Supplemental File 1

RNA-seq library preparation, sequencing, processing and quality control procedures

Total RNA from 384 samples (i.e., n=375 patients, with n=9 replicate samples) were sent to the University of California Davis Genomics Core Facility (Davis, CA, USA) for library preparation and sequencing. Prior to library preparation, 600 nanograms (ng) of total RNA was treated with the Globin-Zero Gold rRNA Removal Kit (Illumina Inc., San Diego, CA) to deplete cytoplasmic ribosomal RNA(1) and human globin mRNA. (2, 3) The globin/ribo depleted RNA was cleaned with Agencourt RNAClean XP (Beckman Coulter, Indianapolis, IN) and the sequencing libraries were prepared with KAPA RNA HyperPrep Kit (Roche Diagnostics Corp., Indianapolis, IN) according to the manufacturer's protocol. Fourteen cycles of polymerase chain reaction (PCR) amplification were used for double six base pair index addition and library fragment enrichment. Prepared libraries were quantified on a Roche LightCycler 480II (Roche Diagnostics Corp., Indianapolis, IN) using KAPA Illumina library quantitative PCR reagents (Roche Diagnostics Corp., Indianapolis, IN).

Sequencing of the 384 samples was done on an Illumina HiSeq 4000 apparatus (Illumina Inc., San Diego, CA). All 384 samples were multiplexed into four pools of 96 samples each, with each sample labeled with a dual-indexed adapter.(4) The sample pools were sequenced on four lanes for 100 cycles of single-end reads with a 1% PhiX v3 control library spike (Illumina Inc., San Diego, CA). Post-sequencing basecall files (bclfiles) were demultiplexed and converted into a FASTQ file format using the bcl2fastq v2.17 software (Illumina Inc., San Diego, CA). Data were posted and retrieved from a secure FTP site hosted by the Core Facility.

RNA-seq data processing was performed based on best practices (5, 6) and our previous experience.(7, 8) Illumina adapters and leading or trailing low quality bases were removed and reads with an average quality per base below 15 in a 4-base sliding window or below a minimum length of 36 bases were removed using Trimmomatic.(9) Individual samples were inspected with FASTQC (10) and in

aggregate with MultiQC.(11) After initial QC, 10 bases were trimmed from the beginning of all reads and reads were re-inspected with FASTQC.

The reference genome was prepared using the GRCh38 assembly (gencode.v24.GRCh38.p5.fa).(12) Transcriptome annotations (n=60,725) were obtained from the Gencode v24 primary assembly (gencode.v24.primary_assembly.annotation.gtf).(12) Trimmed reads were aligned to the annotated reference genome using the STAR aligner.(13) Output alignment files were validated using ValidateSam. Read groups were added to the alignment file using the Picard tool AddOrReplaceReadGroups. Sorted alignment files were inspected using RNA-SeQC(14) and joined for each sample. Abundance of RNA was estimated from the combined aligned reads using featureCounts.(15)

Replicate count data were processed in edgeR.(16) Ensembl transcripts (17) were annotated with Entrez gene ID and symbol.(18) Lowly expressed tags were filtered out by retaining only those tags with $\geq$10/L reads per million (where L is the minimum library size in millions) in at least N samples (where N is the smallest group size). Count estimates were normalized with the trimmed means of M-values (TMM) method.(19) TMM normalization was applied to the dataset in edgeR using calcNormFactors. Data were explored using multi-dimensional scaling (MDS) plots for all samples to identify sample outliers and potential batch effects due to technical artifacts (i.e., RNA integrity number (RIN), date of RNA extraction). The same technician performed all of the RNA extractions in one laboratory. Associations between technical variables and CIN group were assessed using Fisher's Exact Test or a generalized linear model in R. Significance was assessed at a p-value of 0.05.

