UC DAVIS HEALTH | MIND INSTITUTE

Kyoungmi Kim, Ph.D., Professor

MIND Institute Intellectual and Developmental Disabilities Research Center
Director of Biostatistics, Bioinformatics, and Research Design Core

UC Davis Precision Medicine and Data Sciences
Director of Biostatistics

UC Davis Environmental Health Science Center
Director of Biostatistics Core

UC Davis School of Medicine

An Overview of Metabolomics Approaches and Their Application to Clinical and Translational Research

Applied Statistics for Translational Researchers Seminar Series
November 9, 2022

Sponsored by Biostatistics Programs affiliated with the NIH-funded Centers

- Clinical and Translational Science Center
- MIND Institute Intellectual and Developmental Disability Center (IDDRC)
- Comprehensive Cancer Center
- Environmental Health Sciences Center

UCDAVIS HEALTH | SCHOOL OF MEDICINE
Clinical and Translational Science Center

UCDAVIS HEALTH | MIND INSTITUTE

UCDAVIS HEALTH | COMPREHENSIVE CANCER CENTER

UCDAVIS
ENVIRONMENTAL HEALTH SCIENCES CENTER

Overview

- What is the metabolomics and what are its applications?
- Data preprocessing and quality control
- Statistical analysis

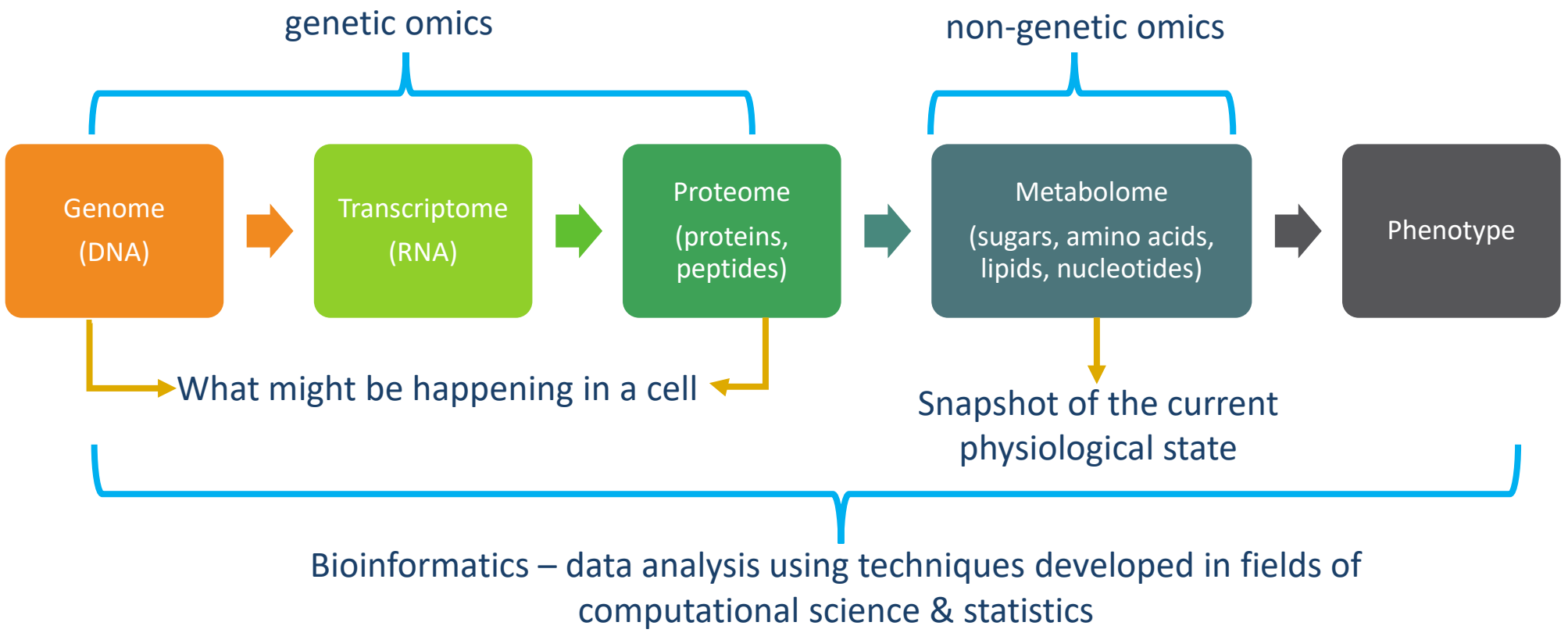
What is the metabolomics?

- Identification and quantification of the complete set of small-molecule metabolites in a biological sample (blood, tissue, cell, organ, or organism)
- Quantitative measurement of the dynamic metabolite response in cells or organs and ways they are altered in disease states and their changes over time as consequence of stimuli (including biological perturbation such as diet, disease, or intervention) or genetic modification
- Provides important insights into physiological and disease states and facilitate in depth understanding of underlying biochemical pathways

