MECP2pedia: A Comprehensive Transcriptome Portal for MECP2 Disease Research

Alexander J. Trostle (1,2,*), Lucian Li (1,2,*), Seon-Young Kim (1,3), Jiasheng Wang (1), Rami Al-Ouran (1,2), Hari Krishna Yalamanchili (1,4), Zhandong Liu (1,2,+), Ying-Wooi Wan (1,3,+)

1. Jan and Dan Duncan Neurological Research Institute at Texas Children's Hospital, Houston, TX 77030, USA.
2. Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030, USA.
3. Department of Molecular and Human Genetics, Baylor College of Medicine, Howard Hughes Medical Institute, Houston, TX 77030, USA.
4. USDA/ARS Children's Nutrition Research Center, Department of Pediatrics, Baylor College of Medicine, Houston, TX, USA

* = these authors contributed equally to this work

+ = corresponding author

Summary

Mutations in MeCP2 result in a crippling neurological disease, but we lack a lucid picture of MeCP2s molecular role. Focusing on individual transcriptomic studies yields inconsistent differentially expressed genes. We have aggregated and homogeneously processed modern public MeCP2 transcriptome data, which we present in a web portal. With this big data, we discovered a commonly perturbed core set of genes that transcends the limitations of any individual study. We then found distinct consistently up and downregulated subsets within these genes. We observe enrichment for this mouse core in other species MeCP2 models and see overlap between this core and ASD models. Analysis of signal to noise finds that many studies lack enough biological replicates. By integrating and examining transcriptomic data at scale, we have generated a valuable resource and insight on MeCP2 function.

Keywords

MeCP2, data portal, Rett syndrome, MeCP2 duplication syndrome, RNA-seq, differential expression analysis, meta-analysis, mouse models

Introduction

Big data integration and analyses can uncover valuable patterns and insights otherwise missed in individual studies (Costa, 2014), but curating, processing, and analyzing the vast amount of data involved is challenging. This is especially true in the biomedical sciences, where the data is highly variable and complex. Researchers can aggregate publicly available processed results, but doing so will inevitably yield inconsistencies between data sets that can be misleading. However, handling a meaningful quantity of raw data requires extensive time, experience, and computational resources unavailable to most researchers.

Databases with abundant biological data do exist, particularly those compiling gene expression profiles. However, these databases either fail to focus on transcriptome perturbation (GTEx) (Lonsdale et al., 2013) or they require significant time and energy to extract and format data specific to a particular disease (ARCHS4) (Lachmann et al., 2013). Similarly, many databases focus on specific model organisms and allow filtering by disease, such as Flybase (Drysdale et al., 2005) and the Rat Genome Database (Smith et al., 2020), but they fail to offer substantial disease-focused analysis. While there are some molecular-focused databases for well-studied diseases such as cancer (TCGA) (Tomczak et al., 2015), cBio Portal (Cerami et al., 2012) and Alzheimer's disease (AMP-AD) (Hodes and Buckholtz, 2016), there are no analogous databases for less common diseases. The massive success of TCGA, in particular, makes the utility of more disease databases clear.

One such overlooked disease is Rett syndrome (RTT), a severe neurodevelopmental disorder in girls caused by mutations in the X-linked gene for methyl-CpG binding protein 2 (*MECP2*). Development proceeds normally until 6-18 months, at which point it stalls and then regresses (Amir et al., 1999). Symptoms and progression can vary substantially between individuals, and despite recent advances, we still do not completely understand MeCP2's molecular role. MECP2 duplication syndrome (MDS) is an overexpression in the same gene, and patients have substantial overlap in phenotype with RTT (Sandeweiss, Brandt and Zoghbi, 2020).

There are numerous mouse models for both diseases, and public data is available for a variety of tissues, ages, and characteristics. This pool of sequencing data will only grow with time, yet currently there is no centralized resource for it. Existing MeCP2-related disease databases are either primarily patient registries (Rett Database Network) (Sampieri et al., 2007) or they focus on mutation information like IRSA North American Database (Percy et al., 2007) and RettBASE (Krishnaraj, Ho, and Christodoulou, 2017). No current database curates MeCP2 expression data in the focused manner that would be most useful to researchers.

To fill this need, we have created MECP2pedia, a database for molecular MeCP2. MECP2pedia is a uniformly processed and expansive collection of MeCP2 transcriptomic data, with readily accessible processed data (expression, quality information, genomic tracks, and differential expression) that researchers can compare across any set of studies or data characteristics. The data is primarily from mice, comprising 20 studies' worth of transcriptomic data. Processing took roughly 2,340 compute hours. The MECP2pedia portal is an efficient and intuitive way for researchers and clinicians to understand and investigate the universal role of MeCP2 function, which we hope will help guide and inform future work in the field.

Results

Comprehensive resource of MeCP2 transcriptomes

The MECP2pedia portal can be accessed at http://www.mecp2pedia.org/. To generate this resource, we queried the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) (Edgar, Domrachev, and Lash, 2002) on August 7, 2019, for "MeCP2" and then filtered the results for "expression profiling by high throughput sequencing." This search resulted in a preliminary list of 47 GEO entries. We filtered this list for entries dated 2015 or later, and then further filtered for entries with at least two RNA-Seq wild-type samples and a treatment labeled either "knockout", "Rett", "point mutation", "transgenic", "overexpression", or "MeCP2 duplication." We retained 26 GEO entries, which we downloaded and processed with a uniform, streamlined Python pipeline (Figure 1A). In total, our processing yielded 534 Sequence Read Archive (SRA) files, 729 FASTQ files, and 481 BAM (alignment) files. The total disk space used for processing was about 6 TB. The number of mouse samples per study is detailed in SuppFigure 1. For each of the processed samples, we generated raw read quality, alignment quality, and track information. For each study, we collected meta information and data characteristics (Figure 1B), and then we aggregated all samples with matching characteristics into "contrasts" for differential gene analysis. Studies typically contained one to two contrasts, with the exception of four studies that respectively contained 3, 4, 6, and 10 contrasts (bar width in Figure 1B). The 26 GEO entries provided a total of 55 contrasts, with 43 from mouse studies, seven from humans, and five from other species. A list of mouse contrasts and their metadata is provided in TableS1.

Portal users can quickly and easily compare across studies the expression of specific genes of interest. Queries can be carried out individually or as a multi-gene search. Bar and scatter plots are available for each gene to show significance and fold change, and these results can be filtered across the uniformly annotated data characteristics. TPM (Transcripts Per Kilobase Million) is shown to allow the comparison of expression between contrasts (StatQuest, 2015). Users can browse genome tracks by individual study, and studies can be compared to one another. Metadata for each contrast, raw read quality, and alignment quality are all available for each study. A "significant genes" tab allows users to filter genes using an FDR (false discovery rate) and log2 fold change thresholds for each contrast, and this information is also downloadable.

MeCP2 transcriptomics in mice reveals a common core of misregulated genes

Our collected data is noteworthy for its comprehensiveness and heterogeneity (SuppFigure 2A). Tissue, mutation, and cell fraction are all highly variable across the breadth of collected RNA-Seq, giving users the most complete picture possible of MeCP2's transcriptomic role. To understand these changes per study, we examined fold change and FDR from the differential gene expression (DEG) analysis as shown in Figure 2A, the number of significantly upregulated and downregulated genes is generally similar across the 43 contrasts. This finding is consistent with MeCP2's reported role in both gene activation and repression (Chahrour et al., 2008). Furthermore, changes in the majority of the dysregulated genes are less than two-fold. This magnitude of change is low when compared to other mouse disease models, which have substantial quantities of genes with changes greater than two-fold (Al-Ouran et al., 2007).

In a traditional RNA-Seq study with one contrast (a comparison of the expression between two groups), a gene is considered a DEG based on FDR thresholding and/or fold change cutoffs. However, there are many contrasts in our data, and a truly biologically important DEG should be observed consistently in several contrasts. We examined the common FDR thresholds of 10% (0.1), 5% (0.05), and

1% (0.01), and due to MeCP2's low magnitude of gene dysregulation, no fold change cutoff was used. We then examined number of contrast cutoffs of 1 (~2.5% of total contrasts), 4 (~10%), 12 (~30%), and 20 (~50%). When examining the FDR thresholds for a specific contrast cutoff (Figure 2B, SuppFigure 2B, per row), DEGs did not differ much. However, when comparing different numbers of contrasts for a specific FDR threshold (Figure 2B, per column), we observed large differences in DEG numbers. Moreover, when a gene is a DEG in many contrasts, it is likely to be either consistently upregulated or consistently downregulated. This indicates that the number of contrast threshold is a critical filter in establishing a robust set of DEGs. Thus, going forward, we set a strict FDR of 1% and a number of contrasts cutoff (10%) for common core DEGs, resulting in 2,971 genes. Using average fold change, the common core DEGs were sorted into either core up (1,666) or core down (1,305) for further analysis. Mouse core DEG lists are in TableS2.

We next investigated properties unique to the common core DEGs. First, we examined the annotations of these genes in relation to all the genes in the genome (Frankish et al., 2019) and to all expressed genes. We observed that the common core DEGs have a notably high proportion of protein-coding genes. Specifically, while protein-coding genes constitute only 40.8% (21,922 out of 53,661) of all the genes, and 57.6% (18,748 out of 32,551) of the expressed genes, they spike to 94% (2,794 out of 2,971) in the common core DEGs (Figure 2C). Next, due to MeCP2's role in both chromatin structure (Gulmez Karaca, Brito, and Oliveira, 2019) and various epigenetic features (Picard and Fagiolini, 2019), we examined the common core DEGs for positional bias across the mouse genome. Within each chromosome, there are differences in positional distribution between the up and down core DEGs (Figure 2D, SuppFigure 2C). One striking example is on chromosome 8, where Circular Binary Segmentation (CBS) (Olshen et al., 2004) identified a stretch at the beginning of the chromosome (4,375,343 – 49,522,639) with many upregulated genes.

To further examine MeCP2's regulatory role on these common core DEGs in an unbiased manner, we carried out unsupervised Leiden clustering (Traag, Waltman, and van Eck, 2019). From the nine clusters obtained, the two largest clusters (clusters 0 and 1) consisted mainly of the consensus core up and core down genes, respectively (Figure 3A). The clear directional separation of these clusters validates our core DEG selection methodology. Subsequent gene ontology (GO) analysis of these two clusters showed an enrichment in RNA Polymerase II (Pol2) and other transcription-related terms for the upregulated cluster (cluster 0) and an enrichment in neuronal and general nervous system-related terms for the downregulated cluster (cluster 1) (Figure 3B). Both up and downregulated clusters displayed significant enrichment for cell differentiation, signal transduction, and general developmental terms. Our computational approach therefore confirmed not only the roles of MeCP2 as both activator and repressor, but also established the core genes and functions involved.

