



Spectra Processing, Compound Annotation, Functional Insight and Causal Analysis using **MetaboAnalyst 6.0**

Jianguo (Jeff) Xia, Associate Professor

Canada Research Chair in Bioinformatics & Big Data Analytics

jeff.xia@mcgill.ca | www.xialab.ca

McGill University, Canada



XiaLab.ca

Empowering researchers through trainings, tools, and AI



McGill
UNIVERSITY

Schedule

Part I: 2:15 p.m. – 4:15 p.m

- **2:15 – 3:00:** Background
 - ✓ General introduction
 - ✓ LC-MS & MS/MS spectral processing
 - ✓ From peaks to functions
- **3:00 – 3:20:** Live demo
- **3:20 – 4:15:** Hands on practice

Part II: 4:30 p.m. – 6:30 p.m.

- **4:30 – 5:10:** Background
 - ✓ Data processing
 - ✓ Statistical analysis
 - ✓ Causal analysis
- **5:10 – 5:40:** Live demo
- **5:40 – 6:15:** Hands on practice
- **6:15 – 6:30:** Summary & discussion

Github Repository

- https://github.com/xia-lab/Metabolomics_2024
- Slides (in PDF format);
- Example data;
- Reference literatures;
- Contact information.

Causal Analysis

- **mGWAS**
- **Two-sample Mendelian Randomization (2SMR)**

MetaboAnalyst 6.0 Modules

Input Data Type	Available Modules (click on a module to proceed, or scroll down to explore a total of 18 modules including utilities)				
LC-MS Spectra (mzML, mzXML or mzData)			Spectra Processing [LC-MS w/wo MS2]		
MS Peaks (peak list or intensity table)		Peak Annotation [MS2-DDA/DIA]	Functional Analysis [LC-MS]	Functional Meta-analysis [LC-MS]	
Generic Format (.csv or .txt table files)	Statistical Analysis [one factor]	Statistical Analysis [metadata table]	Biomarker Analysis	Statistical Meta-analysis	Dose Response Analysis
Annotated Features (metabolite list or table)		Enrichment Analysis	Pathway Analysis	Network Analysis	
Link to Genomics & Phenotypes (metabolite list)			Causal Analysis [Mendelian randomization]		

How can we detect causal effects?

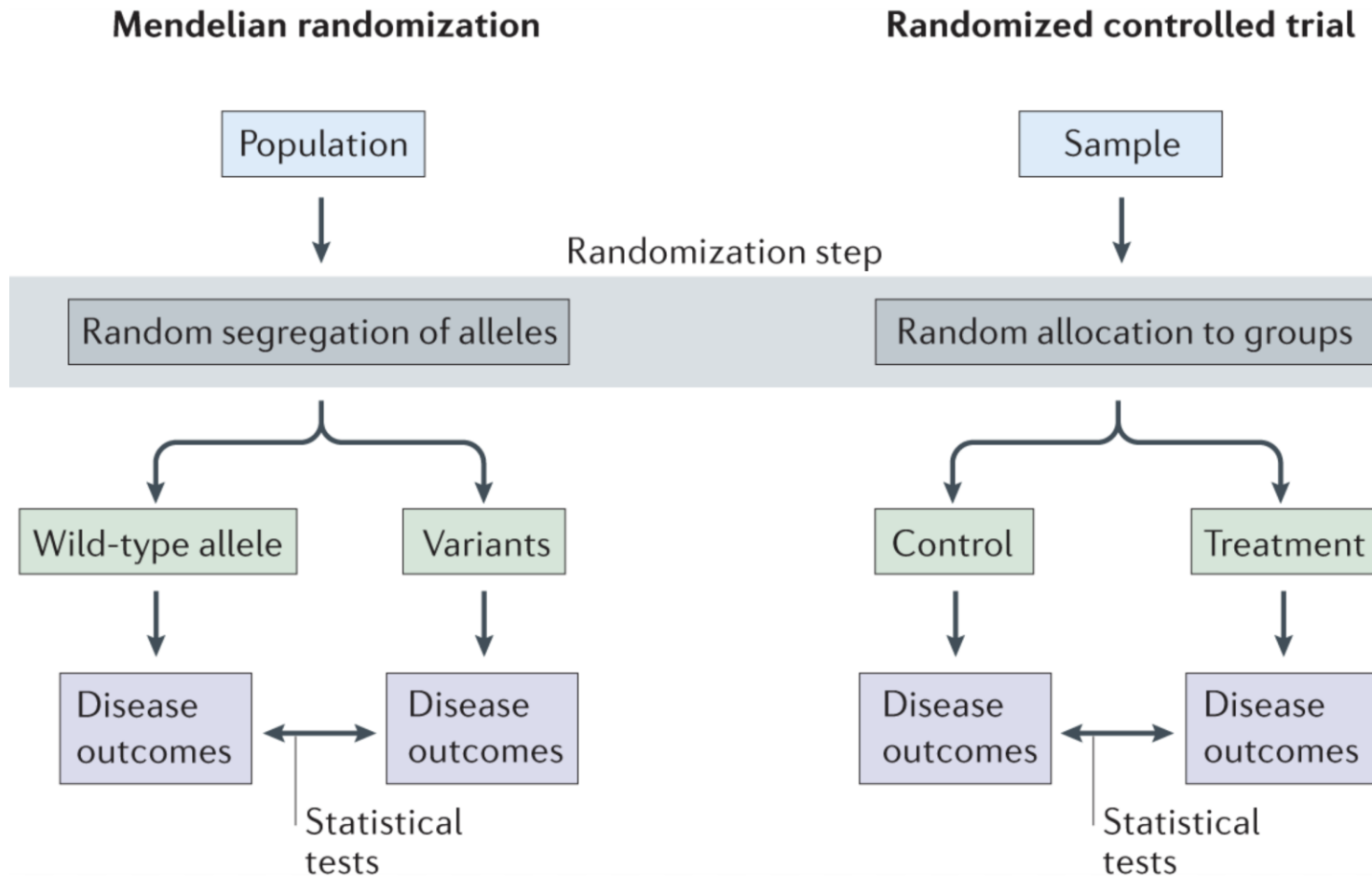
- **Key concept:**

- If a biomarker / metabolite is causal for a disease / phenotype, the genetic variants which influence the levels of the biomarker should result in a higher risk of the disease
- Leverage known genetic variants (i.e. SNPs) to eliminate confounders to help detect causal links

- Mendelian Randomization (MR) analysis

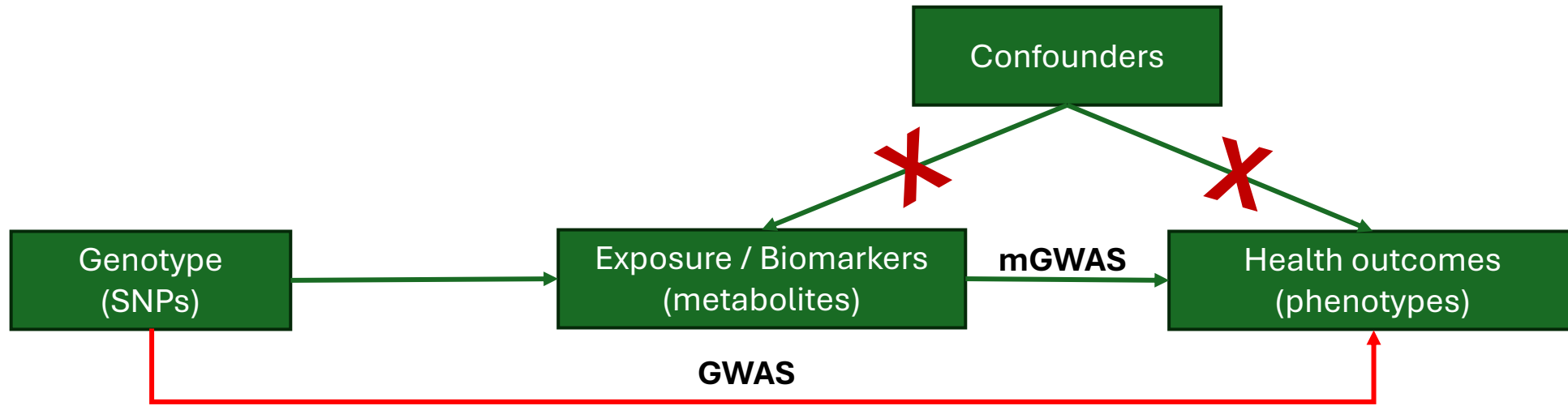
- These genetic variants are called **instrument variables**