Applications

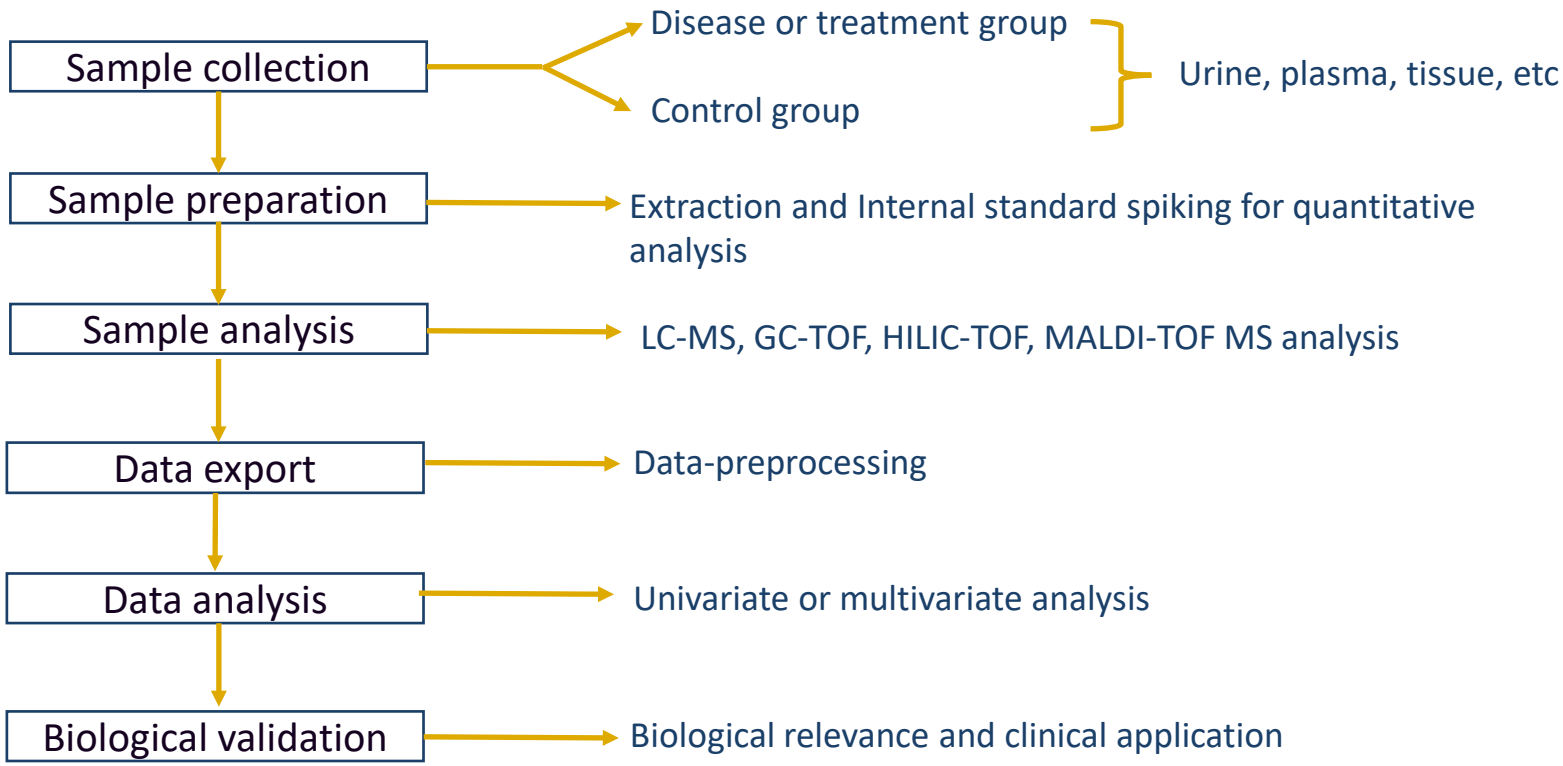
- Study of metabolome started decades ago with early applications in fields of toxicology, inborn metabolic errors, and nutrition.
- Biomarker in disease diagnosis, prognosis, and therapeutic response evaluation
- Pharmacology & pre-clinical drug trials – as pharmacodynamic marker of drug effect, including search for new drug targets
- Personalized medicine – screening/monitoring tool for initiation and progression of disease
- Nurtigenomics – study effect of diet on disease prevention as well as response to diet intervention

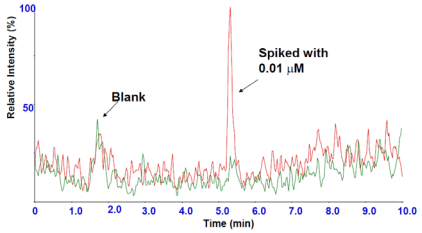
Metabolomics in the context of other omics



- Metabolome is complementary to the other –ome and is the link between genotype and phenotype.

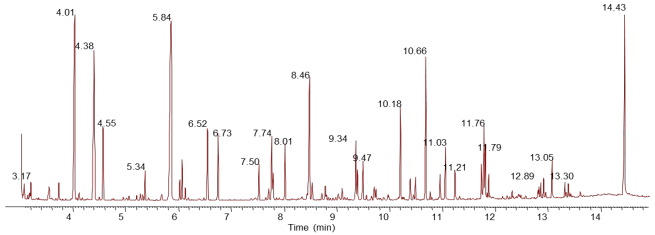
Workflow for metabolomics analysis



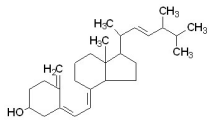


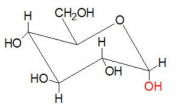
Steps involved in metabolomics analysis

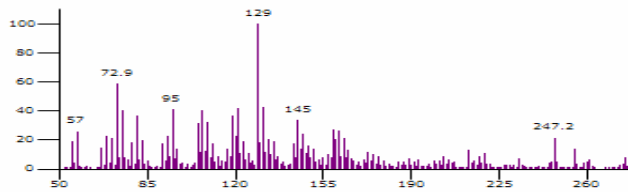
- Profiling involves finding of all metabolites detectable to a selected analytical technique with statistically significant variations in abundance within a set of experimental and control groups
- Identification of chemical structures of metabolites of interest after profiling
- Quantification and validation
- Interpretation of data making connections between the metabolites discovered and the biological conditions



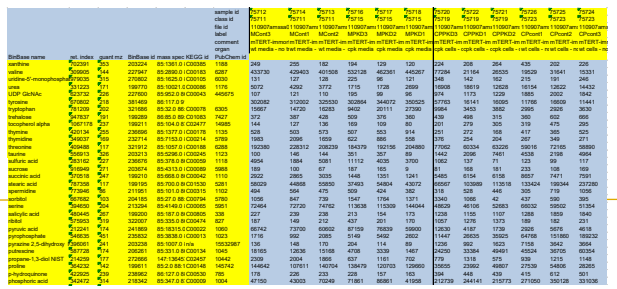
Chromatogram







MS/MS Fragmentation



UCDAVIS HEALTH | MIND INSTITUTE

Targeted vs. Untargeted metabolomics

	Targeted	Untargeted
Metabolites	pre-selected	all
Motivation	very focused, specific biochemical pathways	Open-end, global profiling
Pros	Sensitive, absolute quantification	novel (unknown) metabolites
Cons	Limited to pre-defined metabolites	relative quantification
Study	Hypothesis testing	Hypothesis generating

