcriptomic changes due to
161 MeCP2 in specific cell types, in both male and female mice (Johnson et al., 2017). The samples were
162 derived from the cortical cells of *Mecp2*-mutant mice bearing either of two common Rett-causing
163 mutations: T158M or R106W, which are among the most common mutations found in RTT patients
164 (Cuddapah et al., 2014).
165 We reanalyzed the data using overlap plots and observed no significant downregulation of long
166 genes in wildtype or *Mecp2*-mutant excitatory neurons from 18-week old T158M or R106W female mice
167 (Figures 1D and S4E). In excitatory neurons bearing the R106W mutation, we observed few bins that are
168 significantly different from WT expression levels. Notably, bins with significant gene expression changes
169 were not due to the downregulation of long genes in mutant samples. Rather, these bins were significant
170 due to the downregulation of long genes in control (WT) samples, as indicated by the downward slopes of
171 the running average plots comparing WT vs. WT samples (blue lines in Figures 1D and S4E). Similarly,
172 we observed no significant repression of long genes in nuclear RNA-Seq datasets of excitatory and
173 inhibitory neurons from 6-week old male mice with the same mutation type (Figures S4A-S4D). Finally,
174 when we examined downregulation of long genes from the GRO-Seq (global nuclear run-on with high-
175 throughput sequencing) data collected from these mice, we confirmed a marginal significance in the
176 downregulation of long genes (Figure S3A), but upregulation of long genes was not observed in whole
177 cell RNA-Seq data (Figure S3B). These results suggest that the transcriptome changes in long genes that
178 appear in RNA isolation-based methods are independent of the sex, age, or mutation type of the mouse.
179 Together, these results suggest that when the fold-change difference is 50% or more, as it is in the
180 topotecan datasets, there is likely to be a genuine long gene bias. When the fold-change effect is small
181 (<15%), however, as it is with the long genes observed in the *Mecp2* datasets, it is more likely that the
182 observed long gene trend is due to inherent variation among samples. The reported long gene trend in the
183 *Mecp2* datasets is in the same range as the noise that we derived from the intra-sample comparison in the
184 control groups, and this effect was seen in all the *Mecp2* datasets that we assessed. This further suggests
185 that the length-dependent variability estimated from microarray and RNA-Seq platforms is not sensitive
186 enough to capture small transcriptional changes. We, therefore, recommend that baseline gene length
187 dependency should be evaluated from the control group first to understand the statistical significance of
188 observed long gene trends in any sequencing dataset.
189
190 **Human MeCP2 datasets: the importance of age**

191 To determine whether preferential dysregulation of long genes occurs in *in vitro* human Rett
192 datasets, we computed overlap plots on samples from isogenic human iPSCs (hiPSCs), neural progenitor
193 cells (NPCs), and neurons from the fibroblasts of two independent patients, with and without the *MECP2*
194 mutation. We found no preferential upregulation of long genes (Figures 2A) but did see a trend toward
195 downregulation of long genes among human *in vitro* RTT neuron samples, which is contrary to reports
196 from *Mecp2*-null mouse models (Gabel et al., 2015; Sugino et al., 2014).
197 Although long genes do not appear to be upregulated above the level of background noise in
198 murine *Mecp2* datasets, they have been reported to be preferentially upregulated in human RTT samples
199 (Gabel et al., 2015) as well, and we wondered if a more robust signal would be observed in post-mortem
200 human datasets. Three RTT and three normal control samples from the superior frontal gyrus were
201 obtained from a previous study (Deng et al., 2007). These samples were from three different ages: RTT
202 samples were obtained from donors aged 8, 6, and <4 years (pooled samples from a 2- and a 4-year old),
203 with approximately age-matched normal control samples obtained from donors aged 10, 5 and 2 years,
204 respectively. The long gene trend was observed (Gabel et al., 2015) in a comparison of the three RTT
205 samples to the three control samples (Figures 2B). Because stages of brain development and disease
206 progression in RTT patients change markedly from ages 1 to 5 years before stabilizing (Chahrour and
207 Zoghbi, 2007), we reanalyzed the data by comparing each sample to its age-matched control separately.
208 Dysregulation of long genes was observed only in the 2- and 4-year old RTT samples (Figure 2B left
209 panel), but not in either the 5- or 8-year old RTT samples (Figure 2B right panel). Unfortunately, the
210 statistical significance of this observation cannot be established because of the small sample size (n = 1
211 each).
212 To determine whether length-dependent misregulation of long genes occurs in other human
213 datasets, we analyzed samples from another study (Lin et al., 2016) and in-house generated RNA-Seq
214 RTT datasets. Lin et al. dataset (Lin et al., 2016) consist of postmortem brain samples from the frontal
215 and temporal cortex of RTT patients with age-matched controls (age = 17-20 years, n = 3 each). Because
216 the phenotypes are similar for RTT patients in this age range (Chahrour and Zoghbi, 2007), we grouped
217 these RTT samples together and compared them to the pooled age-matched controls. We computed
218 running average plots on the normalized dataset (Experimental Procedures, Figure S1) and did not
219 observe overrepresentation of long genes in these samples (Figure 2C). Similar results were reported by
220 the original study (Lin et al., 2016). Consistent with our previous results, there was no long gene trend in
221 the running average plot of the RNA-Seq RTT dataset collected from a postmortem frontal cortex sample
222 obtained from an 18-year-old RTT female (Figure 2D, left panel) when it was compared to its age-
223 matched control (age = 18 years, n = 1 each). To further probe whether the long-gene trend might be
224 present in the early stages of the disease, we compared a RTT postmortem male sample from frontal

225 cortex (age = 1 year, n = 1) to an age-matched control sample (age = 2 days, n = 1) and again could find
226 no significant upregulation of long genes (Figure 2D, right panel).
227 One possible explanation for the lack of a long gene trend in human RTT samples is
228 heterogeneity among the various samples (including differences in the genetic background), which
229 increases the inherent variability in gene expression among biological replicates. Such variability could
230 obscure the effects of a subtle bias in the sequencing process. Nevertheless, the present findings suggest
231 that long genes are not preferentially misregulated in human RTT datasets.
232
233 **Differential gene expression analysis for Topotecan and Mecp2 datasets**
234 Our previous analysis suggests that the current transcriptome profiling technologies are limited in
235 their ability to detect subtle differences in gene expression. We hypothesize that long gene effects, if
236 genuine, should be apparent in both binning analysis and the traditional differential gene expression
237 analysis. We, therefore, decided to focus our attention on only the differentially expressed genes that were
238 reported by previous studies (Baker et al., 2013; Ben-Shachar et al., 2009; Chahrour et al., 2008; Chen et
239 al., 2015; Huang et al., 2011; King et al., 2013; Mabb et al., 2016). We divided the entire list of
240 differentially expressed genes into four groups based on gene length (> or < 100kb) and fold-change
241 direction (either up or down). Consistent with our overlap plots, we found long genes to be substantially
242 overrepresented and downregulated in Topotecan datasets (Figure 3A). This result proves that our
243 approach does detect long gene trends in gene expression studies. In the MeCP2 datasets, however, we
244 did not find a preferential upregulation of long genes (Figures 3B-3D) except in the hippocampal dataset
245 (Figure S5) (Baker et al., 2013). Moreover, in contrast to previous studies, we found that more long genes
246 were upregulated than downregulated in the cerebellum of *Mecp2* over-expressing mice (Figure 3C, right
247 panel). Another important difference between the Topotecan and *Mecp2* datasets was that short genes
248 dominated among all differentially expressed genes in *Mecp2* datasets (Figures 3B-3D; Figures S5). This
249 further supports the notion that a preference for long gene misregulation is not an inherent feature of gene
250 expression following the *Mecp2* disruption. This is not to say that MeCP2 does not regulate a subset of
251 long genes, only that our analysis found no preferential misregulation of long gene trend in MeCP2
252 mouse models.
253
254 **RNA-Seq and microarray benchmark datasets are prone to length-dependent bias**
255 To investigate whether length dependent bias might be a function of amplification-based
256 platforms, we next performed running average analysis on the samples from the phase-III
257 Sequencing/Microarray Quality Control (SEQC) project (Consortium, 2014). SEQC was designed to
258 evaluate the performance of various sequencing platforms, sources of bias in gene expression samples,

259 and various methods for downstream analysis. The consortium generated benchmark datasets using four
260 different types of RNA samples: A (Universal Human Reference RNA), B (Human Brain Reference
261 RNA), C (a mixture of A and B at a ratio of 3:1), and D (a mixture of A and B at a defined ratio of 1:3).
262 The RNA-Seq datasets generated using the Illumina HiSeq 2000 platform across six different sites were
263 used for quality control analyses (Experimental Procedures), and the raw read counts were normalized
264 using the DESeq2 method (Love et al., 2014).

265 To determine whether the dataset showed nominal batch effects or other non-biological
266 variability, we used multidimensional scaling (MDS) plots to see if the samples clustered according to
267 RNA sample type. To ascertain whether or not the samples were consistently titrated, we calculated the β
268 ratio of observed gene expression in the samples, which is obtained from the following equation: ((B-
269 A)/(C-A)) (Consortium, 2014). The value of the β ratio (Shippy et al., 2006) is 4:1 (or $\log_2(4) = 2$). In
270 theory, the β ratio should be independent of gene length in the brain and non-brain tissues. After assessing
271 various SEQC datasets, we found that the Novartis dataset had nominal batch effects and the β ratio was
272 close to 2. Therefore, this dataset would be ideal, as it would not bias downstream analyses (Figure 4A).

273 We separated Human Brain Reference (sample type B) RNA-Seq samples into two groups of 32
274 samples each, based on their y-axis coordinates of the MDS plot, and computed a running average plot.
275 Since these samples were technical replicates of the same reference RNA sample type, we expected the
276 mean $\log_2$ fold-change to be a horizontal line along the x-axis with a y-intercept equal to zero (i.e., y=0 on
277 an xy plane). Instead, we found that long genes deviated from the expected pattern, with the fold-changes
278 of long genes being overestimated (Figure S6A, left panel).

279 We then investigated whether the fold-change of long genes is constant for the β ratio samples.
280 The expected average $\log_2$ fold-change should be a horizontal line along the x-axis with a y-intercept
281 equal to two. We found, however, that the expected ratio was not maintained for long genes and was
282 overestimated (Figure 4B). Moreover, we observed a similar bias in the β ratio with respect to transcript
283 length, with longer transcripts being overrepresented (Figure S6A, right panel). Overall, the range of
284 overestimation in the RNA-Seq dataset was between 3% and 40%. Consistent with our findings, another
285 study (using a different dataset) previously reported that long genes were more likely to be identified as
286 statistically significant in RNA-Seq datasets (Oshlack and Wakefield, 2009).

287 To determine whether the long gene bias was unique to the RNA-Seq datasets or could be
288 detected on other platforms, we investigated the MAQC-III microarray Affymetrix dataset generated by
289 the SEQC consortium (Consortium, 2014). Human Brain Reference samples (B) were separated into two
290 groups based on y-axis location on the MDS plot (Figure 4C). The running average plots were computed
291 against their average gene length using the same parameters as described for the RNA-Seq analysis
292 above. As with the RNA-Seq samples, the average fold-change for long genes deviated from the expected

293 value of zero (Figure S6B, left panel). When the β ratio was plotted against the mean gene length (Figure
294 4D) or mean transcript length (Figure S6B right panel), we found that long genes were overrepresented.
295 Further, long gene bias was observed in both RNA-Seq and microarray datasets in a comparison of two
296 groups of universal human reference (Figure S6A-S6B, middle panel). The overestimation in the
297 microarray dataset ranged from 1.5% to 23%—lower overall than for the RNA-Seq dataset, but indicating
298 that microarray datasets are also predisposed to gene and transcript length-dependent bias.
299
300 **Long gene bias is independent of normalization methods**
301 To ensure that the long gene bias we observed was not due to our normalization methods, we
302 compared the mean $\log_2$ fold-change using three different normalization techniques: Total Count, DESeq
303 (Anders and Huber, 2010), and edgeR/TMM (Robinson et al., 2010; Robinson and Oshlack, 2010). We
304 normalized the raw read counts from four different RNA sample types using each of the three
305 normalization methods and computed running average plots of the β ratios against gene and transcript
306 length. In all cases, long genes were still overestimated, regardless of the normalization method (Figures
307 S7A-S7B). This lends support to the notion that the overrepresentation of long genes is independent of the
308 normalization technique.
309
310 **Long gene bias is not observed in NanoString datasets, which are not based on amplification**
311 We hypothesized that PCR amplification, a process shared by both microarray and RNA-Seq
312 technologies, might introduce the observed bias in long gene expression. We, therefore, performed
313 NanoString nCounter gene expression quantification, a technique that does not use amplification, with the
314 SEQC reference RNA samples (A, B, C, and D) (n = 6 each). The MDS plot on normalized data showed
315 that the samples clustered based on sample type (Figure S8A); the effect of batches was minimal
316 (Experimental Procedures). The code set consisted of ~ 184 long genes, out of which ~132 long genes
317 were expressed in brain samples (Figure S8B). We again computed the running average plots against their
318 average gene length, and we did not observe any long gene bias between the brain samples or when
319 computing the β ratio of the samples (Figures S8C- S8D).
320 We next compared the mean expression levels of all the common genes across the RNA-Seq,
321 microarray and nCounter datasets. Our analysis shows that fold-changes of long genes are overestimated
322 in the RNA-Seq (P-value < 2.7 e-07; Figure 4E) and microarray datasets (P-value < 0.021; Figure 4F); in
323 contrast, the nCounter dataset showed no difference in the average expression of long and short genes (P-
324 value = 0.86; Figure 4G). Although it is possible that the smaller number of genes (~680) might make it
325 more difficult to detect a preference, the proportion of long genes in this dataset (~180 out of ~680 genes,
326 or 26%) is twice that found in the human transcriptome (~3200 long genes out of ~ 24,000 genes, or

327 13%). Any preference for long genes should thus be revealed even more strongly in this dataset. These
328 results lead us to posit that the long gene overestimation we observed in RNA-Seq and microarray
329 datasets might be caused by a length-dependent bias in PCR amplification.
330
331 **PCA plot confirms the reciprocal relationship of Mecp2 gain- and loss-of-function datasets**
332 One of the most intriguing components of the long gene story in RTT is the presence of a
333 reciprocal pattern in the *Mecp2*-overexpression model, where a reported preference for downregulation of
334 long genes complements the upregulation of long genes reported in *Mecp2*-null mice (Gabel et al., 2015).
335 To understand this reciprocal relationship, we divided Human Brain Reference samples (B) into 3 groups
336 (n = 16 each) based on different library preparation ID numbers from the Novartis SEQC dataset. The
337 PCA plot clearly clustered the brain samples based on the library preparation group to which they
338 belonged (Figure S9A). Comparing the brain samples of library preparation ID 2 (green) to library
339 preparation ID 1 (red) and ID 3 (blue) separately reversed the running average plot (Figures S9B-S9C).
340 These results show that a reciprocal relationship can be observed in the gene expression data between any
341 groups that form three distinct clusters on a PCA plot.
342 We next assessed the influence of the fold-change threshold on differential expression analysis
343 using brain samples. Although we did not expect to see a trend between replicates, preferential regulation
344 of long genes was observed (Figure S9D) when the fold-change was small (<10%, or log2FC ~ 13%).
345 The bias was similar to the trend observed in previously published *Mecp2*-null and overexpression (Tg)
346 models when library preparations ID 2 (red) and ID 1 (green), or library preparations ID 3 (blue) and ID 1
347 (green), were compared (Gabel et al., 2015; Sugino et al., 2014).
348 In this analysis, all the samples were technical replicates of the same reference RNA and were
349 expected to have identical gene expression levels, but variation associated with library preparation
350 resulted in the samples not clustering together and allowed us to observe an inverse trend in long genes
351 (Figure S9A). Just as biological variation can lead to separation on a PCA plot, so can technical variation,
352 and both can result in the same apparent long gene bias observed in *Mecp2* datasets. Furthermore, our
353 analysis suggests that differentially expressed genes can be highly variable with small fold-changes,
354 which underscores the importance of proper fold-change cut-offs in differential gene expression analysis.
355
356 **Differentially expressed genes with small fold-changes identified by RNA-Seq are not reproducible**
357 **by NanoString in the *Mecp2* dataset**
358 To determine whether a long gene trend is present only in the *Mecp2* RNA-Seq dataset and not in
359 the NanoString dataset, we generated RNA-Seq (> 90 million paired-end sequencing reads per sample; n
360 = 3 each; Table 3) and NanoString (n = 3 each; Table 4) datasets on cerebellar tissue from wild-type and

