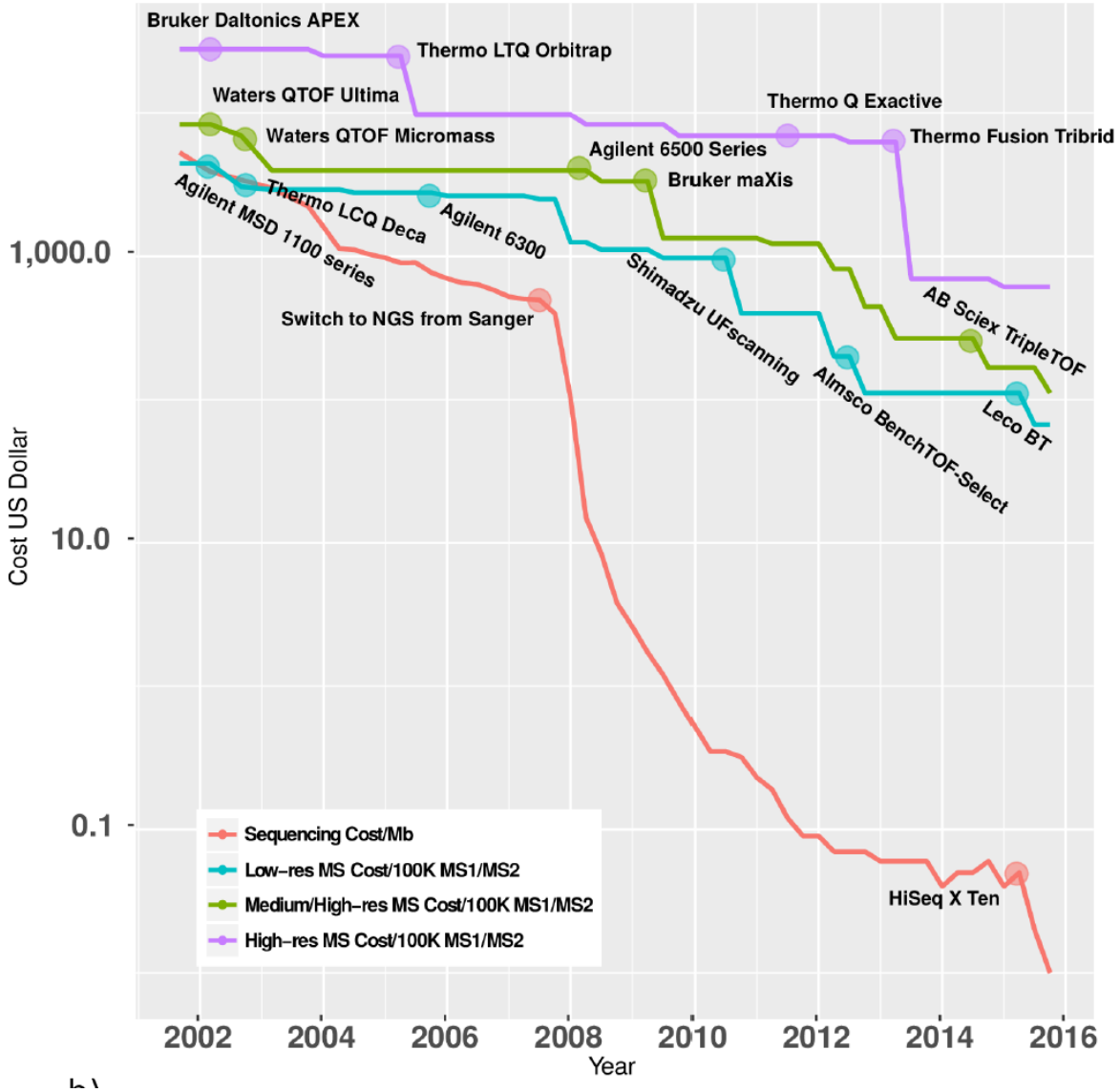
The past, present and future of data in small molecule analysis, including metabolomics, and what impact standards will have.

Pieter C Dorrestein
Twitter: Pdorrestein1

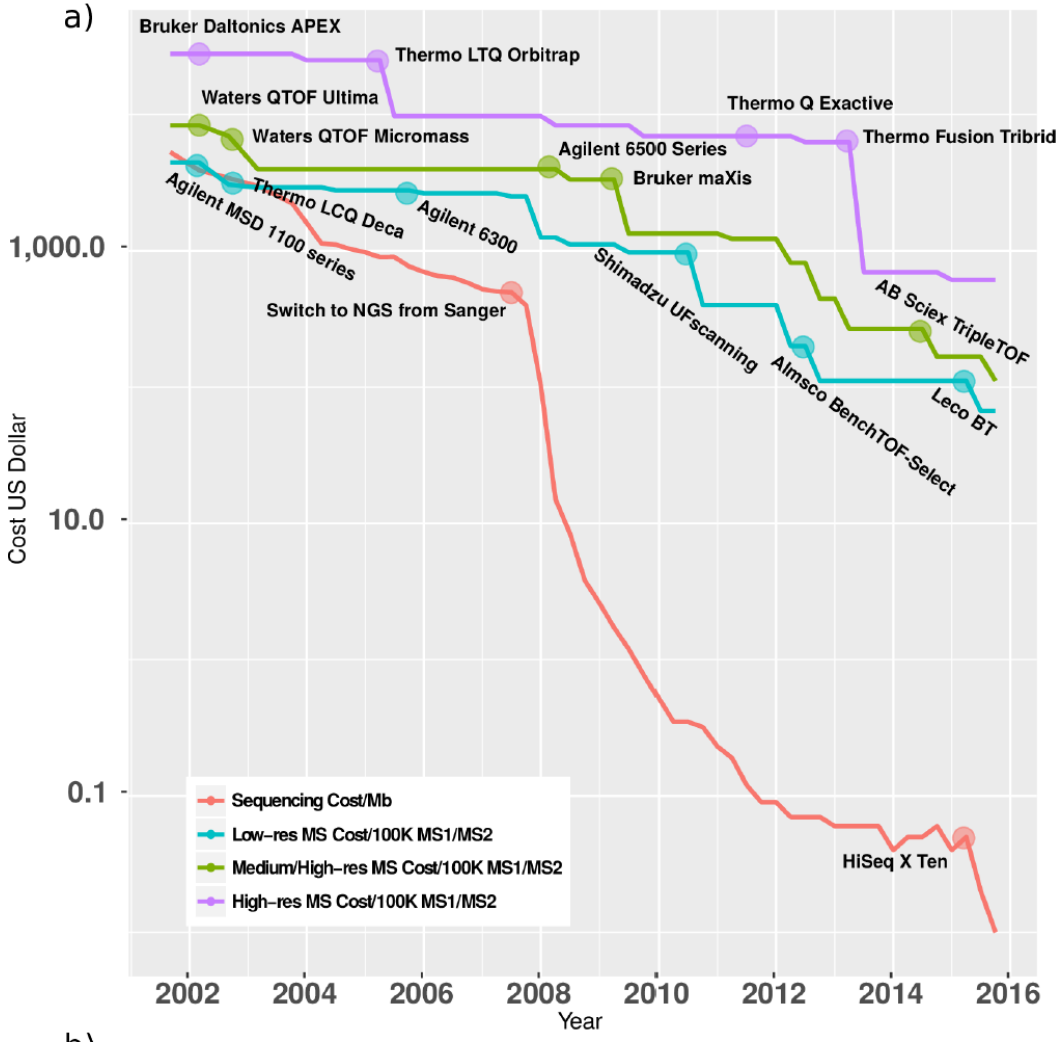
You are welcome to take screenshots and share the work that is presented

We are entering an amazing time in mass spectrometry



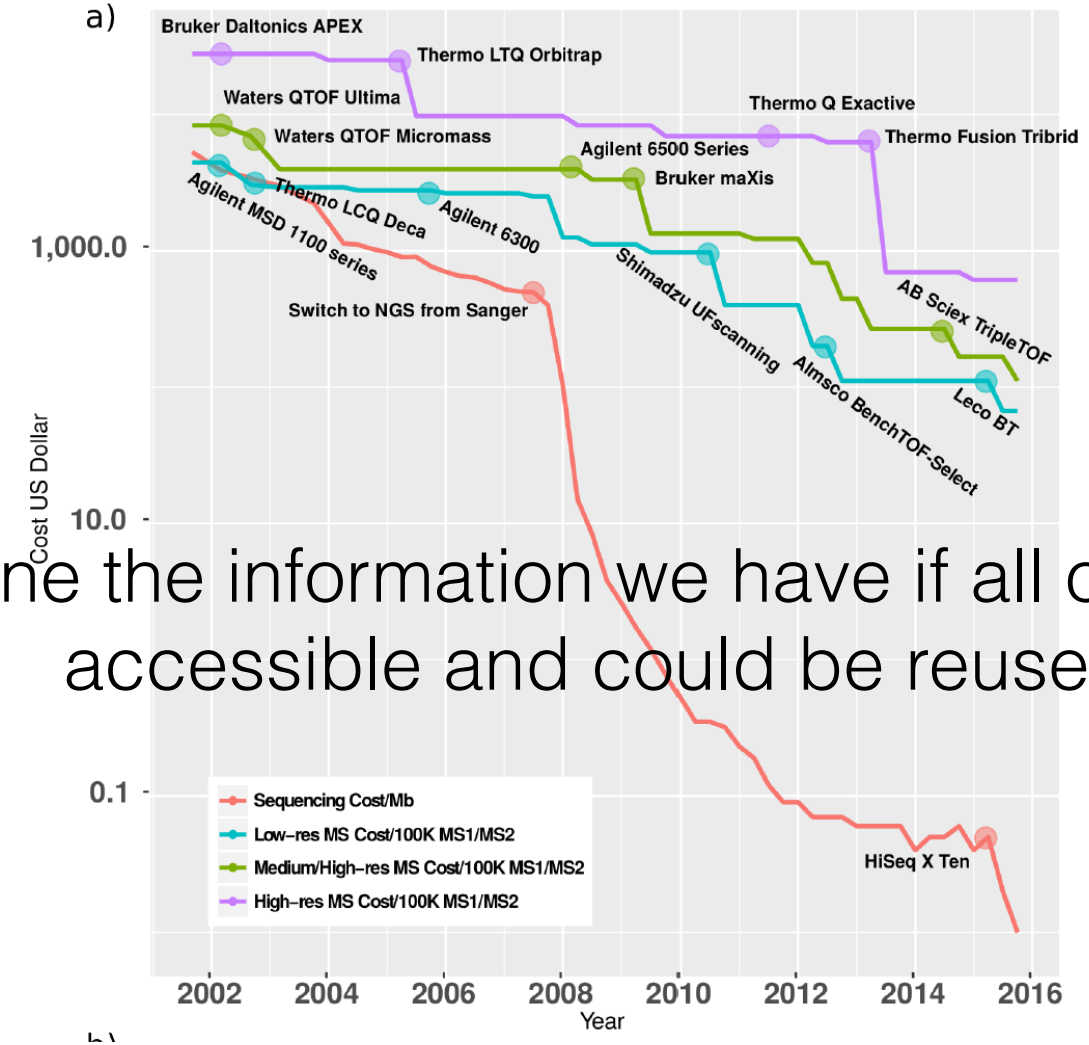
Aksenov et al Nat Rev Chemistry

We are entering an amazing time in mass spectrometry



We can collect more high quality, high resolution data than ever before

We are entering an amazing time in mass spectrometry



Just imagine the information we have if all of the data was accessible and could be reused?

We can collect more high quality, high resolution data than ever before

There are ~ a dozen papers in the literature that reuse small molecule mass spectrometry data sets, and ~50% of them are for my own lab.

Why??

Different communities have different standards and are far behind compared to sequencing.

Comparatively IMO still stuck in…

1.44Mb

What if we ask this of molecules? where do we go?

Retrieval infrastructure

What if we ask this of molecules? where do we go?

https://www.kqed.org/news/10622688

Different communities have different standards and are far behind compared to sequencing.

Comparatively IMO still stuck in…

Basic Local Alignment Search Tool

Stephen F. Altschul[1], Warren Gish[1], Webb Miller[2]
Eugene W. Myers[3] and David J. Lipman[1]

[1]*National Center for Biotechnology Information*
National Library of Medicine, National Institutes of Health
Bethesda, MD 20894, U.S.A.

[2]*Department of Computer Science*
The Pennsylvania State University, University Park, PA 16802, U.S.A.

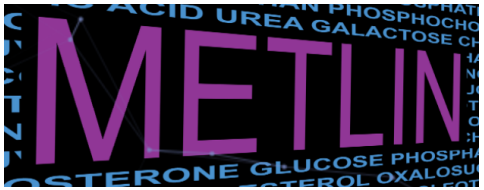
[3]*Department of Computer Science*
University of Arizona, Tucson, AZ 85721, U.S.A.

(Received 26 February 1990; accepted 15 May 1990)

A new approach to rapid sequence comparison, basic local alignment search tool (BLAST), directly approximates alignments that optimize a measure of local similarity, the maximal segment pair (MSP) score. Recent mathematical results on the stochastic properties of MSP scores allow an analysis of the performance of this method as well as the statistical significance of alignments it generates. The basic algorithm is simple and robust; it can be implemented in a number of ways and applied in a variety of contexts including straightforward DNA and protein sequence database searches, motif searches, gene identification searches, and in the analysis of multiple regions of similarity in long DNA sequences. In addition to its flexibility and tractability to mathematical analysis, BLAST is an order of magnitude faster than existing sequence comparison tools of comparable sensitivity.

..and there were 1000s algorithms developed (and are still being developed.)

There are a variety of endeavors (not an exhaustive list below) that are aiming to tackle data analysis and/or the data sharing and analysis reproducibility problem.



MetaboLights

MoNA - MassBank of North America

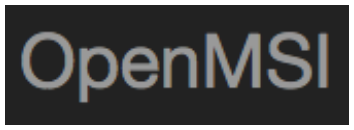
PhenoMeNal

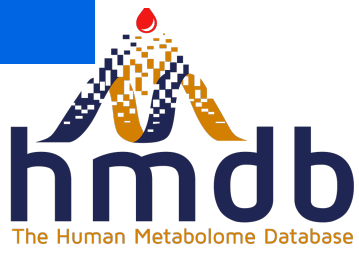
ReSpect

Large-Scale Computing for Medical Metabolomics

KBase ?

PREDICTIVE BIOLOGY





No expectation for data and or knowledge sharing in digital format.

Of ~20 labs I have asked to share data, 0 have been able to.

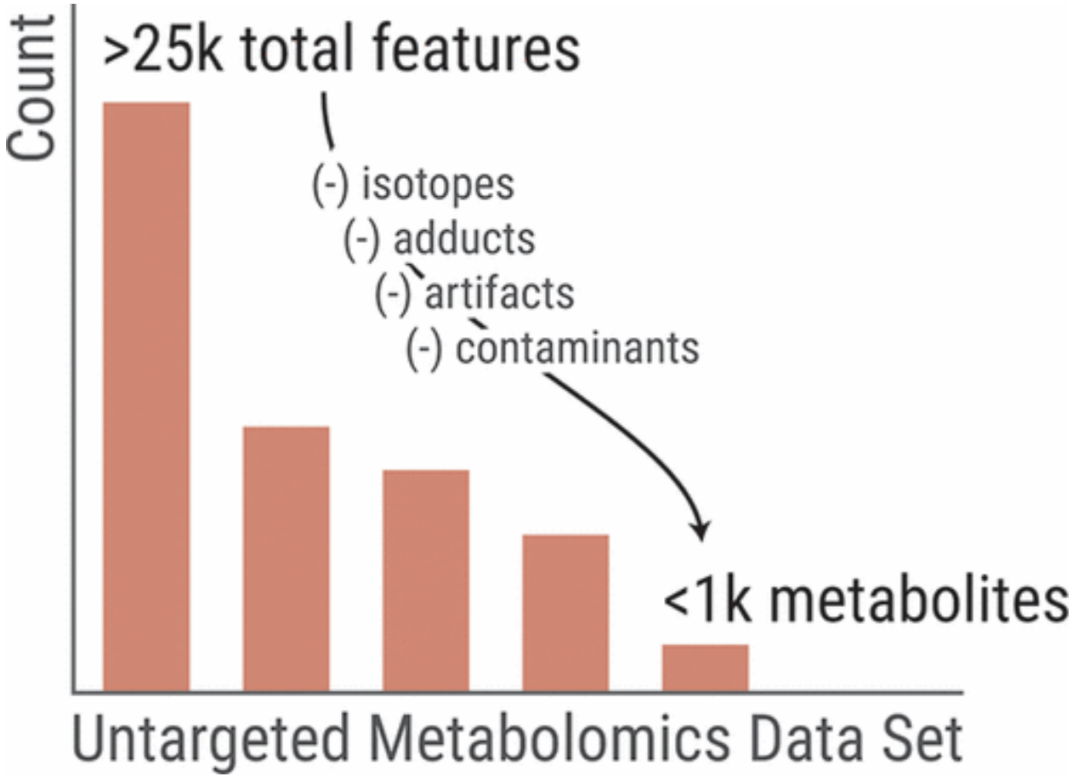
Comparatively IMO still stuck in…

1.44Mb

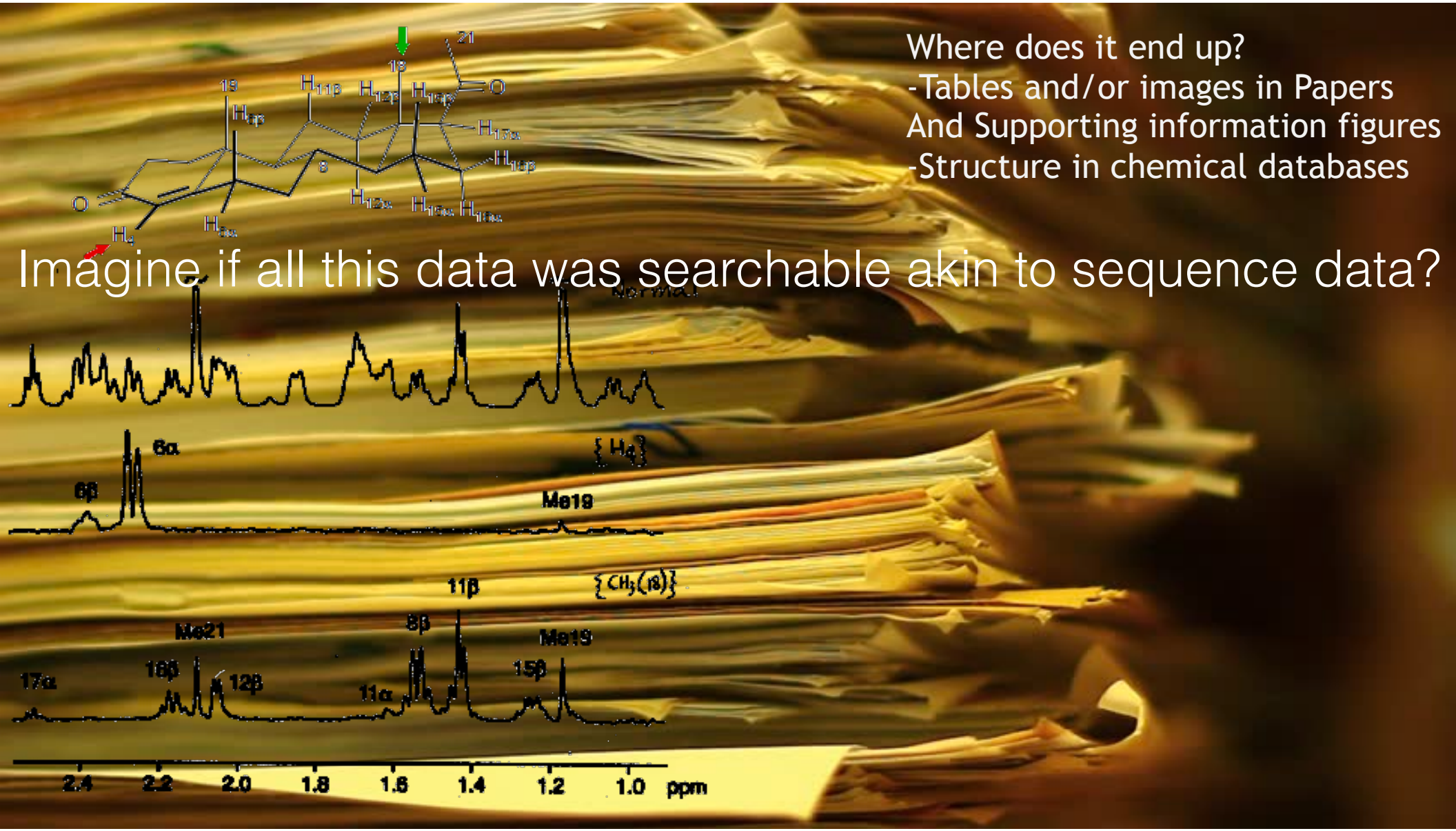
analytical chemistry

Systems-Level Annotation of a Metabolomics Data Set Reduces 25000 Features to Fewer than 1000 Unique Metabolites

Anal. Chem., **2017**, *89* (19), pp 10397–10406
DOI: 10.1021/acs.analchem.7b02380
Publication Date (Web): September 15, 2017



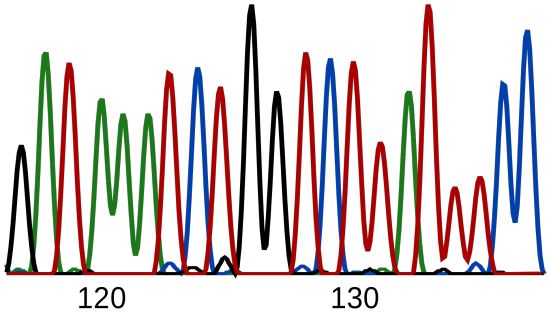
What do you get when you ask a natural product scientist on data from their extract or molecule?



I think raw data deposition is critical-not everyone agrees in the MS community. People that disagree rationalize what is done in sequencing.

Databases

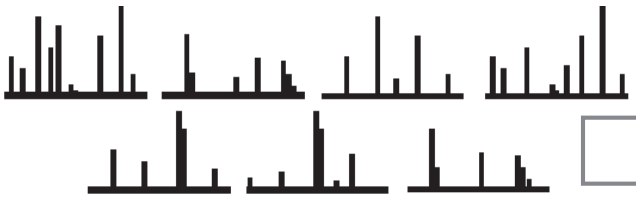
Knowledge bases



GAT AAAT CT GGTCTT ATTTCC

ORF
Gene name
Organism name
etc.
Annotation

"Image" Base Calling
Quality Estimation



Molecule name
Molecular Formula/RT
etc.

Annotation
w/o Quality Estimation

On the Future of Genomic Data

Scott D. Kahn
+ See all authors and affiliations

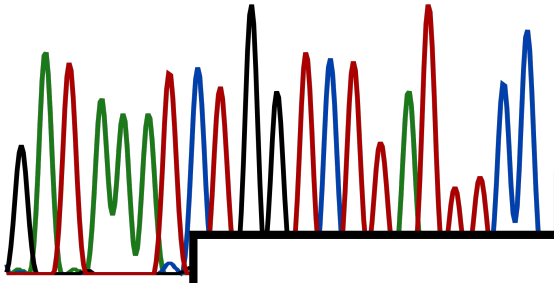
Science 11 Feb 2011:

Raw Genomic Data: Storage, Access, and Sharing

Mahsa Shabani,[1,*]
Danya Vears,[1] and
Pascal Borry[1]

Cell press reviews 2018

I think raw data deposition is critical-not everyone agrees in the MS community. People that disagree rationalize what is done in sequencing.



Databases

Knowledge bases

GAT AAAT CT GG TCTT ATTTCC

ORF

Gene name

n name

ation

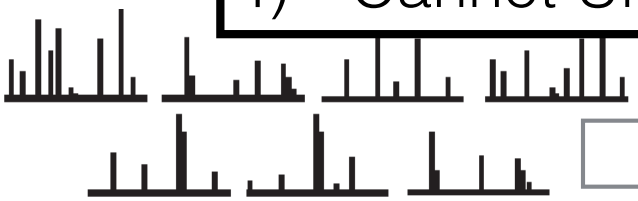
Disagree

1) No quality est.

2) Not all MS signals have an annotation.

3) Annotations are the best knowledge in time.

4) Cannot Understand where annotations come from.



Molecule name

Molecular Formula/RT

etc.

Annotation

w/o Quality Estimation

On the Future of Genomic Data

Scott D. Kahn

+ See all authors and affiliations

Science 11 Feb 2011:

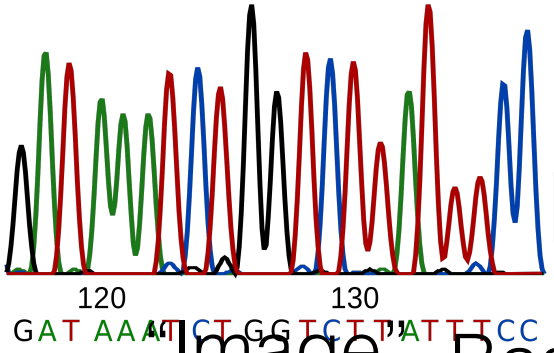
Raw Genomic Data: Storage, Access, and Sharing

Mahsa Shabani,[1,*] Danya Vears,[1] and Pascal Borry[1] Cell press reviews 2018

I think raw data deposition is critical-not everyone agrees in the MS community. People that disagree rationalize what is done in sequencing.

Databases

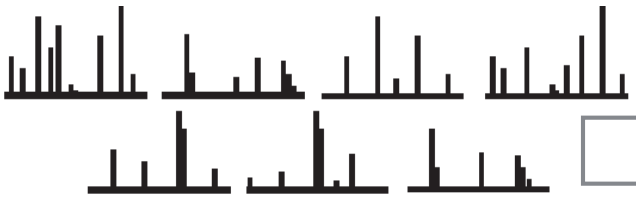
Knowledge bases



GAT AAAT CT GG TCT T ATT TCC

"Image" Base Calling
Quality Estimation

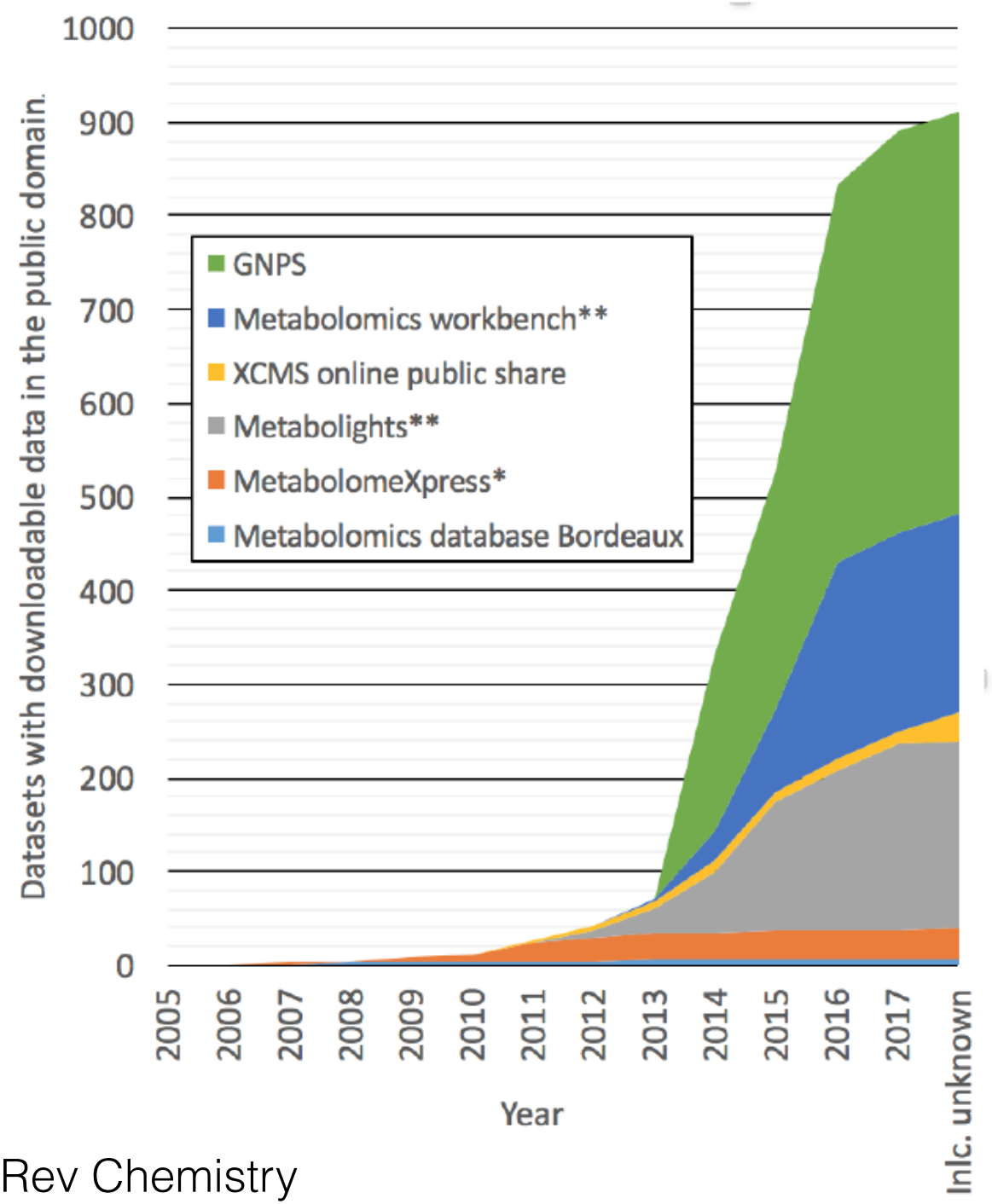
ORF
Gene name
Organism name
etc.
Annotation



Molecule name
Molecular Formula/RT
etc.

Annotation
w/o Quality Estimation

Metabolomics Projects with Downloadable Data in the Public Domain March 7, 2017



Aksenov et al Nat Rev Chemistry

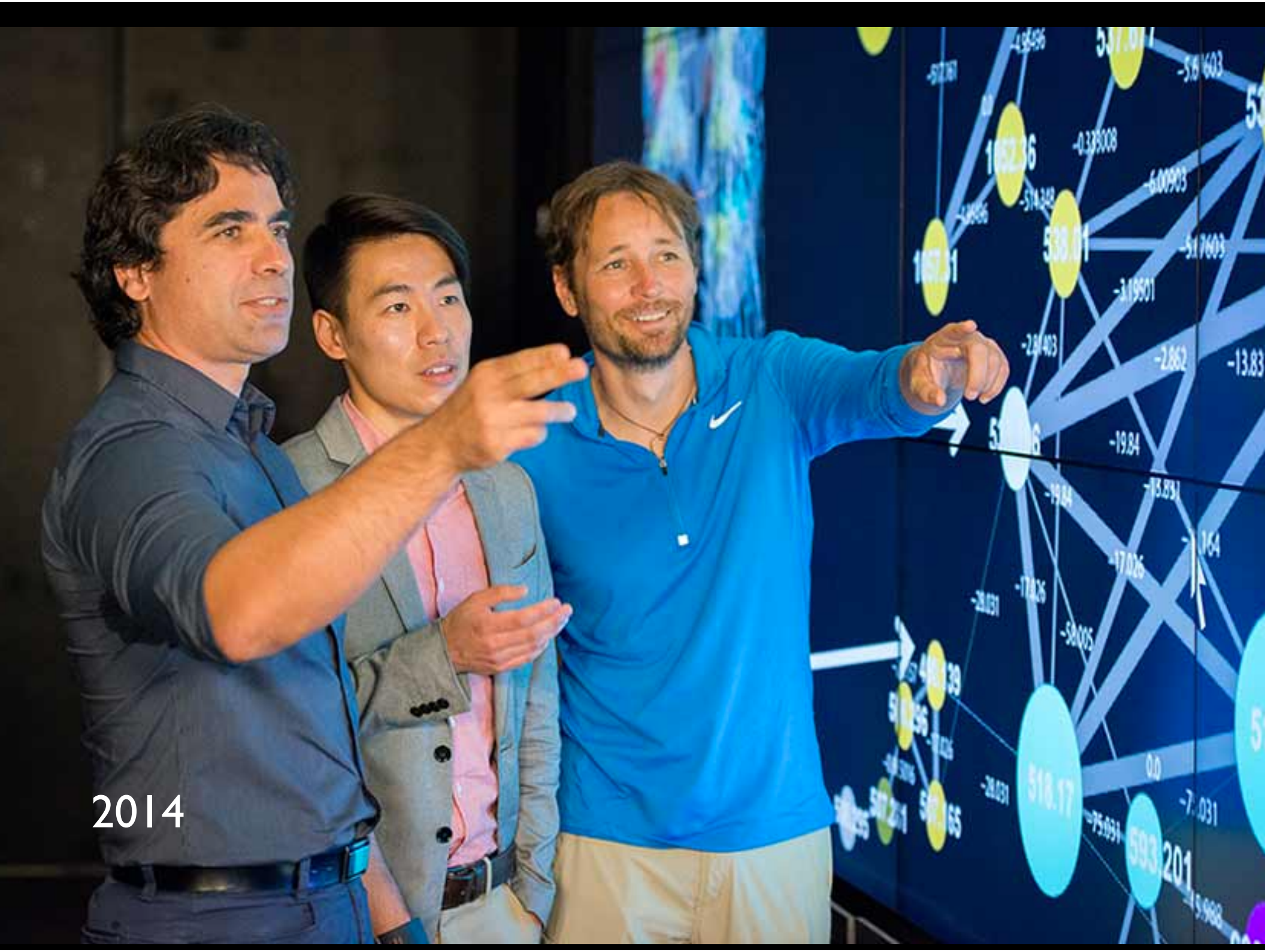
Where is this molecule found? any analogs?

