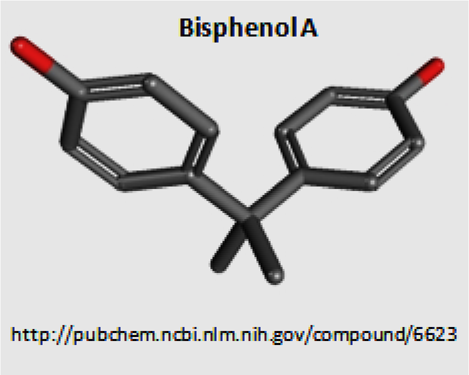
Using High-Throughput Sequencing to Investigate the Transgenerational Effects of Endocrine-Disrupting Compound Bisphenol A on Gene Transcription in Sexually Dimorphic Mouse Brain Regions MPOA and BNST

Anne D. Henriksen Ph.D.

in collaboration with Jennifer T. Wolstenholme Ph.D., Philip S. Lambeth, Jessica A. Goldsby, and Emilie R. Rissman Ph.D.

I am currently engaged in an NIH-funded research project with faculty at The University of Virginia and North Carolina State University that investigates the transgenerational effects of the plastic component, bisphenol A (BPA), on gene transcription in two regions of the brain: the medial preoptic area (MPOA) of the hypothalamus and the bed nucleus of the stria terminalis (BNST). My participation in this work has been generously supported by both the James Madison University and University of Virginia 4-VA programs.



Bisphenol A

http://pubchem.ncbi.nlm.nih.gov/compound/6623

BPA is a recognized endocrine-disrupting compound (EDC) and as such, is known to interfere with receptor interactions that have the sex steroid hormone 17β-estradiol (E2) as a natural ligand [1]. This happens because the three-dimensional structure of BPA is similar enough to E2 that BPA is able to bind to endogenous E2 receptors [2-3]. BPA can then go on to participate in the same signal transduction pathways as 17β-estradiol, or alternatively, it can block the true physiological actions of the endogenous hormone [4-6]. This can even happen at what are considered physiologically low doses [7-9]. One of the most important questions regarding EDCs such as BPA is whether or not endocrine-disrupting effects can be passed on to future generations that are themselves not exposed to BPA at any point in their own lifetime. If a pregnant mouse is exposed to BPA (e.g., through food or water), then the offspring of that pregnancy will also be exposed, as will the offspring of the offspring because they were exposed *in utero* as gametes. The pregnant mouse is considered the F0 generation, the exposed fetuses are considered the F1 generation, and the mice exposed as gametes *in utero* are considered the F2 generation. The first generation to not be directly exposed to BPA is the F3 generation. If the F3 generation displays any alterations to gene transcription processes that the control samples do not display, these changes must be attributable to the BPA exposure that occurred in the previous generations [10-14].

The only way that the effects of BPA on unexposed generations can be explained is if the F0 or F1 or F2 exposure resulted in some kind of epigenetic modification that was transmitted to the F3 epigenome [15-17]. Epigenetic changes include, but are not limited to, DNA methylation and histone tail modifications (e.g., methylation or acetylation). BPA-induced alterations to the epigenome such as DNA methylation and histone tail modifications can either abnormally activate or abnormally repress gene transcription. The outcome of this is that the organism could produce either more or less than the normal complement of proteins, or even a variable complement of proteins, owing to affected promotor sites or post-translational control sites. Such epigenetic alterations could have serious negative consequences, particularly if those alterations occur in the brain and affect signaling pathways that normally interact with sex steroid hormones during crucial periods of brain sexual differentiation or during puberty [18-20].

In the absence of any *a priori* knowledge about potentially affected genes, the most informative and expedient approach for determining if differences exist in the rate of gene transcription between two conditions is high-throughput, next-generation sequencing (NGS) [21]. High-throughput, next-generation sequencing of mRNA is known as RNA-Seq [22-23]. In RNA-Seq, mRNA is collected, cDNA is generated, and the cDNA is put on a flow cell in an instrument where it is amplified and then undergoes a chemically controlled base-by-base addition to form complementary "reads" that replicate the mRNA sequence. Reads consist of long strings of mRNA bases in "FASTQ" files that can be aligned to an annotated genome or transcriptome. The fundamental idea is that the number of reads that align to a particular gene under one set of conditions, relative to the number of reads that align to that same gene under a different set of conditions, will be statistically representative of the *in vivo* differential gene transcription rates [24]. This technique is ideal for trying to determine if transcription rates in BPA-exposed mice differ from transcription rates in controls; this is especially valuable when no preliminary information about differentially expressed genes is available.