361 *Mecp2*-null mouse models (KO). The PCA plot on normalized datasets (Experimental Procedures)
362 showed that the samples clustered based on sample type (Figures S10A-S10B, left panel). Transcriptome
363 analysis was performed using DESeq2 (Love et al., 2014) on both datasets. We first analyzed RNA-Seq
364 data to estimate the strength of the long gene trend. Although there appeared to be a long gene trend in the
365 KO/WT comparison, an overlap plot confirmed there was no significant upregulation of long genes
366 (Figure S10A, middle panel). Consistent with our previous findings, there was no preferential
367 upregulation of long genes in our differential expression analysis (Figure S10A, right panel; absolute
368 $\log_2 FC > 1.2$ & $FDR < 0.05$).
369 We performed further analysis using a list of 750 (~159 long and ~591 short) genes common to
370 both RNA-Seq and nCounter NanoString (Experimental Procedures). Comparison of the log fold-changes
371 using the classic method (i.e., log2((mean(group1) + 1)/(mean(group2) + 1)) and using shrunken log fold-
372 changes by DESeq2 (i.e., obtaining reliable variance estimates by pooling information across all the
373 genes) suggested that the latter method yields more highly correlated fold-changes (Figures S10C). This
374 is consistent with previous findings showing that shrunken log fold-changes are more reproducible (Love
375 et al., 2014; Robinson et al., 2010). Even with this method, however, we observed high variability among
376 genes with low fold-changes between the two datasets, regardless of whether they were long or short
377 (Figures 5A and 5B). Moreover, genes with high fold-changes in expression (~ FC > 20%) were
378 consistently called as differentially expressed in both the datasets (Figures 5A and 5B).
379 This analysis suggests that the genes identified as differentially expressed by RNA-Seq at lower
380 fold changes are not reproducible by NanoString. To determine whether fold-changes are inflated in
381 RNA-Seq, we compared the absolute difference of $\log_2$ fold-change between the RNA-Seq and
382 NanoString datasets. We observed fold-changes of long genes to be overestimated by RNA-Seq
383 technology (Figure 5C; Chi-Square test; p-value <7.44e-3), which further supports our hypothesis that
384 artefactual long gene trends are more likely to appear in amplification-based expression datasets.
385
386 **DISCUSSION**
387 Several recent papers have suggested that diseases associated with synaptic dysfunction tend to
388 preferentially involve misregulation of long genes (>100 Kb) (Gabel et al., 2015; King et al., 2013;
389 Sugino et al., 2014; Zylka et al., 2015). To establish a statistical baseline for the length-dependent gene
390 regulation analysis, we took advantage of a large number of SEQC consortium datasets where the relative
391 gene expression fold-change has been measured using RNA-Seq and microarray. We demonstrated the
392 power of big data analysis by uncovering major sources of technical variation such as intra-sample
393 variation and PCR amplification bias that can affect the analysis of long gene expression. By contrast,
394 NanoString nCounter technology, which does not rely on amplification, revealed no long gene bias. Our

395 results demonstrate that amplification-based transcriptomic technologies can lead to overestimations of
396 long gene expression changes.
397 This is not to say that there is never a bias toward expression changes in long genes. The
398 topotecan dataset showed an authentic long gene trend even after accounting for baseline variability. This
399 sizeable effect on long gene expression is consistent with the biological function of topotecan inhibiting
400 topoisomerase I; long genes should, in theory, be more dependent on proper unwinding during
401 transcription elongation (King et al., 2013). By contrast, we found no bias toward long gene
402 dysregulation in the MeCP2 datasets after baseline correction, even when we focused on only those genes
403 that are differentially expressed to a statistically significant degree. The sole exception was the one
404 infantile RTT case, but a single case does not allow us to draw any firm conclusions. Again, this does not
405 rule out that MeCP2 regulates some long genes; it simply does not support a preferential misregulation of
406 long genes by mutant MeCP2.
407 Apparent expression changes in long genes are clearly liable to exaggeration by biases in
408 microarray and RNA-Seq. We recommend eliminating confounds such as batch effects and properly
409 estimating both inter- and intra-sample variations; the control datasets must be carefully analyzed in order
410 to reveal the degree of baseline variability, which then can inform further analyses of the size of the signal
411 required to overcome background noise in sequencing datasets (Figure S1). These findings are applicable
412 to all research that utilizes current microarray and sequencing technologies. We hope that revealing the
413 influence of protocols and technologies on RNA sequencing data will lead to improved technologies and
414 more reliable analyses for amplification-based sequencing data.
415
416 **AUTHOR CONTRIBUTIONS**
417 Conceptualization: ZL, HYZ, ATR, AEP; Methodology and Investigation: ATR, ZL; Software,
418 Validation and Analysis: ATR, YWW, AEP; Data Curation: ATR, BL, AEP, YWW, HKY; Writing –
419 Original Draft: ATR, ZL, AEP; Writing – Review & Editing: ATR, YWW, AEP, ZL; Visualization:
420 ATR, YWW; Supervision: ZL, HYZ; Funding Acquisition: ZL, HYZ.
421
422 **ACKNOWLEDGMENTS**
423 We thank Laura Lavery, Rami Al-Ouran, Laura Lombardi, Ezequiel Sztainberg, Aya Ishida, and Vicky
424 Brandt for helpful discussions and suggestions. This project was supported by the Genomic and RNA
425 Profiling Core at Baylor College of Medicine and the expert assistance of the Core Director, Lisa D.
426 White, Ph.D.
427
428 **ACCESSION NUMBERS**

429 The GEO accession numbers for NanoString and RNA-seq datasets reported in this paper are as
430 follows: GSE94073, GSE105047 (includes GSE105045 and GSE105046) and GSE107399.
431

432 **REFERENCES**

433 Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome
434 biology 11, 1.

435 Baker, S.A., Chen, L., Wilkins, A.D., Yu, P., Lichtarge, O., and Zoghbi, H.Y. (2013). An AT-hook domain
436 in MeCP2 determines the clinical course of Rett syndrome and related disorders. Cell 152, 984-996.

437 Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A.,
438 Phillippy, K.H., Sherman, P.M., and Holko, M. (2013). NCBI GEO: archive for functional genomics data
439 sets—update. Nucleic acids research 41, D991-D995.

440 Ben-Shachar, S., Chahrour, M., Thaller, C., Shaw, C.A., and Zoghbi, H.Y. (2009). Mouse models of
441 MeCP2 disorders share gene expression changes in the cerebellum and hypothalamus. Human molecular
442 genetics 18, 2431-2442.

443 Chahrour, M., Jung, S.Y., Shaw, C., Zhou, X., Wong, S.T., Qin, J., and Zoghbi, H.Y. (2008). MeCP2, a
444 key contributor to neurological disease, activates and represses transcription. Science 320, 1224-1229.

445 Chahrour, M., and Zoghbi, H.Y. (2007). The story of Rett syndrome: from clinic to neurobiology. Neuron
446 56, 422-437.

447 Chen, L., Chen, K., Lavery, L.A., Baker, S.A., Shaw, C.A., Li, W., and Zoghbi, H.Y. (2015). MeCP2
448 binds to non-CG methylated DNA as neurons mature, influencing transcription and the timing of onset for
449 Rett syndrome. Proceedings of the National Academy of Sciences 112, 5509-5514.

450 Consortium, S.M.-I. (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and
451 information content by the Sequencing Quality Control Consortium. 32, 903-914.

452 Cuddapah, V.A., Pillai, R.B., Shekar, K.V., Lane, J.B., Motil, K.J., Skinner, S.A., Tarquinio, D.C., Glaze,
453 D.G., McGwin, G., Kaufmann, W.E.*, et al.* (2014). Methyl-CpG-binding protein 2 (MECP2) mutation
454 type is associated with disease severity in Rett syndrome. J Med Genet 51, 152-158.

455 Deng, V., Matagne, V., Banine, F., Frerking, M., Ohliger, P., Budden, S., Pevsner, J., Dissen, G.A.,
456 Sherman, L.S., and Ojeda, S.R. (2007). FXYD1 is an MeCP2 target gene overexpressed in the brains of
457 Rett syndrome patients and Mecp2-null mice. Human molecular genetics 16, 640-650.

458 Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot,
459 G., Castel, D., and Estelle, J. (2013). A comprehensive evaluation of normalization methods for Illumina
460 high-throughput RNA sequencing data analysis. Briefings in bioinformatics 14, 671-683.

461 Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and
462 Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15-21.

463 Eppig, J.T., Blake, J.A., Bult, C.J., Kadin, J.A., Richardson, J.E., and Group, M.G.D. (2015). The Mouse
464 Genome Database (MGD): facilitating mouse as a model for human biology and disease. Nucleic acids
465 research 43, D726-D736.

466 Gabel, H.W., Kinde, B., Stroud, H., Gilbert, C.S., Harmin, D.A., Kastan, N.R., Hemberg, M., Ebert, D.H.,
467 and Greenberg, M.E. (2015). Disruption of DNA-methylation-dependent long gene repression in Rett
468 syndrome. Nature 522, 89-93.

469 Gautier, L., Cope, L., Bolstad, B.M., and Irizarry, R.A. (2004). affy—analysis of Affymetrix GeneChip
470 data at the probe level. Bioinformatics 20, 307-315.

471 Hansen, K.D., Brenner, S.E., and Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused
472 by random hexamer priming. Nucleic acids research 38, e131-e131.

473 Huang, H.-S., Allen, J.A., Mabb, A.M., King, I.F., Miriyala, J., Taylor-Blake, B., Sciaky, N., Dutton, J.W.,
474 Lee, H.-M., Chen, X.*, et al.* (2011). Topoisomerase inhibitors unsilence the dormant allele of Ube3a in
475 neurons. 481, 185-189.

476 Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P.
477 (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data.
478 Biostatistics 4, 249-264.

479 Johnson, B.S., Zhao, Y.T., Fasolino, M., Lamonica, J.M., Kim, Y.J., Georgakilas, G., Wood, K.H., Bu, D.,
480 Cui, Y., Goffin, D.*, et al.* (2017). Biotin tagging of MeCP2 in mice reveals contextual insights into the
481 Rett syndrome transcriptome. Nat Med.

482 Katz, D.M., Bird, A., Coenraads, M., Gray, S.J., Menon, D.U., Philpot, B.D., and Tarquinio, D.C. (2016).
483 Rett Syndrome: Crossing the Threshold to Clinical Translation. 39, 100-113.

484 Kinde, B., Wu, D.Y., Greenberg, M.E., and Gabel, H.W. (2016). DNA methylation in the gene body
485 influences MeCP2-mediated gene repression. Proceedings of the National Academy of Sciences of the
486 United States of America.

487 King, I.F., Yandava, C.N., Mabb, A.M., Hsiao, J.S., Huang, H.-S., Pearson, B.L., Calabrese, J.M.,
488 Starmer, J., Parker, J.S., and Magnuson, T. (2013). Topoisomerases facilitate transcription of long genes
489 linked to autism. Nature 501, 58-62.

490 Kishi, N., MacDonald, J.L., Ye, J., Molyneaux, B.J., Azim, E., and Macklis, J.D. (2016). Reduction of
491 aberrant NF-κB signalling ameliorates Rett syndrome phenotypes in Mecp2-null mice. Nat Commun 7,
492 10520.

493 Lahens, N.F., Kavakli, I.H., Zhang, R., Hayer, K., Black, M.B., Dueck, H., Pizarro, A., Kim, J., Irizarry,
494 R., and Thomas, R.S. (2014). IVT-seq reveals extreme bias in RNA sequencing. Genome Biol 15, 1.

495 Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey,
496 V.J. (2013). Software for computing and annotating genomic ranges. PLoS computational biology 9,
497 e1003118.

498 Li, J., Jiang, H., and Wong, W.H. (2010). Modeling non-uniformity in short-read rates in RNA-Seq data.
499 Genome Biol 11, 1.

500 Lin, P., Nicholls, L., Assareh, H., Fang, Z., Amos, T.G., Edwards, R.J., Assareh, A.A., and Voineagu, I.
501 (2016). Transcriptome analysis of human brain tissue identifies reduced expression of complement
502 complex C1Q Genes in Rett syndrome. BMC Genomics 17, 427.

503 Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for
504 RNA-seq data with DESeq2. Genome biology 15, 1.

505 Mabb, A.M., Simon, J.M., King, I.F., Lee, H.M., An, L.K., Philpot, B.D., and Zylka, M.J. (2016).
506 Topoisomerase 1 Regulates Gene Expression in Neurons through Cleavage Complex-Dependent and -
507 Independent Mechanisms. PLoS One 11, e0156439.

508 Oshlack, A., and Wakefield, M.J. (2009). Transcript length bias in RNA-seq data confounds systems
509 biology. 4, 14.

510 Ouwenga, R.L., and Dougherty, J. (2015). Fmrp targets or not: long, highly brain-expressed genes tend to
511 be implicated in autism and brain disorders. 6, 16.

512 Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). GC-Content Normalization for RNA-Seq
513 Data. 12, 480.

514 Robert, C., and Watson, M. (2015). Errors in RNA-Seq quantification affect genes of relevance to human
515 disease. 16.

516 Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for
517 differential expression analysis of digital gene expression data. Bioinformatics 26, 139-140.

518 Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression
519 analysis of RNA-seq data. Genome Biol 11, 1.

520 Samaco, R.C., Mandel-Brehm, C., McGraw, C.M., Shaw, C.A., McGill, B.E., and Zoghbi, H.Y. (2012).
521 Crh and Oprm1 mediate anxiety-related behavior and social approach in a mouse model of MECP2
522 duplication syndrome. 44, 206-211.

523 Shippy, R., Fulmer-Smentek, S., Jensen, R.V., Jones, W.D., Wolber, P.K., Johnson, C.D., Pine, P.S.,
524 Boysen, C., Guo, X., Chudin, E*., et al.* (2006). Using RNA sample titrations to assess microarray platform
525 performance and normalization techniques. Nature biotechnology 24, 1123-1131.

526 Sugino, K., Hempel, C.M., Okaty, B.W., Arnson, H.A., Kato, S., Dani, V.S., and Nelson, S.B. (2014).
527 Cell-Type-Specific Repression by Methyl-CpG-Binding Protein 2 Is Biased toward Long Genes. 34,
528 12877-12883.

529 Sullivan, J.M., Badimon, A., Schaefer, U., Ayata, P., Gray, J., Chung, C.-w., von Schimmelmann, M.,
530 Zhang, F., Garton, N., Smithers, N*., et al.* (2015). Autism-like syndrome is induced by pharmacological
531 suppression of BET proteins in young mice. 212, 1771-1781.

532 Waggott, D., Chu, K., Yin, S., Wouters, B.G., Liu, F.-F., and Boutros, P.C. (2012). NanoStringNorm: an
533 extensible R package for the pre-processing of NanoString mRNA and miRNA data. Bioinformatics 28,
534 1546-1548.

535 Wan, Y.-W., Mach, C.M., Allen, G.I., Anderson, M.L., and Liu, Z. (2014). On the reproducibility of
536 TCGA ovarian cancer microRNA profiles. PloS one 9, e87782.

537 Zhao, Y.-T., Goffin, D., Johnson, B.S., and Zhou, Z. (2013). Loss of MeCP2 function is associated with
538 distinct gene expression changes in the striatum. 59, 257-266.

539 Zoghbi, H.Y., and Bear, M.F. (2012). Synaptic dysfunction in neurodevelopmental disorders associated
540 with autism and intellectual disabilities. Cold Spring Harbor perspectives in biology 4, a009886.