We further examined the common core DEGs' expression changes using a heatmap with annotation of contrast with cell fraction (Figure 3C). The unsupervised clustering shown on the heatmap categorizes the contrasts into three groups: 1) a mixture of all three types of cell fractions, 2) mainly with nucleus, and 3) mainly with whole cell. The common core DEGs are concordantly changed in about half of the contrasts, which fall into the first group of mixed cell fractions. The two largest gene clusters identified from Leiden clustering are strongly up- and downregulated in this concordant set. Genes have lower expression changes in the second (nucleus) group. Notably, contrasts from the whole cell are categorized into two separate groups. The expression changes of common core DEGs are stronger and concordant in the first group, and then weaker but still concordantly changed in the 17 contrasts of the third group. This is consistent with our findings in SuppFigure 3, in which we observed solid overlap between common core DEGs, and the common cores redefined separately by their sequenced cell fraction. Although the bulk of our data is comprised of whole cell, this picture of the transcriptome is not drastically different from MeCP2 dependent expression in the chromatin or the nucleus.

Cross-species and cross-disease comparisons of MeCP2's transcriptomic signature

As we seek to understand the broad and complex role of MeCP2, mouse data alone is insufficient. Accordingly, we uniformly processed three human data sets yielding seven contrasts. Human data yields fewer DEGs on average than mouse data, but the DEGs have similar ranges in fold change (Figure 4A) and a similar proportion of upregulation versus downregulation. The overlap between DEGs from these seven contrasts is low (SuppFigure 4A), suggesting limited homogeneity in the molecular signature of human MeCP2 dysregulation. This heterogeneity may reflect the fact that four (GSE51607_1-4) out of the seven contrasts are from iPSC cells, whereas the other three are from postmortem brain tissues. This separation is also seen in the heatmap of expression changes of human data on the mouse common core DEGs using unsupervised hierarchical clustering (Figure 4B). This heatmap shows limited qualitative correlation between the direction of human and mouse common core DEG dysregulation.

We also found transcriptome data for rat (*Rattus norvegicus)*, macaque (*Macaca fascicularis)*, and zebrafish (*Danio rerio)* MeCP2 models, which we processed and compared to the mouse common core. We found two rat studies with one contrast each, one monkey study with two contrasts, and one zebrafish study with one contrast. This data yields more DEGs than the human data, with roughly even proportions of up and down gene dysregulation. After associating DEGs to mouse orthologs, we found some overlap between these cross-species models and the mouse common core, with more overlap seen between mouse and rat than with other species (SuppFigure 4B). Figure 4B qualitatively shows a correlation between the mouse core and the rat data as well as some of the monkey data.

To provide a quantitative correlation between these data sets, we performed Gene Set Enrichment Analysis (GSEA). Pre-Ranked analysis was run with our up and down mouse cores as gene sets (Figure 4C). Ranked lists from the 12 non-mouse contrasts were checked for overrepresentation in both up and down cores. We expect contrasts to be positively enriched in the up core and negatively enriched in the down core. The rat model has the best enrichment concordance, with both contrasts enriched as expected. Overall, these contrasts were more significantly negatively enriched in the down core than significantly positively enriched in the up core.

RTT and Autism Spectrum Disorder (ASD) share a range of similar symptoms, including loss of social, cognitive, and language skills. Altered MeCP2 expression is also commonly detected in autism brain samples (Costa, 2014). Therefore, we hypothesized that our MeCP2 common core would display significant overlap with perturbed genes of other established ASD models. To generate an autism common core for comparison to our MeCP2 common core, we explored the expression changes in eight ASD models selected from the Simons Foundation Autism Research Initiative (SFARI) (Siegel et al., 2015) (Figure 4D, TableS3). All experiments involving knockdown or modification of the SFARI mouse model genes were retrieved from the ARCHS4 database. Across the eight target model genes, we processed 223 samples in 18 studies from 15 authors, from which we generated 28 contrasts. In our initial attempts to generate an overall ASD core, we found that expression changes were not generally concordant across contrasts (SuppFigure 4C), which was expected as the contrasts contained a wide range of model genes and experimental procedures.

We thus analyzed the contrasts individually and performed Fisher's exact test to determine the significance of overlap between each contrast and the MeCP2 core. We observed significant overlap in 5 contrasts (Figure 4E), representing 5 studies and 4 model genes (ADNP. ARID1B, CHD8, and SHANK3). We plot the fold change of MeCP2 core genes in these 5 contrasts (SuppFigure 4E). We focused further on three of these contrasts (2 CHD8 and 1 ADNP) with the largest DEG counts and most significant overlap and found that the significant overlap persisted even when considering only genes perturbed in the same direction in the contrast and the MeCP2 core (Figure 4F).

We then carried out GO analysis on the significantly overlapping gene sets. GO analysis of genes upregulated in both the MeCP2 core and a CHD8 contrast reveals significant enrichment for Pol2 related terms, while downregulated genes enrich for nervous system development terms (SuppFigure 4D). This is consistent with our observations for up and down genes in the MeCP2 core. Even with the much smaller set of genes (~400 up and ~500 down in the overlap set compared to ~1600 up and ~1400 down in the MeCP2 core), we observe a similar gene ontology signal.

Sample size has a major impact on DEG detection

MeCP2 interacts with other genes in an expansive manner (Tillotson and Bird, 2019), which lends our transcriptome data a low signal-to-noise ratio (SNR). This contributes to the limited number of clear consensus expression targets, as well as the high degree of discordance in many individual data sets. To increase data detection sensitivity with low SNR, therefore, the number of samples plays an important role. Yet published MeCP2 studies often fail to meet to the conventional recommendation of at least six biological replicates (Schurch et al., 2016). To learn whether this problem limited the DEGs delineated from MeCP2 studies, we performed differential gene analysis on replicate down-sampled subsets of the data set with the highest number of replicates (GSE128178 Contrast 1). We found that the number of replicates has a negative correlation with the number of DEGs detected (Figure 5A). With no fold change cutoff, we could not saturate the number of detected DEGs with as many as ten replicates, which is far more sequencing than found in most published MeCP2 data sets. With a mild fold change cutoff (10% changed), the number of additionally detected DEGs in higher sample counts still did not appear close to saturation. Some saturation and flattening of the power curve began to occur with a 20% fold change cutoff.

To confirm our findings, we repeated the analysis using an RNA-Seq dataset in a psoriatic skin disease model (GSE63979) (Greenberg et al., 2020). We chose this dataset for its high sample size and to understand how different transcriptomic SNRs affect the optimal number of replicates. We found relatively similar patterns (Figure 5B), but a reduced DEG loss effect from fold change thresholding. This finding confirms that the impact of replicate number on detected DEG is not unique to MeCP2 or to disease models with low SNRs.

Trends are also verified across common FDR thresholds, and Rand index is computed between full and down sampled gene sets to understand how much the down sampled results differ from the full results (SuppFigure 5A-B). SuppFigure 5C shows the differences between SNR in MeCP2 and psoriatic skin disease. The low power in MeCP2 data also supports our choice to require our common core DEGs to appear in just 10% of the contrasts.

To examine bias in and created by undetected DEGs in smaller n studies, we first found the supersets of genes comprising each DEG cutoff number. For each DEG, we then computed average absolute log2 fold change (across contrasts). The higher sample number (n) analyses detected DEGs at lower fold changes than analyses with lower sample numbers (Figure 5, last column), demonstrating that DEG sets from different sample sizes are affected differently by fold change cutoffs. A cutoff which removes many genes from a high sample size experiment may remove very few genes from a low sample size experiment. Moreover, the DEGs missed because they have too few biological replicates are those with subtle perturbation, which is especially problematic for disease models with a low SNR. Researchers should be aware of this phenomenon when choosing fold change cutoffs and evaluating results.

Batch and technical variation must be overcome in order to integrate and understand data

Relevant non-biological factors, commonly called batch factors, are often present in a given researcher's data. These batch factors could reflect the use of particular tissues, mouse litters, sequencing platforms, or other variables. Therefore, it is important to integrate and analyze transcriptomic data across years of work with dozens of meta-characteristics. We saw extreme batch effects on the raw count values for all samples included in MECP2pedia, in that the samples were initially segregated by study (Figure 6A). After batch correction and normalizing the raw counts, the segregation was reduced, but samples remained grouped by study. This finding indicates that batch correction and normalization failed to fully resolve this batch effect. When we examined all available meta-characteristics, we observed that the clustering weakly overlapped with the studies' prominent meta-characteristics, such as cell fraction, tissue, and gender (SuppFigure 6A). As a basis for comparison, we performed the same analysis on a set of 316 samples from nine neurological degeneration studies analyzed in Wan et al. (2020). We observed a similar outcome: an extreme batch effect on raw data and an inability of batch effect correction and normalization algorithms to fully remove this batch effect (Figure 6A, SuppFigure 6B).

To better understand the concordance between sexes in MeCP2 models, we compared their molecular signatures. RTT occurs almost exclusively in females, but most MeCP2 studies are carried out in male mice due to their relative ease of use and availability (Ribeiro and MacDonald, 2020). Hence, we had only one fair comparison between male and female models of similar age and tissue (Figure 6B). We found that the DEGs from the male mouse model had no overlap with the female model at one time point, and minimal overlap at a second time point. Furthermore, the direction of dysregulated genes did not show a concordant trend. More data is needed to understand these differences, but researchers should be mindful of sex in their experimental design, especially in RTT.

Discussion

With the low cost and high quality of modern sequencing, the scope of publicly available data is rapidly expanding. But to make the leap from big data to big insights, curation and automation are essential. In addition to gleaning insights from the portal's data, researchers can compare their own novel data to this convenient aggregation of publicly available transcriptome profiles. We plan to add new data sets to the portal and also add a feature allowing users to upload their own processed data for comparison.

MeCP2 transcriptomics from published studies often seems contradictory, perhaps due to low SNR and disparities in experimental design. But by bringing a robust approach to the integration of big data, we uncovered a common core of MeCP2 DEGs with high concordance across studies, suggesting that MeCP2's core function is universal across the examined breadth of tissues, cell fractions, mutations, and mouse strains. The positional bias in common core distribution demonstrates MeCP2's importance to the epigenome, while unbiased clustering further underscores the concordance in core genes, providing insight into their regulatory relationship. When the clusters were enriched to GO terms associated with Pol2 activation and neuronal function, respectively, the two main clusters from the unsupervised clustering correspond to up and downregulated genes. Exploration of the smaller mixed clusters may similarly reveal insights into other proposed mechanisms of MeCP2 action (Lyst and Bird, 2015) (Ip, Mellios, and Sur, 2018).

