Supporting Information

Exposome Profiling of Environmental Pollutants in Seminal Plasma and Novel Associations with Semen Parameters.

Haotian Wu[1$*], Vrinda Kalia[1$], Katherine E. Manz[2], Lawrence Chillrud[1], Nathalie Hoffmann Dishon[3], Gabriela L. Jackson[1], Christian K. Dye[1], Raoul Orvieto[3], Adva Aizer[3], Hagai Levine[4], Marianthi-Anna Kioumourtzoglou[1], Kurt D. Pennell[2], Andrea A. Baccarelli[1], Ronit Machtinger[3, 5]

[$]H.W. and V.K. contributed equally to this paper

[1] Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, NY, 10032, USA

[2] School of Engineering, Brown University, Providence, Rhode Island, 02912, USA.

[3] Infertility and IVF Unit, Department of Obstetrics and Gynecology, Chaim Sheba Medical Center (Tel Hashomer), Ramat Gan, 5262000, Israel

[4] Braun School of Public Health and Community Medicine, Hadassah Medical Center, The Faculty of Medicine, Hebrew University of Jerusalem, Israel.

[5] School of Medicine, Tel-Aviv University, Tel Aviv, 6997801, Israel.

Summary: 14 pages, including 11 supplemental figures and 6 supplemental tables

Additional Methods

Supplemental Figures

Supplemental Tables

Additional Methods

QC Spikes and Samples. Three types of QC samples were included: an extraction blank, a spike sample, and NIST Standard Reference Material (SRM) 1958. Extraction blanks were prepared using charcoal stripped fetal bovine serum (FBS) (Millipore Sigma, St. Louis, MO). NIST Standard reference material (SRM) 1958 was purchased from Millipore Sigma (St. Louis, MO). Spike samples were prepared by spiking a known concentration (0.40 – 40 ppb) of a calibration standard into charcoal stripped FBS. The recoveries were between 85 and 120%. The intended use of the QC samples was to monitor instrument performance.

The calibration standard contained 13 certified references standard mixtures. The following certified reference standards were purchased from AccuStandard: Furan Mix, Dioxin Mix, PBDE Congeners of Primary Interest Calibration Mix, Pesticide Mix 1, Pesticide Mix 2, AccuGrand 8270 Semi-Volatile Standard (AG01), Method 525.2 Organochlorine Pesticides, Pesticide/Herbicide Mix, Triphenyl phosphate, WHO/NIST/NOAA Congener List, tris(2-Chloroethyl) phosphate (TCEP), PCB Congeners Mix 2, and Pesticide/Herbicide Mix. NDEA was a component of AccuGrand 8270 Semi-Volatile Standard.

Further, each sample was spiked with internal standard mix (Phenanthrene-d10 and Chrysene-d12), and Carbon Distribution Marker (retention time marker) purchased from AccuStandard (DRH-TX-003-CNM). The Carbon Distribution Marker contains n-hexane, n-heptane, n-octane, n-decane, n-dodecane, n-hexadecane, n-heneicosane, n-octacosane, and n-pentatriacontane.The internal standard was used to evaluate retention time consistency and monitor peak area. The Carbon Distribution Marker was used for processing non-targeted data.

Analytical Sequence. Samples were extracted in 6 batches of 20 or less samples in a randomized order. QC samples (one blank, spike sample, and NIST SRM) were included in each extraction batch. Three analytical runs were performed, each with a calibration curve. The first analytical batch contained extraction batches 1 and 2, the second contained extraction batches 3 and 4, and the final analytical batch contained extraction batches 5 and 6. A calibration standard was analyzed every 10 injections to evaluate changes in the target analytes peak area.

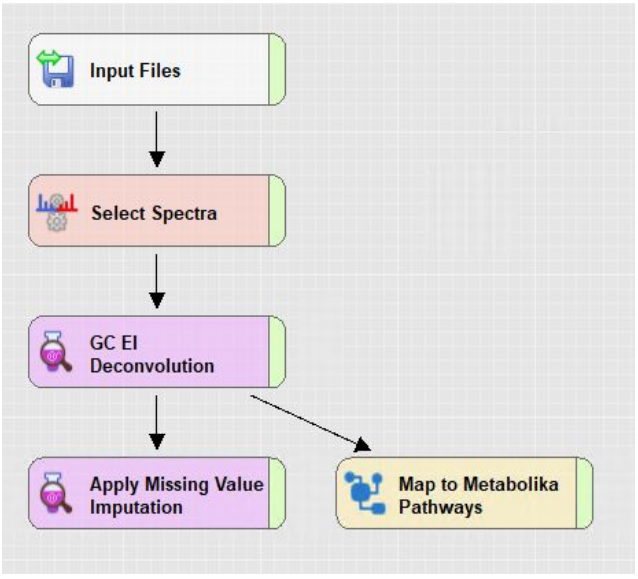
We used Thermo TraceFinder 4.0 software to monitor performance (e.g., standard area and retention time) throughout the analytical sequence.

Chromatography and Mass spectrometry. The table below describes the chromatography and mass spectrometry specifications used in this study. The GC-HRMS instrument was tuned and calibrated prior to each analytical sequence.

Inlet Parameters	
Injection Volume	4 µL
Inlet	Split/Splitless
Inlet Mode	Splitless
Inlet Temperature	290 °C
Splitless Time	1 min
Carrier Gas Flow Rate	1 mL/min
Oven Parameters	
Analytical Column	Restek Rxi-35Sil MS column (30 m x 0.25 mm inner diameter x 0.25 µm film thickness)
Carrier Gas	Helium (99.9999% purity)
Carrier Gas Flow Rate	1 mL/min
Oven Temperature Program	40°C for 0.5 min, increased 7°C/min to 240°C and held for 3 minutes, increased 10 °C/min to 295 °C and held for 3 minutes, and finally 10°C/min increased to 350 °C and held for 7 min
Transfer Line	300 °C
Total Run Time	53 min
Electron Ionization (EI) Source and Full Scan Parameters	
Filament Delay	4.6 minutes
Source Temperature	230°C
Source Voltage	70 eV
Resolution	60,000
AGC target	$1x10^6$
Maximum IT	auto
Scan Range	50 to 750 m/z

Data Processing. All spectral data files were saved in the .RAW file format and NTA/SSA was performed in Thermo Compound Discoverer (CD) 3.2 software. To evaluate CD settings, we analyzed the calibration samples to evaluate workflow parameters (e.g., mass accuracy, retention time, peak area). The CD data processing workflow and nodes are provided below. Peaks were detected with 10 ppm mass tolerance, 10,000 total ion chromatogram (TIC) threshold, signal to noise ratio of 3, and 99% allowable ion overlap. Each chromatogram was retention time aligned using the carbon distribution marker (contains 9 alkanes; only compounds containing greater than 8 carbons were used since the compounds smaller than this eluted during the solvent delay) spiked into each sample and retention indices (RIs) were calculated for each peak detected. The peak area for each putatively identified compound detected was exported to Microsoft Excel after processing the raw data and prior to data filtering.