MR analysis: nature's randomized controlled trials

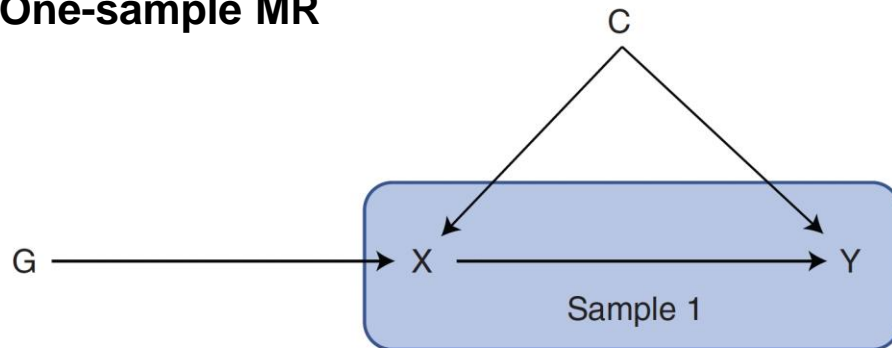


Nat Rev Methods Primers **2**, 6 (2022)

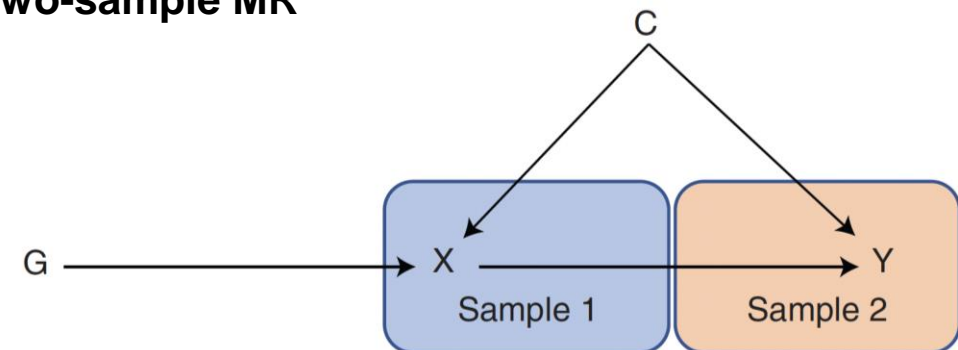
Mendelian randomization concept



One-sample MR



Two-sample MR



Metabolite Genome-Wide Association Study (mGWAS)

Linking the genomics with metabolomics to identify genetic variants affecting metabolite levels

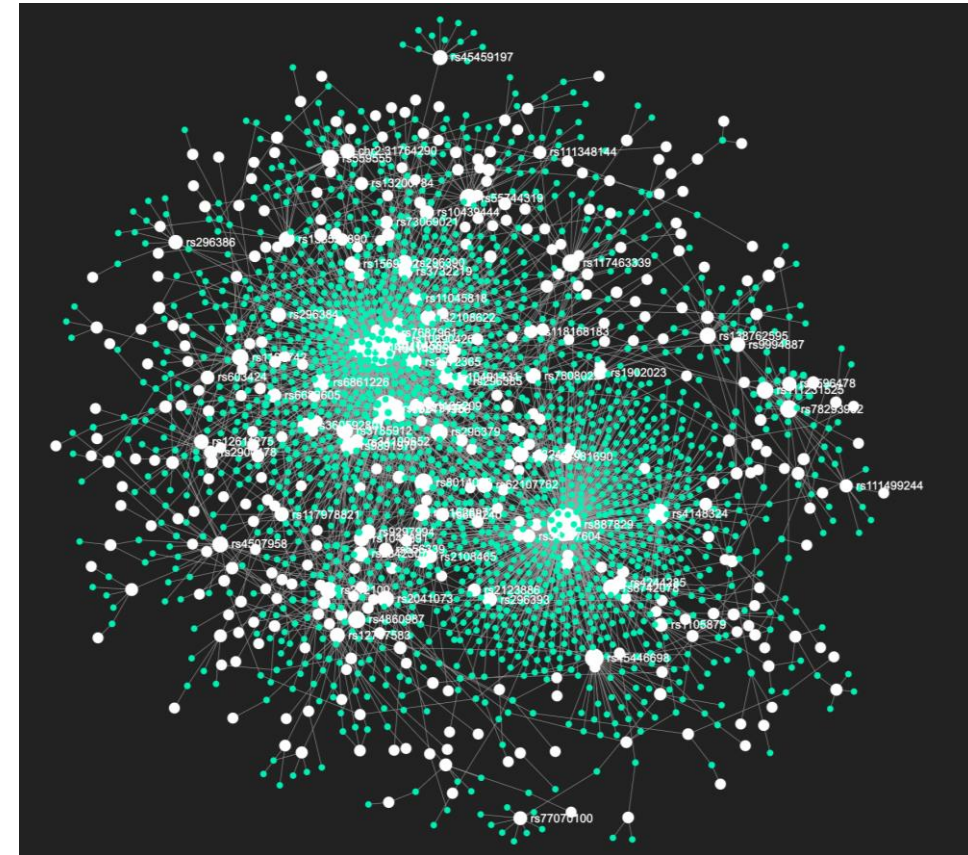
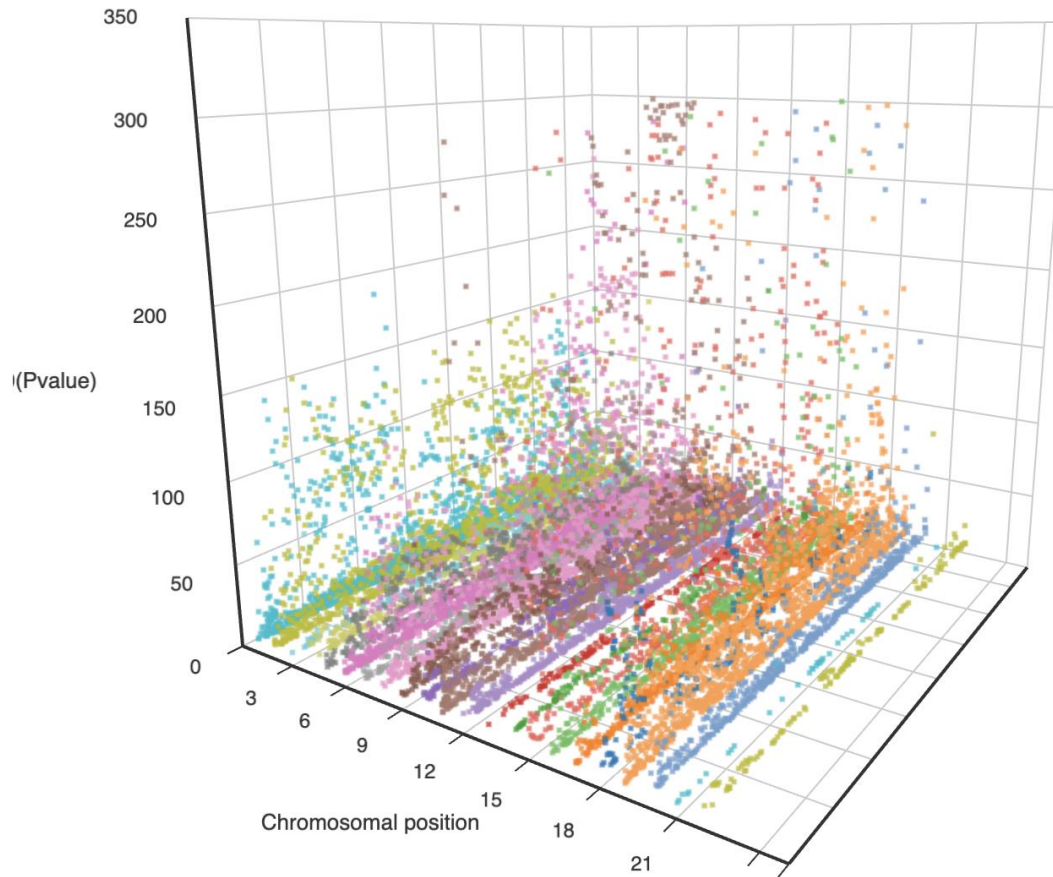
Sample Type	Study #	* Metabolite #	** Metabolite Ratio #	SNP #	SNP–Metabolite Associations #
Blood	57	3992	1265	67,570	30,3090
Urine	5	271	1123	6877	9647
Saliva	1	14	0	1364	1454
Cerebrospinal fluid (CSF)	1	15	0	1178	1182
Mitochondria	1	0	390	194	404
Sum (unique)	65	4147	2388	73,737	313,720