Examination of diverse MeCP2 models provides quantitative comparisons of MeCP2's dysregulatory molecular signature across species. Robust enrichment cross species in the downregulated DEGs supports the core we have derived, and this finding can be further explored in the context of our delineation on up versus down core genes. Comparison across disease models revealed links between MeCP2 disorders and common ASD models. Specifically, we observed highly significant concordant overlap between genes perturbed in the MeCP2 core and two ASD models: ADNP and CHD8. The RNA Pol2 function in the upregulated genes and neuronal development in the downregulated genes we observed in the MeCP2 core are also enriched genes common to MeCP2, ADNP, and CHD8 models. These overlaps could provide the basis for deeper exploration of the relationship between MeCP2 disorders and other autistic spectrum disorders.

Since small sample size leads to low statistical power, the validity, specificity, and robustness of the DEGs delineated from a single study may be unreliable (Schurch et al., 2016). We validated this concern in our analysis of multiple data sets, and the problem is exacerbated when the SNR is low. Researchers who are interested in perturbations with a small effect should therefore aim to generate large data sets, with 10 or more samples per condition. If resources are limited, six samples per condition would be a good compromise. But most studies failed to meet even this lower threshold, which may explain the lack of consensus conclusions across independent studies and their failure to capture the complete picture of transcriptomic perturbation.

Our expansive study sheds light on the high variability in transcriptomic profiles of a disease model across different tissues, ages, sexes, species, and other biological and technical artifacts. The specific experimental conditions of a single study therefore cannot capture the complete picture of transcriptomic changes. Individual researchers will always be limited in the data that they can personally generate. Our big data integration platform solves this problem, making it invaluable for scientists studying complex diseases such as MeCP2.

Acknowledgments

We are grateful to members of the Liu lab and Dr. Hyun-Hwan Jeong for helpful discussions. We are also grateful to members of the Zoghbi lab and Dr. Huda Y. Zoghbi for helpful discussions. This project is supported by the Jan and Dan Duncan Neurological Institute at Texas Children's Hospital, and a grant from National Institute of General Medical Sciences R01-GM120033. Z.L. and L.L is supported by the Chao Foundation and Hamill Foundation. Funded in part with federal funds from the United States Department of Agriculture (USDA/ ARS) under Cooperative Agreement No. 58-3092-0-001 and NRI Zoghbi Scholar Award (to H.K.Y.).

Author Contributions

Conceptualization, A.J.T., S.-Y.K., Z.L., Y.-W.W., H.K.Y., R.A.-O.; Methodology and Data Curation, A.J.T., R.A.-O., L.L., J.W., H.K.Y., Y.-W.W., Z.L.; Formal Analysis, A.J.T., L.L., J.W.; Writing – Original Draft, A.J.T., L.L., Y.-W.W.; Writing – Review & Editing, A.J.T., L.L., Y.-W.W., S.-Y.K., J.W., R.A.-O., H.K.Y., Z.L.; Portal development, S.-Y.K., A.J.T.

Declaration of Interests

The authors declare no competing interests.

STAR Methods:

Data Collection

To generate this resource, we queried NCBI GEO (https://www.ncbi.nlm.nih.gov/gds) on August 7, 2019 for "MeCP2" and then filtered the results for "expression profiling by high throughput sequencing." This search resulted in a preliminary list of 47 GEO entries. We filtered this list for entries dated 2015 or later, and then further filtered for entries with at least two RNA-Seq wild-type samples each and a treatment labeled either "Knock out", "Rett", "Point Mutation", "Transgenic", "Overexpression", or "MeCP2 Duplication." We retained 26 GEO entries. Data sets with no associated publications were included. We subsequently added three more studies, GSE123941, GSE128178, and GSE123372, based on the scope and relevance of their data. All preprocessing was carried out with a uniform, streamlined Python pipeline. SRA files were downloaded with prefetch from SRAtoolkit.2.9.6-1-

centos_linux64 (Sherry, 2012) and then converted to fastq with fasterq-dump version 2.3.5, using the --split-files option. Fastq files were then checked for quality with FastQC version 0.11.7 (Andrews, 2010).

Mouse Data Processing

Mouse samples were aligned to GENCODE GRCm38p6 primary assembly, version 18 (https://www.gencodegenes.org/mouse/release_M18.html), with STAR version 2.6.0a (Dobin et al., 2013) at default parameters. The assembly also contained an appended copy of human MeCP2 from hg38. BigWig files were generated with bamCoverage version 3.3.1 from deepTools (Ramírez et al., 2016). We assessed alignment quality with RSeQC geneBody_coverage and read_distribution, both version 3.0.0. Overall quality per study was examine with MultiQC v1.7 (Ewels et al., 2016). DEG analysis was performed in R version 3.5.2 (Eggshell Igloo) with DESeq version 1.24.0, after loose expression filtering (per contrast, a gene must have a sum of 10 counts in at least half the samples).

Data was not trimmed except for samples SRR3679844, SRR3679845, SRR3679848, SRR3679849, SRR3679852, and SRR3679853 from GSE83474, due to slight anomalies in their raw sequences. The trim was performed with Trimmomatic-0.36 (Bolger, Lohse, and Usadel, 2014) using the following parameters: PE ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 HEADCROP:8

Data Annotation

We downloaded SRA run tables for each GEO entry. Data characteristics of interest were: genotype, organism, experiment, run, sample name, cellular fraction, strain, age, cell line, cell type, tissue, sex, mutation, and disease. Incomplete run tables were filled in from the contents of their publications, if available. After processing, we annotated samples for sequencing depth and contrasts for number of DEG at FDR < 0.01 with no fold change cutoff.

Data Visualization

Unless otherwise specified, plots were made in R with ggplot2 version 3.2.1 (Wickham, 2016).

Data and Code Availability

All datasets are available from GEO under their specified accession numbers. Processed mouse results are available through our portal, and other processed results are available as supplements.

Portal Development

Python pipeline analysis results and GEO sample information were parsed and saved using the MongoDB NoSQL database. The web server was written in JavaScript and serves an API that gives access to the data and the web portal application. Data visualization uses the D3.js library and IGV.js (Robinson et al., 2020) for genomics tracks.

Core Gene Identification and Clustering

First, fold changes with FDR > 0.01 were set to zero. Contrasts with all non-significant fold changes were removed. To generate our set of core genes, we kept only genes with non-zero fold change in four or more contrasts (10%). This resulted in a set of 2,971 genes for further analysis. For consistency in the analysis and the visualization of gene regulation direction, we inverted the direction of fold change for the four contrasts of the MDS model (GSE123372_3, GSE66870_2, GSE71235_1, GSE71235_2).

We also identified alternative core genes based on contrast metadata characteristics. For each type of cell fraction (chromatin, nucleus, whole cell) and the cortex and forebrain tissue types, we performed the same core gene identification filtering as we did for the main core. We considered only the contrasts with the metadata feature of interest and kept only genes with significant and non-zero fold change in at least 10% of the contrasts. We have included an upset plot generated with the R package UpSetR (Conway, Lex, Gehlenborg, 2017) as SuppFigure 3.

After identifying the set of core genes, we assigned unsupervised clusters using the Scanpy (Wolf, Angerer, and Theis, 2018) implementation of the Leiden algorithm with the parameters: number of neighbors=45 and resolution=0.5. Then, we generated UMAP coordinates with the parameters: number of neighbors=45, minimum distance=0.1, and spread=10. We then used UMAP's Scanpy implementation to generate plots of the unsupervised clusters, as well as up and downregulation.

Core Gene Characteristics and Location

We annotated mouse genes with GRCm38p6 primary, version 18 from GENCODE. Genes were sorted into eight super-categories to show broad function. Expressed genes (32,539) is a superset of the genes that pass the expression filter in any contrast.

Core genes were plotted by their TSS. Chromosome 8 was plotted with an equivalent number of randomly drawn non-core genes (150) to show the strength of its regional core up DEG enrichment. We validated this trend with the CBS algorithm, implemented with R Package PSCBS version 0.65.0. The core up, core down, and non-core genes were respectively assigned values of 6, 0, and 3 for segmentation detection and plotting.

GO Analysis

GO analysis was performed using the Python GOAtools package (Klopfenstein et al., 2018). We performed an enrichment analysis on each MeCP2 Leiden cluster, using all NCBI protein coding mouse genes as the background set. For each Leiden cluster, we retained the top 6 biological process GO terms by frequency. We used the same methodology to perform GO enrichment on overlapping MeCP2 and ASD core genes, also retaining the top 6 biological process GO terms by frequency.

Human Data Processing and Comparative Analysis

Human data was aligned to GRCh38p12 primary assembly, version 28 from GENCODE with STAR. BigWig generation, assembly quality metrics, and DEG analysis were performed identically on mouse data. Human genes were then queried for their orthology to mouse genes with DIOPT 8.0 (Hu et al., 2011), using the "return only best match" option. Upset plots were made with function upset from Package UpSetR, version 1.4.0. Human and mouse metadata is available in TableS5.

Other Model Data Processing and Comparative Analysis

Rat data was aligned to the Rnor_6.0 toplevel assembly and annotated with Rnor_6.0.99 from Ensembl (Yates et al., 2020). Zebrafish data was aligned to Danio_rerio.GRCz11 toplevel assembly, and annotated with Danio_rerio.GRCz11.100 from Ensembl. The orthology tables for rat and zebrafish genes were retrieved from DIOPT as done for human data. Macaque data was aligned to Macaca_fascicularis_5.0 and annotated with Macaca_fascicularis_5.0.100 from Ensembl. Macaque gene orthology data was retrieved with the function getBM from R package biomaRt, version 2.38.0 (Durinck et al., 2009). The only data trimmed was GSE57974, using Trimmomatic-0.36 with the following parameters: LEADING:3 TRAILING:20 MINLEN:50. Genotype labels for GSE87855 were inferred based on MeCP2 level.

Heatmaps are made with pheatmap version X (Kolde and Kolde, 2015) with log2 fold change no clustering on rows and clustering_distance_cols = "euclidean". For consistency in the analysis and the visualization of gene regulation direction, we inverted the direction of log2 fold change for the MDS contrast (GSE57974_1)

GSEA

GSEA (Subramanian et al., 2005) (Mootha, Lindgren, and Eriksson, 2003) version 4.1.0 Pre-Ranked was run with default parameters besides set_max of 100,000 and set_min of 1. Ranking values were computed per gene as –log10(adjusted P value) * log2 fold change. For consistency in the analysis and the visualization of gene regulation direction, we inverted the direction of normalized enrichment score fold change for the MDS contrast (GSE57974_1). GSEA results are available as TableS6.

ASD Model Comparison

All experiments involving knockdown or modification of the SFARI mouse model genes were retrieved from the ARCHS4 database. There were 18 such studies, which we processed using DESeq2 to generate DEGs across 28 contrasts. We retained only DEGs with FDR <0.01. With the DEGs for each ASD contrast, we used the hypergeometric and Fisher's exact tests to determine the significance of overlap with all the set of all MeCP2 core genes and also both up and downregulated subsets. Computed ASD contrast fold change is available as TableS4.

