The American Journal of Human Genetics, Volume *108*

Supplemental information

An integrated approach to identify environmental modulators of genetic risk factors for complex traits

Brunilda Balliu, Ivan Carcamo-Orive, Michael J. Gloudemans, Daniel C. Nachun, Matthew G. Durrant, Steven Gazal, Chong Y. Park, David A. Knowles, Martin Wabitsch, Thomas Quertermous, Joshua W. Knowles, and Stephen B. Montgomery

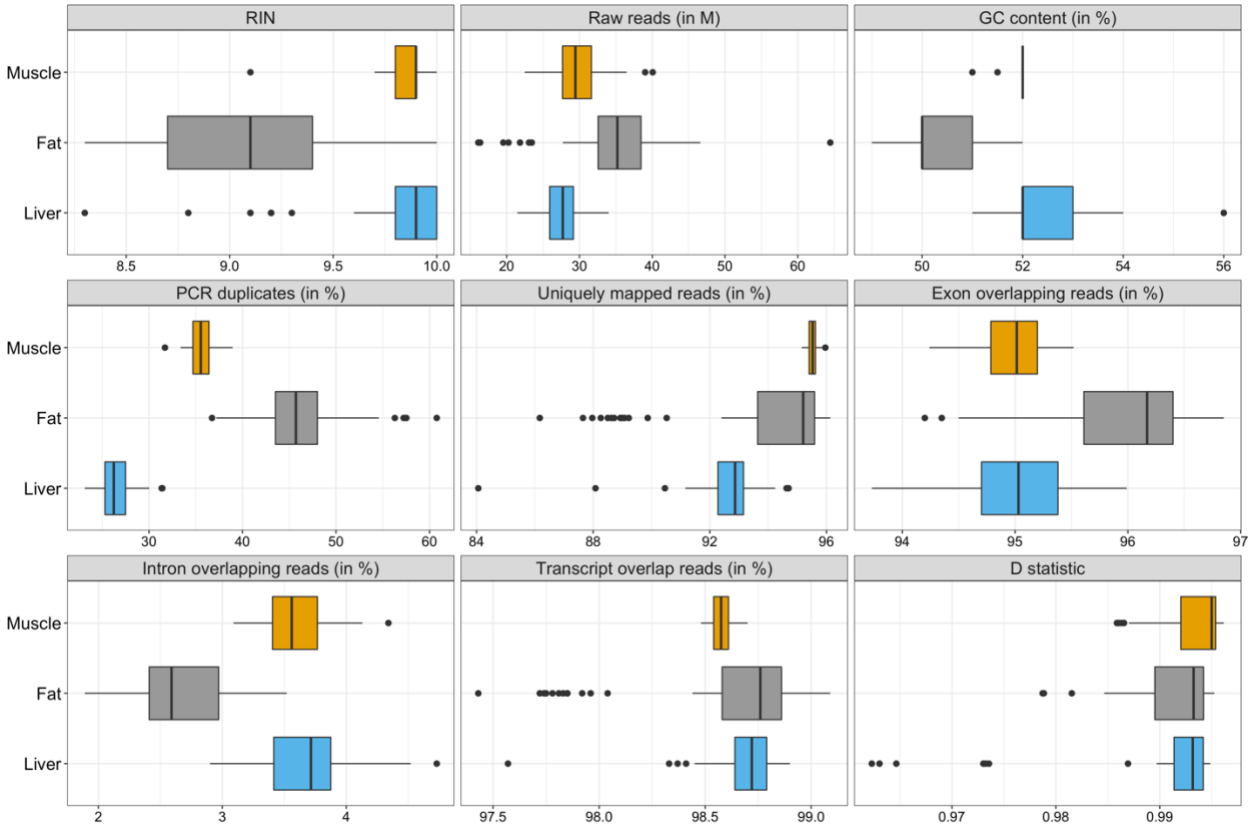

Figure S1: RNA-Seq data quality control (QC). Boxplot of RNA integrity number (RIN), number of sequenced reads (in M), % GC content, % of reads marked as PCR duplicates, % of uniquely mapped reads, % of exon, intron, and transcript overlapping reads, and median Spearman expression correlation (D-statistic) across samples that passed QC. All samples had RIN above 8 (mean = 9.5), at least 16M reads (mean = 30M), an average of 51% GC content, an average of 36% of reads marked as PCR duplicates, at least 84% of their reads mapped uniquely (mean = 95%), an average of 95%, 4%, and 98% exon, intron, and transcript overlapping reads. For all samples, their D-statistic was at least .96. M: millions.