Chang et al. *Metabolites* **2022**, 12(6), 526

Browse mGWAS studies

ID	Biofluid	Study	Publication	Sample Size	Population	Genotyping Platform	Metabolomics Platform	Cutoff Threshold	Browse
65b	Blood	Viñuela_medRxiv_2021_targeted	Genetic analysis of blood molecular phenotypes reveals regulatory networks affecting complex traits: a DIRECT study	3029	European	Illumina HumanCore array (HCE24 v1.0)	BIOCRATES (AbsoluteIDQ™ p150 kit)	5e-08	 View
65a	Blood	Viñuela_medRxiv_2021_untargeted	Genetic analysis of blood molecular phenotypes reveals regulatory networks affecting complex traits: a DIRECT study	3029	European	Illumina HumanCore array (HCE24 v1.0)	Metabolon (LC-MS/MS)	5e-08	 View
64	Blood	Qin_medRxiv_2020	Genome-wide association and Mendelian randomization analysis prioritizes bioactive metabolites with putative causal effects on common diseases	8738	European	Illumina genome-wide SNP arrays (HumanCoreExome BeadChip, Human610-Quad BeadChip and HumanOmniExpress)	Thermo Q Exactive Orbitrap	4.5e-12	 View
63	Blood	Borges_UKBB_2020	Metabolic biomarkers in the UK Biobank measured by Nightingale Health 2020	500000	European	Affymetrix genome-wide genotyping array	Nightingale NMR	5e-08	 View
62	Blood	Montasser_bioRxiv_2021	Leveraging a founder population to identify novel rare-population genetic determinants of lipidome	650	Old Order Amish founder population	Affymetrix 500K array	Agilent (6550 Q-TOF LC/MS)	5e-08	 View
61	Mitochondria	Aboulmaouahib_HMG_2021	First mitochondrial genome wide association study with metabolomics	2718	European	Illumina MiSeq	BIOCRATES (AbsoluteIDQ™ p150 kit)	1e-05	 View
60	Blood	Harshfield_BM_2021	Genome-wide analysis of blood lipid metabolites in over 5000 South Asians reveals biological insights at cardiometabolic disease loci	13814+5662	European+South Asian	Illumina 660-Quad, Illumina HumanOmniExpress, Affymetrix	Thermo Q Exactive Orbitrap	8.9e-10	 View

Associations between SNPs and compounds/peaks

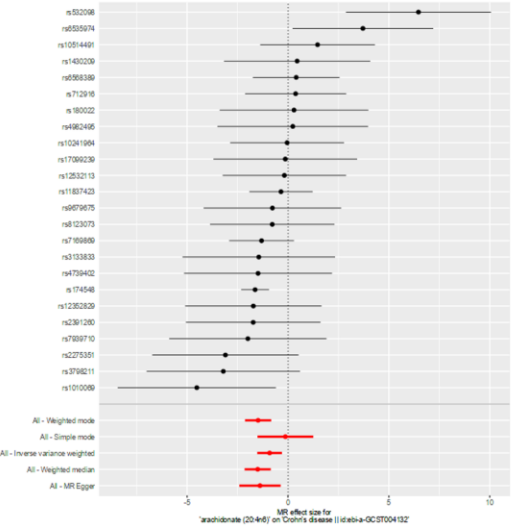


Explore causally links and evidences

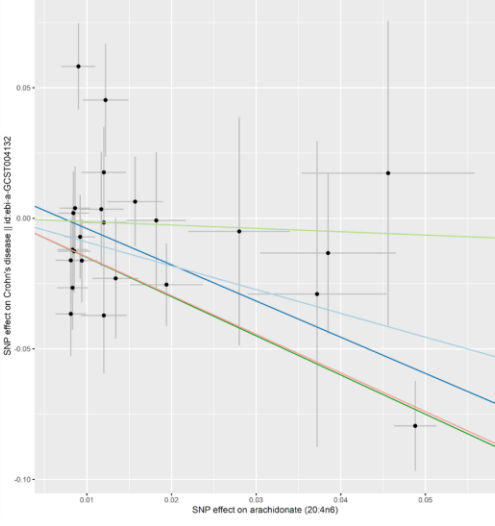
(a)

Methods	MR Results				Heterogeneity Tests			Horizontal Pleiotropy		
	Number of SNPs	Beta	SE	P value	Q	Q_df	Q_pval	Egger Intercept	SE	P value
Inverse variance weighted	24	-0.91	0.313	0.0036	42.1	23	0.00893	-	-	-
MR Egger	24	-1.39	0.523	0.0146	39.7	22	0.0116	0.00993	0.00877	0.27
Simple mode	24	-0.128	0.711	0.858	-	-	-	-	-	-
Weighted median	24	-1.5	0.323	3.46e-06	-	-	-	-	-	-
Weighted mode	24	-1.48	0.315	9.73e-05	-	-	-	-	-	-

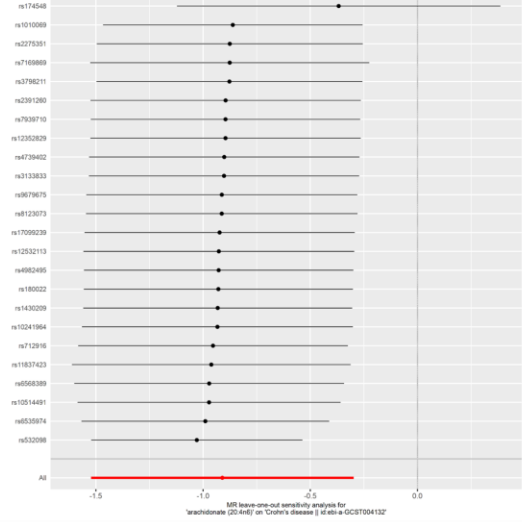
(b)



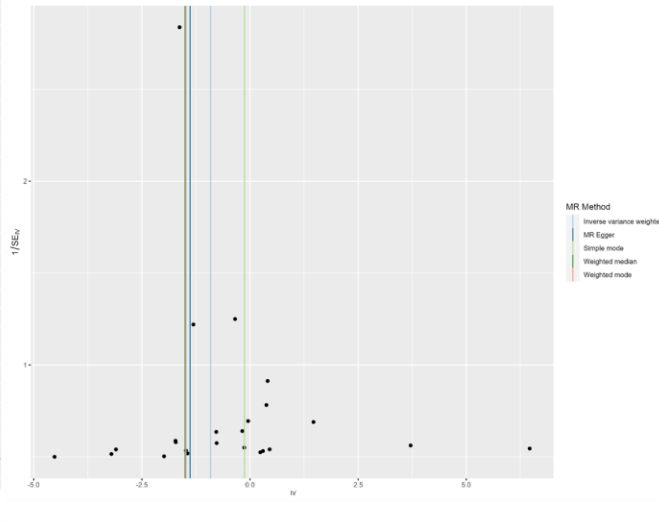
(c)



(d)



(e)



Potential causal associations between ~4000 metabolites and >200 diseases

Diseases	Exposure / Metabolites	SNP	Effect size	SE	Pval
Atherosclerotic heart disease	serine	5	-0.0279495	0.00203363	5.5606E-43
	sm c24:0	5	0.02832925	0.00217185	6.8912E-39
	sm (oh) c22:2	5	0.02907336	0.0022289	6.8912E-39
Inflammatory bowel disease	x-11381	3	-0.8251061	0.04109667	1.1679E-89
	nervonoylcarnitine (c24:1)*	3	-0.9093196	0.04529115	1.1679E-89
	margaroylcarnitine (c17)*	2	-1.0991236	0.06922023	8.9079E-57
	linoleoylcarnitine (c18:2)*	2	-0.8909735	0.05611142	8.9079E-57
	x-24241	1	-1.2770696	0.09285142	4.8253E-43
	histidine betaine (hercynine)*	1	1.24068493	0.09046718	8.3483E-43
	pc aa c32:2	5	0.46386518	0.03590836	3.5615E-38
	pc ae c34:2	5	0.4087556	0.03164226	3.5615E-38
	1-oleoyl-2-eicosapentaenoyl-gpc (18:1/20:5)	4	-0.2323055	0.0195418	1.3726E-32
	arachidoylecarnitine (c20)*	1	-0.9189255	0.07831706	8.5934E-32
	nisinate (24:6n3)	4	-0.3856061	0.03335131	6.4236E-31
Asthma	x-11381	3	0.04781557	0.0027428	4.6252E-68
	pc ae c36:3	8	-0.021023	0.00120984	1.2411E-67
	1-oleoyl-2-eicosapentaenoyl-gpc (18:1/20:5)	4	0.02003149	0.00134398	3.0763E-50
	linoleoylcarnitine (c18:2)*	2	0.05415016	0.00374454	2.1338E-47
	docosatrienoate (22:3n6)*	3	0.0276776	0.00206776	7.3688E-41
	1-eicosapentaenoyl-gpc (20:5)	3	0.02112966	0.00162433	1.0981E-38
	1-palmitoyl-2-eicosapentaenoyl-gpc (16:0/20:5)	3	0.02159433	0.00166005	1.0981E-38
	1-dihomo-linolenoyl-gpe (20:3n3 or 6)	3	-0.016672	0.00128165	1.0981E-38