Microarray hybridization, preprocessing, normalization and quality control procedures

For each sample, (n=360 patients in sample 2) approximately 100 ng of total RNA were labeled using the Illumina Total Prep RNA Amplification Kit (Thermo Fisher Scientific, Waltham, MA) and hybridized to the HumanHT-12 v4.0 Expression BeadChip (46,538 probes) (Illumina, San Diego, CA).(20, 21) The BeadChips were scanned using the iScan system (Illumina, San Diego, CA) at the University of California, San Francisco Genomics Core Facility. Each HumanHT-12 BeadChip contained 12 sample

BeadArrays. Initial quality assessment was performed using BeadArray.(22) Summary level data were calculated from the uncorrected, non-normalized, and non-transformed summary intensities at the probe level with GenomeStudio (Illumina, San Diego, CA). Data preparation and analyses were performed in R (Version 3.3.3) using well-established protocols (23-26) and our previous experience.(20, 21)

Assessment of array performance and data quality are critical for accurate analysis of microarray data.(27) By evaluating array intensities relative to other arrays and to expected controls, outliers are identified and excluded from further analysis. Sample performance was evaluated using quality assessment plots. The R package 'arrayQualityMetrics' version 3.16.0 was employed to identify potential outlier arrays on a quantitative basis by evaluating inter-array and intra-array intensities using expected controls.(28) We examined for potential outlier arrays under three criteria: by quantifying the distance between arrays; the individual array signal intensity distributions; and the individual array quality relative to the median across arrays. Individual arrays were defined as outliers if they met at least two of three criteria.(21)

Background correction is a procedure that attempts to remove signals (e.g., instrument noise) that are not attributable to the signal of interest.(26, 29) Background correction was performed with a normal+exponential convolution model using Illumina's negative control probes to estimate the parameters.(30) Normalization attempts to compensate for differences between sample preparations without knowledge of the actual differences.(26, 29) Quantile normalization was performed using Illumina's negative and positive control probes in addition to regular probes to control for variations in total mRNA production across samples.(30) Finally, log2 transformation was done to facilitate the comparisons of variations in transcript intensity from different genes.(21, 29)

Background correction and quantile normalization were performed with the neqc function from the R package 'limma', that produces a matrix of log2 transformed expression intensities with the control probe-sets removed.(31) The detection p-value provided by GenomeStudio represents the confidence that sufficient expression measurements occur above background for a probe. Probes with a detection p-value

0.05 were excluded. Finally, potential clustering of samples was evaluated by principal components analysis.(21)

Surrogate variable analysis (SVA)

For both the RNA-seq and microarray data, SVA was used to identify technical variations that contributed to heterogeneity in the sample (e.g., batch effects) that were not due to the variable of interest (i.e., nausea group membership) or significant demographic covariates.(32) The "be" method was used to identify surrogate variables.(32, 33) Any surrogate variable that was significantly associated with the phenotype was excluded.

Differential GE

For the RNA-seq data, differential GE tests were performed using our previous protocol. (7, 8) DE was determined under a variance modeling strategy that addressed the over-dispersion observed in GE count data using edgeR.(34) For this analysis, the overall dispersion, as well as the gene-wise and tag-wise dispersion, were estimated using general linear models estimated using the Cox-Reid (CR)-adjusted likelihood method.(35, 36) Differences in GE between the two CIN groups were tested using likelihood ratio tests. Demographic and clinical characteristics that differed between the two CIN groups, as well as surrogate variables, were included as covariates in the model.

For the microarray data from sample 2, differential GE tests were performed using our previously published protocol. (20, 37) Briefly, a linear model was fit using the "ls" method which included array weights and significant demographic, clinical, and surrogate variables using limma.(38) The "eBayes" method was used to evaluate for differential expression (DE).(39)

Fisher's Combined Probability test was used to combine the differential GE tests from both datasets using the uncorrected p-values.(19, 20) The two datasets (i.e., sample 1 and sample 2) were merged at the gene level using the ENTREZ gene identifier. The significance of the combined transcriptome-wide GE analysis was assessed using a strict false discovery rate (FDR) of 5% under the

Benjamini-Hochberg (BH) procedure.(42) No minimal fold-change was evaluated using the p.adjust R function.