Example of raw metabolomics data

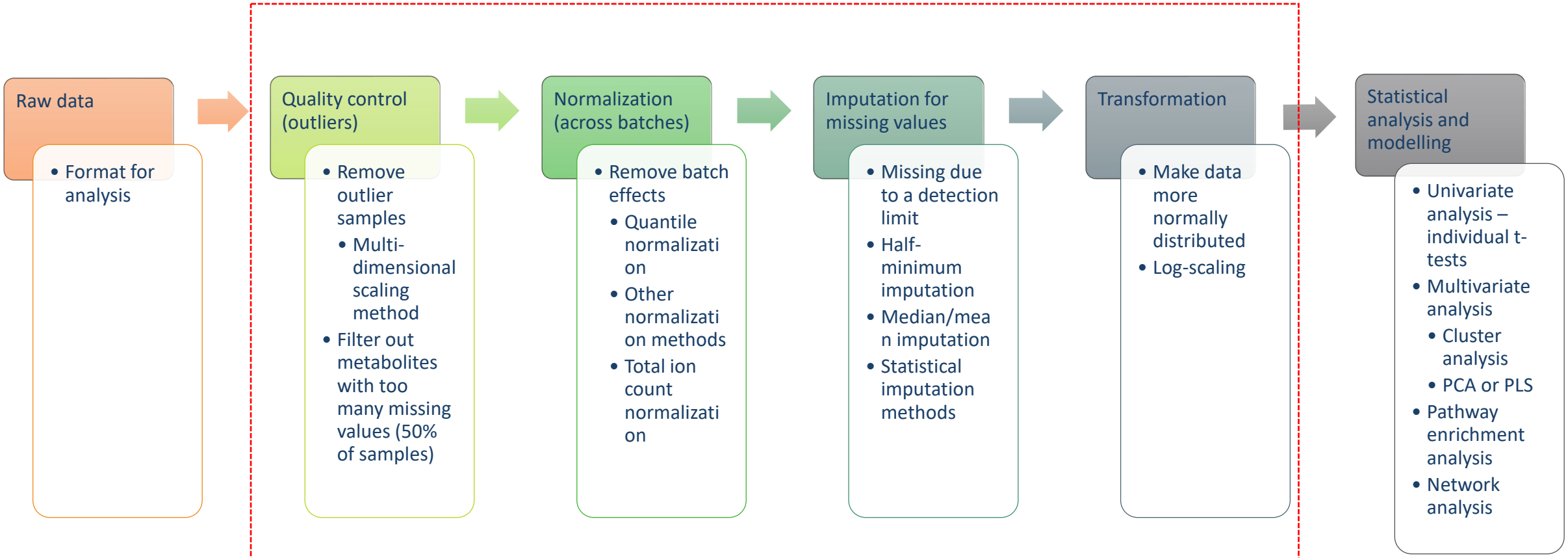
Metabolite name

Sample label (disease, control)

