



Research Analysis  
Platform

Enabled by **DNA**nexus®

# Integrative Analysis of UK Biobank Proteomics Data

APRIL 2023

# Speakers

---



**Prof Naomi Allen**  
UK Biobank

---



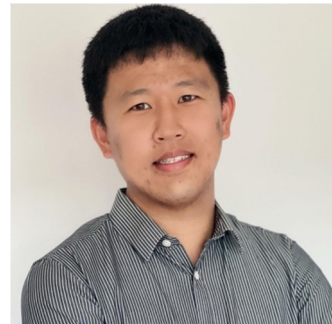
**Cindy Lawley, PhD**  
Olink Proteomics

---



**Chris Whelan, PhD**  
Janssen Pharmaceutical  
Companies of Johnson &  
Johnson

---



**Dr Benjamin Sun, MD  
PhD**  
Biogen

---



**Dr. Karsten Suhre**  
Weill Cornell Medicine -  
Qatar

---



**Ben Busby, PhD**  
DNAnexus

---

# Upcoming Webinars - Links in Related Content Section

- ▶ Oncology Researcher Roundtable:  
**May 25th**
- ▶ Analyzing the UK Biobank  
Proteomics Data on the UK  
Biobank Research Analysis  
Platform: **June 1st**
- ▶ [Find all Event announcements on  
the Community Forum](#)

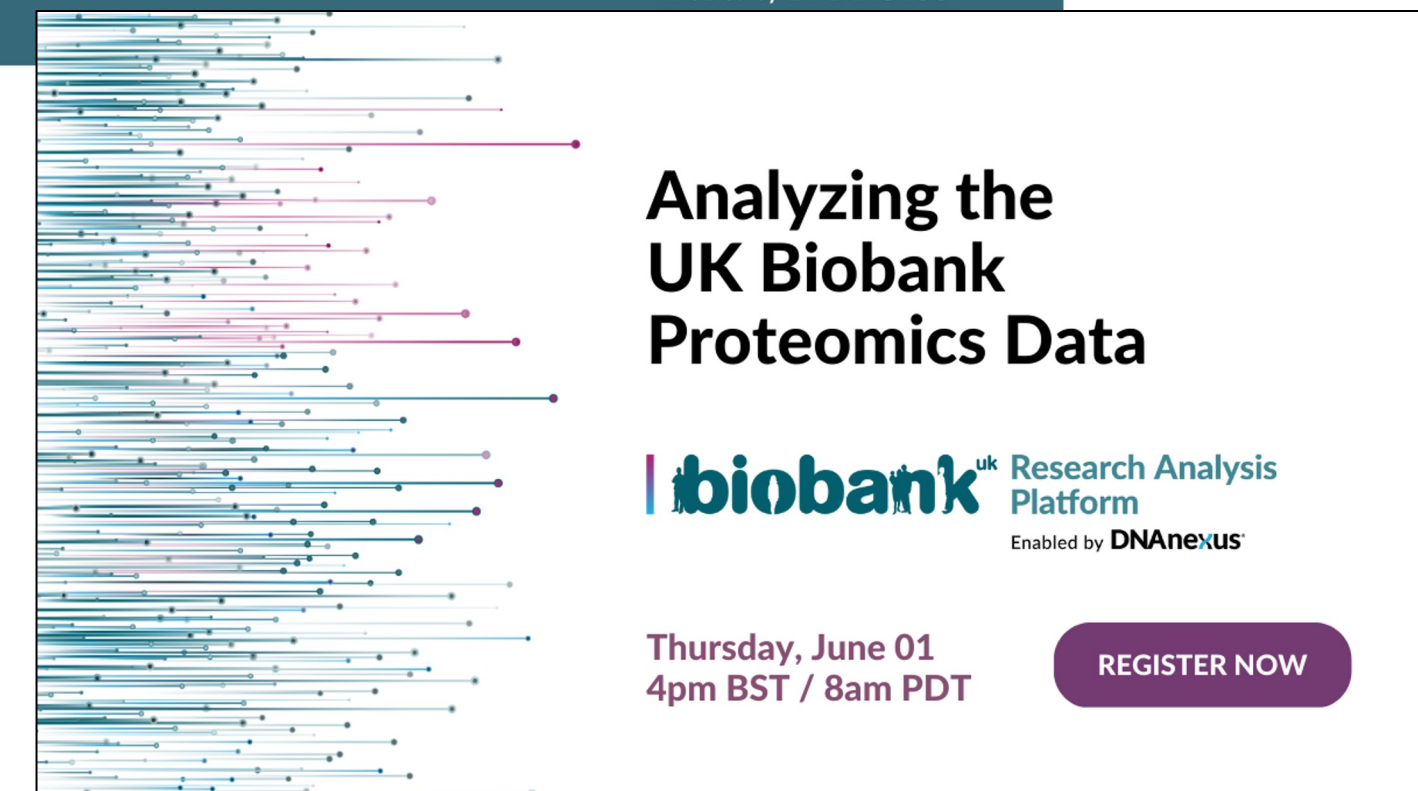


WEBINAR

**Oncology Researcher Roundtable:  
Working with Large-Scale Datasets  
to Enable Discovery**

Thursday, May 25th  
4:00 pm BST/8:00 am PDT

**biobank<sup>uk</sup>**  
Research Analysis  
Platform  
Enabled by **DNAexus**



**Analyzing the  
UK Biobank  
Proteomics Data**

**biobank<sup>uk</sup>** Research Analysis  
Platform  
Enabled by **DNAexus**

Thursday, June 01  
4pm BST / 8am PDT

**REGISTER NOW**

## Join the conversation to:



**Collaborate and connect** with your peers and colleagues and experts from the UK Biobank and DNAexus

Click [Here](#) to Join 

OR

## On Community, you can:



**Search and Discuss:** You can browse specific topics, keywords, or questions and exchange helpful tips and ideas with your peers and colleagues



**Get Early Access:** As a Community member, you get first and early access to all DNAexus webinars, trainings, and roundtable discussions



**Stay Informed:** You can learn the latest information and news on DNAexus and the Research Analysis Platform



# UK BIOBANK PLATFORM CREDITS PROGRAM

APPLY  
TODAY

Visit the  
[Program FAQs](#)

Visit  
[Website](#)

Apply here:  
[Application](#)



## ABOUT THE UK BIOBANK PLATFORM CREDITS PROGRAM

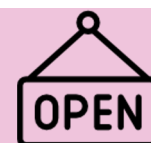
The UK Biobank Platform Credits Program is a courtesy of AWS. The program is available to all eligible researchers and is designed to allow researchers to explore the UKB-RAP in detail, develop and test tools and methods, and undertake analysis to support their research project. Credits can be used to cover costs of compute and storage above £40 credits provided by DNAnexus.

### AM I ELIGIBLE?

UK Biobank defines early career researchers as “an individual within an academic institution within four years of the award of their PhD or equivalent professional training, or within four years of starting their first academic appointment (full-time or part-time), excluding career breaks”. Early career researchers also include those bona fide students eligible for reduced Access fees.

Researchers in low- and middle income countries eligible for reduced Access Fees will also be able to participate in this program.

SCAN TO APPLY!

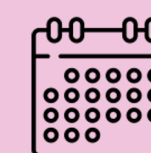


Open to all eligible early career researchers and researchers from [low/low-middle income countries](#)



Two Types of Available Funds

- 1: **Getting Started Grants**-Available for all researchers (one per application)
- 2: **Grant Enhancements**- Requested when compute plan is defined



Credit availability period:  
2022-2024

# UKB-RAP Accelerator

- ▶ Allows UKB researchers to leverage the specific UKB-RAP expertise of DNAnexus to help **navigate the rich data** of UK Biobank, **develop tools** to generate hypotheses and insights, and get the most out of the UKB-RAP
- ▶ Packages of professional service credits, live training, 1:1 consulting and customer support delivered by the **world's leading UKB-RAP experts** who are also experts in bioinformatics, data science, and data engineering
- ▶ Three **flexible package options** to meet the needs of small or pilots projects up through large, complex, enterprise endeavors
- ▶ **Customize packages** to support: GWAS, clinical data analysis, imaging studies, multi-modal data analysis, machine learning and more

## ▶ Managing UKB-RAP Research Just Got Easier



Interested to learn more? Let us know and we'll send you the details

<https://hubs.ly/Q01Fwd9D0>

# Promote UKB-RAP Community Research!

Get involved in upcoming UKB-RAP Events

Submit Your Application: [bit.ly/ukbrap-research](https://bit.ly/ukbrap-research)

Opportunities to:

- ▶ Present in Upcoming Webinars & Researcher Roundtables
- ▶ Highlight your work in the Newsletter Spotlight
- ▶ Organize a Meetup or workshop related to your interest
- ▶ & More!

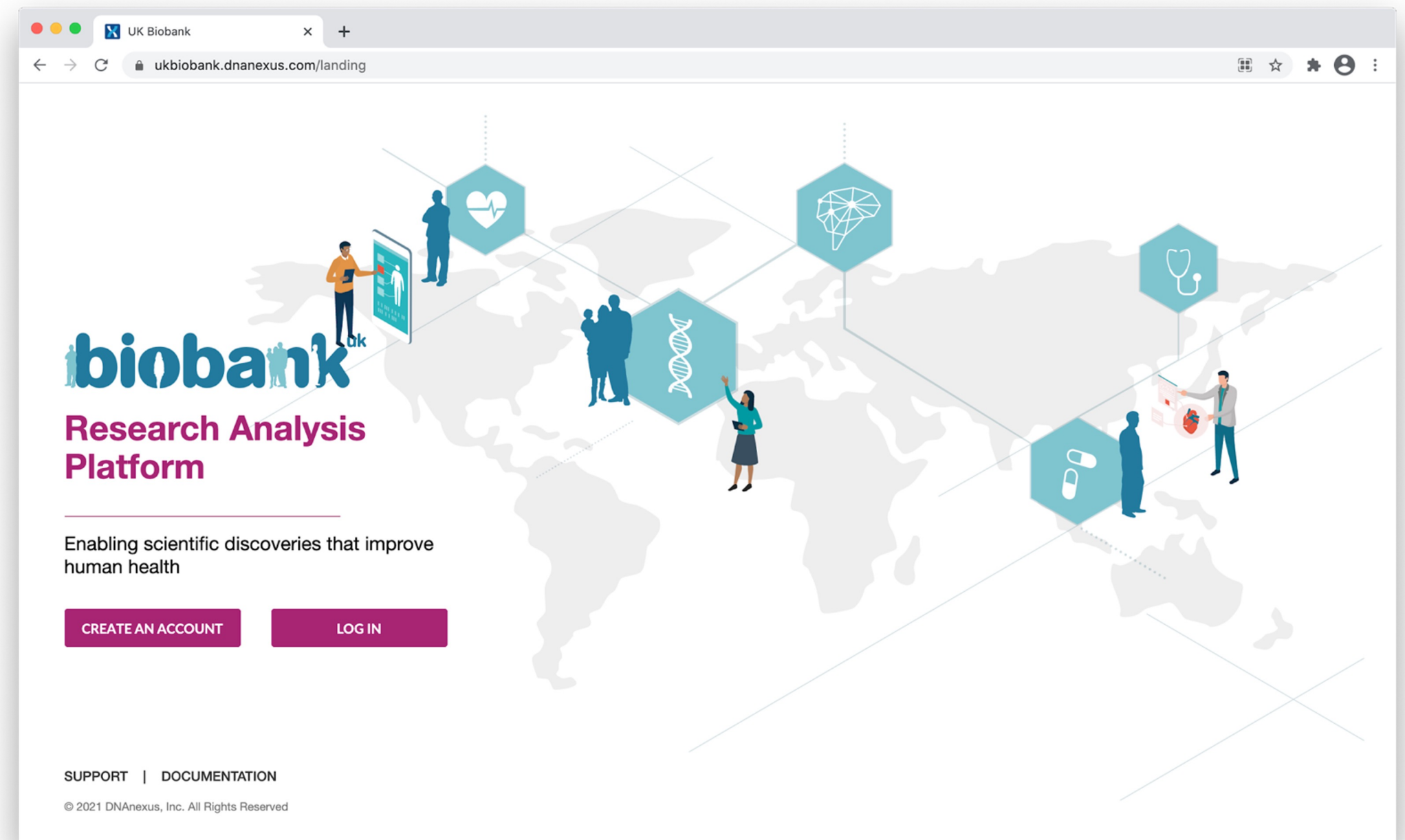


# UK Biobank Research Analysis Platform



Join Now & Receive £40 Credit for Working on the Platform!

- ▶ Register: [bit.ly/ukbrap](https://bit.ly/ukbrap)
- ▶ An all-in-one platform that comprises cloud infrastructure, analysis tools, and UK Biobank data
- ▶ Cloud-based infrastructure allows democratization of access to the data
- ▶ Differentiated, secure data access for users from anywhere in the world





# Overview of UK Biobank

Prof Naomi Allen  
Chief Scientist, UK Biobank  
[naomi.allen@ndph.ox.ac.uk](mailto:naomi.allen@ndph.ox.ac.uk)

## UK Biobank... A unique combination of...



**Size:** 500,000 participants

**Depth:** genomics, biomarkers, lifestyle, imaging, etc.

**Duration:** 15 years of follow-up of health outcomes

**Accessibility:** +30,000 researchers worldwide

**Perception...?**

**Reality**



## Lifestyle questionnaire



*N~500,000*

## Biomarkers



*N~500,000*

## Imaging



*N~100,000*

## Genomics



*N~500,000*

## Physical measures



*N~500,000*

## Follow-up of health via linkage



*N~500,000*





- Easy accessibility of the data to the global research community



interest and opportunities to enhance the resource further

- High scientific added-value of including sequencing and other –omic data at-scale
  - E.g., accelerating drug discovery and development; risk prediction and stratification
- Cohort-wide assays require deep pockets: need for funding consortia
- Establishment of new models of pre-competitive collaboration
- Short period of exclusive access to assay data before they are made available to the broader research community

# Exome and whole-genome sequencing for 500,000 participants



## Whole Exome Sequencing: all 500,000

- Pre-competitive consortium of industry partners
- First 50,000 made available in 2019
- Full cohort made available mid-2022



## Whole Genome Sequencing: all 500,000

- Public-private partnership
- First 200,000 made available Q4 2021
- Full cohort to be made available Nov 2023

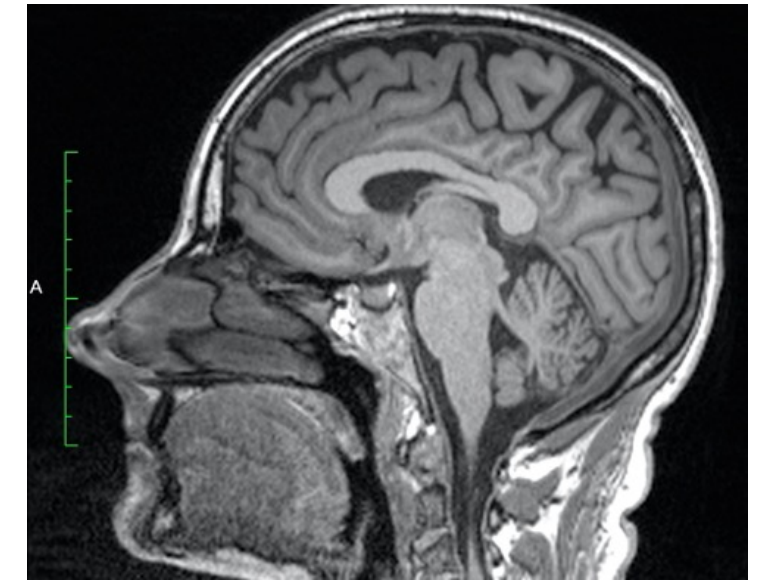


- Key clinical biomarkers available
- Transcriptome, proteome and metabolome; assays have started in each of these areas
- Increasing momentum due to an increasingly wide range of potential sources of external investment
- Successful pilots leading to whole cohort projects
- Combination of complementary **targeted** -omics
- Progression to untargeted -omics, and possibly epigenetics, immune system biomarkers...

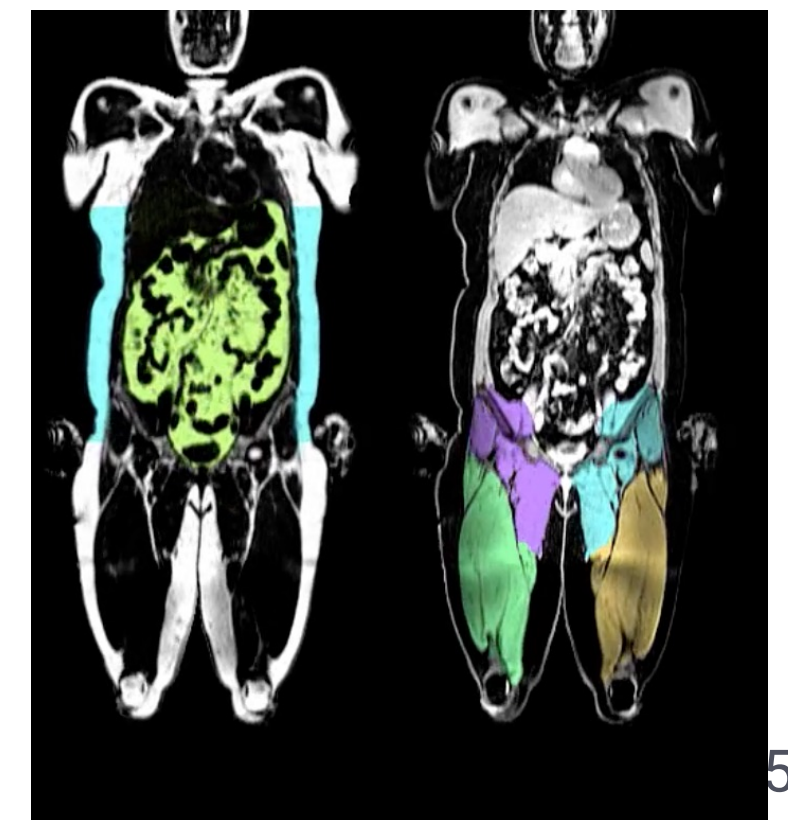
	Now	Future?
Transcriptome	 5,000 scRNA Sequencing (in progress)	60,000 participants scRNA sequencing
Proteome	 57,000 1.5k proteins <i>3k proteins (Nov)</i>	500,000 Targeted <i>Olink, Somalogic...</i>
Metabolome	 500,000 ~200 metabolites (in progress)	500,000 Targeted <i>Biocrates, Lipidyzer...</i>
		500,000 Untargeted MS



- ~100k participants to be imaged by end-2024
- Apr 2023: 70,000 participants scanned
- Multi-modal protocols
  - Brain, cardiac and abdominal MRI
  - Carotid ultrasound
  - DEXA



- Repeat imaging study of 60,000 participants now started
- Full repeat of baseline + additional samples + OCT eye measures
- Longitudinal measures to assess changes over time





15 years of follow-up via cohort-wide linkage to key NHS electronic health care records :

- Deaths
- Cancers
- Inpatient hospital admissions (including critical care)
- SARS-CoV-2 testing data
- Primary care (for 45% of cohort until 2016/7)
  - Ongoing efforts to obtain data

Condition	Year of diagnosis		
	Observed	Predicted	
	2020	2027	2032
Diabetes	31,000	54,000	70,000
Myocardial infarction	15,000	30,000	46,000
Stroke	12,000	25,000	37,000
COPD	25,000	47,000	65,000
Depression	25,000	39,000	47,000
Breast cancer	9,000	14,000	18,000
Colorectal cancer	5,000	8,000	11,000
Lung cancer	4,000	6,000	8,000
Prostate cancer	10,000	16,000	20,000
Hip fracture	5,000	13,000	22,000
Rheumatoid arthritis	4,000	6,000	8,000
Parkinson's disease	4,000	10,000	14,000
Alzheimer's disease	5,000	17,000	37,000





- **SCALE:** 500,000 diverse individuals aged 40-69 years when they joined the study in 2006-10
- **DEPTH:** Exquisite detail about lifestyle, environment and medical history supplemented by an extensive range of biological assays (haematology, biochemistry, genetics, -omics) as well as imaging
- **DURATION:** ~15 years of follow-up has already yielded very large numbers of many different health outcomes
- **ACCESSIBILITY:** Rapidly increasing number of different types of researcher globally (already ~30,000) are using UK Biobank for a wide range of discovery science (1900+ papers in 2022 alone)

***.... and the best is yet to come!***

# Acknowledgements

**UK Biobank:** Executive team and Coordinating centre staff, Strategic Oversight Committee, International Scientific Advisory Board, Scientific Working Groups, Oxford University team