541 Zylka, M.J., Simon, J.M., and Philpot, B.D. (2015). Gene Length Matters in Neurons. 86, 353-355.

542

543 **FIGURES and TABLES**

544

545 **Figure 1. Establishment of baselines and comparison of Mecp2 microarray and RNA-Seq datasets.**
546 **A)** *Topotecan datasets:* The top half of each subgraph shows the comparison of cultured cortical neurons
547 treated with vehicle (V) from C57BL/6J (B6) × CASTEi/J (CAST) F1 hybrid mice with other vehicle-
548 treated samples (V/V, blue line) and comparison of topotecan-treated cortical neurons (D) with vehicle-
549 treated samples (D/V, red line). The red and blue lines diverge only for genes over 100kb in size. (**B-D)**
550 *Mecp2 datasets:* Note the change in the scale of y-axis; these changes are much smaller than in the
551 topotecan studies. Unlike the topotecan results in row A, gene bins with statistical significance are
552 sporadic for both long and short genes in row B. The top half of each subgraph in row **B)** shows the
553 comparison of WT male C57BL samples with other WT male C57BL samples (blue line) and Mecp2

554 male KO samples compared with WT male littermates (red line) in the amygdala (Samaco et al., 2012),
555 cerebellum (Ben-Shachar et al., 2009) and hypothalamus (Chahrour et al., 2008). **C)** Comparison of three
556 different Mecp2 Tg/WT male mouse models. The top half of each subgraph shows the comparison
557 between WT FVB samples and other WT FVB samples (blue line) within the same genotype and Tg
558 samples with their WT littermates (red line) in amygdala (Samaco et al., 2012), cerebellum (Ben-Shachar
559 et al., 2009) and hypothalamus (Chahrour et al., 2008). Note that we observe few long gene bins as well
560 as short gene bins with significant preferential upregulation in *Mecp2*-null and *Mecp2*-overexpression
561 (Tg) mice datasets. **D)** Cortical excitatory neurons from three different Mecp2 KO/WT female mouse
562 models. The top half of each subgraph shows the comparison between two sets of WT C57BL samples
563 (blue line), and between WT littermates and mutant mice bearing either the R106W or T158M mutations
564 (Johnson et al., 2017). Note that the magnitude of length dependent gene misregulation was more
565 substantial in control samples rather than *Mecp2-mutant* samples (blue curve). The blue or red line
566 represents fold-change in expression for genes binned according to gene length (bin size of 200 genes
567 with shift size of 40 genes) as described in (Gabel et al., 2015). The blue and red shaded areas correspond
568 to one-half of one standard deviation of each bin for the comparison of WT/WT and KO/WT (or
569 MUT/WT) or Tg/WT, respectively. The bottom half of each subgraph is the p-value from the two-sample
570 t-test between KO/WT (or MUT/WT) or Tg/WT and WT/WT. Bins with FDR < 0.05 are shown as a red
571 dot. The red dashed line at the bottom of the subgraphs indicates the minimum $-Log_{10}$(p-value) that
572 corresponds to a FDR < 0.05. Please refer to Table 1 for the total number of samples used for the
573 comparison between two random sets of WT (or vehicle-treated) samples and between WT littermates
574 and KO/Tg/mutant mice.
575
576 **Figure 2. No bias toward long genes in *MECP2* human datasets**. **(A)** RNA-Seq analysis of isogenic
577 human Rett *in vitro* models. Overlap plots were used to compare WT and KO samples, where the top half
578 of each subgraph shows the comparison of WT samples with other WT samples (blue line), and RTT
579 samples compared with WT samples (red line) in iPSC (left panel), Neural progenitor cells or NPC
580 (middle panel), and neurons (right panel). **(B)** Microarray analysis of human RTT brain samples
581 compared to age-matched control for Frontal Cortex (Deng et al., 2007). Comparison of gene trends in the
582 pooled sample from 2- and 4-year old patients (left panel) and whole dataset (left panel). Observed long
583 gene trend in the sample from 5-year old (right panel) and 8-year old patients (right panel). **(C)**
584 Microarray analysis of RTT human frontal cortex samples (Lin et al., 2016) compared to controls (left
585 panel) and RTT human temporal cortex samples (Lin et al., 2016) compared to controls (right panel). **(D)**
586 RNA-Seq analysis of RTT human (female) frontal cortex samples compared to controls (left panel) and
587 RTT human (male) frontal lobe samples compared to controls (right panel). The lines in A-D represent

588 fold-change in expression for genes binned according to gene length (bin size of 200 genes with shift size
589 of 40 genes) as described in Gabel, Kinde et al. *Nature* 2015. The blue and red ribbons in **(A)** correspond
590 to one-half of one standard deviation of each bin for the comparison of WT/WT and MUT/WT
591 respectively. The bottom half of each subgraph is the p-value from the two-sample t-test between
592 MUT/WT and WT/WT. Bins with FDR < 0.05 are shown as a red dot. The red dotted line in the bottom
593 of the subgraphs indicates the minimum -Log_{10}(p-value) that corresponds to a FDR < 0.05. Please refer to
594 Table 1 for the total number of samples used for the comparison between two random sets of WT samples
595 and between WT and RTT samples.
596
597 **Figure 3. Differentially expressed genes show length-dependent misregulation in Topotecan**
598 **datasets but not in Mecp2 studies**. **(A)** Scatter plot of log fold-change in expression between topotecan
599 and vehicle-treated cultured cortical neurons (y-axis) against its gene length (x-axis) in RNA-Seq dataset
600 from (King et al., 2013) (left panel; n = 5 each; FDR < 0.05) and (Mabb et al., 2016) RNA-Seq dataset
601 (right panel; n = 3 each; FDR < 0.01). **(B)** Scatter plot of log fold-change in expression (microarray)
602 between C57BL KO and its C57BL WT littermates (y-axis) against its gene length (x-axis) in
603 hypothalamus (left panel; n = 4 each; FDR < 0.05 and log2FC > 0.2; (Chahrour et al., 2008)) and
604 cerebellum (right panel; n = 4 each; FDR < 0.05 and log2FC > 0.2; (Ben-Shachar et al., 2009)). **(C)**
605 Scatter plot of log fold-change in expression (microarray) between FVB Tg to its FVB WT littermates (y-
606 axis) against its gene length (x-axis) in hypothalamus (n = 4 each; FDR < 0.05 and log2FC > 0.2;
607 (Chahrour et al., 2008)) and cerebellum (n = 4 each; FDR < 0.05 and log2FC > 0.2; (Ben-Shachar et al.,
608 2009)). **(D)** Scatter-plot of log fold-change in expression between KO/Tg and WT littermates (y-axis)
609 against gene length (x-axis) in RNA-Seq datasets: Hypothalamus KO/WT comparison (left panel; n = 3
610 each; FDR < 1e-5; (Chen et al., 2015)) and Hypothalamus Tg/WT comparison (right panel; n = 3 each;
611 FDR < 1e-5; (Chen et al., 2015)). Red dot represents long genes and blue dot represents short genes.
612 Differentially expressed genes were obtained from the published gene lists.
613
614 **Figure 4. Long gene bias in SEQC RNA-Seq and microarray, but not NanoString, datasets. (A)**
615 MDS plot using Euclidean distance on the SEQC (Consortium, 2014) NVS count dataset. **(B)** Mean Log2
616 Fold Change plot against gene length using β ratio samples ((B-A/C-A); n =64 each) in RNA-Seq dataset.
617 **(C)** MDS plot using Euclidean distance on the SEQC microarray dataset. **(D)** Mean Log2 Fold-Change
618 plot against gene length using β ratio samples in microarray dataset (n = 4 each). Each blue dot is a bin of
619 200 genes with shift size of 40 genes (Gabel et al., 2015). Box plot of the genes across three different
620 platforms that are present in NanoString codeset. The distributions of the mean fold-changes for β ratio

621 samples for long and short genes are compared across three different platforms: **E)** RNA-Seq, **F)**
622 Microarray, and **G)** NanoString. P-values were computed using the Wilcoxon Mann Whitney test.
623
624 **Figure 5. Expression changes are overestimated in RNA-Seq datasets.** Comparison of log fold-change
625 in expression between RNA-Seq and Nanostring for Short Genes (**A**) and Long Genes (**B**). Here, we used
626 FDR < 0.05 for a gene to be considered differentially expressed. A Red dot represents genes that are
627 called as differentially expressed by both platforms. The Green and Blue dot represents genes that are
628 called differentially expressed by Nanostring and RNA-Seq respectively. **C)** Absolute log fold-change
629 difference between RNA-Seq and Nanostring (y-axis) against gene length (x-axis). A Red dot represents
630 long gene and blue dot represents short genes. P-values were computed using chi-square test.
631
632 Table 1: List of Comparisons used in overlap or average plots

633

Brain region	Mouse strain/Human samples compared	Reference
Fig 1A. Cultured Cortical Neurons (left panel)	BL: Hybrid Vehicle vs Hybrid Vehicle (n = 2 each) RL: Hybrid Topotecan vs Hybrid Vehicle (n = 5 each)	King et al. *Nature* 2013
Cultured Cortical Neurons (middle panel)	BL: Hybrid Vehicle vs Hybrid Vehicle (n = 1 each) RL: Hybrid Topotecan vs Hybrid Vehicle (n = 3 each)	King et al. *Nature* 2013
Cultured Cortical Neurons (right panel)	BL: Hybrid Vehicle vs Hybrid Vehicle (n = 1 each) RL: Hybrid Topotecan vs Hybrid Vehicle (n = 3 each)	Mabb et al. *PLoS One* 2016
Fig 1B. Amygdala (left panel)	BL: C57BL WT vs C57BL WT (n = 2 each) RL: C57BL KO vs C57BL WT (n = 5 each)	Samaco et al. *Nature Gene*tics 2012
Cerebellum (middle panel)	BL: C57BL WT vs C57BL/6J WT (n = 2 each) RL: C57BL KO vs C57BL/6J WT (n = 5 each)	Ben-Shachar et al., *Human Mol. Genet.* 2009
Hypothalamus (right panel)	BL: C57BL WT vs C57BL/6J WT (n = 2 each) RL: C57BL KO vs C57BL/6J WT (n = 4 each)	Chahrour et al., *Science* 2008
Fig 1C. Amygdala (left panel)	BL: FVB WT vs FVB WT (n = 2 each) RL: FVB KO vs FVB WT (n = 5 each)	Samaco et al. *Nature Gene*tics 2012
Cerebellum (middle panel)	BL: FVB WT vs FVB WT (n = 2 each) RL: FVB KO vs FVB WT (n = 5 each)	Ben-Shachar et al. *Human Mol. Genet.* 2009
Hypothalamus (right panel)	BL: FVB WT vs FVB WT (n = 2 each) RL: FVB KO vs FVB WT (n = 4 each)	Chahrour et al., *Science* 2008
Fig 1D. Cortical Excitatory Neurons R106W_{WT} (left panel)	BL: C57BL WT vs C57BL WT (n = 1 each) RL: C57BL R106W_{WT} vs C57BL WT (n = 2 each)	Johnson et al., *Nature Medicine* 2017
Cortical Excitatory Neurons R106W_{MUT} (middle panel)	BL: C57BL WT vs C57BL WT (n = 1 each) RL: C57BL R106W_{MUT} vs C57BL WT (n = 2 each)	Johnson et al., *Nature Medicine* 2017
Cortical Excitatory Neurons T158M_{MUT} (right panel)	BL: C57BL WT vs C57BL WT (n = 1 each) RL: C57BL T158M_{MUT} vs C57BL WT (n = 2 each)	Johnson et al., *Nature Medicine* 2017
Fig 2A. iPSC (left panel)	BL: iPSC WT vs iPSC WT (n = 2 each) RL: iPSC RTT (n = 4) vs iPSC WT (n = 5 each)	GSE#
NPC (middle panel)	BL: NPC WT vs NPC WT (n = 2 each) RL: NPC RTT (n = 4) vs NPC WT (n = 5 each)	GSE#
Neuron (right panel)	BL: Neuron WT vs Neuron WT (n = 2 each) RL: Neuron RTT vs Neuron WT (n = 4 each)	GSE#
Fig 2B. Frontal Cortex (left panel)	RL: Post mortem RTT vs Controls (n = 3 each) BL: Post mortem pooled sample from 2- and 4-year old patient vs Control (n = 1 each)	Deng et al. *Human Mol. Genet.* 2007
Frontal Cortex (right panel)	GL: Post mortem pooled sample from 5-year old patient vs age matched control (n = 1 each) PL: Post mortem pooled sample from 8-year old patient vs age matched control (n = 1 each)	Deng et al. *Human Mol. Genet.* 2007
Fig 2C. Frontal Cortex (left panel)	RTT female samples compared to age matched controls (ages 17-20 years; n = 3)	Lin et al. *BMC Genomics* 2016
Temporal Cortex (right panel)	RTT female samples compared to age matched controls (ages 17-20 years; n = 3)	Lin et al. *BMC Genomics* 2016
Fig 2D. Frontal Cortex (left panel)	RTT female samples compared to age matched controls (ages 18 years; n = 1 each)	GSE#
Frontal Cortex (right panel)	RTT male samples (age 1 year) to compared to age matched (age 2 day) controls (n = 1 each)	GSE#

634 *Note*: Hybrid is mouse line is C57BL/6J (B6) × CASTEi/J (CAST) F1 hybrid mice. BL, RL, GL and PL stands for
635 Blue line, Red line, Green line and Purple line respectively.
636
637 **Supplementary Figure Legends**
638 **Figure S1 Schematic diagram of rigorous assessment of long gene trends**
639 **(Related to Figure 1).**
640
641 **Figure S2 Long gene trend is not present in Mecp2 datasets (Related to Figure 1).** (A-L) Analysis of
642 Intra-sample variation in WT *Mecp2* dataset shows a bias toward long genes across different brain
643 regions. The Blue line (BL) represents the comparison of permuted WT/WT samples from a respective
644 dataset (as mentioned in the comparison table). The Red line (RL) represents the comparison of
645 KO/MUT/Tg samples to its WT littermates from a respective dataset (as mentioned in the comparison
646 table). The top half of each subgraph shows the lines that represent fold-change in expression for genes
647 binned according to gene length (bin size of 200 genes with shift size of 40 genes) as described (Gabel et
648 al., 2015; Zhao et al., 2013). Note that we observe few long gene bins as well as short gene bins with
649 significant preferential upregulation in *Mecp2*-null mice datasets. The blue and red ribbon correspond to
650 one-half of one standard deviation of each bin for the comparison of WT/WT and KO/WT or Tg/WT,
651 respectively. The bottom half of each subgraph is the p-value from the two-sample t-test between KO/WT
652 or Tg/WT and WT/WT. Bins with FDR < 0.05 are showed in red. The red dotted line indicates the
653 minimum $-Log_{10}$(p-value) that corresponds to a FDR < 0.05.
654
655 Here is the list of comparisons for Figure S2:
656