Status for small molecule mass spectrometry analysis

1) No centralized retrieval infrastructure
2) Not enough data in the public domain

Where is this molecule found? any analogs?

MERIB


2014

Being able to scale analysis is important.

>1,000 MS data files normal, must be possible to process
100,000s to
1,000,000s of files but ultimately billions of files.
(currently almost 500,000 files in our infrastructure)

Scaling enables to disseminate knowledge to the world.

Written with 127 authors, nature biotechnology

http://www.nature.com/nbt/journal/v34/n8/full/nbt.3597.html

Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking

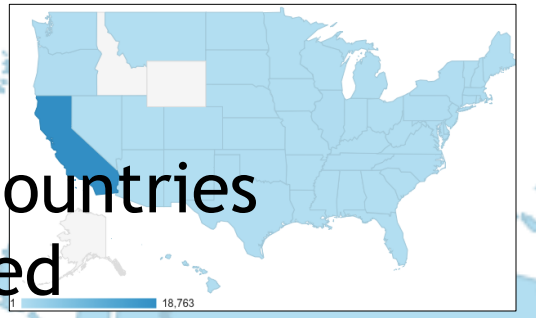
The potential of the diverse chemistries present in natural products (NP) for biotechnology and medicine remains untapped because NP databases are not searchable with raw data and the NP community has no way to share data other than in published papers. Although mass spectrometry (MS) techniques are well-suited to high-throughput characterization of NP, there is a pressing need for an infrastructure to enable sharing and curation of data. We present Global Natural Products Social Molecular Networking (GNPS; http://gnps.ucsd.edu), an open-access knowledge base for community-wide organization and sharing of raw, processed or identified tandem mass (MS/MS) spectrometry data. In GNPS, crowdsourced curation of freely available community-wide reference MS libraries will underpin improved annotations. Data-driven social-networking should facilitate identification of spectra and foster collaborations. We also introduce the concept of 'living data' through continuous reanalysis of deposited data.

NP from marine and terrestrial environments, including their inhabiting microorganisms, plants, animals, and humans, are routinely analyzed using MS. However, a single MS experiment can collect thousands of MS/MS spectra in minutes[1], and individual projects can acquire millions of spectra. These data sets are too large for manual analysis. Furthermore, comprehensive software and proper computational infra-

and UniProt KnowledgeBase (UniProtKB), which provide robust platforms for data sharing and knowledge dissemination[9,10]. Recognizing the need for an analogous community platform to analyze NP MS data, we present GNPS. GNPS is a data-driven platform for the storage, analysis, and knowledge dissemination of MS/MS spectra that enables community sharing of raw spectra, continuous annotation of deposited data, and collaborative curation of reference spectra (referred to as spectral libraries) and experimental data (organized as data sets).

GNPS provides the ability to analyze a data set and to compare it to all publicly available data. By building on the computational infrastructure of the University of California San Diego (UCSD) Center for Computational Mass Spectrometry (CCMS; http://proteomics.ucsd.edu/), GNPS provides public data set deposition and/or retrieval through the Mass Spectrometry Interactive Virtual Environment

January 2014 to November 2015, GNPS grew to serve 9,267 users

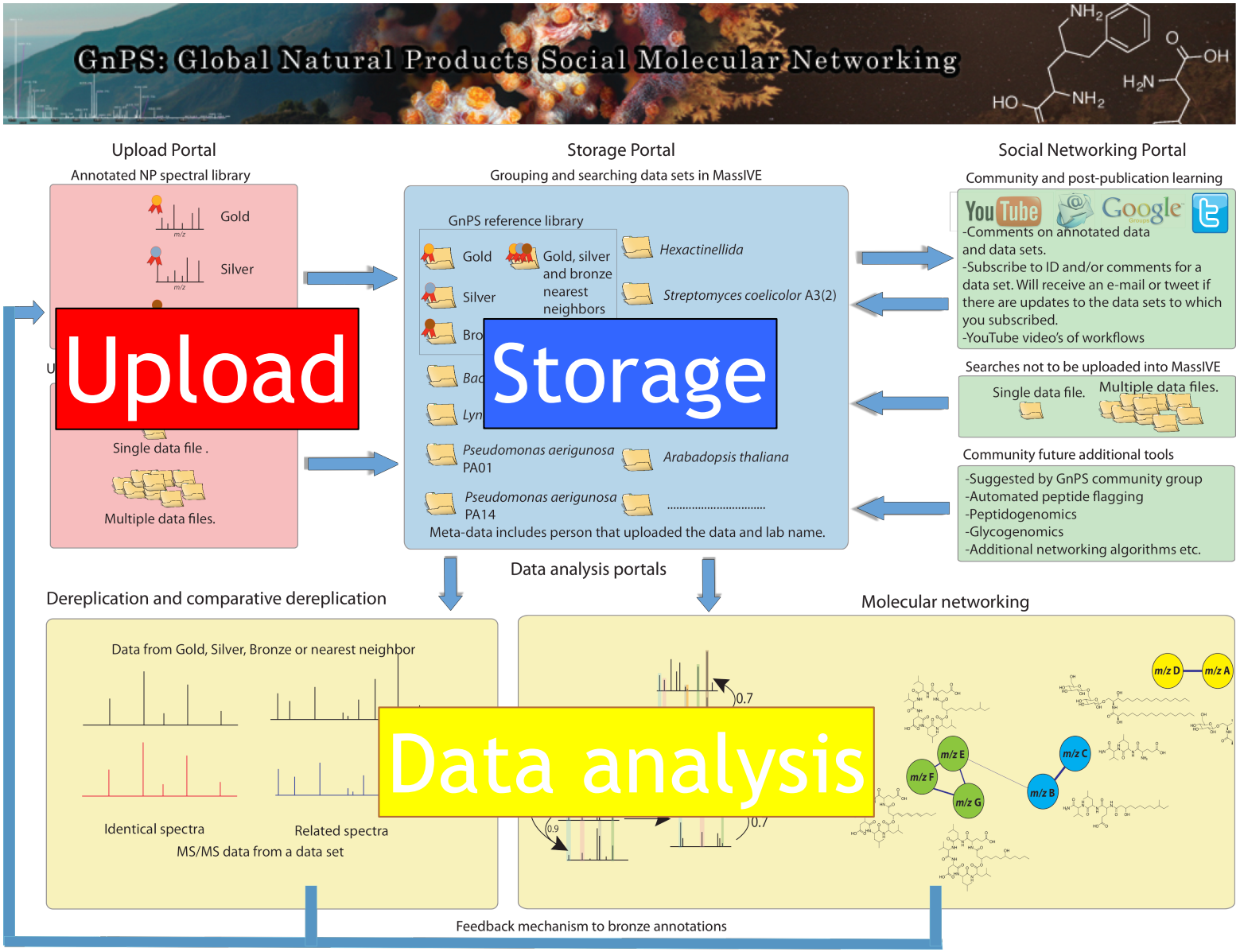


>40,000 users from 130+ countries
>20 billion spectra searched
Alaska is the only state in the US that has not used GNPS

GNPS is a metabolomics analysis and knowledge capture platform (analysis is crowd sourced). All data and reference spectra are free to download.

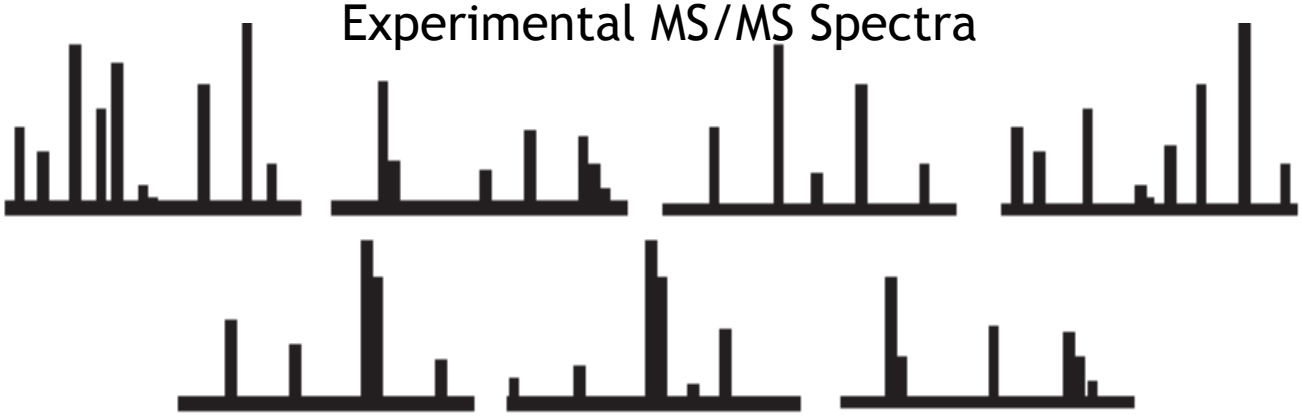
1 23,524

Nature Biotech 2016



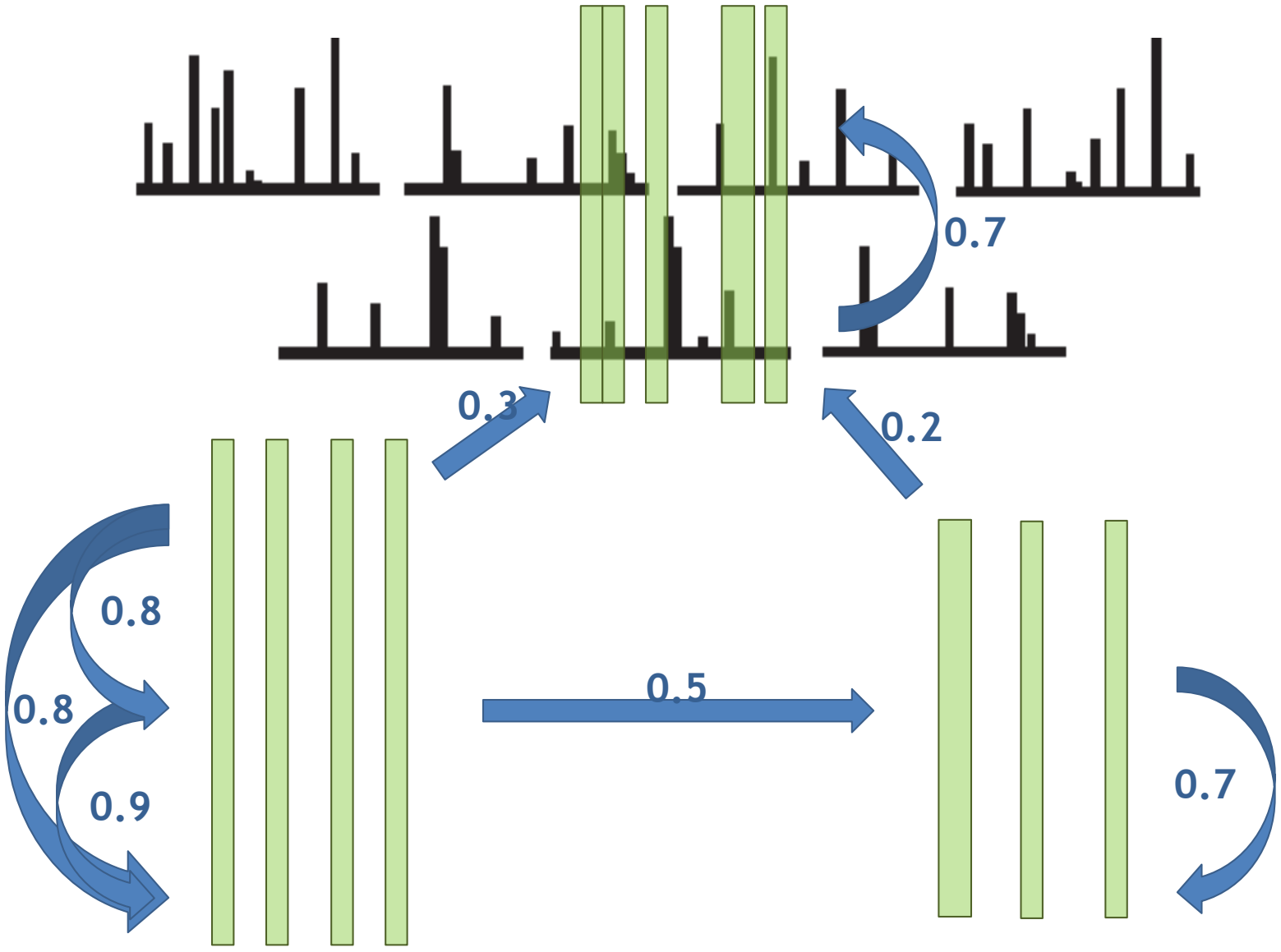
Nature Biotech 2016

Molecular networking

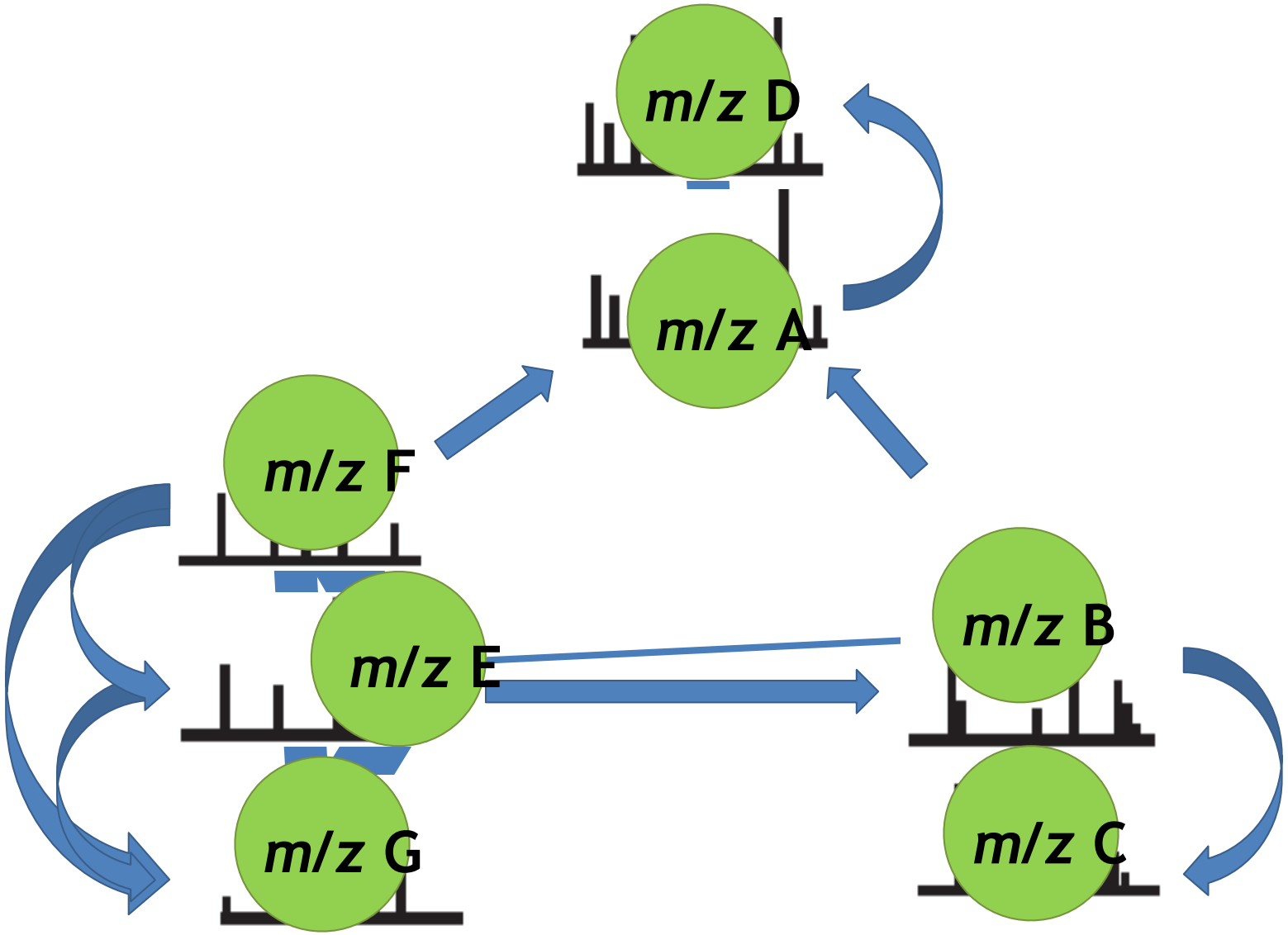


Watrous, JD *et al*. “Mass spectral molecular networking of living microbial colonies” *PNAS* (2012)

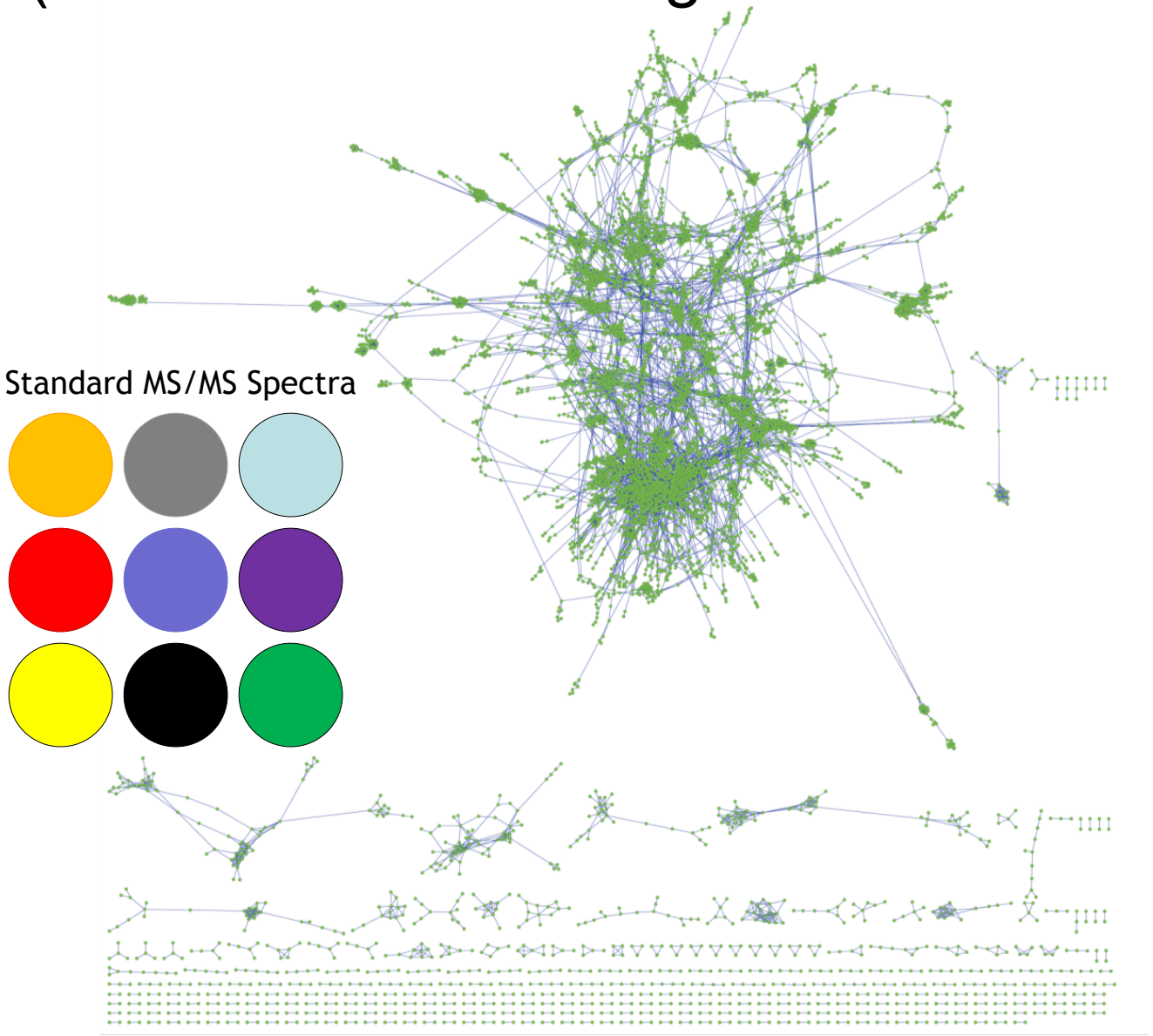
Molecular networking

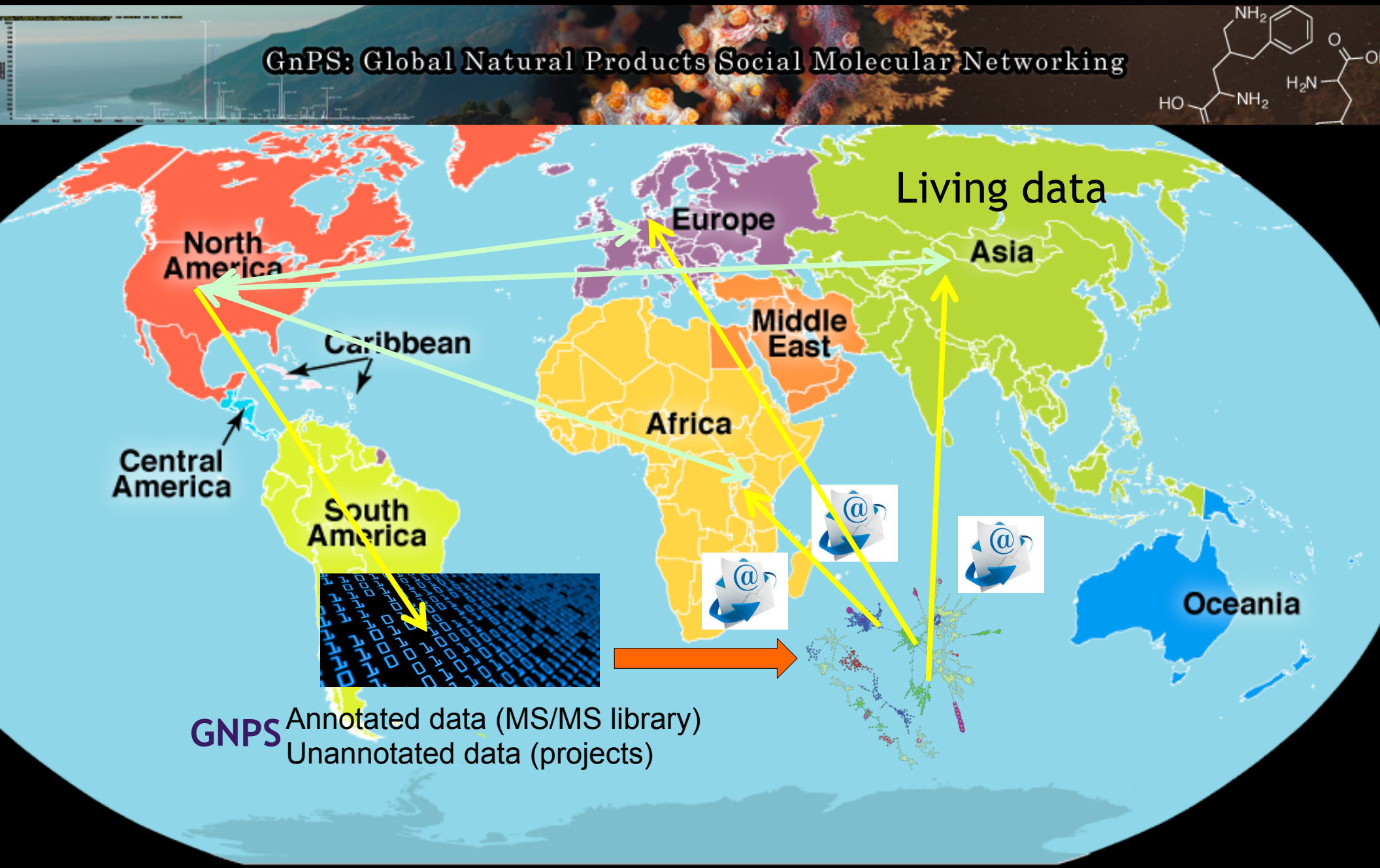


Molecular networking



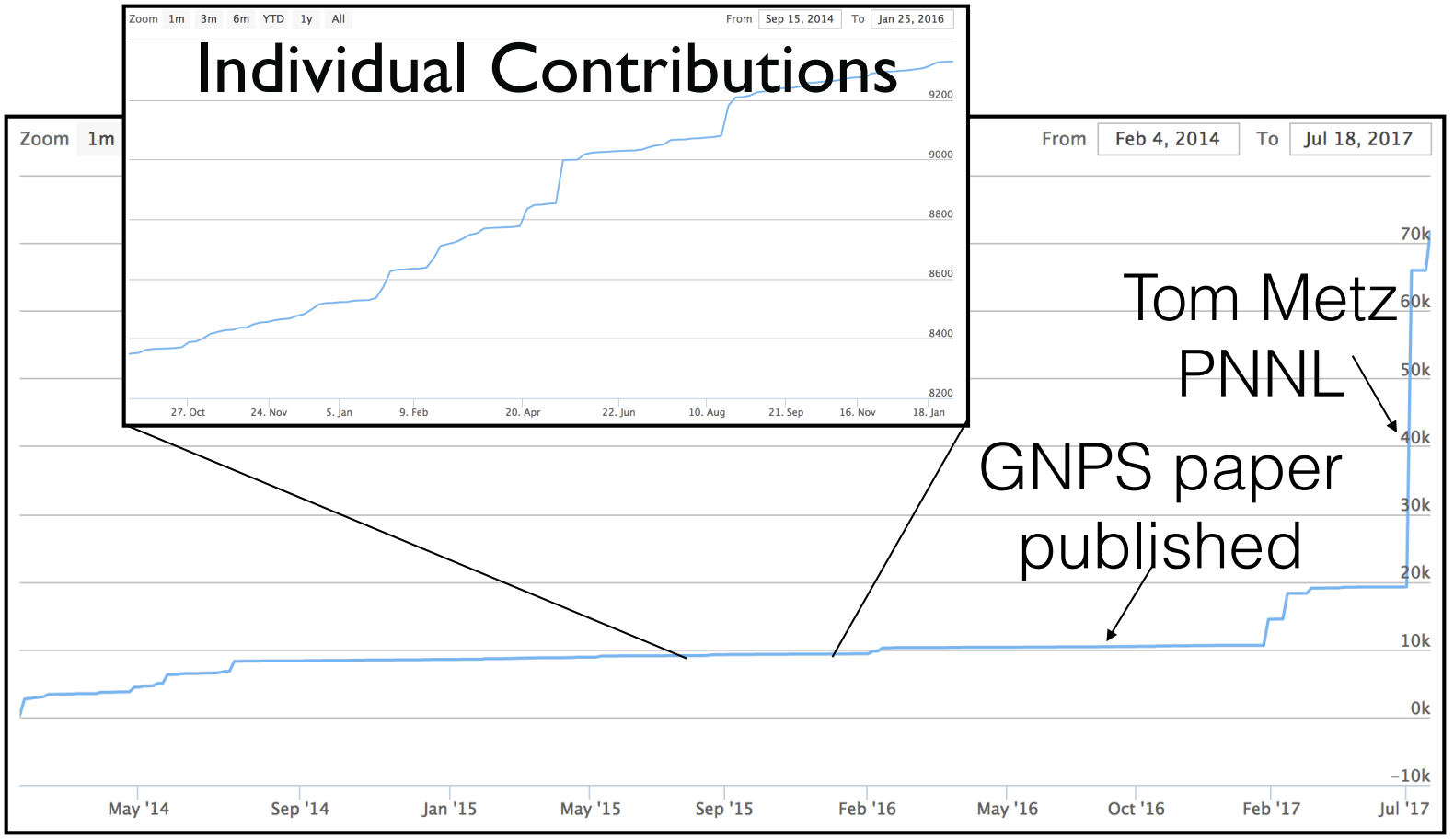
Dereplication and comparative dereplication (also known as finding known unknowns)





Nature Biotech 2016

Growth of knowledge in GNPS (contributed by the GNPS community of >40,000 people)



Data analysis reproducibility

Back to main page

ProteoSAFe Workflow Tasks Hits 1 ~ 30 out of 329 Go to Go

Select columns

	Filter Description	User	Workflow	Site	Status	Create Time	Elapsed Time
1	cheese ID=a9bef02315e546949f3c59181b4eaa33	pdorrestein	METABOLOMICS-SNETS	GNPS	DONE	Nov. 2, 2017, 4:48 PM	9 day 23:38:00
2	fungi vs food ID=8254edfe3528434a93dcda0419f7f275	pdorrestein	METABOLOMICS-SNETS	GNPS	DONE	Jul. 28, 2017, 8:04 AM	1 day 04:17:55
3	food vs skin ID=19d3b6b5f3864fb3b0976ee400379e12	pdorrestein	METABOLOMICS-SNETS	GNPS	DONE	Jul. 28, 2017, 8:02 AM	0 day 22:58:11
4	lululemon 5 ion cos 0.7 min cluster size 3 parent mass 0.1 ans MS/mS match 0.1 ID=8f1b79bcca4a4478972f22bf4f8491bf	pdorrestein	METABOLOMICS-SNETS	GNPS	DONE	Jul. 26, 2017, 8:54 PM	1 day 18:55:49
5	lululemon dereplication ID=69422dc3ac6c453188ff395caeff21b8	pdorrestein	MOLECULAR-LIBRARYSEARCH	GNPS	DONE	Jul. 26, 2017, 8:47 AM	0 day 03:55:49
6	lululemon dereplication ID=4b2b0712650f4807af6e37450b905250	pdorrestein	MOLECULAR-LIBRARYSEARCH	GNPS	FAILED	Jul. 26, 2017, 8:46 AM	0 day 00:03:53
7	ID=958b362704ee4bfeb63d3c0a71c8906e	pdorrestein	UPDATE-SINGLE-ANNOTATED-BRONZE	GNPS	DONE	Jul. 24, 2017, 5:59 PM	0 day 00:01:21

Data analysis reproducibility

METABOLOMICS-SNETS

DONE
[Clone] [Restart][Delete]
[View All Library Hits | View All Clusters With IDs | View All Compounds]

Methods and Citation for Manuscripts
[Networking Parameters and Written Network Description]

Experimental Views
[Reanalyze Cluster Spectra | View Raw Spectra | Topology Signatures | Topology Signatures Histogram]

Auxiliary Views
[View Network, Node Centric | View Network Pairs | Networking Statistics | View Compounds and File Occurrence]

Advanced Views - Networking Graphs
[Nodes, MZ Histogram | Edges, MZ Delta Histogram | Edges, Score vs MZ Delta Plot | Library Search, PPM Error Histogram]

