Comets-Analytics: a centralized computational framework for consortia level meta-analyses

Ewy Mathé [1,*], Steven C. Moore [2,*], Krista Zanetti [3], Kai-Ling Chen [4], Dave Ruggieri [5], Ella Temprosa [6,*] on behalf of the Data Harmonization Working Group of the Consortium of Metabolomics Studies * equal contributors

[1] Department of Biomedical Informatics, The Ohio State University, Columbus, OH, USA; [2] Division of Cancer Epidemiology & Genetics, National Cancer Institute, Rockville, MD, USA; [3] Division of Cancer Control & Population Sciences, National Cancer Institute, Rockville, MD, USA; [4] Essential Software Inc. Contractor at Center for Biomedical Informatics and Information Technology, National Cancer Institute, Bethesda, MD, USA; [5] Information Management Services, Inc. Contractor at Center for Biomedical Informatics and Information Technology, National Cancer Institute, Bethesda, MD, USA; [6] Milken Institute School of Public Health and Health Services, George Washington University, Rockville, MD, USA.

***Aim:* To provide a streamlined process for conducting standardized and reproducible consortia research**

Abstract

Metabolomics is increasingly applied in large-scale epidemiological studies to uncover metabolites associated with physiological states (e.g. age, disease). The National Cancer Institute-led "Consortium of Metabolomics Studies" (COMETS) includes > 45 international prospective cohorts with serum metabolomics profiles and detailed phenotypic data. To support meta-analysis of these studies at a consortia level, we created a centralized computational infrastructure, Comets-Analytics.

With Comets-Analytics, cohorts analyze their own data using a common data format and standardized results are sent centrally for meta-analysis. This streamlined approach greatly facilitates large consortia studies and helps ensure integrity, and reproducibility of results.

Availability

Server:
http://comets-analytics.org

R package:
https://github.com/CBIIT/R-cometsAnalytics

Tutorial:
https://ellatemprosa.github.io/cometstutorial/index.html

Methods

Guiding Principles:

1) Minimal burden on analyst time
2) Reproducibility
3) Data privacy
4) Adherence with FAIR guidelines
5) Usability

Analyses Supported:

- Harmonization of metabolite names across different platforms used in COMETS
- Partial correlations

Software Testing:

1) Devtools Rcheck
2) Web interface results using test data
3) Statistical output against SAS and Stata output on same test data

Detailed Documentation:

- R package has vignette and functions with working examples
- Extensive tutorial to walk through analyses pipeline
- Sample input file available for testing

Server run modes

Interactive:

Users can build models using the user-friendly interface

Batch:

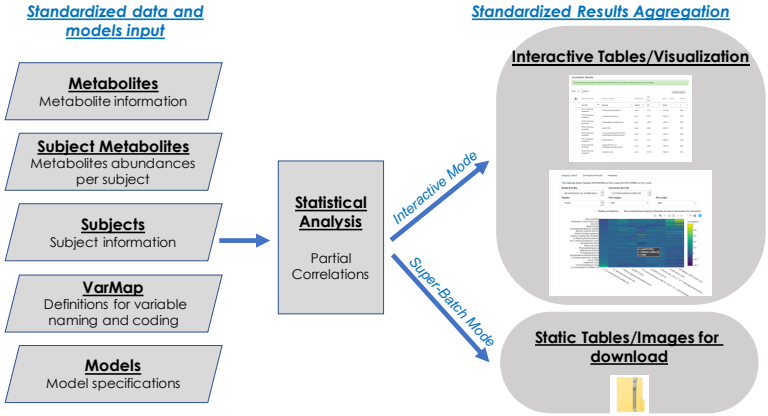
Users define models in input data file (Models sheet) and call each model by name one by one using the interface

Super-Batch:

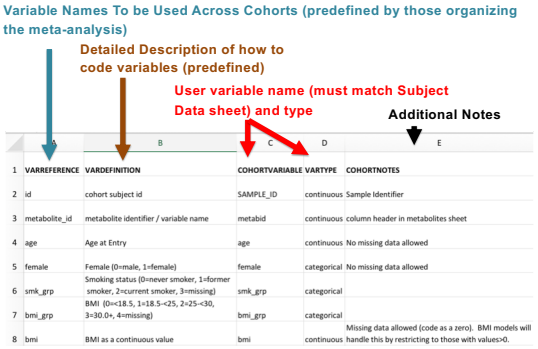
Users can run all models specified in input data file by a single click