Pathway Impact Analysis (PIA)

Most pathway analyses consider pathways as lists of genes and ignore the additional information available in the pathway representation (e.g., topology). However, PIA includes potentially important biological factors (e.g., gene-gene interactions, flow signals in a pathway, pathway topologies) as well as the magnitude (i.e., log fold-change) and the p-values from the DE analysis.(43) Using Pathway Express,(44) the PIA included p-values and log fold changes for all genes that had DE results to determine the probability of a pathway perturbation (pPERT). By including all genes in the analysis, and using the DE analysis results to represent the biological differences between the groups, we are able to capture the adjustments made for the demographic, clinical, and technical (i.e. surrogate variables) variations in the sample. A total of 208 signaling pathways were defined using the KEGG database.(45) Sequence loci data were annotated with Entrez gene IDs. The gene names were annotated using the HUGO Gene Nomenclature Committee resource database.(46) PIA was performed independently for each dataset (i.e., microarray and RNA-seq).

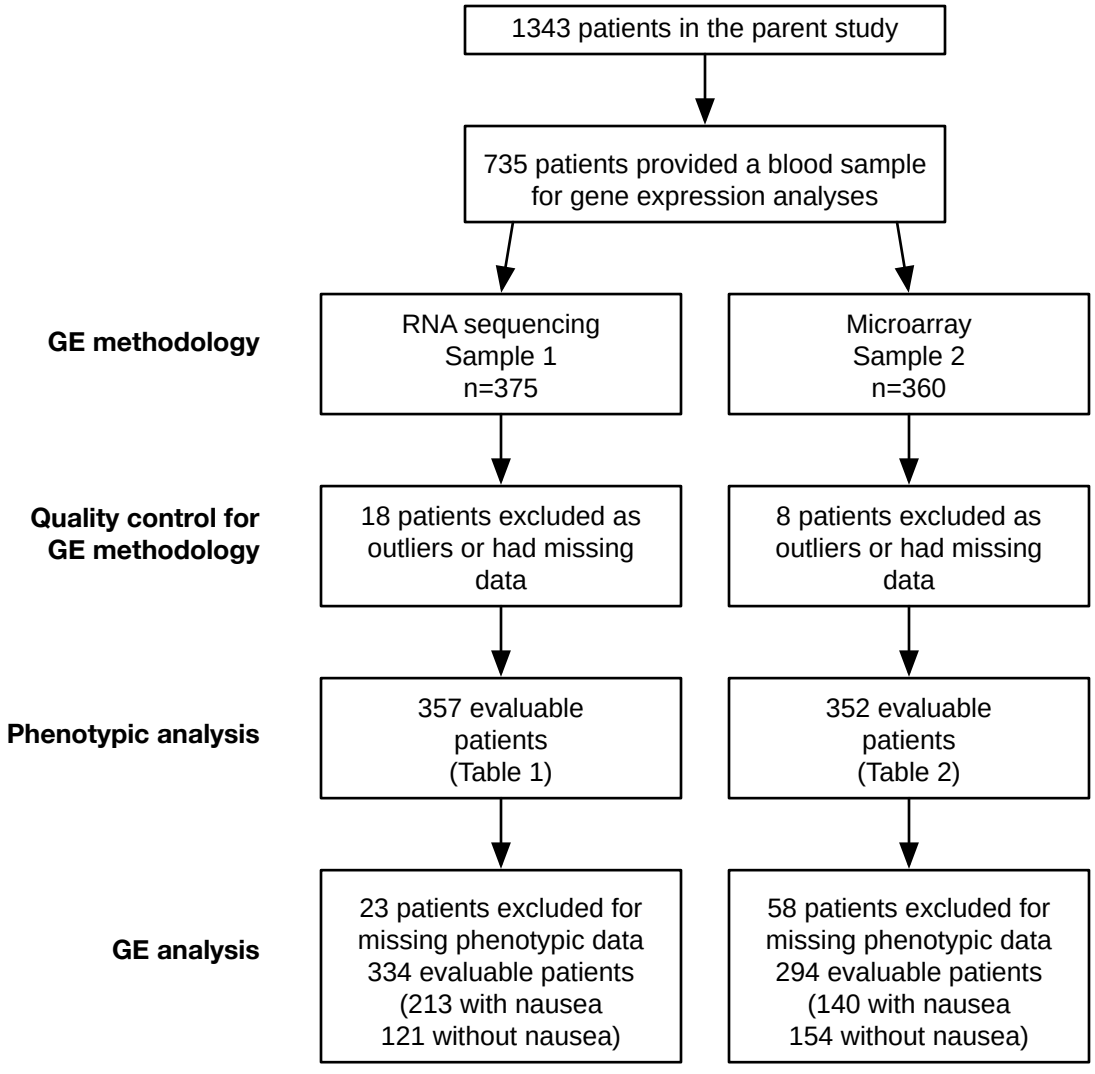
Fisher's Combined Probability test was used to determine the overall number of significantly perturbed pathways by combining the uncorrected p-values (i.e., pPERT) from the PIA tests for both samples.(40, 41) Significance of the combined transcriptome-wide PIA analysis was assessed using a family wise error rate (FWER) of 1% under the Bonferroni method.(44)