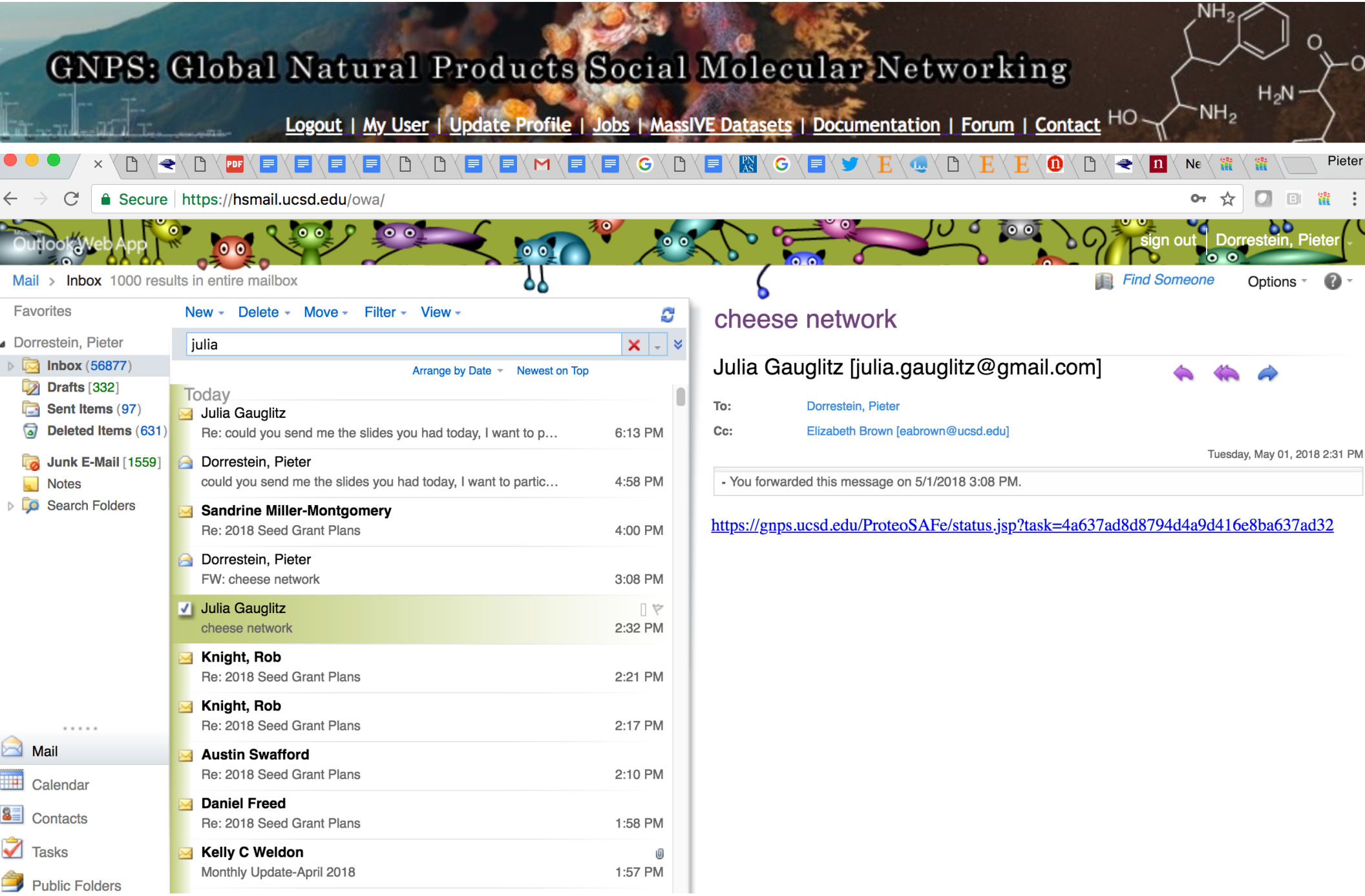
Advanced Views - Third Party Visualization
[View Emporer PCoA Plot in GNPS | View ili in GNPS]

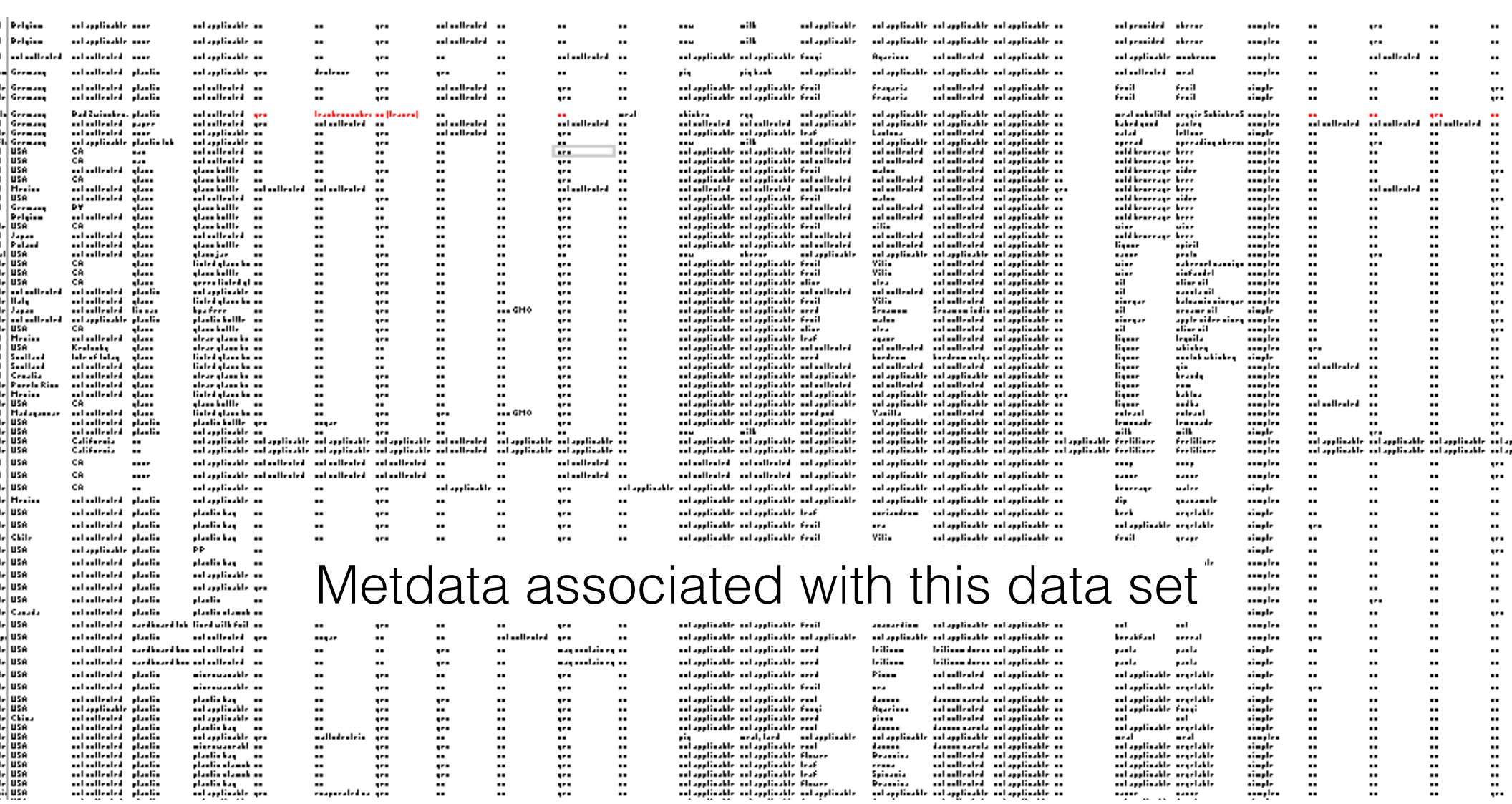
Community Matches
[Dataset Matches]

Network Visualizations
[View Spectral Families (In Browser Network Visualizer) | Network Summarizing Graphs]

Export
[Download Clustered Spectra as MGF | Download Cytoscape Data | Download Bucket Table | Download BioM For Qiime/Qiita | Download Metadata For Qiime | Make Public Dataset | Download ili Data]

Data analysis reproducibility and role of digitizing metadata

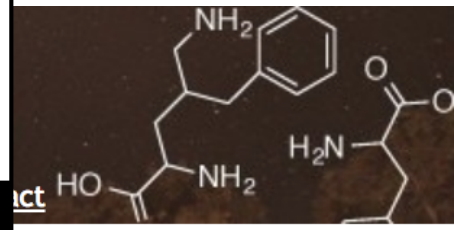




Metdata associated with this data set

Not yet done but IMO we need to capture metadata at the time of deciding to run the experiment, not add afterwards.

https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=4a637ad8d8794d4a9d416e8ba637ad32



METABOLOMICS-SNETS

DONE
[Clone]
[View All Library Hits | View

Methods and Citation for Ma
[Networking Parameters and

Experimental Views
[Reanalyze Cluster Spectra |

Auxiliary Views
[View Network, Node Centri

Advanced Views - Networkin
[Nodes, MZ Histogram | Edge

Advanced Views - Third Part
[View Emporer PCoA Plot in

Community Matches
[Dataset Matches]

Network Visualizations
[View Spectral Families (In B

Export
[Download Clustered Spectra
Metadata For Qiime | Make P

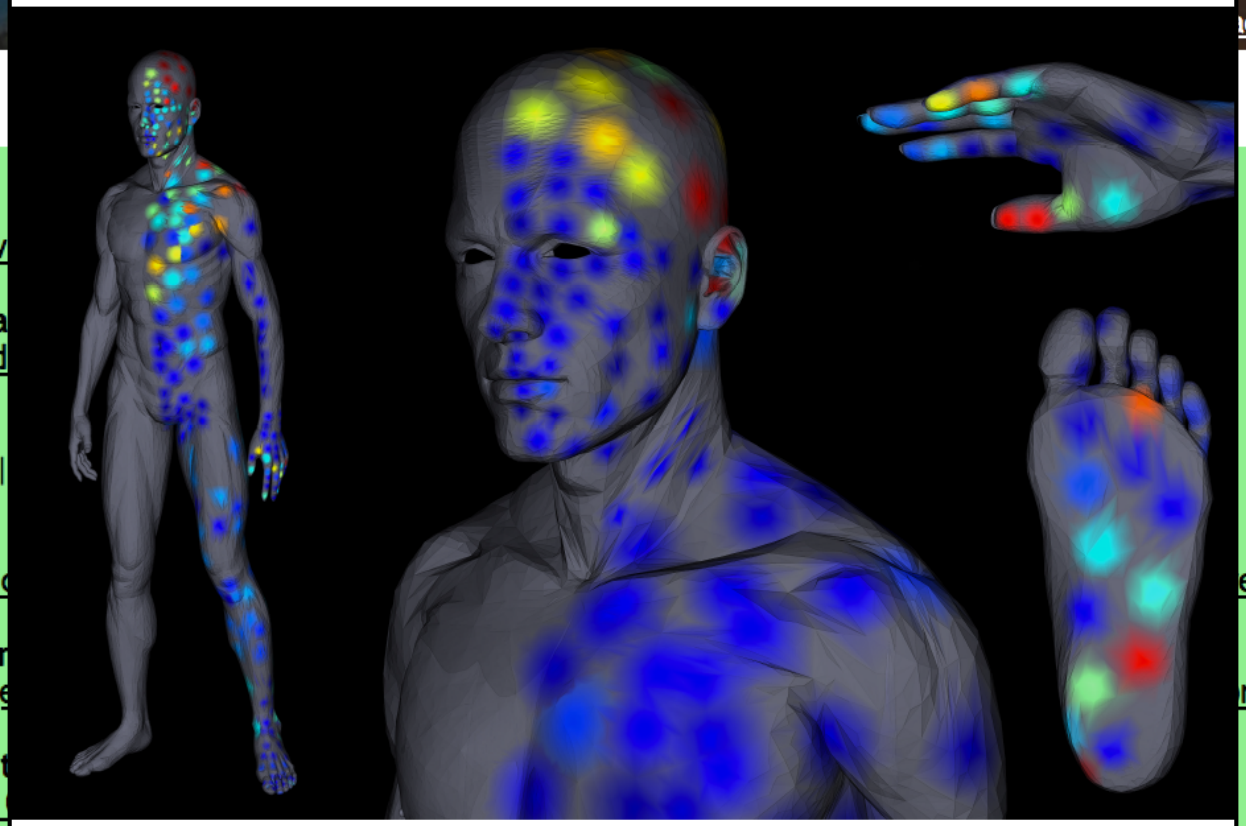
[Restart][Delete]

ence]

r Histogram]

for Qiime/Qiita | Download

nature protocols
Recipes for Researchers
Vol. 12 Issue 12
www.natureprotocols.com



- **Mouse cerebellar window**
- **HPLC detection of monoamines and their cofactors**
- **Probing virus–host interactions with AFM**
- **Massively parallel cancer-mutation stratification**
- **Chemical synthesis of histones**

SPRINGER NATURE

Examples of data reuse from my own lab.

Example Data reuse 1.

Bioactivity Based Molecular Networking for the Discovery of Active Natural Products

Louis-Félix Nothias,[1,2,3] Mélissa Nothias-Esposito,[3,4] Ricardo da Silva,[1,2] Ivan Protsyuk,[5] Pieter Leyssen,[6] David Touboul,[3] Julien Paolini,[4] Theodore Alexandrov,[5] Marc Litaudon,[3] & Pieter C. Dorrestein[1,2,**]

ABSTRACT

It is a common problem in natural product therapeutic lead discovery programs that despite good bioassay results in the initial extract, the active compound(s) could not be isolated during subsequent bioassay-guided purification. Herein, we present the concept of bioactive molecular networking to find the candidate active molecules directly from fractionated bioactive extracts. Mass spectrometry based molecular networking, bioactivity score prediction to create bioactive molecular networks enables the discovery and the targeted isolation of lead molecules. We highlight the feasibility of this approach

In press

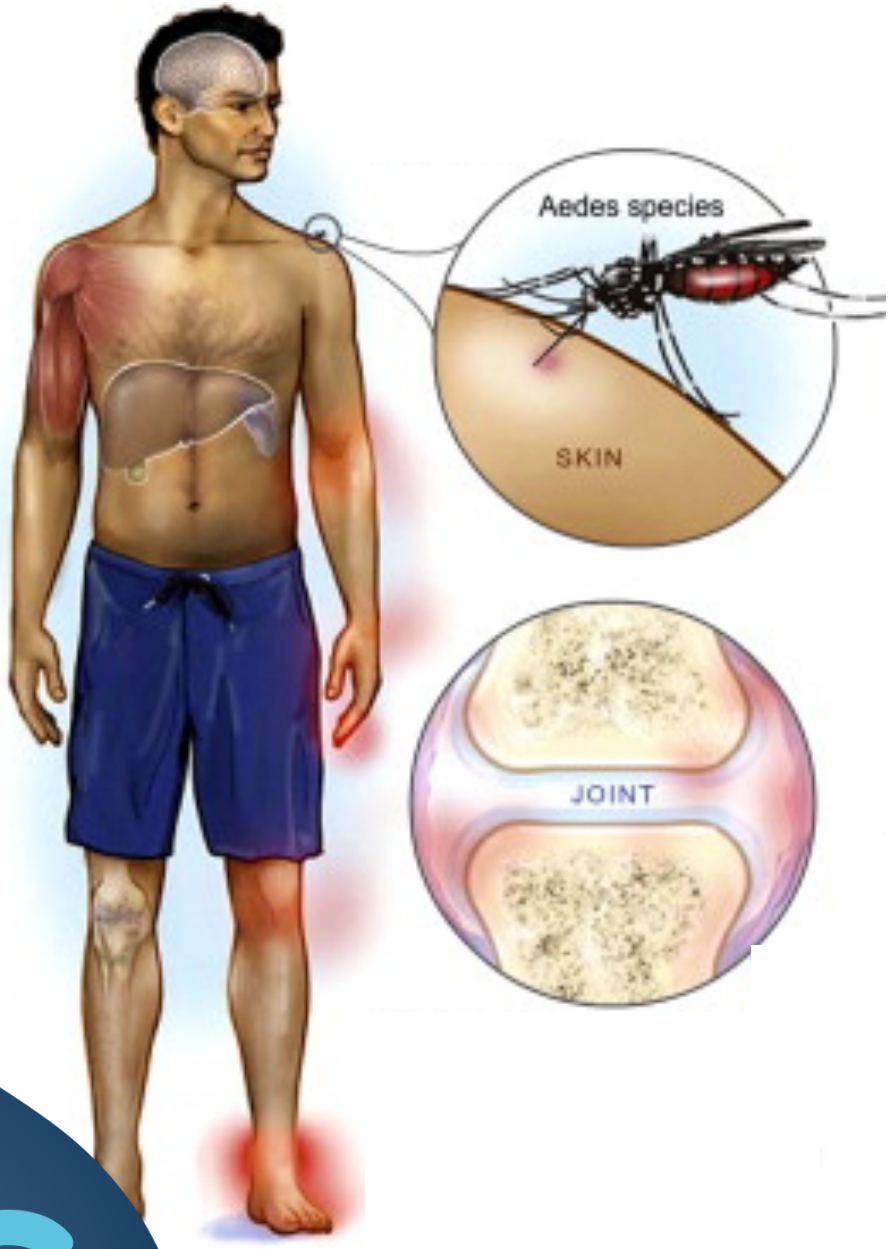
Why study Euphorbia dendroides for anti-viral activities?

Why study Euphorbia dendroides for anti-viral activities?

://ceanothus.blogspot.com.br/2008/10/euphoric-about-euphorbia.html

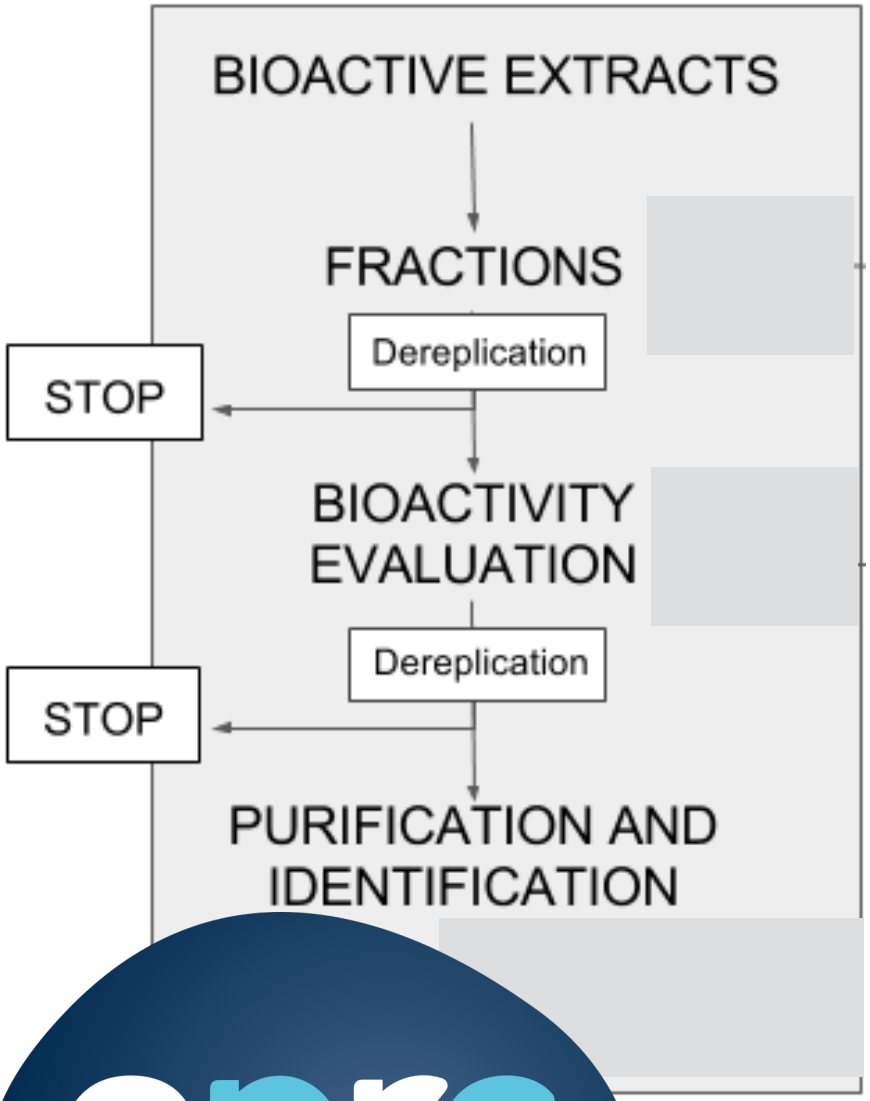
healthline.net

Marc Litaudon and Julien Paolini asked - Does the extract of this plant have anti-viral activity?



Tested extracts against chikungunya virus in monkey cell lines

"Classical" bioassay-guided fractionation



Euphorbia dendroides

↓

Activity against the chikungunya virus

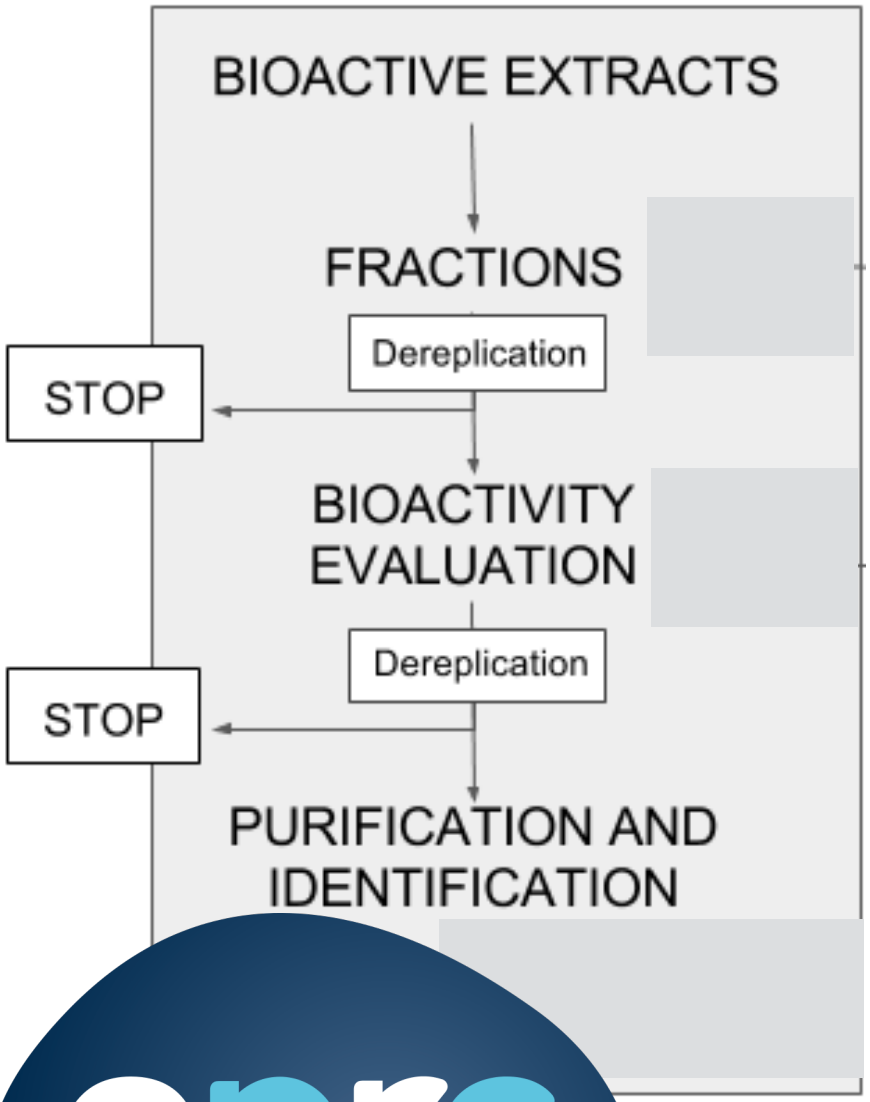
↓

Fractionated into 20 fractions

↓

3 fractions active EC50= 0.268, 0.169 and 1.130 µg/ml

"Classical" bioassay-guided fractionation



Activity against the chikungunya virus

↓

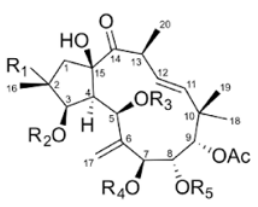
Fractionated into 20 fractions

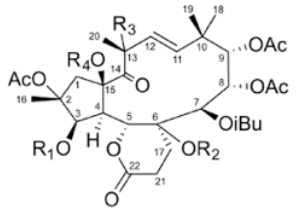
↓

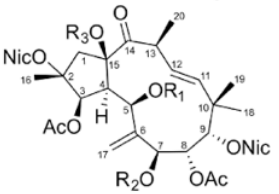
3 fractions active EC50= 0.268, 0.169 and 1.130 µg/ml

↓

***PREVIOUSLY ISOLATED COMPOUNDS* (1-19)**



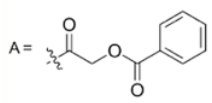




		R_1	R_2	R_3	R_4	R_5
III	1	H	Ac	iBu	Bz	Ac
	2	OH	Ac	iBu	Bz	Ac
	3	OH	Bz	H	Bz	Ac
	4	OAc	H	iBu	Bz	Ac
I	5	OAc	iBu	iBu	Bz	Ac
	6	OAc	Ac	iBu	Bz	Ac
	7	OAc	A	Ac	iBu	Ac
	8	OAc	Bz	iBu	iBu	Ac
III	9	OAc	H	MeBu	Bz	Ac
	10	OAc	H	Bz	Bz	Ac
	11	OAc	H	Bz	Bz	H

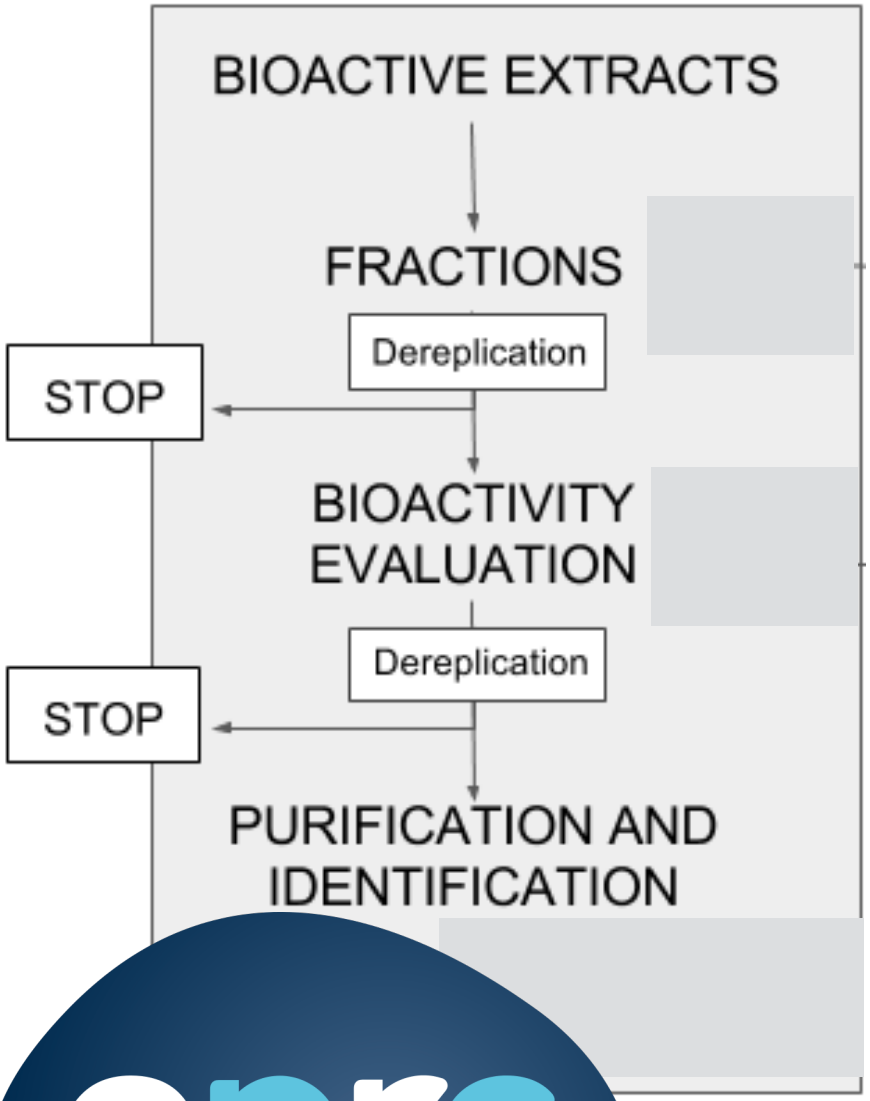
		R_1	R_2	R_3	R_4
II	12	Ac	Ac	OH	H
	13	Ac	Ac	OH	Ac
	14	Ac	iBu	H	H
	15	H	Ac	H	Ac

		R_1	R_2	R_3
IV	16	Ac	Bz	H
	17	Ac	Bz	Ac
	18	Ac	iBu	Ac
	19	iBu	Bz	H



Compounds **1-15** did not show significant bioactivity against CHIKV replication [20]

"Classical" bioassay-guided fractionation

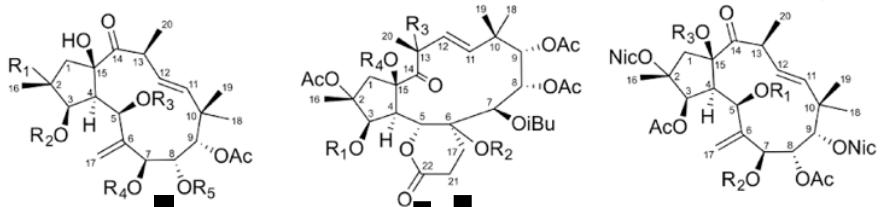


Activity against the chikungunya virus

Fractionated into 20 fractions

3 fractions active EC50= 0.268, 0.169 and 1.130 µg/ml

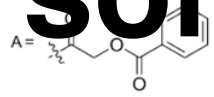
***PREVIOUSLY ISOLATED COMPOUNDS* (1-19)**



		R_1	R_2	R_3
III	1	H	Ac	iBu
	2	OH	Ac	iBu
	3	OH	Bz	H
	4	OAc	H	iBu
	5	iBu	iBu	Bz
I	6	Ac	Ac	Bz
	7	OAc	A	Ac
	8	Bz	Bz	iBu
III	9	OAc	H	eBu
	10	OAc	H	Bz
	11	OAc	H	Bz

		R_1	R_2	R_3	R_4
II	13	Ac	Ac	OH	Ac
	14	Ac	iBu	H	H
	15	H	Ac	H	Ac

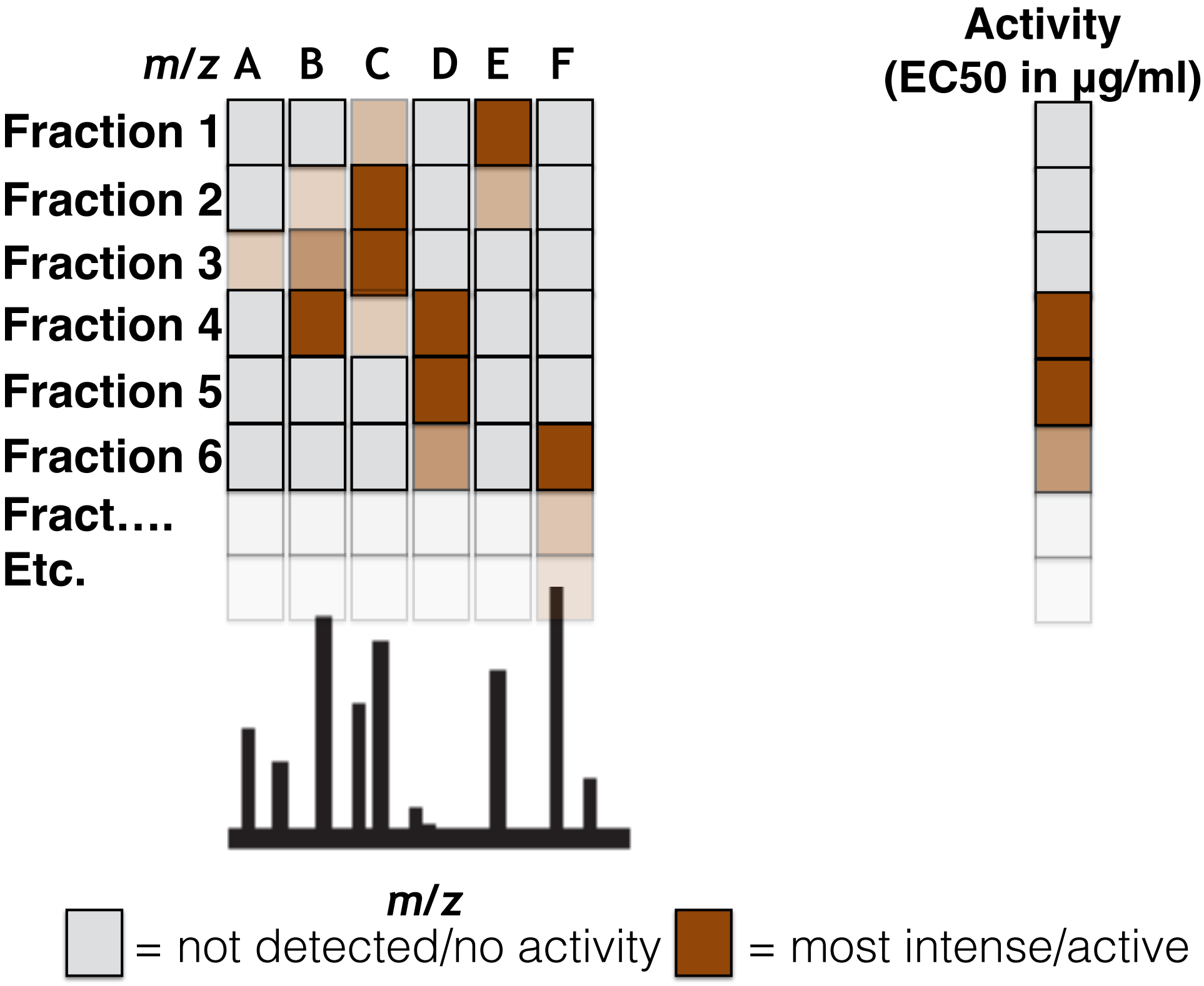
		R_1	R_2	R_3
IV	16	Ac	Bz	H
	17	Ac	Bz	Ac
	18	Ac	iBu	Ac
	19	iBu	Bz	H