Funding: Division of Cancer Epidemiology & Genetics and Division of Cancer Control & Population Sciences, National Cancer Institute, National Institutes of Health

Workflow of individual cohort-level analysis



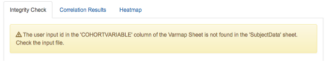
Robust Analytics on Data Input



	A	B	C	D	E
1	VARREFERENCE	VARDEFINITION	COHORTVARIABLE	VARTYPE	COHORTNOTES
2	id	cohort subject id	SAMPLE_ID	continuous	Sample Identifier
3	metabolite_id	metabolite identifier / variable name	metabid	continuous	column header in metabolites sheet
4	age	Age at Entry	age	continuous	No missing data allowed
5	female	Female (0=male, 1=female)	female	categorical	No missing data allowed
6	smk_grp	Smoking status (0=never smoker, 1=former smoker, 2=current smoker, 3=missing)	smk_grp	categorical	
7	bmi_grp	BMI (0=<18.5, 1=18.5-<25, 2=25-<30, 3=30.0+, 4=missing)	bmi_grp	categorical	
8	bmi	BMI as a continuous value	bmi	continuous	Missing data allowed (code as a zero). BMI models will handle this by restricting to those with values>0.

✧ Extensive checks of input data, returning a **meaningful error** when:

- Software cannot find appropriate variables (incorrect VarMap sheet)
- Metabolites or samples have missing or non-matching meta data
- Adjustment and stratification or exposure and stratification variables are the same



✧ ***If you get a non-meaningful error, please email us!***

Robust Analytics on Data Models

	A	B	C	D	E	F
1	MODEL	OUTCOMES	EXPOSURE	ADJUSTMENT	STRATIFICATION	WHERE
2	1 Gender adjusted	All metabolites	age	female		
3	2 Multivariable adjusted	All metabolites	age	female smk_grp bmi_grp race_grp educ_grp alc_grp multivitamin horm_curr fasted		
4	2.1 Gender stratified	All metabolites	age	smk_grp bmi_grp race_grp educ_grp alc_grp multivitamin horm_curr fasted	female	
5	2.2 BMI stratified	All metabolites	age	female smk_grp race_grp educ_grp alc_grp multivitamin horm_curr fasted	bmi_grp	
6	2.3 Smoking stratified	All metabolites	age	female bmi_grp race_grp educ_grp alc_grp multivitamin horm_curr fasted	smk_grp	
7	2.4 Cancer stratified	All metabolites	age	female smk_grp bmi_grp race_grp educ_grp alc_grp multivitamin horm_curr fasted	prev_cancer	
8	2.5 Heart-disease stratified	All metabolites	age	female smk_grp bmi_grp race_grp educ_grp alc_grp multivitamin horm_curr fasted	prev_heart_dx	
9	2.6 Race stratified	All metabolites	age	female smk_grp bmi_grp educ_grp alc_grp multivitamin horm_curr fasted	race_grp	
10	2.7 HT stratified	All metabolites	age	female smk_grp bmi_grp race_grp educ_grp alc_grp multivitamin fasted	horm_curr	female==1
11	2.8 Age stratified	All metabolites	age	female smk_grp bmi_grp race_grp educ_grp alc_grp multivitamin horm_curr fasted	age_grp	
12	2.9 Case selection stratified	All metabolites	age	female smk_grp bmi_grp race_grp educ_grp alc_grp multivitamin horm_curr fasted	nested_case	
13	3.0 All pairwise metabolites	All metabolites	All metabolites			

✧ ***If you get a non-meaningful warning, please email us!***

✧ Will give a **meaningful warning** when:

- Check that adjustment variables have at least 2 unique values (if not, model runs without adjustment)
- Check that all covariate have non-zero variance (those with zero-variance are dropped)
- Check for correlated covariates (this will remove the first "factor" that is highly correlated with another)
- Check for linear dependencies

Data and Privacy

Interactive mode:

- Uploaded input data file is deleted immediately after integrity check (< 1 min)
- Output files are deleted from server after 24 hrs or end of user.

Super-batch mode:

- Input data file is deleted once the job is picked up by the Comets queue processor immediately after integrity check
- Result files are stored in a secure, private S3 bucket in AWS, and can only be accessed via a unique, encrypted URL sent to user after job completion
- Result files are deleted after 7 days

Future directions

❖ Generalized linear modeling
❖ Pathway analysis

COMETS Analytics
Reproducibility, usability, input/output consistency