						sample id	75712	75714	75713	75716	75717	75718	75720	75722	75721	75726	75725	75724
						class id	75711	75711	75711	75715	75715	75715	75719	75719	75719	75723	75723	75723
						file id	110907amssa0	110907am	110907am	110907am	110907am	110907am	110907am	110907am	110907am	110907am	110907am	110907ams
						label	MCont3	MCont1	MCont2	MPKD3	MPKD2	MPKD1	CPPKD3	CPPKD1	CPPKD2	CPcont1	CPcont2	CPcont3
						comment	mTERT-immor	mTERT-im	mTERT-im	mTERT-im	mTERT-im	mTERT-im	mTERT-im	mTERT-im	mTERT-im	mTERT-im	mTERT-im	mTERT-im
						organ	wt media - no t	wt media -	wt media -	cpk media	cpk media	cpk media	cpk cells - (	cpk cells - (	cpk cells - (	wt cells - n	wt cells - n	wt cells - n
BinBase name	ret. index	quant mz	BinBase id	mass spec	KEGG id	PubChem id												
xanthine	702391	353	203224	85:1361.0	C00385	1188	249	255	182	194	129	120	224	208	264	435	202	226
valine	309905	144	227947	85:2890.0	C00183	6287	433730	429403	401508	532128	462361	445267	77284	21164	26535	19529	31641	15331
uridine-5'-monophospha	979035	315	270802	85:1625.0	C00105	6030	131	127	128	225	96	121	348	142	162	215	191	246
urea	331223	171	199770	85:10021.0	C00086	1176	5072	4292	3772	1715	1728	2699	16908	18619	12628	16154	12622	14432
UDP GlcNAc	623732	226	227600	85:952.0 8	C00043	445675	107	121	110	195	99	96	974	1173	1229	1885	2002	1842
tyrosine	670802	218	381469	86:117.0 9			302082	312002	325530	302864	344072	350525	57763	16141	16095	11766	16609	11441
tryptophan	781209	202	321686	85:32.0 86:	C00078	6305	15667	14720	16283	9402	20111	27390	9984	3453	3882	2995	2926	3630
trehalose	947837	191	199289	86:85.0 89:	C01083	7427	372	387	428	509	376	360	439	498	315	360	602	666
tocopherol alpha	1067178	237	199211	85:104.0 8	C02477	14985	144	127	136	169	109	80	201	279	305	319	475	295
thymine	420134	255	236696	85:1377.0	C00178	1135	528	503	573	507	553	914	251	272	168	417	365	525
thymidine	349037	169	232714	85:7153.0	C00214	5789	1983	2096	1659	622	886	558	376	254	204	267	349	271
threonine	409488	117	321912	85:1057.0	C00188	6288	192380	228312	208239	184379	192156	204880	77062	60334	63226	59016	72165	58890
taurine	556913	326	203213	85:5296.0	C00245	1123	100	146	144	351	357	89	1442	2096	7461	4538	2198	4964
sulfuric acid	283162	227	236676	85:378.0 8	C00059	1118	4954	1884	5081	11112	4035	3700	1062	137	71	123	99	117
sucrose	916949	271	203674	85:4313.0	C00089	5988	189	100	67	187	165	9	81	168	181	233	108	169
succinic acid	370518	247	199210	85:668.0 8	C00042	1110	2922	2865	3035	1448	1351	1241	5485	6154	6158	8657	7471	7591
stearic acid	787358	117	199195	85:700.0 8	C01530	5281	58029	44868	55850	37493	54804	43072	66567	103989	113518	133424	199344	237280
spermidine	773946	86	211951	85:101.0 8	C00315	1102	494	564	475	509	424	382	318	528	446	305	719	1056
sorbitol	667682	103	204185	85:27.0 88:	C00794	5780	1056	847	739	1547	1764	1371	3340	1066	42	437	590	395
serine	394650	204	213294	85:4149.0	C00065	5951	72464	72720	74762	113638	115309	144044	48629	46106	52683	66032	59502	51354
salicylic acid	480445	267	199200	85:187.0 8	C00805	338	227	239	238	213	154	173	1238	1155	1107	1288	1859	1840
ribitol	575953	319	322007	85:335.0 8	C00474	827	187	149	212	437	201	170	1057	1276	1373	189	182	231
pyruvic acid	212241	174	241869	85:18315.0	C00022	1060	66742	73700	60602	87159	76839	59900	12630	4187	1739	2926	5676	4618
pyrophosphate	546635	451	235832	85:3838.0	C00013	1023	1716	992	2085	5149	5492	2602	11447	26635	35925	64768	151860	189232
pyrazine 2,5-dihydroxy	396061	241	203238	85:1007.0	n/a	15532987	136	148	170	204	114	89	1236	992	1623	7158	3642	3664
putrescine	587728	174	206261	85:331.0 8	C00134	1045	18165	12636	15168	4708	3339	1467	24250	33384	49491	45524	36705	60354
propane-1,3-diol NIST	214259	177	272666	147:13645	C02457	10442	2309	2004	1866	637	1161	702	779	1318	575	939	1215	1148
proline	364232	142	199611	85:2.0 88:1	C00148	145742	144642	107611	140704	138479	120703	129660	35655	23992	49807	27539	54806	28265
p-hydroquinone	422925	239	238962	86:127.0 8	C00530	785	178	226	233	228	157	163	394	448	439	415	612	501
phosphoric acid	342472	314	218342	85:347.0 8	C00009	1004	47150	43003	70249	71861	86861	41958	212739	244141	215773	271050	350128	331036

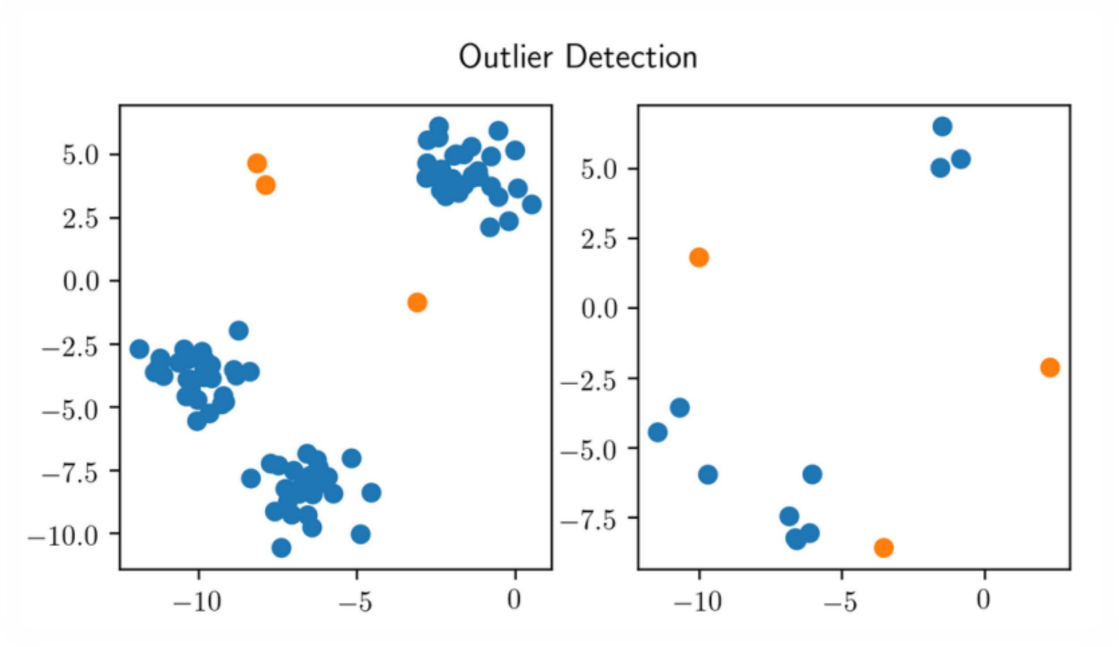
Metabolite concentration level (numeric)