In this research, RNA-Seq was performed on poly-A-tailed mRNA and long noncoding RNA (lncRNA) obtained from the medial preoptic area (MPOA) of the hypothalamus and the bed nucleus of the stria terminalis (BNST) for both F1 and F3 generations of BPA-exposed and control mice. Detailed bioinformatic procedures were then performed to analyze and interpret the RNA-Seq read data into biologically relevant information about differentially expressed genes and affected pathways. Once the statistically significant differentially expressed genes between BPA and control mice were characterized, confirmatory experiments using RT-qPCR were done to validate transcription fold changes for select genes.

In addition to RNA-Seq of mRNA and lncRNA, it is also informative to investigate epigenetic features of the BPA-exposed vs. control mice for differences in histone tail modifications, as this would be a factor contributing to differential expression. One process for probing sites of histone tail modifications is chromatin immunoprecipitation combined with high-throughput sequencing, known as ChIP-Seq. ChIP-Seq is used to analyze the binding-site interactions between protein transcription factors and DNA, in particular, the relationship between the specific binding site of those proteins and relative gene expression levels [25-26]. Modifications to the histone tails, particularly the addition of methyl groups or acetyl groups to key lysine residues, is known to either activate or repress gene transcription by altering interactions of the transcription factors with the DNA. Information about DNA-protein interactions can be used to deduce possible mechanisms of differential expression between treatment and control conditions. In this study of the transgenerational effects of BPA, we will perform ChIP-Seq at three known sites of methylation or acetylation of specific H3 histone protein lysine residues: H3K27me3, which is a known repressor of gene transcription, and H3K4me3 and HeK27ac, which are known activators of gene transcription. We hope to correlate RNA-Seq differential expression profiles with ChIP-Seq results to understand how BPA may achieve any transgenerational effects in the F3 generation.

My role in this study is the analysis and interpretation of the RNA-Seq data for the BPA vs. control F1 and F3 generations of C57BL/6J mice. Initially I partnered with Stephen Turner, Ph.D. of the Bioinformatics Core at The University of Virginia, but subsequently modified and then re-created all the analyses using different platforms and strategies. Analysis of RNA-Seq data is accomplished in a series of stages: 1.) alignment of the raw FASTQ read files to the appropriate genome; 2.) quantification of the aligned reads with respect to the annotated transcriptome; 3.) downstream exploratory data analysis of the resulting aligned read data; 4.) differential expression analysis of the contrast conditions; and 5.) gene ontology and pathway enrichment analysis of the differentially expressed genes. Each of these stages requires dedicated software, and any one task has a plethora of strategies and software alternatives available to carry out such analyses [27-29].

Once the FASTQ files are made available by the facility that does the actual sequencing, the first step of the upstream analysis is to inspect the FASTQ files and trim the reads if necessary to eliminate low-quality base calls [30]. This was done initially by the company that did the sequencing, Expression Analysis, using ea-utils, which is a proprietary tool of that company [31]. I subsequently performed inspection and trimming in Galaxy using FASTQC and FASTX Trimmer [32]. Galaxy is a comprehensive bioinformatics workflow management platform comprising all the software necessary to do high-throughput, next-generation sequencing data analysis [33-35].

The next step is to align the quality-trimmed reads to the mouse genome, and then to assign those aligned reads to gene transcripts and/or splice variants. The alignment step was done initially to the mm9 mouse genome using STAR, which is an ultrafast, universal read alignment tool running on the University of Virginia 24-core UNIX platform [36]. Within Galaxy, the alignment was done to the mm9 mouse genome using TopHat2, which is part of the Tuxedo RNA-Seq analysis tool set. TopHat2 aligns reads to a genome based on the ultrafast and memory-efficient, short-read aligner Bowtie2, but Tophat2 has the added capability of being able to identify splice junction sites [37-38]. This capability is of critical importance when mapping RNA-Seq reads because spliced mRNA is missing large exonic sequences with respect to the genome. Both upstream analyses approaches produce, as their primary final product, a matrix of quantified reads organized by gene and by experiment, which is then input to the downstream processes of exploratory data analysis.