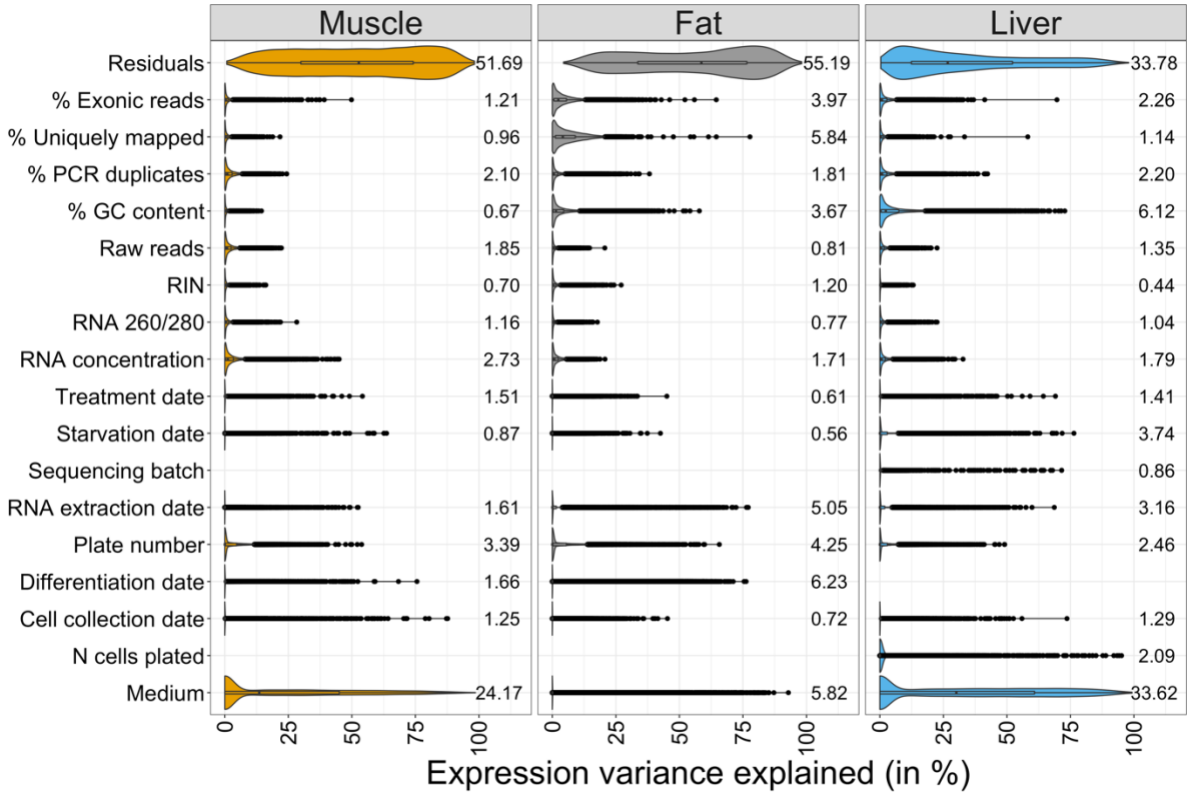

Figure S2: Identifying major components of variability in RNA-Seq data. Proportion of gene expression variance explained (VE) by technical metadata. Numbers next to the violin plots show the mean proportion of VE across all genes. To get the % of VE by each metadata for each gene, we used a linear mixed model with effect of medium, number of cells plated, plate number, sequencing batch, cell collection, differentiation, RNA extraction, starvation, and treatment date as random and the effects of all other variables as fixed effects. Sequencing batch and number of cells plated only differed for liver samples while differentiation date only differed for muscle and fat samples.