We also used the previously described process to perform GO analysis on genes in the intersection of the ASD contrasts and MeCP2 cores.

Down Sampling Analysis

MeCP2 data is from GSE128178. All 10 samples of wild-type and knockout whole cell data were randomly selected to create 100 random drawings at each different sample number. (For instance, sample 9 was one random sample removed from each genotype, and so on.) Once selected, samples were normalized with each other and analyzed with the same DEG methodology detailed above. The Rand index was then computed using the rand.index function in R (fossil, version 0.4.0) (Vavrek, 2011). Vectors indicating if each gene was a DEG for a particular run were compared to the vectors of DEGs for the complete contrast of 10 wild-type and 10 knockout samples, respectively.

Psoriatic skin data is from GSE63979 (SRP050971). This study contains the total RNA-Seq data of nine normal skin samples, nine lesional psoriatic samples, and 27 uninvolved psoriatic samples. In order to conduct the downsampling analysis between two groups with the same sample number, only normal skin samples and lesional psoriatic samples were chosen. For each phenotype, all nine samples were randomly selected to create 100 random drawings at each different sample number. (For instance, sample 8 was one random sample removed from each phenotype, and so on.) The DEG analysis and Rand index comparison on psoriatic skin data was the same used for MeCP2 data.

Technical Variation / Batch Effect Analysis

All raw MeCP2 mouse expression value data was dimensionally reduced using UMAP version 0.2.6.0 in R and plotted with color for contrast of origin. ComBat_seq from R package sva, version 3.36.0, was then run with contrast as batch to deconvolute the data. ComBat_seq-normalized data was then size factor-normalized with DESeq and plotted again.

To provide an alternate dataset to evaluate batch effect, we used a set of Alzheimer's disease data stored on Synapse from Wan, et al. (2020). We retrieved raw count data and plotted the samples using UMAP. Count data was normalized with DESeq, and then we used the ComBat_seq function to attempt to control for batch effect.

Sex Comparison

Compared contrasts are GSE90736_1, GSE90736_2, and GSE66211_1. Genes are considered DEGs if they pass FDR < 0.01. Plot is made with R package *VennDiagram* version 1.6.20.

Works Cited

1. Lonsdale, J., Thomas, J., Salvatore, M. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45,** 580–585 (2013). https://doi.org/10.1038/ng.2653
2. Lachmann, A., Torre, D., Keenan, A.B. *et al.* Massive mining of publicly available RNA-seq data from human and mouse. *Nat Commun* **9,** 1366 (2018). https://doi.org/10.1038/s41467-018-03751-6
3. Drysdale, R. A., Crosby, M. A., & FlyBase Consortium (2005). FlyBase: genes and gene models. *Nucleic acids research*, *33*(Database issue), D390–D395. https://doi.org/10.1093/nar/gki046
4. Smith, J. R., Hayman, G. T., Wang, S. J., Laulederkind, S., Hoffman, M. J., Kaldunski, M. L., Tutaj, M., Thota, J., Nalabolu, H. S., Ellanki, S., Tutaj, M. A., De Pons, J. L., Kwitek, A. E., Dwinell, M. R., & Shimoyama, M. E. (2020). The Year of the Rat: The Rat Genome Database at 20: a multi-species knowledgebase and analysis platform. *Nucleic acids research*, *48*(D1), D731–D742. https://doi.org/10.1093/nar/gkz1041
5. Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology*, *19*(1A), A68–A77. https://doi.org/10.5114/wo.2014.47136
6. Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C., & Schultz, N. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery*, *2*(5), 401–404. https://doi.org/10.1158/2159-8290.CD-12-0095
7. Amir, R. E., Van den Veyver, I. B., Wan, M., Tran, C. Q., Francke, U., & Zoghbi, H. Y. (1999). Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nature genetics*, *23*(2), 185–188. https://doi.org/10.1038/13810
8. Sandweiss, A. J., Brandt, V. L., & Zoghbi, H. Y. (2020). Advances in understanding of Rett syndrome and MECP2 duplication syndrome: prospects for future therapies. *The Lancet. Neurology*, *19*(8), 689–698. https://doi.org/10.1016/S1474-4422(20)30217-9
9. Sampieri, K., Meloni, I., Scala, E., Ariani, F., Caselli, R., Pescucci, C., Longo, I., Artuso, R., Bruttini, M., Mencarelli, M. A., Speciale, C., Causarano, V., Hayek, G., Zappella, M., Renieri, A., & Mari, F. (2007). Italian Rett database and biobank. *Human mutation*, *28*(4), 329–335. https://doi.org/10.1002/humu.20453
10. Percy, A. K., Lane, J. B., Childers, J., Skinner, S., Annese, F., Barrish, J., Caeg, E., Glaze, D. G., & MacLeod, P. (2007). Rett syndrome: North American database. *Journal of child neurology*, *22*(12), 1338–1341. https://doi.org/10.1177/0883073807308715
11. Krishnaraj, R., Ho, G., & Christodoulou, J. (2017). RettBASE: Rett syndrome database update. *Human mutation*, *38*(8), 922–931. https://doi.org/10.1002/humu.23263

12. Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research*, *30*(1), 207–210. https://doi.org/10.1093/nar/30.1.207
13. Chahrour, M., Jung, S. Y., Shaw, C., Zhou, X., Wong, S. T., Qin, J., & Zoghbi, H. Y. (2008). MeCP2, a key contributor to neurological disease, activates and represses transcription. *Science (New York, N.Y.)*, *320*(5880), 1224–1229. https://doi.org/10.1126/science.1153252
14. Al-Ouran, R., Wan, Y. W., Mangleburg, C. G., Lee, T. V., Allison, K., Shulman, J. M., & Liu, Z. (2019). A Portal to Visualize Transcriptome Profiles in Mouse Models of Neurological Disorders. *Genes*, *10*(10), 759. https://doi.org/10.3390/genes10100759
15. Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., García Girón, C., … Flicek, P. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic acids research*, *47*(D1), D766–D773. https://doi.org/10.1093/nar/gky955
16. Gulmez Karaca, K., Brito, D., & Oliveira, A. (2019). MeCP2: A Critical Regulator of Chromatin in Neurodevelopment and Adult Brain Function. *International journal of molecular sciences*, *20*(18), 4577. https://doi.org/10.3390/ijms20184577
17. Picard, N., & Fagiolini, M. (2019). MeCP2: an epigenetic regulator of critical periods. *Current opinion in neurobiology*, *59*, 95–101. https://doi.org/10.1016/j.conb.2019.04.004
18. Olshen, A. B., Venkatraman, E. S., Lucito, R., & Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics (Oxford, England)*, *5*(4), 557–572. https://doi.org/10.1093/biostatistics/kxh008
19. Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports*, *9*(1), 5233. https://doi.org/10.1038/s41598-019-41695-z
20. Siegel, M., Smith, K. A., Mazefsky, C., Gabriels, R. L., Erickson, C., Kaplan, D., Morrow, E. M., Wink, L., Santangelo, S. L., & Autism and Developmental Disorders Inpatient Research Collaborative (ADDIRC) (2015). The autism inpatient collection: methods and preliminary sample description. *Molecular autism*, *6*, 61. https://doi.org/10.1186/s13229-015-0054-8
21. Tillotson, R., & Bird, A. (2019). The Molecular Basis of MeCP2 Function in the Brain. *Journal of molecular biology*, S0022-2836(19)30595-9. Advance online publication. https://doi.org/10.1016/j.jmb.2019.10.004
22. Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G. G., Owen-Hughes, T., Blaxter, M., & Barton, G. J. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?. *RNA (New York, N.Y.)*, *22*(6), 839–851. https://doi.org/10.1261/rna.053959.115
23. Greenberg, E. N., Marshall, M. E., Jin, S., Venkatesh, S., Dragan, M., Tsoi, L. C., Gudjonsson, J. E., Nie, Q., Takahashi, J. S., & Andersen, B. (2020). Circadian control of interferon-sensitive gene

expression in murine skin. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(11), 5761–5771. https://doi.org/10.1073/pnas.1915773117

24. Zhang, Y., Parmigiani, G., & Johnson, W. E. (2020). *ComBat-seq*: batch effect adjustment for RNA-seq count data. *NAR genomics and bioinformatics*, *2*(3), lqaa078. https://doi.org/10.1093/nargab/lqaa078
25. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8
26. Wan, Y. W., Al-Ouran, R., Mangleburg, C. G., Perumal, T. M., Lee, T. V., Allison, K., Swarup, V., Funk, C. C., Gaiteri, C., Allen, M., Wang, M., Neuner, S. M., Kaczorowski, C. C., Philip, V. M., Howell, G. R., Martini-Stoica, H., Zheng, H., Mei, H., Zhong, X., Kim, J. W., … Logsdon, B. A. (2020). Meta-Analysis of the Alzheimer's Disease Human Brain Transcriptome and Functional Dissection in Mouse Models. *Cell reports*, *32*(2), 107908. https://doi.org/10.1016/j.celrep.2020.107908
27. Ribeiro, M. C., & MacDonald, J. L. (2020). Sex differences in Mecp2-mutant Rett syndrome model mice and the impact of cellular mosaicism in phenotype development. *Brain research*, *1729*, 146644. https://doi.org/10.1016/j.brainres.2019.146644
28. Lyst, M. J., & Bird, A. (2015). Rett syndrome: a complex disorder with simple roots. *Nature reviews. Genetics*, *16*(5), 261–275. https://doi.org/10.1038/nrg3897
29. Ip, J., Mellios, N., & Sur, M. (2018). Rett syndrome: insights into genetic, molecular and circuit mechanisms. *Nature reviews. Neuroscience*, *19*(6), 368–382. https://doi.org/10.1038/s41583-018-0006-3
30. Sherry, S., Xiao, C., Durbrow, K., Kimelman, M., Rodarmer, K., Shumway, M., & Yaschenko, E. (2012, January). NCBI sra toolkit technology for next generation sequence data. In *Plant and Animal Genome XX Conference (January 14-18, 2012). Plant and Animal Genome*.
31. Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
32. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635
33. Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F., & Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic acids research*, *44*(W1), W160–W165. https://doi.org/10.1093/nar/gkw257
34. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170
35. Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