References

1. O'Neil D, Glowatz H, Schlumpberger M. Ribosomal RNA depletion for efficient use of RNA-seq capacity. Curr Protoc Mol Biol. 2013;Chapter 4:Unit 4 19. doi: 10.1002/0471142727.mb0419s103. PubMed PMID: 23821444.
2. Mastrokolias A, den Dunnen JT, van Ommen GB, t Hoen PA, van Roon-Mom WM. Increased sensitivity of next generation sequencing-based expression profiling after globin reduction in human blood RNA. BMC Genomics. 2012;13:28. doi: 10.1186/1471-2164-13-28. PubMed PMID: 22257641; PMCID: PMC3275489.
3. Choi I, Bao H, Kommadath A, Hosseini A, Sun X, Meng Y, Stothard P, Plastow GS, Tuggle CK, Reecy JM, Fritz-Waters E, Abrams SM, Lunney JK, Guan le L. Increasing gene discovery and coverage using RNA-seq of globin RNA reduced porcine blood samples. BMC Genomics. 2014;15:954. doi: 10.1186/1471-2164-15-954. PubMed PMID: 25374277; PMCID: PMC4230834.
4. Sinha R, Stanley G, Gulati GS, Ezran C, Travaglini KJ, Wei E, Chan CKF, Nabhan AN, Su T, Morganti RM, Conley SD, Chaib H, Red-Horse K, Longaker MT, Snyder MP, Krasnow MA, Weissman IL. Index Switching Causes "Spreading-Of-Signal" Among Multiplexed Samples In Illumina HiSeq 4000 DNA Sequencing. bioRxiv. 2017. doi: 10.1101/125724.
5. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczesniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13. doi: 10.1186/s13059-016-0881-8. PubMed PMID: 26813401; PMCID: PMC4728800.
6. Kukurba KR, Montgomery SB. RNA Sequencing and Analysis. Cold Spring Harb Protoc. 2015;2015(11):951-69. doi: 10.1101/pdb.top084970. PubMed PMID: 25870306; PMCID: PMC4863231.
7. Carrico AW, Flentje A, Kober K, Lee S, Hunt P, Riley ED, Shoptaw S, Flowers RNE, Dilworth SE, Pahwa S, Aouizerat BE. Recent Stimulant Use and Leukocyte Gene Expression In Methamphetamine Users with Treated HIV Infection. Brain Behav Immun. 2018. doi: 10.1016/j.bbi.2018.04.004. PubMed PMID: 29679637.
8. Flentje A, Kober KM, Carrico AW, Neilands TB, Flowers E, Heck NC, Aouizerat BE. Minority stress and leukocyte gene expression in sexual minority men living with treated HIV infection. Brain Behav Immun. 2018. doi: 10.1016/j.bbi.2018.03.016. PubMed PMID: 29548994.
9. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20. doi: 10.1093/bioinformatics/btu170. PubMed PMID: 24695404; PMCID: PMC4103590.
10. FASTQC - A quality control tool for high throughput sequence data: Brabraham Institute; 2018. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
11. Ewels P, Magnusson M, Lundin S, Kaller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32(19):3047-8. doi: 10.1093/bioinformatics/btw354. PubMed PMID: 27312411; PMCID: PMC5039924.
12. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigo R, Hubbard TJ. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22(9):1760-74. doi: 10.1101/gr.135350.111. PubMed PMID: 22955987; PMCID: PMC3431492.
13. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21. doi: 10.1093/bioinformatics/bts635. PubMed PMID: 23104886; PMCID: PMC3530905.