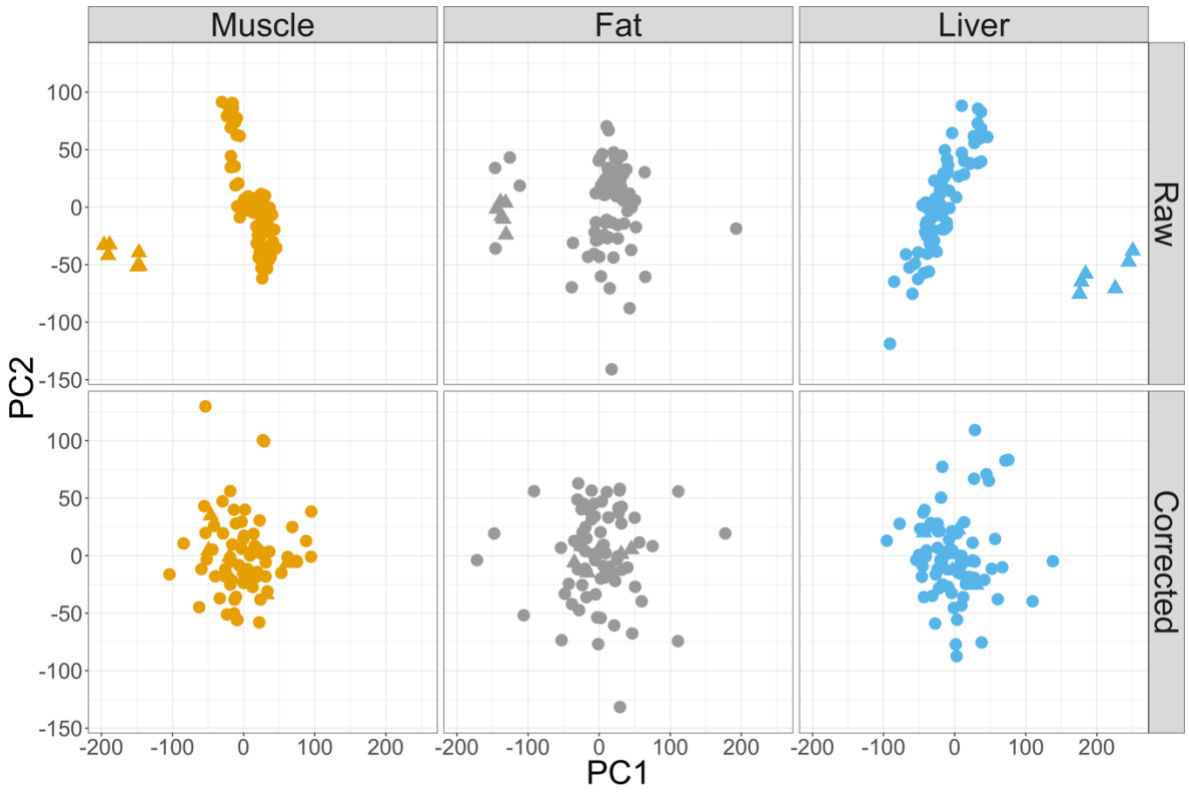

Figure S3: Detecting outliers in RNA-Seq data using Principal Component Analysis (PCA). Scatter plot of first two principal components (PCs). Shape indicates if the sample was glucose-related (triangle = glucose control or treated with glucose) or non-glucose-related (circle). PCA was applied separately to muscle, fat, and liver samples based on raw gene expression data (top panels) and expression residuals (bottom panel) after correcting for major components of expression variability as defined in section S6 (See Figure S2).