**Funders:**



**And, of course, our 500k Participants:**



# Olink Overview



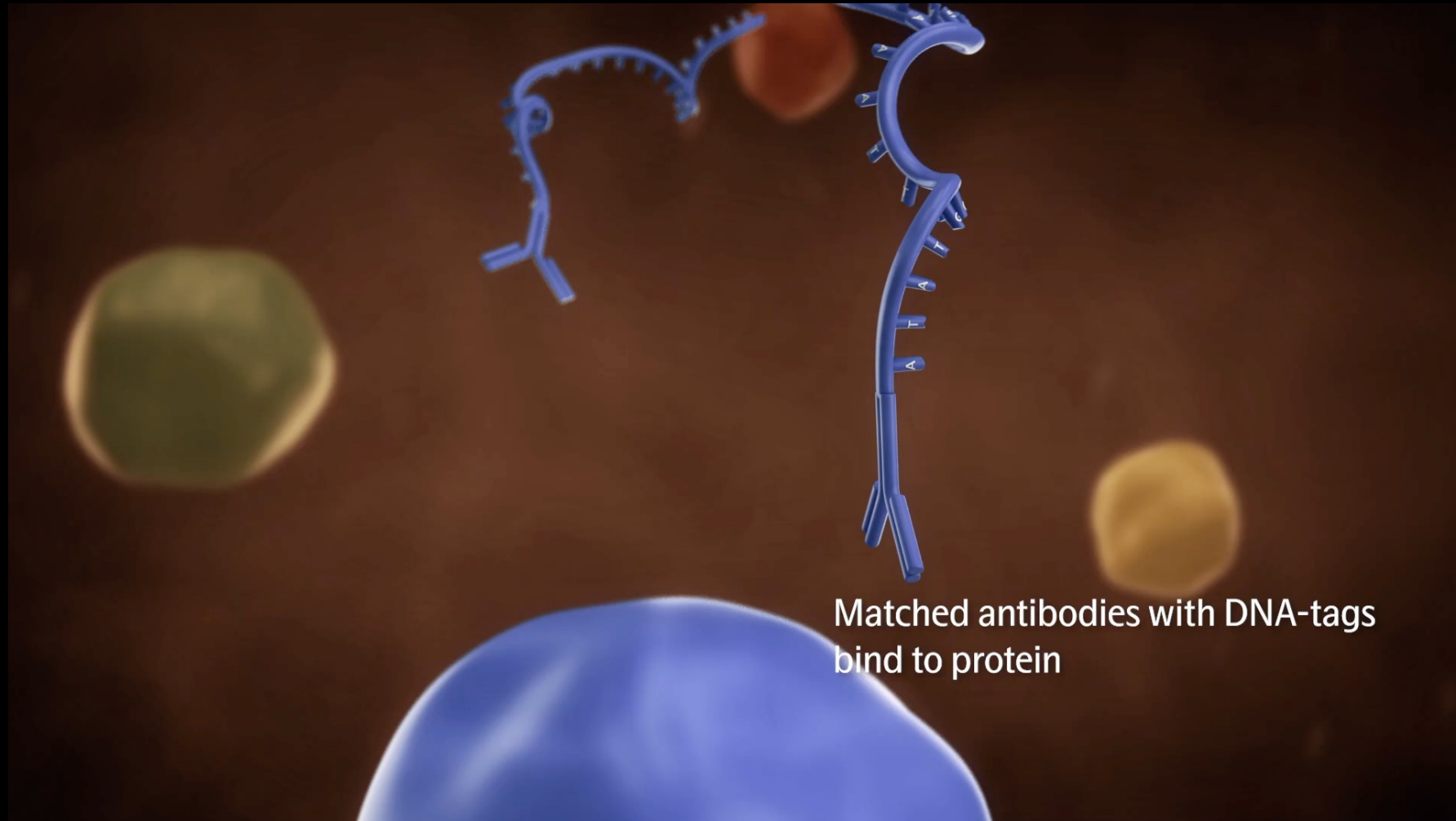
Cindy Lawley, PhD  
Sr Director, Population Health  
Olink Proteomics





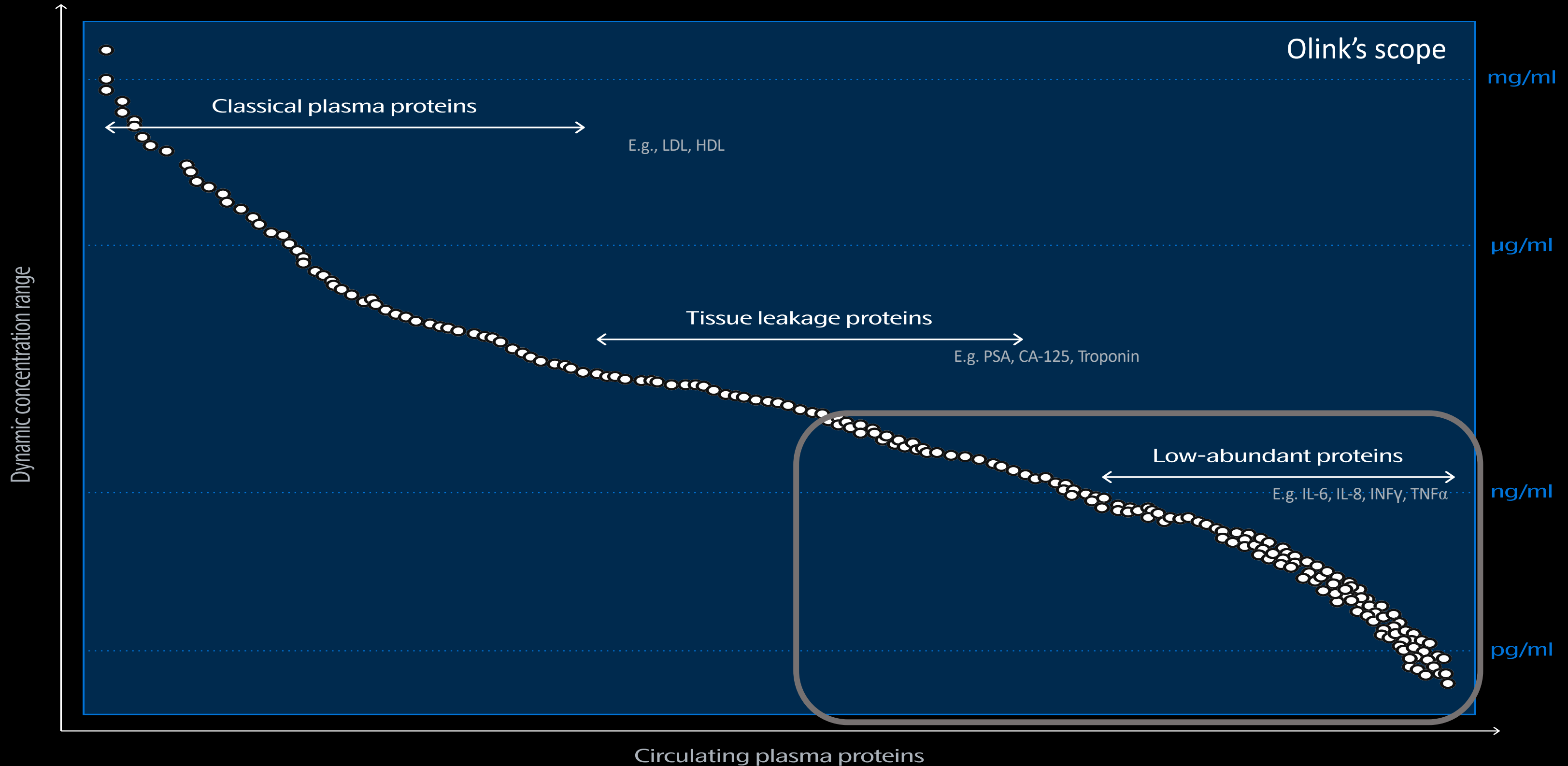
Olink®  
Accelerating proteomics together

# The Platform: Proximity Extension Assay





# Covering the broad range of the plasma proteome





# Output File – NPX Data File

Flagged sample (did not pass QC)

	A	B	C	D	E	F	G	H	I	J	K	L
1	SampleID	Index	OlinkID	UniProt	Assay	MissingFreq	Panel	Panel_Version	PlateID	QC_Warning	LOD	NPX
2	A1	1	OID20321	O00584	RNASSET2	0.03	CARDIOMETABOLIC	1	Project1_Plate1_INF	PASS	1.4223	4.3509
3	A1	1	OID20206	Q07108	CD69	0.04	CARDIOMETABOLIC	1	Project1_Plate1_INF	PASS	1.0986	3.0382
4	A1	1	OID20195	P35754	GLRX	0.06	CARDIOMETABOLIC	1	Project1_Plate1_INF	PASS	2.0558	3.5265
5	A1	1	OID20094	Q9H5Y7	SLITRK6	0.05	CARDIOMETABOLIC	1	Project1_Plate1_INF	PASS	1.5126	2.5526
6	A1	1	OID20216	P31431	SDC4	0.03	CARDIOMETABOLIC	1	Project1_Plate1_INF	WARN	0.95	3.2493
7	A1	1	OID20106	P07585	DCN	0.03	CARDIOMETABOLIC	1	Project1_Plate1_INF	PASS	0.508	1.7733
8	A1	1	OID20324	O75023	LIURB5	0.03	CARDIOMETABOLIC	1	Project1_Plate1_INF	PASS	0.8072	6.3634
9	A1	1	OID20299	P48745	CCN3	0.03	CARDIOMETABOLIC	1	Project1_Plate1_INF	PASS	0.7146	1.7109
10	A1	1	OID20381	P07359	GP1BA	0.03	CARDIOMETABOLIC	1	Project1_Plate1_INF	PASS	-0.1536	3.9579
11	A1	1	OID20187	P41159	LEP	0.13	CARDIOMETABOLIC	1	Project1_Plate1_INF	PASS	0.7591	3.6306
12	A1	1	OID20202	Q9UK05	GDF2	0.04	CARDIOMETABOLIC	1	Project1_Plate1_INF	PASS	1.1128	3.2225
13	A1	1	OID20180	Q9GZM7	TINAGL1	0.03	CARDIOMETABOLIC	1	Project1_Plate1_INF	PASS	0.6614	4.0904
14	A1	1	OID20099	Q8NC01	CLEC1A	0.09	CARDIOMETABOLIC	1	Project1_Plate1_INF	PASS	1.0432	1.2994
15	A1	1	OID20281	Q86U17	SERPINA11	0.03	CARDIOMETABOLIC	1	Project1_Plate1_INF	PASS	0.1092	7.3305

Frequency of data < LOD

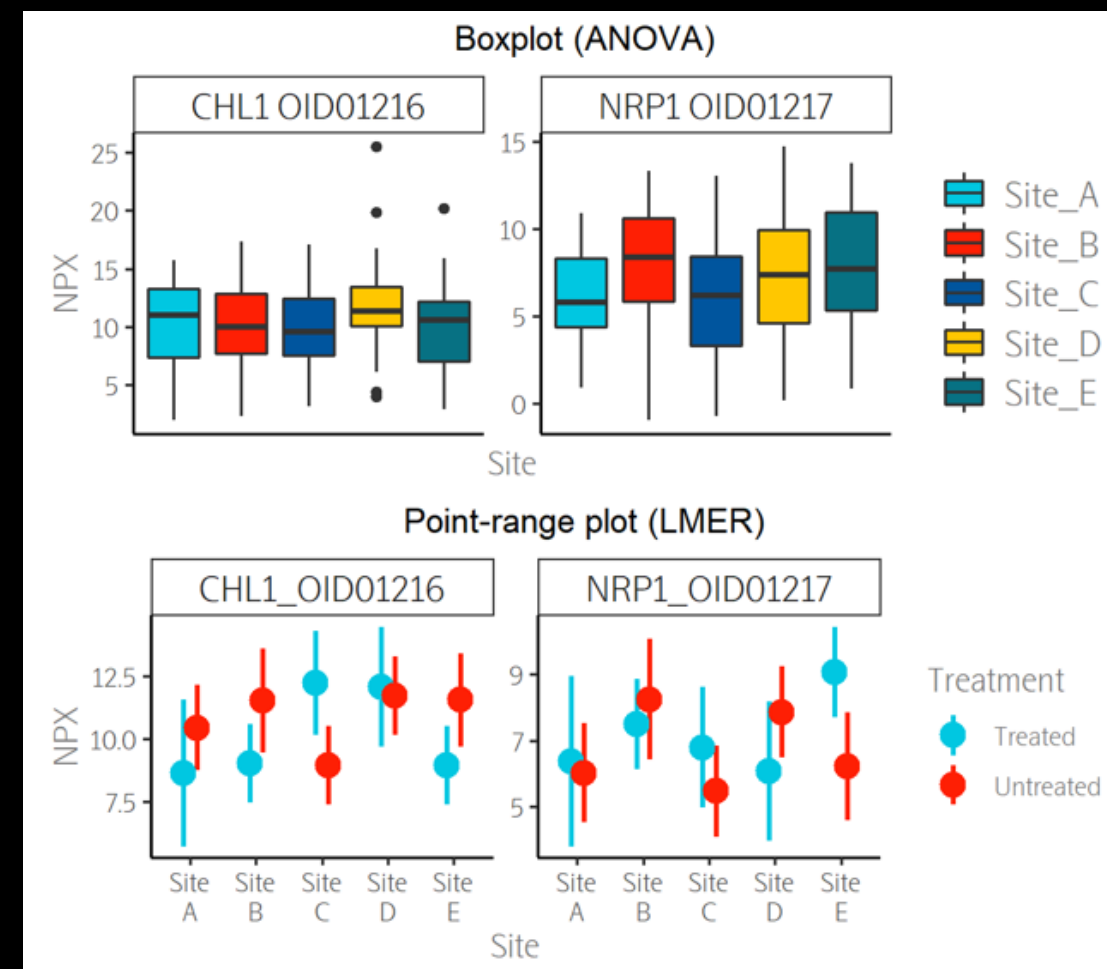
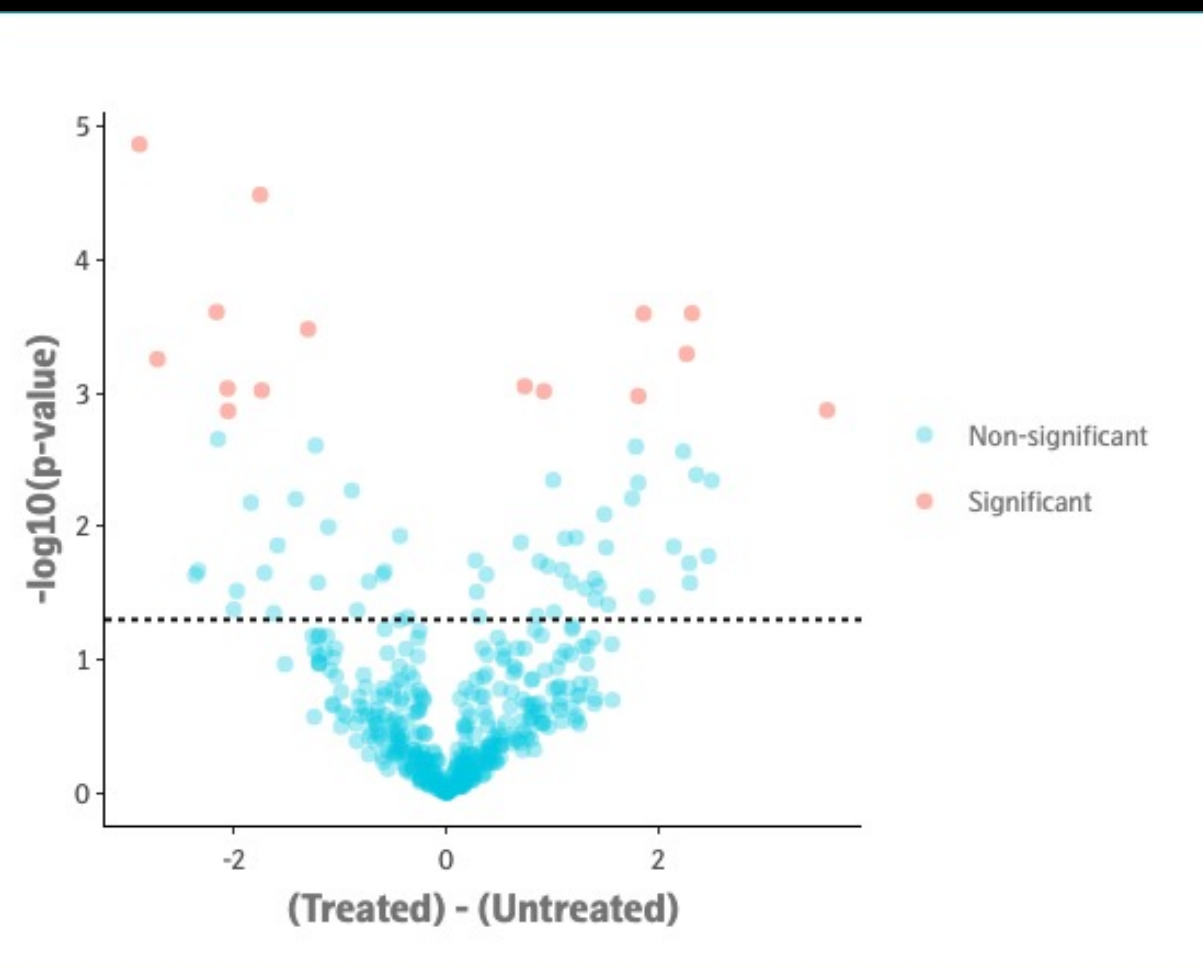
LOD values for the assay

NPX value for sample/assay



Olink®  
Accelerating proteomics together

# Proteomics Analysis Support Tools



### Proteins (471)

UNIPROT ID	GENE	PANELS	LIST
P14210	HGF	●●●●●	TruSight500
P15692	VEGFA	●●●●●	TruSight500
P22301	IL10	●●●●●	TruSight500
O75144	ICOSLG	●●●●	TruSight500
P05412	JUN	●●●●	TruSight500
P09038	FGF2	●●●●	TruSight500

Olink® Statistical Analysis App

<https://olink.com/products-services/data-analysis-products/olink-statistical-analysis-app/>

The Olink R Package

<https://cran.r-project.org/package=OlinkAnalyze>

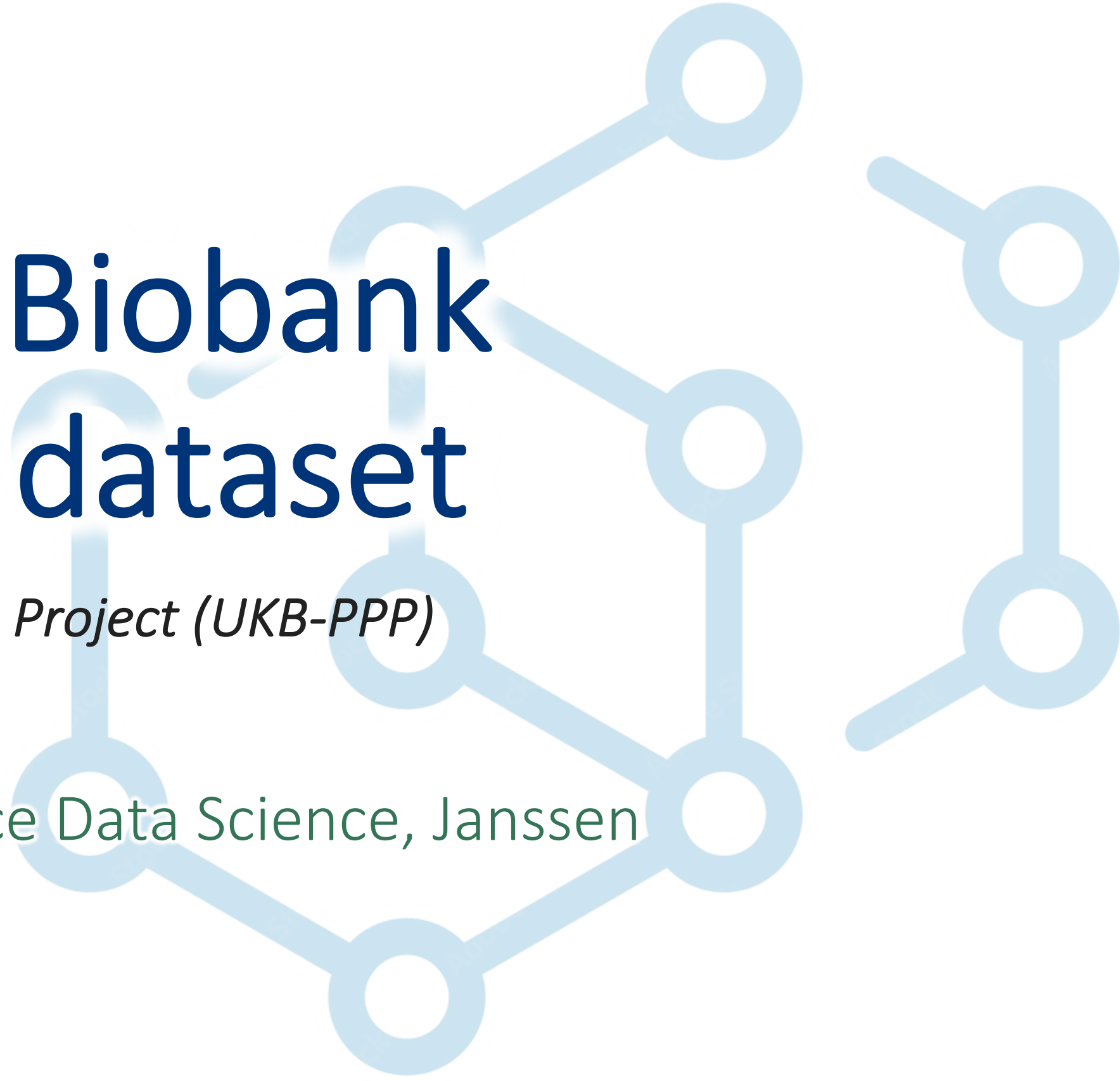
Insight Pathway Browser

<https://insight.olink.com/pathway-browser>

# Introducing the UK Biobank plasma proteomics dataset

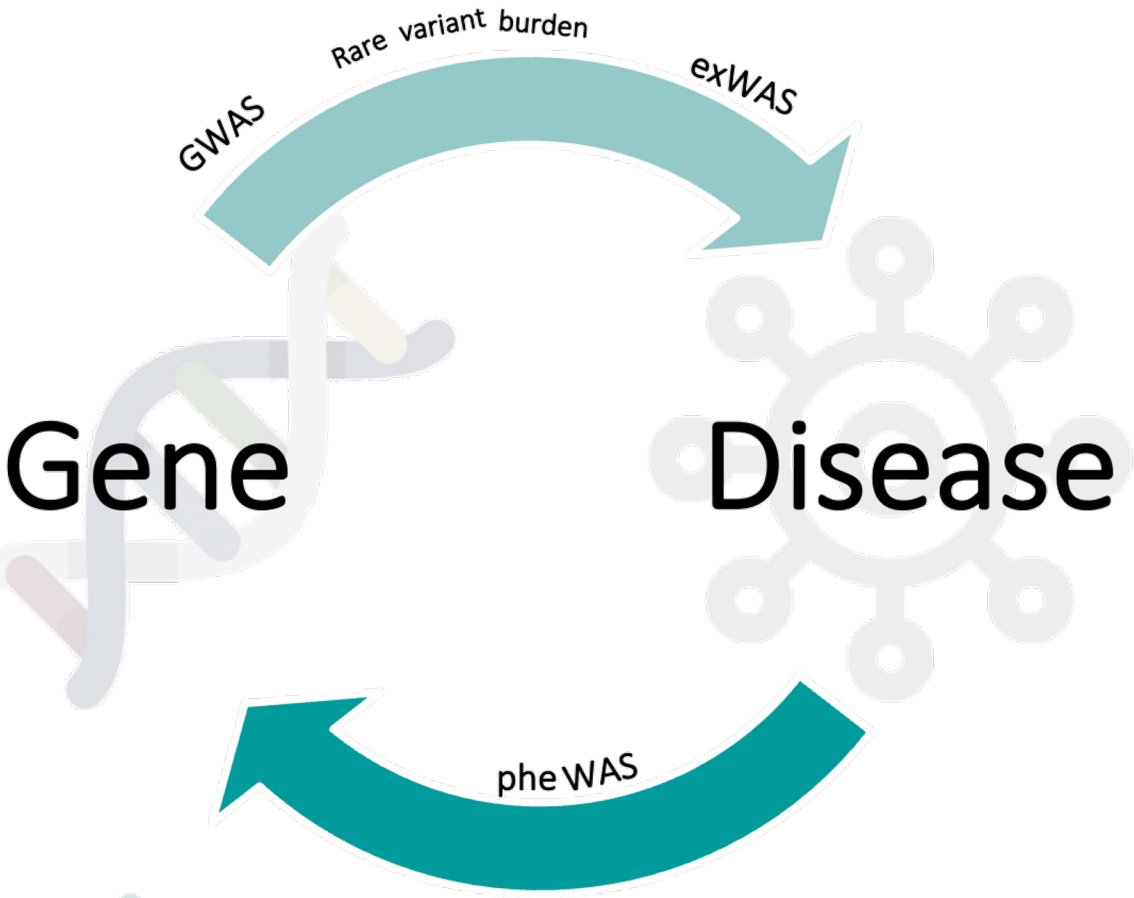
*Facilitated via the UK Biobank Pharma Proteomics Project (UKB-PPP)*