Live Demo

Specify metabolite and phenotype of interest

Please specify metabolites (exposure) and outcome of interest

Causal analysis is based on Mendelian randomization (MR) which leverages genetic variants such as single-nucleotide polymorphisms (SNPs) as instrumental variables (IV) to estimate exposure-outcome associations. The growing number of mGWAS studies and two-sample MR method permit causal analysis between metabolite and outcome of interest as described below:

1. Identify SNPs that are significantly associate with a metabolite of interest from our large collections of the recent [mGWAS studies](#) (covering > 4000 metabolites including their ratios);
2. Obtain the estimates of associations between these same SNPs with an outcome of interest from public repository. We use [Open GWAS Project](#).
3. Perform SNP filtering and harmonize the effect sizes for SNPs on the exposures and the outcomes to be for the same reference allele.
4. Conduct MR analysis, sensitivity analyses, and explore the graphical outputs

Please note you may not be able to perform causal analysis in some cases when no suitable SNPs are found in the two repositories.

1. Select a metabolite of interest (exposure):

Due to its complex and computing intensive nature, MR analysis is typically performed with one metabolite/exposure at a time, to make sure that each step is performed properly as well as to avoid performance issue (max 5).

Search bar: cys

Filter selected metabolites: L-Cystathionine

Metabolites list:

- ☐ Cysteineglutathione disulfide
- ☐ Methylycysteine
- ☐ 3-(Cystein-S-yl)acetaminophen
- ☐ S-N-Methylcysteine
- ☐ 3-Indolepropionic acid/S-(5-Adenosyl)-L-homocysteine
- ☐ 3-Indolepropionic acid/S-Sulfo-L-cysteine
- ☐ L-Cystine/3-Indolepropionic acid

Users should first select an exposure (i.e., metabolites) and an outcome (i.e., diseases) of interest.

Search to choose the metabolite of interest (e.g. Cystathionine) from the left box. Once you select it, it will be automatically added into the right box.

2. Specify an outcome of interest:

Enter a key word to see available options from the public repository.

Type 2 diabetes | finn-b-E4_DM2

Proceed

For instance, we are interested in Type 2 diabetes. Type the name to see a list of matched studies. Here we choose finn-b-E4_DM2

SNP Filtering & Harmonization

- Multiple SNPs could be identified as potential instrumental variables (IV) from the mGWAS and GWAS studies.
- To perform proper 2SMR, the IVs should be
 - Independent (i.e. not correlated with each other)
 - Showing strong effect (i.e. significant p-values)
 - No horizontal pleiotropy (i.e. affect the outcome only through the metabolite).
- Users need to carefully examine SNPs and apply different filtering and harmonization methods for each criterion

SNP Filtering and harmonization

To properly conduct two-sample MR analysis, the instrumental variables (IV) should be **independent** (i.e. not correlated with each other), showing **strong effect** (i.e. significant p-values), and **no horizontal pleiotropy** (i.e. affect the outcome only through the metabolite). The step provides following procedures to facilitate proper MR analysis:

- Acquisition of independent IVs by performing linkage disequilibrium (LD) clumping.
- In cases where the SNP query is absent in the outcome GWAS, a proxy SNP in LD with the input SNP, utilizing the 1000 Genomes Project (phase 3).
- Harmonizing exposure and outcome data to make sure that the effects of the SNPs on exposure and outcome are associated with the same allele. You should also review the table below to perform further harmonization based on other metadata (such as population, study info, etc)
- To control horizontal pleiotropy, you should manually exclude SNPs that are associated with multiple metabolites.

1. LD Clumping

- ☒ Do not check for LD between SNPs
☐ Use clumping to prune SNPs for LD

2. LD Proxies

- ☒ Do not use proxies
☐ Use proxies and allow palindrome SNPs (advanced settings)

3. Allele Harmonization

- ☐ Assume all alleles are presented on the forward strand
☒ Try to infer the forward strand alleles using allele frequency information
☐ Correct the strand for non-palindromic SNPs, but drop all palindromic SNPs

Submit

▶ Proceed

Harmonization steps require intensive computing and also access via remote server.

It could take a long time or time out.
Please be patient

Details from the
mGWAS studies

SNP ID ↑↓	Associated Metabolites ↑↓	Nearest Gene ↑↓	P-value ↑↓	Biofluid ↑↓	Population ↑↓	Study ↑↓	
▼ L-Cystathionine (7)							
rs117782586	L-Cystathionine	JRK1	4.637e-08	Blood	European	28263315	✓
rs146276253	L-Cystathionine	ANKRD13C	3.436e-08	Blood	European	28263315	✓
rs150320192	L-Cystathionine	SRSF11	3.566e-08	Blood	European	28263315	✓

Different methods for MR analysis

MR analysis methods are based on the *TwoSampleMR* and *MRInstruments* R packages. Among these methods, the *median estimator* and *MR Egger regression* allow for genetic pleiotropy. You can use mouse-over of the corresponding question marks to learn more about each method.

<input checked="" type="checkbox"/> Wald ratio [?]	<input type="checkbox"/> Maximum likelihood [?]	<input checked="" type="checkbox"/> MR Egger [?]
<input type="checkbox"/> Simple median [?]	<input checked="" type="checkbox"/> Weighted median [?]	<input type="checkbox"/> Inverse variance weighted radial [?]
<input type="checkbox"/> Inverse variance weighted (MRE) [?]	<input type="checkbox"/> Inverse variance weighted (FE) [?]	<input checked="" type="checkbox"/> Simple mode [?]
<input checked="" type="checkbox"/> Weighted mode [?]	<input type="checkbox"/> Weighted mode (NOME) [?]	<input type="checkbox"/> Simple mode (NOME) [?]
<input type="checkbox"/> Sign concordance test [?]	<input type="checkbox"/> Unweighted regression [?]	

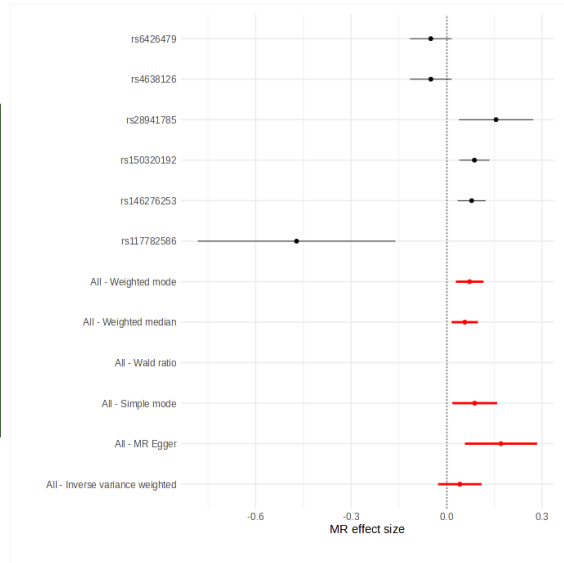
A total of 14 MR methods are offered currently. Some of them are more robust and can better tolerate violations of the assumptions to certain degree

- Mouse over the question marks for each method to see their main features.
- You can also find more detailed introduction on the forum:

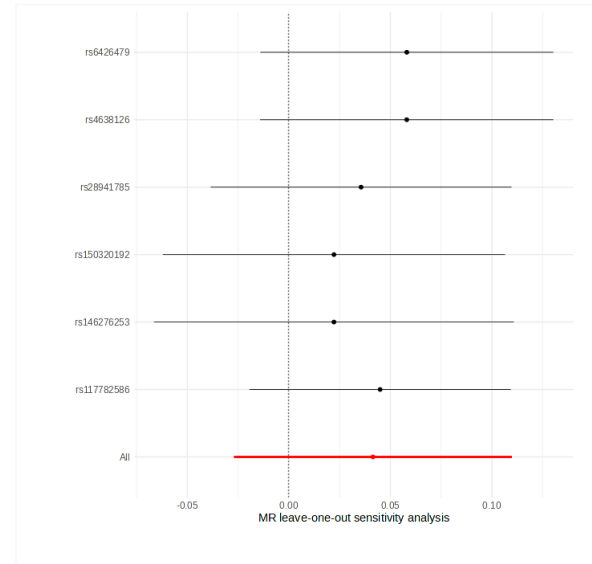
<https://omicsforum.ca/t/what-are-the-differences-between-the-mr-analysis-methods/1045>