The downstream analyses of the count matrix data were conducted initially by the UVA bioinformatics core using DESeq2, a package in Bioconductor [39-41]. Bioconductor is an open source compilation of genomic analysis packages written for the R Statistical Programming Language [42]. DESeq2 performs differential gene expression analysis based on the negative binomial distribution and it includes a variety of exploratory data analysis procedures such as count matrix normalization, principal component analysis (PCA), MA plots, volcano plots, and dispersion estimates. Following initial DESeq2 analyses, one set of mouse data was determined to be a severe outlier. That mouse data had to be eliminated and the revised count matrix had to be completely re-analyzed. I performed a second complete DESeq2 analysis of the STAR-aligned, revised count matrix without the outlier mouse. This included re-writing and re-running the DESeq2 script and writing R code for boxplots, PCA plots, and volcano plots.

Contrasts involve the formal comparison of expression levels between different treatment conditions of an experiment to determine which genes are differentially expressed. The main contrasts of interest here are the control vs. BPA-exposed mice within the F1 and F3 generations, denoted BCF1 and BCF3. Contrasts were done using the R package DESeq2, first by the UVA Bioinformatics Core, and then re-done again in DESeq2 with the outlier mouse removed. Contrasts between conditions produce a measure of effect size, log fold change (lfc), and a measure of statistical significance, *p* value adjusted for multiple comparisons. Genes can be ranked by both lfc and adjusted *p* value to determine those which are differentially expressed between two conditions. In this experiment, genes that had a lfc of greater than 0.5 and an adjusted *p* value of less than 0.05 were considered differentially expressed. Volcano plots were then done to visualize the DE genes and to select candidates for further investigation, including clustering, heatmapping, and RT-qPCR. A volcano plot of the 77 DE genes in the BPA vs. Control F1 generation contrast is shown in Figure 1. Fourteen genes were determined to be more than twice as highly expressed in the F1 generation under direct exposure *in utero* to BPA, and thirteen genes were determined to be less than half as expressed in the F1 generation under direct exposure *in utero* to BPA.

In order to explore for genes whose expression levels may be correlated, genes can be “clustered” based on a mathematical algorithm that groups together genes with similar expression profiles across biological replicates. Genes whose expression levels cluster together may be involved in related biological pathways and their transcription processes may be similarly affected by BPA. Clustered genes are visualized in a heatmap, where colors are used to indicate relative expression levels of a gene across experiments. Clusters of genes can then be put through a gene ontology analysis to identify specific pathways and related biological functions. A clustered heatmap is shown in Figure 2 for the 77 genes that are differentially expressed between BPA and control mice in the F1 generation. This heatmap was done in R using the package *pheatmap* [43]. The color of the cells indicates the standardized value of normalized mRNA reads of the twelve samples for that gene. In the bottom half of the heatmap, the BPA F1 generation is more highly expressed (shades of blue) than the control F1 generation, which implies that exposure to BPA results in upregulation of these genes relative to controls. In the top half of the heatmap, the BPA F1 generation is less expressed (shades of orange) than the control F1 generation, which implies that exposure to BPA results in downregulation of these genes relative to controls. Note that the three biological replicates for BPA F1 and the three biological replicates for control F1 cluster together.

Gene ontology (GO) enrichment analysis is a process for investigating gene sets that express similarly to determine if those genes participate in common biological pathways or have related functions [44]. If we know how genes that are similarly DE between two conditions map to specific biological pathways or functional groupings, we may be able to reason how transcriptional alterations bring about the physiological consequences of a treatment. We are currently in the process of GO enrichment analysis of the DE genes between BPA and control for the F1 and F3 generations. We are using DAVID 6.7, a GO analysis tool made available by the NIH National Institute of Allergy and Infectious Diseases (NIAID) [45]. DAVID is an acronym for Database for Annotation, Visualization and Integrated Discovery. DAVID takes as input a list of gene symbols or identifications and returns a variety of tables and charts detailing biological processes, cellular components, molecular function, diseases, protein interactions, tissue expression, pathways, annotation clustering, and functional classifications of annotated genes. This work is ongoing and will be reported with the results of the RNA-Seq and behavior studies.

Also continuing is the analysis of the RNA-Seq data in the Galaxy pipeline. Gene expression counts (in FPKMs) and differential expression profiles that result from the Galaxy pipeline will be compared to the UNIX platform results [24]. One advantage of the Galaxy platform is that it is easy to align reads and perform differential expression analysis with the Tuxedo suite under a variety of different scenarios and parameter settings [46]. This will provide supplementary information about differentially expressed genes using different alignment software and different read quantification approaches.