Data pre-processing



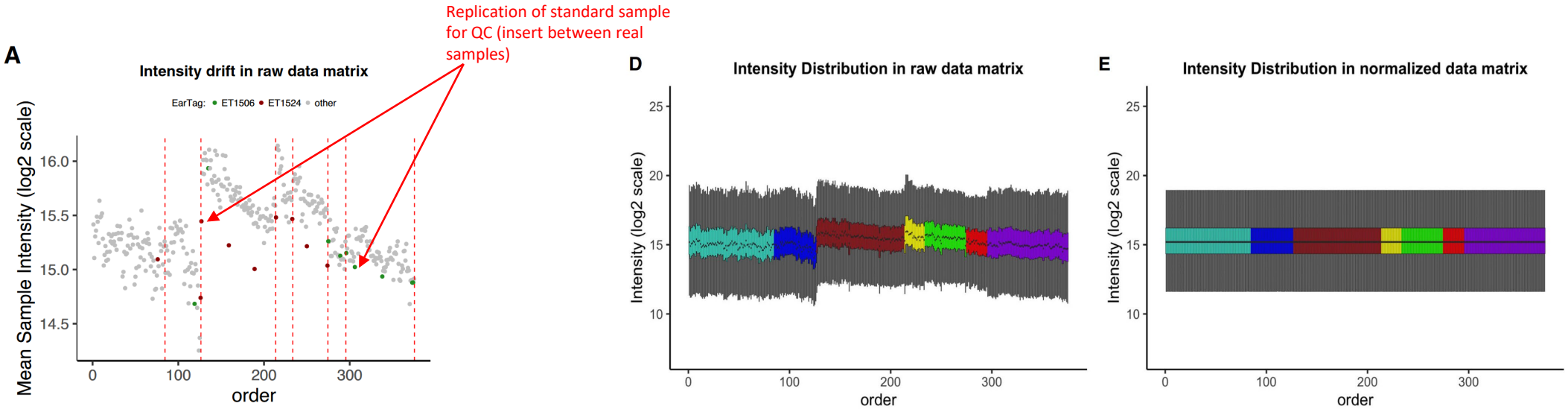
Quality control

- Remove sample outliers
  - **Multi-dimensional scaling method**
- Filter out metabolites with too many missing values (> 50% (??) of samples missing)



Normalization

- Remove batch effects
- Quantile normalization
- Other normalization methods



Mean intensity vs. sample running order (Repeatedly replicated samples are shown in red or green color. Vertical dotted lines indicate MS batch boundaries.)

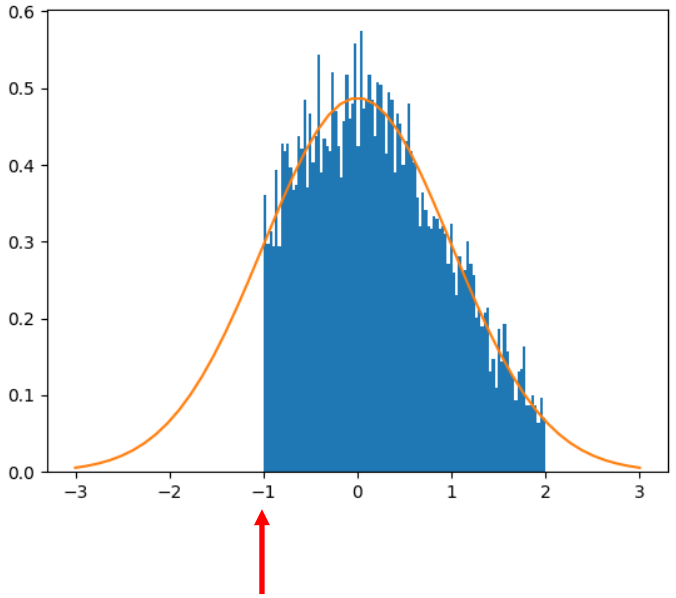
Before normalization

After normalization

Figures from Cuklina et al. Molecular Systems Biology (2021) 17:e10240

Imputation for missing values

- Missing due to a detection limit
- Half-minimum imputation
- Median/mean imputation
- Other sophisticated statistical imputation methods (e.g., left-censoring model)

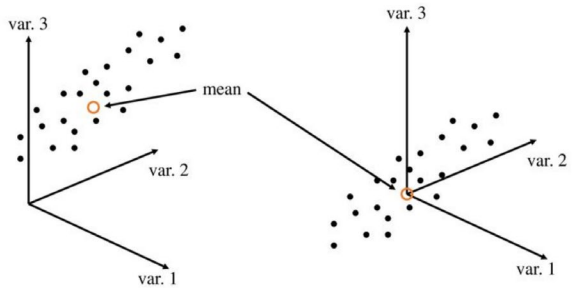


Lower limit of detection (LLD*) = -1

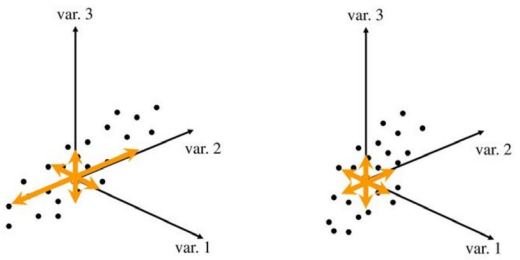
*LLD = the smallest amount of an analyte that can reliably be detected.

Transformation

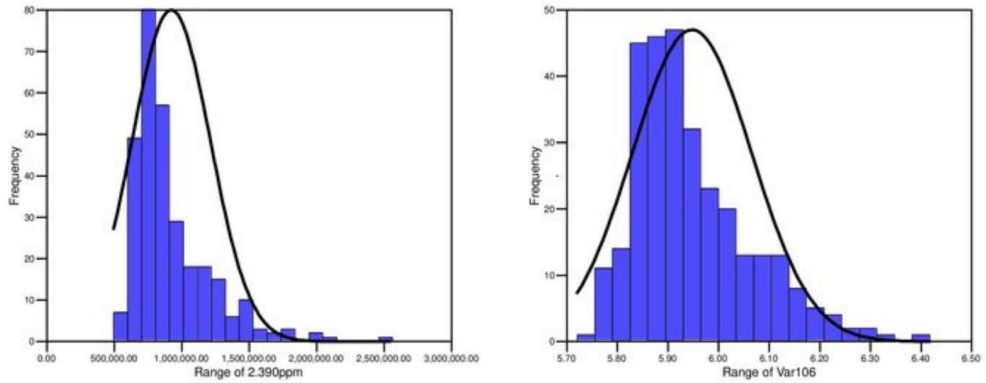
- Data is not normally distributed
- No scaling (mean-centering only)
- Pareto scaling (root square SD)
- Autoscaling (to unit variance)