36. James T. Robinson, Helga Thorvaldsdóttir, Douglass Turner, Jill P. Mesirov. igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). bioRxiv 2020.05.03075499.
37. Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome biology*, *19*(1), 15. https://doi.org/10.1186/s13059-017-1382-0
38. Olshen, A. B., Bengtsson, H., Neuvial, P., Spellman, P. T., Olshen, R. A., & Seshan, V. E. (2011). Parent-specific copy number in paired tumor-normal studies using circular binary segmentation. *Bioinformatics (Oxford, England)*, *27*(15), 2038–2046. https://doi.org/10.1093/bioinformatics/btr329
39. Klopfenstein, D. V., Zhang, L., Pedersen, B. S., Ramírez, F., Warwick Vesztrocy, A., Naldi, A., Mungall, C. J., Yunes, J. M., Botvinnik, O., Weigel, M., Dampier, W., Dessimoz, C., Flick, P., & Tang, H. (2018). GOATOOLS: A Python library for Gene Ontology analyses. *Scientific reports*, *8*(1), 10872. https://doi.org/10.1038/s41598-018-28948-z
40. Hu, Y., Flockhart, I., Vinayagam, A., Bergwitz, C., Berger, B., Perrimon, N., & Mohr, S. E. (2011). An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC bioinformatics*, *12*, 357. https://doi.org/10.1186/1471-2105-12-357
41. R Conway, J, R. Lex, A. Gehlenborg, N. (2017) UpSetR: An R Package for the Visualization of Intersecting Sets and their Properties. https://doi.org/10.1093/bioinformatics/btx364
42. Vavrek MJ (2011). “fossil: palaeoecological and palaeogeographical analysis tools.” *Palaeontologia Electronica*, **14**(1), 1T. R package version 0.4.0. Available online at https://CRAN.R-project.org/package=fossil
43. Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Marugán, J. C., Cummins, C., Davidson, C., Dodiya, K., Fatima, R., Gall, A., Giron, C. G., … Flicek, P. (2020). Ensembl 2020. *Nucleic acids research*, *48*(D1), D682–D688. https://doi.org/10.1093/nar/gkz966
44. Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature protocols*, *4*(8), 1184–1191. https://doi.org/10.1038/nprot.2009.97
45. Costa, F. F. (2014). Big data in biomedicine. *Drug discovery today*, *19*(4), 433-440. https://doi.org/10.1016/j.drudis.2013.10.012
46. Hodes, R. J., & Buckholtz, N. (2016). Accelerating medicines partnership: Alzheimer’s disease (AMP-AD) knowledge portal aids Alzheimer’s drug discovery through open data sharing. *Expert opinion on therapeutic targets*, *20*(4), 389-391. https://doi.org/10.1517/14728222.2016.1135132
47. StatQuest (Ed.). (2015, July 22). *RPKM, FPKM and TPM, clearly explained: RNA-Seq Blog*. rna-seqblog. https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/.
48. Ewels, P. Magnusson, M. Lundin, S. Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report, *Bioinformatics*, Volume 32, Issue 19, 1 October 2016, Pages 3047–3048, https://doi.org/10.1093/bioinformatics/btw354

49. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., ... & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545-15550. https://doi.org/10.1073/pnas.0506580102
50. Mootha, V., Lindgren, C., Eriksson, KF. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* **34,** 267–273 (2003). https://doi.org/10.1038/ng1180
51. Kolde, R., & Kolde, M. R. (2015). Package ‘pheatmap’. *R package*, *1*(7), 790. https://cran.r-project.org/package=pheatmap

Figure Legends

Figure 1: Overview of data and workflow

A. Workflow for portal data and analysis. Processing is uniform and unbiased. Quality, track, and DEG analysis results are available in an intuitive and comparable manner through our portal.

B. Sankey plot on major characteristics per contrast of the collected mouse data (date, cell fraction, strain, tissue, first author). Metadata was collected with sequence data and then standardized.

Figure 2: Mouse transcriptome common core

A. Distribution of log2 fold change across contrasts with significant (FDR < 0.01) DEGs. Dark blue and dark red respectively indicate genes that are core down and core up, and pale blue/pale red respectively indicate down and up DEG. Pie charts with the same annotation colors show what percentage of each contrast's DEGs falls into each category. Stacked bar charts with the same annotation colors show each contrast's DEG quantity.

B. Histograms of significantly up and downregulated genes cut for different FDR thresholds and the number of total contrasts in which a DEG appears. Genes at the extreme ratios of 0 or 1 percent upregulated are highly concordant across contrasts, whereas genes that fall into the middle are discordant. For consistency in this analysis, we inverted the direction of fold change for the four contrasts of the TG model.

C. Genes as annotated by Gencode and condensed into eight broad categories. We consider 53,661 genes from our annotation and find 32,539 that pass our expression filter in at least one contrast. The common core (FDR < 0.01 in at least four contrasts) is comprised of 2,971 genes.

D. Exploration of genome location trends in the common core. All non DEG are plotted in the upper portion of the panel, and violin plots show areas of gene density. Chromosome 8 was selected for further examination in the lower half of the panel, with baseline genes in equal quantity to the core genes (150) also plotted. CBS method is used to identify trends in the up/down/baseline genes. The bands on middle dot plot show these CBS results, and the lower stack plot shows the density of up, down, and baseline genes on chromosome 8.

Figure 3: Unsupervised clustering on core genes

A. The left UMAP plot colors each gene by cluster, assigned through unsupervised Leiden clustering. The right UMAP plot displays the percentage of contrasts in which the gene was upregulated on a spectrum of red (upregulated in all contrasts) to blue (downregulated in all contrasts). We can see that the green and orange clusters roughly encompass the up and downregulated genes.

B. Results of GO analysis on Leiden clusters 0 and 1. The bar colors correspond to cluster. Bar length represents the proportion of genes enriched with the term in the cluster and the line plot represents the FDR of the enrichment.

C. Heatmap of contrasts (columns) by genes (rows). Contrasts are labelled based on the experiment's cell fraction, and genes are labelled based on their Leiden cluster. We can see the general downregulation in the orange cluster and the upregulation in the green cluster. We can also see from this figure that the other clusters are generally caused by extreme deviations in one or two studies.

Figure 4: Mouse transcriptome translation to other models

A. Distribution of log2 fold change across contrasts with significant (FDR < 0.01) DEGs. Blue and red respectively indicate down and up DEGs. Pie charts with the same annotation coloring show the percentage of each contrast's DEGs in each category. Stacked bar charts with the same annotation color show each contrasts' DEG quantity. Upper 7 contrasts are human data, lower 5 are other species.

B. Heatmaps plotted to compare direction of dysregulation to the consensus from mouse data. Genes examined are the mouse common core, and plots are annotated on mouse core down and mouse core up.

C. Per contrast visualization of GSEA normalized enrichment score and FDR. Direction and color of line represents normalized enrichment score and point size represents -log10(FDR). Contrasts are grouped and shaded corresponding to their model of origin. MDS model is annotated with a small star.

D. Sankey plot of ASD contrast metadata characteristics. From left to right: First Author, Tissue, Strain, Target Gene, Experimental Procedure.

E. Fisher's exact test results. Points sized by –log10(p-value), length determined by odds ratio, data colored by gene. Points are opaque and overlap to MeCP2 core is considered significant if the Fisher p-value is less than 0.05.

F. Pie charts show the magnitude of overlap between selected ASD contrasts and the MeCP2 common core. Down and up only show genes changed in the same direction in both sets. P-values beneath each plot show the Fisher's exact test significance of the overlap for each intersection, colored red if the p-value is less than 0.05.

Figure 5: Downsampling analysis

A-B. Large contrasts from MeCP2 and lesional psoriatic skin data was downsampled to smaller and more common experimental n and DEG analysis was run on these subsets. Box plot and jitter points are plotted for resultant DEG numbers under each condition. Cutoffs for MeCP2 are sample numbers 9 through 3. Cutoffs for psoriatic data are sample numbers 8 through 3. Each cutoff number was repeated 100 times, with random samples discarded each time. Results are plotted at FDR < 0.01. MeCP2 data is fold change (FC) cutoff at any FC, FC > 10%, and FC > 20%. Psoriatic skin data is FC cutoff at any FC, FC > 10%, and FC > 200%. Curves indicate the percent of DEGs remaining at continuous |log2 fold change| cutoffs. Horizontal line indicates 50% of genes removed.

Figure 6: Technical variation / batch effect analysis

A. UMAP visualization of raw and corrected data from MeCP2 and AD.

B. Comparison of cell type-matched (microglia), male-to-female contrasts by DEG overlap and direction of misregulation. DEG are FDR < 0.01 and any fold change.

Supp Figure 1: Mouse sample number

Bar plot of number of mouse samples by year and first author.

Supp Figure 2: Expanded examination of common core characteristics

A. Bar plots of number of DEGs from each contrast, colored and annotated by tissue, cell fraction, and model. No strong bias is observed for any of these characteristics by DEG number.

B. Histograms of significantly up- and downregulated genes cut for different FDR thresholds and the number of total contrasts in which a DEG appears. Genes at the extreme ratios of 0 or 1 for percent upregulated are highly concordant across contrasts, whereas genes that fall into the middle are discordant. For consistency in this analysis, we inverted the direction of fold change for the four contrasts of TG model.

C. Exploration of genome location trends in the common core. Box plot, violin plot, and jittered dots are plotted for genome coordinate of TSS for each core DEG. Chromosome 8 was selected for further analysis, with random baseline genes added for detection of robust trends.

Supp Figure 3: Overlap between specific common core designations

Upset plot to examine overlap between alternative core designations (chromatin, nucleus, whole cell, cortex-forebrain) and the common core. There is some overlap with all cores, demonstrating the absence of a strong bias in the types of data that contribute genes to the common core.

Supp Figure 4: Mouse transcriptome translation to other models

A. Upset plot to examine overlap between the DEG lists from human contrasts.

B. Upset plot to examine overlap between the DEG lists from other species contrasts.

C. Heatmap of genes (rows) by contrasts (columns). Contrasts are labelled by ASD model gene.

D. Gene ontology enrichment on specific overlapping gene sets. Bar length represents the proportion of genes enriched with the term in the cluster and the line plot represents the FDR of the enrichment.

E. Heatmaps plotted to compare the direction of dysregulation to the consensus from mouse data. Genes examined are the mouse common core, and plots are annotated on mouse common down and mouse common up. Only ASD contrasts with significant overlap to MeCP2 core are included.

Supp Figure 5: Downsampling analysis

A. DEG analysis run on subsets of n = 10 from the MeCP2 data set. Cutoffs are sample numbers 9 through 3. Each cutoff number was repeated 100 times, with random samples discarded each time. Downsampled MeCP2 data at FDR thresholds < 0.01, 0.05, and 0.1. Rand index calculated for each downsampled result against full (n = 10) sample space.