657

Brain region	Mouse lines compared	Reference
A. Striatum	BL: C57BL WT vs C57BL WT (n = 2 each) RL: C57BL KO vs C57BL WT (n = 5 each)	Zhao et al. *Neuro of Disease* 2013
B. Hippocampus (4 weeks)	BL: FVBx129 WT vs FVBx129 WT (n = 2 each) RL: FVBx129 KO vs FVBx129 WT (n = 4 each)	Baker et al. *Cell* 2013
C. Hippocampus (9 weeks)	BL: FVBx129 WT vs FVBx129 WT (n = 2 each) RL: FVBx129 KO vs FVBx129 WT (n = 4 each)	Baker et al. *Cell* 2013
D. Visual Cortex	BL: WT vs WT (n = 1 each) RL: KO (Mecp2tm1.1Bird) vs WT (n = 3 each)	Gabel, Kinde et al. *Nature* 2015
E. Locus Coeruleus Neurons (TH Young/~P22)	BL: C57BL/6J WT vs C57BL/6J WT (n = 1 each) RL: C57BL/6J KO vs C57BL/6J WT (n = 3 each)	Sugino et al. *J Neurosci.* 2014
F. Locus Coeruleus Neurons (TH)	BL: C57BL/6J WT vs C57BL/6J WT (n = 1 each) RL: C57BL/6J KO vs C57BL/6J WT (n = 3 each)	Sugino et al. *J Neurosci.* 2014
G. Fast Spiking interneurons, Motor Cortex (G42)	BL: C57BL/6J WT vs C57BL/6J WT (n = 2 each) RL: C57BL/6J KO vs C57BL/6J WT (n = 4 each)	Sugino et al. *J Neurosci.* 2014
H. Purkinje Cells, Cerebellum (G42)	BL: C57BL/6J WT vs C57BL/6J WT (n = 1 each) RL: C57BL/6J KO vs C57BL/6J WT (n = 3 each)	Sugino et al. *J Neurosci.* 2014
I. Pyramidal Neurons, Motor Cortex (YPFH)	BL: C57BL/6J WT vs C57BL/6J WT (n = 1 each) RL: C57BL/6J KO vs C57BL/6J WT (n = 3 each)	Sugino et al. *J Neurosci.* 2014
J. Callosal Projection Neurons	BL: C57BL/6J WT vs C57BL/6J WT (n = 1 each) RL: C57BL/6J KO vs C57BL/6J WT (n = 3 each)	Kishi et al. *Nature Comm.* 2016
K. Hypothalamus (KO – RNA-Seq)	BL: FVBx129SvEvTac WT vs FVBx129SvEvTac WT (n =1 each) RL: FVBx129SvEvTac KO vs FVBx129SvEvTac WT (n = 3 each)	Chen et al. *PNAS* 2015
L. Hypothalamus (Tg – RNA-Seq)	BL: FVBx129SvEvTac WT vs FVBx129SvEvTac WT (n =1 each) RL: FVBx129SvEvTac Tg vs FVBx129SvEvTac WT (n = 3 each)	Chen et al. *PNAS* 2015

658

659 **Figure S3 Intra-sample variation bias in WT *Mecp2* datasets is independent of the RNA isolation**
660 **method (Related to Figure 1):** Blue line (BL) represents the comparison of permuted WT/WT samples
661 from a respective dataset (as mentioned in the comparison table). The Red line (RL) represents the
662 comparison of MUT samples to its WT littermates from a respective dataset (as mentioned in the
663 comparison table). The top half of each subgraph shows the lines that represent fold-change in expression
664 for genes binned according to gene length (bin size of 200 genes with shift size of 40 genes) as described
665 (Gabel et al., 2015). The blue and red ribbon correspond to one-half of one standard deviation of each bin
666 for the comparison of WT/WT and MUT/WT respectively. The bottom half of each subgraph is the p-
667 value from the two-sample t-test between MUT/WT and WT/WT. Bins with FDR < 0.05 are shown in
668 red. The red dotted line indicates the minimum $-Log_{10}$(p-value) that corresponds to a FDR < 0.05.
669

Brain region	Mouse lines compared	Reference
A. Male Cortex (GRO-Seq)	BL: C57BL/6J WT vs C57BL/6J WT (n = 1 each) RL: C57BL/6J R106W vs C57BL/6J WT (n = 2 each)	Johnson et al. *Nature Med.* 2017
B. Male Cortex (Whole Cell)	BL: C57BL/6J WT vs C57BL/6J WT (n = 1 each) RL: C57BL/6J R106Wvs C57BL/6J WT (n = 2 each)	Johnson et al. *Nature Med.* 2017

670

671 **Figures S4 Intra sample variation bias in WT *Mecp2* datasets is independent of the sex of mouse of**
672 **the mouse model (Related to Figure 1):** Blue line (BL) represents the comparison of permuted WT/WT
673 samples from a respective dataset (as mentioned in the comparison table). The Red line (RL) represents
674 the comparison of MUT samples to its WT littermates from a respective dataset (as mentioned in the
675 comparison table). The top half of each subgraph shows the lines that represent fold-change in expression
676 for genes binned according to gene length (bin size of 200 genes with shift size of 40 genes) as described
677 (Gabel et al., 2015). The blue and red ribbon correspond to one-half of one standard deviation of each bin
678 for the comparison of WT/WT and MUT/WT respectively. The bottom half of each subgraph is the p-
679 value from the two-sample t-test between MUT/WT and WT/WT. Bins with FDR < 0.05 are shown in
680 red. The red dotted line indicates the minimum -Log_{10}(p-value) that corresponds to a FDR < 0.05.

681

Brain region	Mouse lines compared	Reference
M. Cortical Excitatory Neurons (R106W, **Male**)	BL: C57BL/6J WT vs C57BL/6J WT (n = 2 each) RL: C57BL/6J MUT vs C57BL/6J WT (n = 4 each)	Johnson et al. *Nature Med.* 2017
N. Cortical Excitatory Neurons (T158M, **Male**)	BL: C57BL/6J WT vs C57BL/6J WT (n = 2 each) RL: C57BL/6J MUT vs C57BL/6J WT (n = 4 each)	Johnson et al. *Nature Med.* 2017
O. Cortical Inhibitory Neurons (R106W, **Male**)	BL: C57BL/6J WT vs C57BL/6J WT (n = 2 each) RL: C57BL/6J MUT vs C57BL/6J WT (n = 4 each)	Johnson et al. *Nature Med.* 2017
P. Cortical Inhibitory Neurons (T158M, **Male**)	BL: C57BL/6J WT vs C57BL/6J WT (n = 2 each) RL: C57BL/6J MUT vs C57BL/6J WT (n = 4 each)	Johnson et al. *Nature Med.* 2017
Q Excitatory Neurons ($T158M_{WT}$; **Female**)	BL: C57BL/6J WT vs C57BL/6J WT (n = 1 each) RL: C57BL/6J MUT vs C57BL/6J WT (n = 2 each)	Johnson et al. *Nature Med.* 2017

682

683 **Figure S5 (Related to Figure 3).** Differentially expressed gene analysis using gene list from Baker et al.,
684 2013 on Mecp2 hippocampus dataset. Scatter plot of log fold-change (log2FC > 0.1 and FDR < 0.05) in
685 expression between FVBx129 KO to its FVBx129 WT littermates (y-axis) against its gene length (x-axis)
686 in samples of hippocampus from 4-week old and 9-week old mice (n = 4; (Baker et al., 2013)).

687

688 **Figure S6 Long gene bias in the SEQC dataset (Related to Figure 4). (A)** Brain vs. Brain randomized
689 log fold-change plot against gene length (left panel; n = 32 each), Universal human reference (UHR) vs
690 UHR randomized log fold-change plot against gene length (middle panel; n = 32 each) and log2 fold-
691 change plot against transcript length using β ratio samples in RNA-Seq dataset (right panel; n = 32 each).
692 **(B)** Brain vs. Brain randomized fold-change plot against gene length (left panel; n = 2 each), Universal
693 human reference (UHR) vs UHR randomized log fold-change plot against gene length (middle panel; n=2
694 each) and log2 fold-change plot against transcript length using β ratio samples in microarray dataset (right
695 panel; n = 2 each). Each blue dot is a bin of 200 genes with shift size of 40 genes (Gabel et al., 2015).
696

697 **Figure S7 Long gene bias is independent of normalization methods (Related to Figure 4).** Log2 fold-
698 change plot against gene length using β ratio samples (n =64 each) for all genes using **(A) l**ibrary size
699 normalization (or total count) against gene length (left panel) & transcript length (right panel) and **(B)**
700 TMM (edgeR) normalization against gene length (left panel) & transcript length (right panel). Each blue
701 dot is a bin of 200 genes with shift size of 40 genes (Gabel et al., 2015).
702

703 **Figure S8 Long gene bias is not observed in Nanostring dataset (Related to Figure 4). (A)** PCA plot
704 on the NanoString dataset (n = 6 each sample type). **(B)** Scatter plot for mean gene expression in brain
705 samples against its gene length **(C)** brain vs. brain randomized fold-change plot against gene length (n = 6
706 each). **(D)** Log2 fold-change plot against gene length using (B-A/C-A) = 4:1 samples (n =6 each).
707

708 **Figure S9 Explanation of reciprocal relationship among transcriptional changes between RTT and**
709 ***MECP2* duplication syndrome. (A)** PCA Plot of the B samples in Novartis SEQC dataset using library
710 prep IDs. **(B)** Comparison of brain samples having library preparation 1 vs 2 against gene length (n = 16
711 each). **(C)** Comparison of brain samples having library preparation 3 vs 2 against gene length (n = 16
712 each). **(D)** Differential expression analysis between brain samples having library preparation id 1 vs 2 and
713 3 vs 2 across different fold changes and FDR < 0.05. Each blue dot is a bin of 200 genes with shift size of
714 40 genes (Gabel et al., 2015). The red and blue dot in (D) represent long and short genes, respectively.
715

716 **Figure S10 RNA-Seq and Nanostring analysis of Mecp2 KO and WT samples from the cerebellum**
717 **of male mice (Related to Figure 5)**. A) Analysis using all the genes in the RNA-Seq cerebellum dataset.
718 PCA Plot of Mecp2 KO and WT samples (left panel), overlap plot (middle panel) where, blue line (BL)
719 represents the comparison of permuted WT/WT samples from a respective dataset (n = 1 each). The Red
720 line (RL) represents the comparison of KO samples to its WT littermates (n = 3 each). The top half of

721 each subgraph shows the lines that represent fold-change in expression for genes binned according to
722 gene length (bin size of 200 genes with shift size of 40 genes) as described (Gabel et al., 2015). The blue
723 and red ribbon correspond to one-half of one standard deviation of each bin for the comparison of
724 WT/WT and KO/WT respectively. The bottom half of each subgraph is the p-value from the two-sample
725 t-test between KO/WT and WT/WT. Bins with FDR < 0.05 are shown in red. The red dotted line
726 indicates the minimum $-Log_{10}$(p-value) that corresponds to a FDR < 0.05. Scatter plot of log fold-change
727 in expression between KO and WT samples (right panel; n = 3 each) against gene length. The
728 differentially expressed genes (FDR < 0.05 & absolute log2FC > log2(1.2)) were plotted. B) Analysis
729 using 750 genes common in both RNA-Seq and Nanostring dataset. PCA plot of Mecp2 KO and WT
730 samples (n = 3 each) by RNA-Seq (left panel) and Nanostring (right panel) platforms. C) Comparison of
731 log2 fold changes using classical/standard method (left panel) and shrunken log2 fold changes (right
732 panel) using DESeq2.

STAR METHODS

KEY RESOURCES TABLE

Deposited Data

Table 1 has all details about the datasets used in the analysis with GEO Accession IDs.

Software and Algorithms

Reagent or Resource	Source	Identifier
STAR aligner (v2.4.2a)	Dobin et al., 2013	https://github.com/alexdobin/STAR/releases/tag/STAR_2.4.2a
DESeq2	Love et al., 2014	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
edgeR	Robinson et al., 2010	https://bioconductor.org/packages/release/bioc/html/edgeR.html
NanoStringNorm	Waggott et al., 2012	https://cran.r-project.org/web/packages/NanoStringNorm/index.html
GenomicFeatures	Lawrence et al., 2013	https://bioconductor.org/packages/release/bioc/html/GenomicFeatures.html
ggplot2	Hadley Wickham. ggplot2: Elegant Graphics for Data Analysis (2010)	https://github.com/tidyverse/ggplot2
cowplot	CRAN Package	https://github.com/wilkelab/cowplot

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Zhandong Liu (zhandong.liu@bcm.edu).

EXPERIMENTAL MODELS AND SUBJECT DETAILS

Mice

All mice used in this study were FVB.129 F1-hybrids. They were group-housed with up to five mice per cage. They were maintained on a 14h light:10h dark cycle (light on at 06:00) with standard mouse chow and water *ad libitum* in our AAALAS-accredited facility. All research and animal care

758 procedures were approved by the Baylor College of Medicine Institutional Animal Care and Use
759 Committee.
760
761 **METHOD DETAILS**
762
763 **Analysis of Mecp2 datasets**
764 The transcriptome datasets from *Mecp2* studies generated using microarray (GEO accession ids:
765 GSE50225, GSE11150, GSE15574, GSE33457, GSE42895, GSE42987, GSE8720 and GSE6955) were
766 downloaded from GEO. RMA function (Gautier et al., 2004; Irizarry et al., 2003) in the R "affy" package
767 was used to perform background correction, normalization, and summarization of core probesets. NetAffx
768 annotation files (Release 33 for mm9) was used to map affy probes to its official gene symbols. The
769 expression values for genes with multiple probes were obtained by taking the average $\log_2$ expression
770 value across all the probes corresponding to each gene. The NetAffx annotation file has information about
771 the probe location, length and gene coordinates; we calculated gene length using the gene coordinates,
772 and we specifically used gene length in all our figures where "Gene length in KB" is defined on the x-
773 axis. We also ran our analysis on the transcript length (see figures S5A-B, right panel, and S6A-B, right
774 panel). The extent of length-dependent bias with transcript length was similar to that of gene length. Since
775 gene length information was not available in case of Affymetrix Human Genome U95 version 2 array, we
776 mapped the probe to its gene and gene length using Ensembl Biomart database (version
777 GRCh38.p5/Ensembl Genes 84).
778 The transcriptome dataset of the virtual cortex (Gabel et al., 2015) (GSE60077) was mapped to
779 mm10 genome using STAR aligner v2.4.2a (Dobin et al., 2013) and for hypothalamus RNA-Seq dataset
780 (Chen et al., 2015) (GSE66871), we used a published list of differentially expressed genes and normalized
781 counts. For Johnson et al. (Johnson et al., 2017) RNA-Seq dataset (GSE83474), we used the raw count
782 files provided by the authors in GEO. Similarly, for the transcriptome analysis of frontal and temporal
783 cortex from RTT patients, we used the normalized gene expression profile provided in GSE75303 (Lin et
784 al., 2016). We performed box plot and MDS plot to check for outliers in the sample distribution. The
785 annotation files provided by GPL10558 were obtained to map Illumina probes to official gene symbols
786 and RefSeq hg19 annotation was used to obtain gene length information.
787
788 **Running Average Plots**
789 We used the same method as described in (Gabel et al., 2015) to compute the running average
790 plot. In brief, the genes were sorted by their lengths and partitioned into bins using a sliding window of
791 200 consecutive genes in steps of 40 genes. The $\log_2$ fold-change values for genes within each bin were