The RI of each peak was used to limit suspects during identification; the allowed maximum RI difference was 300. Compounds were identified by searching their mass spectra in the NIST Mass Spectra Library (NIST/EPA/NIH EI and NIST Tandem Mass Spectral Library Version 2.3) and a high-resolution library developed in-house using certified standards containing 354 unique compounds. A minimum Match Factor (SI) and Reverse Match Factor (RSI) score of 500 was used for assigning library matches. Peaks with scores less than 500 were not assigned the identification. Chemicals that matched to our in-house library were assigned Level 1 annotation if they were also detected in our standard mixture.



Parameters of 'Map to Metabolika Pathways'

Show Advanced Parameters

1. Search Settings	
Metabolika Pathways	\(3R)-linalool biosynthesis.metabolika\2-nitrobenzoate
Search Mode	By Formula or Mass
2. By Mass Search Settings	
Mass Tolerance	10 ppm
3. By Formula Search Settings	
Max. # of Predicted Compositions to be	3
4. Display Settings	
Max. # Pathways in 'Pathways' column	20

Parameters of 'Apply Missing Value Imputation'

Show Advanced Parameters

1. General Settings	
Imputation Method	Random Forest Imputation

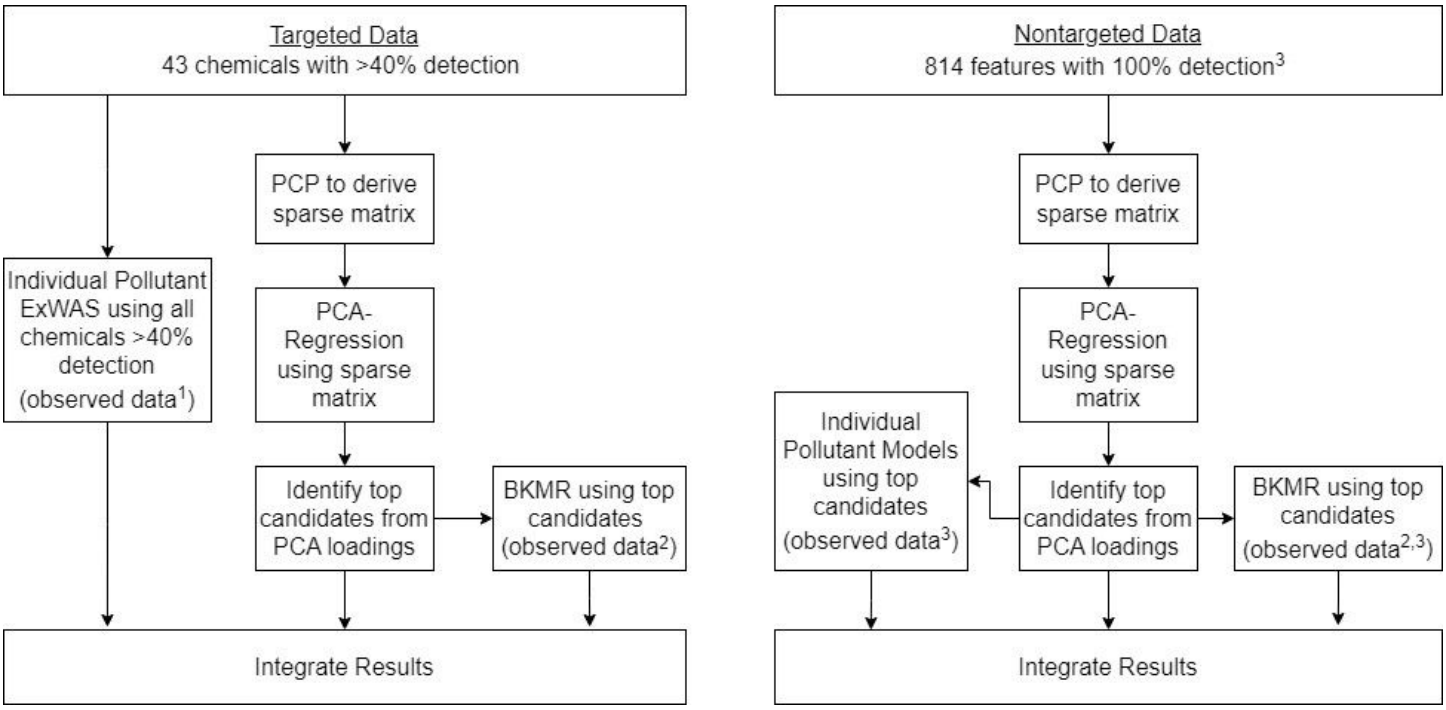
Parameters of 'GC EI Deconvolution'

Show Advanced Parameters

1. Peak Detection Settings	
Mass Tolerance	10 ppm
Spectral S/N Threshold	3
Peak S/N Threshold	3
Smoothing	9
TIC Threshold	10000
Ion Overlap Window [%]	99
Include Reference and Exception Peaks	True
2. RI Settings	
Calculate RIs	True
Column Type	Semi std non polar
n-Alkane Reference Items	8@6.48; 10@10.48; 12@17.96; 16@25.45; 21@35.02; 2
3. Group Compounds Settings	
Group Compounds Across Samples by S	True
RT Tolerance [sec]	10
Dot Product Threshold	500
Composition Threshold [%]	10
4. Library Search Settings	
Do Library Search	True
Library Search Type	Normal
Search Libraries	mainlib; Pennell GCHRMS; Pennell Lab GCHRMS
SI/RSI Threshold	500
Use Reverse Search	True
HRF Score	True
Show fragments	False
Penalize Missing Molecular Mass	False
Show Synonyms	True
Use RI Delta Filter	True
Maximum RI Delta	300