Chris Whelan, Ph.D. | Director, Neuroscience Data Science, Janssen

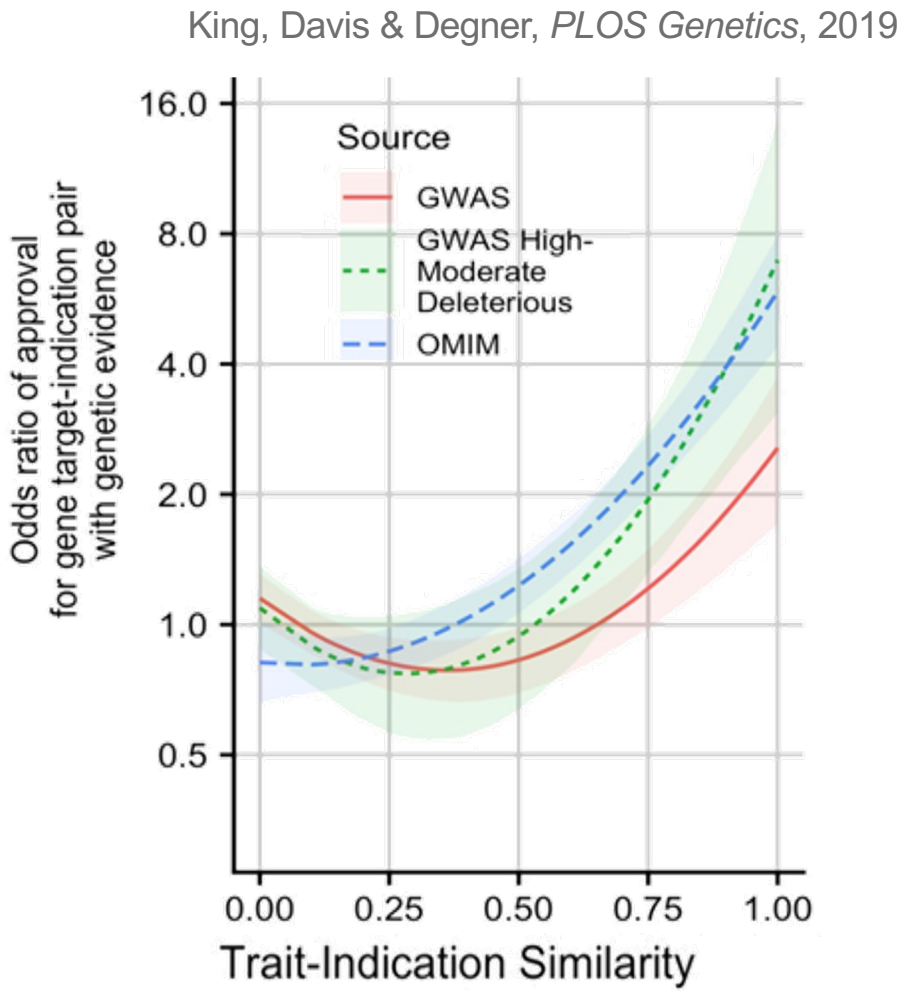




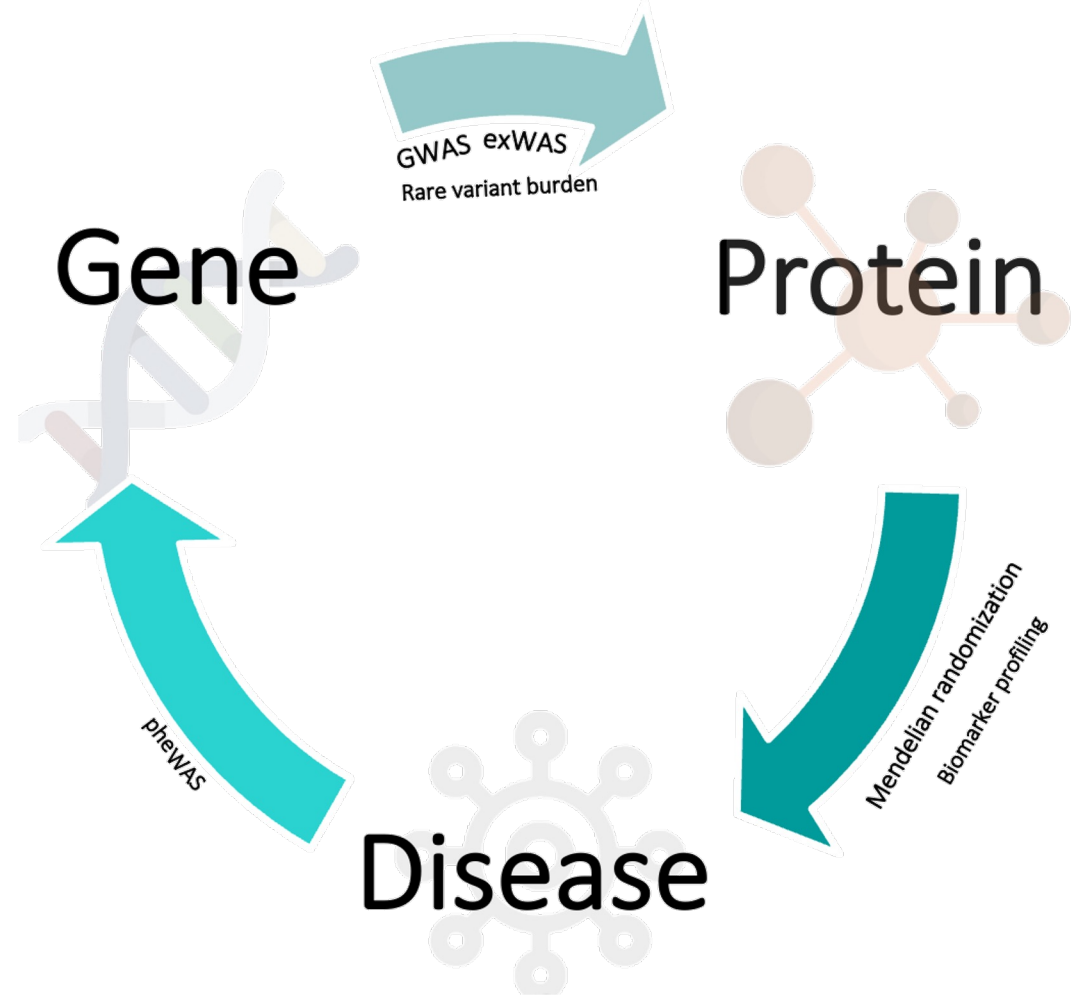
# Proteomics could accelerate **Genetics-Guided Drug Development (G2D2)**



*UK Biobank's breadth & depth has facilitated systematic **G2D2***



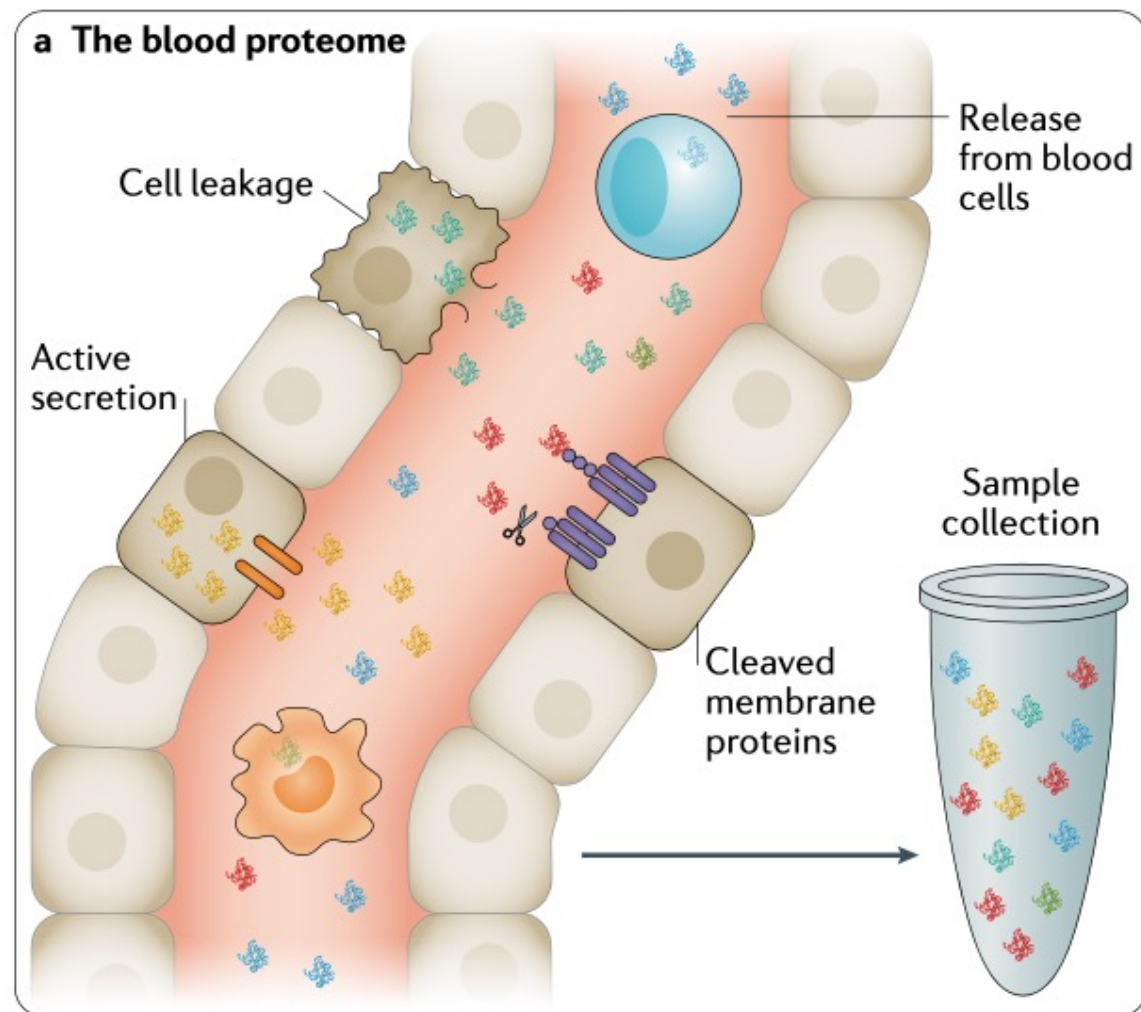
*Genetics is a promising but imperfect drug discovery tool*



*Measuring drug targets (i.e., proteins) directly could enhance **G2D2***

# Proteins are the building blocks of life

(and drug development)



<sup>1</sup>Suhre et al., *Nature Reviews Genetics*, 2020

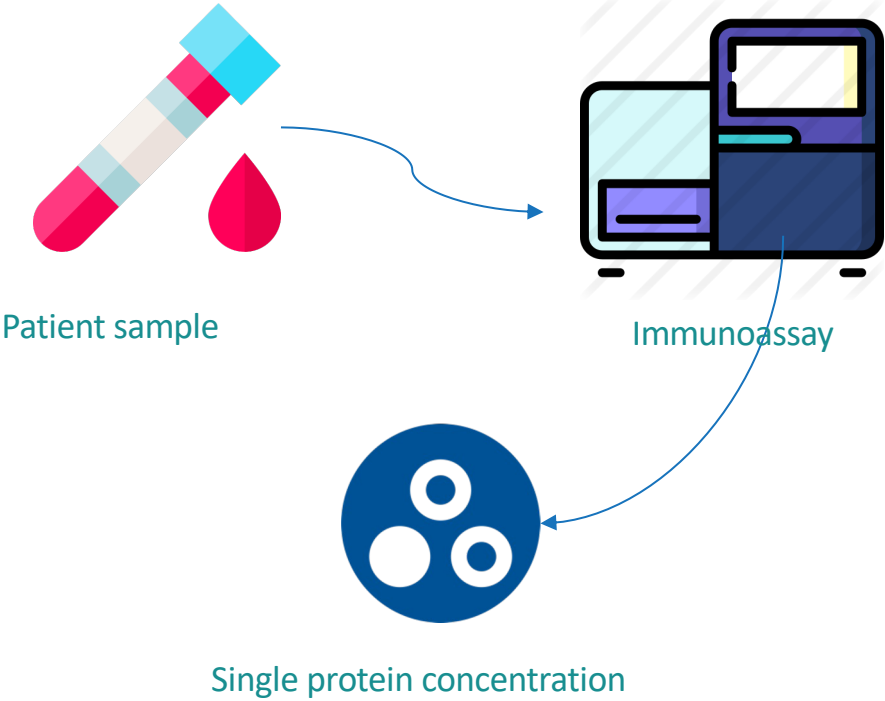
Measuring proteins appearing in circulation, due to active secretion or cellular leakage, can offer a window into the state of human health.

- Provides a basis for **diagnostics** and **therapeutics**
- e.g. **plasma p-tau-217** for Alzheimer's disease; **ApoB** for familial hypercholesterolemia

Recent advances "are allowing proteomics to take its place alongside the comprehensive characterization possible for other omics approaches, such as those focused on genetic variation and RNA expression".<sup>1</sup>

# Multiplex proteomics allows us to agnostically explore the molecular underpinnings of health and disease

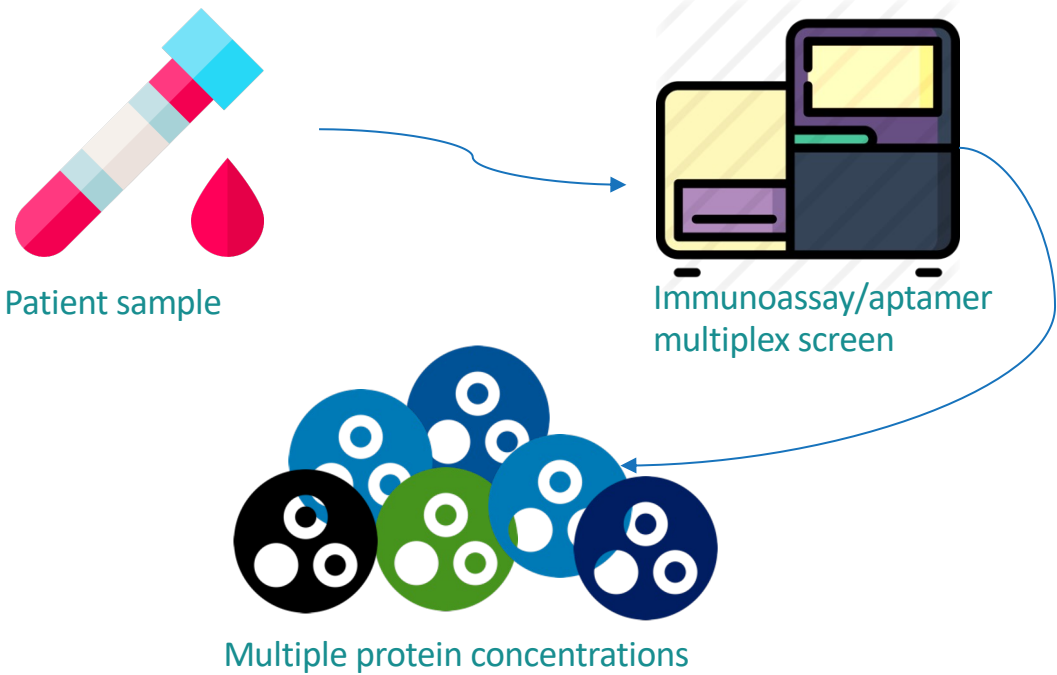
## Targeted proteomics



- **Pros:** Highly sensitive and specific
- **Cons:** Less useful for biomarker discovery, patient stratification

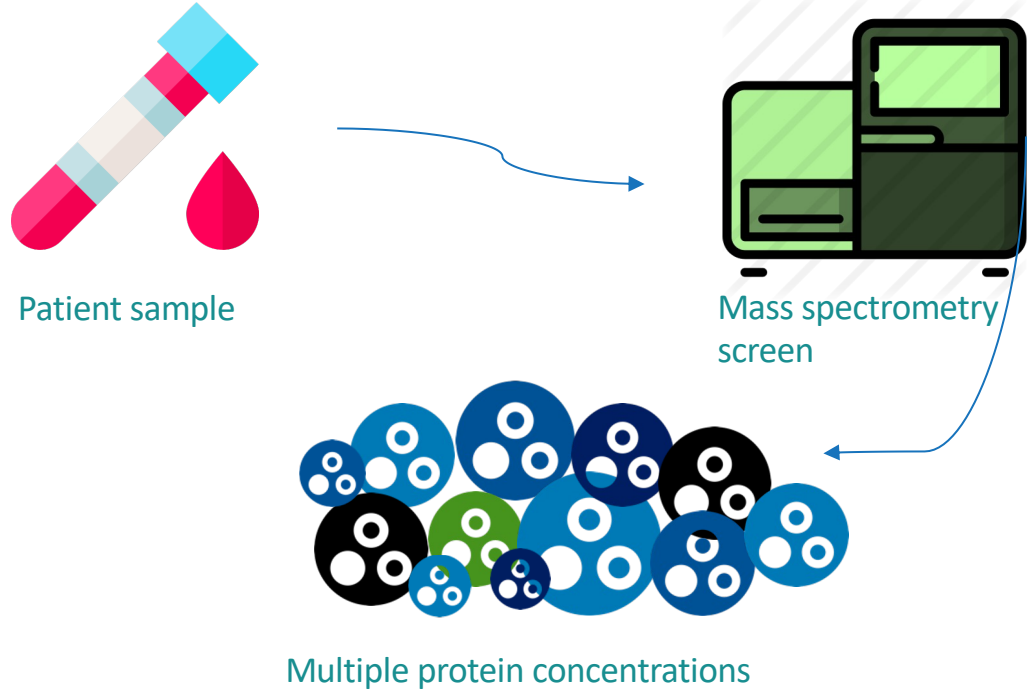
## Multiplex proteomics

### Affinity-based



- **Pros:** Detects multiple proteins with relatively high specificity & sensitivity incl. low-expressing proteins; requires ~1-5µL
- **Cons:** Somewhat biased; expensive

### Mass spectrometry-based



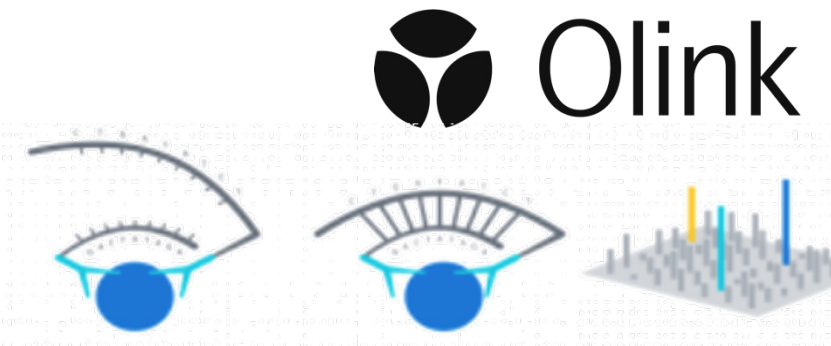
- **Pros:** Unbiased, versatile
- **Cons:** Requires large volumes; often misses low-expressing proteins, less scalable

We formed the UKB pharma proteomics project (**UKB-PPP**) to facilitate *population proteomics* & accelerate G2D2



UKB-PPP is a consortium of...