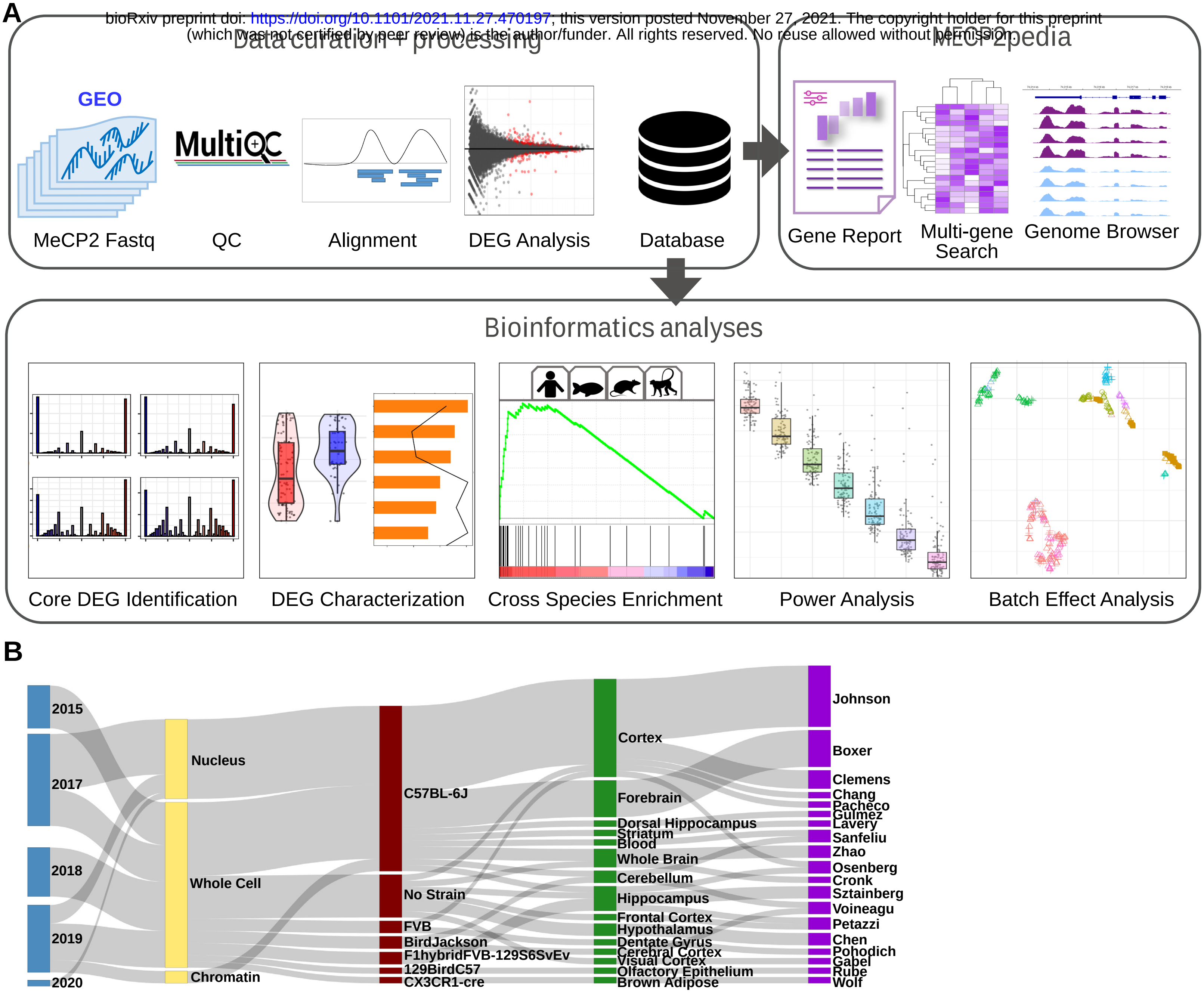
B. DEG analysis run on subsets of n = 9 psoriatic skin data set. Cutoffs are sample numbers 8 through 3. Each cutoff number was repeated 100 times, with random samples discarded each time. Downsampled MeCP2 data at FDR thresholds < 0.01, 0.05, and 0.1. Rand index calculated for each downsampled result against full (n = 9) sample space.

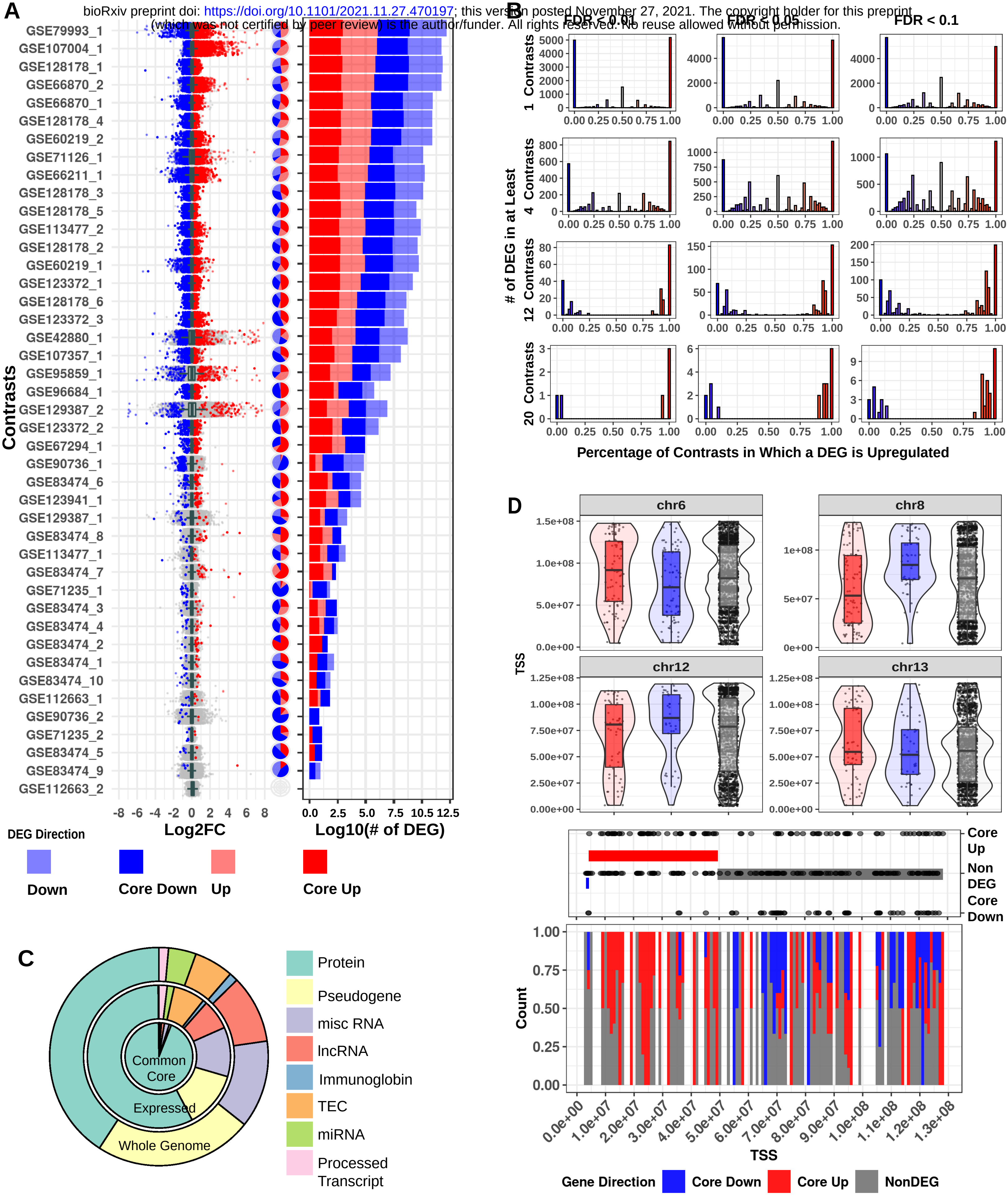
C. PCA, MAplot, eCDF, and distribution of fold change for both MeCP2 and psoriatic skin data.

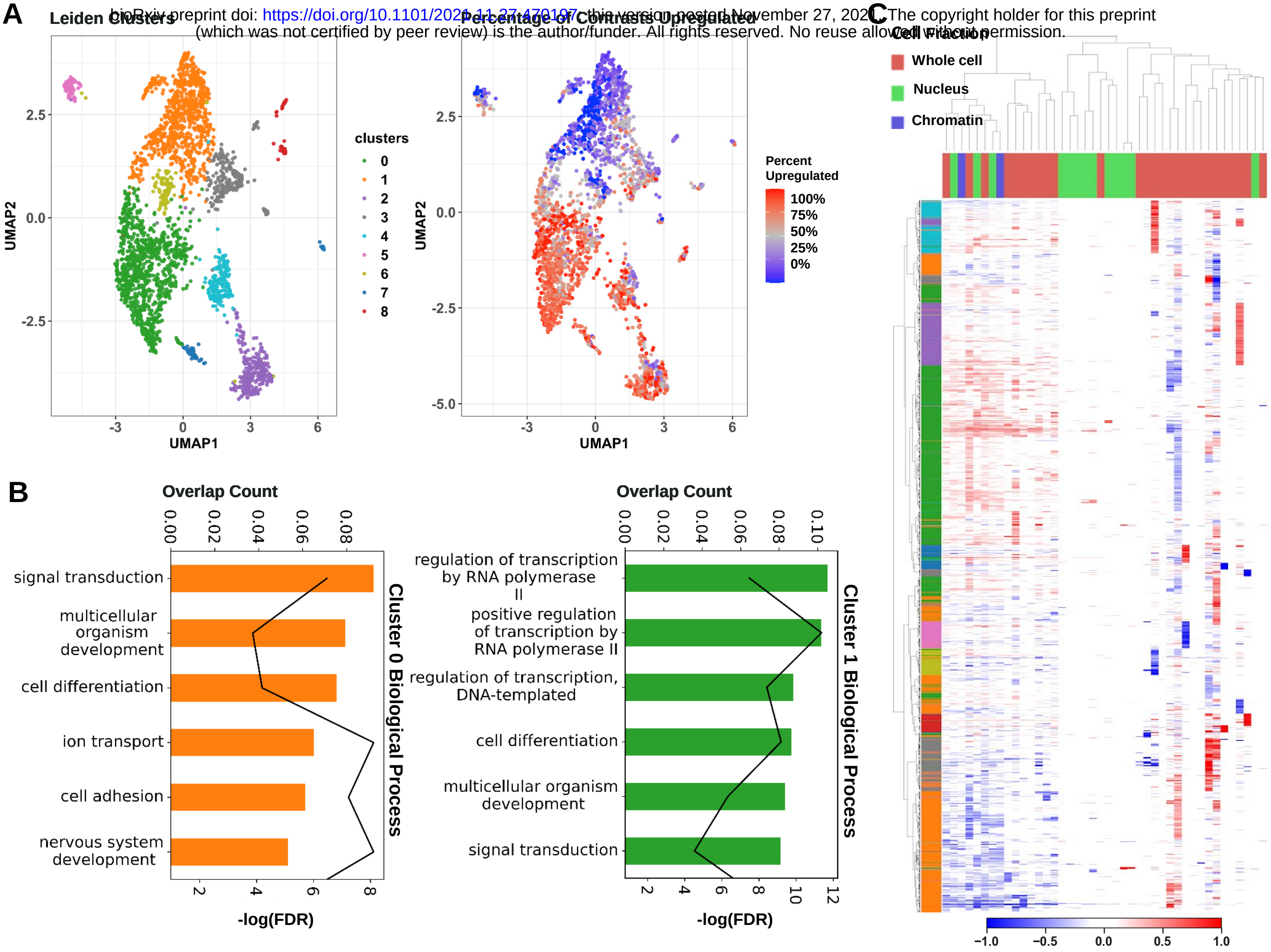
Supp Figure 6: Batch effect UMAP visualization