4. Library Search Settings	
Do Library Search	True
Library Search Type	Normal
Search Libraries	mainlib; Pennell GCHRMS; Pennell Lab GCHRMS
SI/RSI Threshold	500
Use Reverse Search	True
HRF Score	True
Show fragments	False
Penalize Missing Molecular Mass	False
Show Synonyms	True
Use RI Delta Filter	True
Maximum RI Delta	300
Use RI Diff[%] Filter	True
Maximum RI Diff [%]	30
Use Unspecified Column Type	False

Supplemental Figure 1: Overview of Statistical Approach



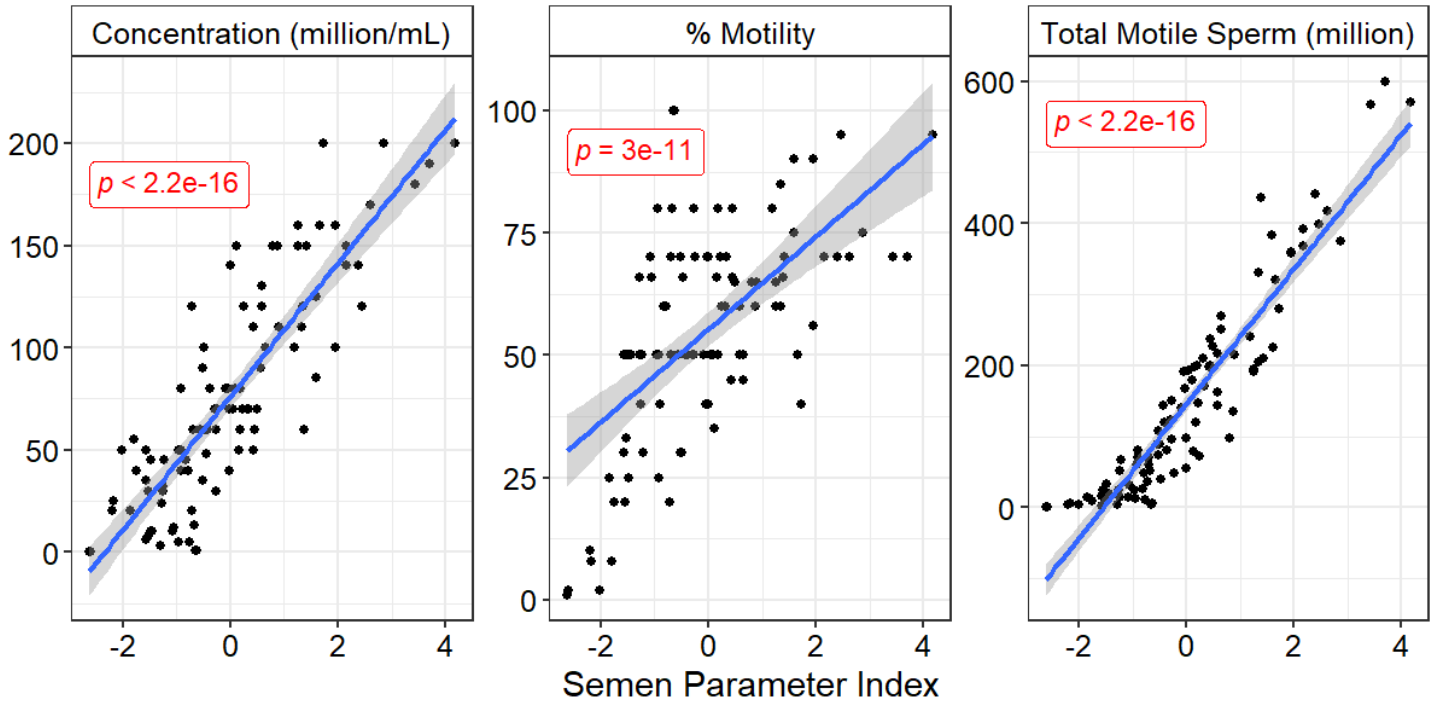
A visual representation of the statistical workflow for both the targeted and non-targeted data. Abbreviations: PCP: Principal Component Pursuit; PCA: Principal Component Analysis; ExWAS: Exposome Wide Association Study; BKMR: Bayesian Kernel Machine Regression.

[1] – The chemical data was modeled as binary (detected vs. non-detected [reference group]) if the detection rate was 40-70% and as continuous (log2-transformed) if the detection rate is ≥70%.

[2] – All chemical data were scaled prior to BKMR to ensure comparability across different mixture components.

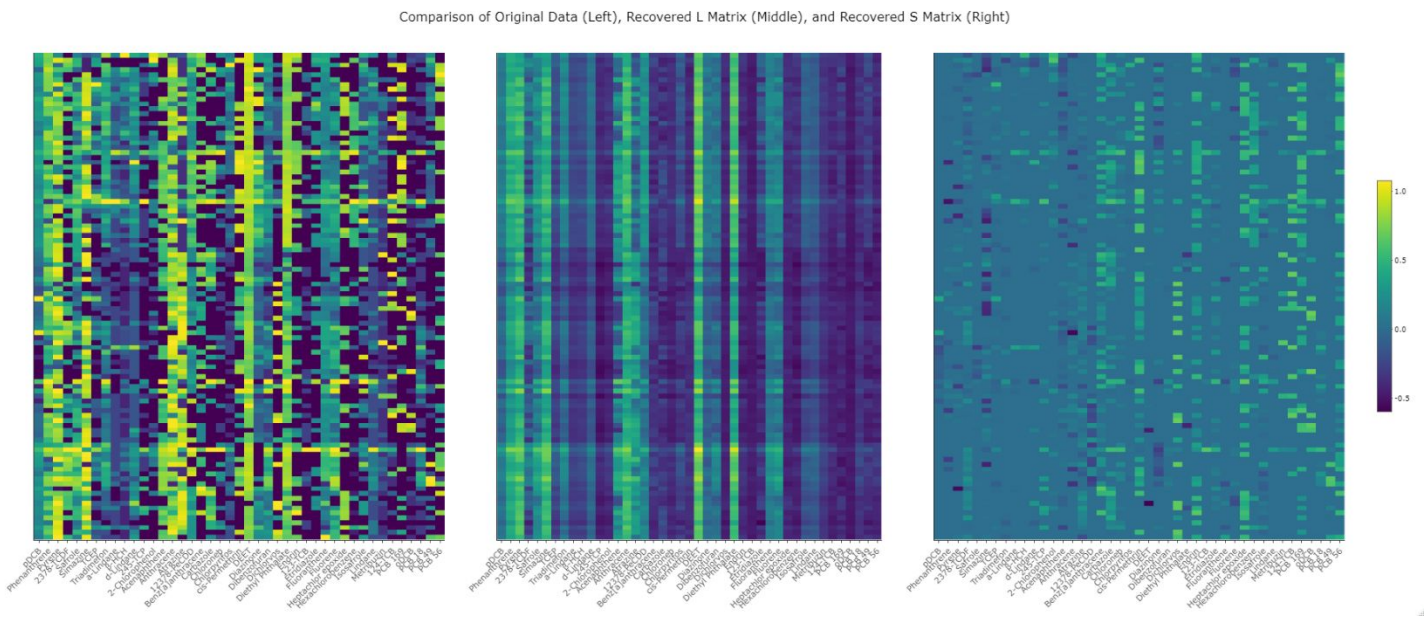
[3] – All nontargeted peaks selected for analysis were universally detected in our sample and treated as linear variables in the ensuing analysis.