Centering data – move center of each metabolite to the origin

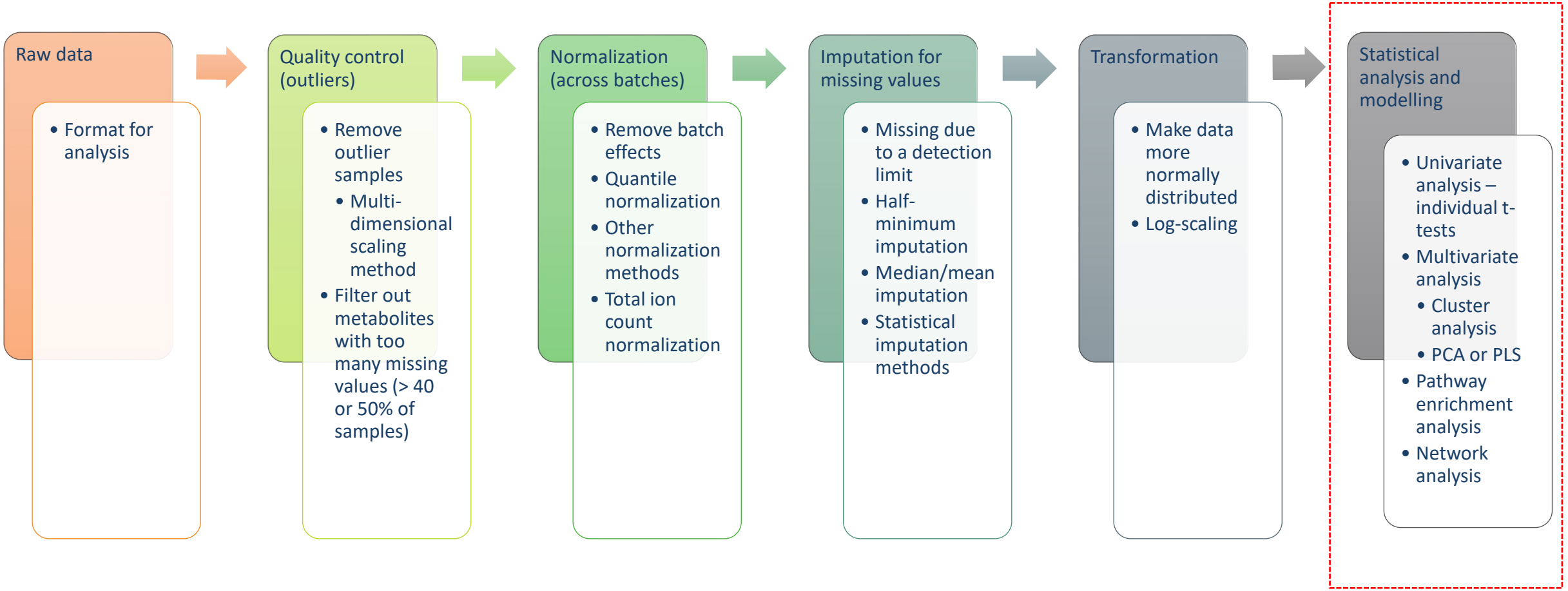


Scaling data– put each metabolite on an equal footing (make SDs equal)



Log-transforming data - convert log-normal distribution to a normal (symmetric) distribution

Statistical analysis and modelling



Univariate analysis

- To identify individual metabolites associated with the outcome
- Differential analysis (e.g., t-test or regression analysis)
- Correlation analysis (e.g., Pearson's or Spearman's correlation)
- Single-metabolite analysis needs to control for multiple testing issues (e.g., false discovery rate)
- Example:
  - 29 kidney cancer patients & 33 controls
  - 298 known metabolites – 274 after exclusion with > 50% missing
  - Half-min imputation, log2 transformation, & normalization

Table 1.

Results of Differential Analysis

		Mean levels[a]		
Top 13 most significant metabolite in order	***Diff. Analysis (cancer vs. control) p-value***	***Cancer***	***Control***	***Direction***
1. quinolinate	0.0026[b]	0.315	−0.571	up
2. *4-hydroxybenzoate*	0.0034[b]	−0.378	0.239	down
3. gentisate	0.0039[b]	−0.605	0.227	down
4. *galactitol (dulcitol)*	0.0085	−0.401	0.234	down
5. N-(2-furoyl)glycine	0.0087	−0.306	0.269	down
6. alpha-ketoglutarate	0.0100	−0.205	−0.960	up
7. *fructose*	0.0140	−0.257	0.226	down
8. *thymol sulfate*	0.0176	−0.578	−0.012	down
9. tryptophan betaine	0.0179	−0.481	0.070	down
10. *hexanoylglycine*	0.0332	0.134	−0.529	up
11. 1,6-anhydroglucose	0.0376	−0.239	0.210	down
12. 4guanidinobutanoate	0.0390	−0.197	0.102	down
13. 3-hydroxyphenylacetate	0.0488	−0.729	−0.250	down

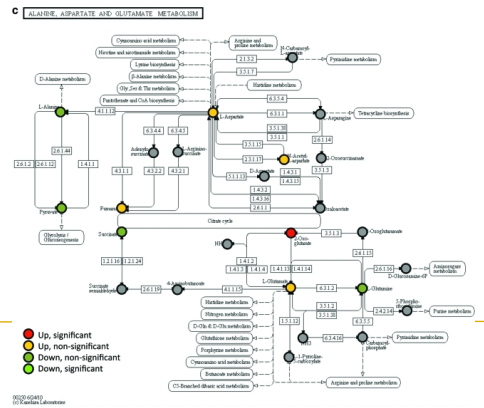
Metabolites whose expression differed significantly (at $p < 0.05$) between cancer and control patients are shown. Direction indicates whether cancer patients have lower (down) or higher (up) levels of the metabolite compared to control patients. Metabolites in *italic* indicate those involved in significant or suggestive pathways.