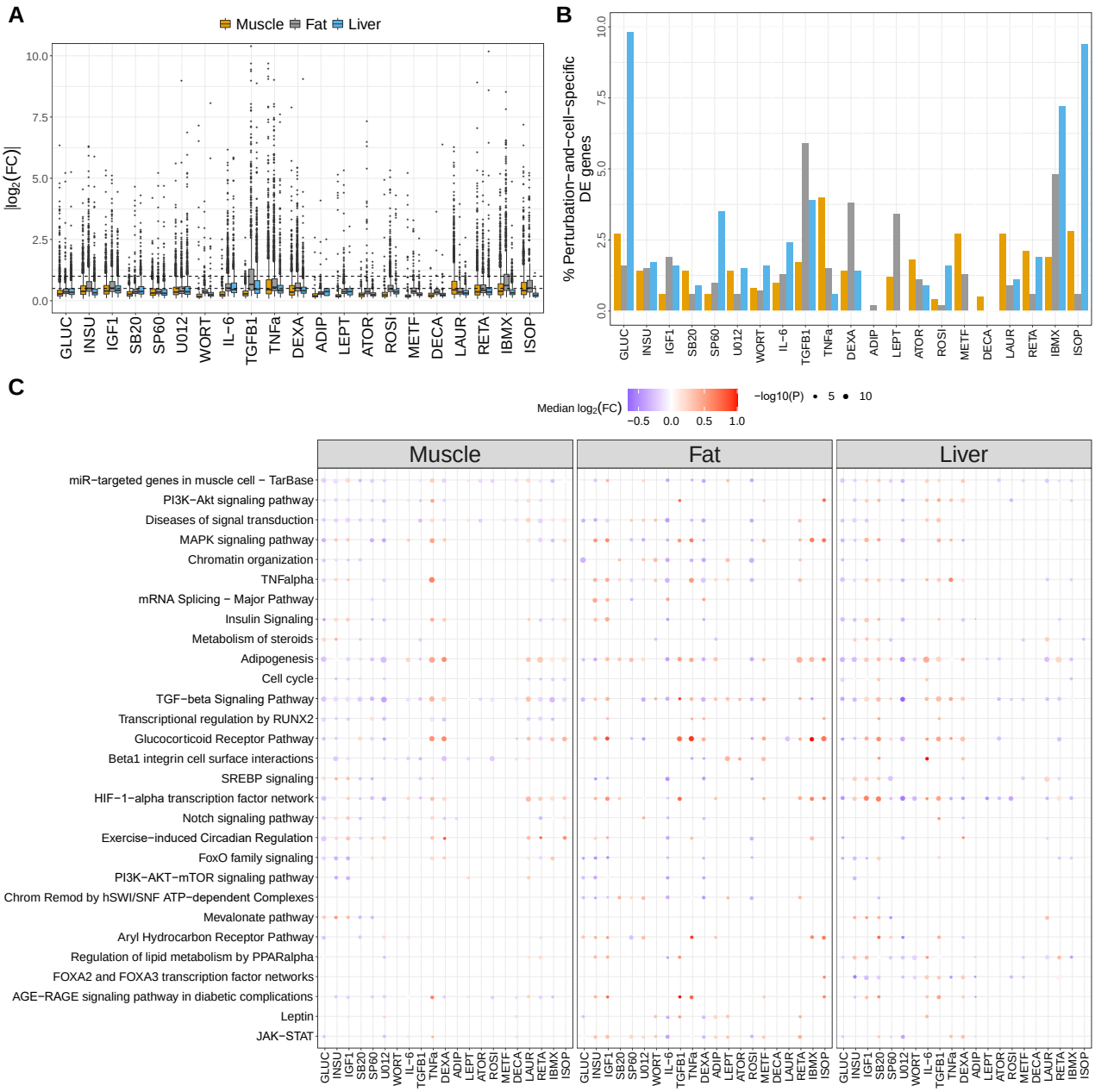

Figure S4: Transcriptome map of 21 perturbations across human skeletal muscle, fat, and liver cell line. (A) Absolute $\log_2$(FC) of DE genes (FDR<5%) for each perturbation and cell line. The two dashed lines indicate $|\log_2(FC)|$ of 0.5 and 1, the latter corresponding to a two-fold increase/decrease in gene expression after perturbation. (B) Sharing and specificity of transcriptional responses to environmental perturbations in muscle, fat, and liver. Percent of genes that show perturbation and cell line-specific expression, i.e., they are DE in a specific perturbation and cell line (FDR<5%). **(C)** Pathway enrichment analysis of DE genes. The dot size represents the significance of enrichment. The color represents the direction of transcriptional regulation of genes in each pathway (blue: downregulated, red: upregulated). Median $\log_2$(FC) has been censored at (-1,1) for ease of visualization. FC: fold change. DE: differentially expressed.

Figure S5: Identifying environmental perturbations impacting significant GWAS loci. GWAS enrichment results for complex traits from the GWAS catalog. Each point represents a perturbation-cell-line combination that passes the FDR<10% cut off; color indicates the cell line. The y-axis represents the -$\log_{10}$(P-value) and the size indicates the odds ratio OR for the enrichment of GWAS hits of each trait from the GWAS catalog. The shading color within each panel indicates the perturbation category from Figure 1A. Numerical results are reported in Table SX.