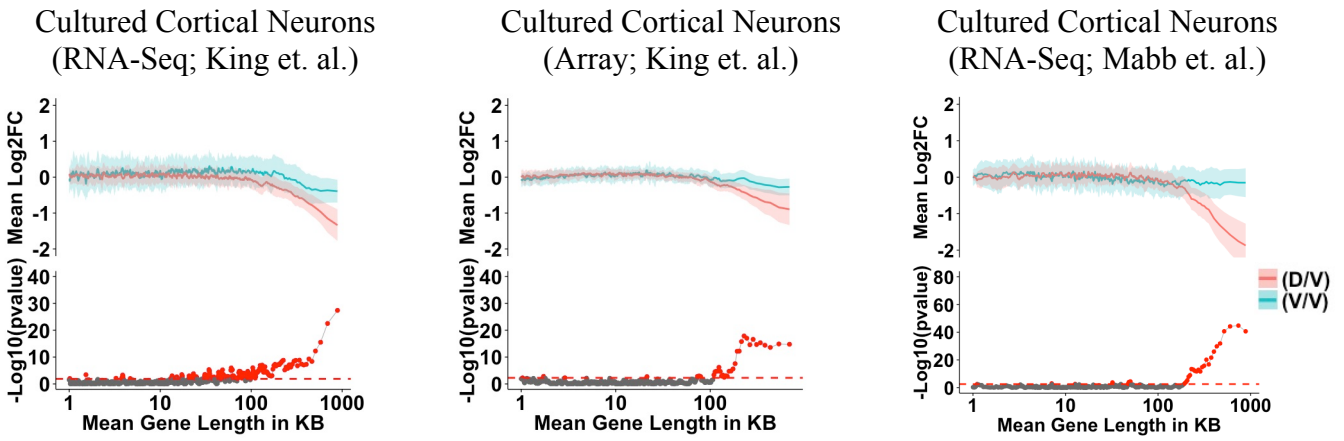
792 averaged. For consistency with the previous studies, we used genes whose lengths are between 1 kb and
793 1000 kb for all the plots. These plots were created using ggplot2 package in R.
794
795 **Confidence interval estimation in Overlap plots**
796 We define the plots used in Figure 1 as "overlap plots," meaning an overlap of two running
797 average plots that shows intra-sample variation between control samples (WT) and inter-sample variation
798 between two genotypes or conditions. To determine the amount of intra-sample variation, we computed
799 the standard deviation of the genes in the same sliding window. By definition, 95% confidence interval
800 for the mean is sample mean plus minus 1.96 times of the standard deviation. In all our overlap plots for
801 the Mecp2 datasets, however, the confidence interval of KO/WT (or Tg/WT or D/V or RTT/WT) and
802 WT/WT completely overlap. For the sake of legibility, we plotted only half of one standard deviation of
803 the mean for each bin in the comparison of WT/WT and KO/WT (or Tg/WT or D/V or RTT/WT), which
804 is denoted by the blue and red ribbon, respectively. Two-sample Student t-test was applied to each of the
805 bins between KO/WT (or Tg/WT or D/V or RTT/WT) and WT/WT, followed by multiple hypothesis
806 adjustment using the Benjamini-Hochberg method (FDR). The significant bins (FDR < 0.05) are denoted
807 by red and non-significant bins are denoted by grey. The overlap plots were created using cowplot
808 package in R.
809
810 **Distribution of differentially expressed genes in *Mecp2* datasets**
811 To measure the distribution of long gene bias among differentially expressed genes, we extracted
812 published lists of genes found to be significantly activated or repressed by *Mecp2* across different brain
813 region. The published lists of differentially expressed genes were downloaded from the supplementary
814 files in each study. Because of the frequent changes in gene name and annotation, we used MGI batch
815 query (Eppig et al., 2015) to facilitate uniform comparison between these gene lists. The genomic
816 locations were obtained for mm10/GRCm38. The original fold-change and FDR thresholds reported by
817 respective publications were used. In case of microarray datasets, genes were plotted against their length.
818 In the case of the RNA-Seq dataset, the calculation was done based on UCSC transcript IDs. Long genes
819 (gene length > 100 Kb) were represented as red and short genes were represented as blue. The numbers of
820 the upregulated and downregulated long/short genes are shown in four different quadrants.
821
822 **Analysis of SEQC dataset**
823 We measured the long gene fold-change bias in RNA-Seq and microarray benchmark datasets,
824 using the RNA-Seq datasets generated by all the Illumina HiSeq 2000 sites and microarray datasets
825 generated by USF using Affymetrix Human Gene 2.0 ST Array in the SEQC consortium. The RNA-Seq

826 raw count files and microarray PrimeView normalized file were accessed from the Gene Expression
827 Omnibus database (GEO) (Barrett et al., 2013). The GEO accession IDs for the RNA-Seq and microarray
828 datasets are GSE47774 and GSE56457, respectively. Raw count files from the Australian Genome
829 Research Facility (AGR), Beijing Genomics Institute (BGI), Weill Cornell Medical College (CNL), City
830 of Hope (COH), Mayo Clinic (MAY) and Novartis (NVS) were normalized using the DESeq2 method.
831 Principal Component Analysis (PCA) and Multidimensional scaling plots (using Euclidean distance) were
832 used to do a sanity check for a nominal amount of batch effects.
833 For further downstream analysis, we decided to use the Novartis dataset, as it had a minimal
834 amount of non-biological variation (data not shown). The Novartis dataset consisted of 64 technical
835 samples each of A (Universal Human Reference RNA), B (Human Brain Reference RNA), C (3A:1B)
836 and D (1A:3B). We did not use sample type E (Ambion ERCC Spike-In Control Mix 1) or F (Ambion
837 ERCC Spike-In Control Mix 2) in our analysis. For consistency with the SEQC consortium, we used
838 hg19 iGenome NCBI/RefSeq annotation (build 27.2). The *transcripts* and *exon* functions in
839 GenomicFeatures Bioconductor package (Lawrence et al., 2013) were used to obtain the gene and
840 transcript length respectively, from the hg19 GTF file. Since a small number of genes or transcripts have
841 multiple different genomic locations, genes or transcripts with the longest length were used. Expression
842 values for genes with multiple transcript clusters were averaged across all transcript clusters
843 corresponding to each gene. Similarly, for the microarray USF PrimeView dataset, sanity checks were
844 performed using boxplot and MDS plots. Boxplots were used to check if the dataset was properly
845 normalized and MDS plots were used to confirm that the dataset had a nominal amount of batch effects or
846 non-biological variation.
847
848 **Library size normalization using Total Count and Trimmed Mean of M-values**
849 To ensure that our normalization methods were not obscuring a genuine long gene bias, we
850 normalized the raw counts from Novartis RNA-Seq dataset based on two other methods apart from
851 DESeq2 (Love et al., 2014): a) Total Counts (Dillies et al., 2013) and b) the Trimmed Mean of M-values
852 (TMM) method implemented in edgeR (Robinson et al., 2010; Robinson and Oshlack, 2010). For Total
853 Counts, scaling factors were computed such that the normalized read counts across all samples are equal.
854 In the case of the TMM method, we used the *calcNormFactors* function in the edgeR Bioconductor
855 package to get the scaling factors and normalized read counts.
856
857 **SEQC NanoString sample preparation and analysis**
858 We purchased Universal Human Reference RNA from Agilent Technologies, Inc., and Human
859 Brain Reference RNA from Life Technologies, Inc. For the nCounter experiments, we used the same

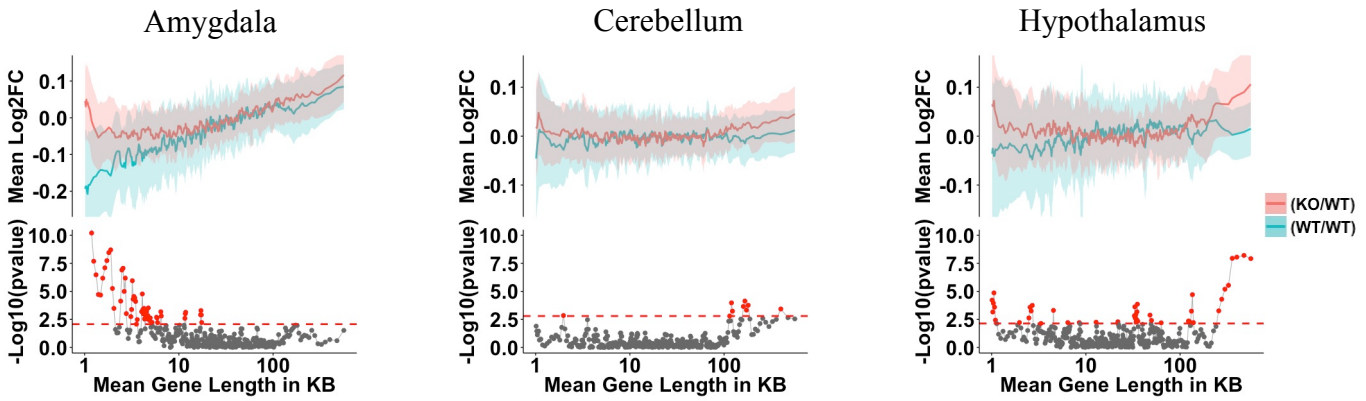
860 RNA sample types as SEQC. We assessed RNA purity and integrity with Bioanalyzer (Agilent
861 Technologies, Inc.) prior to use in the nCounter assays. Sample preparation and analysis were done using
862 a nCounter Prep Station 5s and a nCounter Digital Analyzer 5s. Expression of 770 genes (~730 genes
863 with ~40 housekeeping genes and positive and negative controls) was assessed using the nCounter
864 Human PanCancer Pathways Panel. A second PanCancer Pathways Panel was run using the same samples
865 submitted to the first panel to assess the effect of batches on nCounter results. We used NanoStringNorm
866 function (Waggott et al., 2012)in the R NanoStringNorm package to normalize the dataset. Boxplots and
867 MDS plots were used for sanity checks. The two-sided Wilcoxon rank sum test was used to compare the
868 distribution of the fold-change between long and short genes across the three different platforms.
869
870 **RNA isolation, sequencing and nanostring analysis from mouse cerebellum**
871 We performed RNA extraction and purification from the cerebellum of male mice 8 to 9 weeks of
872 age (three biological replicates of wild-type and Mecp2-null) using the Aurum™ Total RNA Fatty and
873 Fibrous Tissue Kit (Bio-Rad 7326830) per the manufacturer's instructions. Genomic DNA was
874 eliminated using an on-column DNase digestion step. RNA quality was assessed using the Agilent 2100
875 Bioanalyzer system prior to library preparation for deep sequencing or use of the total RNA for
876 Nanostring nCounter quantification.
877 RNA sequencing was performed using Illumina HiSeq 2000. All sequencing was done by the
878 Genomic and RNA Profiling Core at the Baylor College of Medicine. For each sample, about 90 to 110
879 million pairs of 100 bp reads were generated. Raw reads were aligned to the *Mus musculus* genome
880 (Gencode mm10; version M10) using STAR aligner v2.4.2a (Dobin et al., 2013) with default parameters.
881 The overall mappability for all 6 samples was above 90% (Table 2). The read counts per gene were
882 obtained using the *quantMode* function in STAR. These read counts are analogous to the expression level
883 of the gene. Using the obtained raw counts, normalization and differential gene analysis were carried out
884 using the DESeq2 package in the R environment. DESeq2 allows us to test for gene expression changes
885 between samples in different conditions using more robust shrinkage estimation for dispersion and fold
886 changes (Love et al., 2014). The default negative binomial generalized linear model with Wald test
887 implemented in the package was used to identify significant differential expressed genes. Log fold -
888 change was calculated using both the classic method and shrinkage estimates calculated by DESeq2.
889 For the nCounter experiments, sample preparation and quality analysis were done using a
890 nCounter Prep Station 5s and an nCounter Digital Analyzer 5s. Expression of 784 genes (750 endogenous
891 genes with 34 housekeeping genes and positive and negative controls) was assessed using the nCounter
892 Mouse PanCancer Pathways Panel. We used NanoStringNorm function (Waggott et al., 2012) in the R
893 NanoStringNorm package to normalize the dataset and DESeq2 for differential expression analysis.

Figure 1

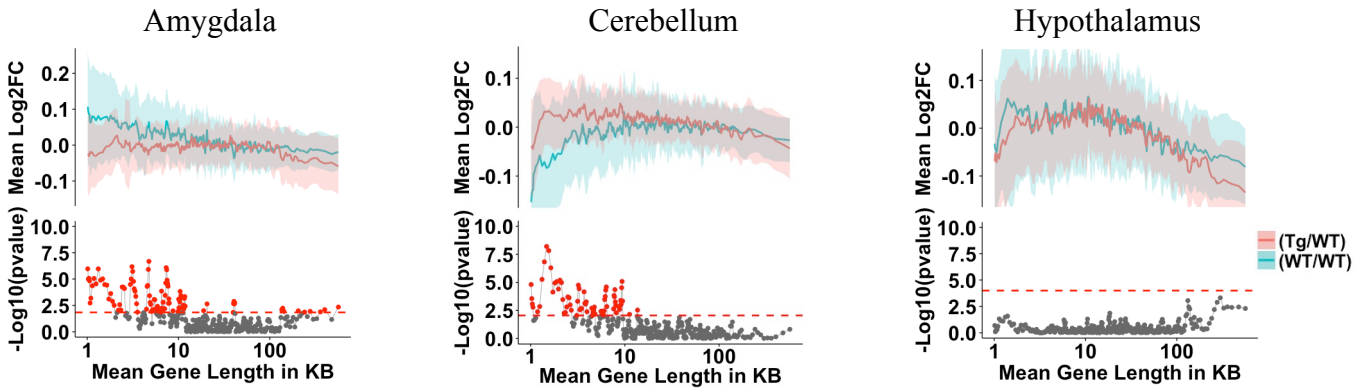
A 300nM Topotecan Treatment



B Mecp2 (KO/WT) Male Mouse Models



C Mecp2 (Tg/WT) Male Mouse Models



D Mecp2 (MUT/WT) Female Mouse Models

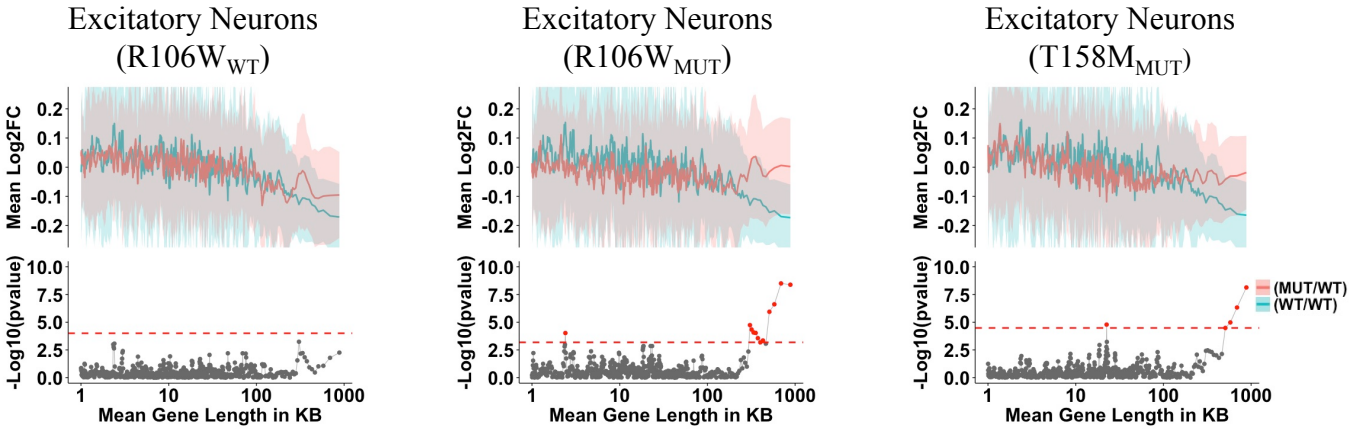
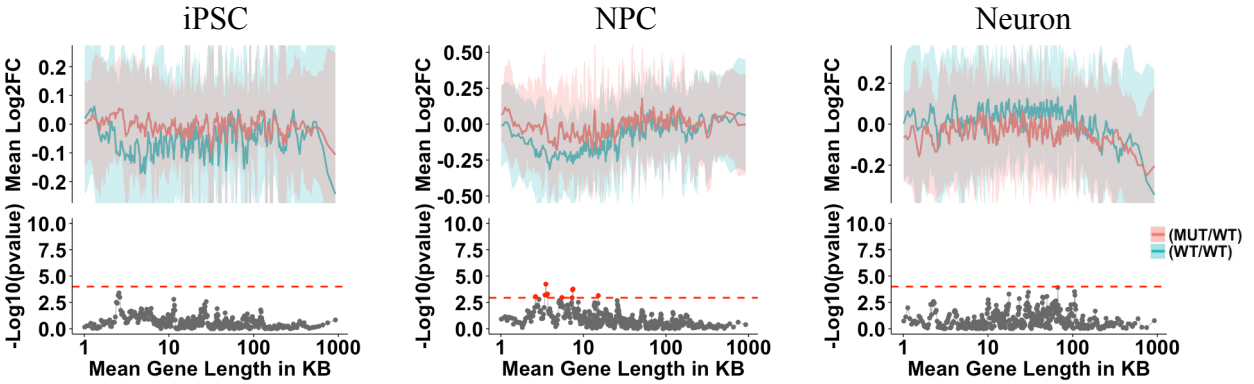
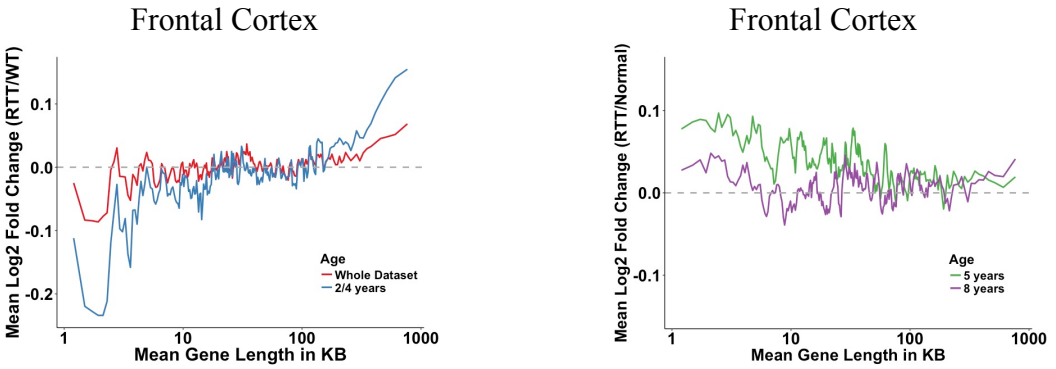