Supplemental Figure 2: Correlation of Semen Parameter Index with Observed Semen Parameters (Volume, Concentration, % Motility, and Total Motile Sperm)



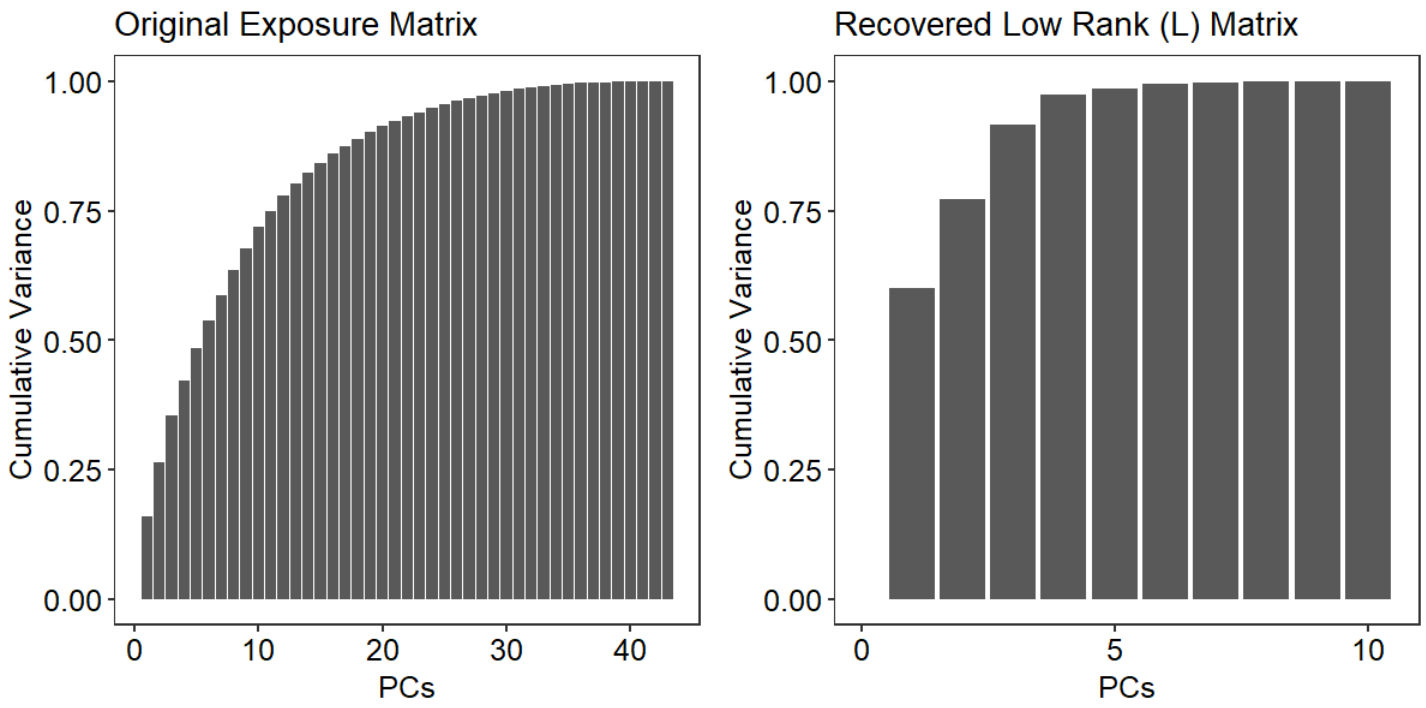
The strong positive correlations with semen concentration, % motility, and total motile sperm indicate that the index is a useful measure that encapsulates all three outcomes in a single composite index.

Supplemental Figure 3: Targeted Exposome Data - Comparison of Before vs. After Principal Component Pursuit