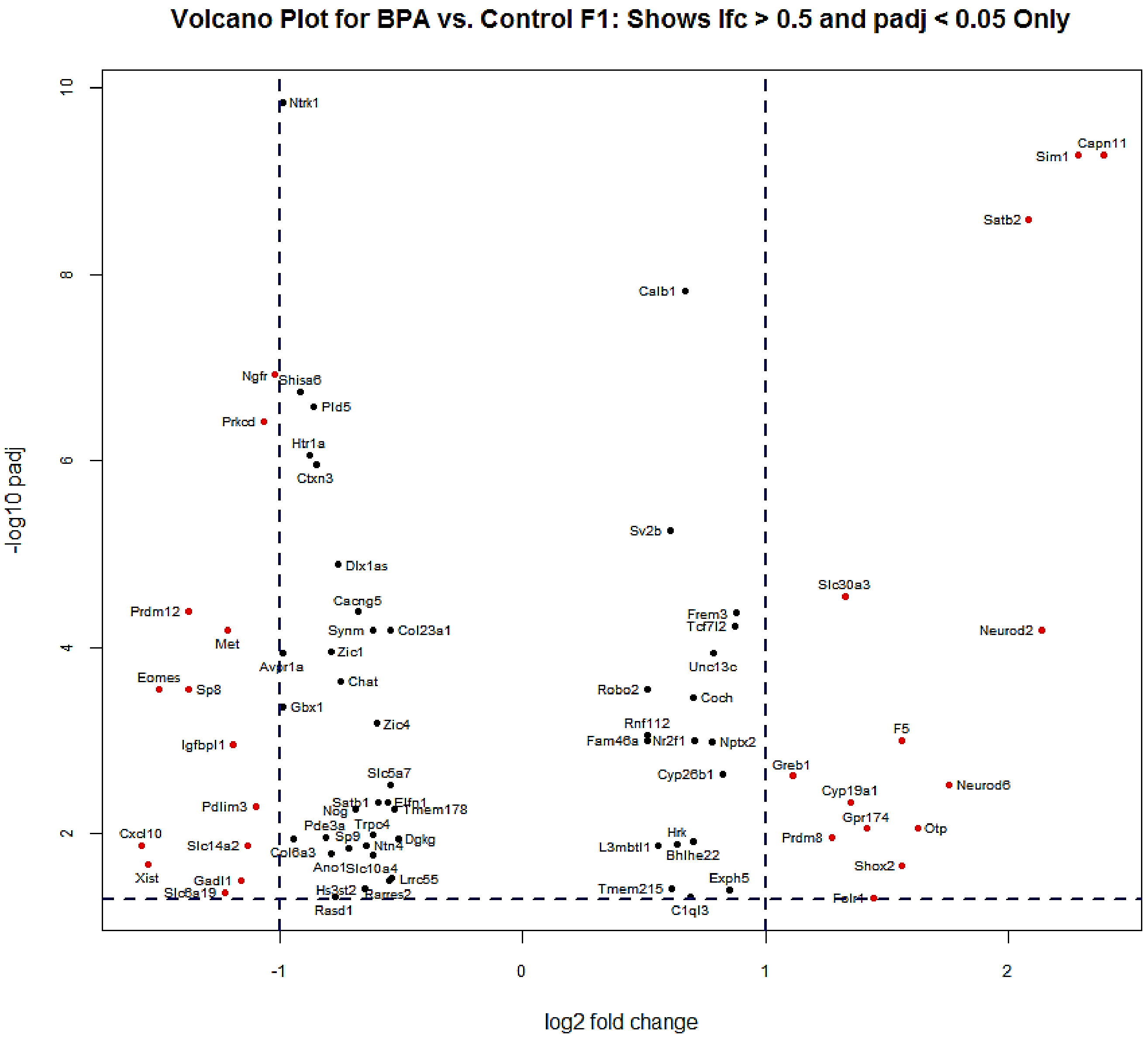


Figure 1. *Volcano plot of the 77 out of 2227 genes that were determined to be differentially expressed (DE) in this RNA-Seq data. This volcano plot is for the BPA vs. control mice in the F1 generation. Only the section of the log-scale y axis for which the adjusted p value is less than 0.05 (-log$_{10}$(0.05) = 1.3) is shown. The higher up and further to the edges that a gene is situated, the greater the effect size (higher log fold change) and the higher the statistical significance (lower adjusted p value). Fourteen genes (red dots on the right) were determined to be more than twice as highly expressed in the F1 generation under direct exposure in utero to BPA, and thirteen genes (red dots on the left) were determined to be less than half as expressed in the F1 generation under direct exposure in utero to BPA. Volcano plot done in R Statistical Programming Language.*