Figure 2

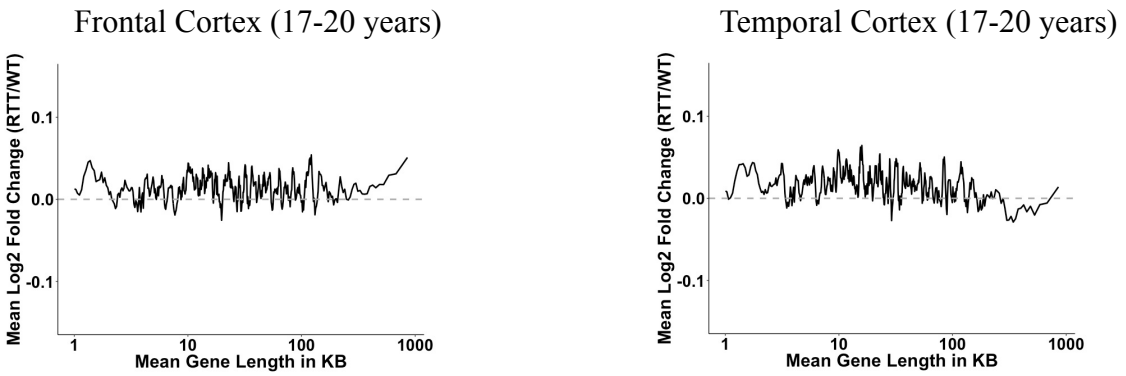
A Lowry's Human RTT *in vitro* Dataset



B Deng's Dataset (RTT/WT)



C Lin's Dataset (RTT/WT)



D Lowry's Human RTT Dataset (RTT/WT)

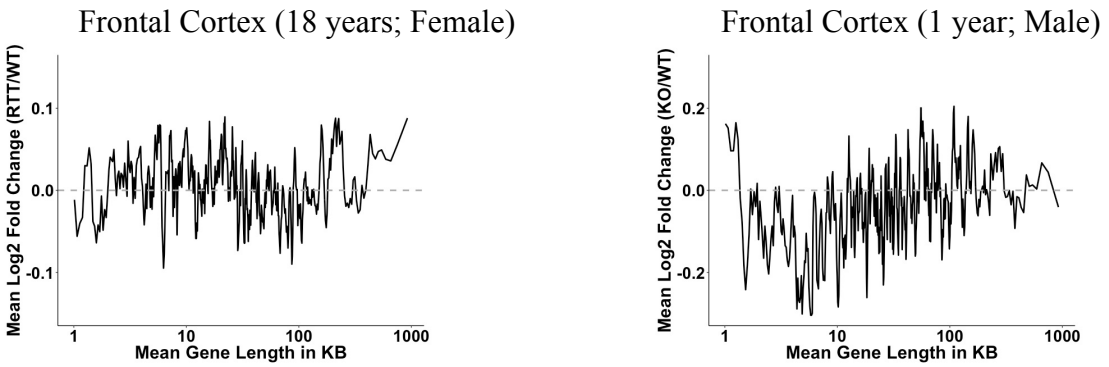
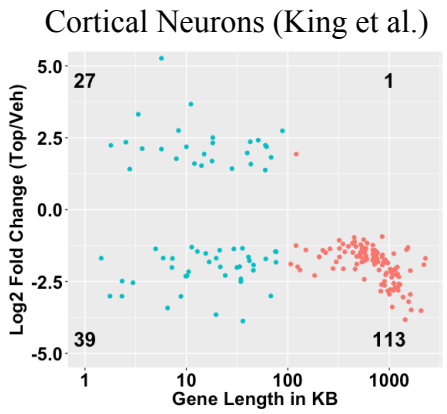
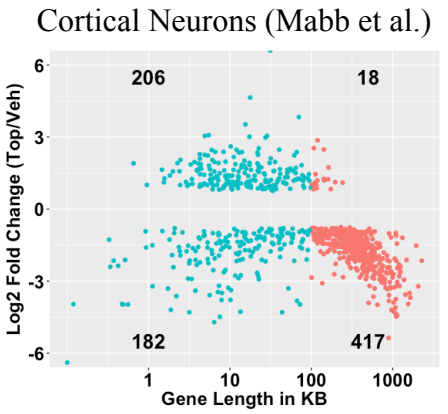


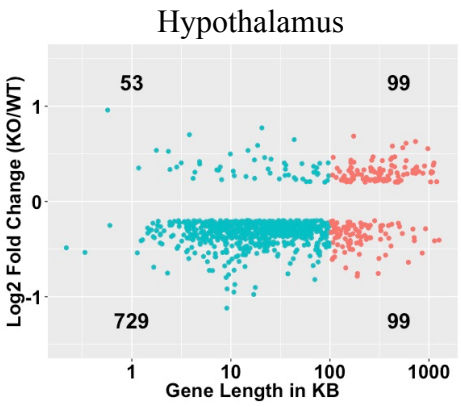
Figure 3

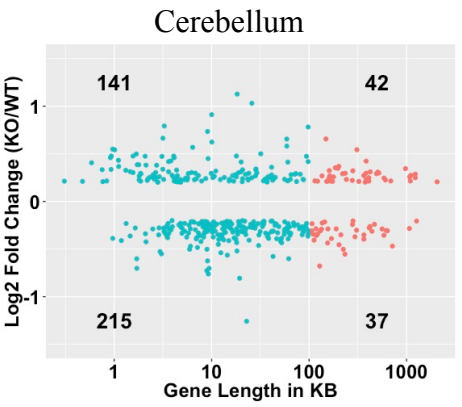
A 300nM Topotecan Treatment RNA-Seq Datasets



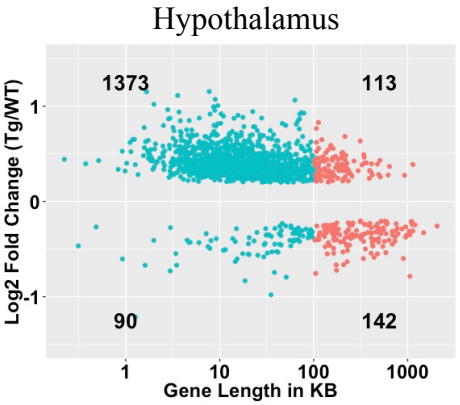


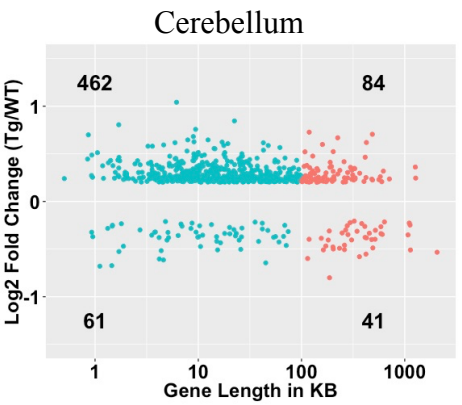
B Mecp2 related Array (KO/WT) Datasets



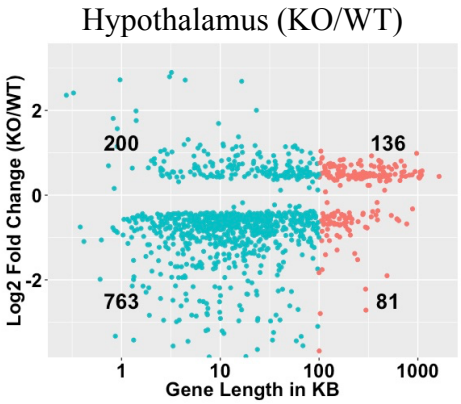


C Mecp2 related Array (Tg/WT) Datasets





D Mecp2 related Seq Datasets



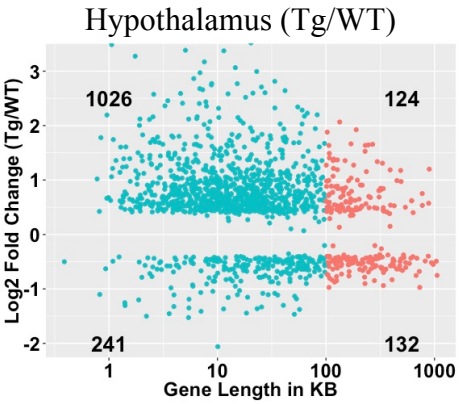
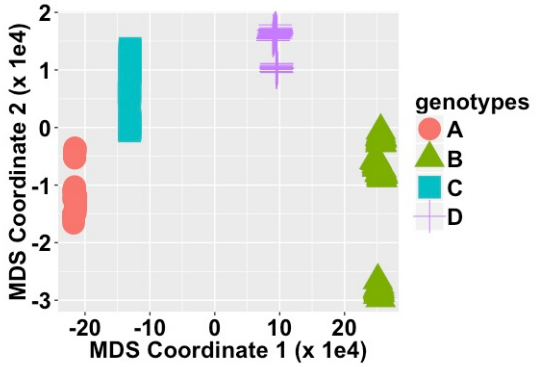


Figure 4

SEQC RNA-Seq

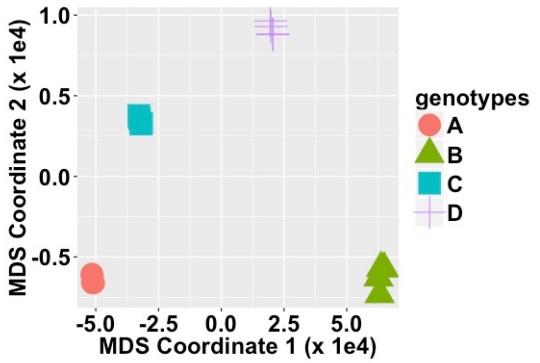
A



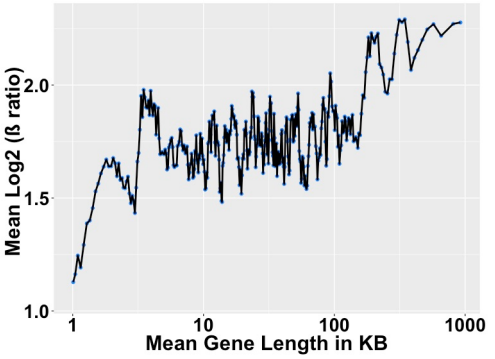
B

SEQC Array

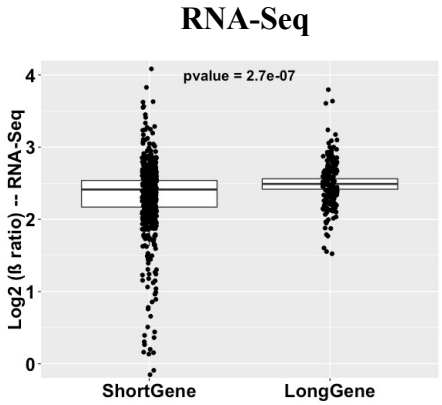
C



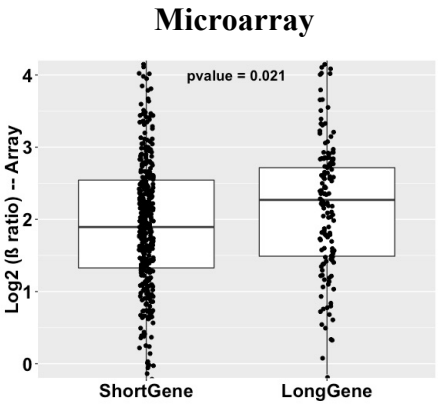
D



E



F



G

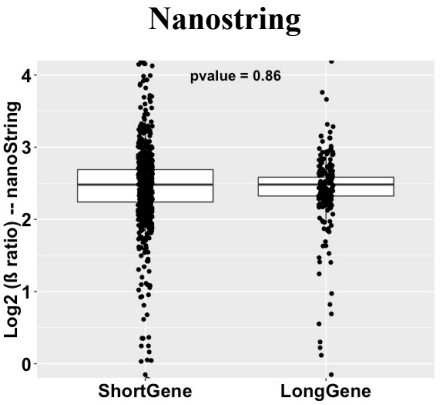
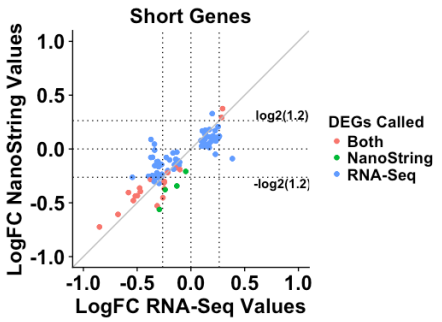
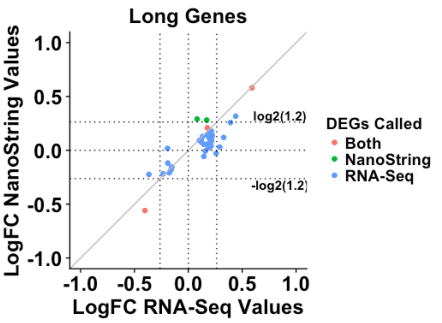


Figure 5

A


Short Genes
LogFC NanoString Values
LogFC RNA-Seq Values
1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0
log2(1.2)
-log2(1.2)
DEGs Called
Both
NanoString
RNA-Seq

B


Long Genes
LogFC NanoString Values
LogFC RNA-Seq Values
1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0
log2(1.2)
-log2(1.2)
DEGs Called
Both
NanoString
RNA-Seq

C

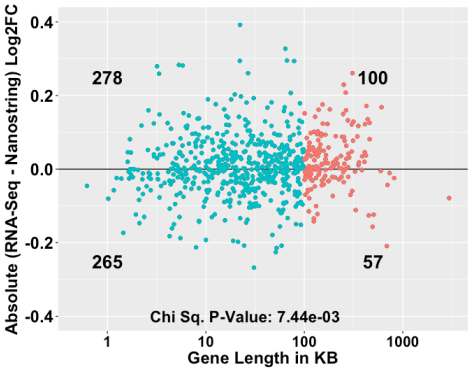

Absolute (RNA-Seq - Nanostring) Log2FC
0.4
0.2
0.0
-0.2
-0.4
278
100
265
57
Chi Sq. P-Value: 7.44e-03
1
10
100
1000
Gene Length in KB