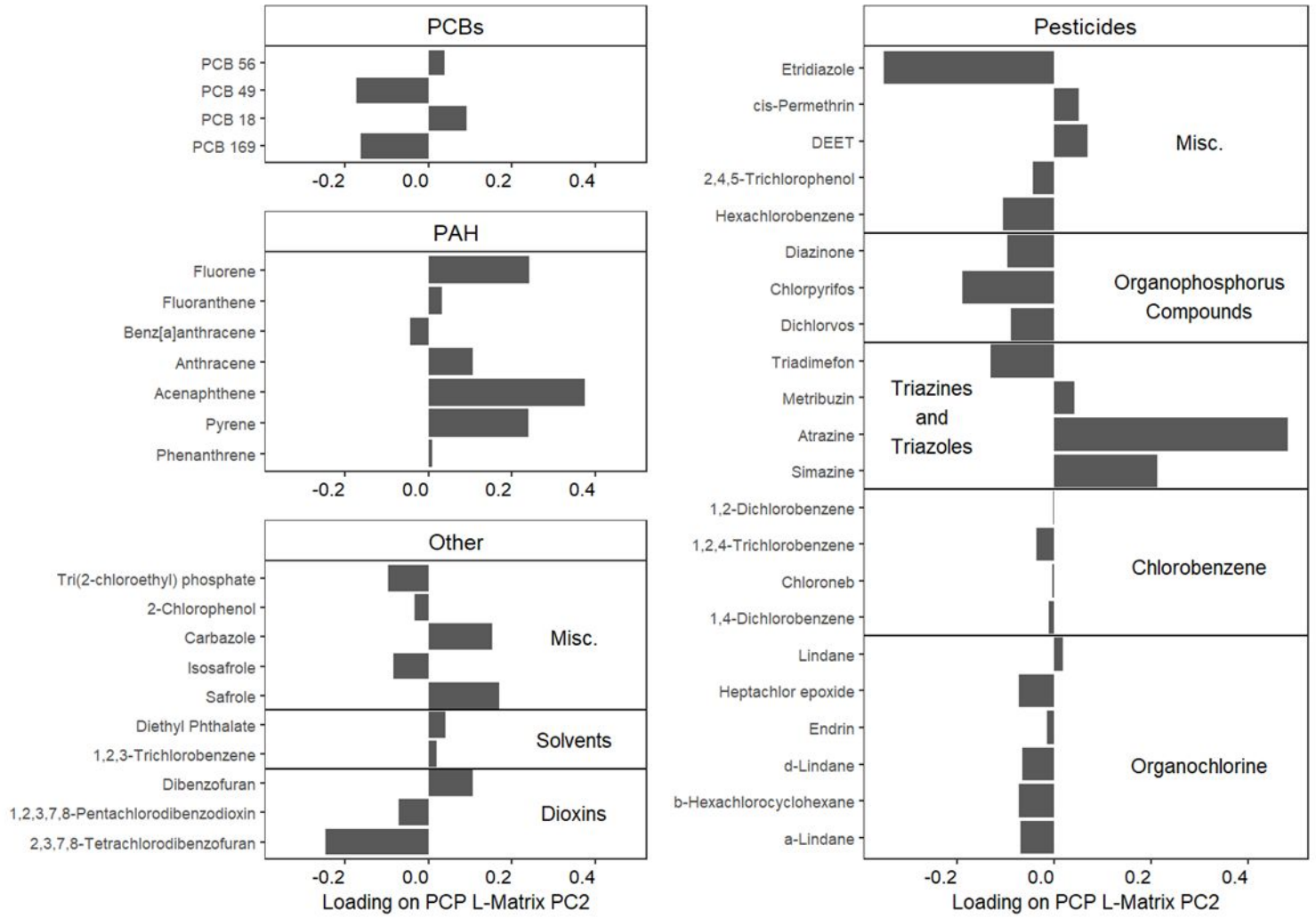
Heatplot illustration of the decomposition of the targeted exposome data comprising 43 chemicals with >40% detection rate. Y-axis (rows) are individuals and X-axis (columns) are each chemical. The original data (scaled), on the left, showed substantial variation that is later decomposed into the low rank L-matrix (middle) and sparse S matrix (right).

Supplemental Figure 4: Targeted Exposome Data - Cumulative Variance Explained Before vs. After Principal Component Pursuit



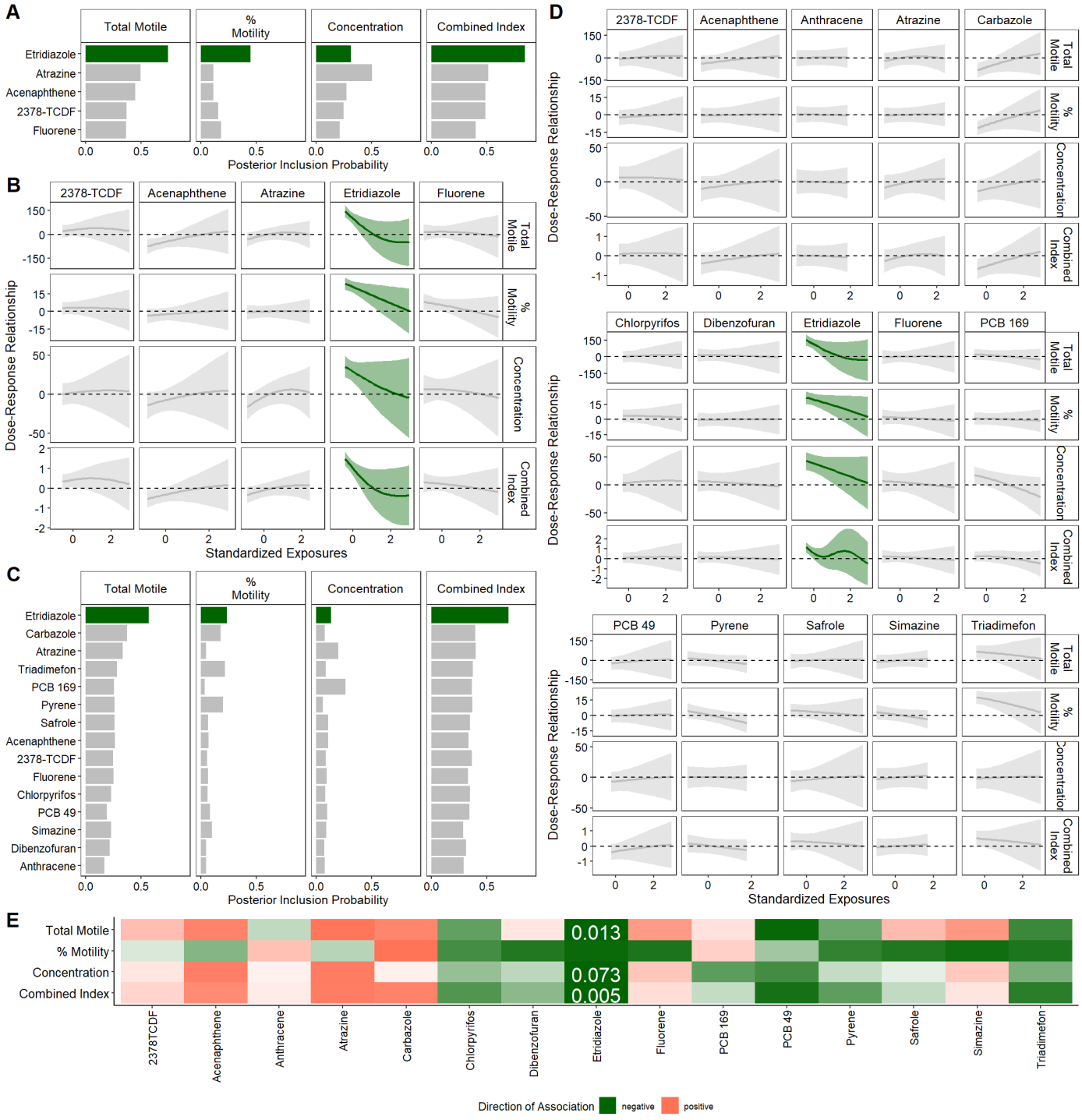
The original exposure matrix is noisy and is not easily amenable to factor analysis approaches to data dimension reduction. As shown here, principal component analysis of the original exposure matrix has resulted in numerous PCs where 34 PCs are required to explain ~99% of the variance in the data. However, after decomposition via Principal Component Pursuit (PCP), the resulting low rank L-matrix has considerably fewer components. Specifically, there are 8 PCs total for the L-matrix, explaining 59.9%, 17.4%, 14.2%, 5.9%, 1%, 0.7%, 0.4%, and 0.1% of variance each, respectively.