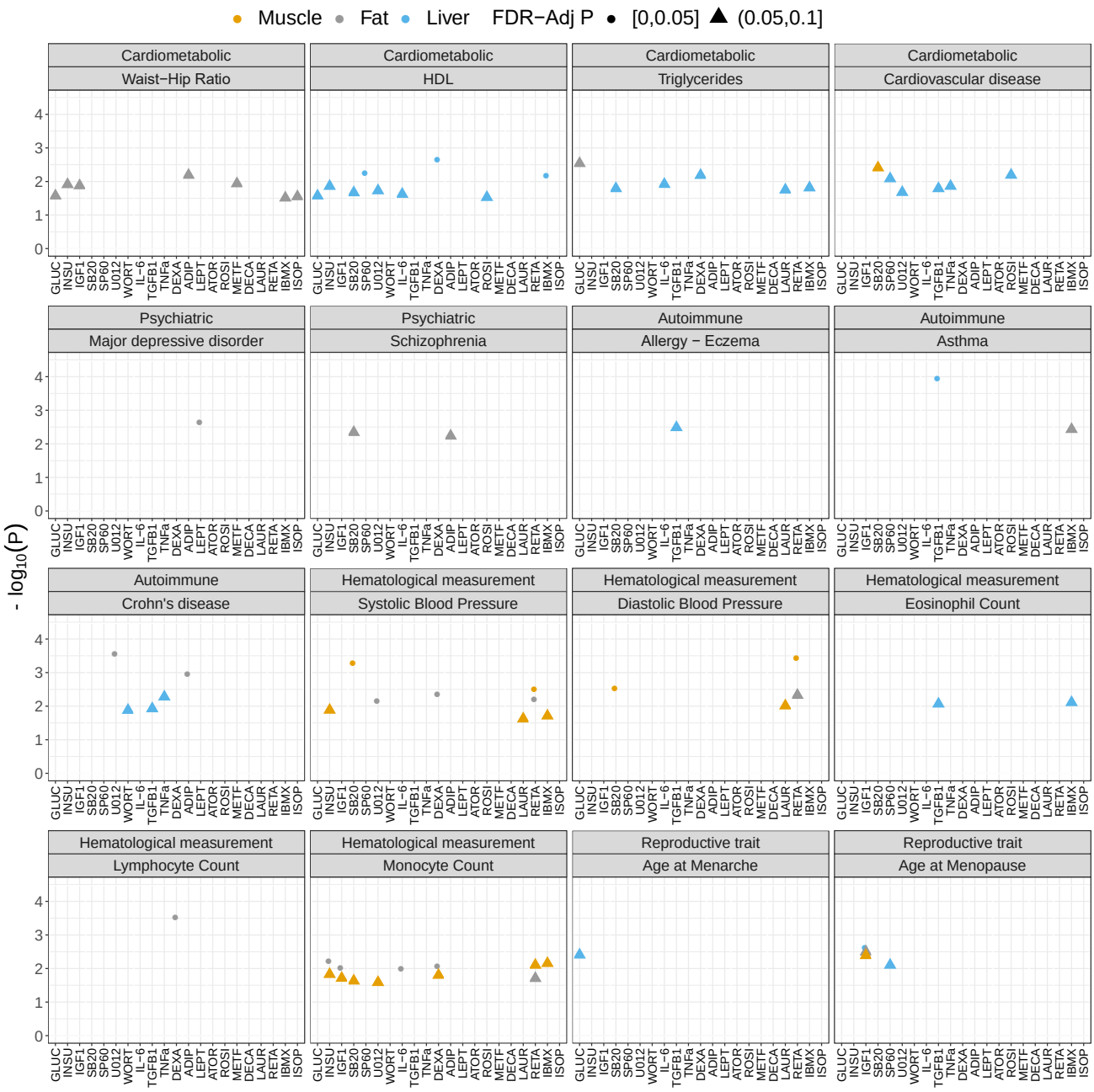

Figure S6. Prioritizing complex disease-relevant environmental perturbations via heritability enrichment analysis. Heritability enrichment results for each complex trait for different FDR thresholds. Each point represents a perturbation-cell-line combination that passes the FDR<10% cut-off. The y-axis represents the -$\log_{10}$(P-value) of heritability enrichment, the x-axis indicates perturbation, the color of the point indicates cell line, and the shading color within each panel indicates the perturbation category from Figure 1A. Finally, the shape indicates if the FDR-adjusted P-value passes the 5% or 10% threshold.

Figure S7. Identifying environmental perturbations impacting significant GWAS loci. GWAS enrichment results for each group of complex traits from the GWAS catalog for different FDR thresholds. Each point represents a perturbation-cell- line combination that passes the FDR<10% cut-off; the color of the point indicates the cell line and the shading color within each panel indicates the perturbation category from Figure 1A. The y-axis represents the -$\log_{10}$(P-value) of Fisher's exact test and the size indicates the odds ratio for the enrichment of GWAS hits of each group of traits from the GWAS catalog. Finally, the shape indicates if the FDR-adjusted P-value passes the 5% or 10% threshold.