14. DeLuca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, Reich M, Winckler W, Getz G. RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics. 2012;28(11):1530-2. doi: 10.1093/bioinformatics/bts196. PubMed PMID: 22539670; PMCID: PMC3356847.
15. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923-30. doi: 10.1093/bioinformatics/btt656. PubMed PMID: 24227677.
16. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139-40. doi: 10.1093/bioinformatics/btp616. PubMed PMID: 19910308; PMCID: PMC2796818.
17. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Giron CG, Gil L, Gordon L, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, To JK, Laird MR, Lavidas I, Liu Z, Loveland JE, Maurel T, McLaren W, Moore B, Mudge J, Murphy DN, Newman V, Nuhn M, Ogeh D, Ong CK, Parker A, Patricio M, Riat HS, Schuilenburg H, Sheppard D, Sparrow H, Taylor K, Thormann A, Vullo A, Walts B, Zadissa A, Frankish A, Hunt SE, Kostadima M, Langridge N, Martin FJ, Muffato M, Perry E, Ruffier M, Staines DM, Trevanion SJ, Aken BL, Cunningham F, Yates A, Flicek P. Ensembl 2018. Nucleic Acids Res. 2018;46(D1):D754-D61. doi: 10.1093/nar/gkx1098. PubMed PMID: 29155950; PMCID: PMC5753206.
18. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. 2011;39(Database issue):D52-7. doi: 10.1093/nar/gkq1237. PubMed PMID: 21115458; PMCID: PMC3013746.
19. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):R25. doi: 10.1186/gb-2010-11-3-r25. PubMed PMID: 20196867; PMCID: PMC2864565.
20. Flowers E, Miaskowski C, Conley Y, Hammer M, Levine J, Mastick J, Paul S, Wright S, Kober K. Differential Expression of Genes and Differentially Perturbed Pathways Associated with Very High Evening Fatigue in Oncology Patients Receiving Chemotherapy. Support Care Cancer. *2018*;26(3):739-50. Epub 2017 Sep 25. PubMed PMID: 28944404; PMCID: PMC5786467
21. Kober KM, Dunn L, Mastick J, Cooper B, Langford D, Melisko M, Venook A, Chen LM, Wright F, Hammer M, Schmidt BL, Levine J, Miaskowski C, Aouizerat BE. Gene Expression Profiling of Evening Fatigue in Women Undergoing Chemotherapy for Breast Cancer. Biol Res Nurs. 2016;18(4):370-85. doi: 10.1177/1099800416629209. PubMed PMID: 26957308.
22. Dunning MJ, Smith ML, Ritchie ME, Tavare S. beadarray: R classes and methods for Illumina bead-based data. Bioinformatics. 2007;23(16):2183-4. Epub 2007/06/26. doi: 10.1093/bioinformatics/btm311. PubMed PMID: 17586828.
23. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5(10):R80. doi: gb-2004-5-10-r80 [pii] 10.1186/gb-2004-5-10-r80. PubMed PMID: 15461798; PMCID: PMC545600.
24. Reimers M. Making informed choices about microarray data analysis. PLoS Comput Biol. 2010;6(5):e1000786. Epub 2010/06/05. doi: 10.1371/journal.pcbi.1000786. PubMed PMID: 20523743; PMCID: 2877726.
25. Ritchie ME, Dunning MJ, Smith ML, Shi W, Lynch AG. BeadArray expression analysis using bioconductor. PLoS Comput Biol. 2011;7(12):e1002276. doi: PCOMPBIOL-D-11-00763 [pii] 10.1371/journal.pcbi.1002276. PubMed PMID: 22144879; PMCID: PMC3228778.
26. Butte A. The use and analysis of microarray data. Nat Rev Drug Discov. 2002;1(12):951-60. Epub 2002/12/04. doi: 10.1038/nrd961. PubMed PMID: 12461517.
27. Kauffmann A, Huber W. Microarray data quality control improves the detection of differentially expressed genes. . Genomics. 2010;95:138-42. doi: 10.1016/j.ygeno.2010.01.003.