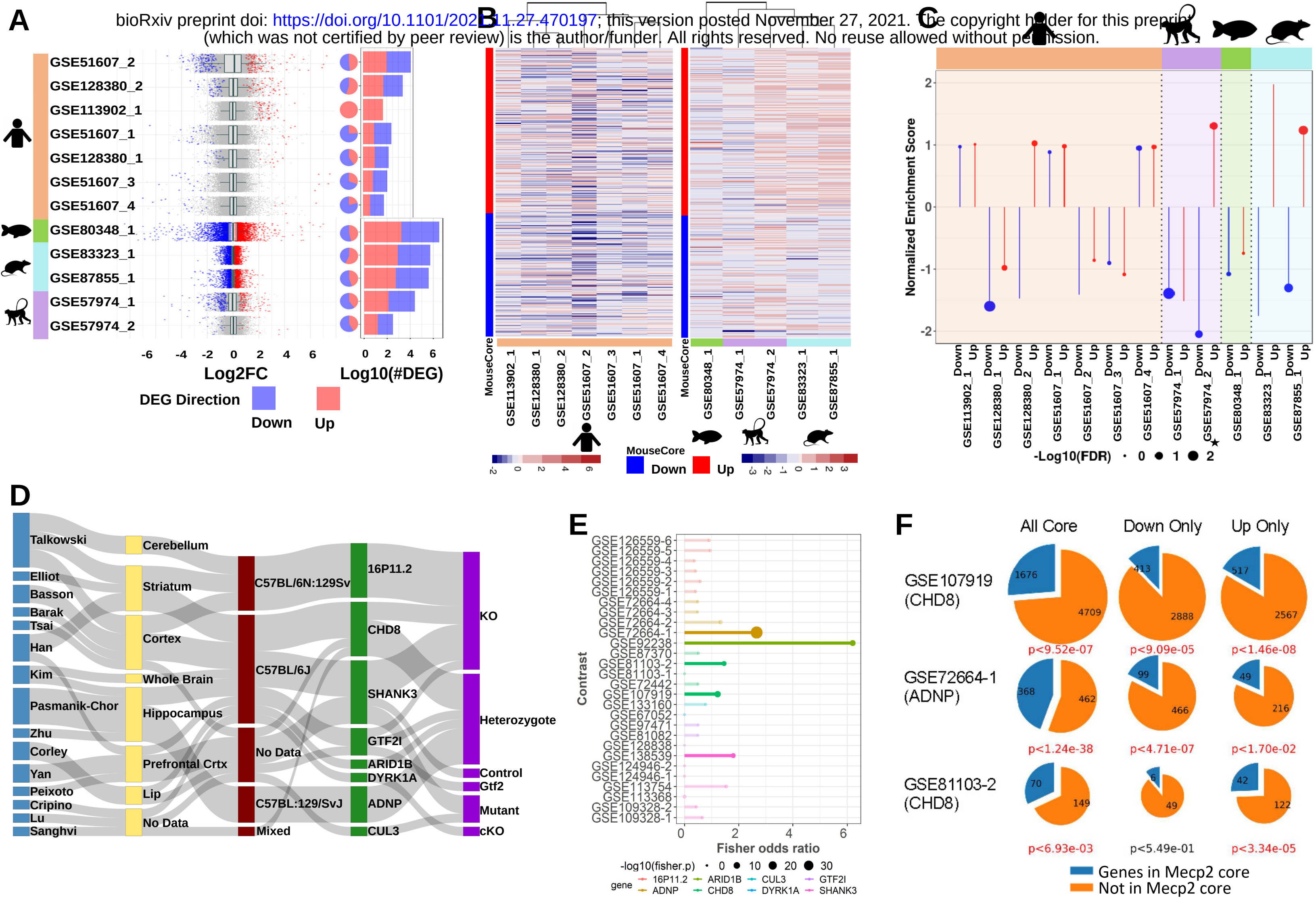
A. UMAP on normalized and batch-corrected data from MeCP2 mouse data, colored for meta-characteristics. Some show good clustering on the characteristic (cell fraction) and some do not (sequencing depth).

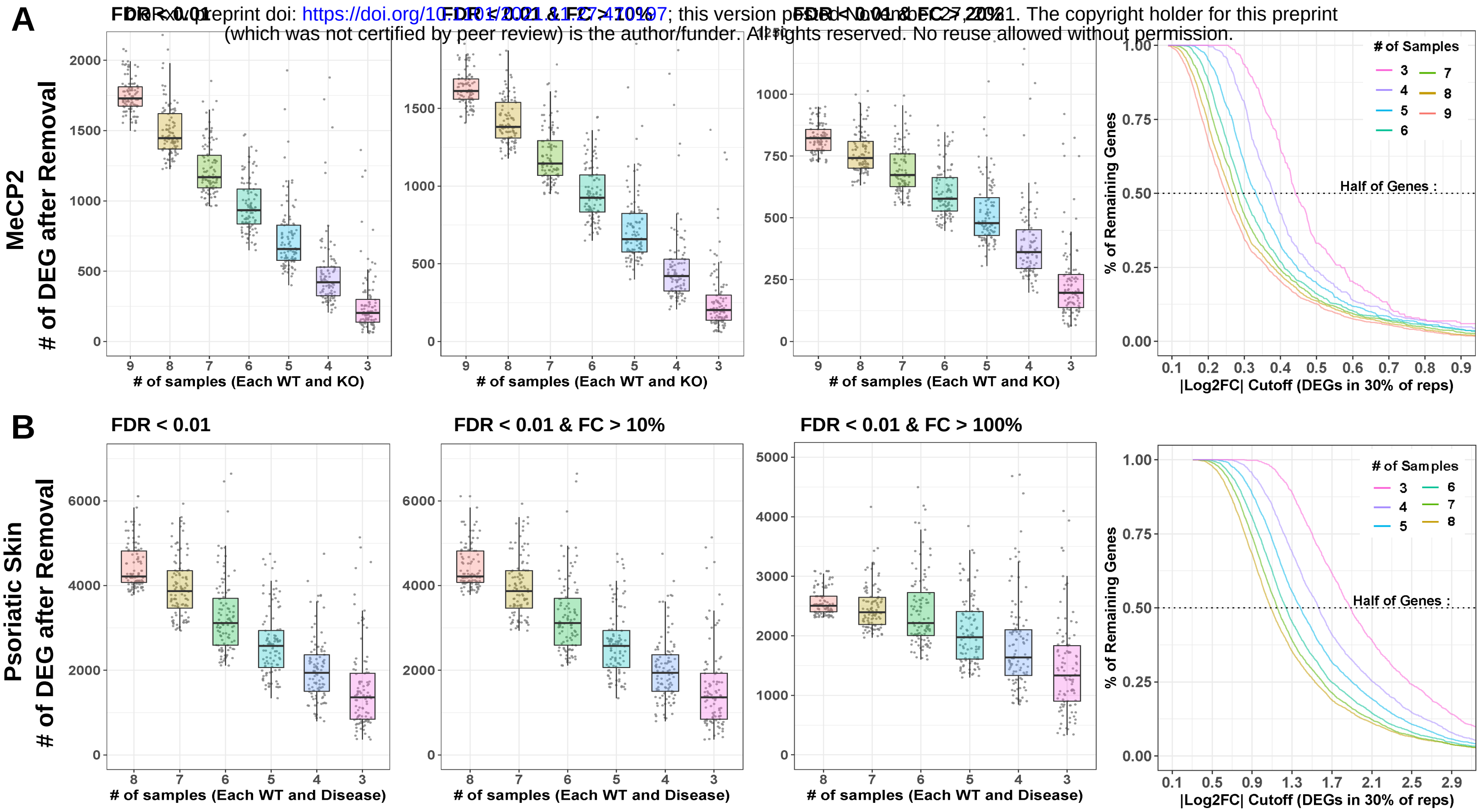
B. UMAP on normalized and batch-corrected on Alzheimer's data, colored for meta-characteristics.