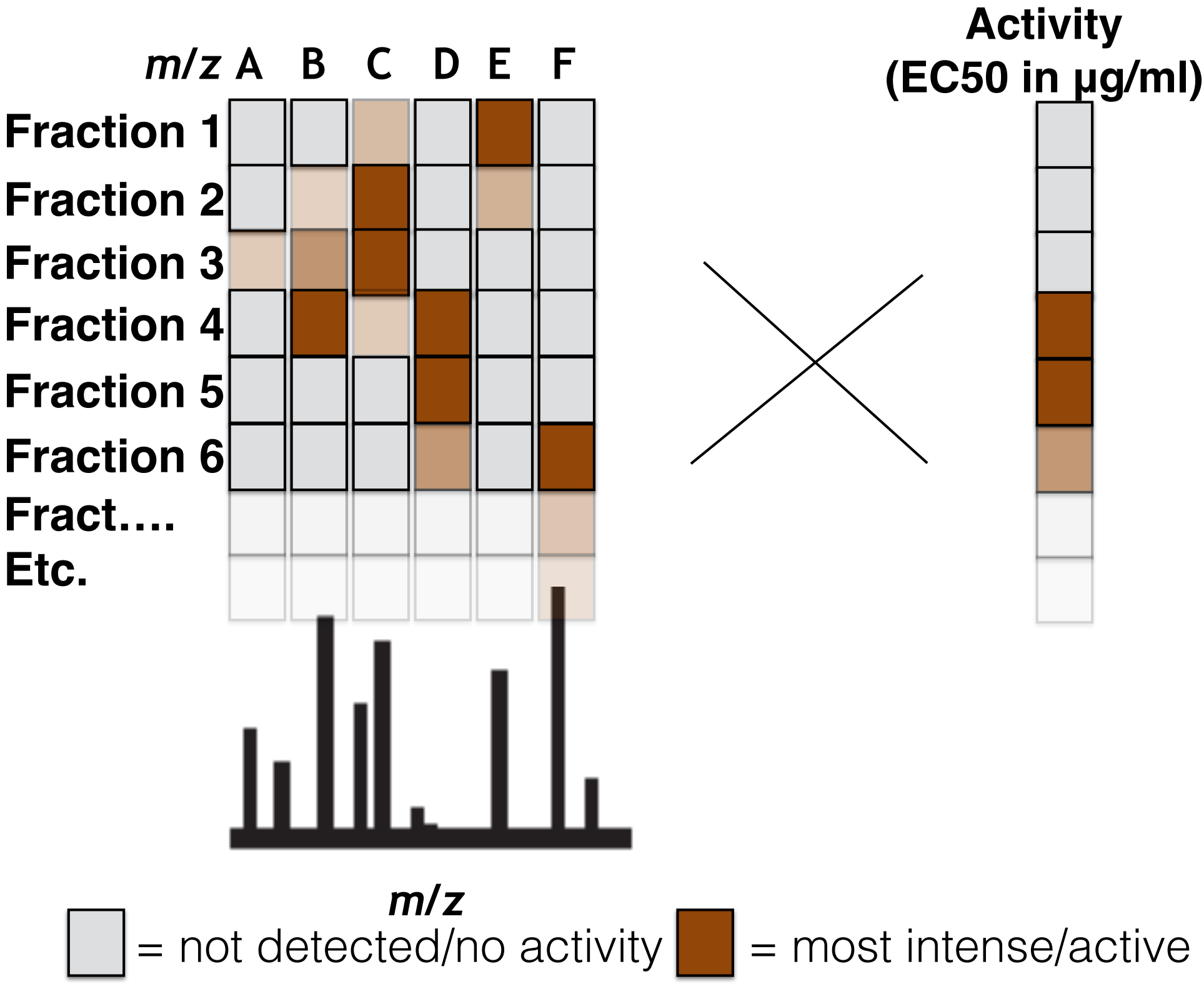


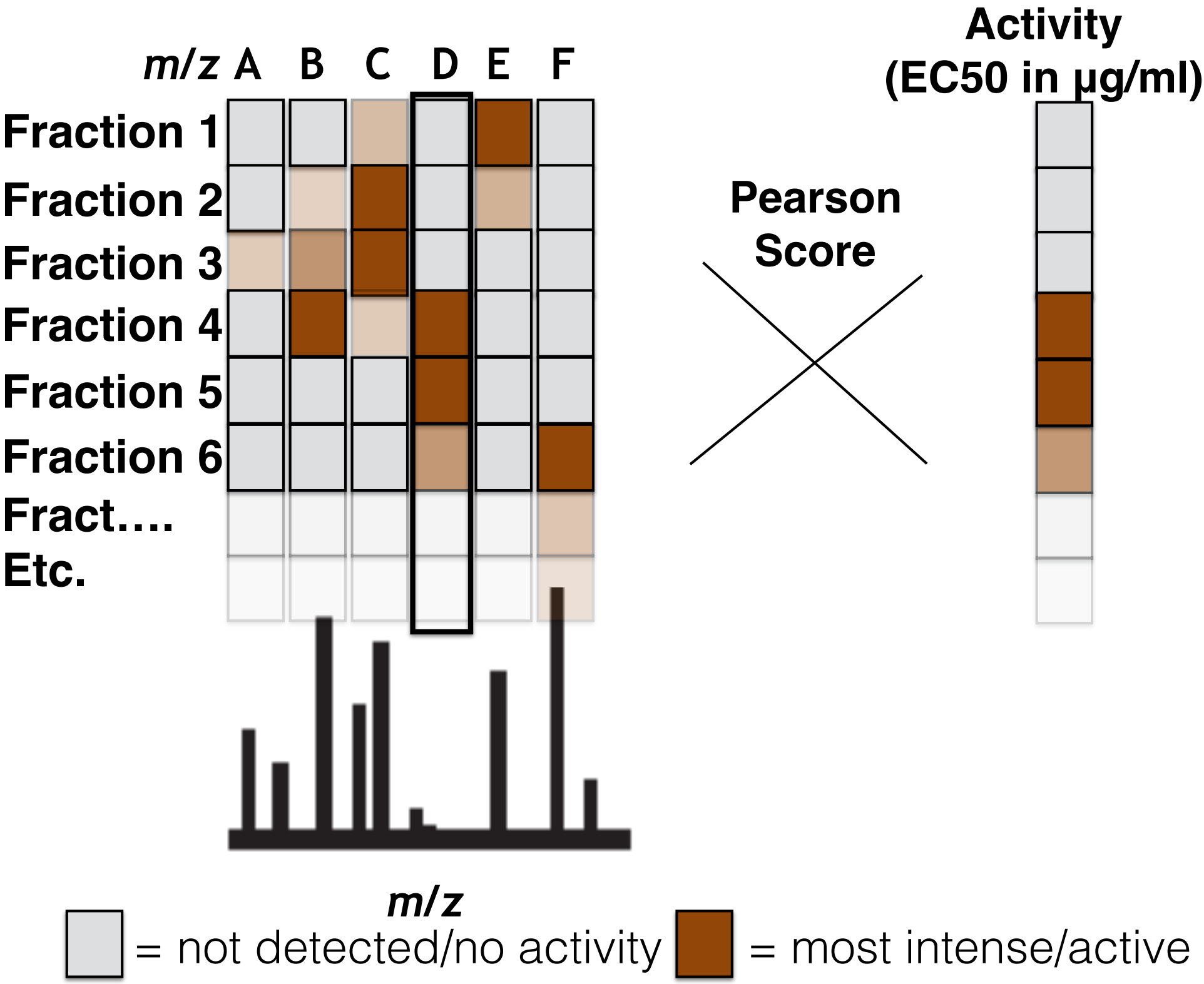
Compounds 1-19 did not show significant bioactivity against CHIKV replication [20]

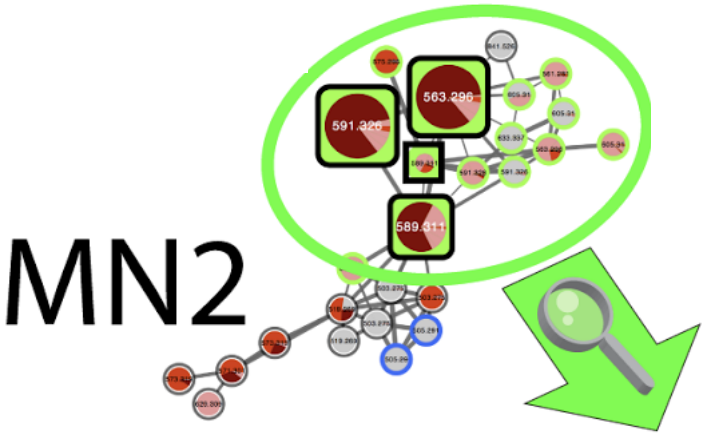
Inactive-

15 person years


m/z A B C D E F
Fraction 1
Fraction 2
Fraction 3
Fraction 4
Fraction 5
Fraction 6
Fract….
Etc.
Activity
(EC50 in µg/ml)
m/z
= not detected/no activity
= most intense/active

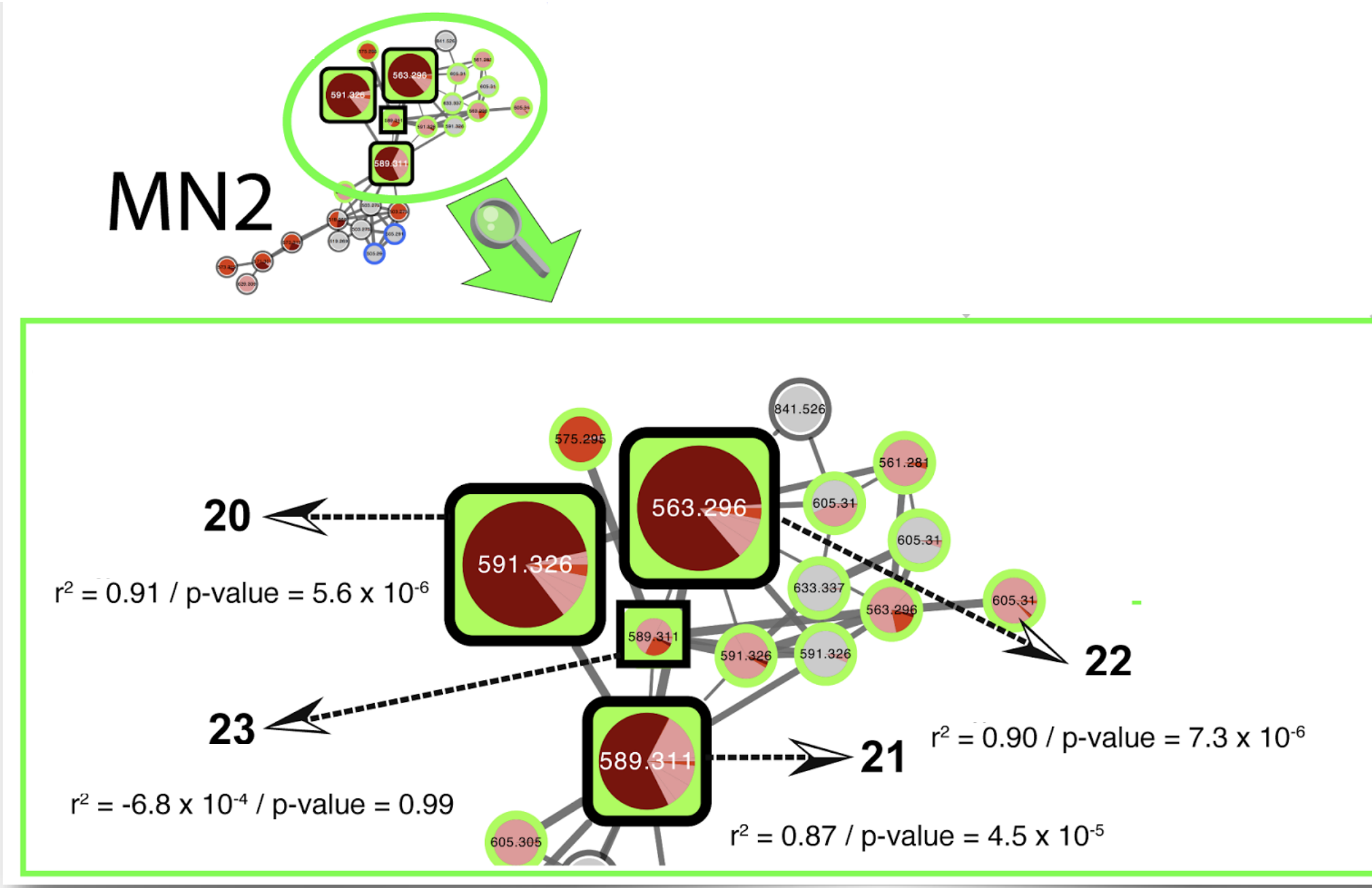

m/z A B C D E F
Fraction 1
Fraction 2
Fraction 3
Fraction 4
Fraction 5
Fraction 6
Fract….
Etc.
Activity
(EC50 in µg/ml)
m/z
= not detected/no activity
= most intense/active


m/z A B C D E F
Fraction 1
Fraction 2
Fraction 3
Fraction 4
Fraction 5
Fraction 6
Fract….
Etc.
m/z
Activity
(EC50 in µg/ml)
Pearson
Score
= not detected/no activity
= most intense/active



Correlation does not mean causation!!

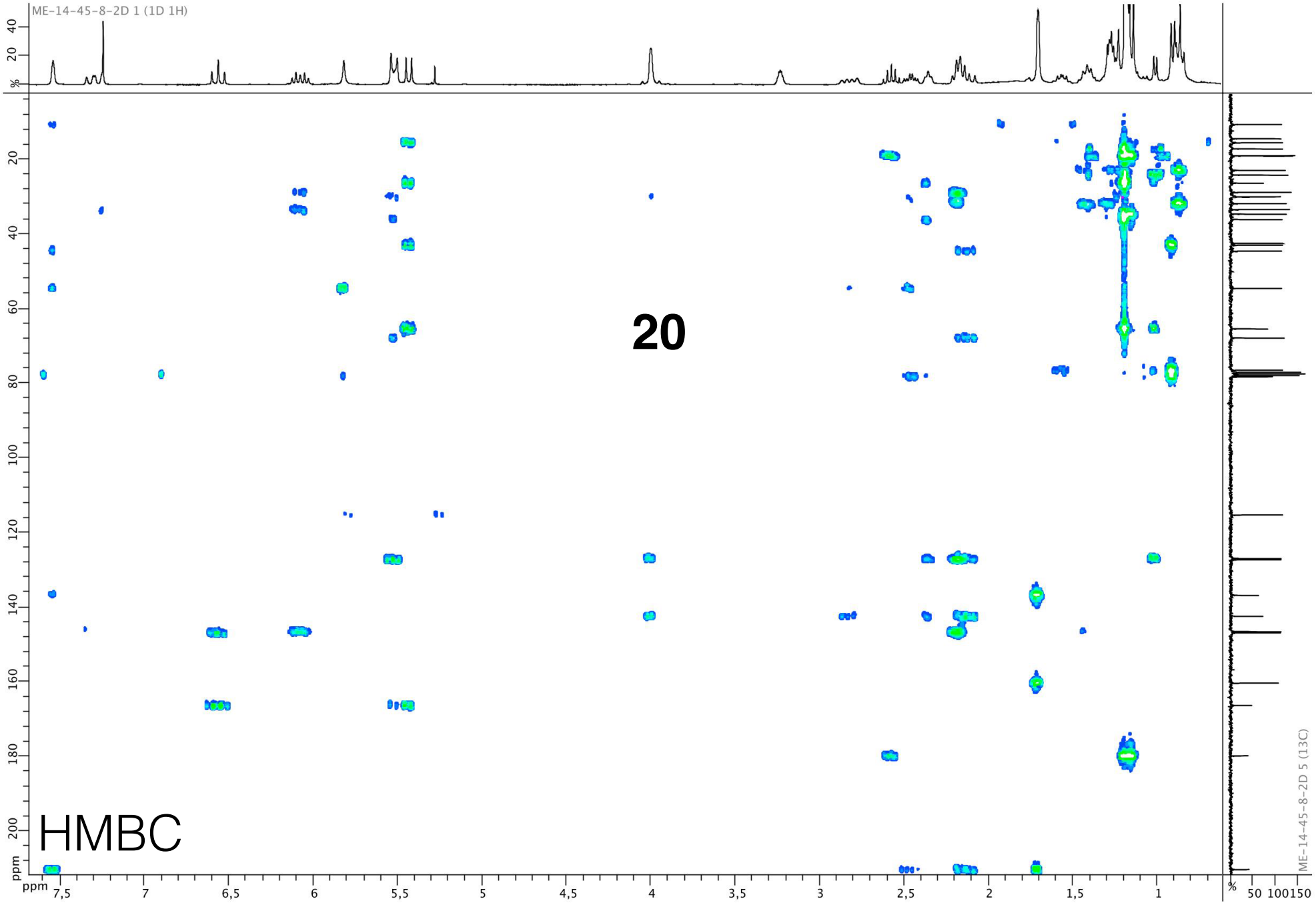
We have potential NP candidates responsible for activity.



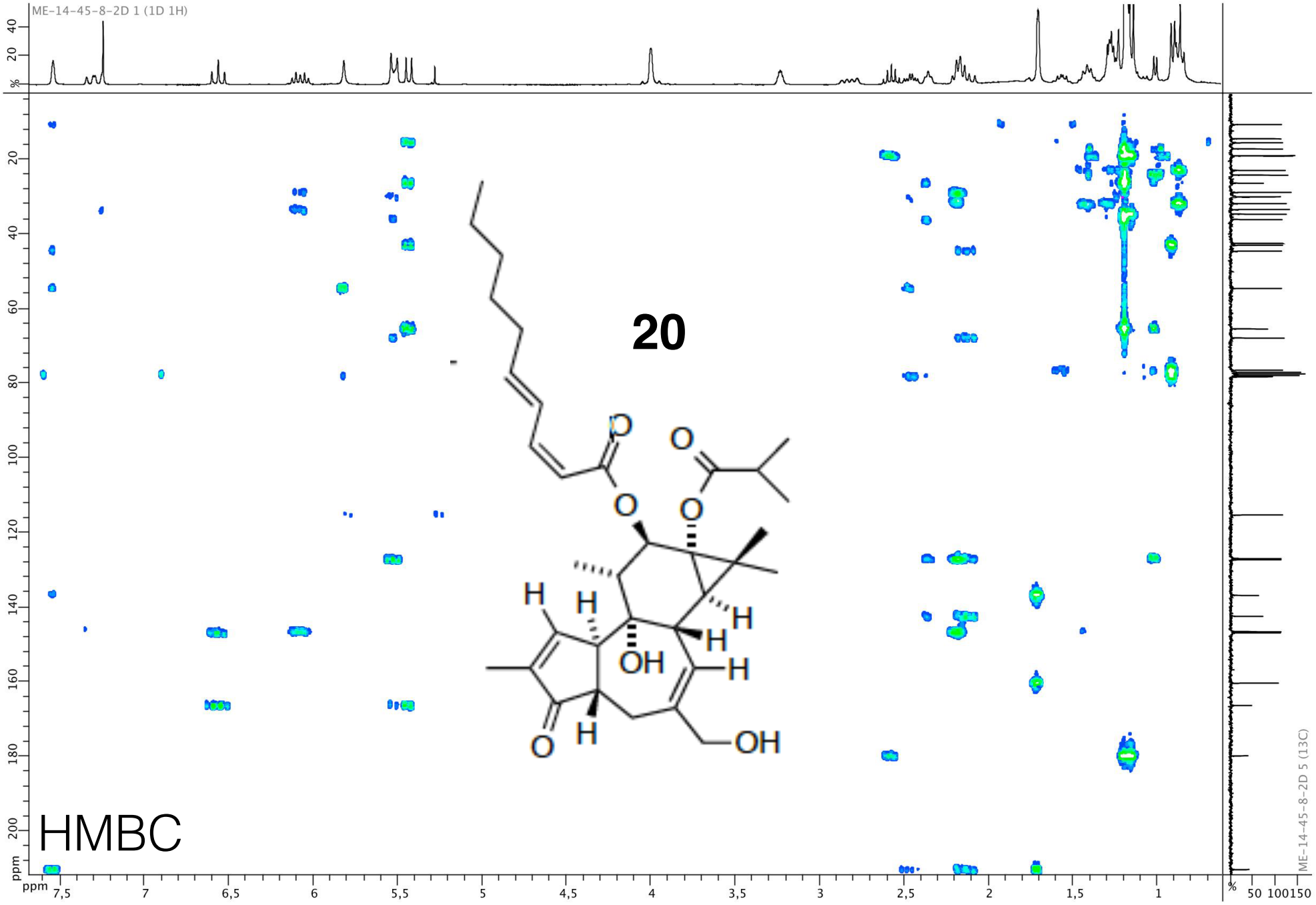
Correlation does not mean causation!!

We have potential NP candidates responsible for activity.

MS based targeted isolation - NMR

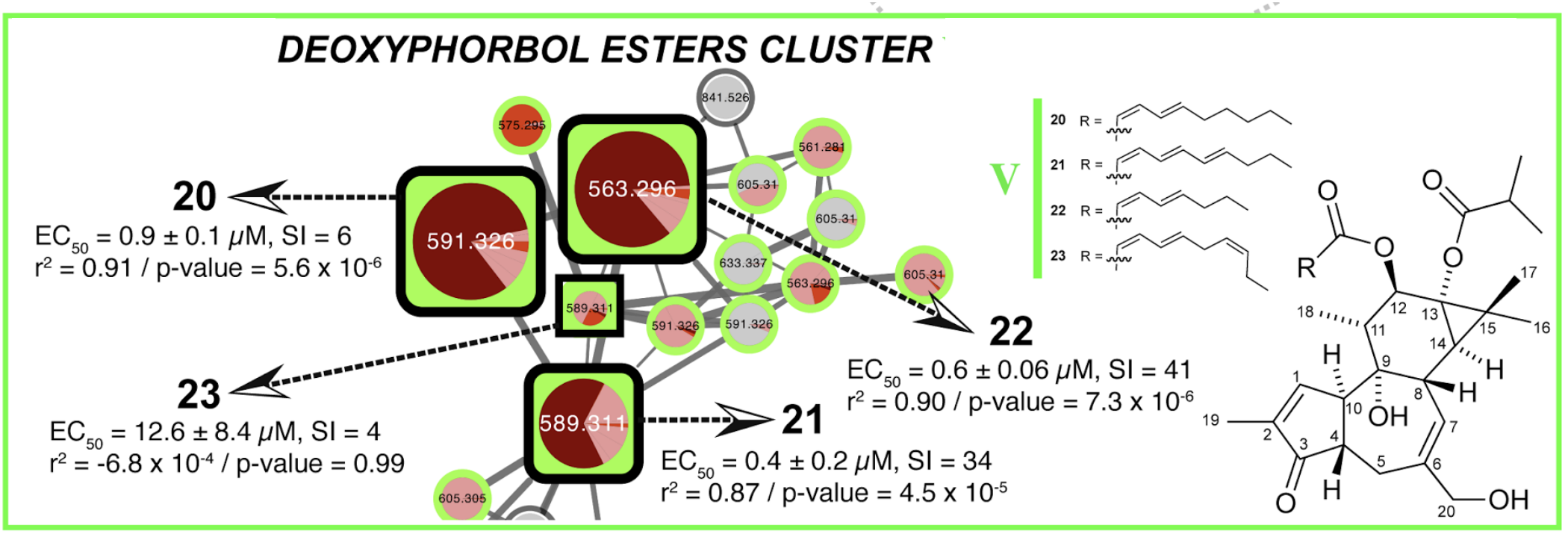


MS based targeted isolation - NMR

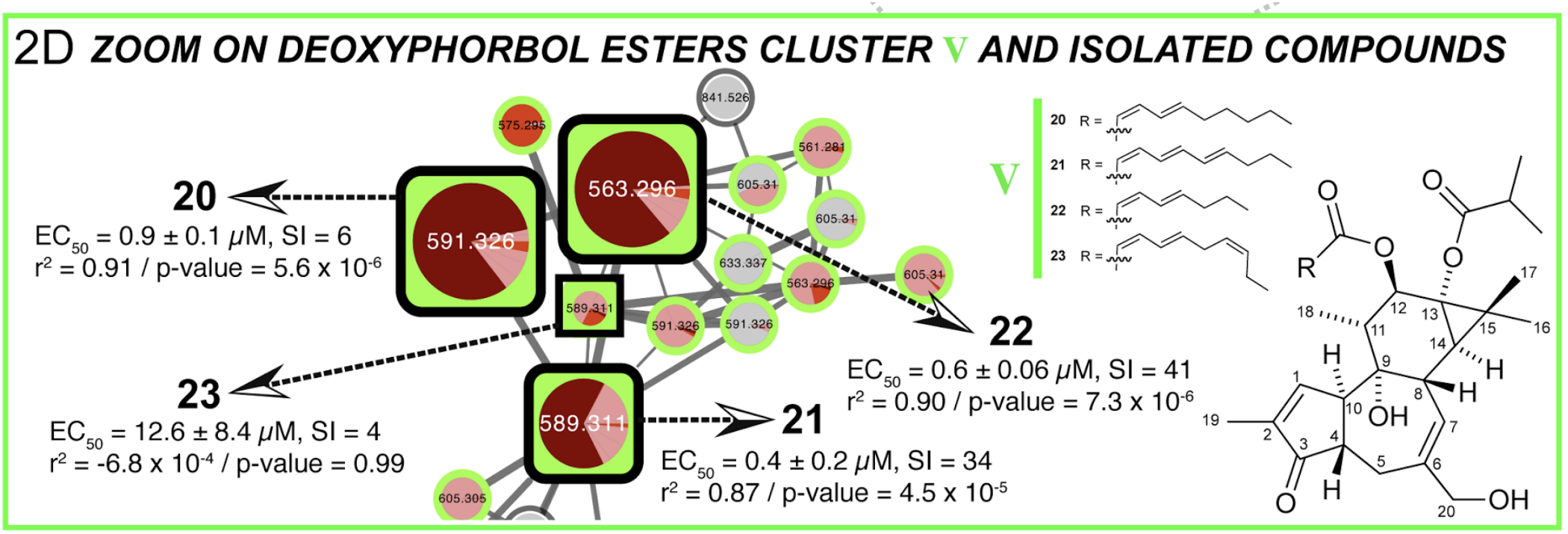


New family of deoxyphorbol esters

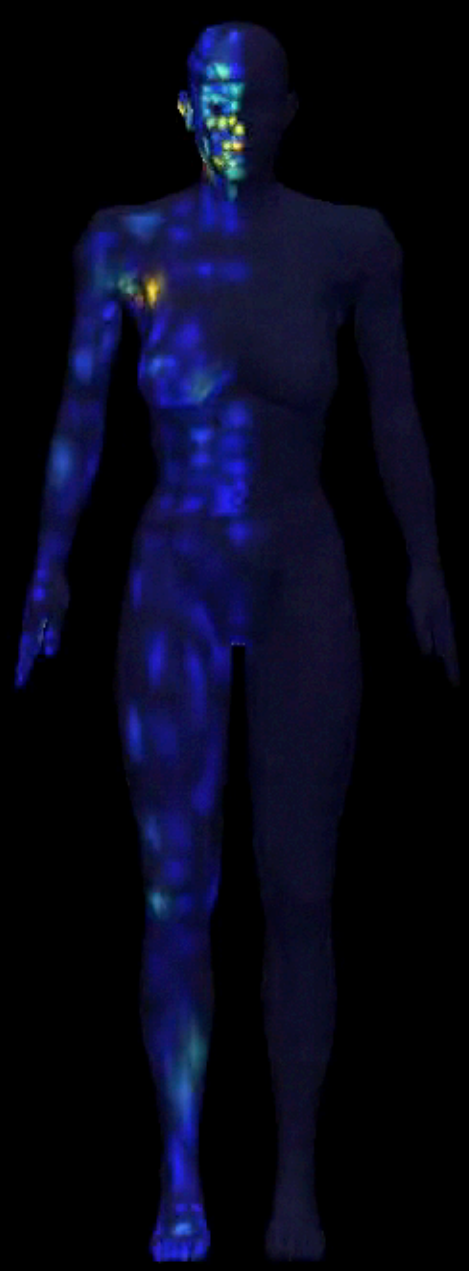
Structures and anti-chikungunya virus activity



This discovery would not have been made if data and metadata was not public



Example Data reuse 2.