**13 pharmaceutical companies**



...funding the measurement of

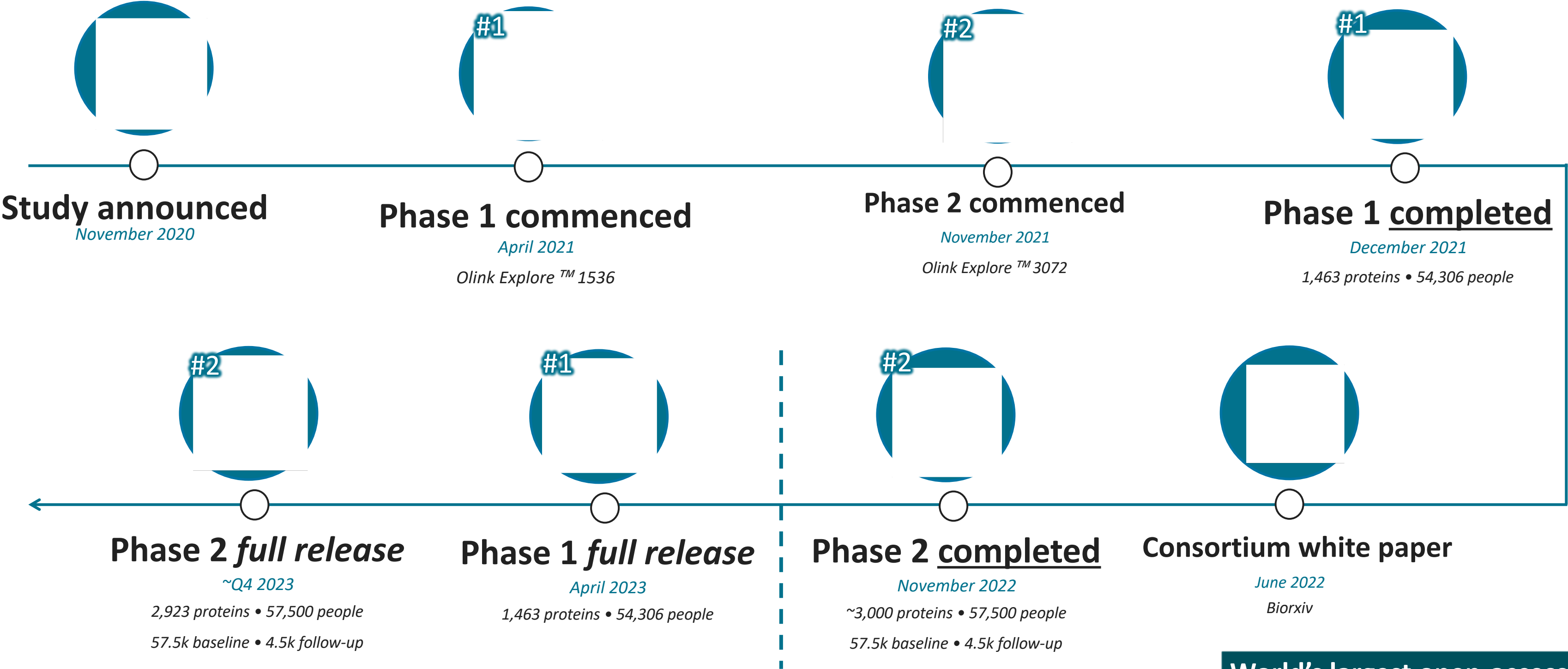
**~3,000 circulating proteins  
in 57,500 people**

...to accelerate the identification of...

**Better genetic drug targets,  
biomarkers & medicines**

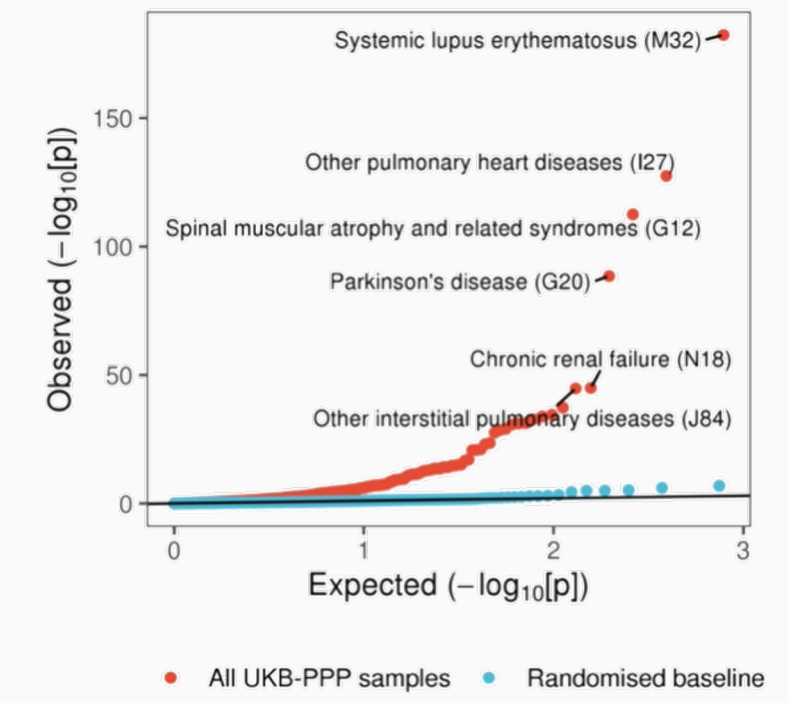
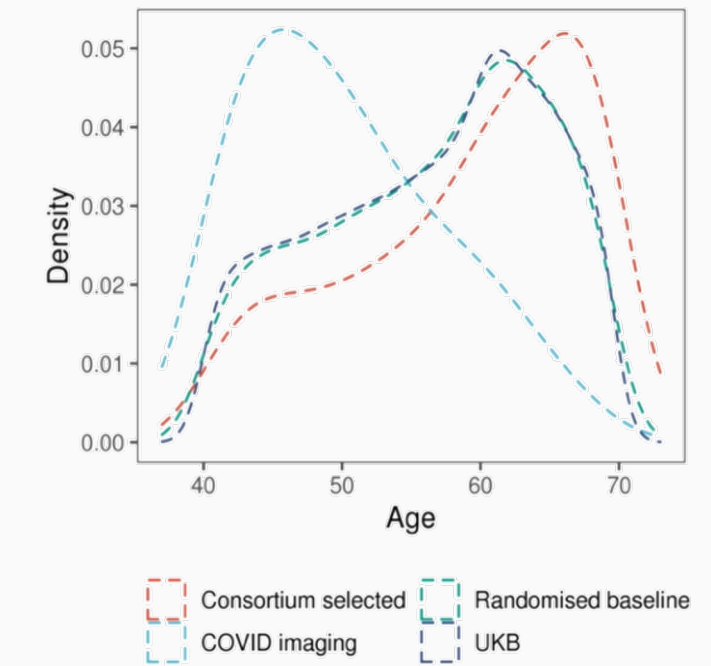
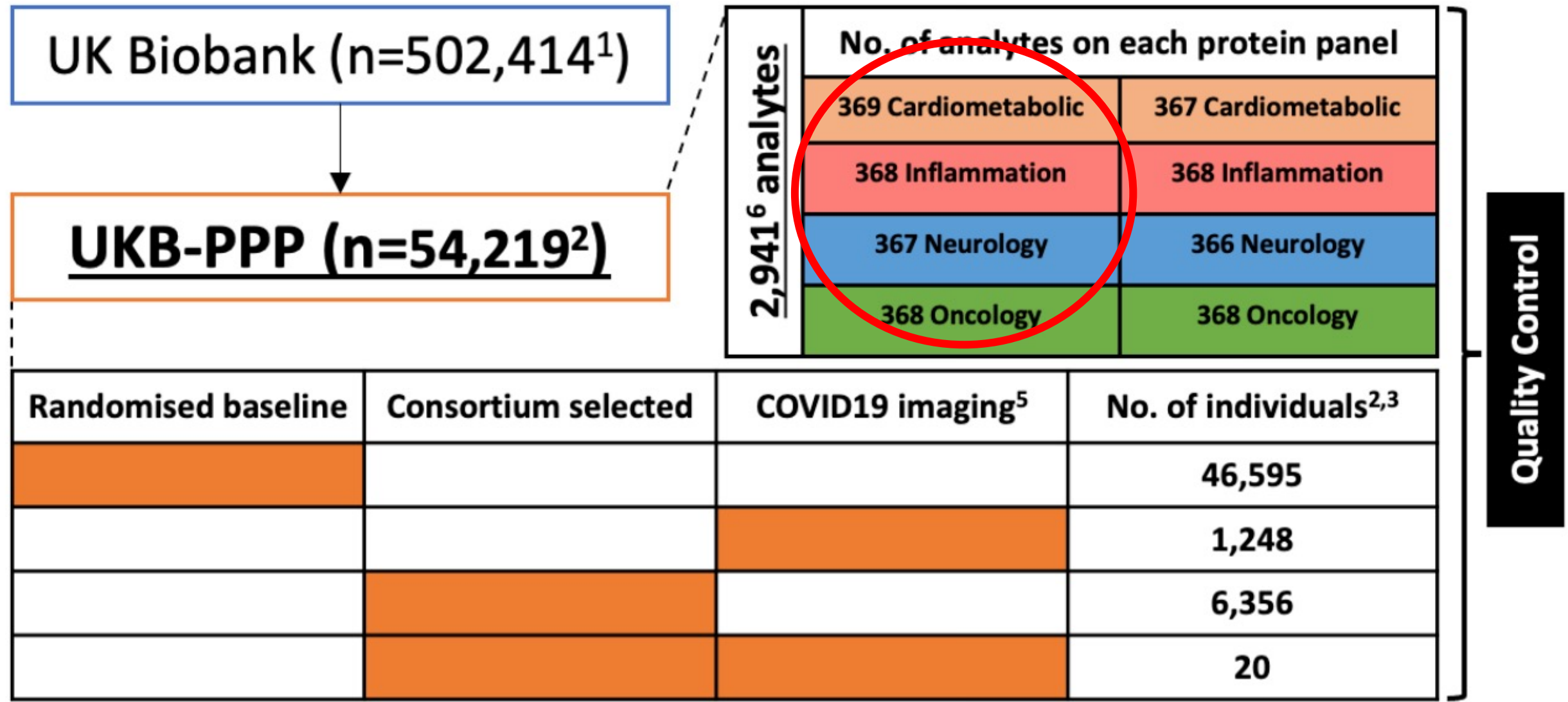


# UKB-PPP was a ~1.5-year project: Commencing in April 2021, completing in November 2022



**World's largest open-access proteogenomics dataset!**

# We applied Olink™ Explore to 54,306 UKB participants, included **randomly-selected & pre-selected samples**



# UKB-PPP partners are using these proteomic data to reveal dozens of new insights into complex diseases

In our flagship consortium manuscript, we identified **10,000+** gene variants influencing protein levels

(i.e., protein quantitative trait loci, 'pQTLs'; Sun, Whelan et al., *Under Review*)



## Genetic regulation of the human plasma proteome in 54,306 UK Biobank participants

- Benjamin B. Sun, Joshua Chiou, Matthew Traylor, Christian Benner, Yi-Hsiang Hsu, Tom G. Richardson, Praveen Surendran, Anubha Mahajan, Chloe Robins, Steven G. Vaszquez-Grinnell, Liping Hou, Erika M. Kvikstad, Oliver S. Burren, Madeleine Cule, Jonathan Davitte, Kyle L. Ferber, Christopher E. Gillies, Åsa K. Hedman, Sile Hu, Tinchu Lin, Rajesh Mikkilineni, Rion K. Pendergrass, Corran Pickering, Bram Prins, Anil Raj, Jamie Robinson, Anurag Sethi, Lucas D. Ward, Samantha Welsh, Carissa M. Willis, Alnylam Human Genetics, AstraZeneca Genomics Initiative, Biogen Biobank Team, Bristol Myers Squibb, Genentech Human Genetics, GlaxoSmithKline Genomic Sciences, Pfizer Integrative Biology, Population Analytics of Janssen Data Sciences, Regeneron Genetics Center, Lucy Burkitt-Gray, Mary Helen Black, Eric B. Fauman, Joanna M. M. Howson, Hyun Min Kang, Mark I. McCarthy, Eugene Melamud, Paul Nioi, Slavé Petrovski, Robert A. Scott, Erin N. Smith, Sándor Szalma, Dawn M. Waterworth, Lyndon J. Mitnau, Joseph D. Szustakowski, Bradford W. Gibson, Melissa R. Miller, Christopher D. Whelan

doi: <https://doi.org/10.1101/2022.06.17.496443>

The teams at AstraZeneca & Biogen characterized rare variants influencing protein abundances, via exome sequencing

(Dhindsa, Petrovski et al., *Under Review*)

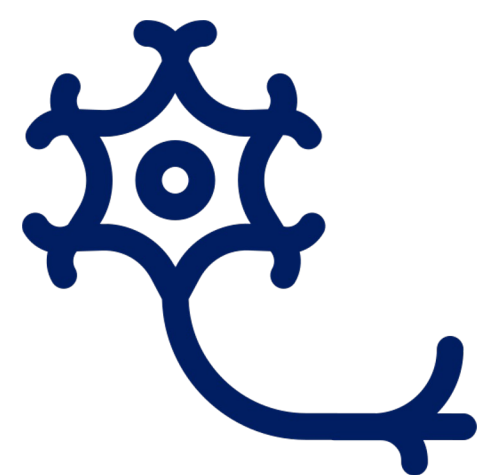


The **Calico** group profiled markers of mortality

(Sethi et al., *Under Review*)

The team at **Takeda** characterized neurofilament light (NfL) as a prognostic marker for ALS

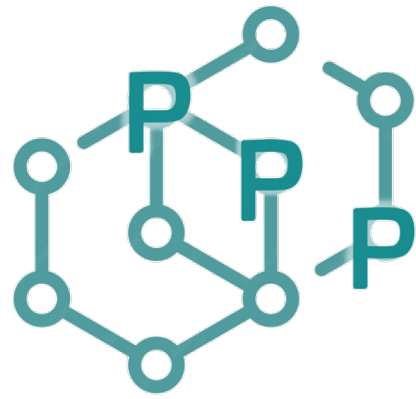
(Smith et al., *Under Review*)



Other companies are revealing new insights into disease causality and pathophysiological mechanisms, as evidenced from **ASHG 2022**

(Six oral presentations; 13 posters)

# Massive-scale proteomics requires massive-scale collaboration!



Work was completed under  
UKB AMS Application 65851

- Amgen • **Brad Gibson**, Kimberly Pohoski, Yi-Hisang Hsu, Kári Stefánsson
- Alnylam • **Luke Ward**, Paul Nioi, Aimee Deaton
- AstraZeneca • **Slavé Petrovski**, Oliver Burren, Ryan Dhindsa
- Biogen • **Ben Sun**, Helen McLaughlin, Danai Chasioti, Tinchu Li, Kyle Ferber
- BMS • **Joe Szustakowski**, Erika Kvikstad, Steve Vasquez-Grinnell
- Calico • **Eugene Melamud**, Endy Inui
- Genentech • **Mark McCarthy**, Anubha Mahajan, Rion Pendergrass
- GSK • **Robert Scott**, Chloe Robins, Praveen Surendran
- Janssen • **Mary Helen Black** (formerly JRD), Letizia Goretti, Dawn Waterworth, Liping Hou, Nasser Doostparast, Yanfei Zhang, Gayle Wittenberg, Shuwei Li
- Novo Nordisk • **Joanna Howson**, Tom Richardson
- Pfizer • **Melissa Miller**, Josh Chiou, Eric Fauman, Craig Hyde
- Regeneron • **Lyndon Mitnaul**, Hyun Min Kang, Thomas Coleman
- Takeda • **Erin Smith**, Sandor Szalma
- Olink • **Evan Mills**, Cindy Lawley, Philippa Pettingiell, Klev Diamanti, Linda Jung, Jon Heimer
- UK Biobank • **Lauren Carson**, John Busby, Dan Fry, Lucy Burkitt-Gray, Naomi Allen, Rory Collins, all 55,000+ participants!

**Pharma Proteomics Project**

**JSC: Erin Smith** (Takeda)

**PI + Consortium chair: Christopher Whelan** (Alnylam)

**JSC: Luke Ward** (Alnylam)

**JSC: Lyndon Mitnaul** (Regeneron)

**JSC: Kári Stefánsson** (Amgen)

**Bradford Gibson** (Amgen)

**JSC: Melissa Miller** (Pfizer)

**JSC: Slavé Petrovski** (AstraZeneca)

**JSC: Ben Sun** (Biogen)

**JSC: Joanna Howson** (Novo Nordisk)

**JSC: Shuwei Li** (Janssen)

**JSC: Eugene Melamund** (Calico)

**JSC: Robert Scott** (GSK)

**JSC: Mark McCarthy** (Genentech)

**JSC: Joe Szustakowski** (Bristol Myers Squibb)





# Working with proteomics data

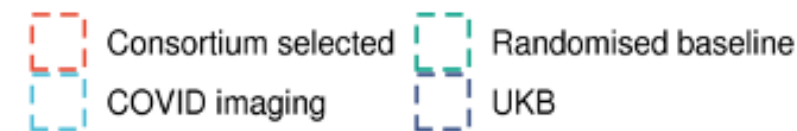
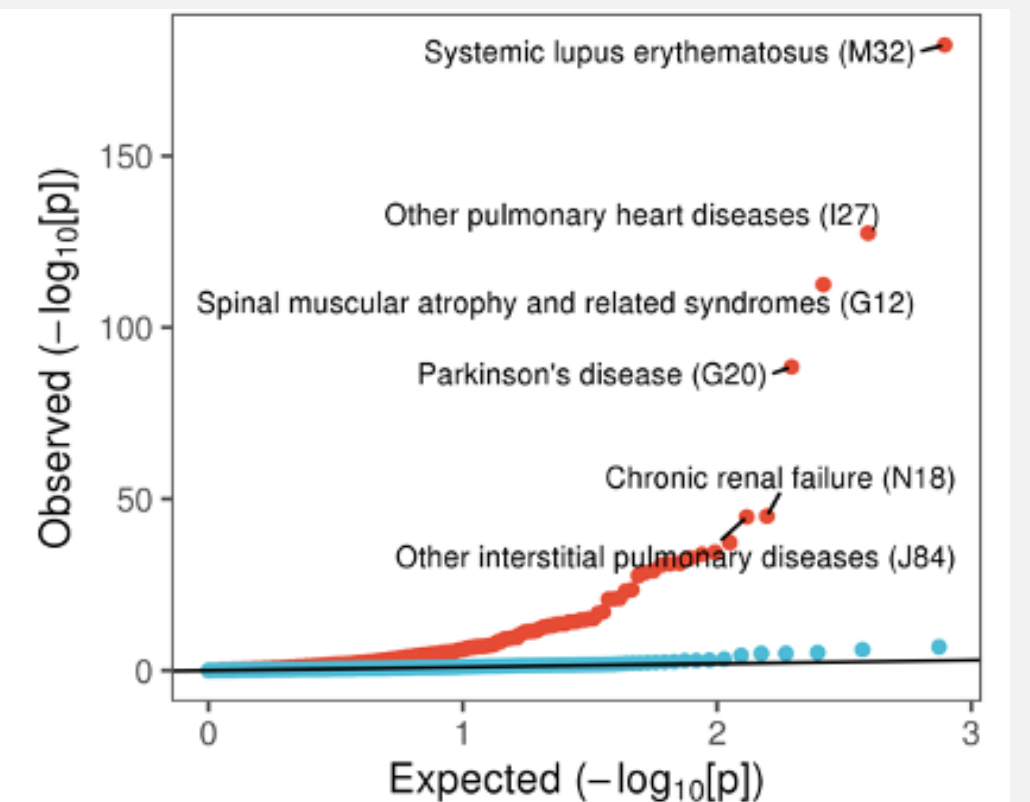
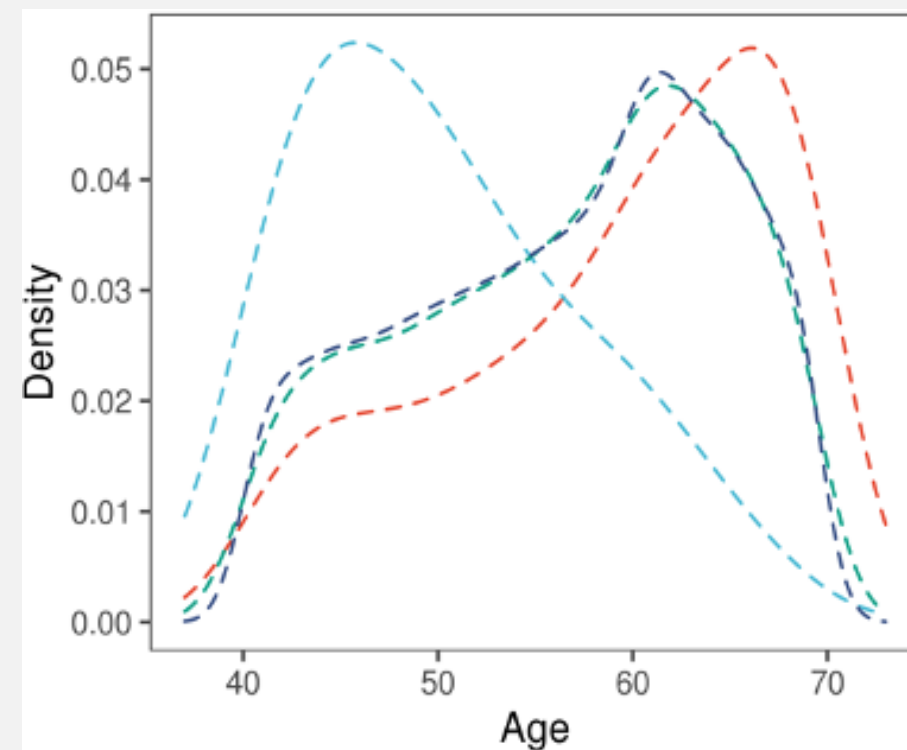
Benjamin Sun