Supplemental Figure 5: Targeted Exposome Data - Chemical Loadings on Principal Component 2



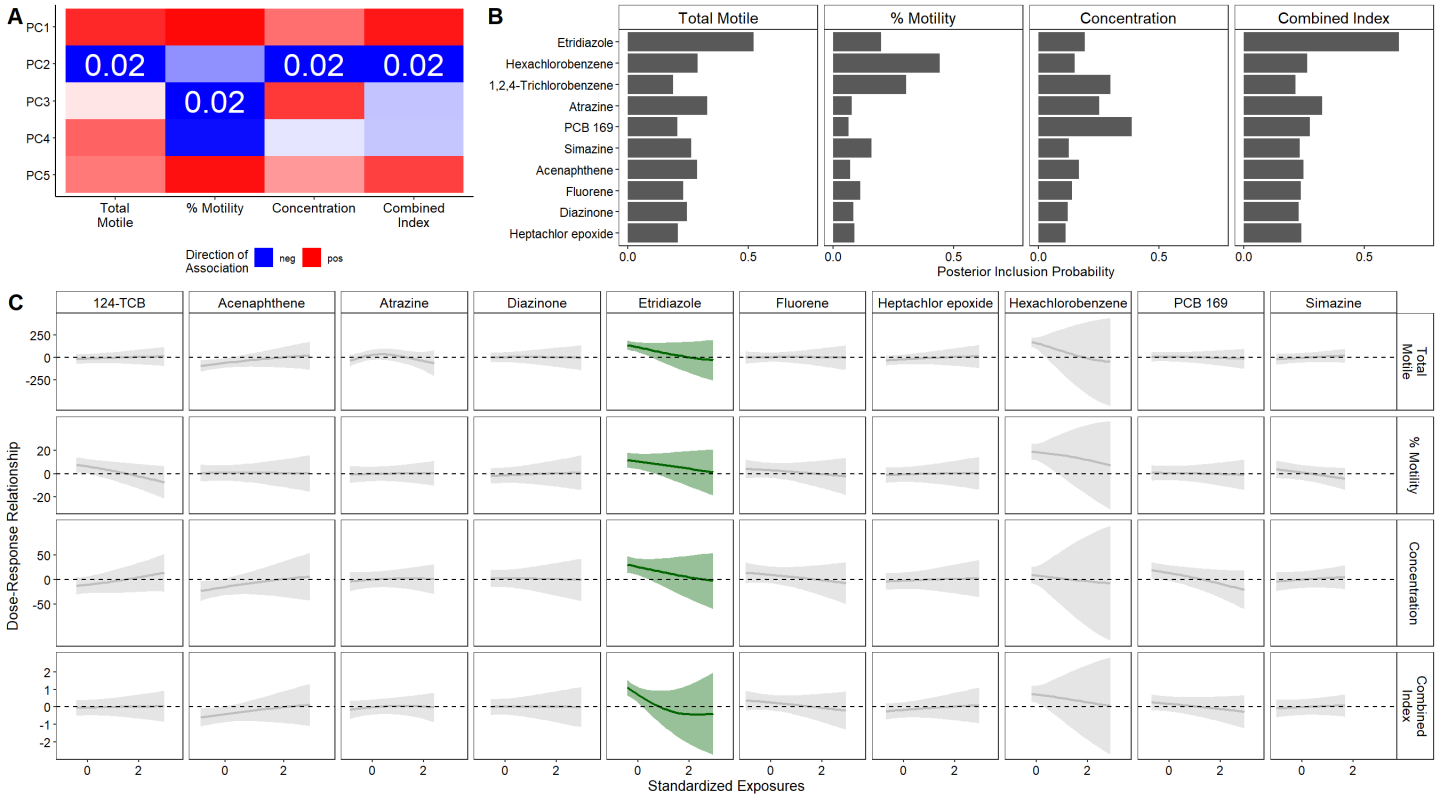
The loadings on PCP-PC2 showed high positive loadings for PAHs and triazine/triazole pesticides and highest negative loadings for Etridiazole and 2,3,7,8-Tetrachlorodibenzofuran. There was also moderately high loadings from several PAHs. Given that these chemicals all come from different classes, there does not appear to be an underlying exposure pattern. The top 5-15 chemicals were modeled jointly as a mixture in subsequent analyses.

Supplemental Figure 6: Targeted Exposome Data - BKMR Mixture Association using 5 (A+B) and 15 (C+D) components.