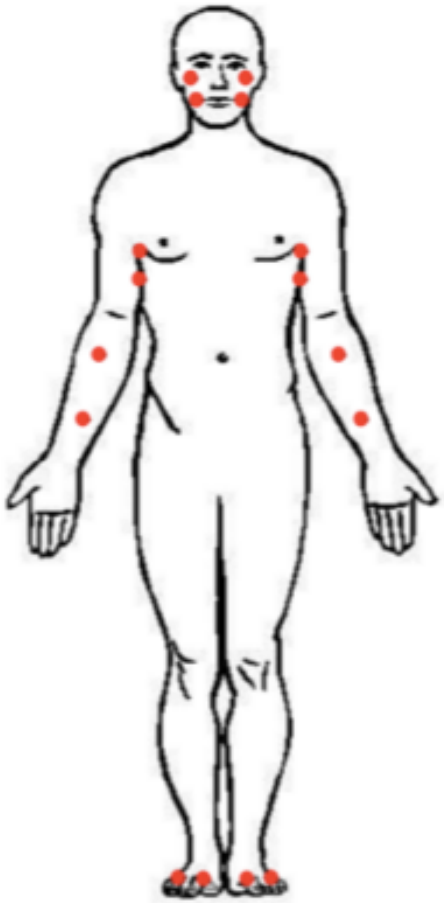


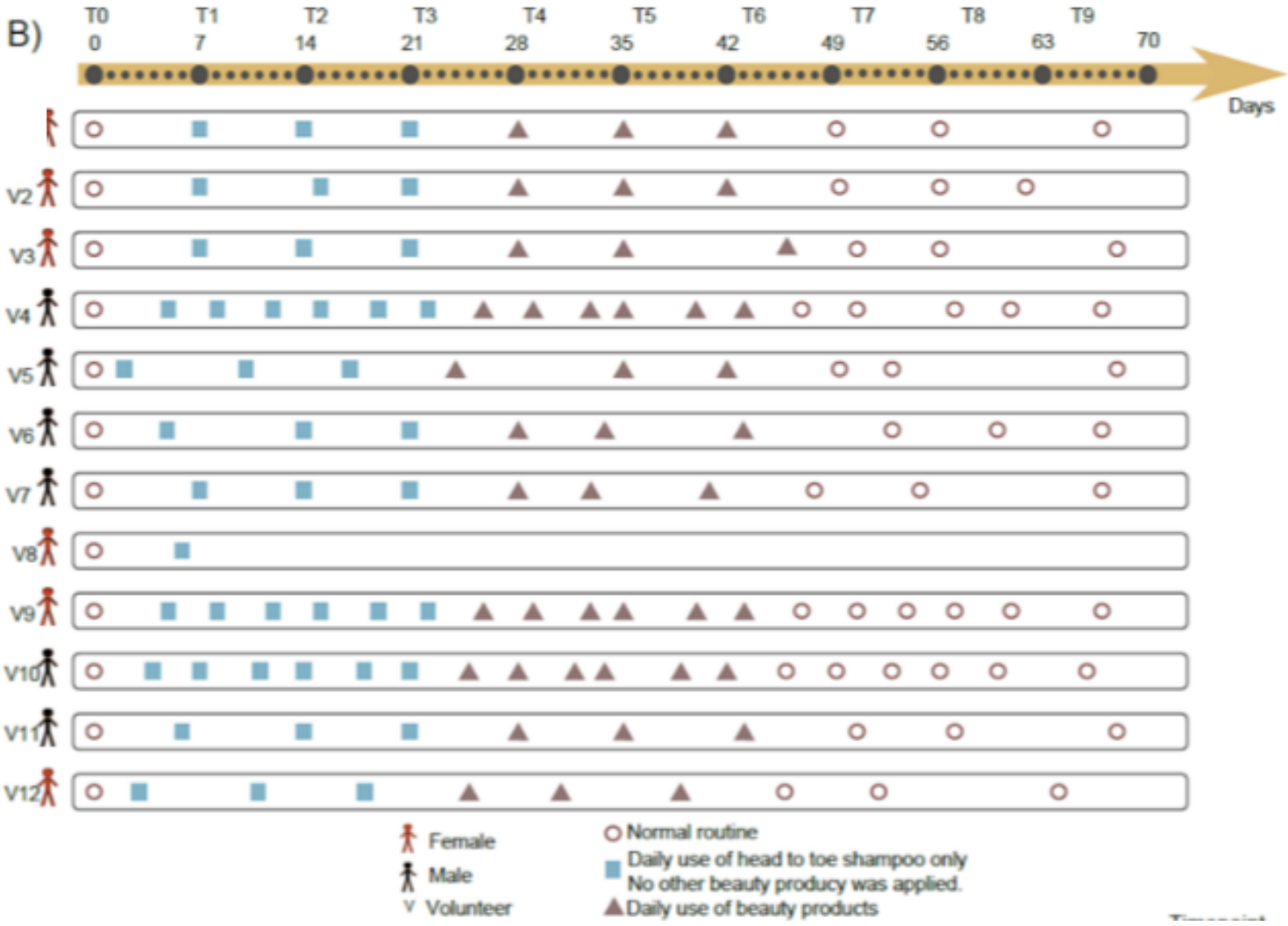
m/z 50

m/*z* value

Red = high intensity
Orange
Yellow
Green
Light Blue
Blue = low intensity

Does personal care impact the skin?

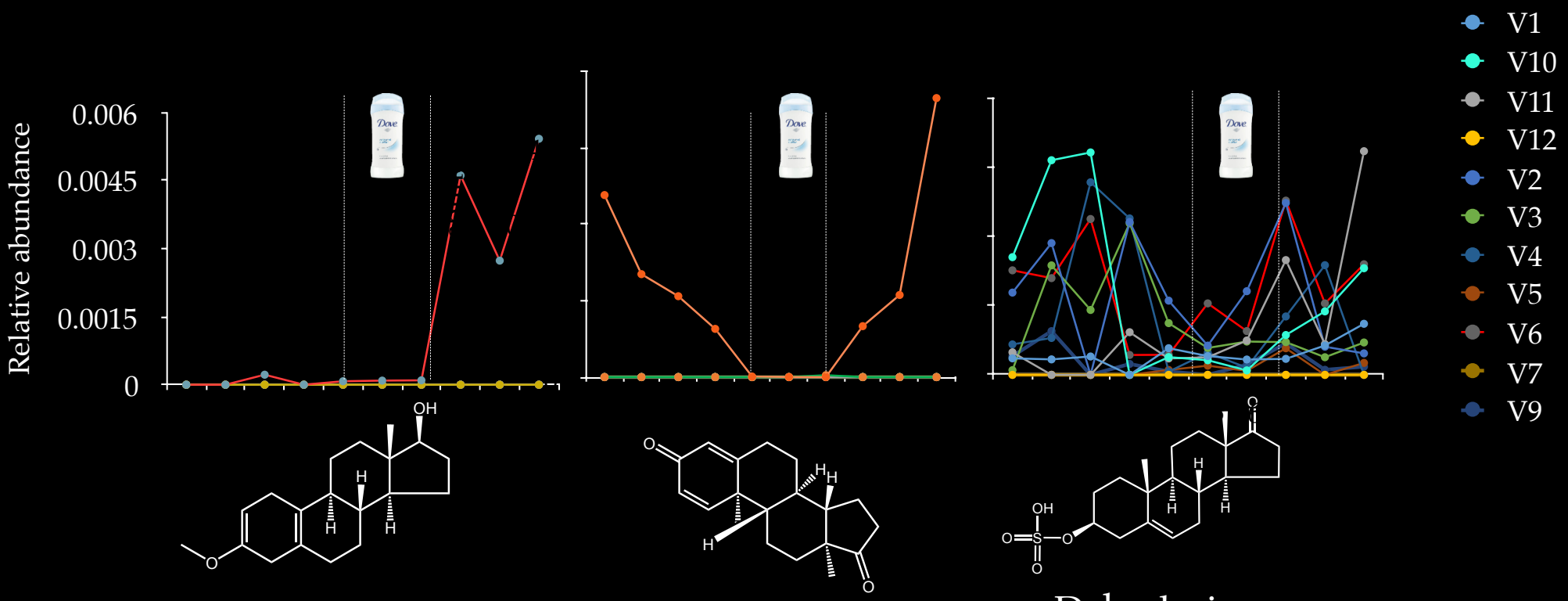




Amina Bouslimani, Louie Felix, Tal Lazutto-Knaan and Kathleen Dorrestein

Steroid molecular family in **Armpits**

Androsterone-response is person specific.



3-Methoxyestra-2,5(10)-dien-17beta-ol

1-Dehydroandrostenedione

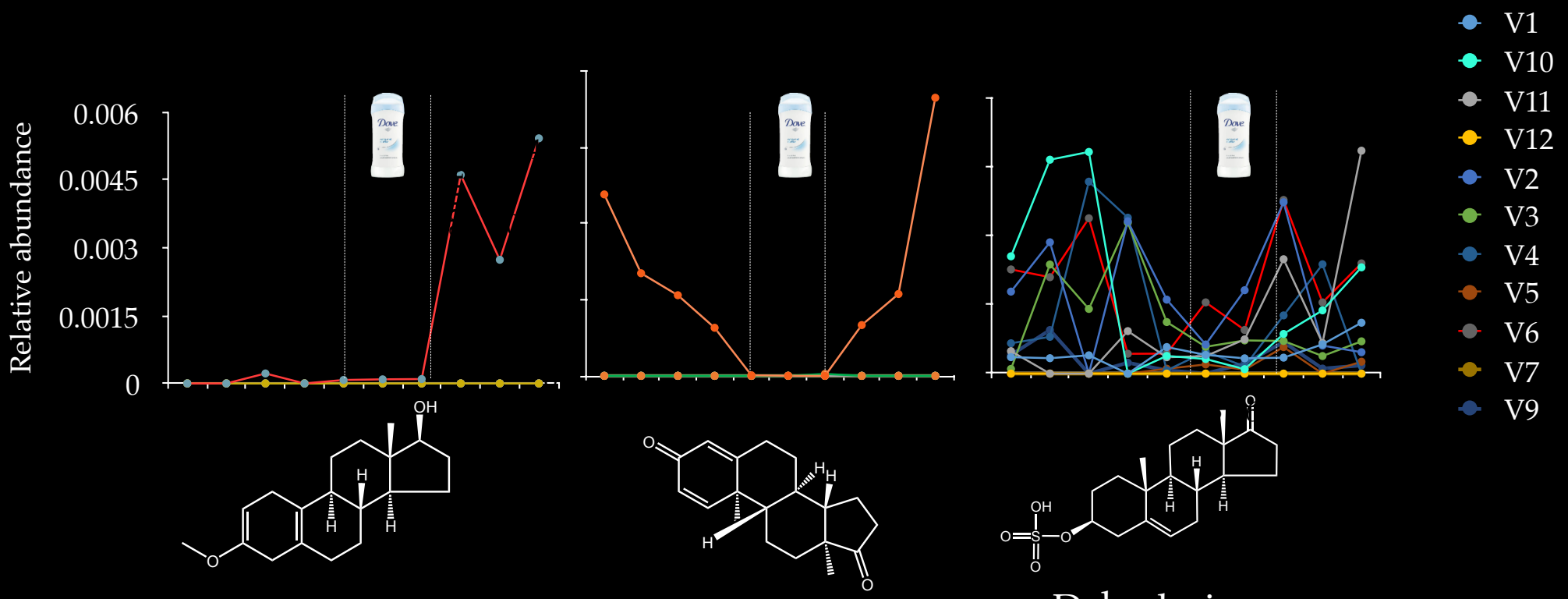
Dehydroisoandrosterone sulfate

Influence of Androstenol and Androsterone on the Evalulation of Men of Varying Attractiveness Levels

R. E. Maiworm, W. U. Langthaler

Steroid molecular family in **Armpits**

Had to get DEA approved, background checks etc. to get the standard so we could validate.



3-Methoxyestra-2,5(10)-dien-17beta-ol

1-Dehydroandrostenedione

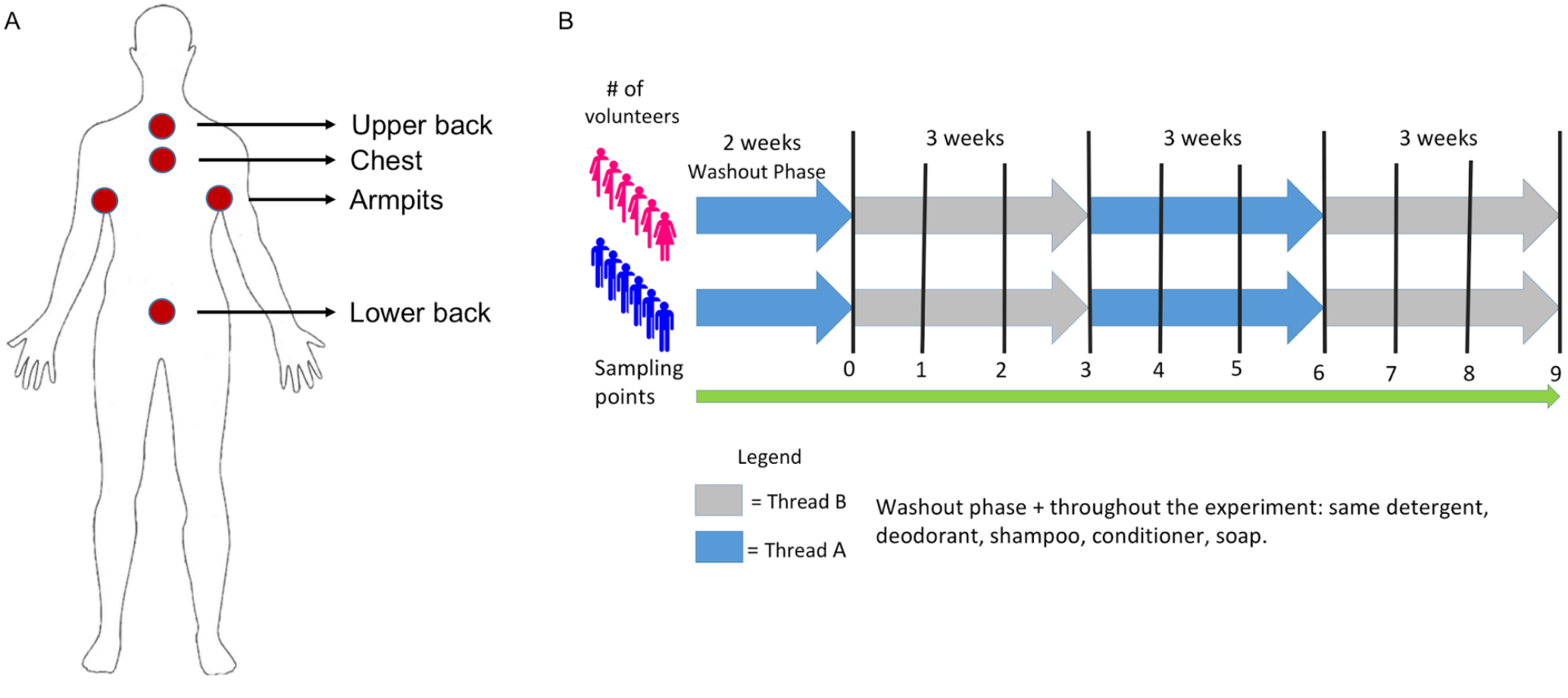
Dehydroisoandrosterone sulfate

Influence of Androstenol and Androsterone on the Evalulation of Men of Varying Attractiveness Levels

R. E. Maiworm, W. U. Langthaler

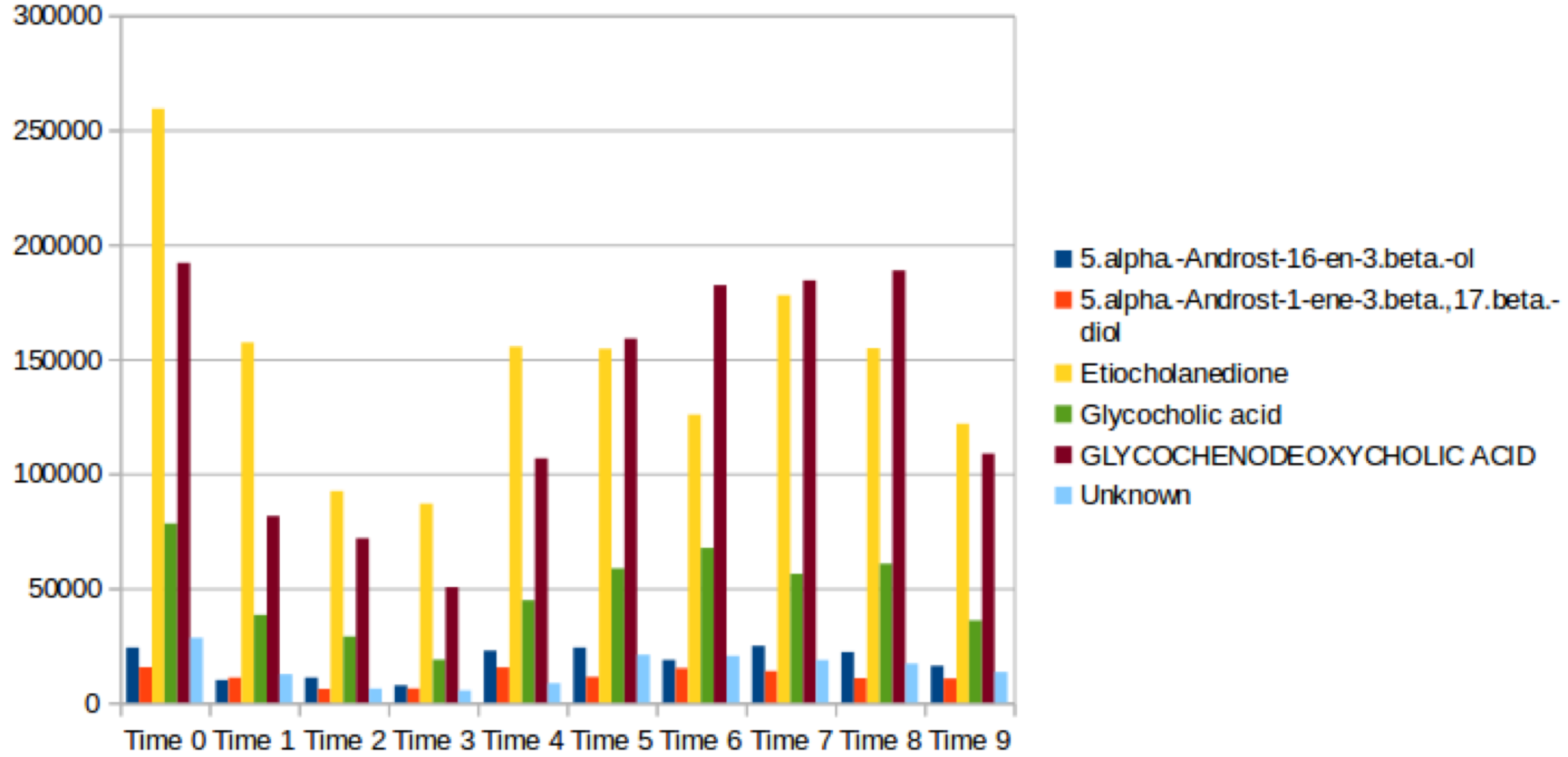
Steroid molecular family in **Armpits**

Had to get DEA approved, background checks etc. to get the standard so we could validate.



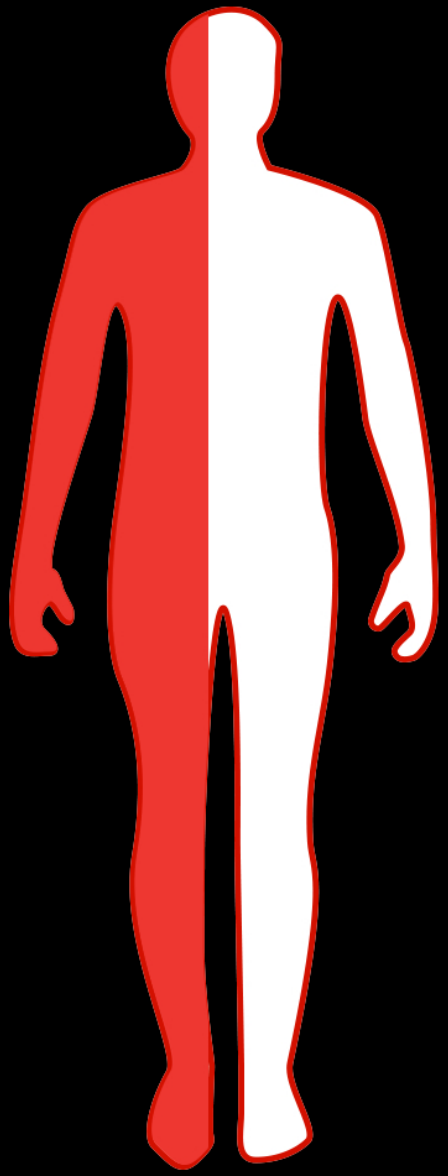
Steroid molecular family in **Armpits**

We could reuse data to validate.



Example Data reuse 3.

30 trillion
human cells



39 trillion
microbial cells

43%

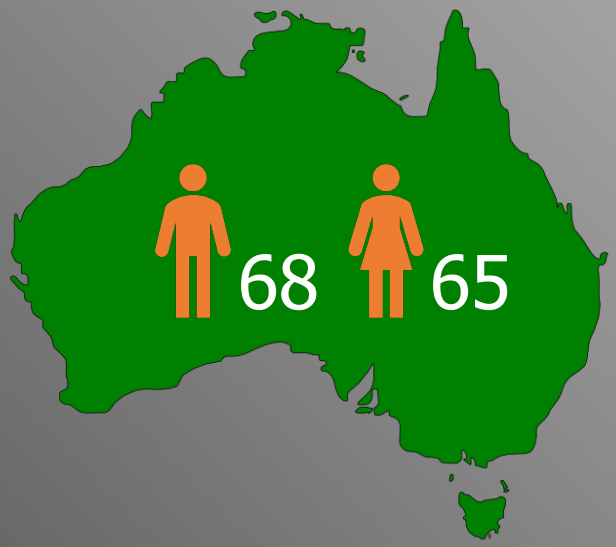
american gut

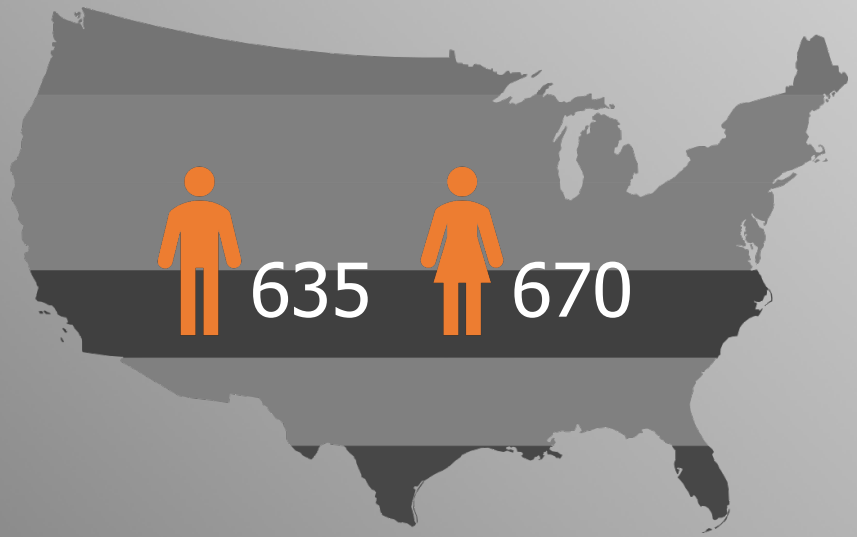
Citizen – scientist

Crowd funded -15,000 sample collection

Paired microbiome & metabolome

Extensive metadata, self-reporting





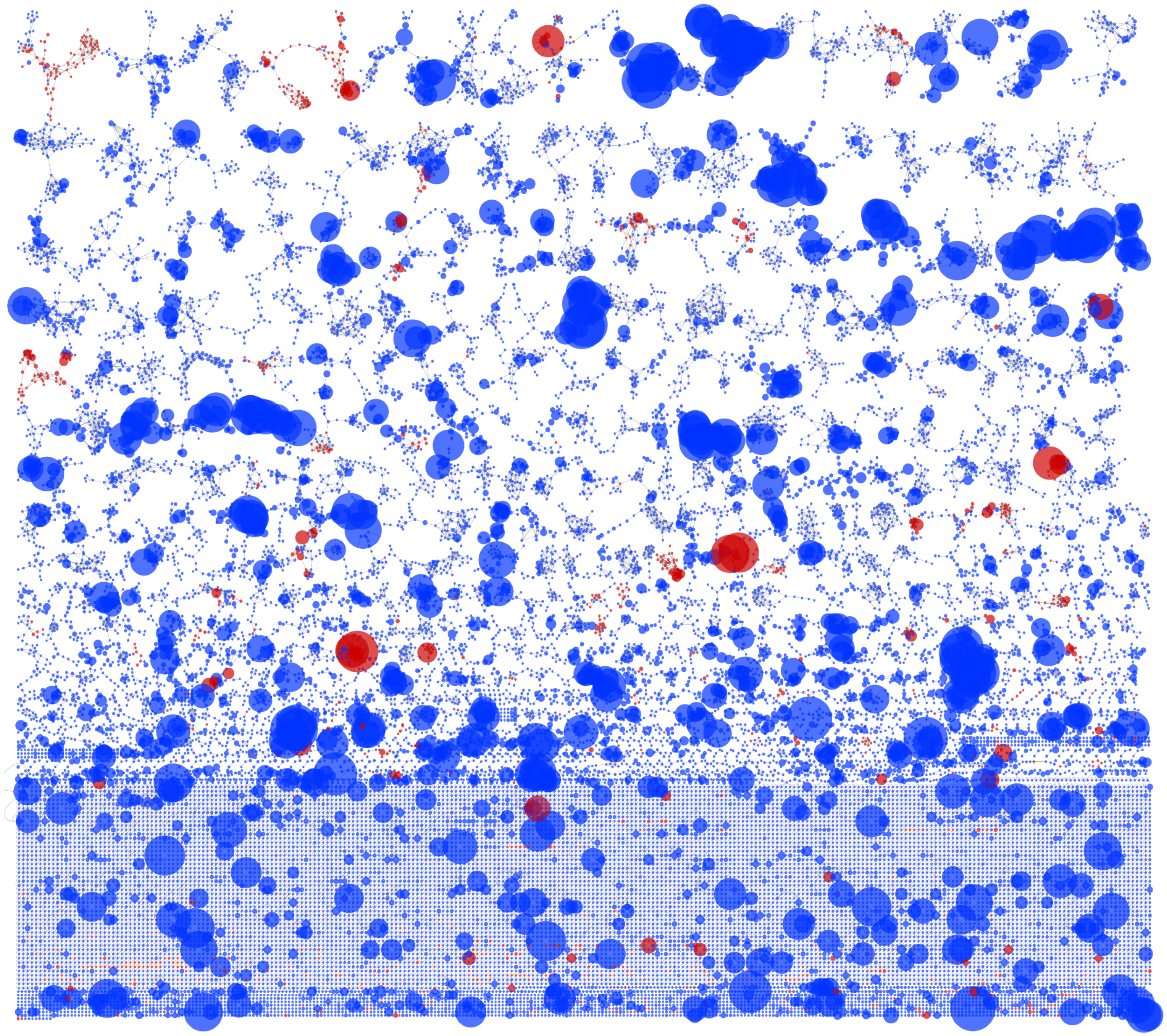
Do we see any microbial molecules?