[a]Expression levels were normalized and log transformed as described in the text.
[b]*p*-value as significant at FDR of 0.26.

From Kim K et al. Urine metabolomic analysis identifies potential biomarkers and pathogenic pathways in kidney cancer. OMICS. 2011 May;15(5):293-303.

Aggregation-based pathway enrichment analysis

- To identify sets of relevant metabolites that act synergistically within functionally defined pathways
- Map to the significant metabolites and their altered directions into biochemical pathways using the KEGG database
- Example:
  - Found xenobiotic or dietary component pathway and fatty acid β-oxidation metabolism pathway to be significantly regulated in cancer patients compared to controls
- Biological validation



Accounted for the direction of regulation

Ignore the direction

Table 2. Results of Functional Score Analysis

Top 13 most significant in order	*Pathway*	*KEGG ID*	*HMDB ID*	# *Met*	# *Up*	# *Down*	% *Up*	p-*Value*[a] *partial* t-*statistics*	p-*Value*[a] *squared partial* t-*statistics*
1. quinolinate	Nicotinate and nicotinamide metabolism	C03722	HMDB00232	4	1	3	25%	0.363	0.403
2. 4-hydroxybenzoate	Benzoate metabolism	C00156	HMDB00500	10	0	10	0%	0.053[b]	0.110
3. gentisate	Phenylalanine and tyrosine metabolism	C00628	HMDB00152	19	2	17	11%	0.225	0.582
4. galacticol (dulcitol)	Fructose, mannose, galactose, starch and sucrose metabolism	C01697	HMDB00107	8	2	6	25%	0.092[b]	0.094[b]
5. N-(2-furoyl)glycine	Glycine, serine and threonine metabolism	NA	HMDB00439	9	1	8	11%	0.223	0.285
6. α-ketoglutarate	Krebs cycle	C00026	HMDB00208	10	7	3	70%	0.437	0.535
7. fructose	Fructose, mannose, galactose, starch and sucrose metabolism	C00095	HMDB00660	8	2	6	25%	0.092[b]	0.094[b]
8. thymol sulfate	Food component/ Plant	C09908	NA	6	1	5	17%	0.007[c]	0.027[c]
9. tryptophan betaine	Tryptophan metabolism	C00078	HMDB00929	11	3	8	27%	0.508	0.654
10. hexanoylglycine	Fatty acid, beta-oxidation metabolism	NA	HMDB00701	1	1	0	100%	0.032[c]	0.032[c]
11. 1,6-anhydroglucose	Glycolysis, gluconeogenesis, pyruvate metabolism	NA	HMDB00640	6	2	4	33%	0.473	0.785
12. 4-guanidinobutanoate	Guanidino and acetamido metabolism	C01035	HMDB03464	3	2	1	67%	0.498	0.654
13. 3-hydroxyphenylacetate	Phenylalanine and tyrosine metabolism	C05593	HMDB0040	19	2	17	11%	0.225	0.582

Pathways associated with the 13 most significant metabolites (Table 1) are shown. #Met is the number of metabolites identified through GC/MS or LC/MS in each pathway. % Up is the percentage of metabolites upregulated in each pathway. *p*-Value partial *t*-statistic is the *p*-value obtained by using unsquared *t*-statistics that accounted for the direction of regulation, and *p*-value squared partial *t*-statistics is the *p*-value irrespective of the direction of regulation in each pathway.

[a]*p*-Values were calculated from a permutation null distribution based on 10,000 permutations.

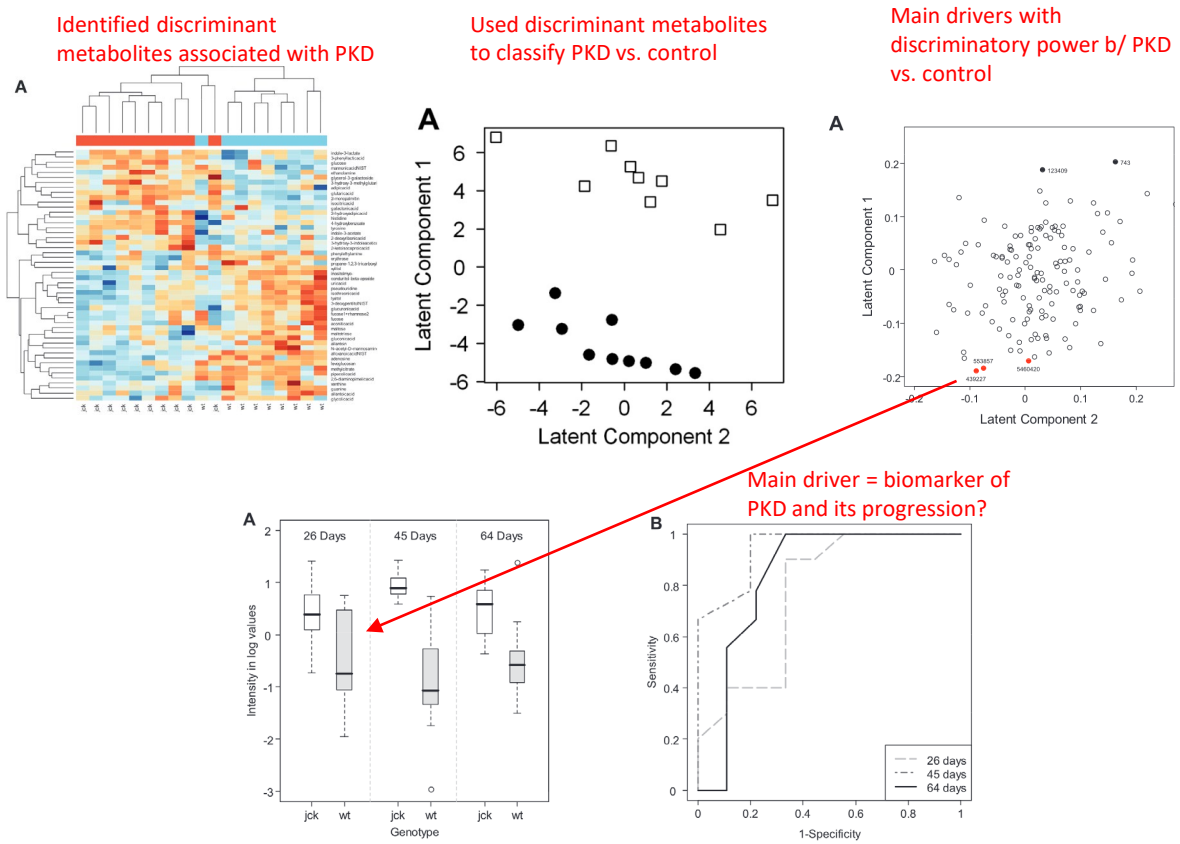
[b]*p*-value significant at 0.1; [c]*p*-value significant at 0.05.