# Structure within the UKB-PPP data

- UKB-PPP contains both random sampled components and non-random selected components based on certain traits
- Consortium samples samples can be quite different to underlying UKB
  - UK population -> UKB -> UKB-PPP (random baseline) – as representative as you get
- UKB-PPP is enriched for diseases that would be otherwise rare by random sampling

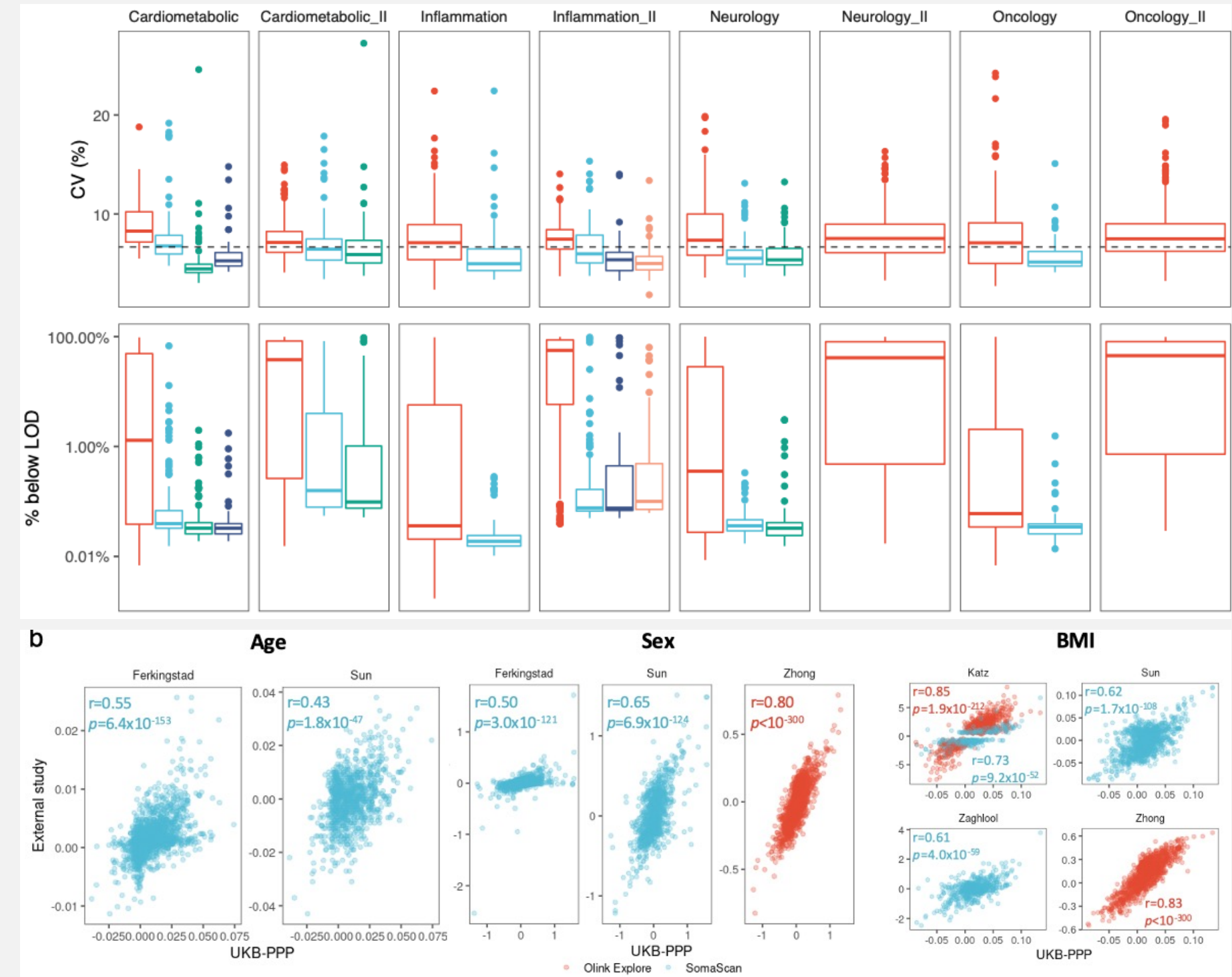
-> Use random baseline if you want the most UKB representative samples, otherwise use combined but be wary of the sample substructure and adjust for if needed

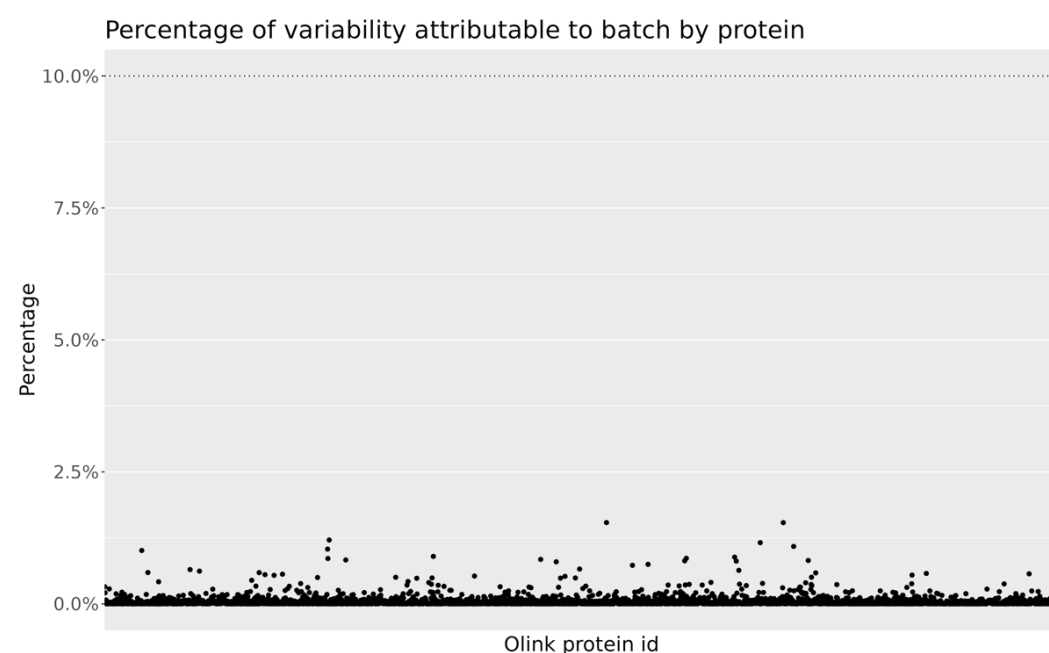
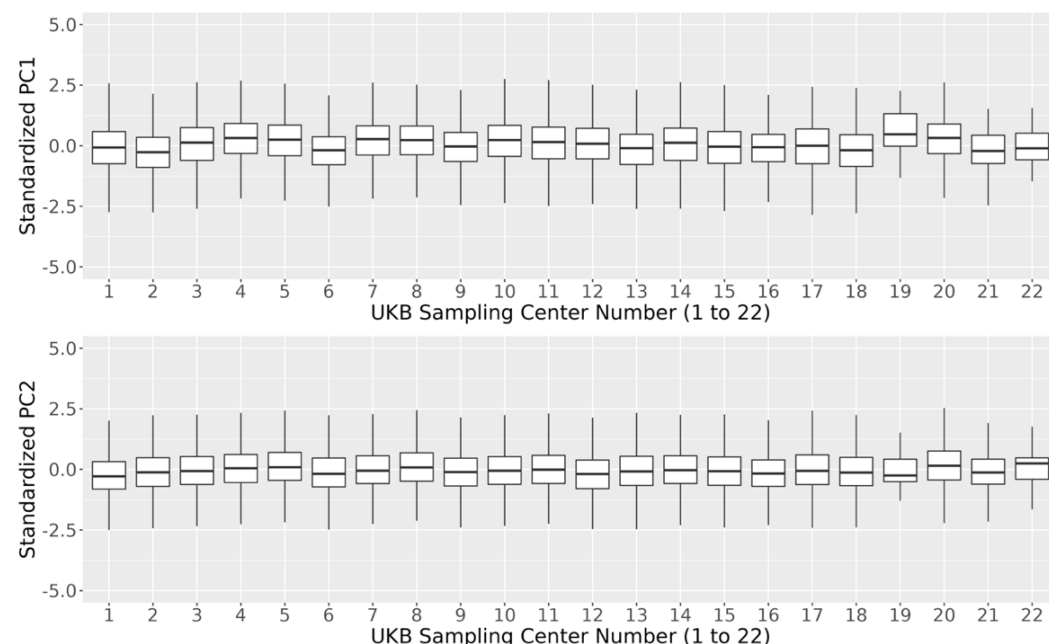
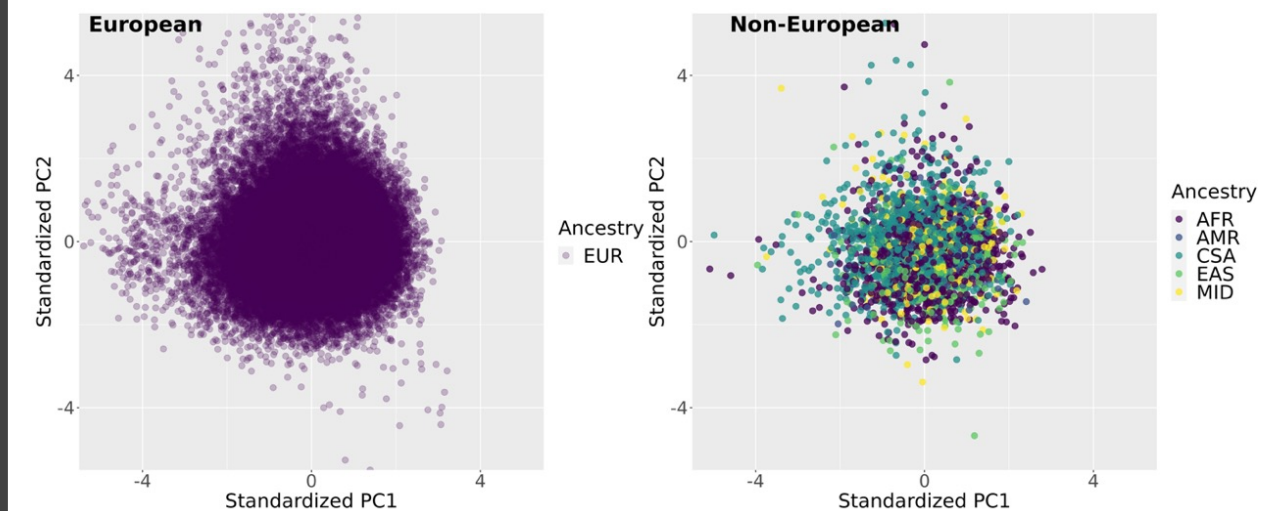
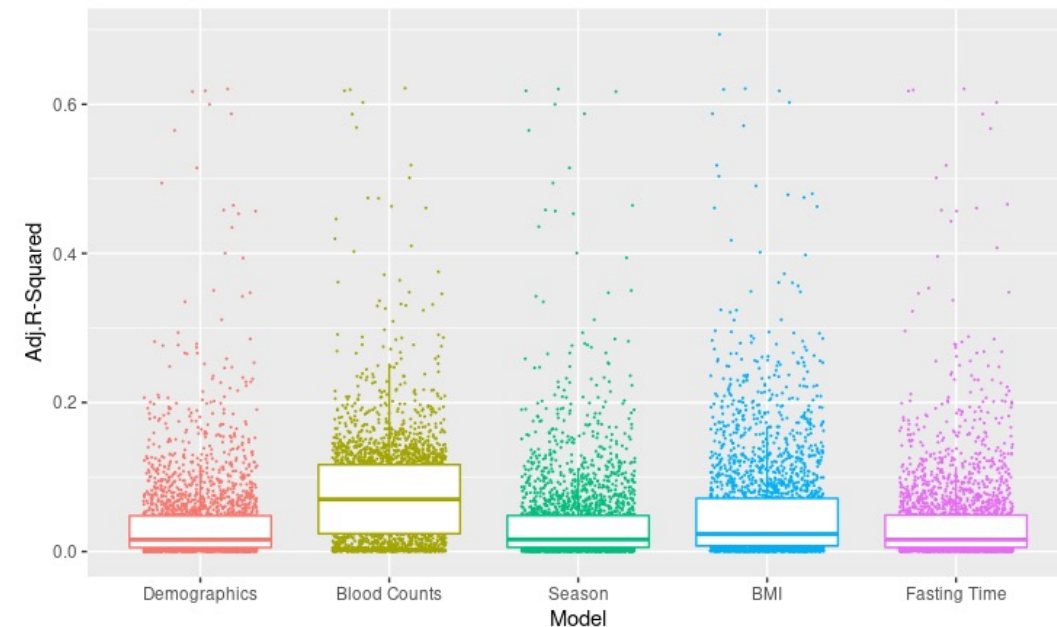
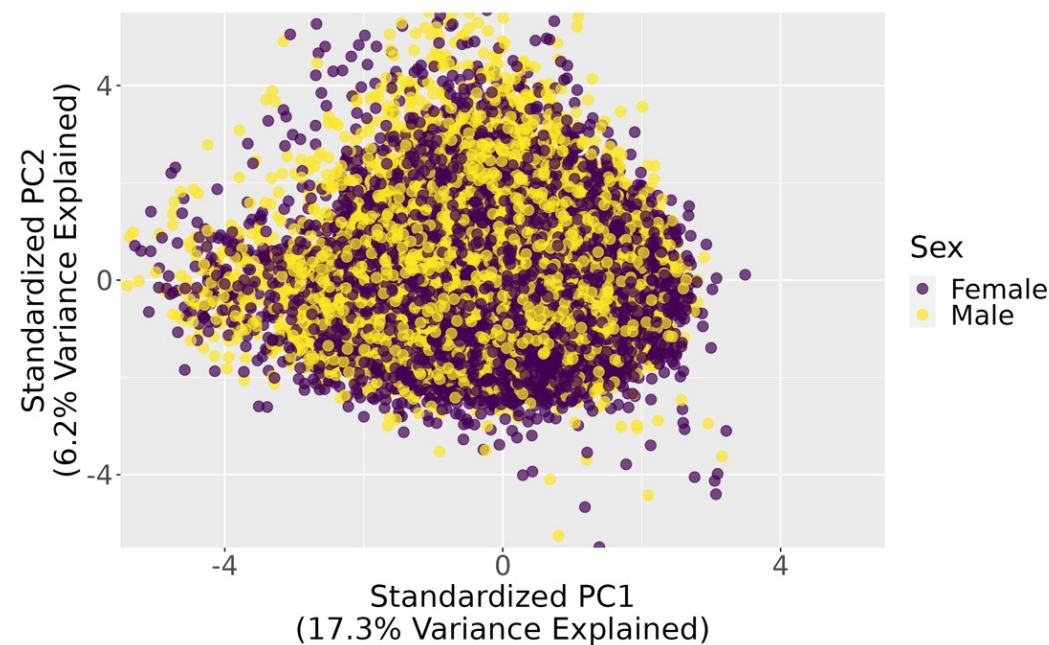
Randomised baseline	Consortium selected	COVID19 imaging <sup>5</sup>	No. of individuals <sup>2,3</sup>
			46,673
			1,248
			6,365
			20



# Be aware of potential sources of proteomics variation

- Olink coefficient of variation (CVs) in line with previous studies and mostly <10%
- Below LOD values are reported
- Both are affected by expected abundance
- Comparisons with UKB assays for overlapping proteins show good concordance
- Associations with demographics show expected trends with previous reports
- Olink -> Olink replication better than Olink -> SomaScan replication





# Sources of variation to be wary of

In addition to age and sex, think about and account for:

- Randomised baseline vs selected samples
- Ethnicities/ancestry -> can be estimated with genetic PCs
- Batches
- UKB centre effects
- Time between blood draw and protein measurement (sample age)

+ if analysing genetic data:

- Genotyping array
- Genotyping vs imputation vs WES vs WGS sequencing data -> not always the same
- Genetic PCs