Graphical outputs from MR analysis

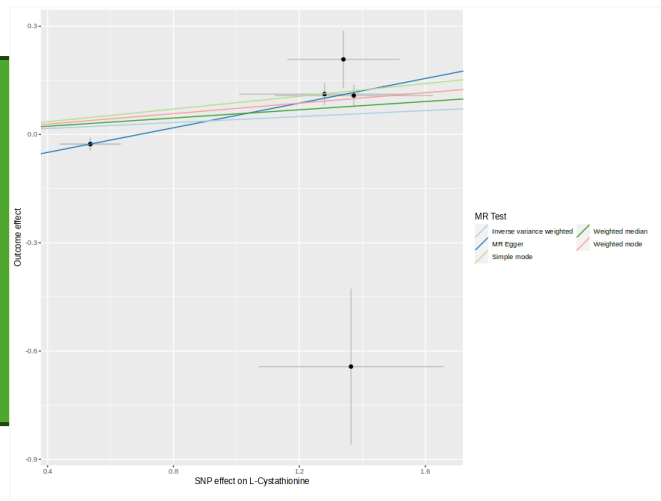
Forest plot compares the causal effect calculated using the methods that include all the SNPs to using each SNP separately.



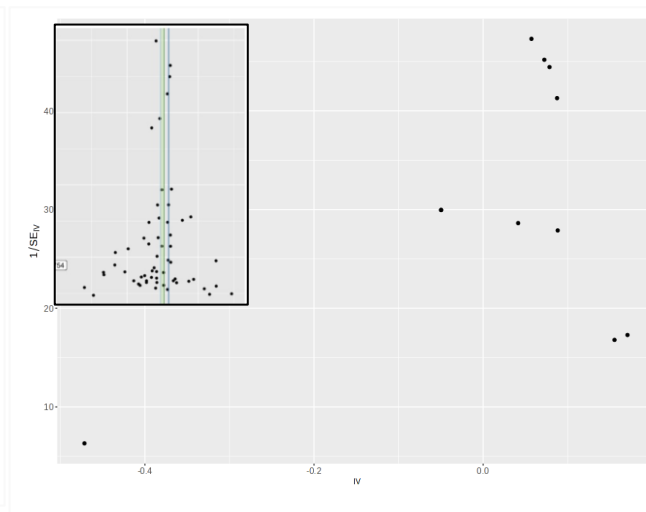
Leave one out sensitivity analysis: assesses whether a single SNP is having a disproportionately larger impact on an association. Each dot represents the MR analysis excluding that specific SNP using IVW method.



Scatter plot shows the relationships between SNP effects on exposure vs on the outcome. The slopes indicating the causal association



Funnel plot: Funnel shape will become more obvious with many SNPs (i.e. green box inset). Its asymmetry and wider spread may suggest horizontal pleiotropy.



Mendelian randomization results

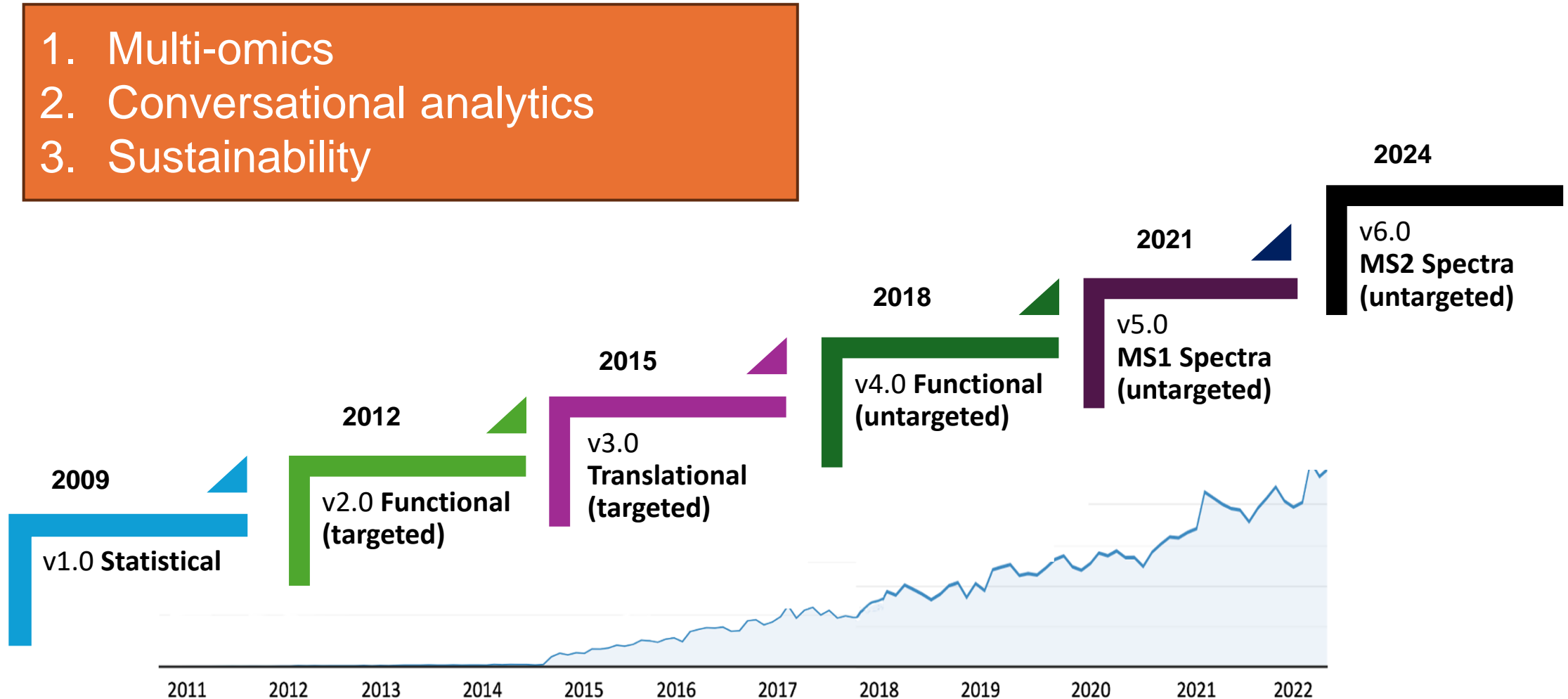
L-Cystathionine										
Methods	SNP Count	Causal Effect Estimates			Heterogeneity Tests			Horizontal Pleiotropy		
		Beta	SE	P value	Q	Q_df	Q_pval	Egger Intercept	SE	P value
Inverse variance weighted	6	0.041396	0.034939	0.23609	35.367	5	1.2709e-06	-	-	-
MR Egger	6	0.17071	0.057839	0.041906	14.006	4	0.0072767	-0.1179	0.047732	0.068948
Simple mode	6	0.088203	0.032068	0.040286	-	-	-	-	-	-
Weighted median	6	0.057145	0.021665	0.0083484	-	-	-	-	-	-
Weighted mode	6	0.072379	0.0219	0.021358	-	-	-	-	-	-

- The MR results are organized per metabolite (exposure).
- For metabolite, it shows the SNPs instrumental variables, along with their corresponding causal effect estimates, standard errors and p-values.
- Key values such as the MR-Egger regression intercept and its corresponding p-value are presented.
- Not all methods selected from the previous page would yield results depending on the data used.