GNPS Networking of American Gut Samples

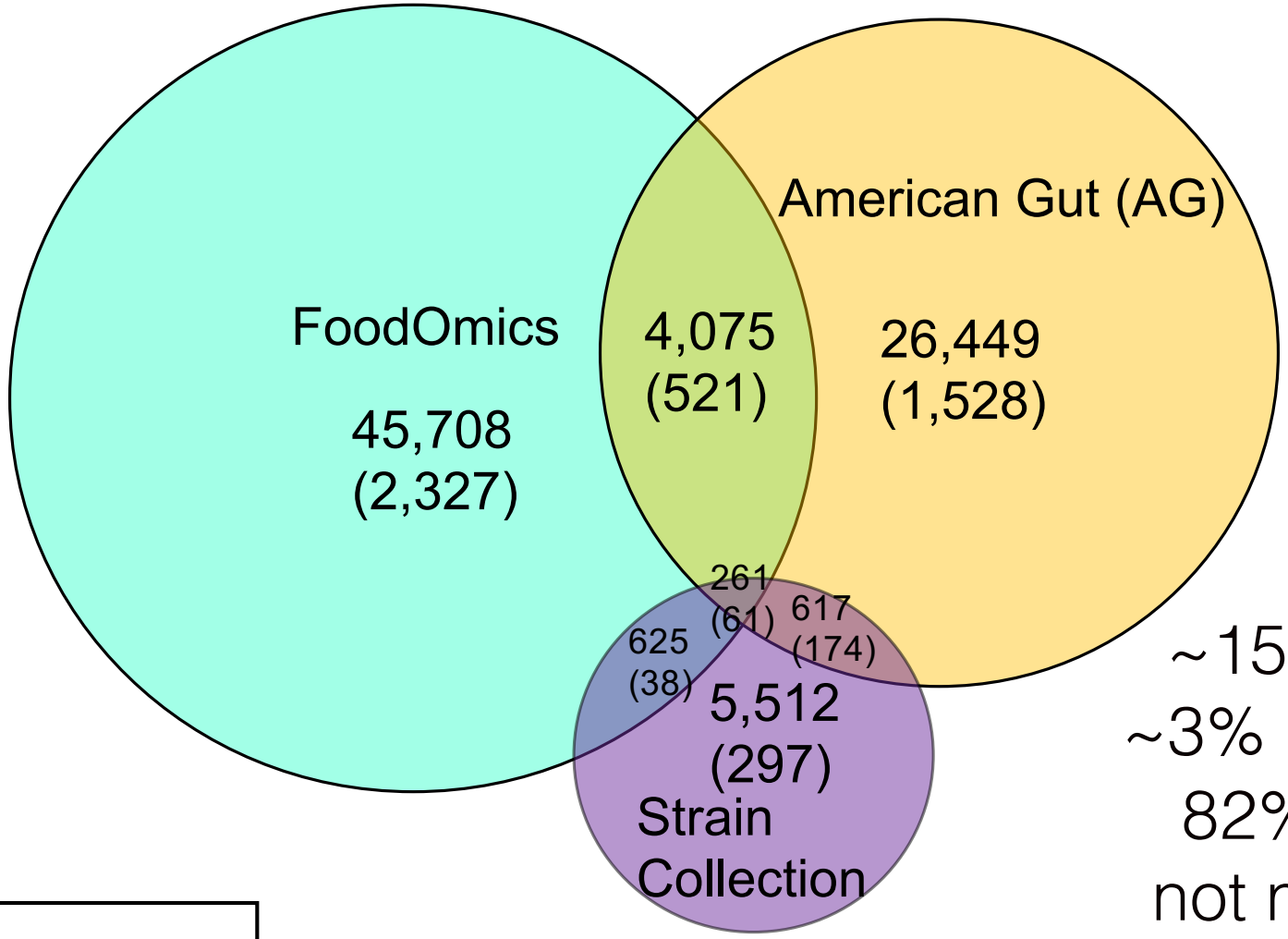
31,511 nodes

unannotated
annotated

less frequent ⟷ more frequent



american gut

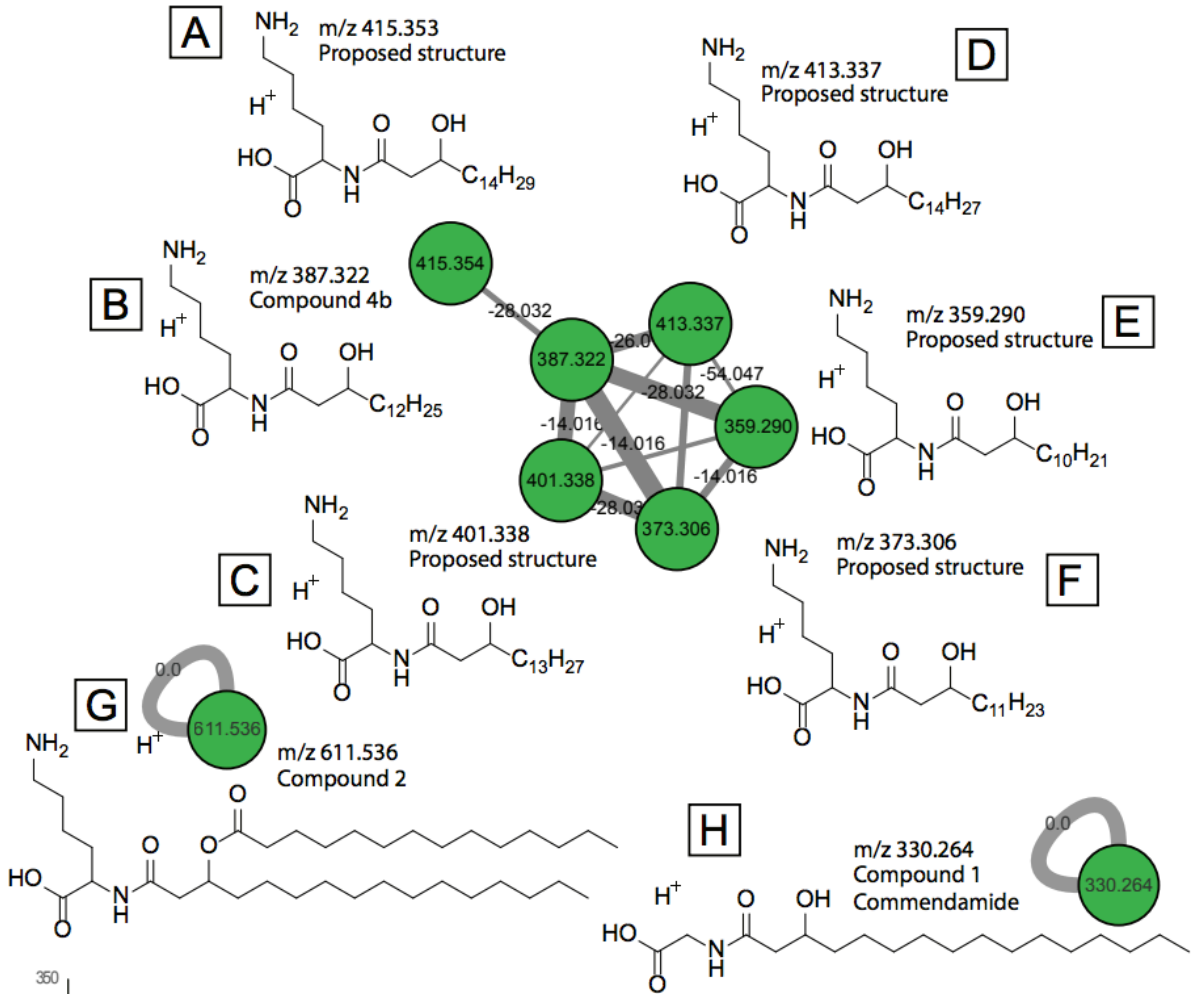


~15% Food
~3% Microbes
82% origin
not matched

GNPS

NIH Human Microbiome Project

Microbial Reference Genomes



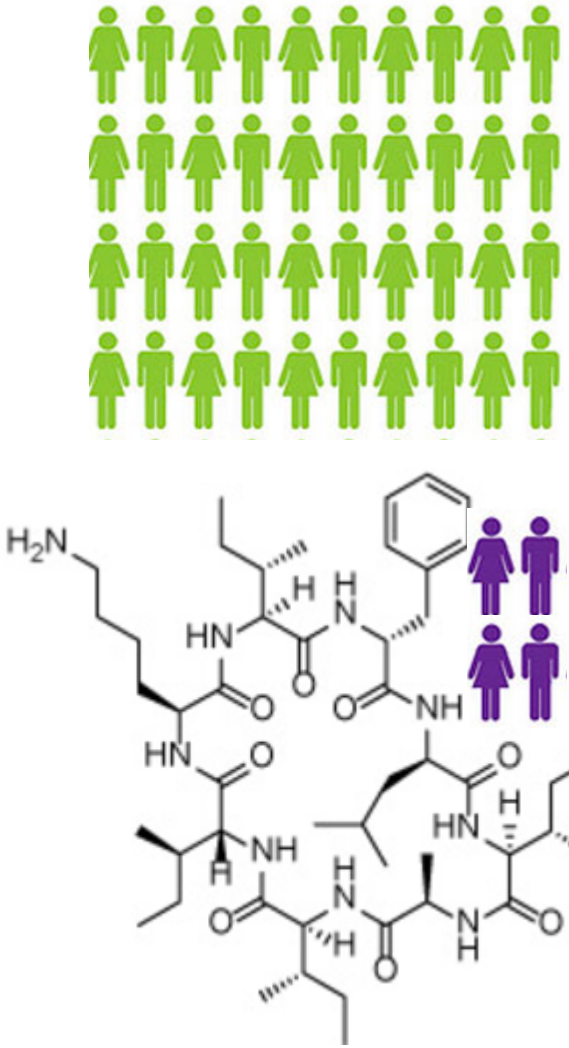
nature
International journal of science

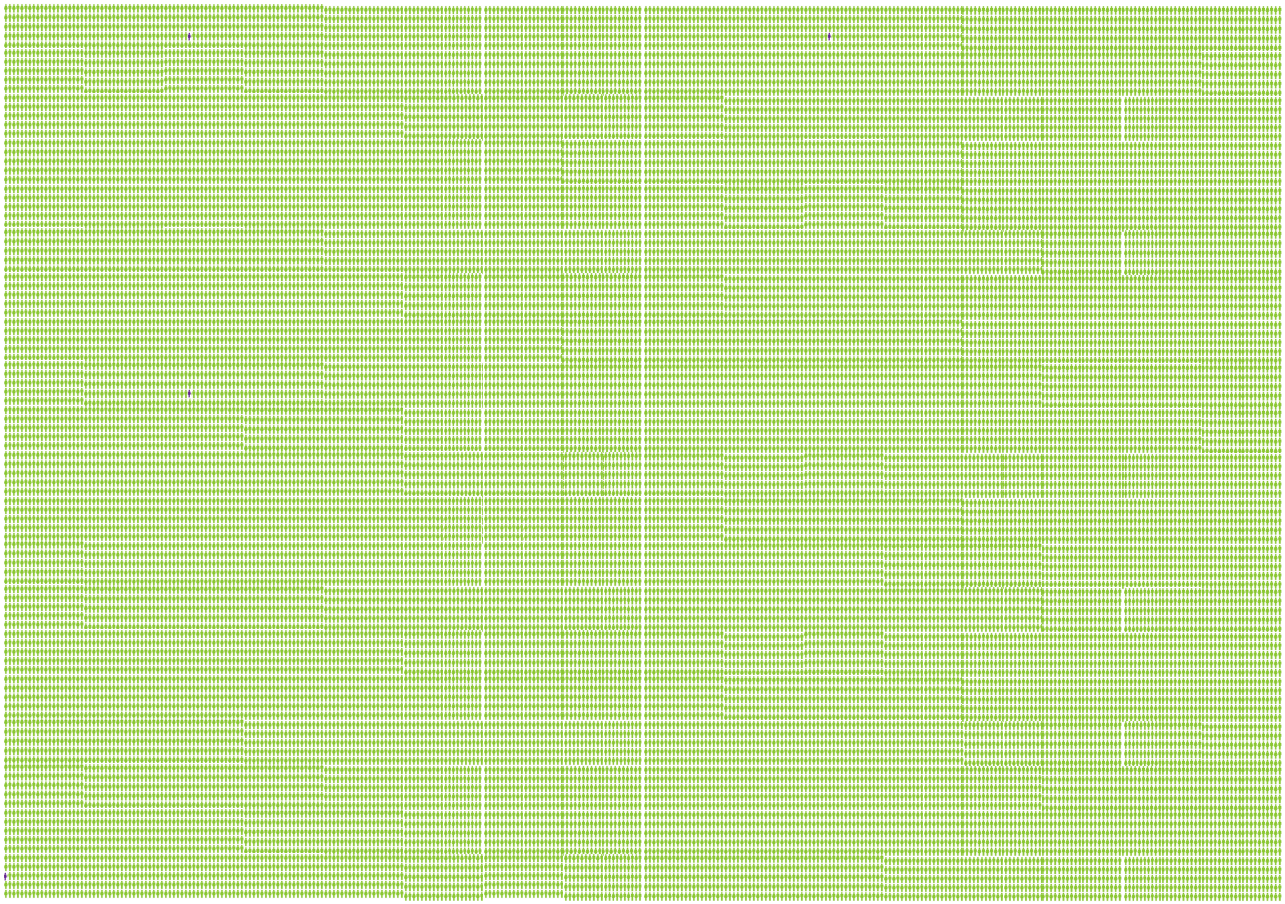
Commensal bacteria make GPCR ligands that mimic human signalling molecules

Louis J. Cohen, Daria Esterhazy, Seong-Hwan Kim, Christophe Lemetre, Rhiannon R. Aguilar, Emma A. Gordon, Amanda J. Pickard, Justin R. Cross, Ana B. Emiliano, Sun M. Han, John Chu, Xavier Vila-Farres, Jeremy Kaplitt, Aneta Rogoz, Paula Y. Calle, Craig Hunter, J. Kipchirchir Bitok & Sean F. Brady

regulates glucose homeostasis

Matches to microbial molecules annotated by the GNPS community



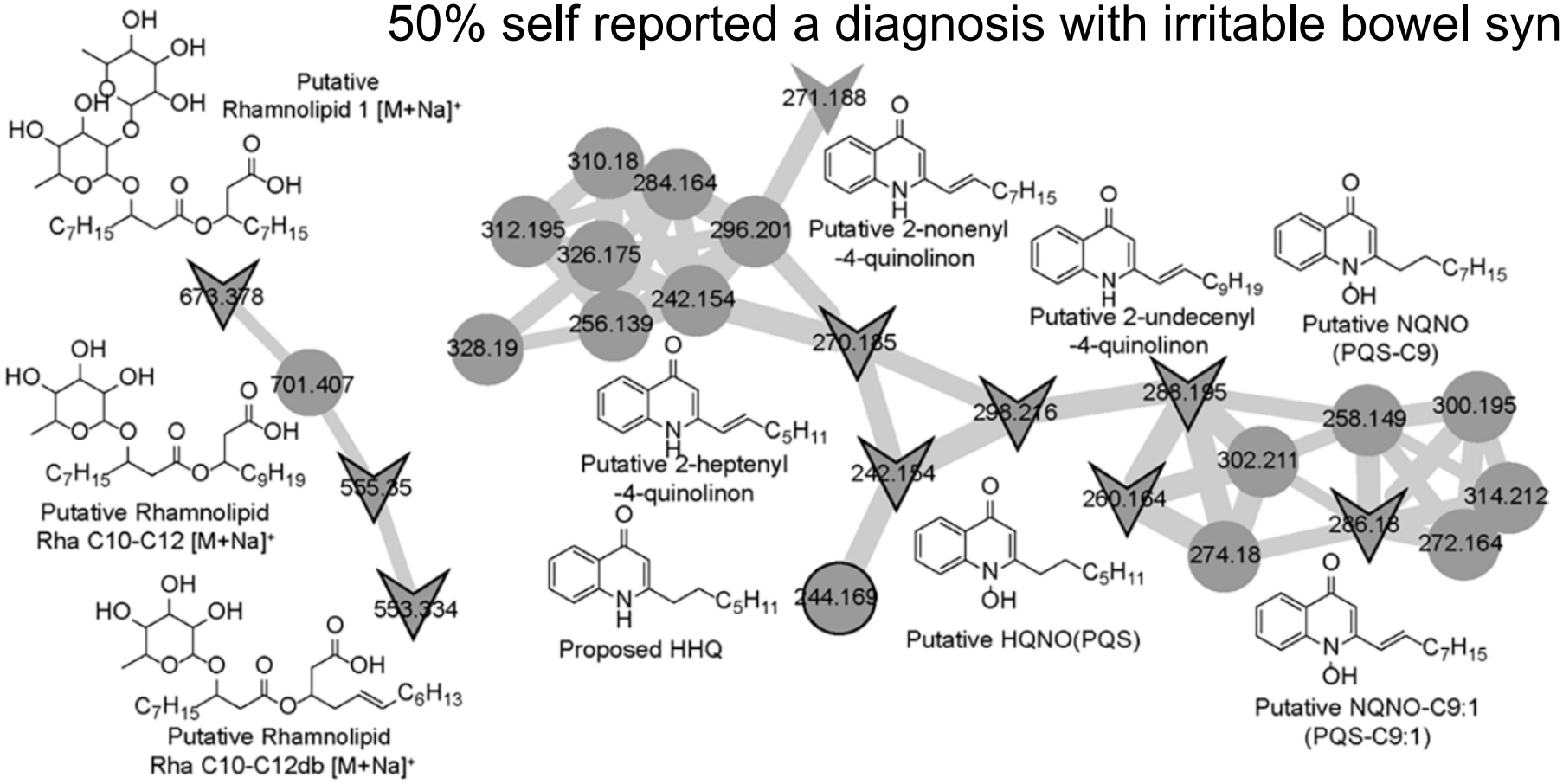


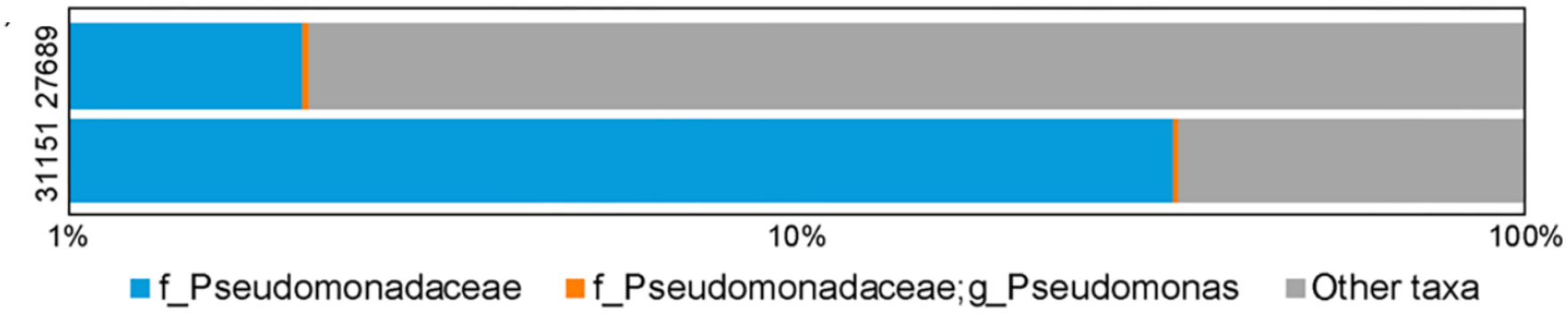
Matched to 4 out of the 2100 samples

Mohimani

Matches to microbial molecules annotated by the GNPS community

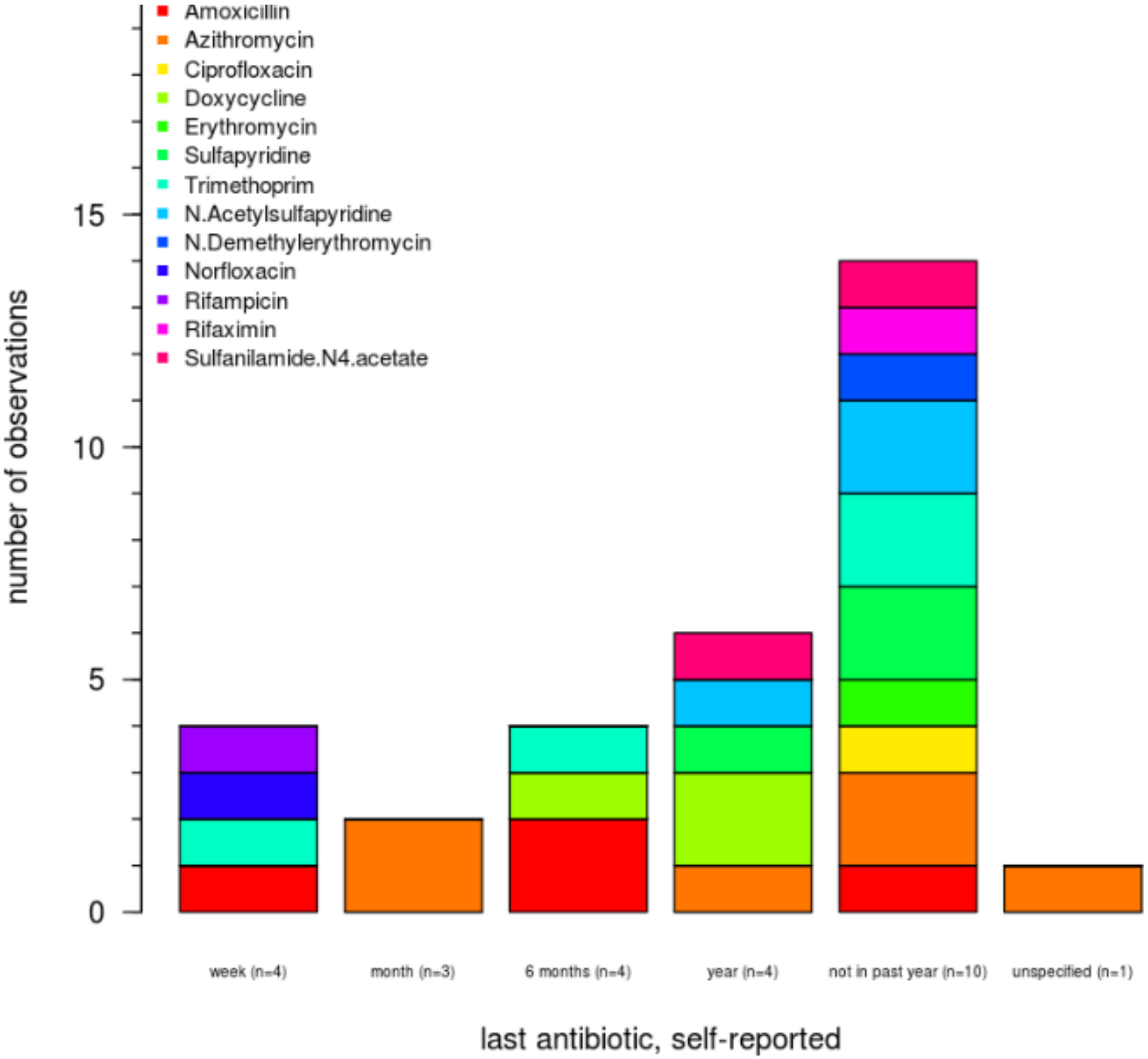
Pseudomonas aeruginosa-associated molecules.
50% self reported a diagnosis with irritable bowel syndrome.





Melnik

>80% of people where we detected antibiotics had not taken antibiotics in >6 months.

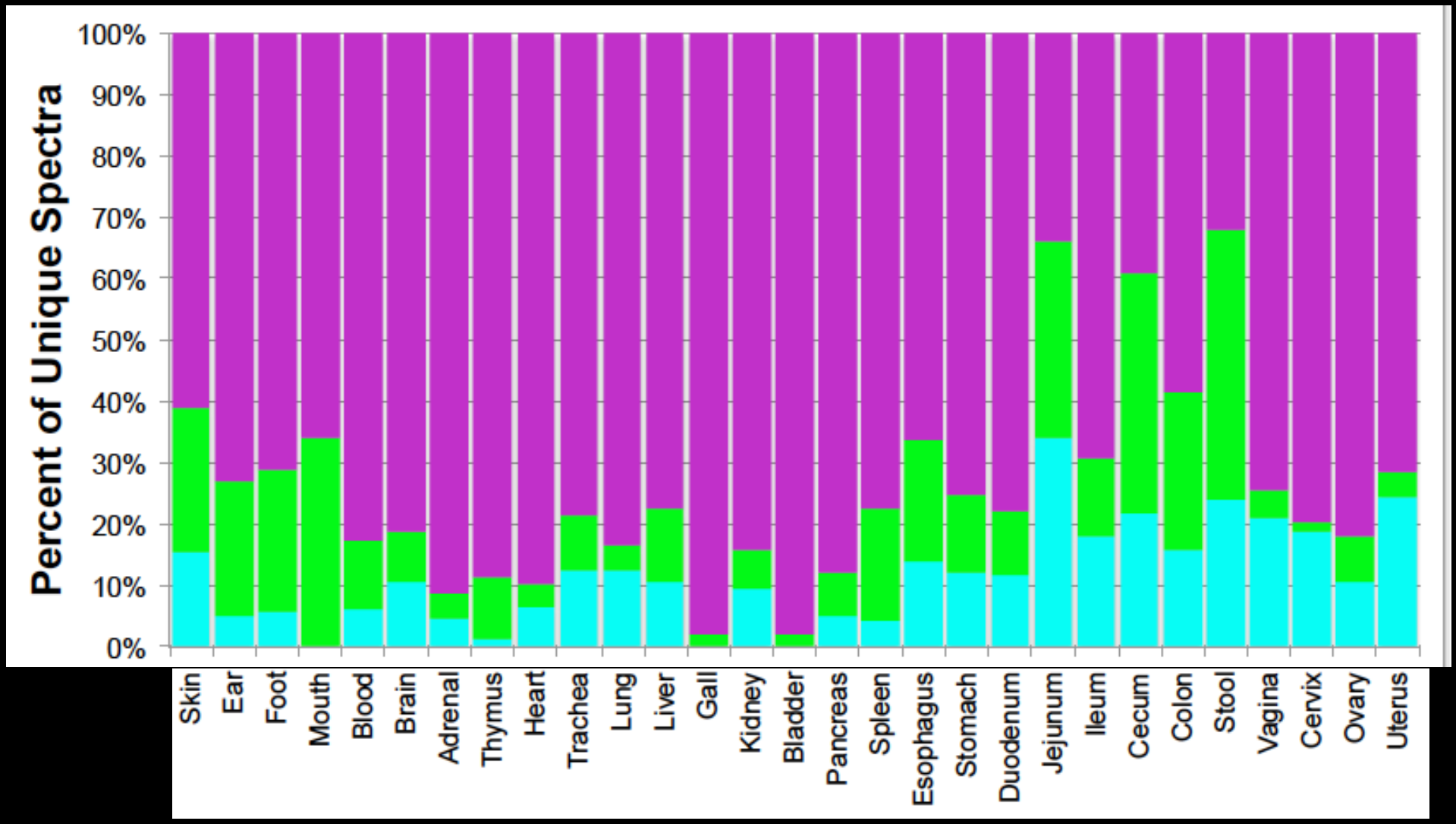


Example Data reuse 4.

We are not mono-colonized, we colonized by 1000s of microbes.

What impact does a “normal” microbiome have?

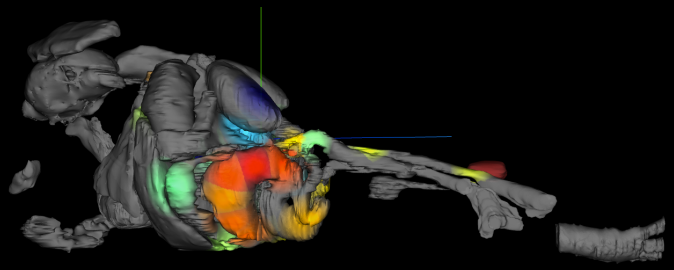
As much as 70% of the chemistry is different due to microbes
What are these microbial molecules?



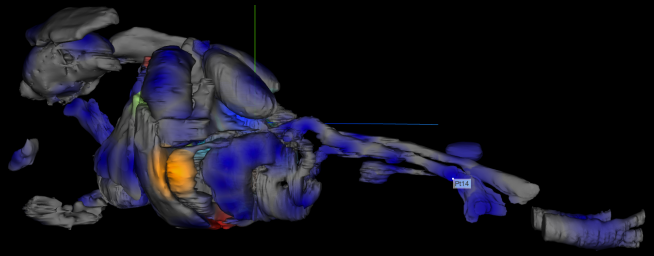
-Microbes +Microbes Shared

Paper will be submitted in a week or so

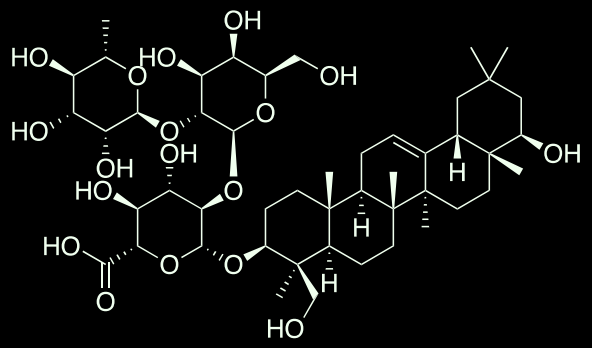
Microbiome-Mediated Metabolism of Food



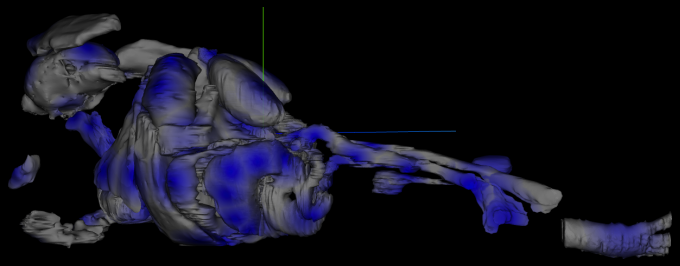
-Microbes



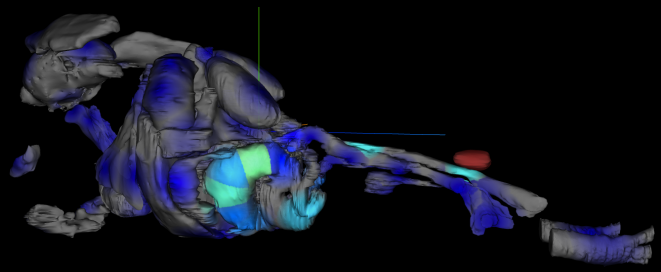
+Microbes



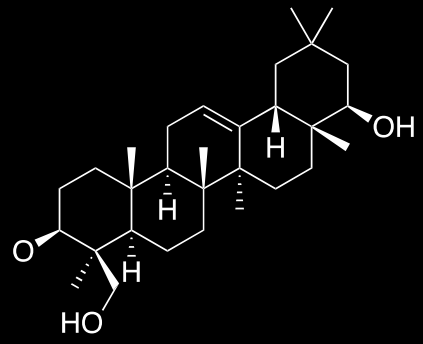
Soyasaponin



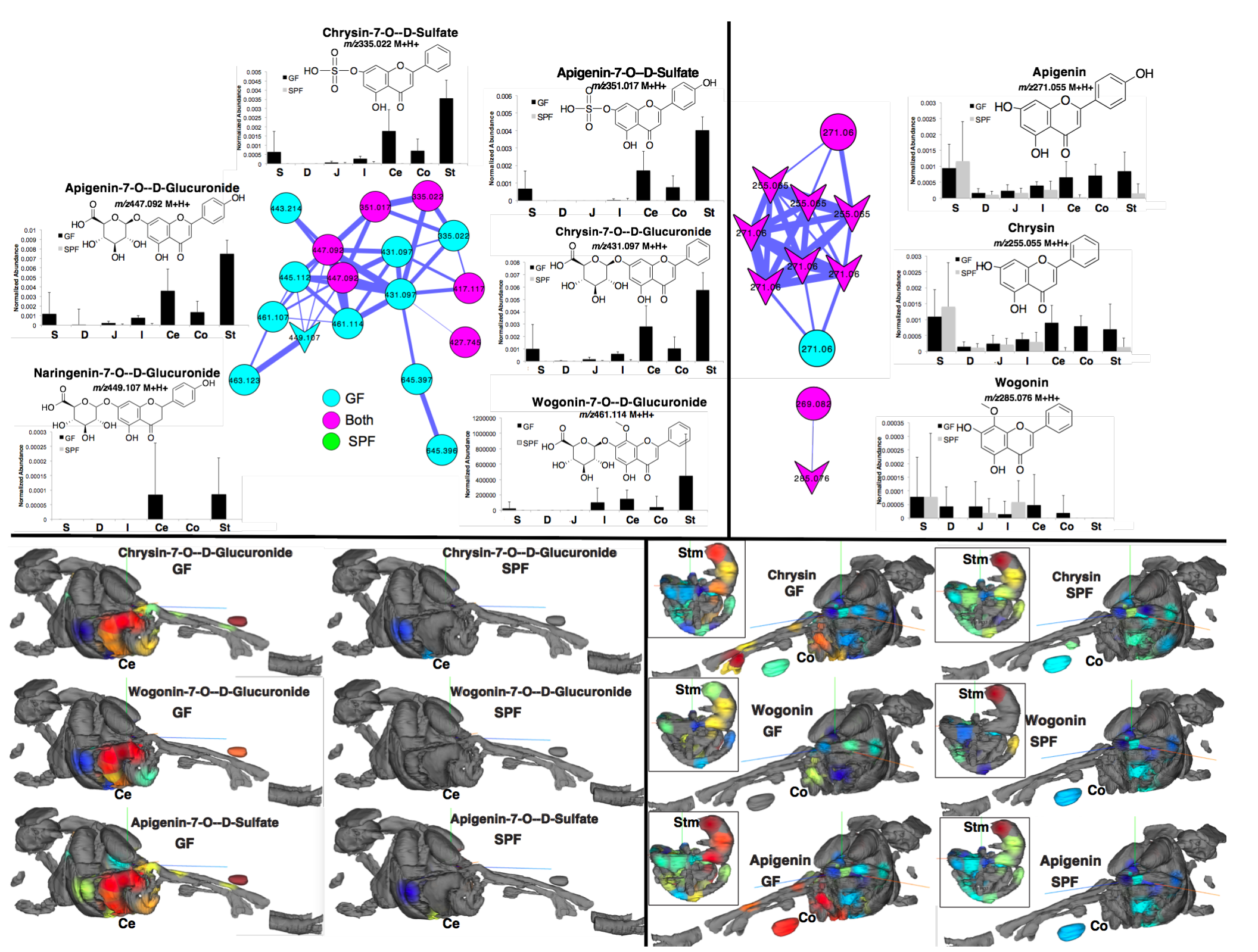
-Microbes

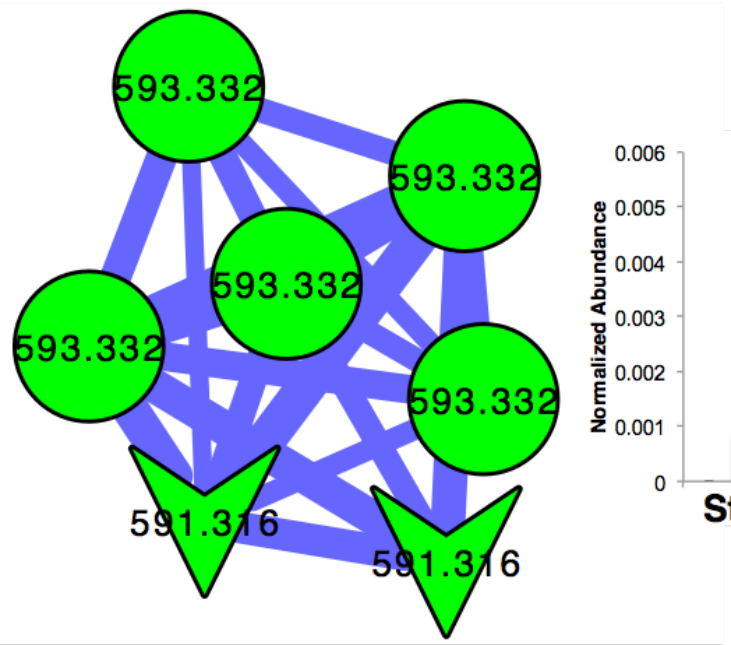


+Microbes



Soyaspongenol

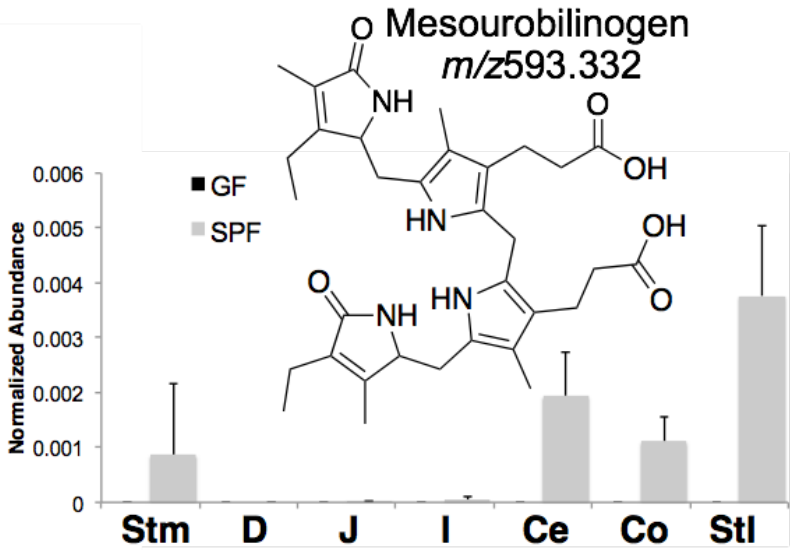

Chrysin-7-O--D-Sulfate
m/z335.022 M+H+
Apigenin-7-O--D-Sulfate
m/z351.017 M+H+
Apigenin
m/z271.055 M+H+
Apigenin-7-O--D-Glucuronide
m/z447.092 M+H+
Chrysin-7-O--D-Glucuronide
m/z431.097 M+H+
Chrysin
m/z255.055 M+H+
Naringenin-7-O--D-Glucuronide
m/z449.107 M+H+
Wogonin-7-O--D-Glucuronide
m/z461.114 M+H+
Wogonin
m/z285.076 M+H+
GF
SPF
S D J I Ce Co St
Normalized Abundance
443.214
351.017
335.022
447.092
431.097
445.112
417.117
461.107
461.114
449.107
427.748
463.123
645.397
645.396
GF
Both
SPF
271.06
255.065
269.082
285.076
Chrysin-7-O--D-Glucuronide
GF
Chrysin-7-O--D-Glucuronide
SPF
Wogonin-7-O--D-Glucuronide
GF
Wogonin-7-O--D-Glucuronide
SPF
Apigenin-7-O--D-Sulfate
GF
Apigenin-7-O--D-Sulfate
SPF
Stm
Chrysin
GF
Chrysin
SPF
Wogonin
GF
Wogonin
SPF
Apigenin
GF
Apigenin
SPF
Ce
Co