28. Kauffmann A, Gentleman R, Huber W. arrayQualityMetrics--a bioconductor package for quality assessment of microarray data. . Bioinformatics. 2009;25(3):415-6. doi: 10.1093/bioinformatics/btn647.
29. Reimers M. Making Informed Choices about Microarray Data Analysis. PLos Computational Biology. 2010;6(5):e1000786.
30. Shi W, Oshlack A, Smyth GK. Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. Nucleic Acids Research. 2010;38(e204). doi: 10.1093/nar/gkq871.
31. Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, Holloway A, Smyth GK. A comparison of background correction methods for two-colour microarrays. Bioinformatics. 2007;23:2700-707. doi: 10.1093/bioinformatics/btm412.
32. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3(9):1724-35. Epub 2007/10/03. doi: 10.1371/journal.pgen.0030161. PubMed PMID: 17907809; PMCID: 1994707.
33. Leek JT, Johnson WE, Parker HS, Jaffe AEa, Storey J, D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012. doi: DOI:10.1093/bioinformatics/bts034.
34. Landau WM, Liu P. Dispersion estimation and its effect on test performance in RNA-seq data analysis: a simulation-based comparison of methods. PLoS One. 2013;8(12):e81415. doi: 10.1371/journal.pone.0081415. PubMed PMID: 24349066; PMCID: PMC3857202.
35. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012;40(10):4288-97. doi: 10.1093/nar/gks042. PubMed PMID: 22287627; PMCID: PMC3378882.
36. Cox D, Reid N. Parameter orthogonality and approximate conditional inference. J Roy Stat Soc Ser B Method. 1987;49:1-39.
37. Flowers E, Flentje A, Levine J, Olshen A, Hammer M, Paul S, Conley Y, Miaskowski C, Kober K. A Pilot Study Using a Multi-Staged Integrated Analysis of Gene Expression and Methylation to Evaluate Mechanisms for Evening Fatigue Biol Res Nurs. *In Press.*
38. Smyth G. Limma: Linear models for microarray data. In: R. C. Gentleman VJC, S. Dudoit, R. Irizarry, & W. Huber (Eds.), editor. Bioinformatics and computational biology. New York, NY: Springer; 2005. p. 397-420.
39. McCarthy DJ, Smyth GK. Testing significance relative to a fold-change threshold is a TREAT. Bioinformatics. 2009;25:765-71.
40. Fisher RA. Statistical Methods for Research Workers. Edinburgh: Oliver and Boyd; 1925.
41. Fisher RA. "Questions and answers #14". The American Statistician. 1948;2(5):30-1. doi: 10.2307/2681650.
42. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. Stat Med. 1990;9(7):811-8.
43. Mitrea C, Taghavi Z, Bokanizad B, Hanoudi S, Tagett R, Donato M, Voichita C, Draghici S. Methods and approaches in the topology-based analysis of biological pathways. Front Physiol. 2013;4:278. doi: 10.3389/fphys.2013.00278. PubMed PMID: 24133454; PMCID: PMC3794382.
44. Draghici S, Khatri P, Tarca AL, Amin K, Done A, Voichita C, Georgescu C, Romero R. A systems biology approach for pathway level analysis. Genome Res. 2007;17(10):1537-45. doi: 10.1101/gr.6202607. PubMed PMID: 17785539; PMCID: PMC1987343.
45. Aoki-Kinoshita KF, Kanehisa M. Gene annotation and pathway mapping in KEGG. Methods in molecular biology (Clifton, NJ. 2007;396:71-91. doi: 1-59745-515-6:71 [pii]. PubMed PMID: 18025687.
46. Gray KA, Daugherty LC, Gordon SM, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2013. Nucleic Acids Res. 2013;41(Database issue):D545-52. Epub 2012/11/20. doi: 10.1093/nar/gks1066. PubMed PMID: 23161694; PMCID: 3531211.



Supplementary Figure 1: Flow diagram of number of patients available for phenotypic and gene expression (GE) analyses.