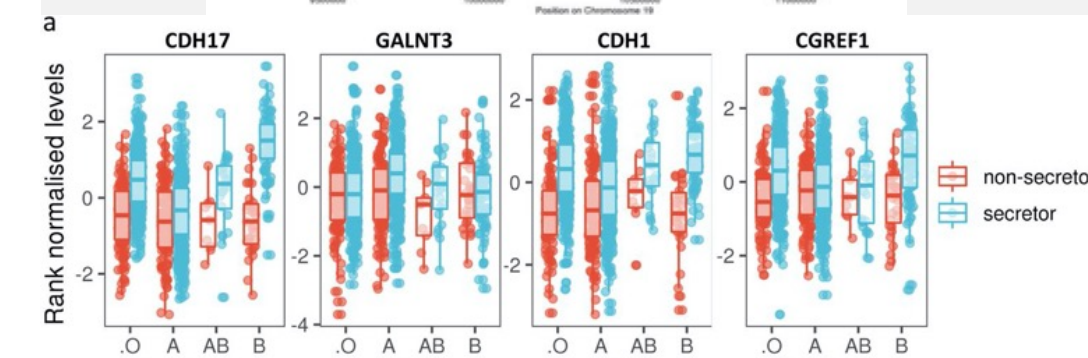
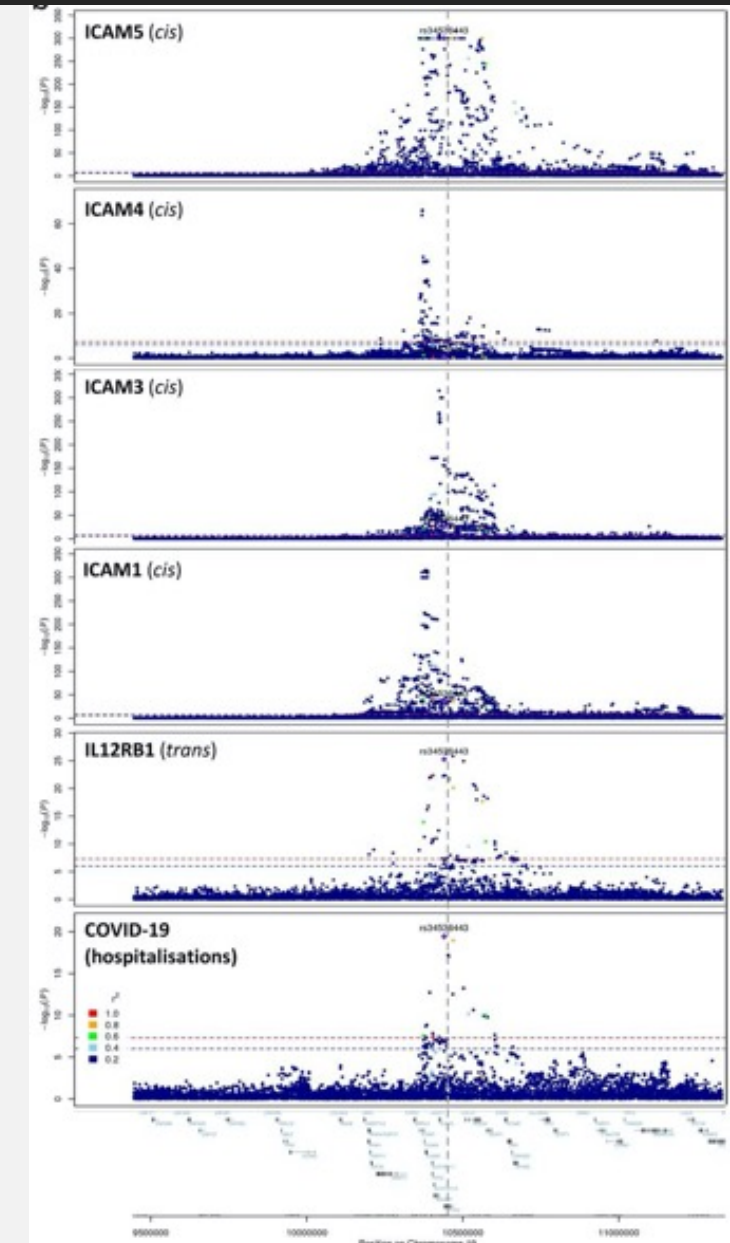
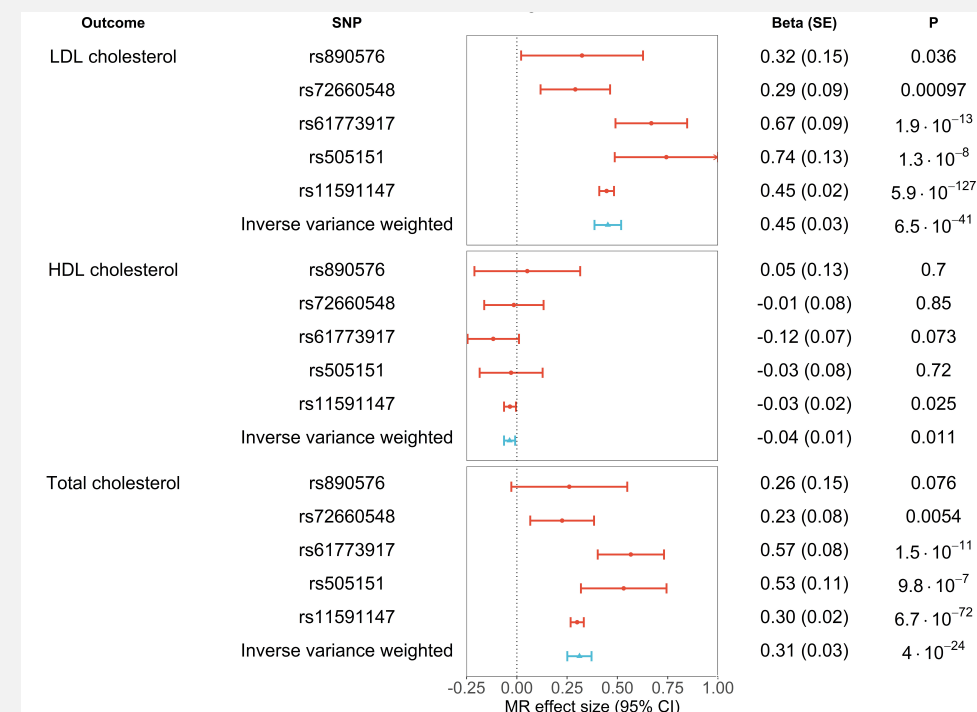
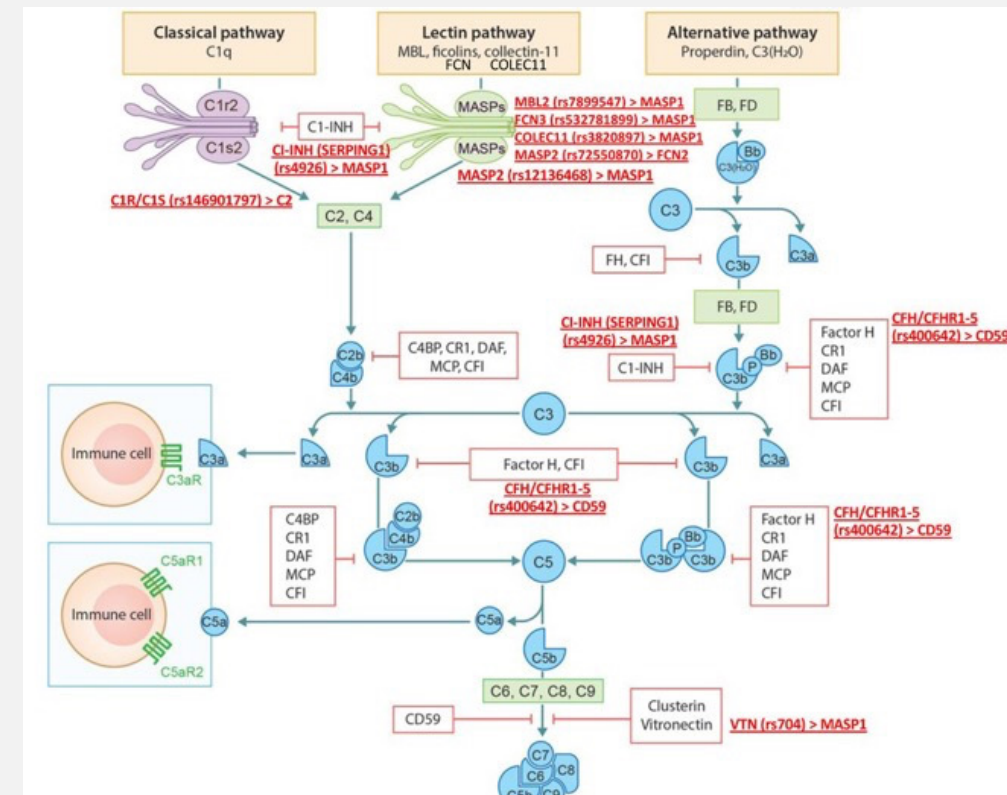
+ many other measured confounders if deep-diving into specific traits

+ protein transformation

- NPX is relative measure and  $\sim$ log scale
- Inverse rank normalization pragmatic if doing proteome-wide scans but doesn't preclude other transformations for specific proteins

# Examples of applications of proteomics

- Proteogenomics – UKB-PPP consortium will be looking to release summary associations for further downstream interrogation via approaches like colocalization and Mendelian randomisation
- Non-genetic proteomics based associations
- Proteomic prediction space
- Combining with other multi-omics datasets in UKB
- Discovery/validation dataset for other studies
- NOTE: plasma proteomics may not necessarily transcend to other tissues/fluids



# Summary

- 
- Consider sources of systematic and sample-based variation/structure which may confound associations
  - Careful framing of the proteomic assay in the right technical and biological context
  - Beware of potential colliders too - pitfalls of over-adjusting – research design and question is important
  - Replication is important as always (UKB-PPP can be thought not only as a discovery platform but also as a replication resource)
  - A wide range of applications for this data – multiple testing is also very important topic

*“with great data comes great scientific responsibility”*

# Crossing Proteomic with other data types

Karsten Suhre

Weill Cornell Medicine - Qatar



As faculty of Weill Cornell Medicine, we are committed to providing transparency for any and all external relationships prior to giving an academic presentation.

I am involved in setting up private companies that aim at bringing omics research to the clinic.



CHYMIA LLC



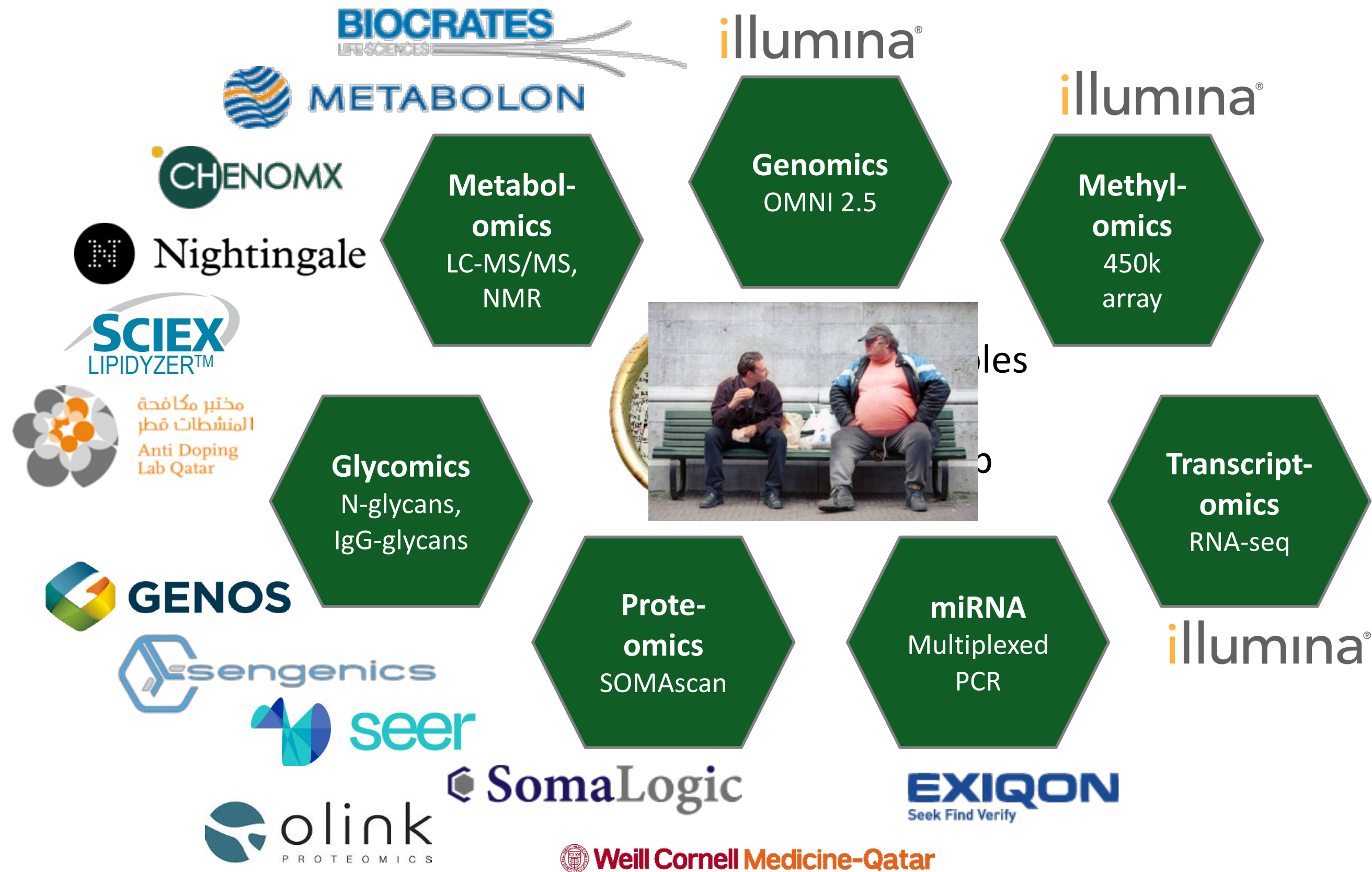
valdia

فالديا

*-Karsten Suhre*

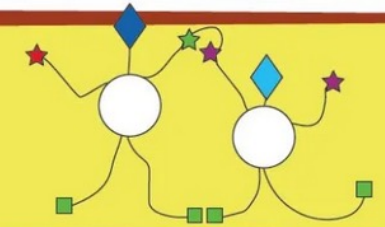


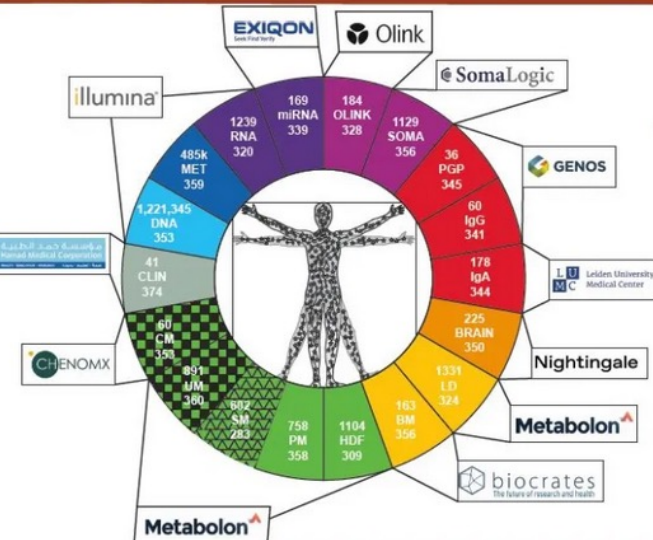
# Multiomics data integration



# Multiomics data integration

## COMics: Connecting Omics






**MULTI-OMICS MEASURED OVER:**

- 18 DIFFERENT PLATFORMS, 3 MATRIXES (SALIVA, URINE, BLOOD) FROM UP TO 374 SUBJECTS!!!**
- TOTAL OF 8,170 MOLECULAR TRAITS**
- GENETIC VARIANTS: 1,221,345**
- METHYLATION: 470,837 CPG SITES**
- GENE EXPRESSION: 57K TRANSCRIPTS**

**WE CONNECTED ALL MULTIOMICS TRAITS USING:**

We connected all multiomics traits using partial correlations to construct mutual best correlation hits (MBH) between molecules within and between different omics layers

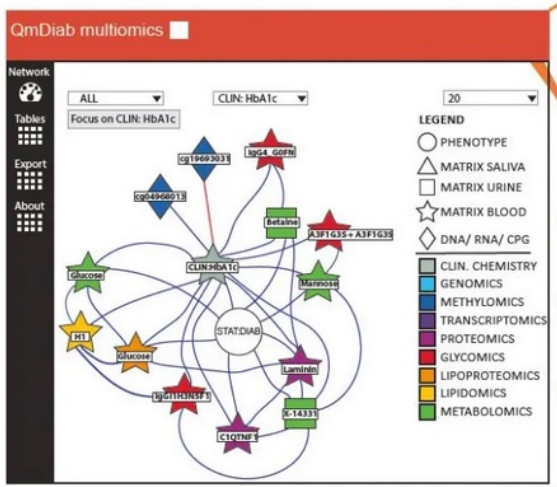


**Gaussian Graphical Models (GGMs) within individual omics-layers** & **genome-wide (GWAS)**

**epigenome-wide (EWAS)** & **transcriptome-wide (TWAS)**

**GWAS Catalog**

**COMICS WEBSERVER GIVE YOU ACCESS TO MULTIOMICS DATA WITH 34,000 STATISTICALLY SIGNIFICANT ASSOCIATIONS**




**USE IT TO FIND OUT:**

- HOW A MOLECULE OF YOUR INTEREST BEHAVES IN MOLECULAR NETWORK;
- HOW MANY ASSOCIATIONS IT HAS;
- &
- HOW IT IS LINKED TO SPECIFIC DISEASE ENDPOINTS.

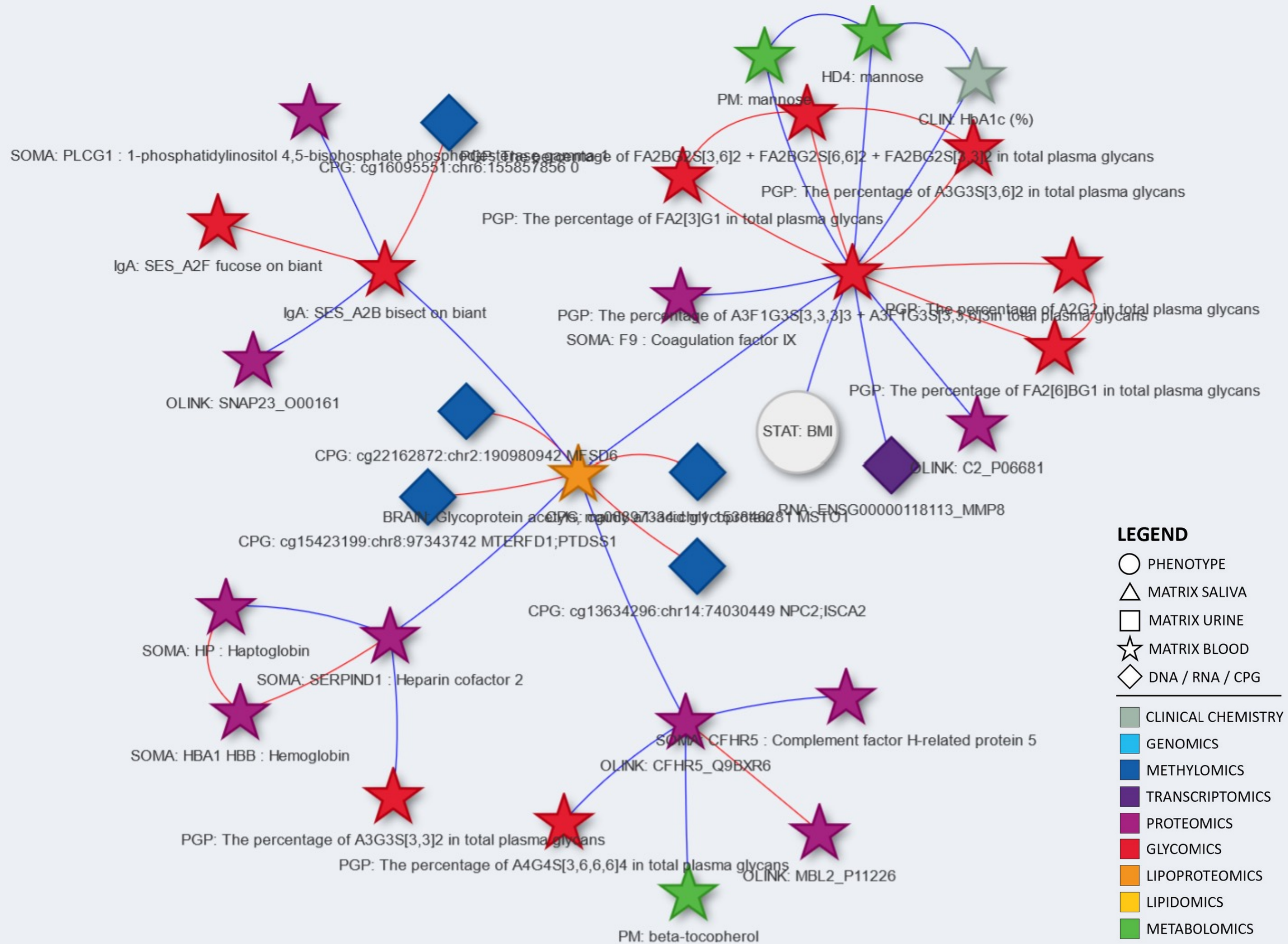
**HowTo** Press if you like to explore COMics and you need some help.

**Use-cases** Press if you like to have a look at some of the examples.

**Network** Press if you like to test it right away.



<http://www.metabolomix.com/comics/>



# First contact with the UKB PPP Olink data



Researcher log in

Participant log in

Contact us

Enable your research

Explore your participation

Learn more about UK Biobank



## Enabling your vision to improve public health

Data drives discovery. We have curated a uniquely powerful biomedical database that can be accessed globally for public health research. Explore data from half a million UK Biobank participants to enable new discoveries to improve public health.

Data Showcase

Future data releases



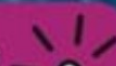
UK Biobank is a large-scale biomedical database and research resource, containing in-depth genetic and health information from half a million UK participants. The database is regularly augmented with additional data and is globally accessible to approved researchers undertaking vital research into the most common and life-threatening diseases. It is a major contributor to the advancement of modern medicine and treatment and has enabled several scientific discoveries that improve human health.



## UK Biobank Winter Scientific Conference

Transforming Future Health

WATCH ON DEMAND



# First contact with the UKB PPP Olink data

The screenshot shows the Biobank UK website interface. At the top, there is a navigation bar with links for Index, Browse, Search, Catalogues, Downloads, Login, and Help. The main content area is titled 'Browse by Primary Category of Origin'. It features a tree view of categories with item counts and a set of buttons for navigating through levels.

Category	Items
Population characteristics	35
Assessment centre	3939
Biological samples	0
Blood assays	0
Blood count	155
Blood biochemistry	210
Infectious Diseases	72
NMR metabolomics	508
Proteomics	0
Protein biomarkers	4
Sample inventory	13
Saliva assays	0
Urine assays	16
Genomics	271
Online follow-up	1107
Additional exposures	366
Health-related outcomes	2646

Summary generated 11 March 2023

Enabling scientific discoveries that improve human health

<https://biobank.ndph.ox.ac.uk/showcase/browse.cgi>

## Data-Field 30900

Description: Number of proteins measured

Category: [Biological samples](#) ▶ [Blood assays](#) ▶ [Proteomics](#) ▶ [Protein biomarkers](#)

Participants	52,749
Item count	55,002
Stability	Accruing

Value Type	Integer, number of proteins
Item Type	Records
Strata	Derived

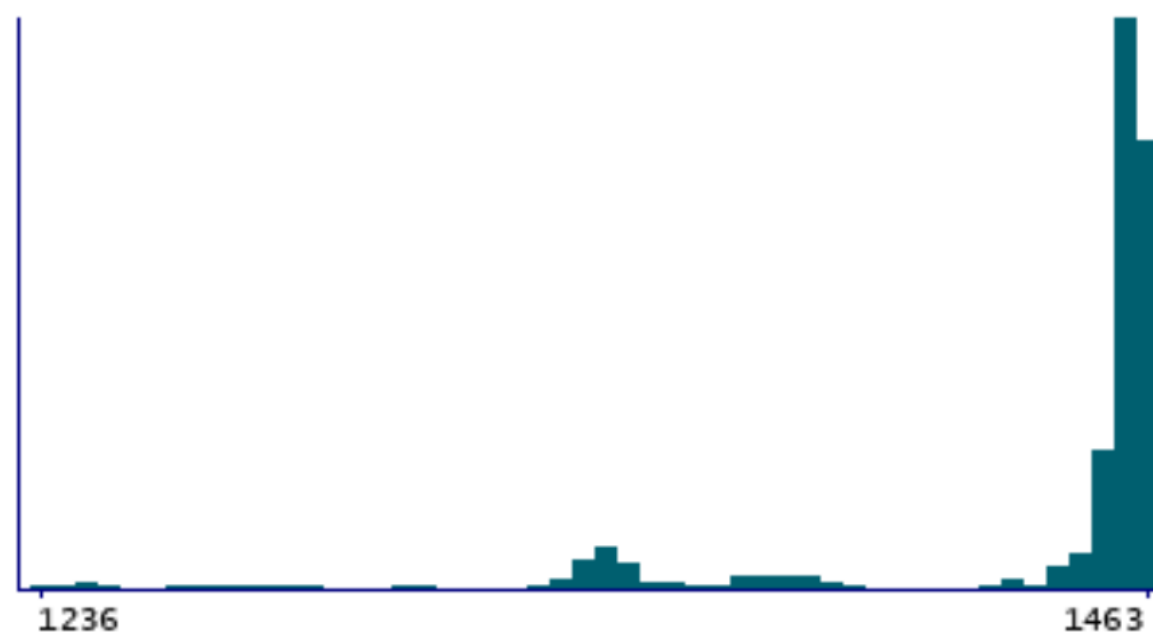
Sexed	Both sexes
Instances	Defined (4)
Array	No

Debut	Jan 2023
Version	Feb 2023
Cost Tier	d2 o2 s2

<b>Data</b>	<b>3 Instances</b>	<b>Notes</b>	<b>1 Record Table</b>	<b>0 Related Data-Fields</b>	<b>12 Resources</b>
-------------	--------------------	--------------	-----------------------	------------------------------	---------------------

55,002 items are available, covering 52,749 participants.  
 Defined-instances run from 0 to 3, labelled using Instancing 2.  
 Units of measurement are number of proteins.