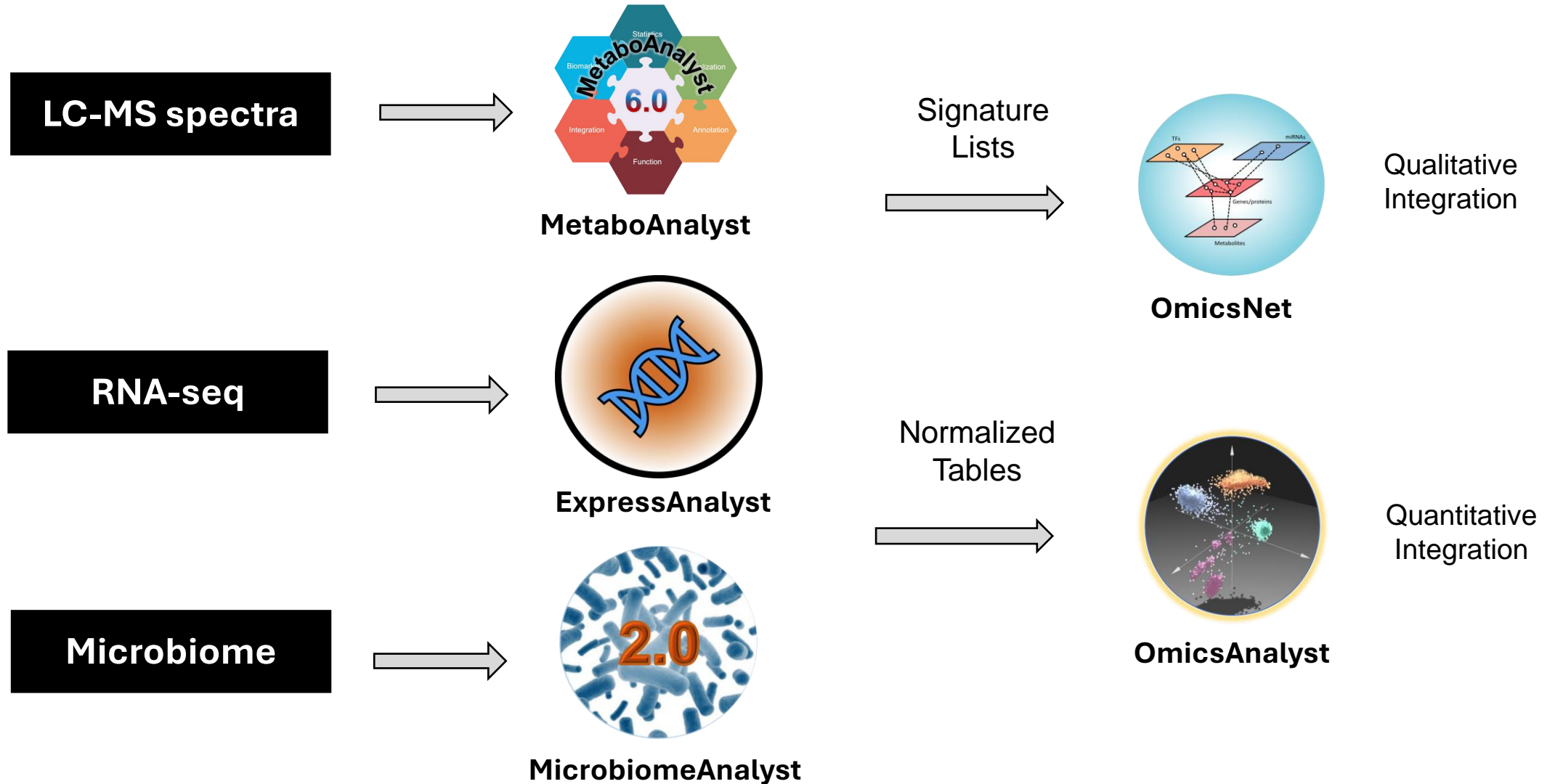
Hands-on time

Summary & Future Perspectives

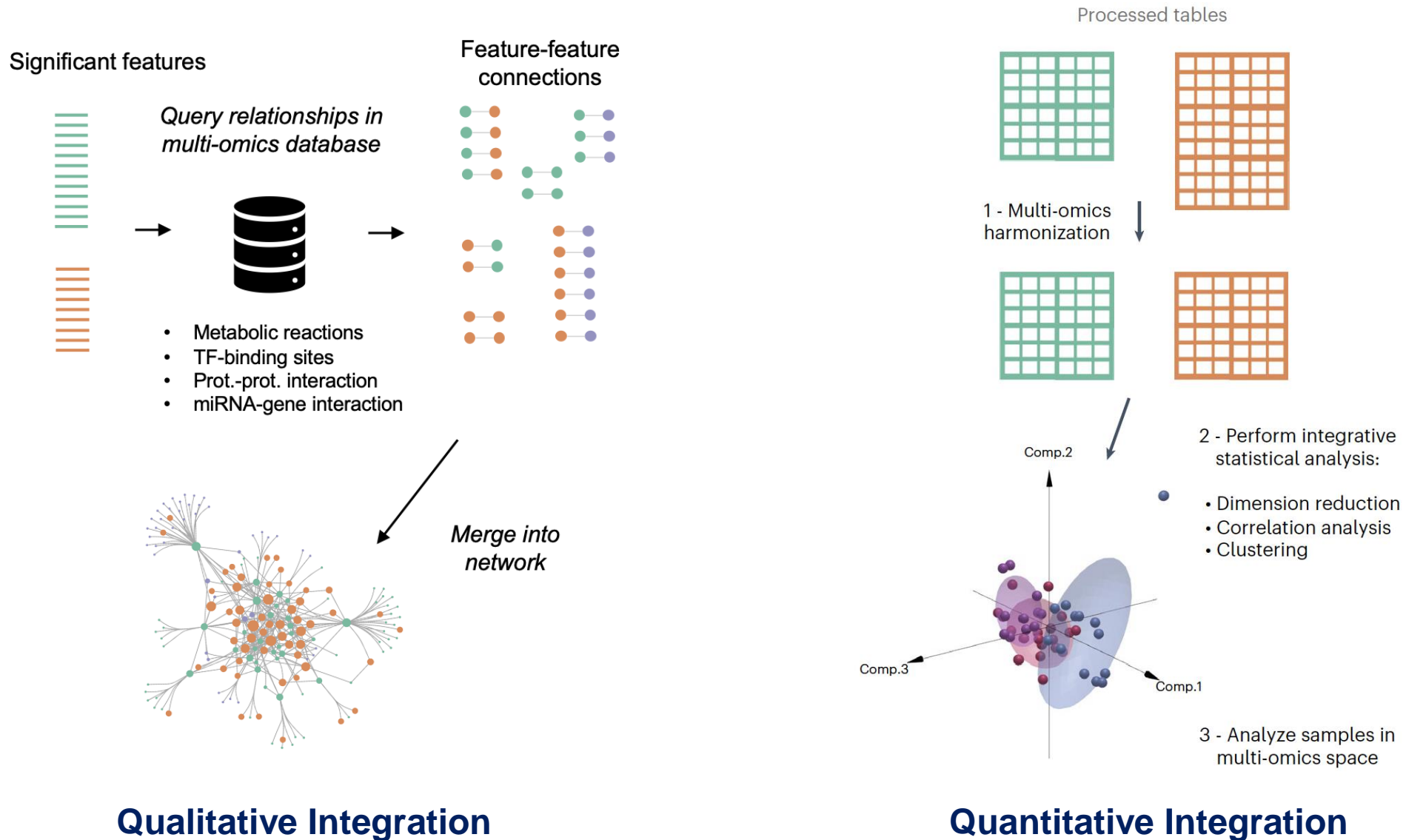
MetaboAnalyst roadmap – what's next?



An ecosystem for omics data analysis



General workflow for multi-omics



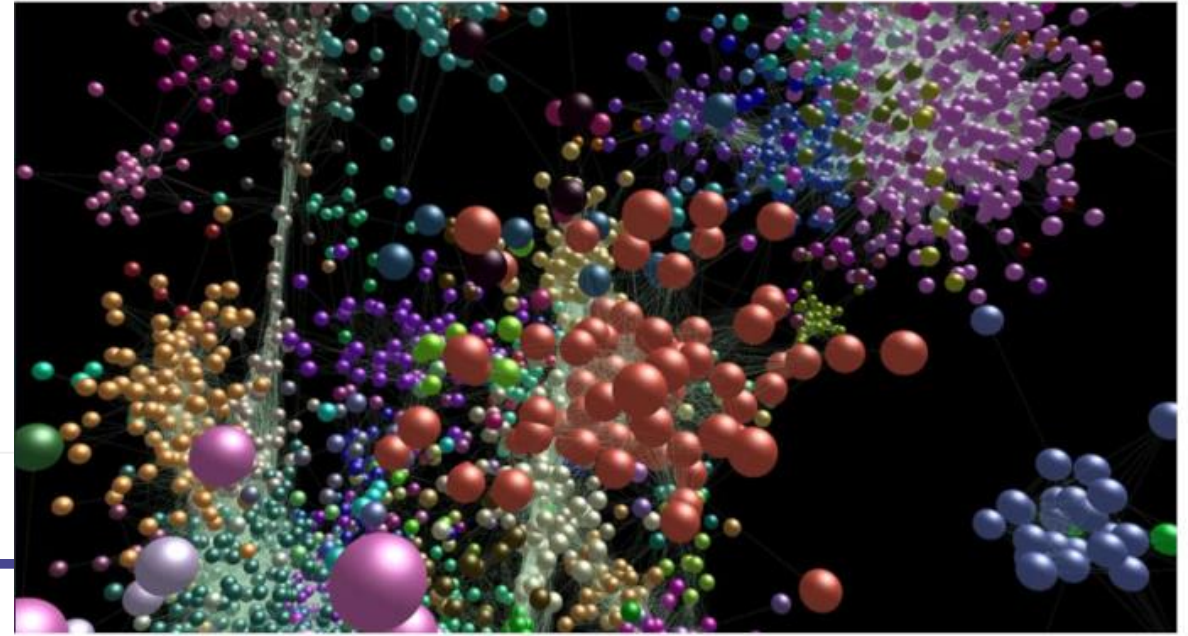
Web-based multi-omics integration using the Analyst software suite