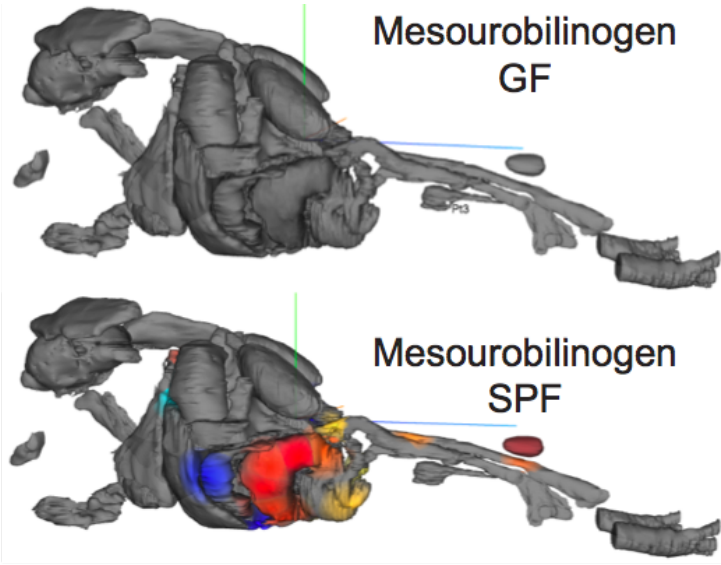

593.332
593.332
593.332
593.332
593.332
591.316
591.316

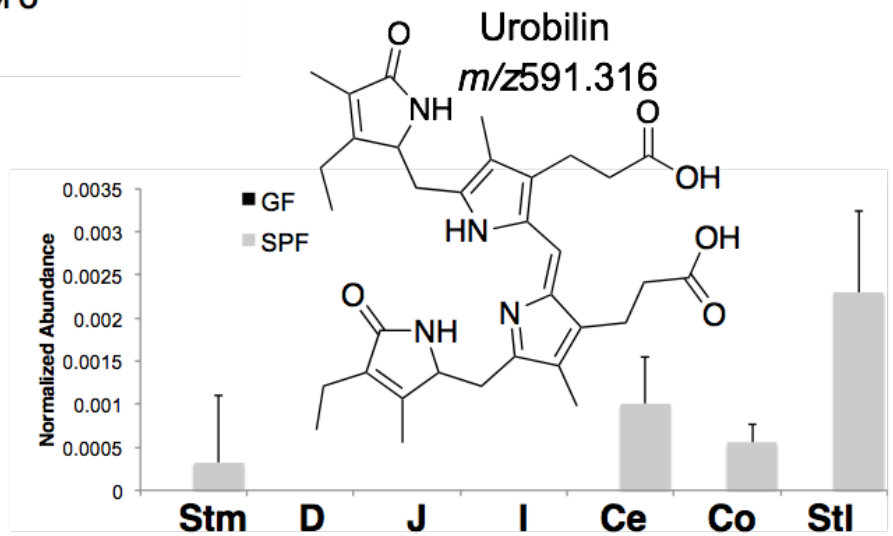
GF

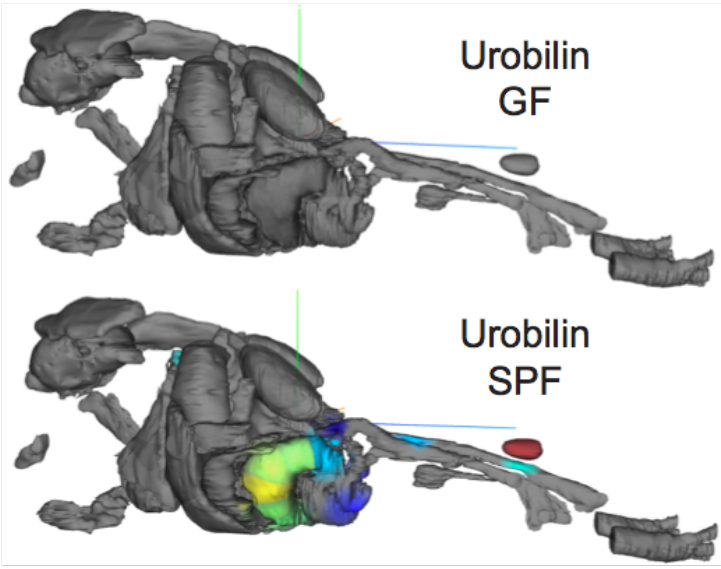
Both

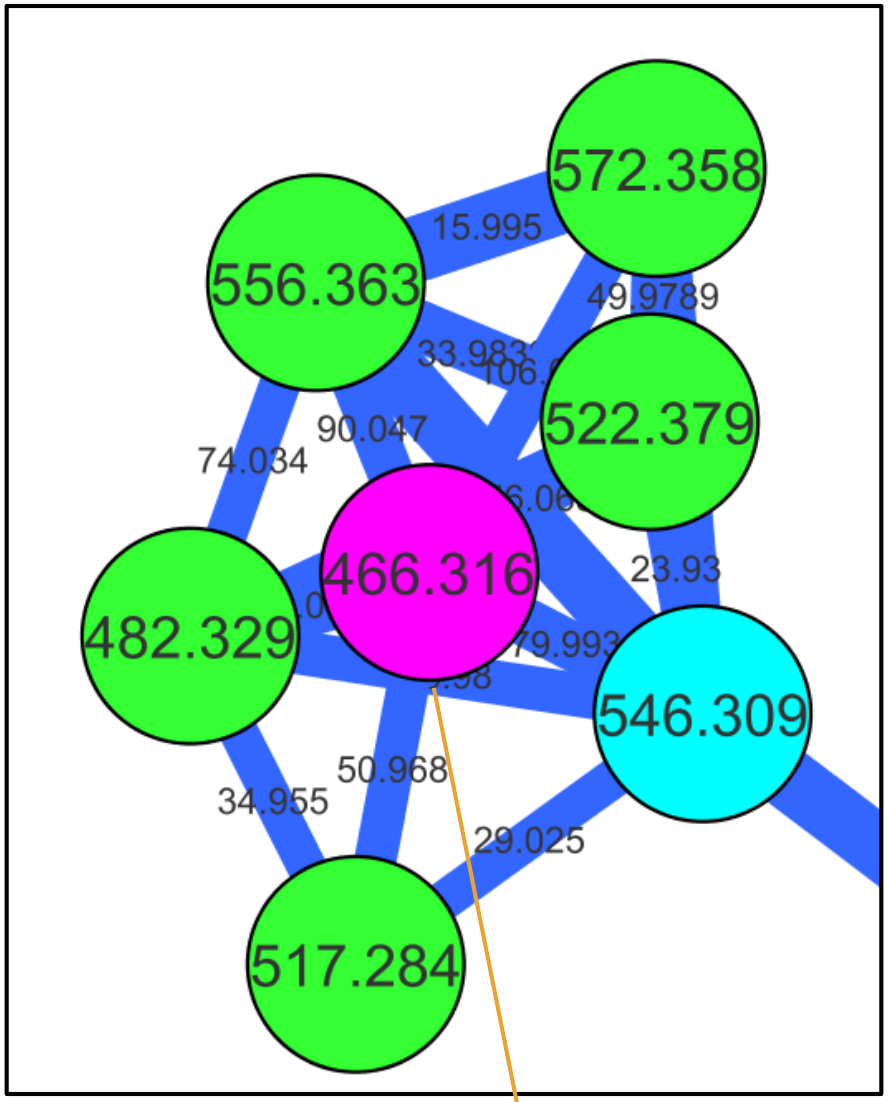
SPF

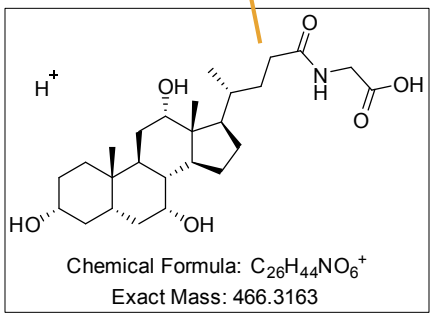

Mesourobilinogen
m/z593.332
GF
SPF
Normalized Abundance
0.006
0.005
0.004
0.003
0.002
0.001
0
Stm
D
J
I
Ce
Co
StI


Mesourobilinogen
GF
Mesourobilinogen
SPF


Urobilin
m/z591.316
GF
SPF
Normalized Abundance
0.0035
0.003
0.0025
0.002
0.0015
0.001
0.0005
0
Stm
D
J
I
Ce
Co
StI

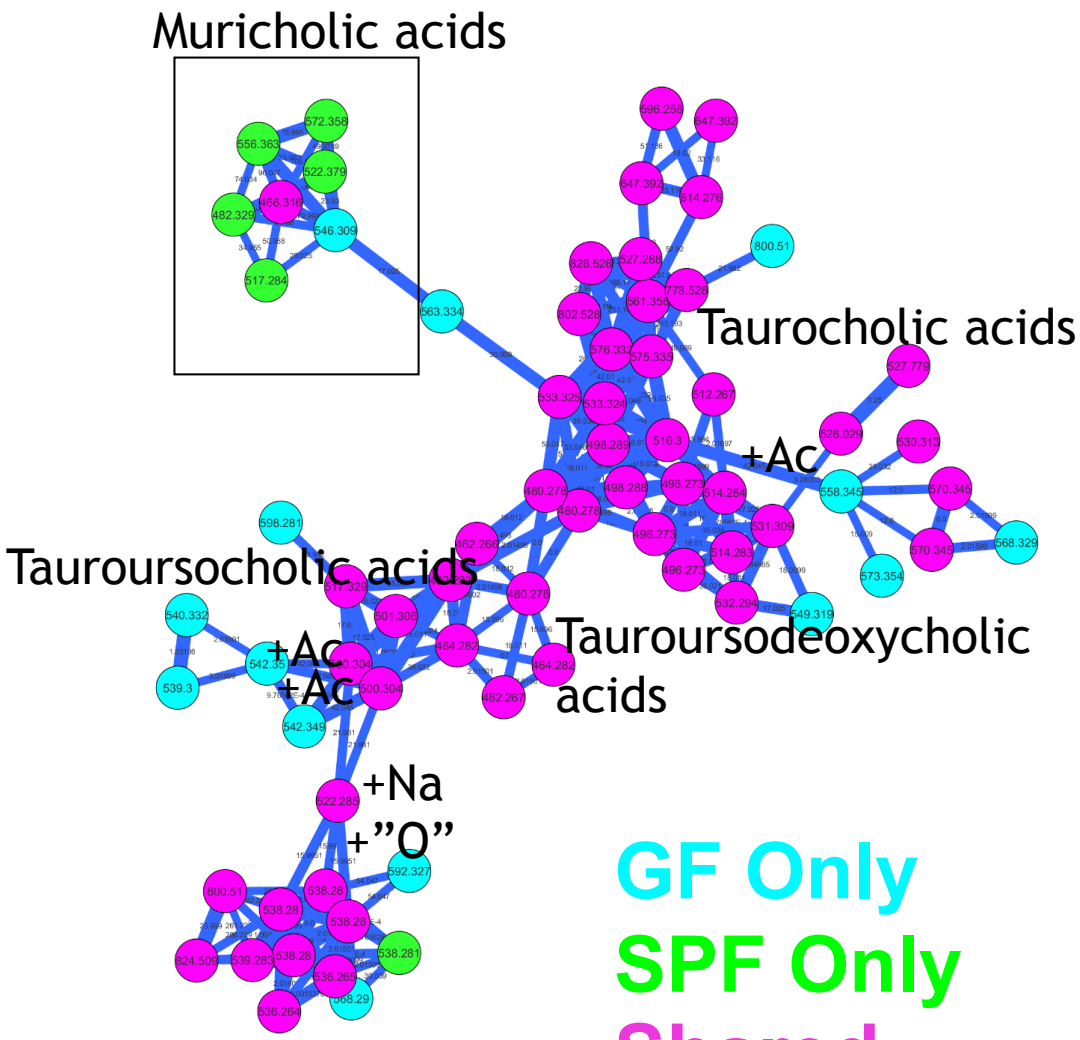

Urobilin
GF
Urobilin
SPF





Muriglycocholic acid

Duodenum

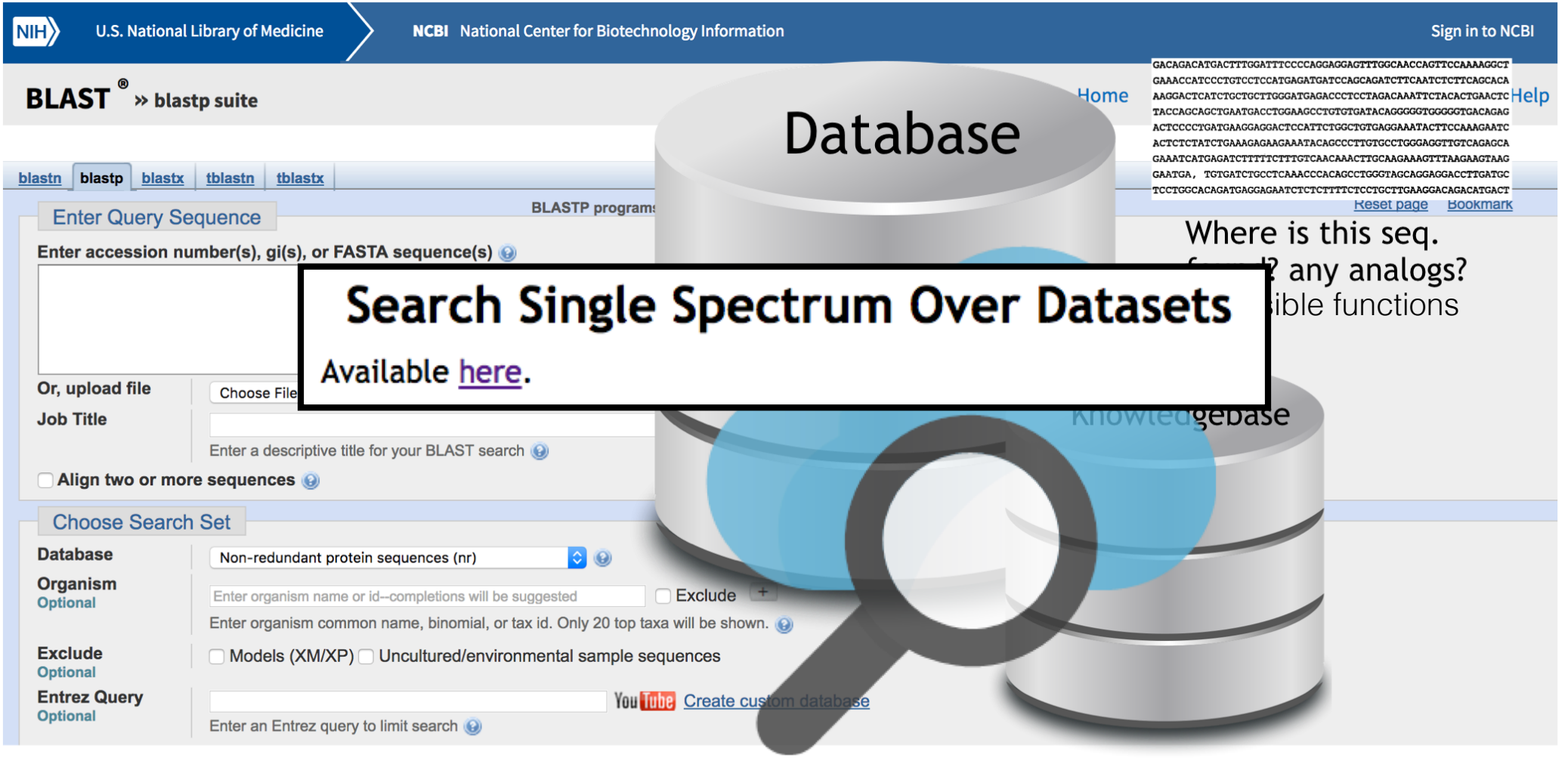


GF Only

SPF Only

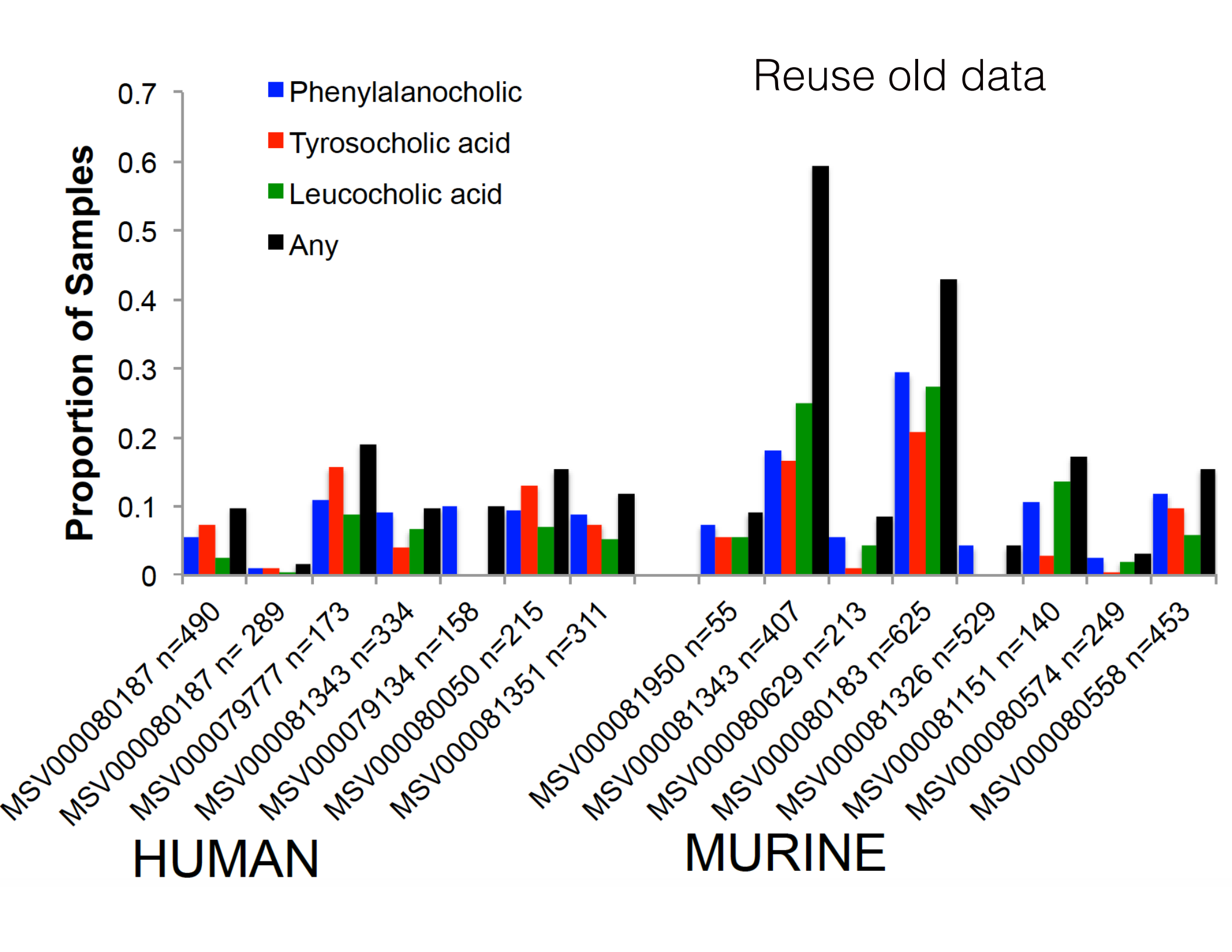
Shared

Are they found in humans?

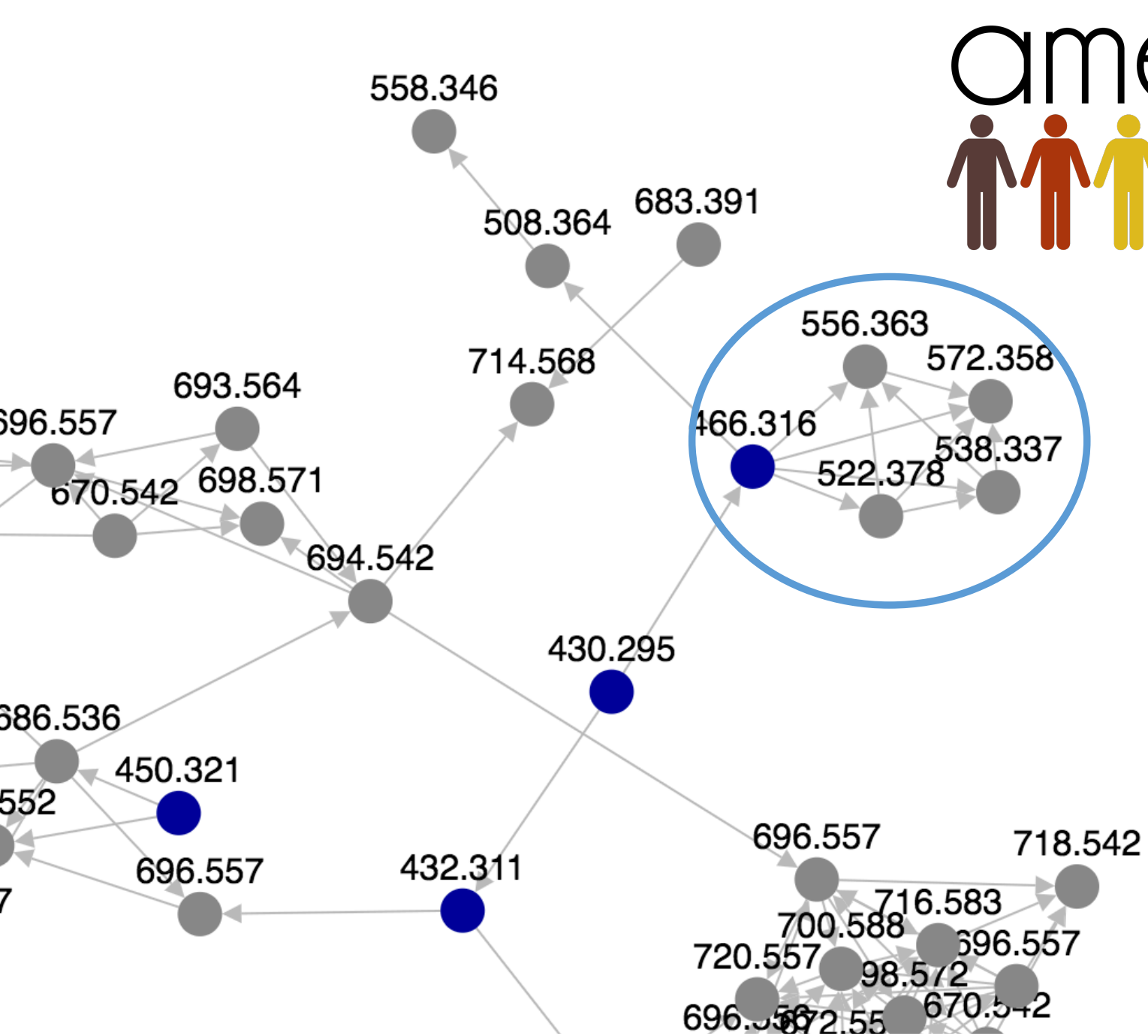


Retrieval infrastructure

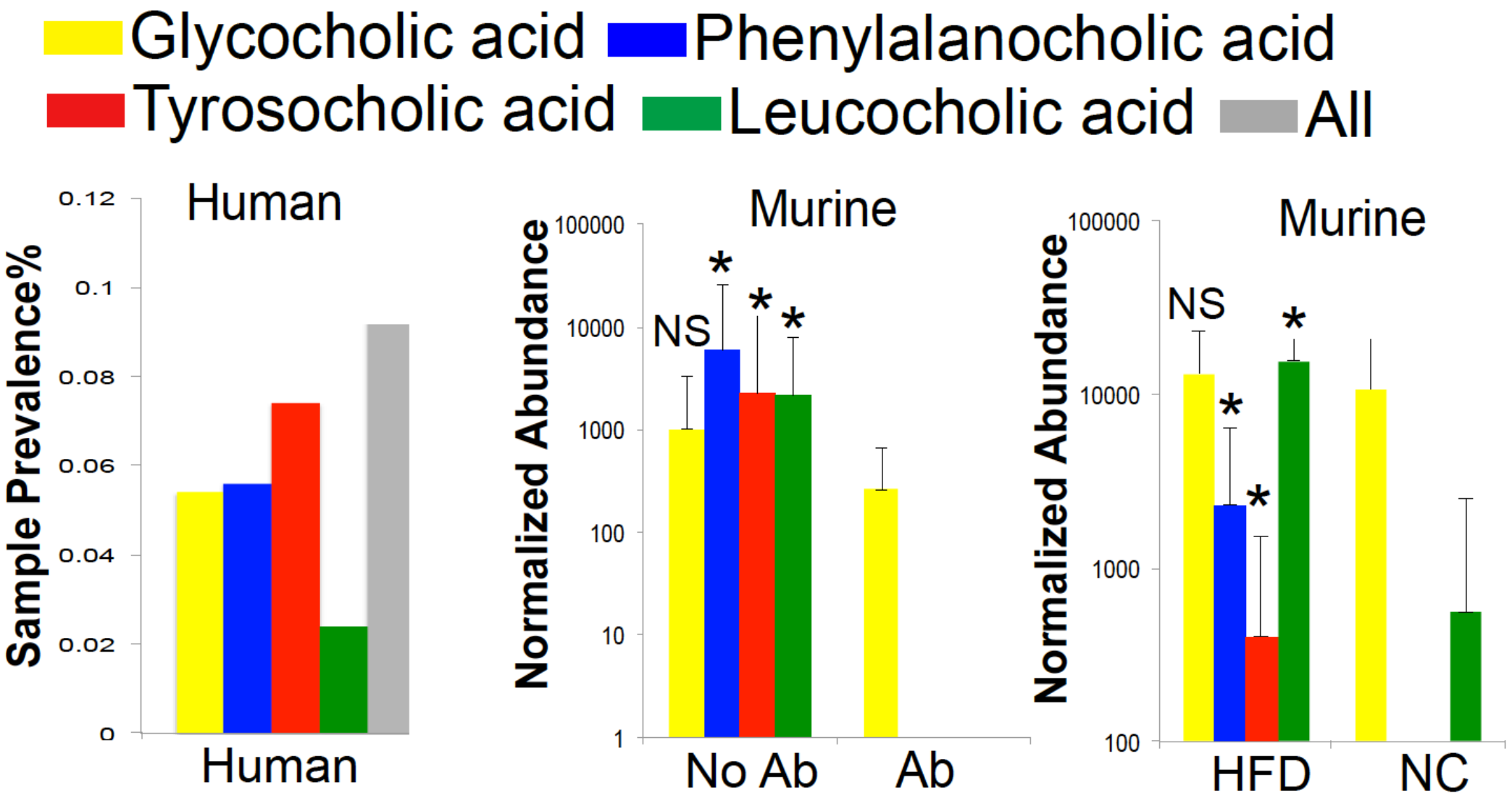
GNPS


Reuse old data
Phenylalanocholic
Tyrosocholic acid
Leucocholic acid
Any
Proportion of Samples
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
MSV000080187 n=490
MSV000080187 n= 289
MSV000079777 n=173
MSV000081343 n=334
MSV000079134 n=158
MSV000080050 n=215
MSV000081351 n=311
MSV000081950 n=55
MSV000081343 n=407
MSV000080629 n=213
MSV000080183 n=625
MSV000081326 n=529
MSV000081151 n=140
MSV000080574 n=249
MSV000080558 n=453
HUMAN
MURINE

Reuse old data: Are they found in humans?



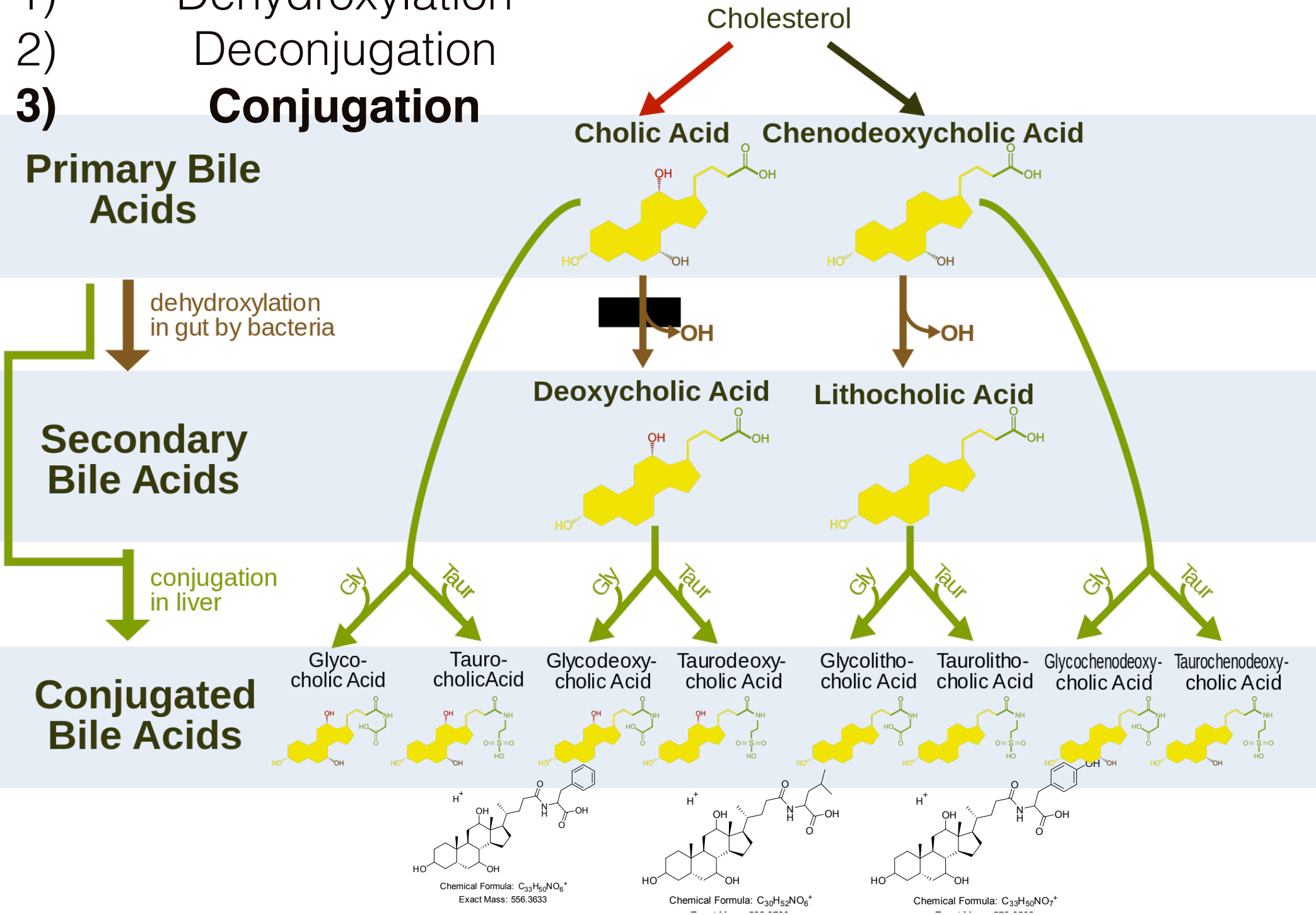
Looking at the metadata of published data



Generally reduced with antibiotic use, diarrhea increased, in high fat diet increased, viral infection increased

Bacteria are responsible for

1) Dehydroxylation
2) Deconjugation
3) **Conjugation**



Vision for the future

and is a form of data reuse

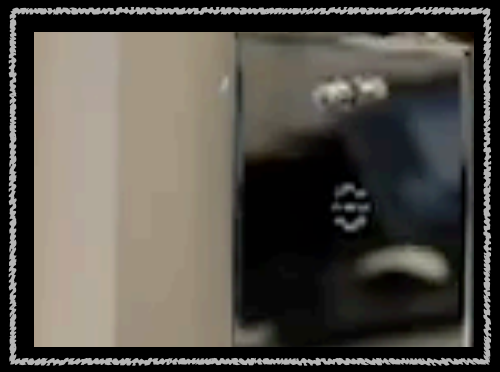

OH
H
H
H
HO
O
CH3
H3C
N
N
O
N
N
CH3
O
OH
H2N
O
OH
NH2
HO

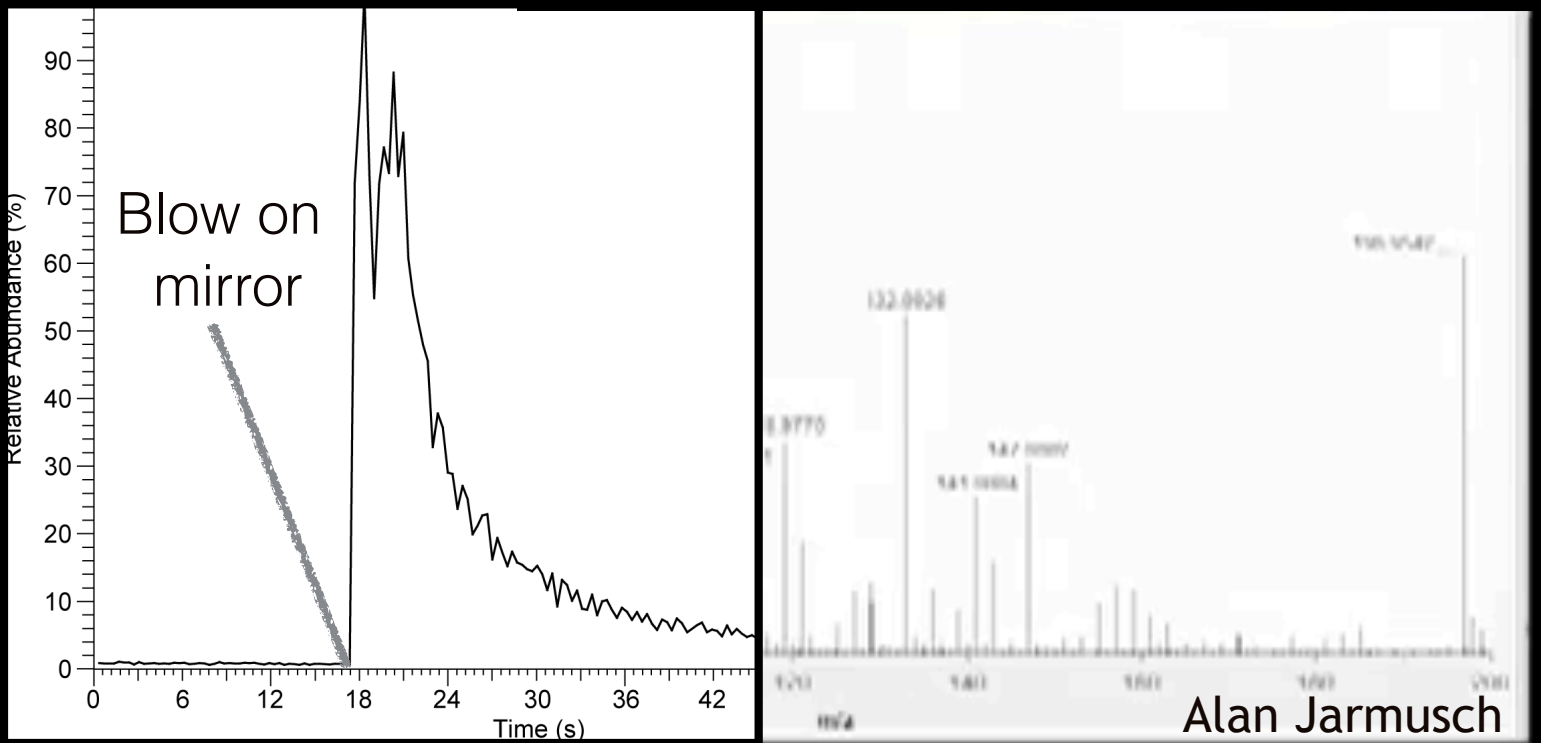

100
80
60
40
20
0
Frequency
Your Sample
Average
Similar Diet
Same Gender
Similar BMI
Similar Age
Michael Pollan
Firmicutes
Bacteroidetes
Proteobacteria
Actinobacteria
Verrucomicrobia
Tenericutes
Cyanobacteria
Fusobacteria
Other


DANGER:
IBD RISK!