Maximum	1463
Decile 9	1462
Decile 8	1461
Decile 7	1461
Decile 6	1460
Median	1459
Decile 4	1458
Decile 3	1455
Decile 2	1400
Decile 1	1349
Minimum	75



- There are 672 distinct values.
- Mean = 1421.15
- Std.dev = 108.591
- 2455 items below graph minimum of 1236

<https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=30900>

## Data-Coding 143

Name: UniProt meaning for OLINK Protein ID

Description: Relates the integer Protein ID presented in the OLINK dataset to the UniProt meaning.

This is a flat (unstructured) list which uses integers to represent categories or special values.

Coding can be downloaded here as a tab-separated file. [Download](#)

### 2923 Categories

Coding	Meaning
1	A1BG;Alpha-1B-glycoprotein
2	AAMDC;Mth938 domain-containing protein
3	AARSD1;Alanyl-tRNA editing protein Aarsd1
4	ABCA2;ATP-binding cassette sub-family A member 2
5	ABHD14B;Protein ABHD14B
6	ABL1;Tyrosine-protein kinase ABL1
7	ABO;Histo-blood group ABO system transferase
8	ABRAXAS2;BRISC complex subunit Abraxas 2
9	ACAA1;3-ketoacyl-CoA thiolase, peroxisomal
10	ACADM;Medium-chain specific acyl-CoA dehydrogenase, mitochondrial
11	ACADSB;Short/branched chain specific acyl-CoA dehydrogenase, mitochondrial
12	ACAN;Aggrecan core protein
13	ACE;Angiotensin-converting enzyme
14	ACE2;Angiotensin-converting enzyme 2

<https://biobank.ndph.ox.ac.uk/showcase/coding.cgi?id=143&nl=1>

2918	ZNF830;Zinc finger protein 830
2919	ZNRD2;Protein ZNRD2
2920	ZNRF4;E3 ubiquitin-protein ligase ZNRF4
2921	ZP3;Zona pellucida sperm-binding protein 3
2922	ZP4;Zona pellucida sperm-binding protein 4
2923	ZPR1;Zinc finger protein ZPR1

## Data-Field 30900

Description: Number of proteins measured

Category: [Biological samples](#) ▶ [Blood assays](#) ▶ [Proteomics](#) ▶ [Protein biomarkers](#)













Participants	52,749
Item count	55,002
Stability	Accruing

Value Type	Integer, number of proteins
Item Type	Records
Strata	Derived

Sexed	Both sexes
Instances	Defined (4)
Array	No

Debut	Jan 2023
Version	Feb 2023
Cost Tier	d2 o2 s2

[Data](#) | [3 Instances](#) | [Notes](#) | [1 Record Table](#) | [0 Related Data-Fields](#) | [12 Resources](#)

Preview Name	Res ID
 Olink Analysis Report	4655
 Olink Explore 1536 - FAQ	4657
 Olink data normalisation strategy	4656
 Olink proteomics data	4654
 Quality control of olink NPX dataset	4658
 olink assay	1013
 olink assay version	1014
 olink assay warning	1015
 olink batch number	1016
 olink limit of detection	1017
 olink panel lot number	1018
 olink processing start date	1019

<https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=30900>



# First contact with the UKB PPP Olink data

UK Biobank Pharma Proteomics Project:

**Olink quality control summary**



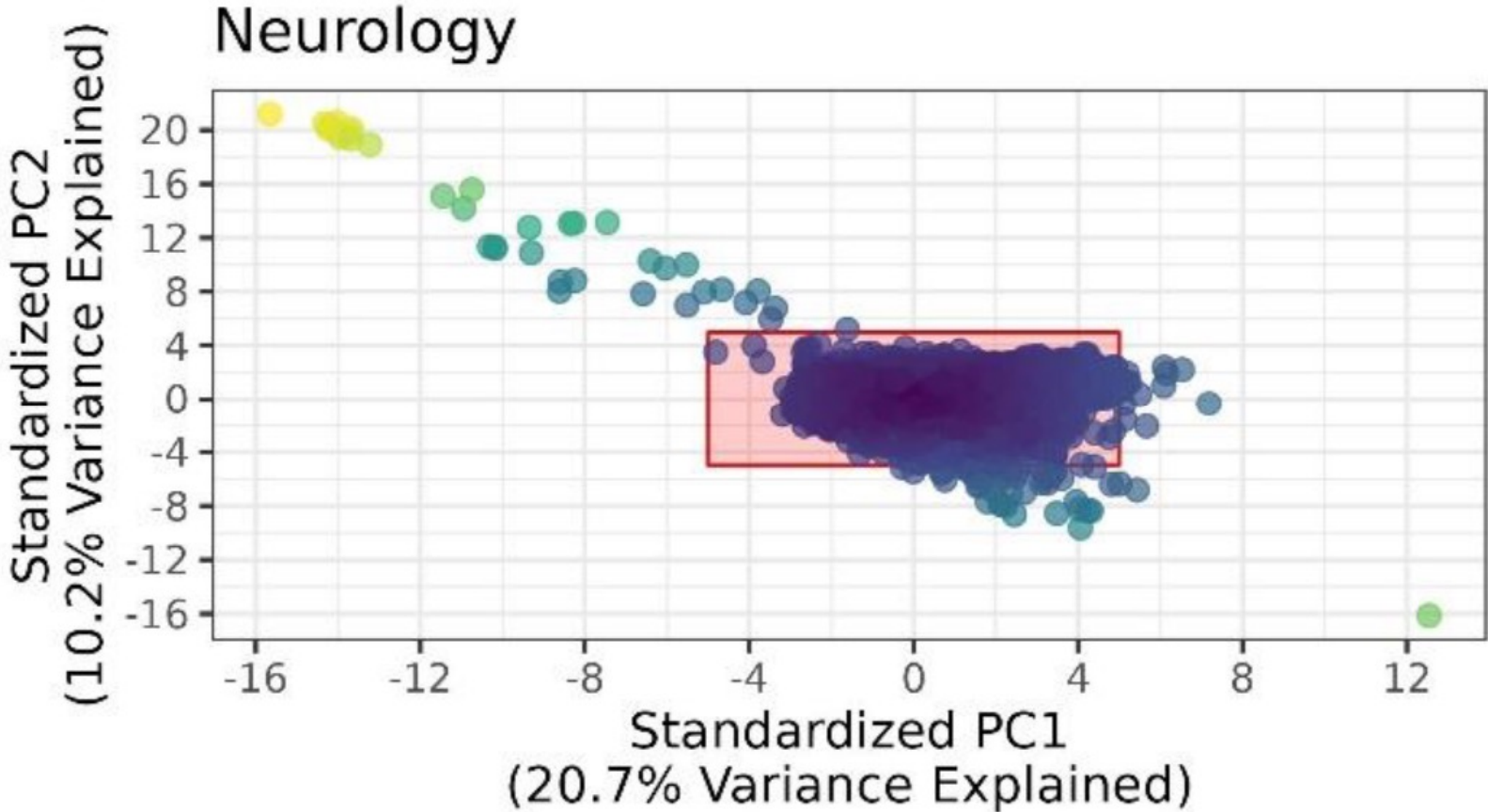
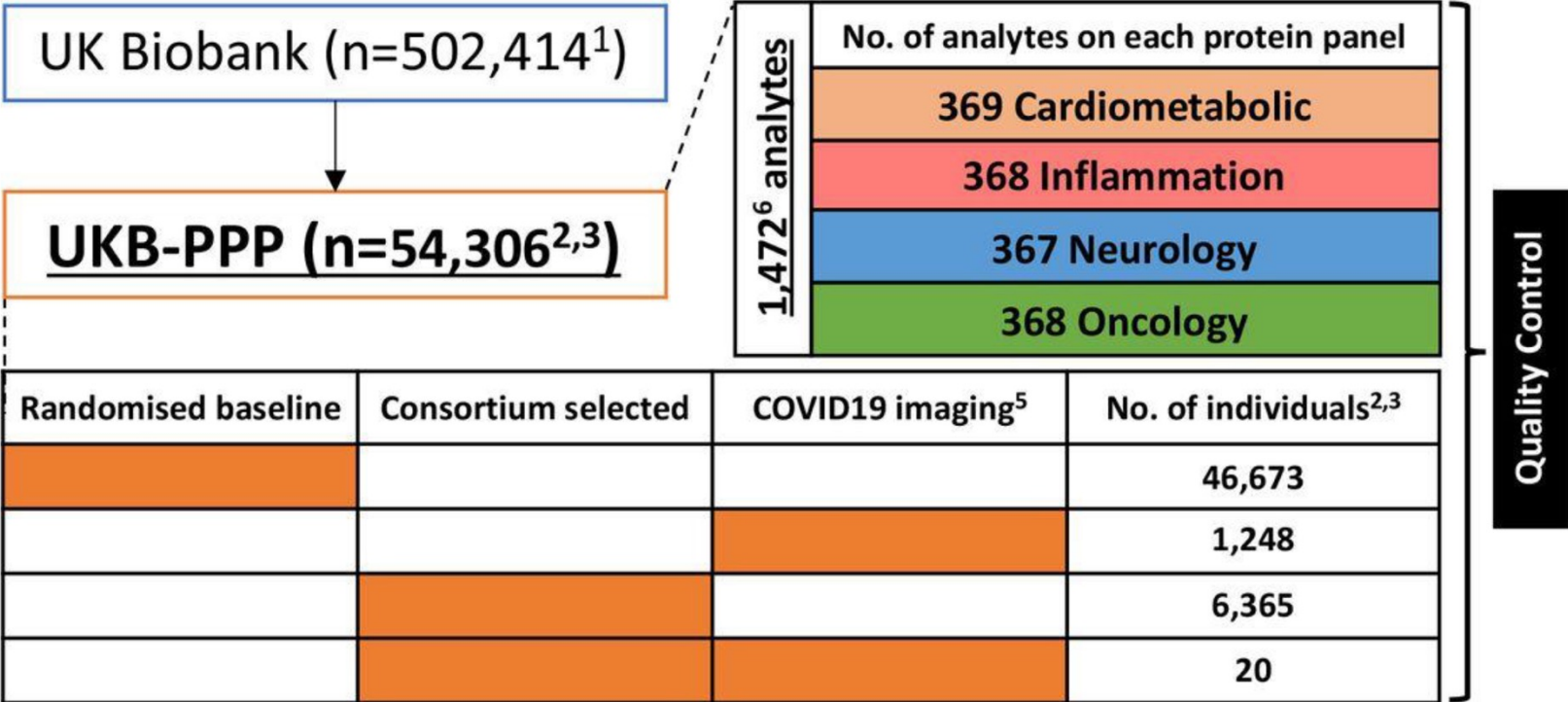
---

## UK Biobank Pharma Proteomics Project

Quality control of Olink NPX dataset • 'Phase 1', Batches 0-7

Benjamin B. Sun, Kyle Ferber and Tinchu Lin (Biogen); Christopher D. Whelan (Janssen)

# First contact with the UKB PPP Olink data

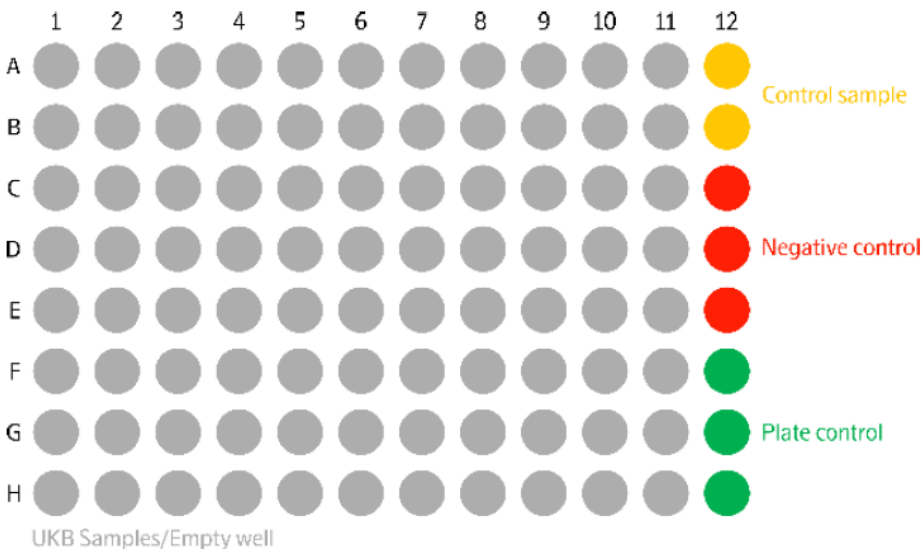


# First contact with the UKB PPP Olink data



## UKB – Olink Explore 1536 - Data Normalization Strategy

Samples from the study were divided into two sets: i) Set 1 – UKB; and ii) Set 2 – COVID; depending on the time point they were randomly selected from the UKB population. Samples were randomly assigned to 96-well plates, and fully randomized within plates. Each plate contained: i) 87 samples from set 1, set 2 or both; ii) 1 empty well, bridge or overlapping sample; iii) 2 Olink control samples used for quality control; iv) 3 Olink negative control samples used to compute the baseline assay level of each plate; and v) 3 Olink plate control samples used for normalization of protein expression. All Olink samples were placed in column 12 of plates while the remaining 87 samples + 1 empty well were randomized across columns 1-11 and rows A to H (Figure 1).



# First contact with the UKB PPP Olink data



## Analysis Report

### 2.1 QC summary

Olink Panel	Samples passed QC	Samples passed QC (%)	Datapoints passed QC	Datapoints passed QC (%)
Cardiometabolic	54523 / 58369	93.41	21152616 / 21538161	98.21
Inflammation	54281 / 58363	93.01	21011339 / 21477584	97.83
Oncology	54153 / 58366	92.78	20961212 / 21478688	97.59
Neurology	54133 / 58366	92.75	20913874 / 21420322	97.64

### 2.2.1 Average %CV

Olink Panel	Intra-plate %CV	Inter-plate %CV
Explore 384 Cardiometabolic	7.59	17.83
Explore 384 Inflammation	7.61	17.59
Explore 384 Neurology	7.81	18.52
Explore 384 Oncology	8.04	18.37

# Accessing the Olink data on DNAnexus / UKB RAP



**biobank**  
uk




**Research Analysis  
Platform**

Enabling scientific discoveries that improve  
human health

**ACCESS PLATFORM**

# Accessing the data on DNAnexus / UKB RAP

The screenshot shows the Biobank UK website interface. At the top, the logo "biobank<sup>uk</sup>" is followed by navigation links: "PROJECTS" (with a dropdown arrow), "TOOLS" (with a dropdown arrow), "ORG ADMIN" (with a dropdown arrow), and "HELP" (with a dropdown arrow). Below this is a horizontal line. Underneath the line, the word "Projects" is displayed in a large font, followed by a tab labeled "ALL" which is underlined. Below the tab are five filter buttons: "Any Name" (with a dropdown arrow), "Any ID" (with a dropdown arrow), "Any Creator" (with a dropdown arrow), "Any Shared With" (with a dropdown arrow), and "Any Billed To" (with a dropdown arrow). Below the filters is a table with the following columns: "Name", "Data Usage", "Access", "Mem...", and "Status" (with an upward arrow). The table contains one row of data: a briefcase icon, the text "UKB\_PPP\_Olink", a teal arrow icon, "5.34 PiB", "Admin", "2", and a refresh icon followed by "Dispensing (10%)".

Name	Data Usage	Access	Mem...	Status ^
 UKB_PPP_Olink 	5.34 PiB	Admin	2	 Dispensing (10%)

# Accessing the data on DNAnexus / UKB RAP



PROJECTS ▾

TOOLS ▾

ORG ADMIN ▾

HELP ▾



app43418\_20230423140958.dataset  
502,368 Participants



Untitled Cohort  
13,351 Participants



Untitled Cohort

13,351 of 502,368 Participants

+ Add Filter

Clear All Filters

Select PARTICIPANT

Number of proteins measured | Instance 0 IS GREATER THAN 0 ⊗

AND

Spectrometer | Instance 0 IS NOT NULL ⊗

# Accessing the data on DNAnexus / UKB RAP

+ Add Column to Table ×

APOM ⊗

- Assessment centre
  - Verbal interview
    - Medications
      - abc Treatment/medication code | Instance 0 1 match
      - abc Treatment/medication code | Instance 1 1 match
      - abc Treatment/medication code | Instance 2 1 match
      - abc Treatment/medication code | Instance 3 1 match
- Biological samples
  - Blood assays
    - Proteomics
      - Protein biomarkers
        - Olink Instance 0
          - 123 **APOM;Apolipoprotein M**
        - Olink Instance 2
          - 123 **APOM;Apolipoprotein M**
        - Olink Instance 3
          - 123 **APOM;Apolipoprotein M**