[Jessica D. Ewald](#), [Guangyan Zhou](#), [Yao Lu](#), [Jelena Kolic](#), [Cara Ellis](#), [James D. Johnson](#), [Patrick E.](#)

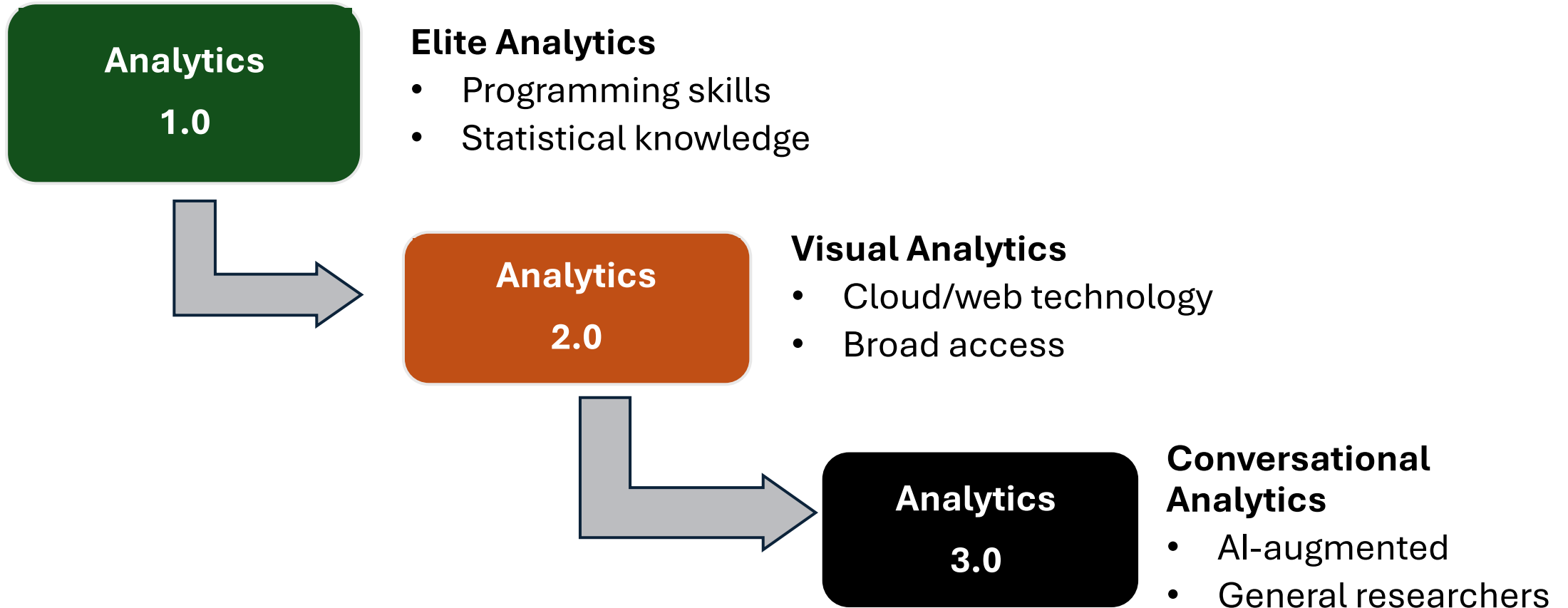
[Macdonald](#) & [Jianguo Xia](#) ✉

[Nature Protocols](#) (2024) | [Cite this article](#)

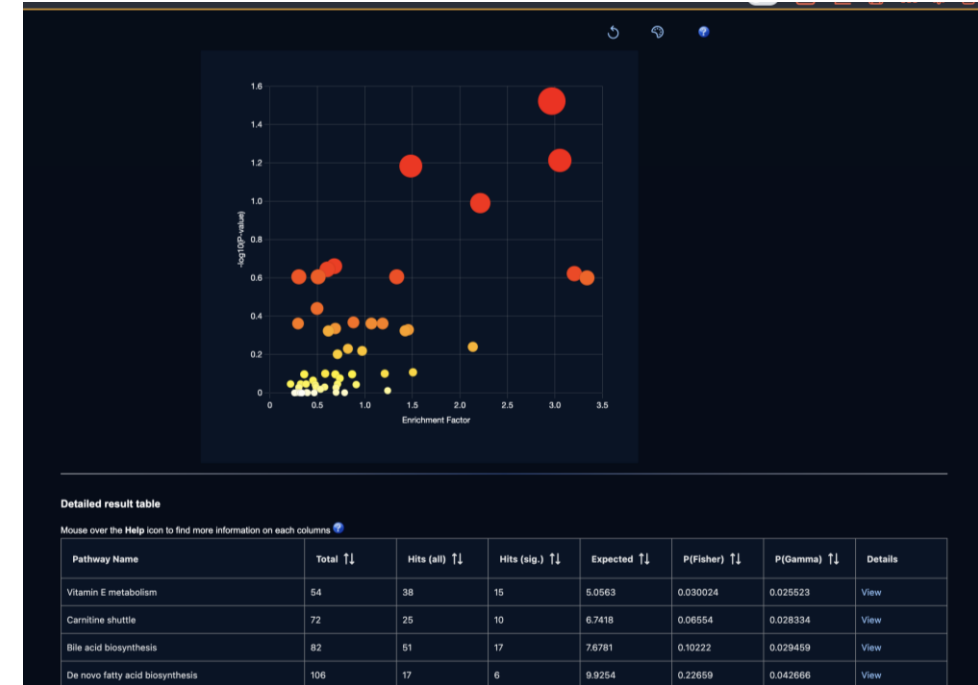
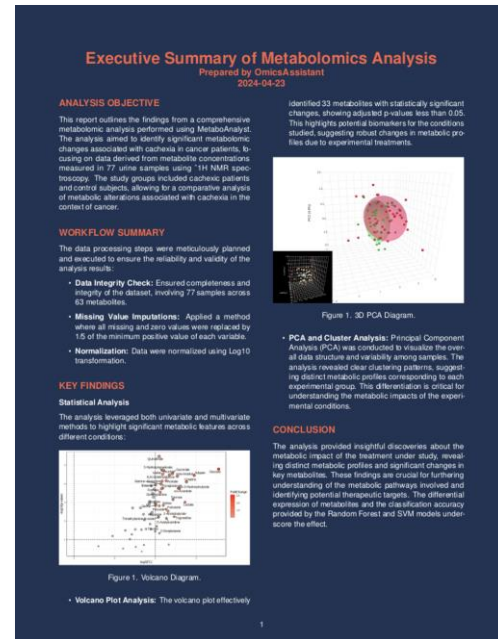
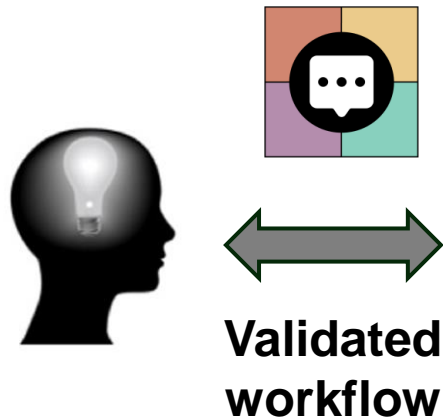
<https://www.nature.com/articles/s41596-023-00950-4>