Figure S1

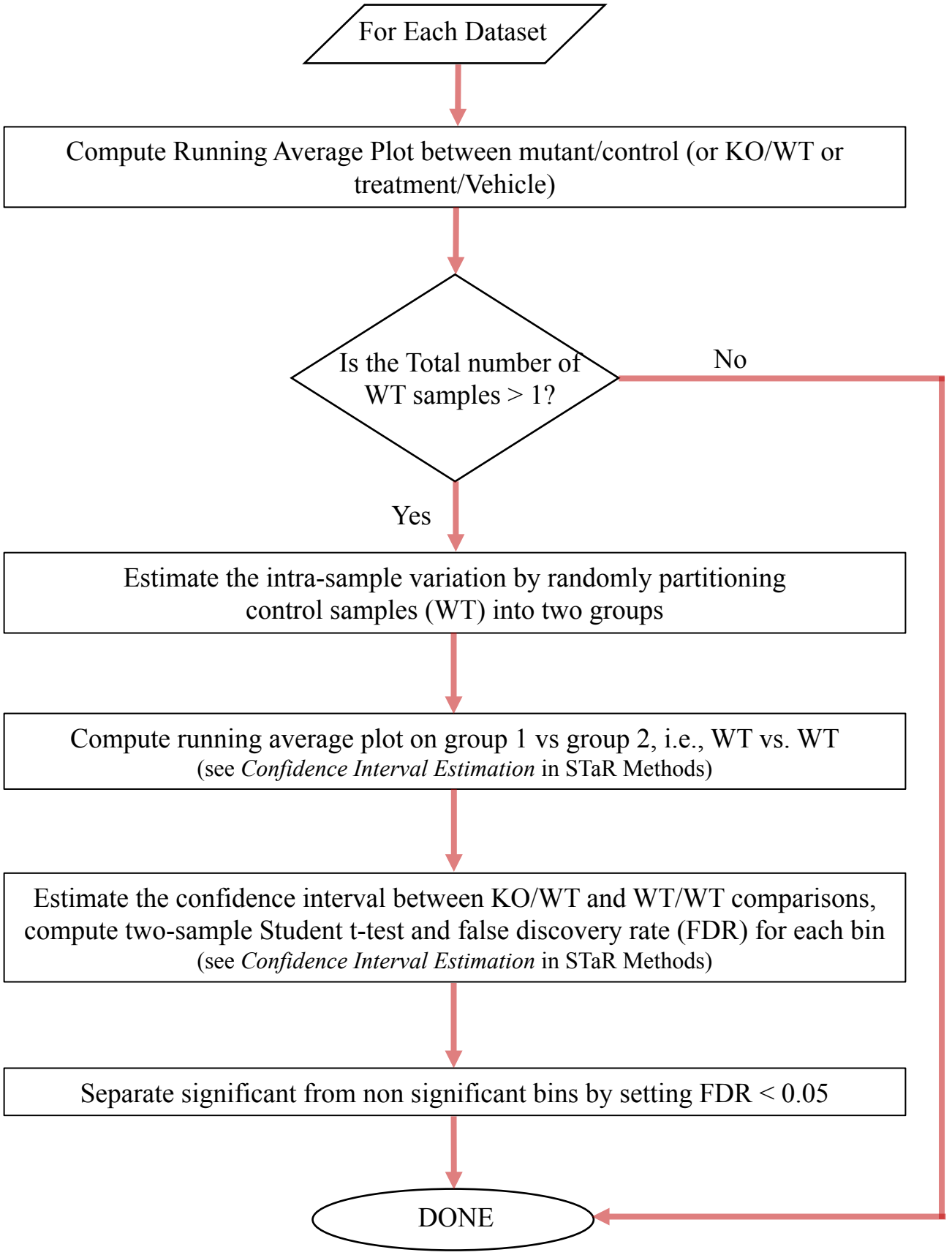


Fig. S1: Schematic for rigorous assessment of long gene trends

Figure S2

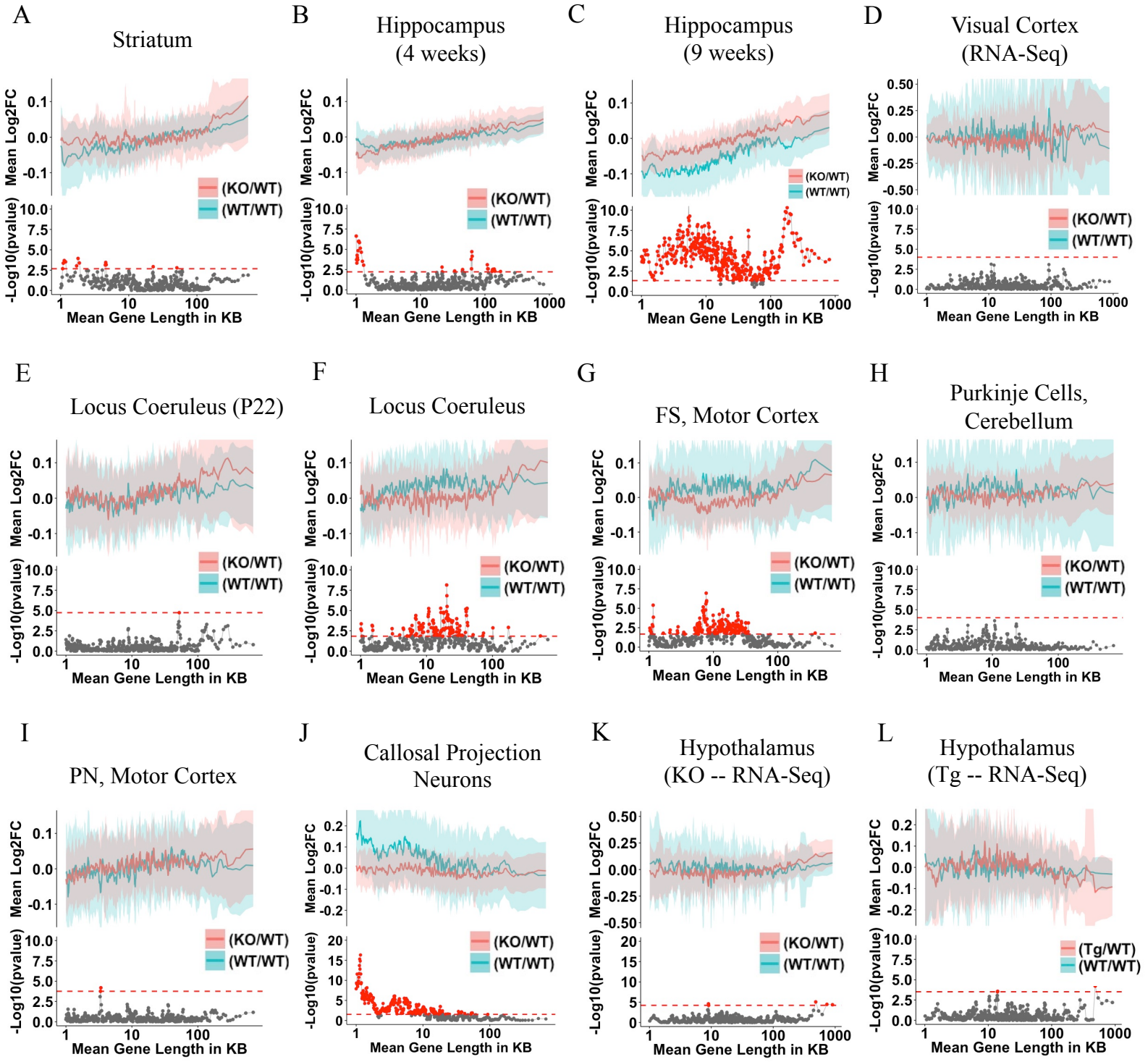

A
Striatum
B
Hippocampus (4 weeks)
C
Hippocampus (9 weeks)
D
Visual Cortex (RNA-Seq)
E
Locus Coeruleus (P22)
F
Locus Coeruleus
G
FS, Motor Cortex
H
Purkinje Cells, Cerebellum
I
PN, Motor Cortex
J
Callosal Projection Neurons
K
Hypothalamus (KO -- RNA-Seq)
L
Hypothalamus (Tg -- RNA-Seq)
Mean Log2FC
-Log10(pvalue)
Mean Gene Length in KB
(KO/WT)
(WT/WT)
(Tg/WT)

Figure S3

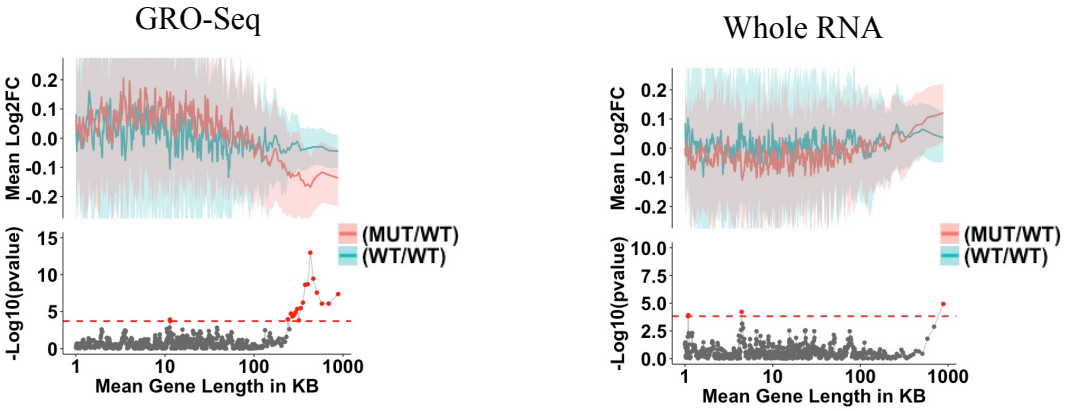


Figure S4

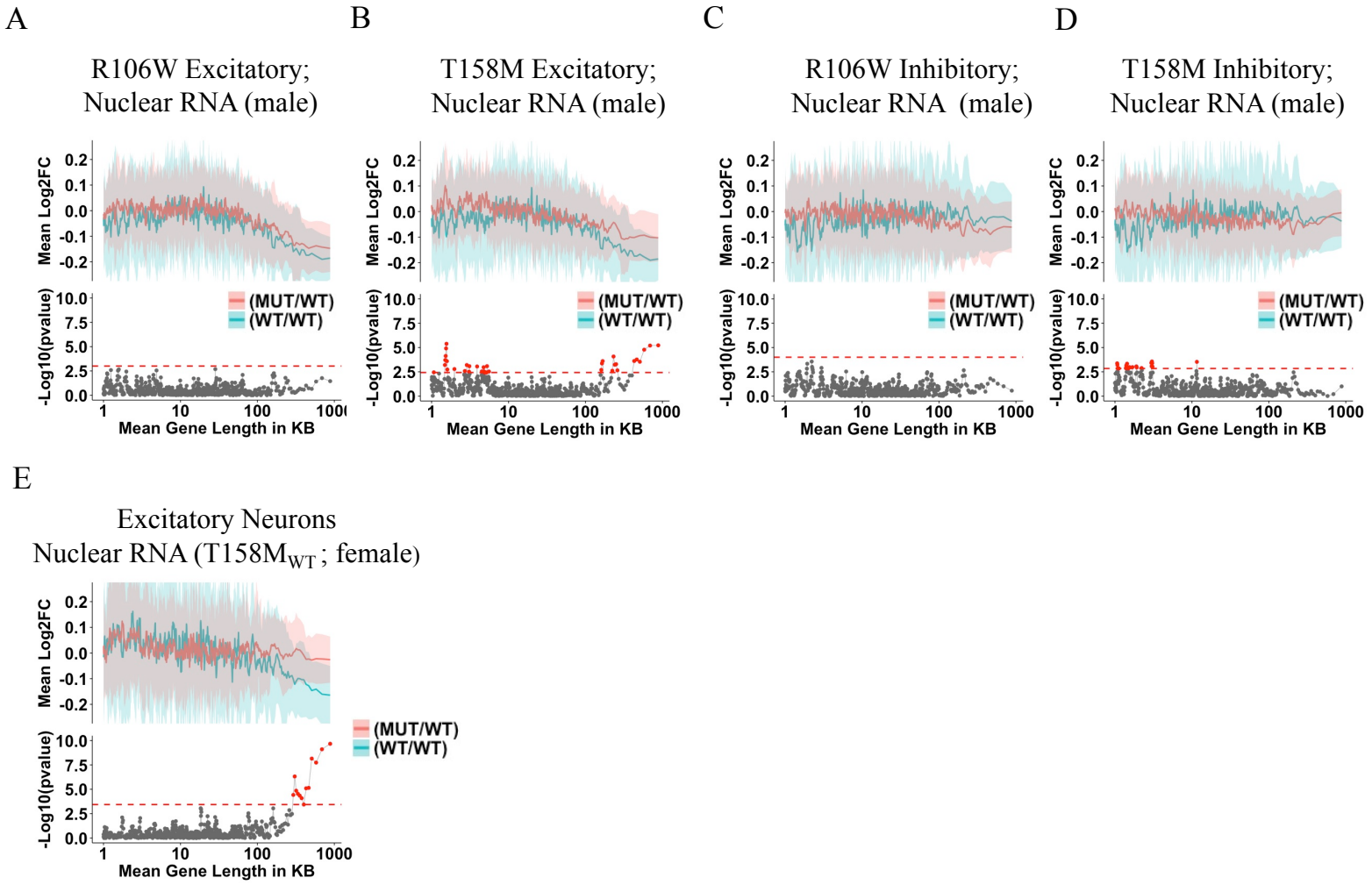


Figure S5

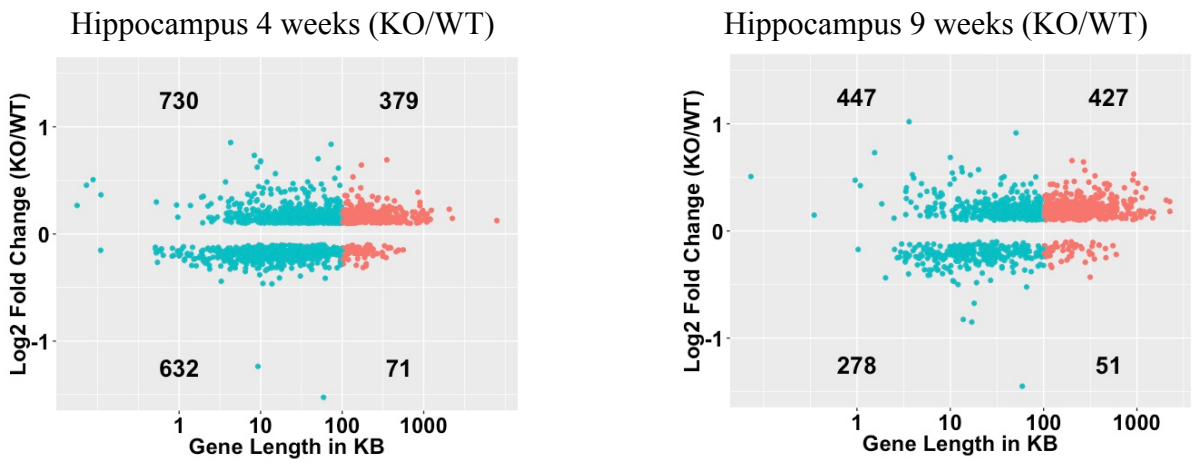
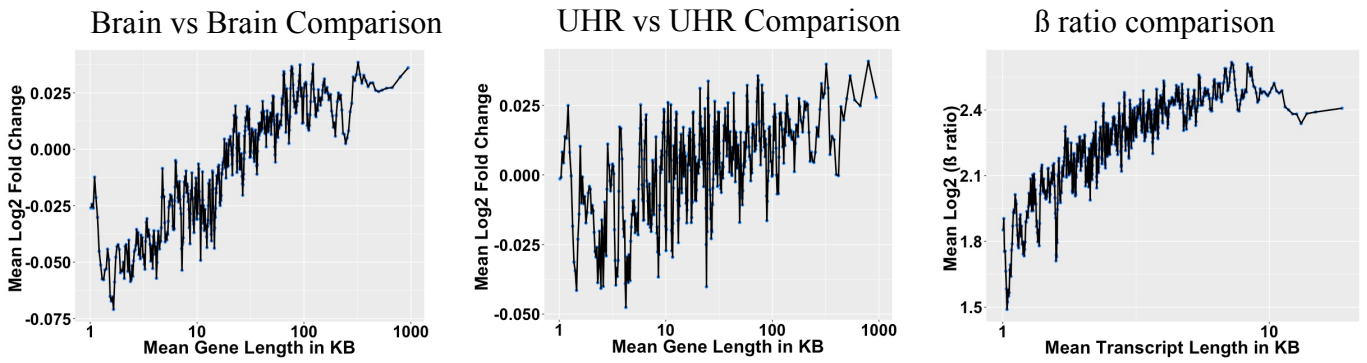