adhenriksen 4/2015

Heatmap of BPA vs. Control F1: Correlation Distance and Average Linkage

Genes with |lfc| > 0.5 and *p* adjusted < 0.05

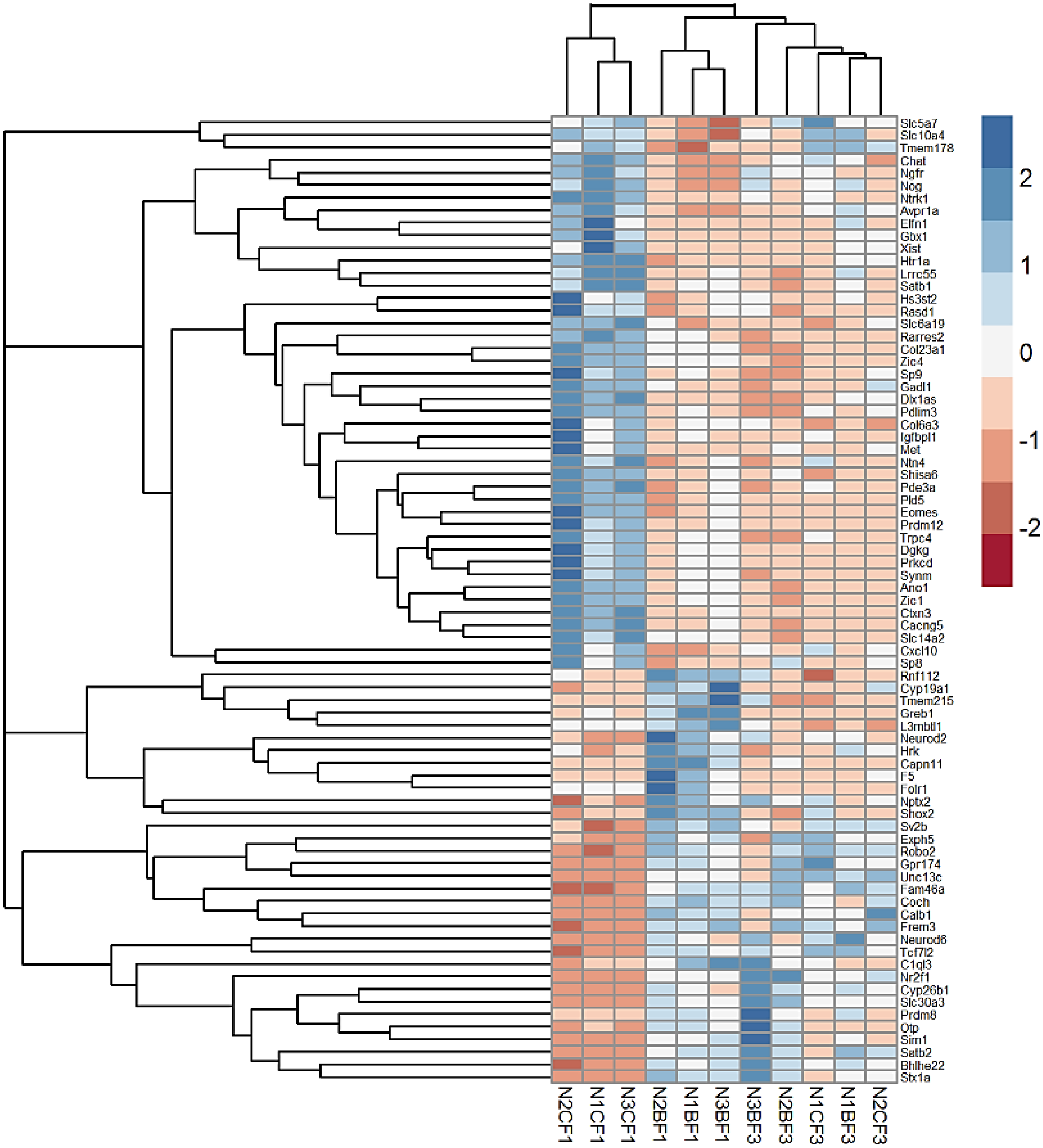


Figure 2. *Heatmap and clustering diagram for the 77 of the 2227 genes that were determined to be differentially expressed (DE) in this RNA-Seq data with adjusted p values of < 0.05 and log fold changes of > 0.5. These are the genes that are DE between BPA and control mice in the F1 generation. The clustering was done in R Statistical Programming Language with the pheatmap package using correlation distance and average linkage. The color of the cells indicates the standardized number of normalized mRNA reads of the twelve samples for that gene. In the bottom half of the heatmap, the BPA F1 generation is more highly expressed (shades of blue) than the control F1 generation, which implies that exposure to BPA results in upregulation of these genes relative to controls. In the top half of the heatmap, the BPA F1 generation is less expressed (shades of orange) than the control F1 generation, which implies that exposure to BPA results in downregulation of these genes relative to controls. Note that the three biological replicates for BPA F1 and the three biological replicates for control F1 cluster together.*

adhenriksen 4/2015

References

[1] Rubin, B. S. (2011). Bisphenol A: an endocrine disruptor with widespread exposure and multiple effects. *The Journal of steroid biochemistry and molecular biology*, *127*(1), 27-34.

[2] Li, L., Wang, Q., Zhang, Y., Niu, Y., Yao, X., & Liu, H. (2015). The Molecular Mechanism of Bisphenol A (BPA) as an Endocrine Disruptor by Interacting with Nuclear Receptors: Insights from Molecular Dynamics (MD) Simulations. *PloS one*, *10*(3).

[3] Li, Y., Luh, C. J., Burns, K. A., Arao, Y., Jiang, Z., Teng, C. T., ... & Korach, K. S. (2013). Endocrine-Disrupting Chemicals(EDCs): In Vitro Mechanism of Estrogenic Activation and Differential Effects on ER Target Genes. *Environmental health perspectives*, *121*(4), 459-466.