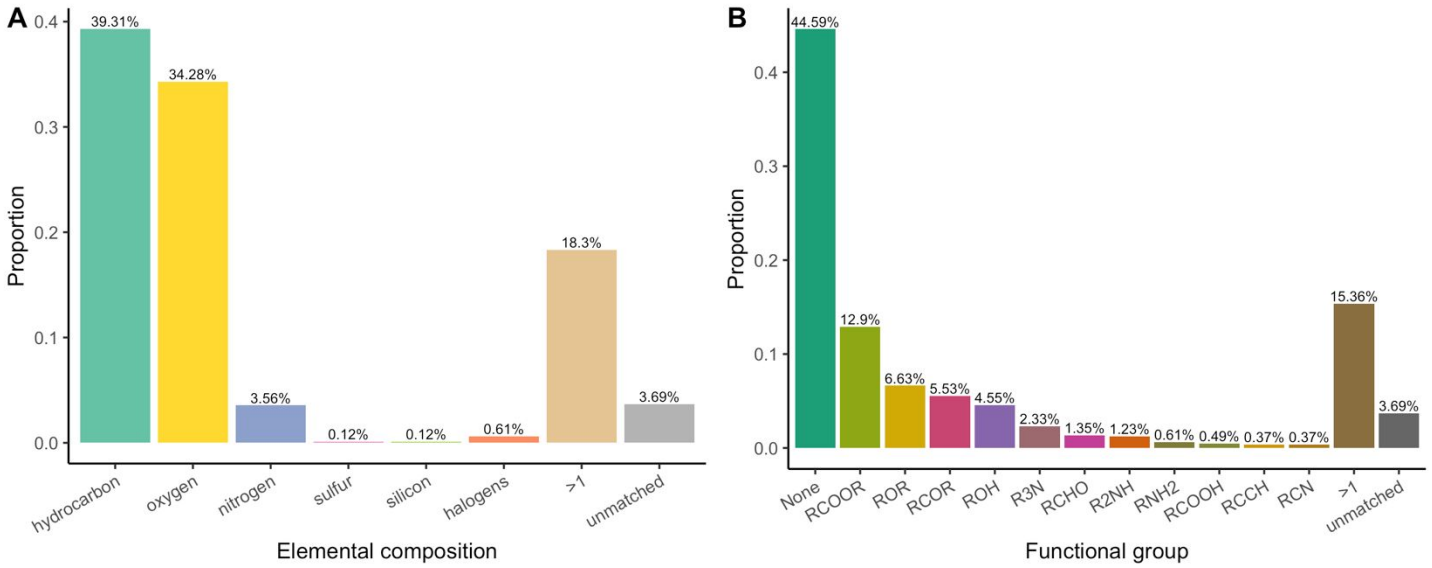
In models with only top 5 components, Etridiazole is the only chemical that displayed the highest posterior inclusion probability (A) and a plausible and statistically significant (i.e. crosses the credible interval) dose-response relationship with our semen parameter outcomes. (B) In the models with top 15 components, three possible significant associations were displayed, including Etridiazole in green and two other chemicals (Carbazole and Triadimefon). Etridiazole once again had the highest posterior inclusion probability for total motile sperm, % motility, and combined index (C). In comparison, Carbazole and Triadimefon had lower posterior inclusion probability and were generally not significant in individual exposure models from the (E) ExWAS analysis. Together, these models suggest our PCP-PCA-BKMR pipeline identifies Etridiazole as associated with semen parameters, regardless of our choice for the number of mixture components.

Supplemental Figure 7: Targeted Exposome Data – Alternative PCP Parameters



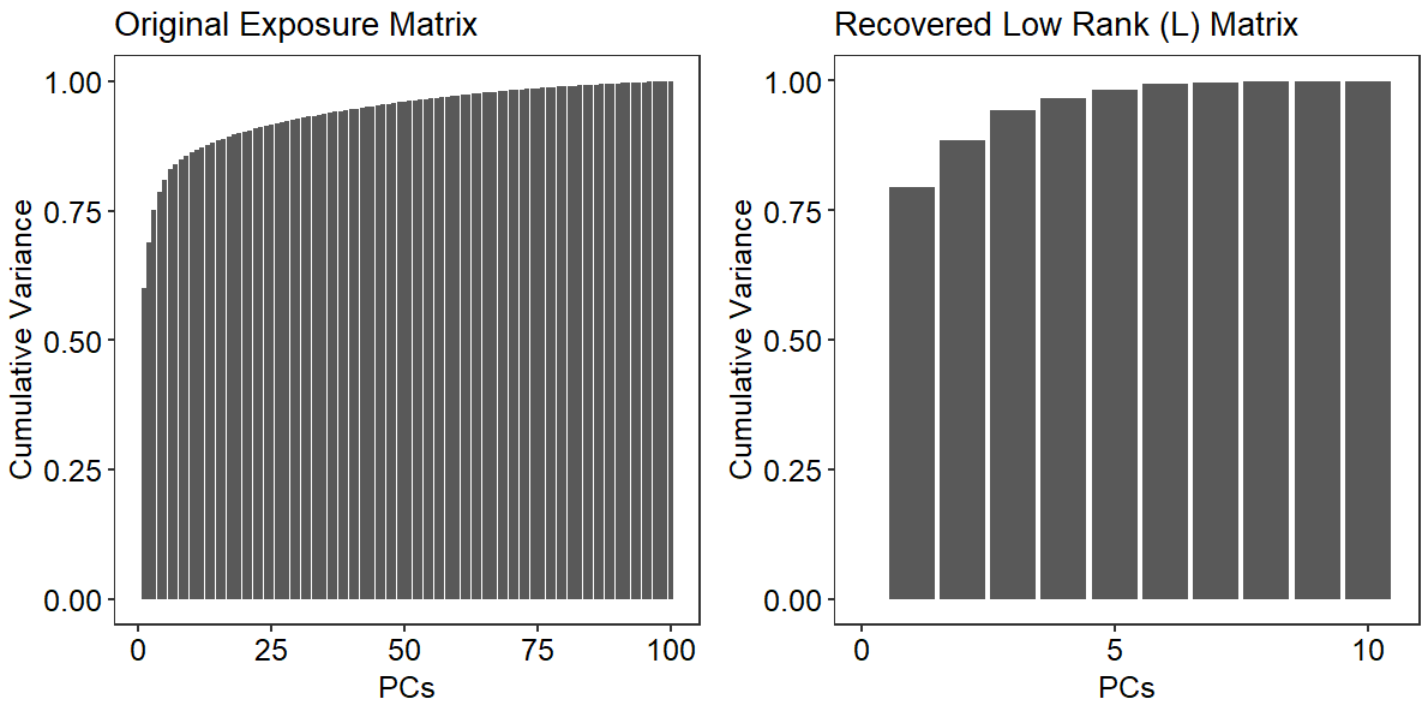
Panel A shows the nominal p-value (number) and direction (color) of the association between PCP-derived principal components and semen parameters. Thus, we again show that using grid-search derived alternative PCP parameters (λ and μ), we also find that PC2 is correlated with semen parameters (A). Then, taking the top 10 loadings from PC2 and putting them in a BKMR model showed that Etridiazole again shows the highest posterior inclusion probability (B) and a negative relationship with all four semen parameters (C). There was also some suggestive evidence that hexachlorobenzene may be related to total motile sperm and % motility, but this was not supported by individual exposure models from the ExWAS analysis (see Supplemental Table 1), making this unreliable and likely spurious. Together, these results show that our results are not sensitive to the starting parameters of PCP.