The evolution of data analytics

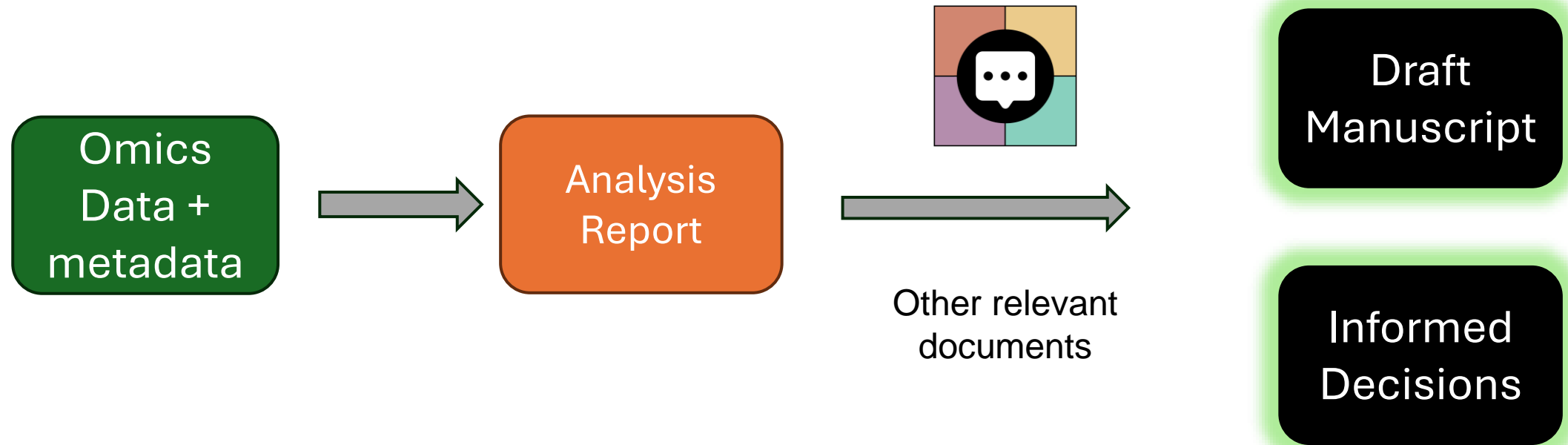


Towards conversational analytics (I)



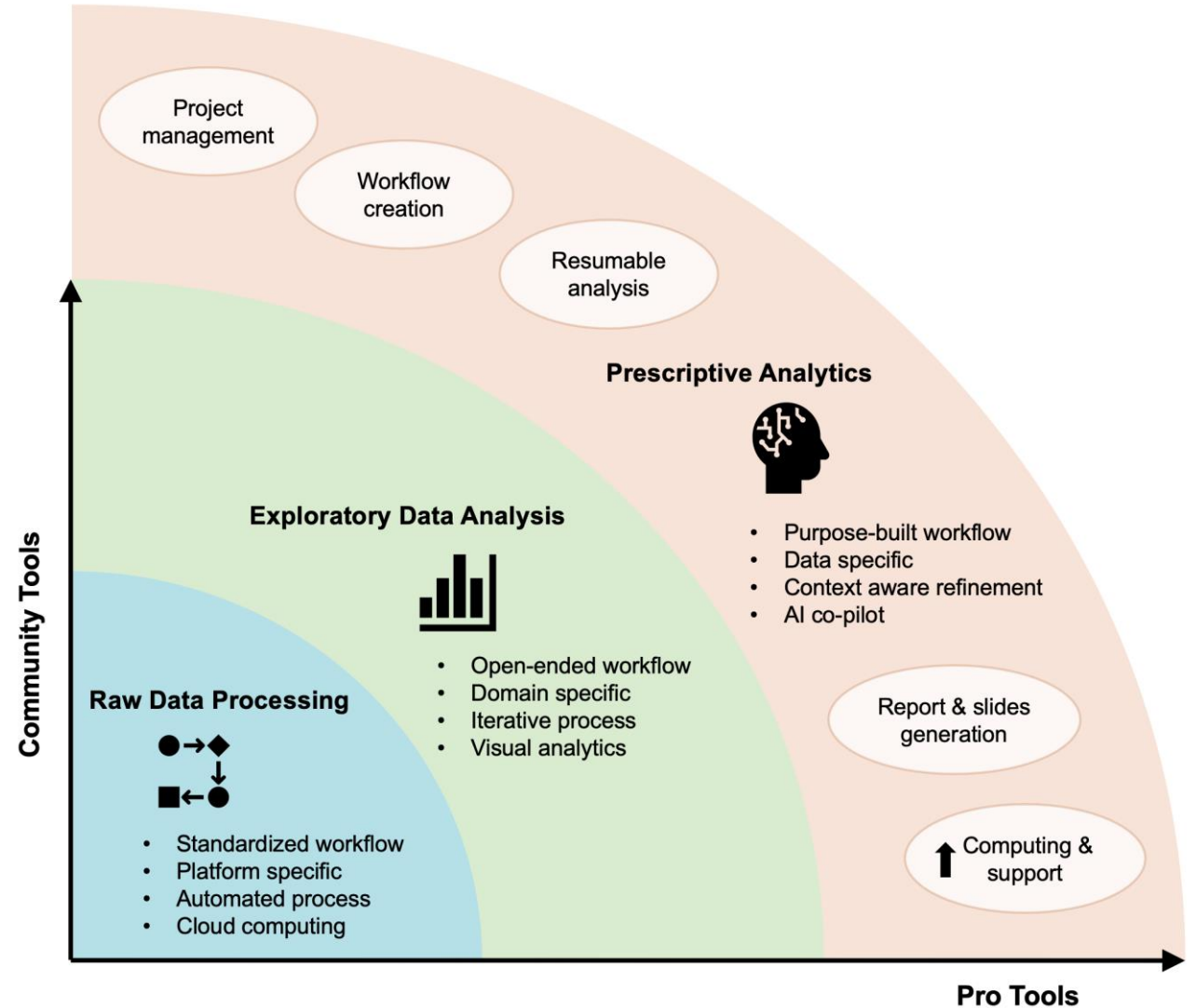
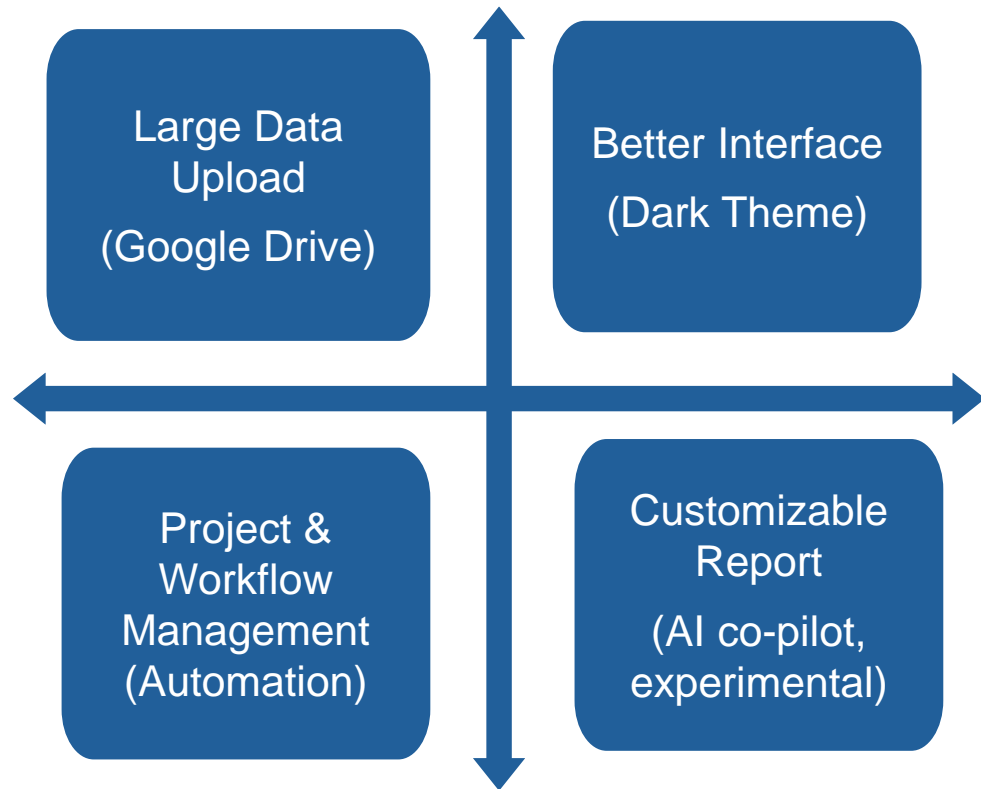
AI for task execution & FAQs

Towards conversational analytics (II)



AI for scientific writing & decision support

Key Features of “Pro” version



Acknowledgements

If you have any questions, please read/post into **OmicsForum** (<https://omicsforum.ca>)

Contact us:

- zhiqiang.pang@xialab.ca
- jeff.xia@xialab.ca

Omics Data Science course

- Summer Bootcamp
 - Aug. 5 - 9, 9:30 - 16:30
- Regular Session
 - Saturday morning 9:30 - 12:00, Sept. - Nov.