[4] Xu, X. B., He, Y., Song, C., Ke, X., Fan, S. J., Peng, W. J., ... & Kato, N. (2014). Bisphenol A regulates the estrogen receptor alpha signaling in developing hippocampus of male rats through estrogen receptor. *Hippocampus*, *24*(12), 1570-1580.

[5] A Alonso-Magdalena, P., Ropero, A. B., Soriano, S., García-Arévalo, M., Ripoll, C., Fuentes, E., ... & Nadal, Á. (2012). Bisphenol-A acts as a potent estrogen via non-classical estrogen triggered pathways. *Molecular and cellular endocrinology*, *355*(2), 201-207.

[6] Rogers, J. A., Metz, L., & Yong, V. W. (2013). Review: Endocrine disrupting chemicals and immune responses: a focus on bisphenol-A and its potential mechanisms. *Molecular immunology*, *53*(4), 421-430.

[7] Welshons, W. V., Nagel, S. C., & vom Saal, F. S. (2006). Large effects from small exposures. III. Endocrine mechanisms mediating effects of bisphenol A at levels of human exposure. *Endocrinology*, *147*(6), s56-s69.

[8] Vandenberg, L. N., Colborn, T., Hayes, T. B., Heindel, J. J., Jacobs Jr, D. R., Lee, D. H., ... & Myers, J. P. (2012). Hormones and endocrine-disrupting chemicals: low-dose effects and nonmonotonic dose responses. *Endocrine reviews*, *33*(3), 378-455.

[9] Teeguarden, J. G., & Hanson-Drury, S. (2013). A systematic review of bisphenol A "low dose" studies in the context of human exposure: A case for establishing standards for reporting "low-dose" effects of chemicals. *Food and Chemical Toxicology*, *62*, 935-948.

[10] Skinner, M. K., Manikkam, M., & Guerrero-Bosagna, C. (2010). Epigenetic transgenerational actions of environmental factors in disease etiology. *Trends in Endocrinology & Metabolism*, *21*(4), 214-222.

[11] Rissman, E. F., & Adli, M. (2014). Minireview: Transgenerational Epigenetic Inheritance: Focus on Endocrine Disrupting Compounds. *Endocrinology*, *155*(8), 2770-2780.