Go and see you doctor!!

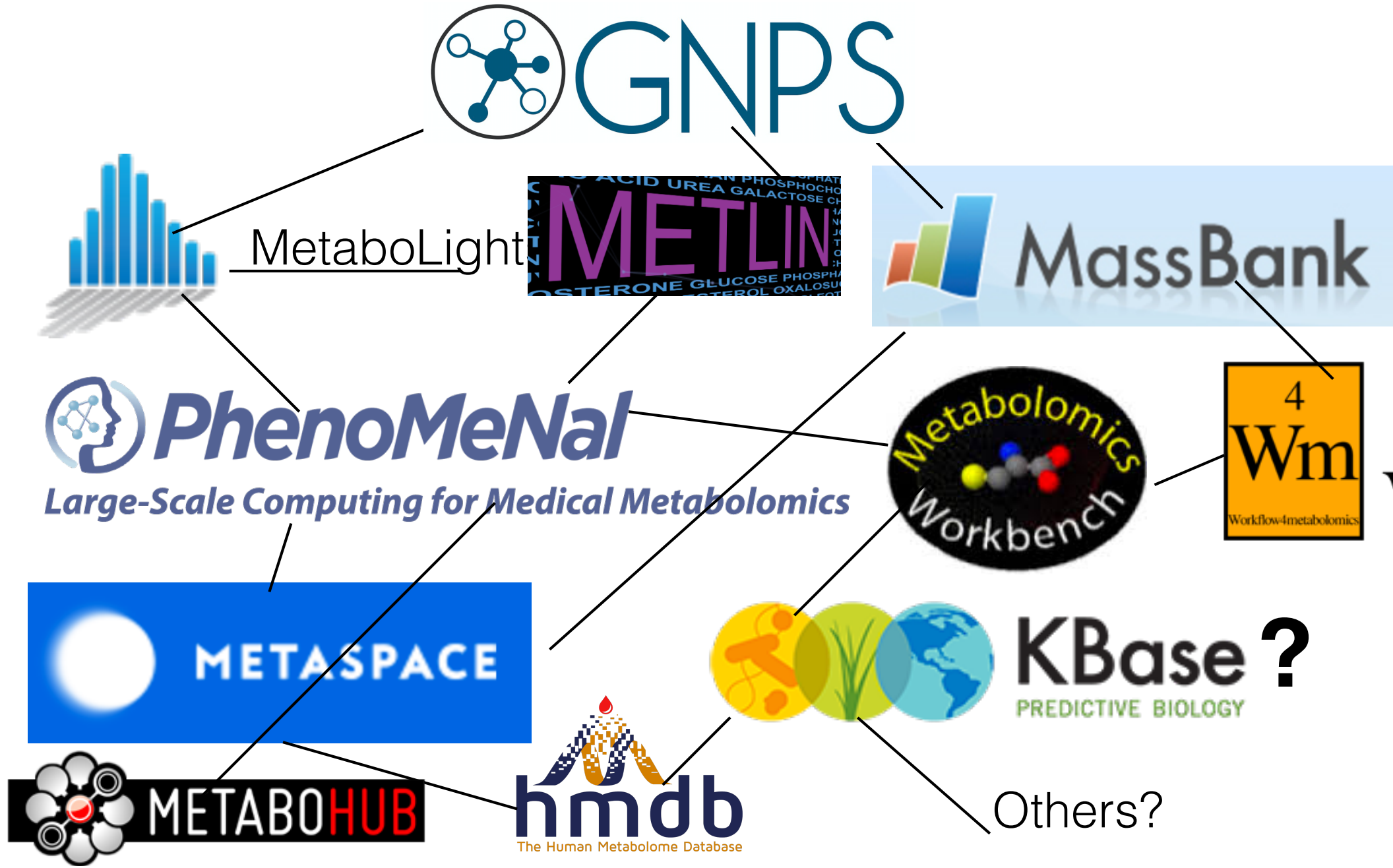
Mirror Mass Spectrometry





Data availability and algorithms are the limitation

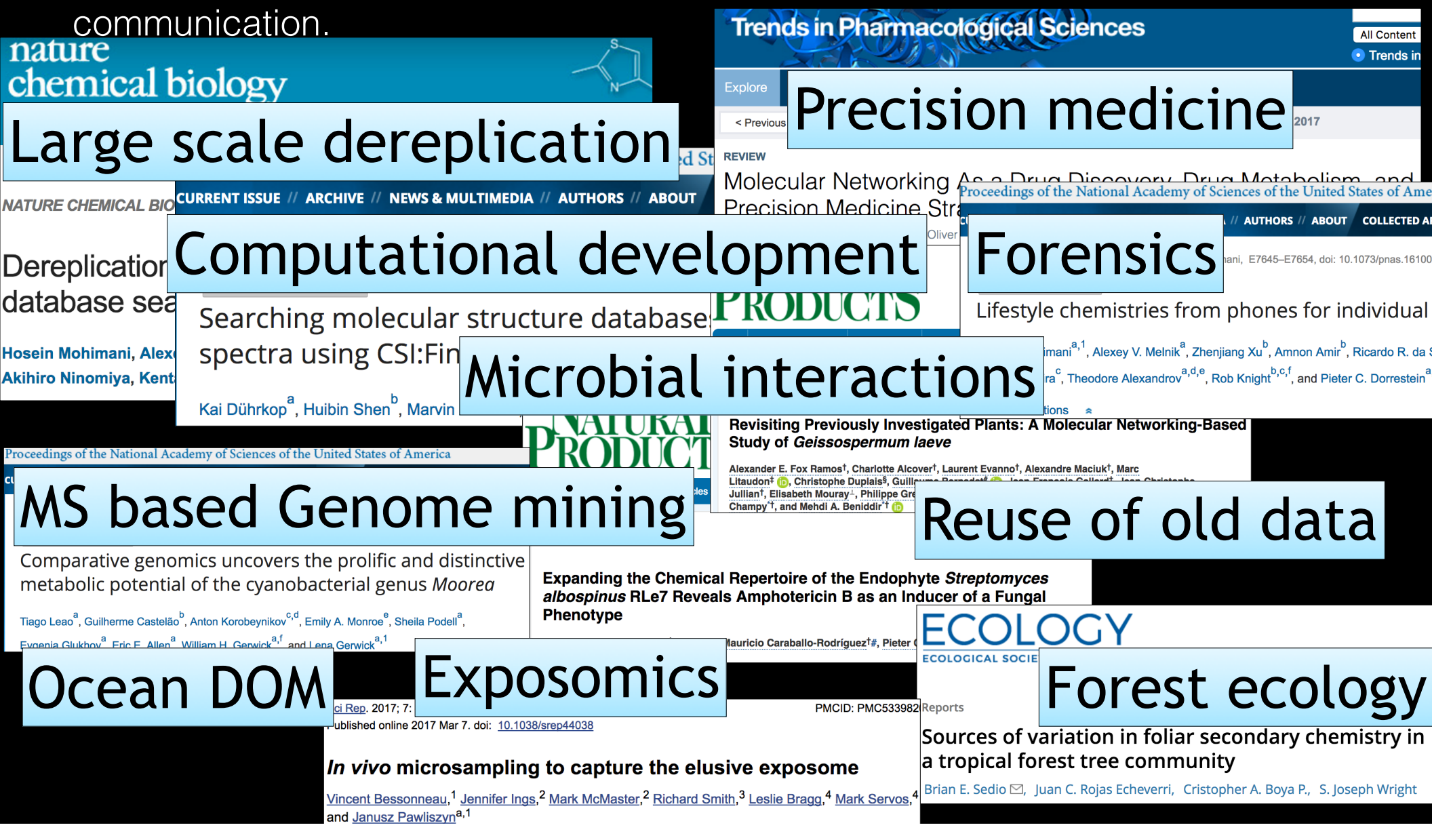
Ultimately a network of interoperation need to be established to analyze MS data- as they all serve different functions. What might this look like?



Ultimately a network of interoperation need to be established to analyze MS data- as they all serve different functions.

Take home messages

1) Raw data has to be shared-this is key.
2) If the same data is used, and same tools with the same time stamp, one should get the same data. In other words cloning of the exact analysis of the method with the same parameters is possible. Metadata collection is critical and will need to be standardized
3) Multiple resources should communicate-data standards enables such cross communication.



THANKS!

The community to crowdsource molecular networking

Mingxun Wang[1,3], Jeremy Carver[1,3], Vanessa V. Phelan[1,5], Laura M. Sanchez[1,5], Neha Garg[1,5], Yao Peng[1,4], Jeramie Watrous[5], Clifford Kopano[4], Don Duy Nguyen[4], Tal Luzzatto-Knaan[5], Carla Porto[5], Amina Bouslimani[5], Alexey V. Melnik[5], Michael J. Meehan[5], Wei-Ting Liu[6], Paul D. Boudreau[8], Eduardo Esquenazi[10], Mario Sandoval-Calderón[11], Roland D. Kersten[12], Robert A. Quinn[13], Katherine Duncan[8,39], Richard Hsu[4,6], Dimitrios J. Floros[4,5], Ronnie G. Gavilan[14], Karin Kleigrewe[8], Norberto Peporine Lopes[16], Trent Northen[17], Rachel J. Dutton[46], Delphine Parrot[19], Erin E. Carlson,[30] Bertrand Aigle[31], Lars Jelsbak[29], Christian Sohlenkamp[11], Charlotte F. Michelsen[29], Pavel Pevzner[3], Anna Edlund[33, 34], Jeffrey McLean[28], Joern Piel[26], Brian T. Murphy[27], Lena Gerwick[8], Chih-Chuang Liaw[22], Yu-Liang Yang[23], Hans-Ulrich Humpf[40], Maria Maansson[24], Robert A Keyzers[25], Amy C Sims[36], Andrew R. Johnson[37], Ashley M. Sidebottom[37], Brian E. Sedio[38], Andreas Klitgaard[29], Charles B. Larson[8,42,49], Cristopher A. Boya P.[14], Daniel Torres-Mendoza[14], David J. Gonzalez[5], Denise Brentan Silva[16], Egle Pociute[10], Ellis O'Neill[8], Enora Briand[8, 19], Eric J. N. Helfrich[26], Eve A. Granatosky[32], Evgenia Glukhov[8], Florian Ryffel[26], Hailey Houson[10], Hosein Mohimani[3], Jenan Kharbush[8], Yi Zeng[4], Julia A. Vorholt[26], Kenji L. Kurita[21], Pep Charusanti[15], Kerry L. McPhail[44], Kristian Fog Nielsen[29], Lisa Vuong[10], Maryam Elfeki[27], Matthew F. Traxler[18], Max Crüsemann[8], Niclas Engene[47], Nobuhiro Koyama[5], Oliver B. Vining[44], Ralph Baric[36], Ricardo Roberto Silva[16], Samantha J Mascuch[8], Sophie Tomasi[19], Stefan Jenkins[17], Venkat Macherla[10], Thomas Hoffmann[51], Vinayak Agarwal[52], Philip G. Williams[54], Jingqui Dai[54], Ram Neupane[54], Joshua Gurr[54], Andrés Mauricio Caraballo Rodríguez[53], Wenyuan Shi[50], Rob Knight[41], Paul R. Jensen[8], Bernhard Ø. Palsson[15], Kit Pogliano[7], Roger G. Linington[21], Marcelino Gutiérrez[14], , William H. Gerwick[5,8], Bradley S. Moore[5,8,52], Nuno Bandeira[2,3,5]

Photo: Nick Pizzo

Nature Biotech 2016, a

A Big Thanks to Friends, Co-workers and Support That Allow Us To Do The Research

Bandeira Lab
Ming Wang

Alexandrov lab
Sergey Ryazanov
Ivan Protsyuk

Knight lab
Antonio Peña,
Donna Berg-Lyons
Gail Ackermann
Embriette Hyde
Gregory Humphrey
James Gaffney
Jeff Dereus
Dominguez lab
Pilau lab

Moore Lab
Gerwick Labs
Jensen lab
Fenical Lab
Nizet Lab
Pevzner lab

Rohwer lab
Conrad lab
Lin lab

Medema lab

Boeker lab

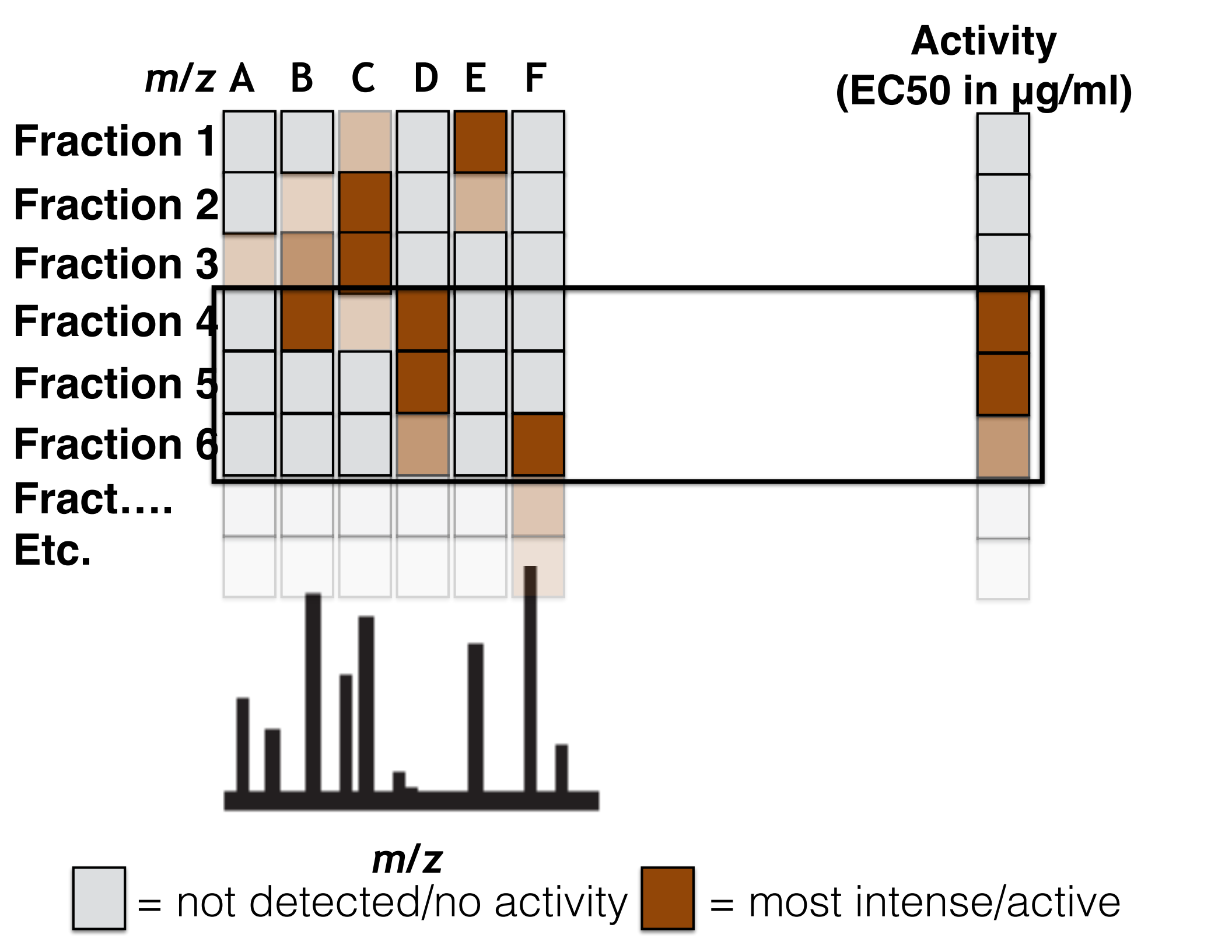
Mazmanian lab

Dorrestein lab (current)
Neha Garg
Robby Quinn
Cliff Kapono
Fernado Vargas
Don Nguyen
Amina Bouslimani
Alexey Melnik
Kathleen Dorrestein
Mike Meehan
Carla Porto
Ricardo da Silva
Alexander Aksenov
Alan Jarmusch
Julia Gauglitz
And visitors

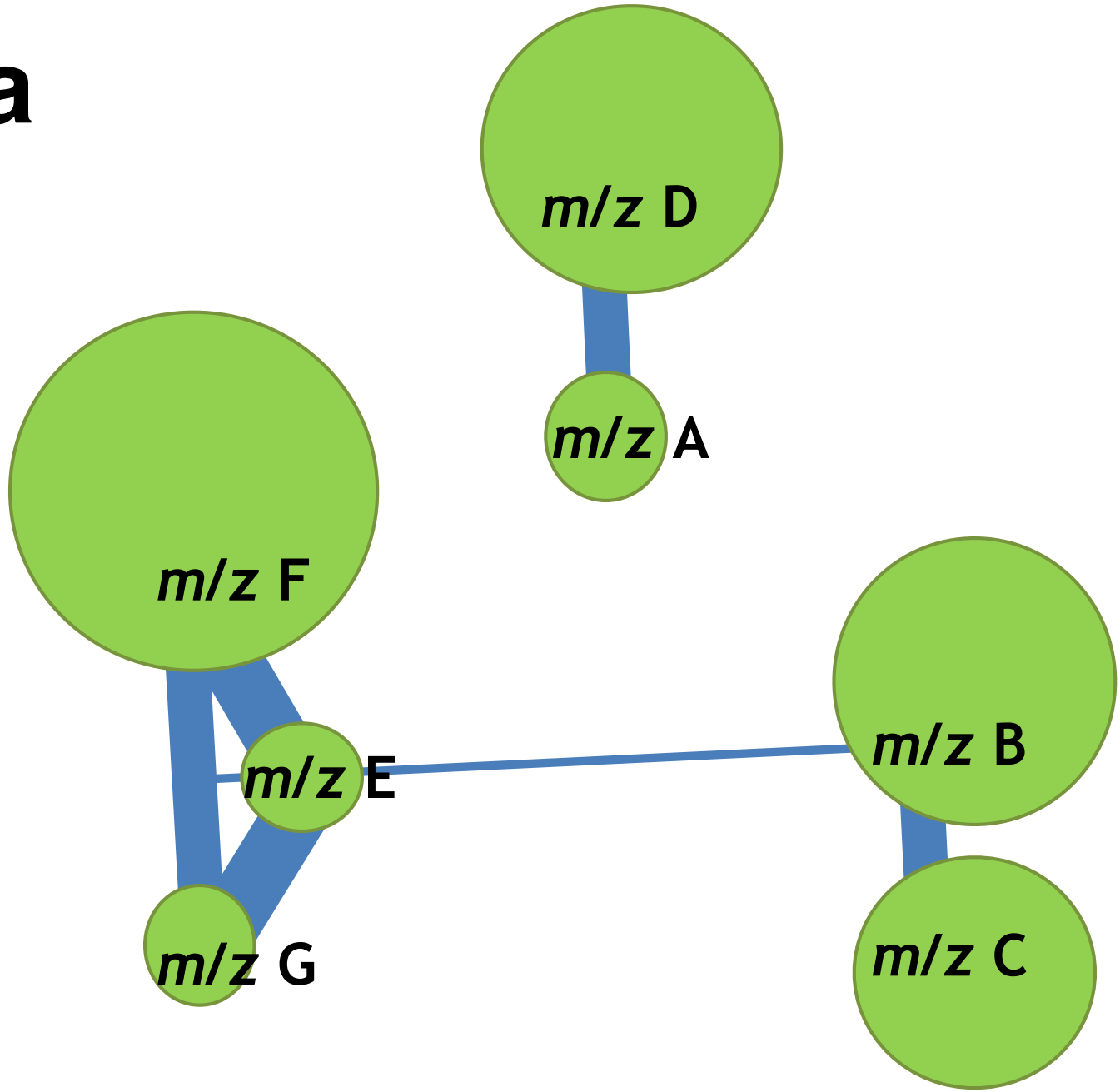
Thanks to reviewers of grants that have supported forward looking ideas

Current Support:
Bruker, NIH, NSF, USDA
EU 7th framework and horizon 2020
Janssen, Bayer, Colgate
Sloan and Moore foundations
UCSD, Department of Justice, Office of Naval Research, Science without Borders, lululemon, CMI

Other collaborators and co-workers were highlighted throughout the talk

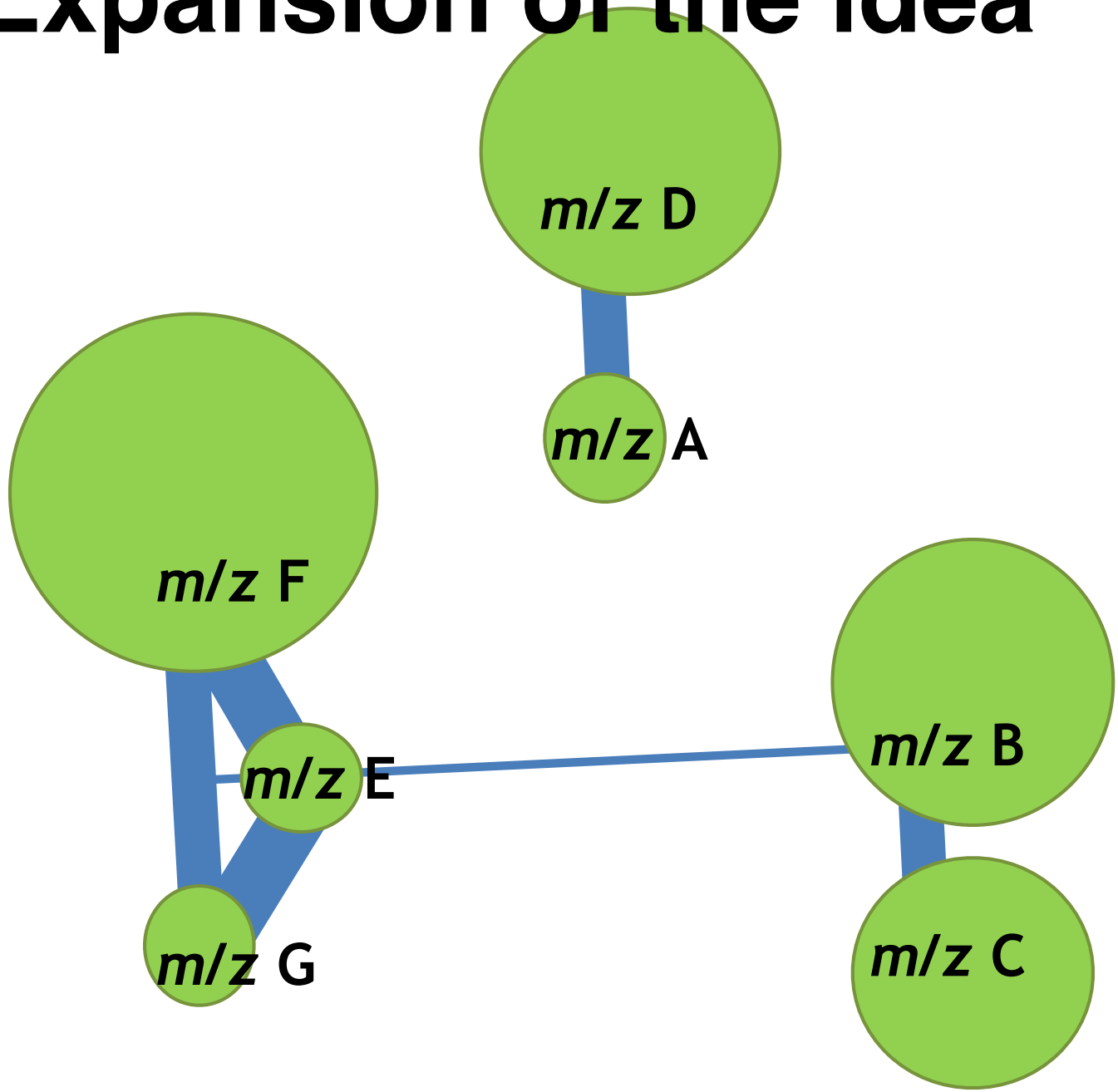

m/z A B C D E F
Activity
(EC50 in µg/ml)
Fraction 1
Fraction 2
Fraction 3
Fraction 4
Fraction 5
Fraction 6
Fract….
Etc.
m/z
= not detected/no activity
= most intense/active

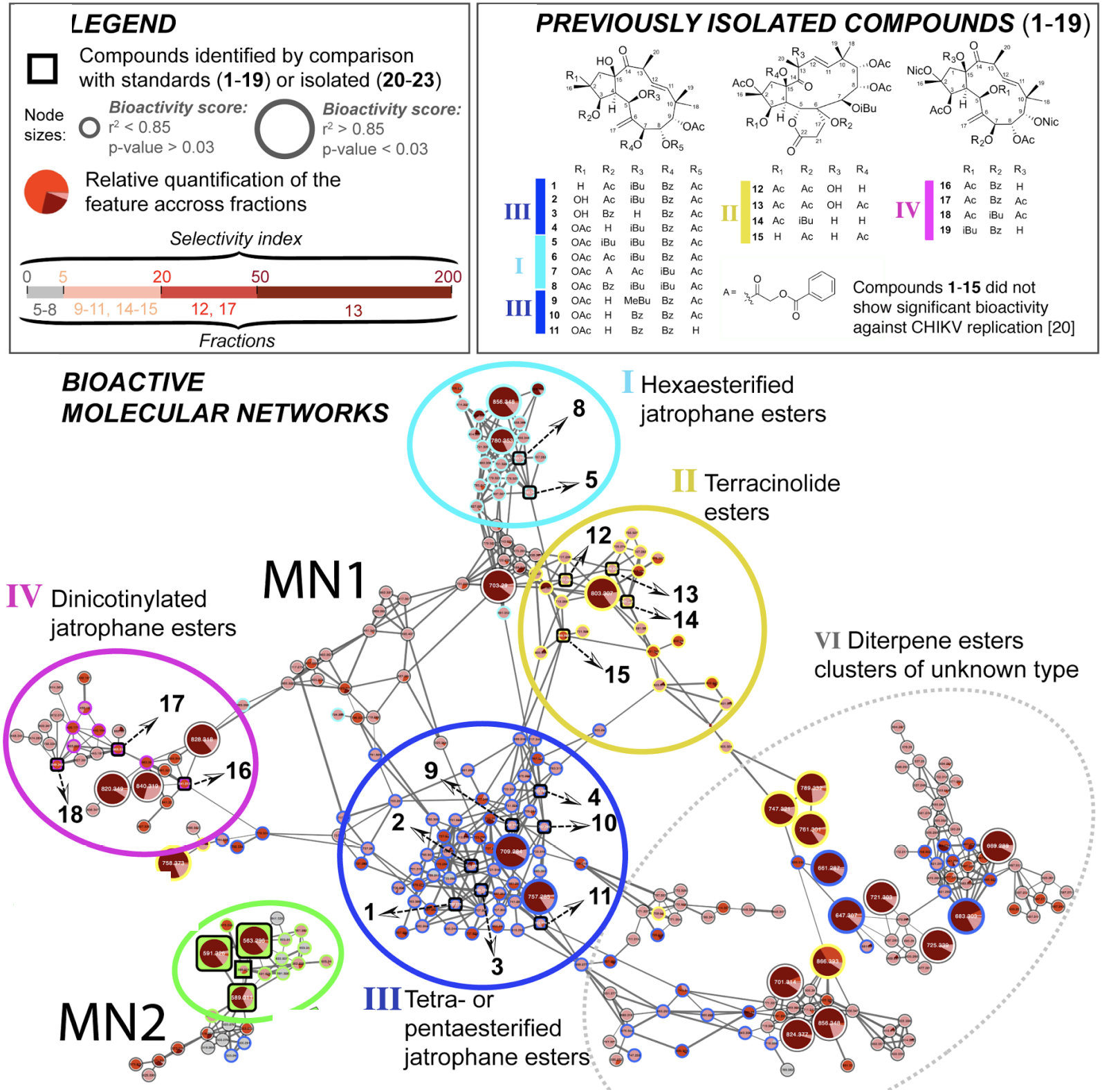
Idea



Problem - many ions observed in the three active fractions

Expansion of the idea




LEGEND
Compounds identified by comparison with standards (1-19) or isolated (20-23)
Node sizes:
Bioactivity score:
r² < 0.85
p-value > 0.03
Bioactivity score:
r² > 0.85
p-value < 0.03
Relative quantification of the feature accross fractions
Selectivity index
0
5
20
50
200
5-8
9-11, 14-15
12, 17
13
Fractions
PREVIOUSLY ISOLATED COMPOUNDS (1-19)
III
I
III
II
IV
Compounds 1-15 did not show significant bioactivity against CHIKV replication [20]
BIOACTIVE MOLECULAR NETWORKS
MN1
MN2
I Hexaesterified jatrophane esters
II Terracinolide esters
III Tetra- or pentaesterified jatrophane esters
IV Dinicotinylated jatrophane esters
VI Diterpene esters clusters of unknown type

Natural products from food