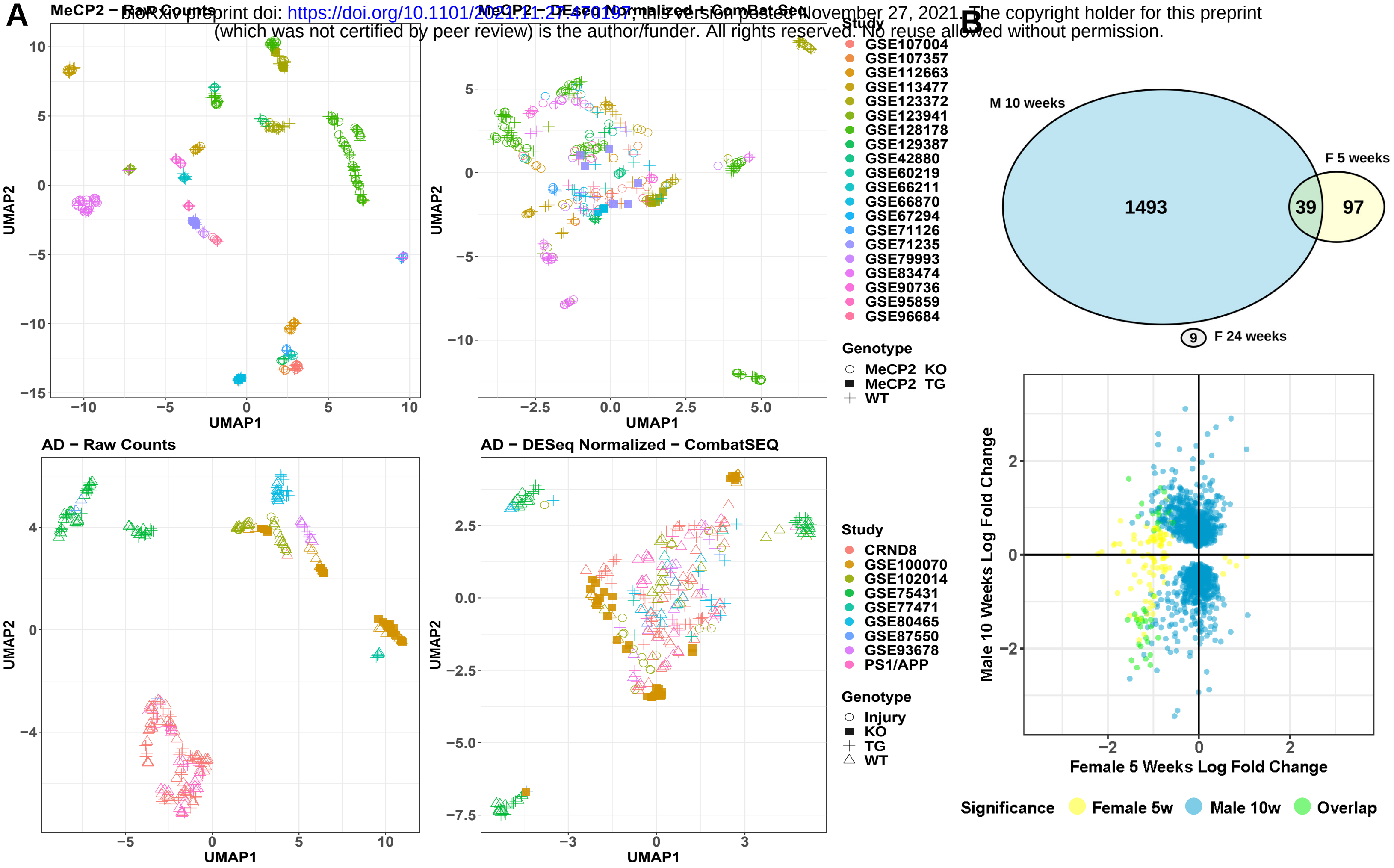

A
Data curation + processing
GEO
MultiQC
MeCP2 Fastq
QC
Alignment
DEG Analysis
Database
MECP2pedia
Gene Report
Multi-gene Search
Genome Browser
Bioinformatics analyses
Core DEG Identification
DEG Characterization
Cross Species Enrichment
Power Analysis
Batch Effect Analysis
B
2015
2017
2018
2019
2020
Nucleus
Whole Cell
Chromatin
C57BL-6J
No Strain
FVB
BirdJackson
F1hybridFVB-129S6SvEv
129BirdC57
CX3CR1-cre
Cortex
Forebrain
Dorsal Hippocampus
Striatum
Blood
Whole Brain
Cerebellum
Hippocampus
Frontal Cortex
Hypothalamus
Dentate Gyrus
Cerebral Cortex
Visual Cortex
Olfactory Epithelium
Brown Adipose
Johnson
Boxer
Clemens
Chang
Pacheco
Gulmez
Lavery
Sanfeliu
Zhao
Osenberg
Cronk
Sztainberg
Voineagu
Petazzi
Chen
Pohodich
Gabel
Rube
Wolf


A
Contrasts
GSE79993_1
GSE107004_1
GSE128178_1
GSE66870_2
GSE66870_1
GSE128178_4
GSE60219_2
GSE71126_1
GSE66211_1
GSE128178_3
GSE128178_5
GSE113477_2
GSE128178_2
GSE60219_1
GSE123372_1
GSE128178_6
GSE123372_3
GSE42880_1
GSE107357_1
GSE95859_1
GSE96684_1
GSE129387_2
GSE123372_2
GSE67294_1
GSE90736_1
GSE83474_6
GSE123941_1
GSE129387_1
GSE83474_8
GSE113477_1
GSE83474_7
GSE71235_1
GSE83474_3
GSE83474_4
GSE83474_2
GSE83474_1
GSE83474_10
GSE112663_1
GSE90736_2
GSE71235_2
GSE83474_5
GSE83474_9
GSE112663_2
-8 -6 -4 -2 0 2 4 6 8
Log2FC
0.0 2.5 5.0 7.5 10.5 12.5
Log10(# of DEG)
DEG Direction
Down
Core Down
Up
Core Up
B
FDR < 0.01
FDR < 0.05
FDR < 0.1
# of DEG in at Least
1 Contrasts
4 Contrasts
12 Contrasts
20 Contrasts
0.00 0.25 0.50 0.75 1.00
Percentage of Contrasts in Which a DEG is Upregulated
C
Common Core
Expressed
Whole Genome
Protein
Pseudogene
misc RNA
lncRNA
Immunoglobin
TEC
miRNA
Processed Transcript
D
chr6
chr8
chr12
chr13
TSS
1.5e+08
1.0e+08
5.0e+07
0.0e+00
1e+08
5e+07
0e+00
1.25e+08
1.00e+08
7.50e+07
5.00e+07
2.50e+07
0.00e+00
Core Up
Non DEG
Core Down
Count
1.00
0.75
0.50
0.25
0.00
0.0e+00 1.0e+07 2.0e+07 3.0e+07 4.0e+07 5.0e+07 6.0e+07 7.0e+07 8.0e+07 9.0e+07 1.0e+08 1.1e+08 1.2e+08 1.3e+08
TSS
Gene Direction
Core Down
Core Up
NonDEG


A
Leiden Clusters
UMAP2
UMAP1
2.5
0.0
-2.5
-3
0
3
6
clusters
0
1
2
3
4
5
6
7
8
Percentage of Contrasts Upregulated
UMAP2
UMAP1
2.5
0.0
-2.5
-5.0
Percent Upregulated
100%
75%
50%
25%
0%
C
Cell Fraction
Whole cell
Nucleus
Chromatin
−1.0
−0.5
0.0
0.5
1.0
B
Overlap Count
0.00
0.02
0.04
0.06
0.08
signal transduction
multicellular organism development
cell differentiation
ion transport
cell adhesion
nervous system development
Cluster 0 Biological Process
2
4
6
8
-log(FDR)
Overlap Count
0.00
0.02
0.04
0.06
0.08
0.10
regulation of transcription by RNA polymerase II
positive regulation of transcription by RNA polymerase II
regulation of transcription, DNA-templated
cell differentiation
multicellular organism development
signal transduction
Cluster 1 Biological Process
2
4
6
8
10
12
-log(FDR)


A
GSE51607_2
GSE128380_2
GSE113902_1
GSE51607_1
GSE128380_1
GSE51607_3
GSE51607_4
GSE80348_1
GSE83323_1
GSE87855_1
GSE57974_1
GSE57974_2
-6 -4 -2 0 2 4 6
Log2FC
0 2 4 6
Log10(#DEG)
DEG Direction
Down
Up
B
MouseCore
GSE113902_1
GSE128380_1
GSE128380_2
GSE51607_2
GSE51607_3
GSE51607_1
GSE51607_4
GSE80348_1
GSE57974_1
GSE57974_2
GSE83323_1
GSE87855_1
-2 0 2 4 6
MouseCore
Down
Up
-3 -2 -1 0 1 2 3
C
Normalized Enrichment Score
2
1
0
-1
-2
Down Up
GSE113902_1
GSE128380_1
GSE128380_2
GSE51607_1
GSE51607_2
GSE51607_3
GSE51607_4
GSE57974_1
GSE57974_2
GSE80348_1
GSE83323_1
GSE87855_1
★
-Log10(FDR) • 0 ● 1 ● 2
D
Talkowski
Elliot
Basson
Barak
Tsai
Han
Kim
Pasmanik-Chor
Zhu
Corley
Yan
Peixoto
Cripino
Lu
Sanghvi
Cerebellum
Striatum
Cortex
Whole Brain
Hippocampus
Prefrontal Crtx
Lip
No Data
C57BL/6N:129Sv
C57BL/6J
No Data
C57BL:129/SvJ
Mixed
16P11.2
CHD8
SHANK3
GTF2I
ARID1B
DYRK1A
ADNP
CUL3
KO
Heterozygote
Control
Gtf2
Mutant
cKO
E
GSE126559-6
GSE126559-5
GSE126559-4
GSE126559-3
GSE126559-2
GSE126559-1
GSE72664-4
GSE72664-3
GSE72664-2
GSE72664-1
GSE92238
GSE87370
GSE81103-2
GSE81103-1
GSE72442
GSE107919
GSE133160
GSE67052
GSE97471
GSE81082
GSE128838
GSE138539
GSE124946-2
GSE124946-1
GSE113754
GSE113368
GSE109328-2
GSE109328-1
Contrast
0 2 4 6
Fisher odds ratio
-log10(fisher.p) • 0 ● 10 ● 20 ● 30
gene
16P11.2
ARID1B
CUL3
GTF2I
ADNP
CHD8
DYRK1A
SHANK3
F
All Core
Down Only
Up Only
GSE107919 (CHD8)
1676
4709
413
2888
517
2567
p<9.52e-07
p<9.09e-05
p<1.46e-08
GSE72664-1 (ADNP)
368
462
99
466
49
216
p<1.24e-38
p<4.71e-07
p<1.70e-02
GSE81103-2 (CHD8)
70
149
6
49
42
122
p<6.93e-03
p<5.49e-01
p<3.34e-05
Genes in Mecp2 core
Not in Mecp2 core


A
FDR < 0.01
FDR < 0.01 & FC > 10%
FDR < 0.01 & FC > 20%
MeCP2
# of DEG after Removal
# of samples (Each WT and KO)
9 8 7 6 5 4 3
0 500 1000 1500 2000
0 250 500 750 1000
% of Remaining Genes
|Log2FC| Cutoff (DEGs in 30% of reps)
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
0.00 0.25 0.50 0.75 1.00
Half of Genes :
# of Samples
3 4 5 6 7 8 9
B
FDR < 0.01
FDR < 0.01 & FC > 10%
FDR < 0.01 & FC > 100%
Psoriatic Skin
# of DEG after Removal
# of samples (Each WT and Disease)
8 7 6 5 4 3
0 2000 4000 6000
0 1000 2000 3000 4000 5000
% of Remaining Genes
|Log2FC| Cutoff (DEGs in 30% of reps)
0.1 0.5 0.9 1.3 1.7 2.1 2.5 2.9
0.00 0.25 0.50 0.75 1.00
Half of Genes :
# of Samples
3 4 5 6 7 8


A
MeCP2 − Raw Counts
MeCP2 − DESeq Normalized + CombatSeq
UMAP1
UMAP2
Study
GSE107004
GSE107357
GSE112663
GSE113477
GSE123372
GSE123941
GSE128178
GSE129387
GSE42880
GSE60219
GSE66211
GSE66870
GSE67294
GSE71126
GSE71235
GSE79993
GSE83474
GSE90736
GSE95859
GSE96684
Genotype
○ MeCP2 KO
■ MeCP2 TG
+ WT
AD − Raw Counts
AD − DESeq Normalized − CombatSEQ
Study
CRND8
GSE100070
GSE102014
GSE75431
GSE77471
GSE80465
GSE87550
GSE93678
PS1/APP
Genotype
○ Injury
■ KO
+ TG
△ WT
B
M 10 weeks
F 5 weeks
1493
39
97
9
F 24 weeks
Male 10 Weeks Log Fold Change
Female 5 Weeks Log Fold Change
Significance
Female 5w
Male 10w
Overlap