[12] Wolstenholme, J. T., Edwards, M., Shetty, S. R., Gatewood, J. D., Taylor, J. A., Rissman, E. F., & Connelly, J. J. (2012). Gestational exposure to bisphenol A produces transgenerational changes in behaviors and gene expression. *Endocrinology*, *153*(8), 3828-3838.

[13] Manikkam, M., Tracey, R., Guerrero-Bosagna, C., & Skinner, M. K. (2013). Plastics derived endocrine disruptors (BPA, DEHP and DBP) induce epigenetic transgenerational inheritance of obesity, reproductive disease and sperm epimutations. *PLoS One*, *8*(1), e55387.

[14] Wolstenholme, J. T., Goldsby, J. A., & Rissman, E. F. (2013). Transgenerational effects of prenatal bisphenol A on social recognition. *Hormones and behavior*, *64*(5), 833-839.

[15] Bollati, V., & Baccarelli, A. (2010). Environmental epigenetics. *Heredity*, *105*(1), 105-112.

[16] Bernal, A. J., & Jirtle, R. L. (2010). Epigenomic disruption: the effects of early developmental exposures. *Birth Defects Research Part A: Clinical and Molecular Teratology*, *88*(10), 938-944.

[17] Skinner, M. K., Manikkam, M., & Guerrero-Bosagna, C. (2010). Epigenetic transgenerational actions of environmental factors in disease etiology. *Trends in Endocrinology & Metabolism*, *21*(4), 214-222.

[18] Wolstenholme, J. T., Rissman, E. F., & Connelly, J. J. (2011). The role of Bisphenol A in shaping the brain, epigenome and behavior. *Hormones and behavior*, *59*(3), 296-305.

[19] Palanza, P., Gioiosa, L., vom Saal, F. S., & Parmigiani, S. (2008). Effects of developmental exposure to bisphenol A on brain and behavior in mice. *Environmental research*, *108*(2), 150-157.

[20] Bohacek, J., Gapp, K., Saab, B. J., & Mansuy, I. M. (2013). Transgenerational epigenetic effects on brain functions. *Biological psychiatry*, *73*(4), 313-320.

[21] Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature Reviews Genetics*, *11*(1), 31-46.

[22] Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57-63.

[23] Marguerat, S., & Bähler, J. (2010). RNA-seq: from technology to biology. *Cellular and molecular life sciences*, *67*(4), 569-579.

[24] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, *5*(7), 621-628.

[25] Pepke, S., Wold, B., & Mortazavi, A. (2009). Computation for ChIP-seq and RNA-seq studies. *Nature methods*, *6*, S22-S32.

[26] Park, P. J. (2009). ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, *10*(10), 669-680.

[27] Oshlack, A., Robinson, M. D., & Young, M. D. (2010). From RNA-seq reads to differential expression results. *Genome biol*, *11*(12), 220.

[28] Soneson, C., & Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC bioinformatics*, *14*(1), 91.

[29] Garber, M., Grabherr, M. G., Guttman, M., & Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature methods*, *8*(6), 469-477.

[30] Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research*, *38*(6), 1767-1771.

[31] Erik Aronesty (2011). *ea-utils* : "Command-line tools for processing biological sequencing data"; http://code.google.com/p/ea-utils

[32] Andrews, S. (2014). FastQC. A quality control tool for high throughput sequence data. Babraham Bioinformatics. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[33] Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., ... & Nekrutenko, A. (2005). Galaxy: a platform for interactive large-scale genome analysis. *Genome research*, *15*(10), 1451-1455.

[34] Goecks, J., Nekrutenko, A., & Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, *11*(8), R86.

[35] Blankenberg, D., Kuster, G. V., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., ... & Taylor, J. (2010). Galaxy: a web-based genome analysis tool for experimentalists. *Current protocols in molecular biology*, 19-10.

[36] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., ... & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15-21.

[37] Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., ... & Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, *7*(3), 562-578.

[38] Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*, *14*(4), R36.

[39] 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.

[40] Love, M., Anders, S., & Huber, W. (2013). Differential analysis of count data–the DESeq2 package. http://140.107.3.20/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.pdf .

[41] Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., ... & Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, *5*(10), R80.

[42] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[43] Kolde, R. (2013). pheatmap: Pretty Heatmaps. version 0.7.

[44] Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, *37*(1), 1-13.

[45] Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, *4*(1), 44-57.

[46] Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., ... & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, *28*(5), 511-515.