Multivariate analysis

- Extract info from data with multiple variables by using all the variables simultaneously
- Determine whether the metabolome can discriminate disease from control (e.g., classification or discrimination analysis)
- Construct multi-metabolite ‘signature’ for disease
- Principal component analysis (PCA) or cluster analysis - unsupervised learning methods
- Partial least squares regression-linear discriminant analysis (PLS-LDA)- supervised learning method
- Other machine learning methods

Multivariate analysis: Example

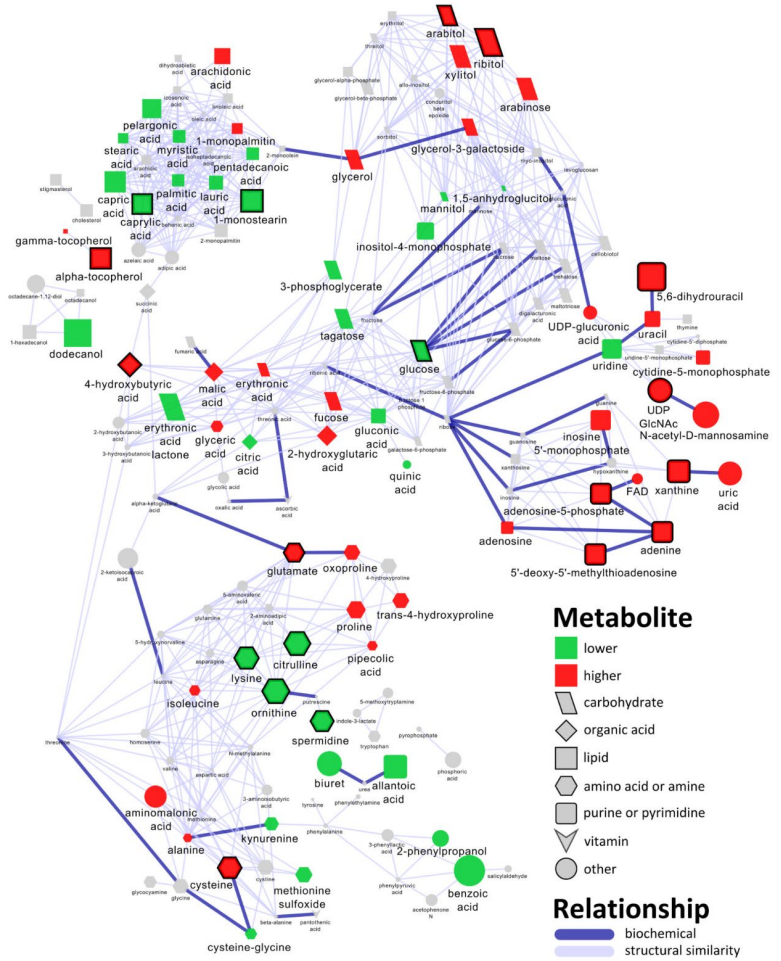
- Example:
  - Autosomal dominant polycystic kidney disease (ADPKD), the most common inherited kidney disease
  - Utilized urinary metabolomics to identify biomarkers for ADPKD and its progression
  - Samples from PKD and controls collected at 26 (before evidence of kidney dysfunction), 45, and 64 days.
  - Conducted PLS-LDA analysis



From Taylor SL et al. A metabolomics approach using juvenile cystic mice to identify urinary biomarkers and altered pathways in polycystic kidney disease. Am J Physiol Renal Physiol. 2010 Apr;298(4):F909-22.

Network analysis

- Metabolomic network of biochemical differences between cancer and control
- Strength of inter-relationships between metabolites
- Node color displays significance and direction of the change in cancer relative to control (green, decrease; red, increase; gray, insignificant change)
- Node size displays PLS-DA loadings for selected discriminants for cancer.
- Node shape denotes the biochemical super class of each molecule.



Wikoff et al. Metabolomic markers of altered nucleotide metabolism in early stage adenocarcinoma. Cancer Prev Res (Phila). 2015 May;8(5):410-8.

Experimental design – need to consider

- Your research hypotheses
- Type of sample
- Need to consider which MS technique will be used for metabolomics analysis (LC, GC-TOF, etc.)
- Statistical analysis method
  - study design
  - hypotheses/objectives
- Sample size
  - signal to noise
  - sample-to-sample variability within and between groups
    - Cell culture (very low), tissue (medium), blood and urine (high)

Biostatistics Support is available

- MIND IDDRC (full support at no charge but available only to IDDRC membership)

- CTSC and Cancer Center Biostatistics Office Hours
  - Every Tuesday from 12-2pm currently via WebEx
  - Sign-up through the CTSC biostatistics website

- EHSC Biostatistics Drop-in Office hours
  - Every Monday 2-4pm or Upon request

- Request Biostatistics Consultations through the center websites
  - MIND IDDRC
  - CTSC
  - Cancer Center Shared Resources
  - EHSC

Questions?

- Thank you for attending the *Applied Statistics for Translational Researchers Seminar Series!*