Supplemental Figure 8: Non-targeted Exposome Data (814 spectral peaks) – Chemical characteristics



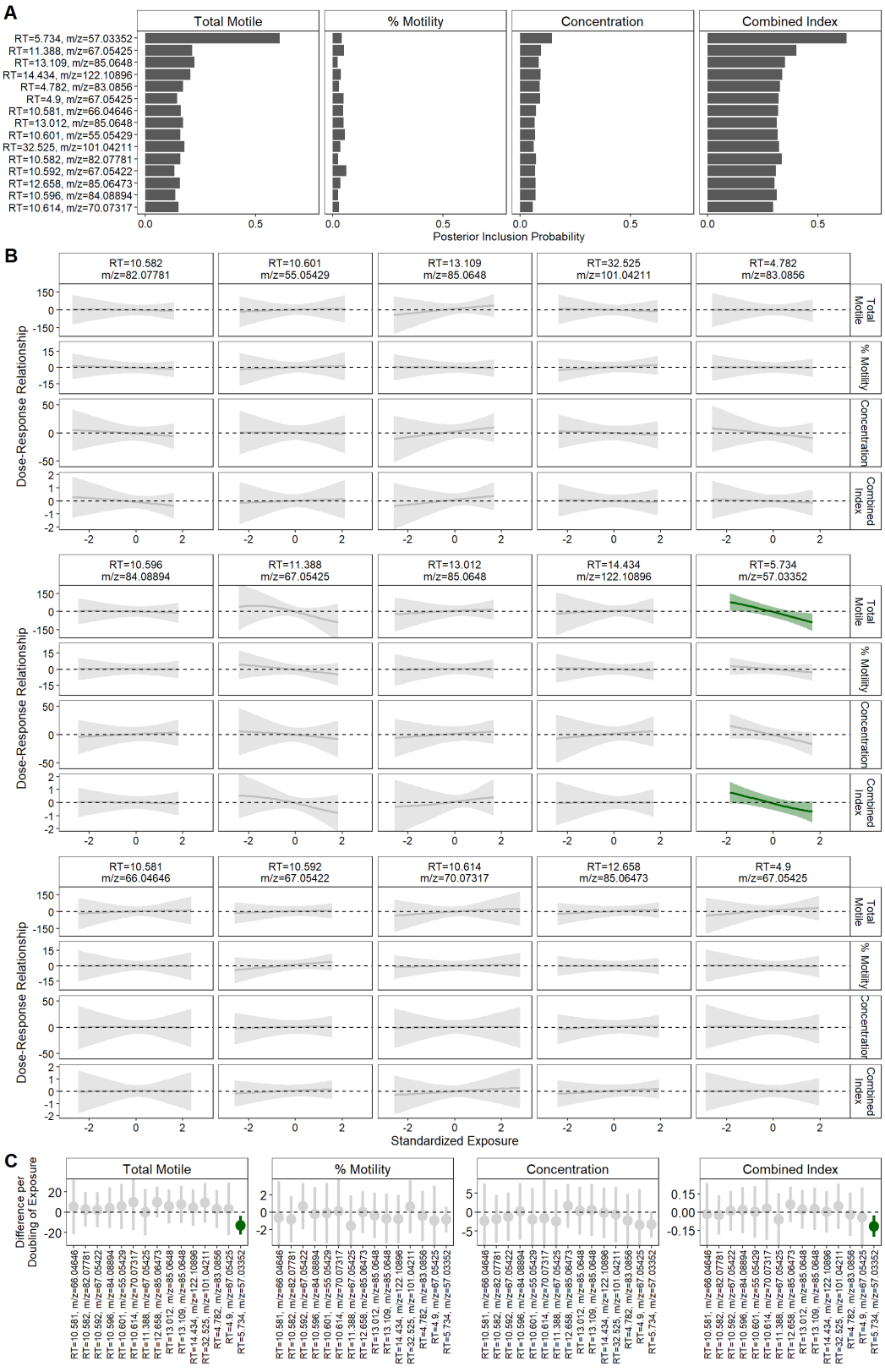
To describe the spectral peaks captured by the non-targeted assay, we used SMILES generated for each peak to perform chemical characterization. In A, 39% of the fragments were categorized as hydrocarbons (containing carbon and hydrogen atoms), with fragments containing an oxygen atom as the second most abundant. We also determined the proportion of fragments that contained nitrogen, sulfur, silicon, halogens (fluorine, bromine, chlorine), boron, or phosphorous, or a combination of elements other than carbon or hydrogen (>1). In B, we characterized the type of functional group present on the fragments, and found that nearly half of the fragments did not contain a function group while 15% of the fragments had more than one functional group present.

Supplemental Figure 9: Non-targeted Exposome Data (814 spectral peaks) - Cumulative Variance Explained Before vs. After Principal Component Pursuit



Similar to Supplemental Figure 5, the original untargeted exposure matrix is noisy and is not easily amenable to factor analysis approaches to data dimension reduction. As shown here, principal component analysis of the original exposure matrix has resulted in numerous PCs where 81 PCs are required to explain ≥99% of the variance in the data. However, after decomposition via Principal Component Pursuit (PCP), the resulting low rank L-matrix has considerably fewer components. Specifically, only 6 PCs were needed to explain ≥99% of the variance in the L-matrix, each explaining 79.5%, 8.8%, 5.9%, 2.4%, 1.7%, and 1.1% of the variance. In total, 18 ranks were identified, but the majority of these (i.e. PCs 7-18) explained ≤0.25% of variance each.

Supplemental Figure 10: Non-targeted Exposome Data- BKMR Mixture Association using 15 components



In models with top 15 components, Peak 10 (identified as N-Nitrosodiethylamine [NDEA] in Supplemental Figure 9 and Supplemental Table 3) had the highest posterior inclusion probability (A) and is the only chemical that displayed a plausible and statistically significant (i.e. crosses the credible interval) dose-response relationship with our semen parameter outcomes (B). Similarly, it is the only peak from PC6 where it was associated with semen parameters in individual exposure models (C).

Supplemental Figure 11: Chromatogram of NDEA in Study Samples (Top) and in Commercially Available Standards (Bottom)