Figure S6

A **SEQC RNA-Seq**



B **SEQC Array**

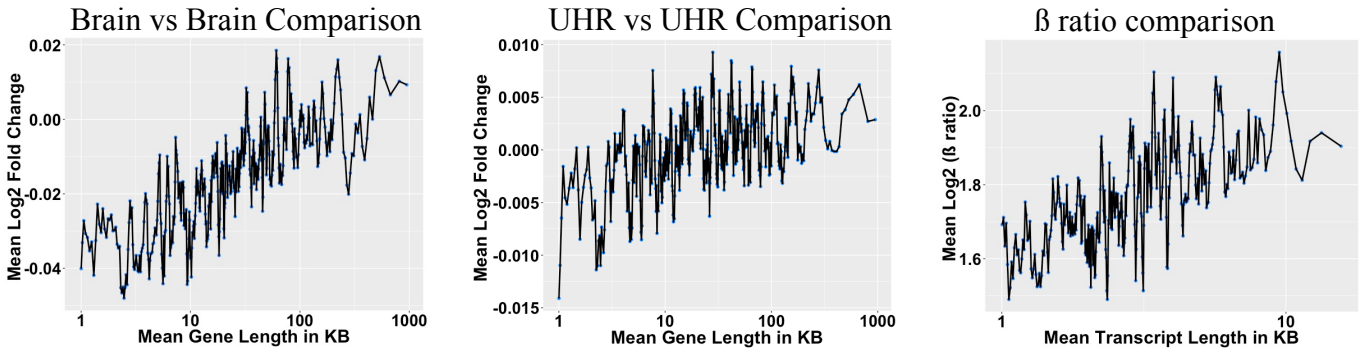
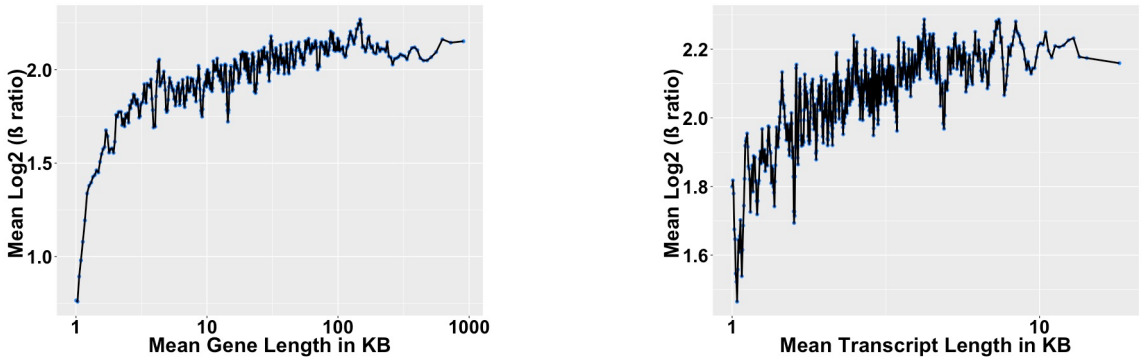


Figure S7

A **Total Count**



B **TMM (edgeR)**

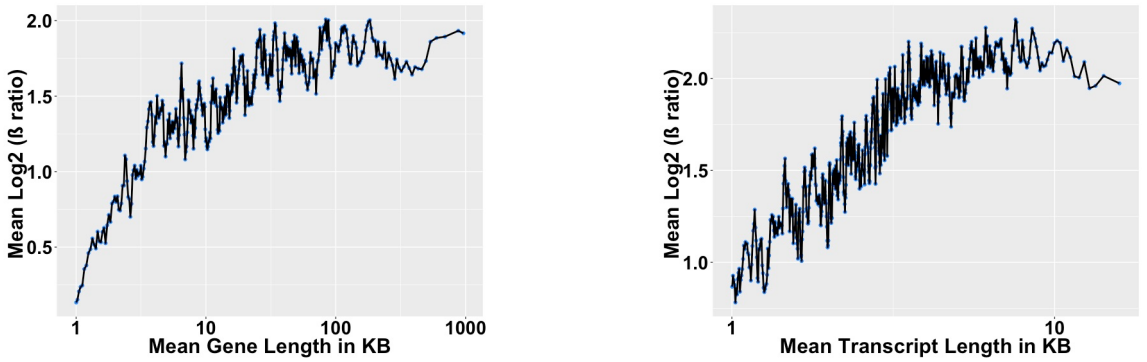
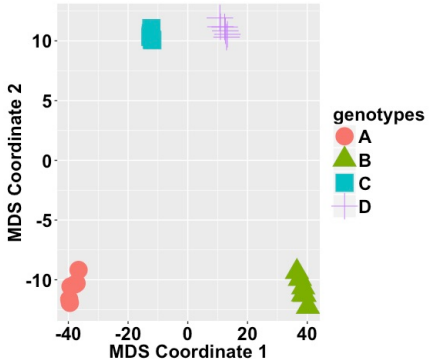
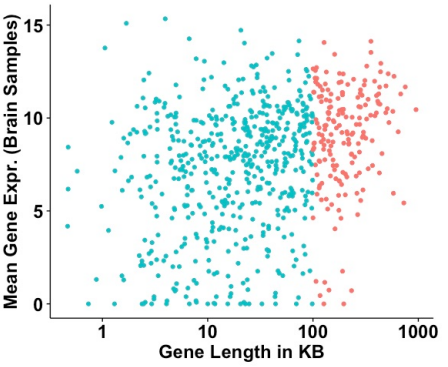


Figure S8

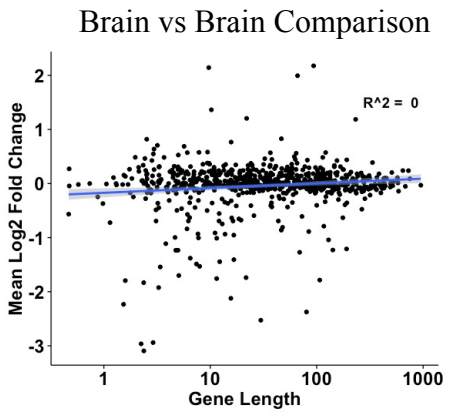
A



B



C



D

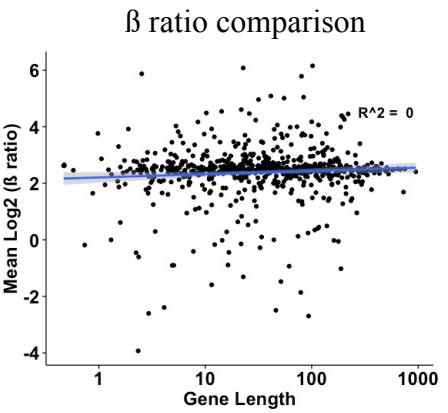


Figure S9

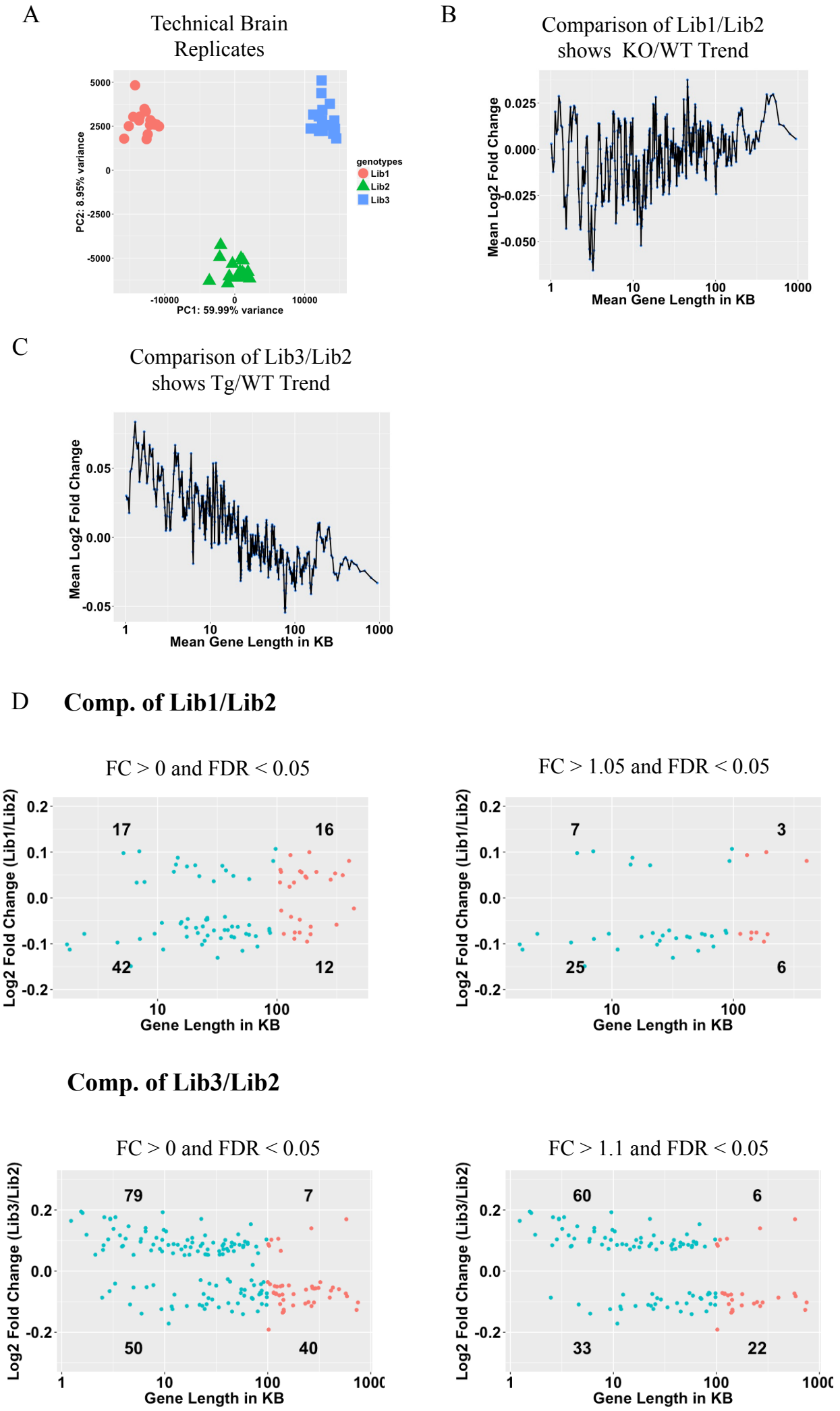
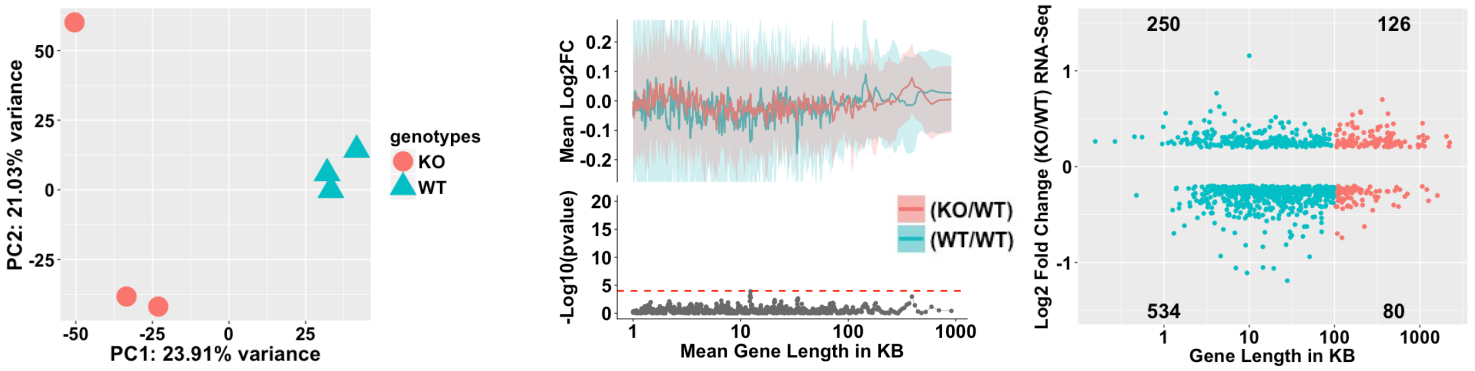
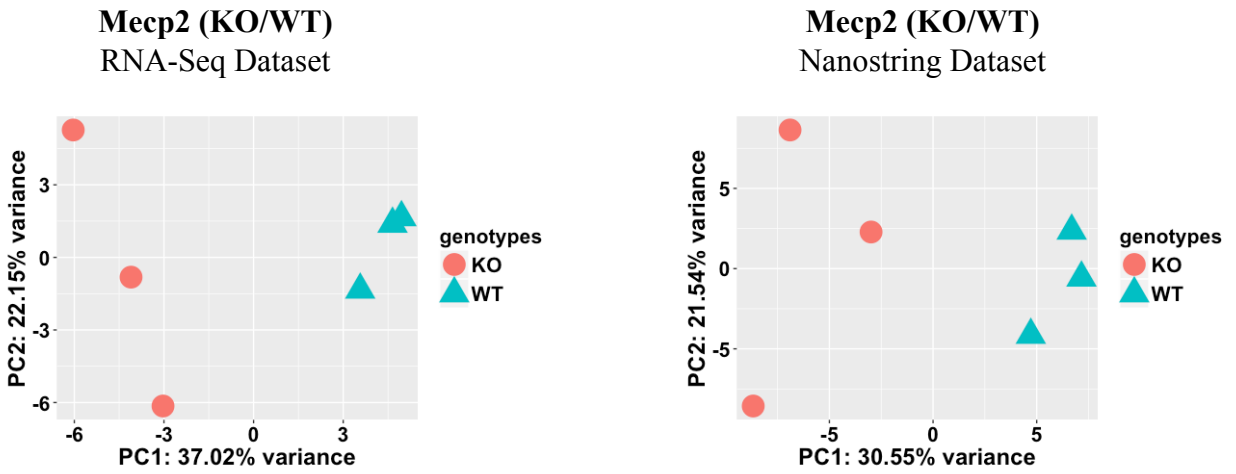
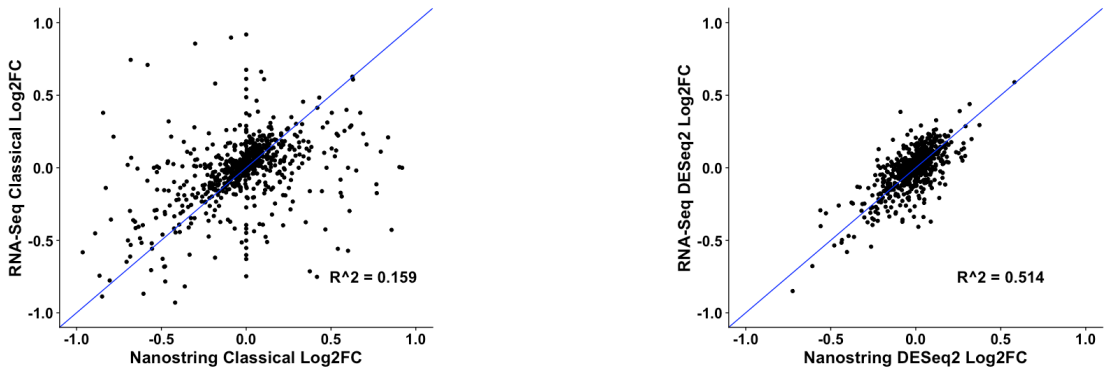

A
Technical Brain Replicates
PC2: 8.95% variance
PC1: 59.99% variance
genotypes
Lib1
Lib2
Lib3
B
Comparison of Lib1/Lib2 shows KO/WT Trend
Mean Log2 Fold Change
Mean Gene Length in KB
C
Comparison of Lib3/Lib2 shows Tg/WT Trend
Mean Log2 Fold Change
Mean Gene Length in KB
D
Comp. of Lib1/Lib2
FC > 0 and FDR < 0.05
17
16
42
12
Log2 Fold Change (Lib1/Lib2)
Gene Length in KB
FC > 1.05 and FDR < 0.05
7
3
25
6
Log2 Fold Change (Lib1/Lib2)
Gene Length in KB
Comp. of Lib3/Lib2
FC > 0 and FDR < 0.05
79
7
50
40
Log2 Fold Change (Lib3/Lib2)
Gene Length in KB
FC > 1.1 and FDR < 0.05
60
6
33
22
Log2 Fold Change (Lib3/Lib2)
Gene Length in KB

Figure S10

A **Mecp2 Cerebellum RNA-Seq KO/WT Dataset (Whole Genome)**



B **750 common genes between RNA-Seq and Nanostring**



C