Table S1: Validation of DE genes using external data sets. Each row lists, for each study, the cell, compound, concentration, exposure time, the assay used, the number of DE genes found in the original study that is expressed and tested in our study, and the validation method and ratio. For method = "% with FDR<5%" the validation ratio corresponds to the proportion of DE genes that are expressed in our study which show a significant response to the same perturbation and cell line in our experiment at FDR < 5%. For method = "π1", the validation method refers to the estimated proportion of true positives using the q-value method. Neither method requires the direction of differential expression to match as direction effect was not available for must studies.

citation	cell	compound	concentration	time	assay	# DE genes expressed in our study	Validation Method	Validation ratio
PMID: 25153832	HepG2	ATOR	10uM	24hrs	RNAseq	90	% with FDR<5%	89.69%
PMID: 26217794	HepG2	Il-6	50ng/ml	6/24hrs	RNAseq	58	% with FDR<5%	65.51%
DOI: 10.2174/1876524600902010005	SGBS	INSU	100nM	24hrs	array	77	π1	76.11%
DOI: 10.2174/1876524600902010005	SGBS	IGF-1	50ng/ml (6.58nM)	24hrs	array	27	% with FDR<5%	74.07%
PMID: 17066518	SGBS	TNFa	50ng/ml	24hrs	array	8	% with FDR<5%	62.25%
PMID: 30893613	Liver	INSU	4/12 mU/kg/min	20min/3h	RNAseq	91	π1	56.48%
PMID: 30893613	Muscle	INSU	4/12 mU/kg/min	20min/3h	RNAseq	79	% with FDR<5%	63.29%