i Data Field Details ×

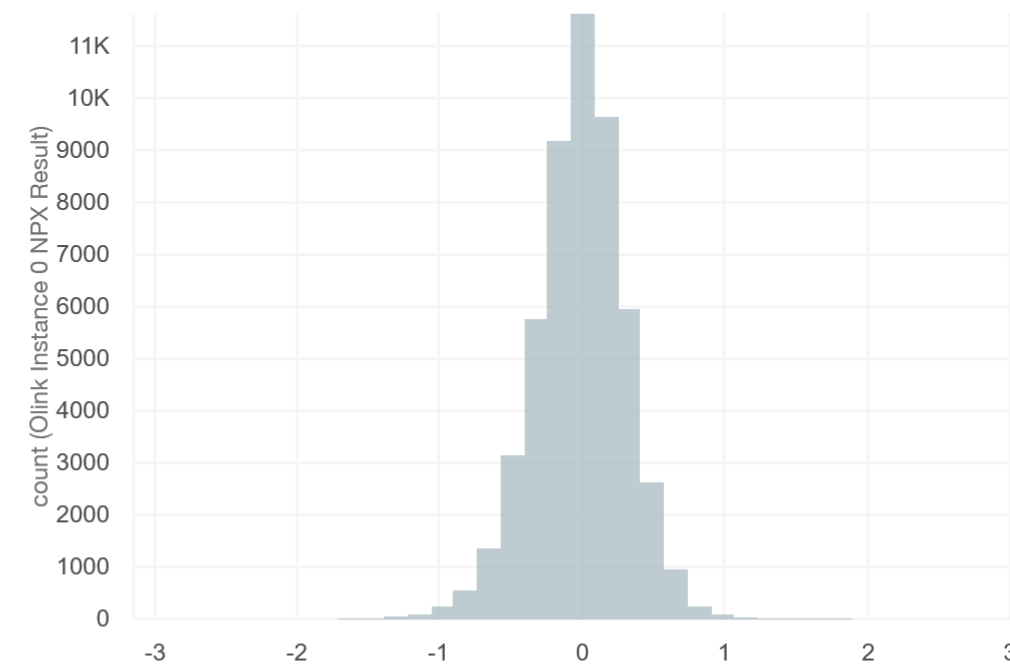
APOM;Apolipoprotein M ×

**Entity** Olink Instance 0 NPX Result  
1:1 to Participant

**Category**

**Value Type** Numerical (Float)

Data Preview: APOM;Apolipoprotein M



Add Cohort Filter

Add Tile

Other actions ▾

Add to Data Preview



# Accessing the data on DNAnexus / UKB RAP

**biobank<sup>uk</sup>** PROJECTS ▾ TOOLS ▾ ORG ADMIN ▾ HELP ▾ 🔔 K ▾

Run Analysis | **Table exporter** ● Required field not configured Start Analysis

🔍 🔍 Reset zoom 🏠

**ANALYSIS SETTINGS** ANALYSIS INPUTS 1 APP SETTINGS

**Execution Name**  
Table exporter

**Execute in Project**  
📁 UKB\_Olink (project-GV2YZB0J8xYPKxkK73PG2Yzb) ...

**Execution Output Folder**  
📁 / ...

**Priority** ⓘ  
Normal ▾

**Spending Limit** ⓘ  
No Limit

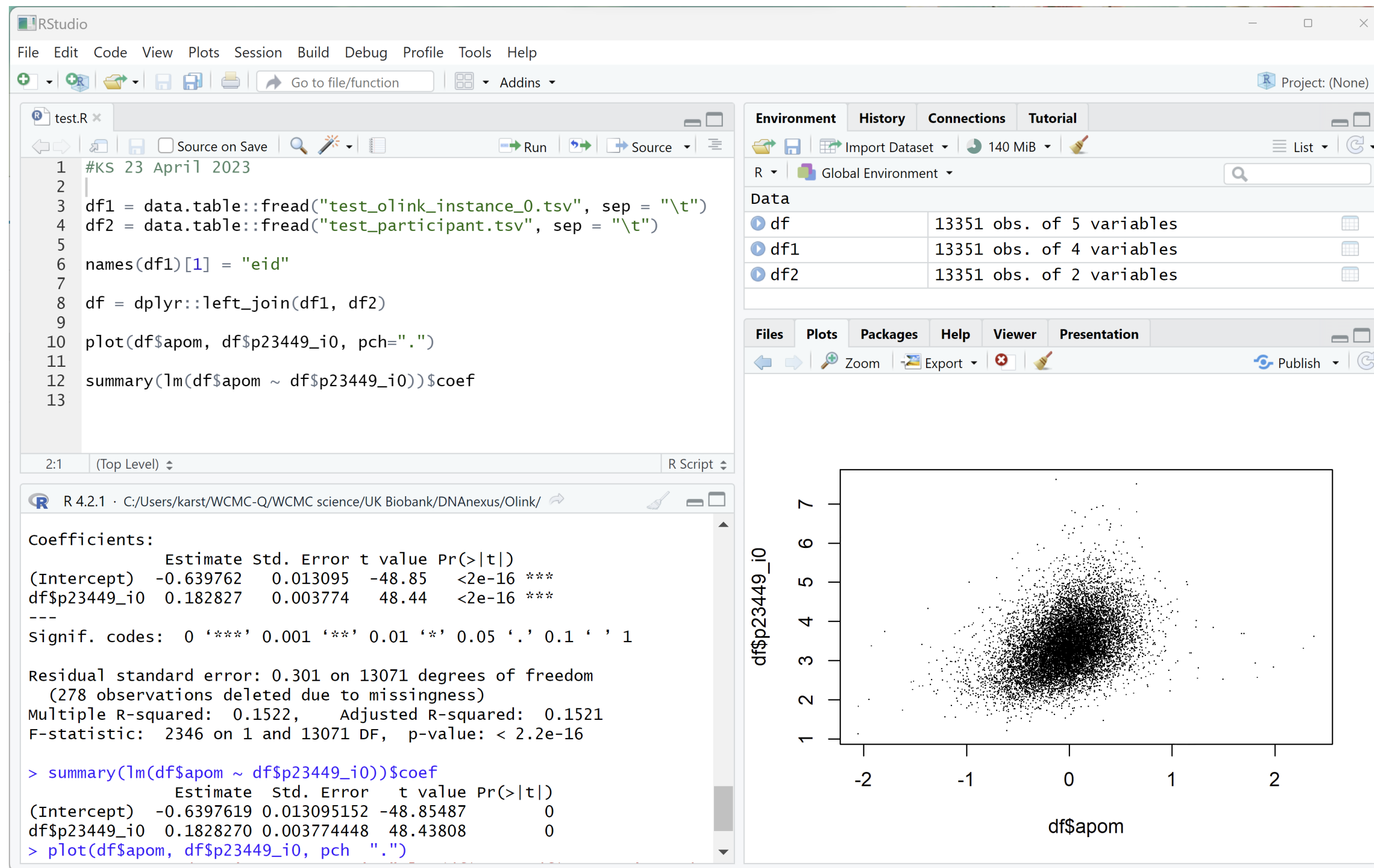
**ADVANCED**

**Allow SSH Access**  OFF

The tool or one of its dependencies does not allow SSH access.

```
graph LR; A((dataset_or_cohort_or_dashboard)) --- B[Table exporter]; C((field_names_file_txt)) --- B; B --- D((CSV *.csv, *.tsv))
```

# Accessing the data on DNAnexus / UKB RAP



The screenshot displays the RStudio interface. The source editor on the left contains the following R code:

```
1 #KS 23 April 2023
2
3 df1 = data.table::fread("test_olink_instance_0.tsv", sep = "\t")
4 df2 = data.table::fread("test_participant.tsv", sep = "\t")
5
6 names(df1)[1] = "eid"
7
8 df = dplyr::left_join(df1, df2)
9
10 plot(df$apom, df$p23449_i0, pch=".")
11
12 summary(lm(df$apom ~ df$p23449_i0))$coef
13
```

The Environment pane on the right shows the following data objects:

Object	Size
df	13351 obs. of 5 variables
df1	13351 obs. of 4 variables
df2	13351 obs. of 2 variables

The console on the left shows the output of the `summary` function:

```
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.639762  0.013095  -48.85  <2e-16 ***
df$p23449_i0  0.182827  0.003774   48.44  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.301 on 13071 degrees of freedom
(278 observations deleted due to missingness)
Multiple R-squared:  0.1522,    Adjusted R-squared:  0.1521
F-statistic: 2346 on 1 and 13071 DF, p-value: < 2.2e-16

> summary(lm(df$apom ~ df$p23449_i0))$coef
      Estimate Std. Error  t value Pr(>|t|)
(Intercept) -0.6397619 0.013095152 -48.85487    0
df$p23449_i0  0.1828270 0.003774448  48.43808    0
> plot(df$apom, df$p23449_i0, pch ".")
```

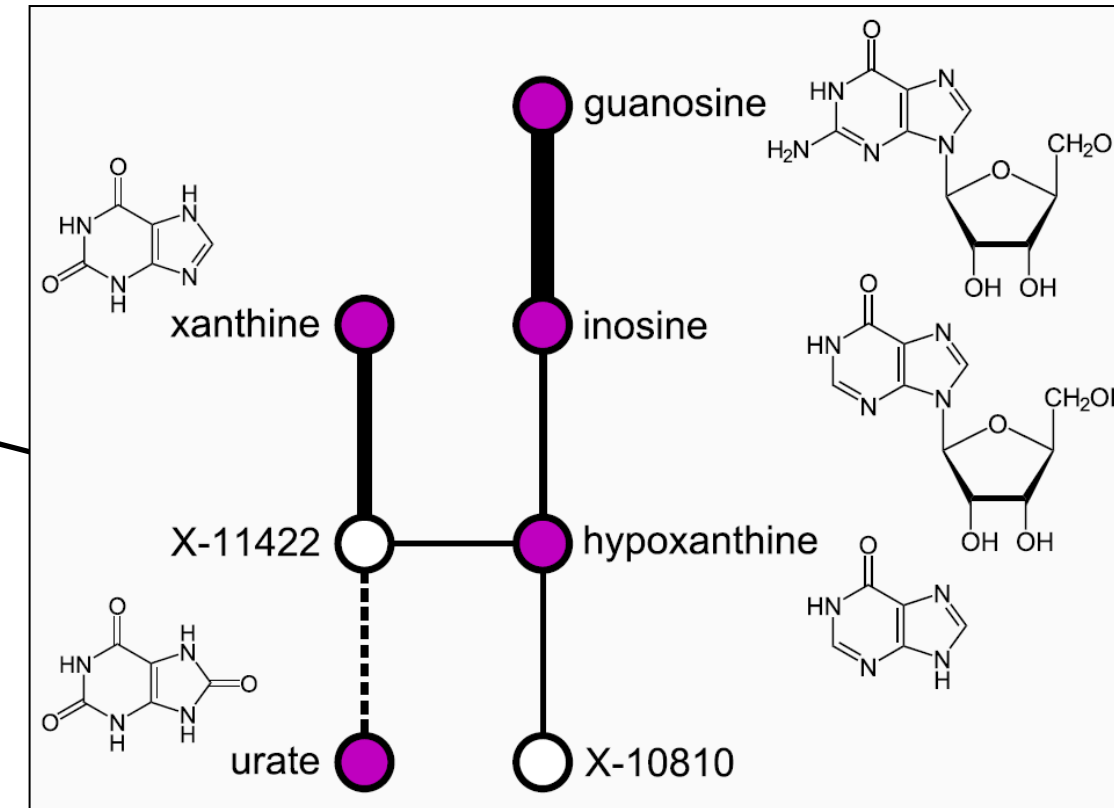
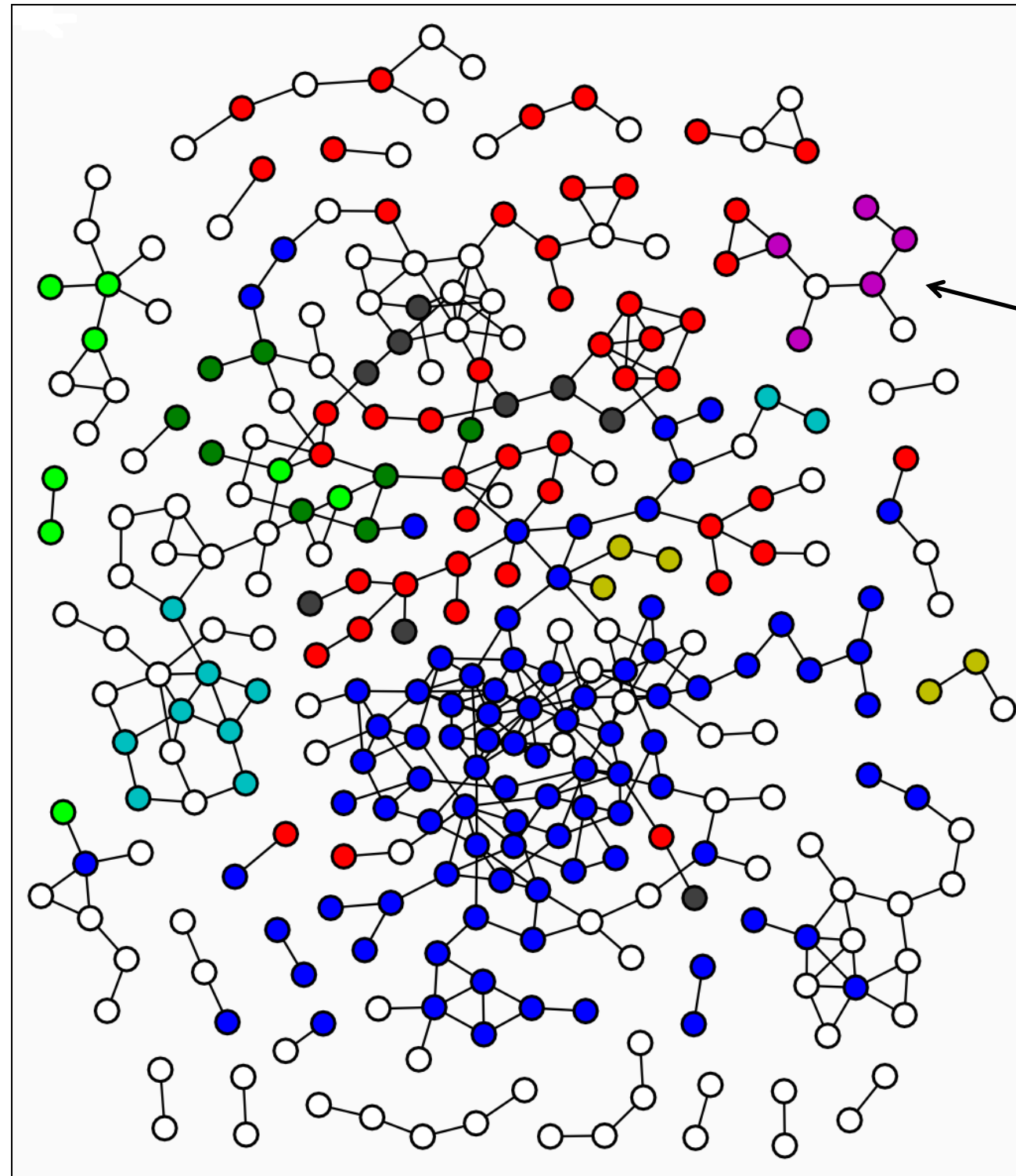
The Plots pane on the right displays a scatter plot of `df$p23449_i0` (y-axis) versus `df$apom` (x-axis). The plot shows a dense cloud of points centered around the origin, with a slight positive correlation. The x-axis ranges from -2 to 2, and the y-axis ranges from 1 to 7.

# Krumsiek Lab



<http://krumsieklab.org>

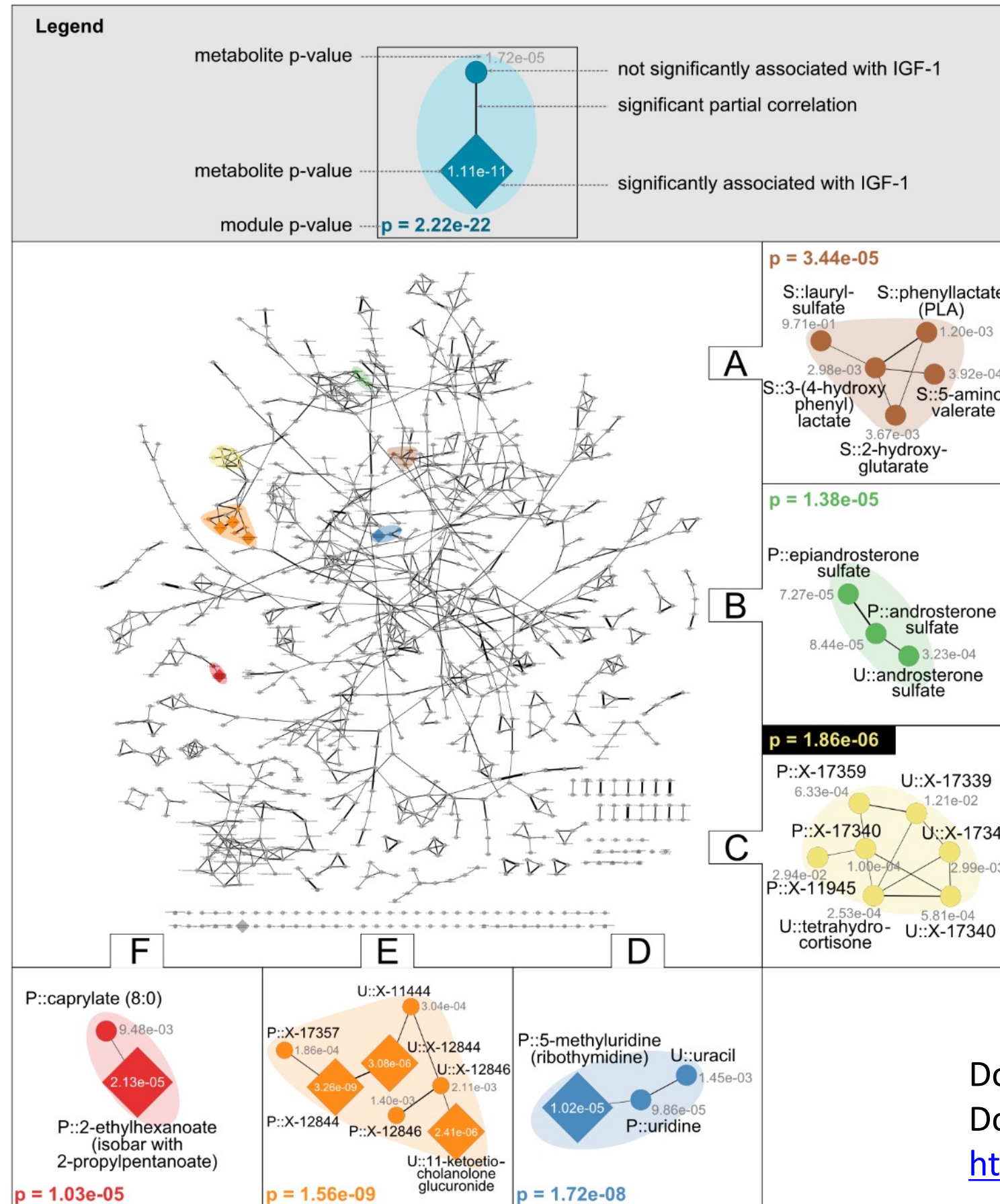
# Gaussian graphical models



- Pathway**
- Lipid
  - Carbohydrate
  - Amino acid
  - Xenobiotics
  - Nucleotide
  - Energy
  - Peptide
  - Cofactors & vitamins
  - Unknowns

Krumsiek et al., BMC Syst Bio, 2011  
 Krumsiek et al., PLoS Genetics, 2012  
 Shin et al., Nat Genetics, 2014  
 Aichler et al., Cell Metabolism, 2017

# Multi-fluid metabolomic modules



Modules for IGF-1 associations at metabolite level

Do et al., *npj Systems Biology and Applications*, 2017  
 Do et al., *Bioinformatics*, 2018  
<https://github.com/krumsieklab/MoDentify>

# AutoFocus



<https://capturetheatlas.com/wp-content/uploads/2020/02/aperture-and-depth-of-field-in-photography-chart.jpg>

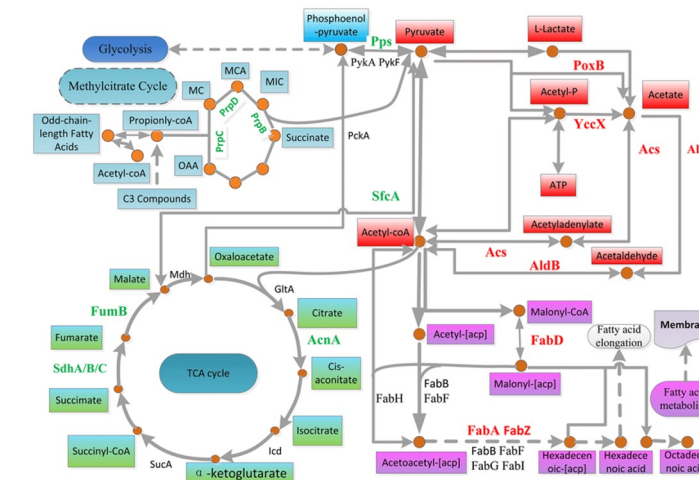
Pyruvate



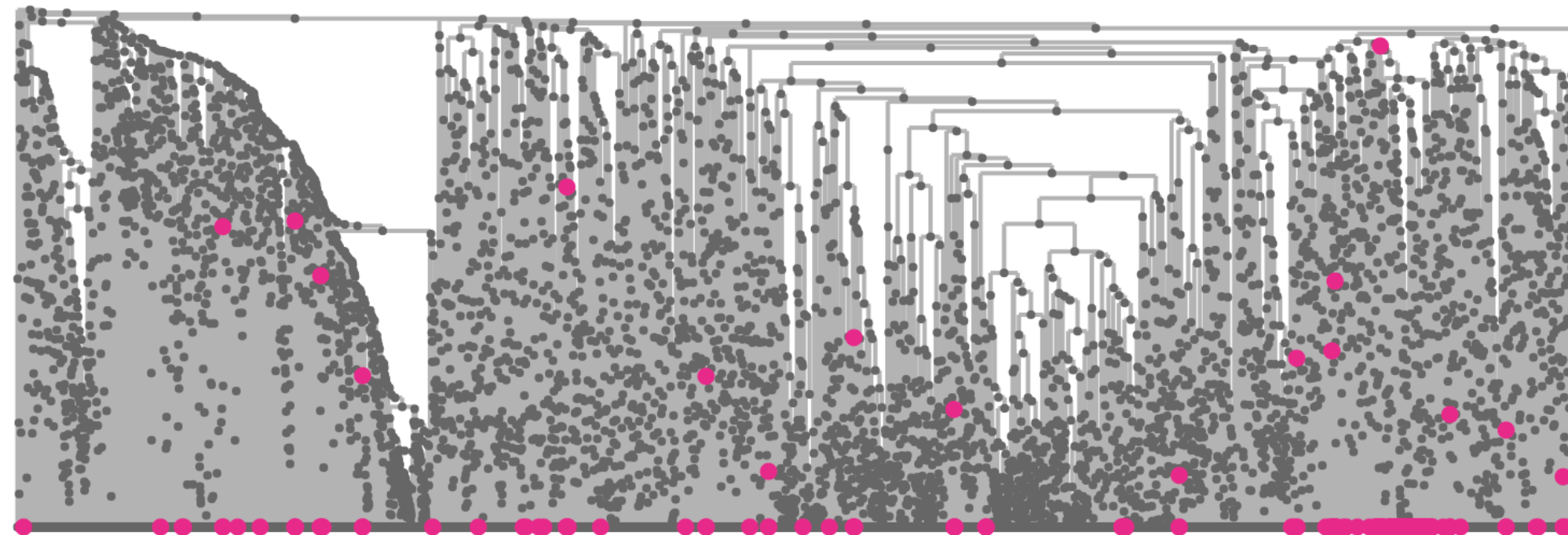
Glycolysis



Central carbon

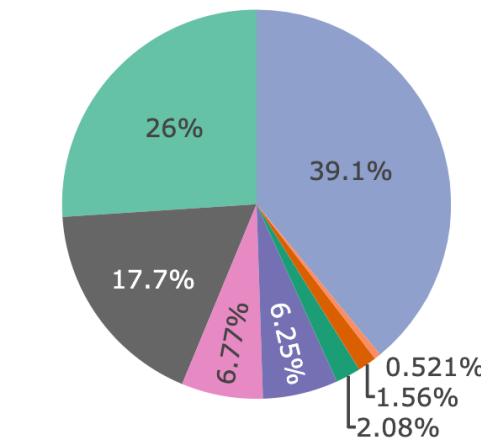


# AutoFocus

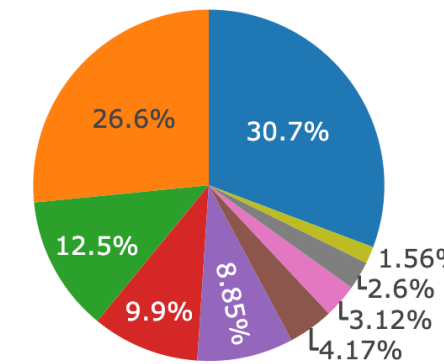


**Size**      **Densities**      **Platform Distribution**      **Pathway Distribution**

192      0.536

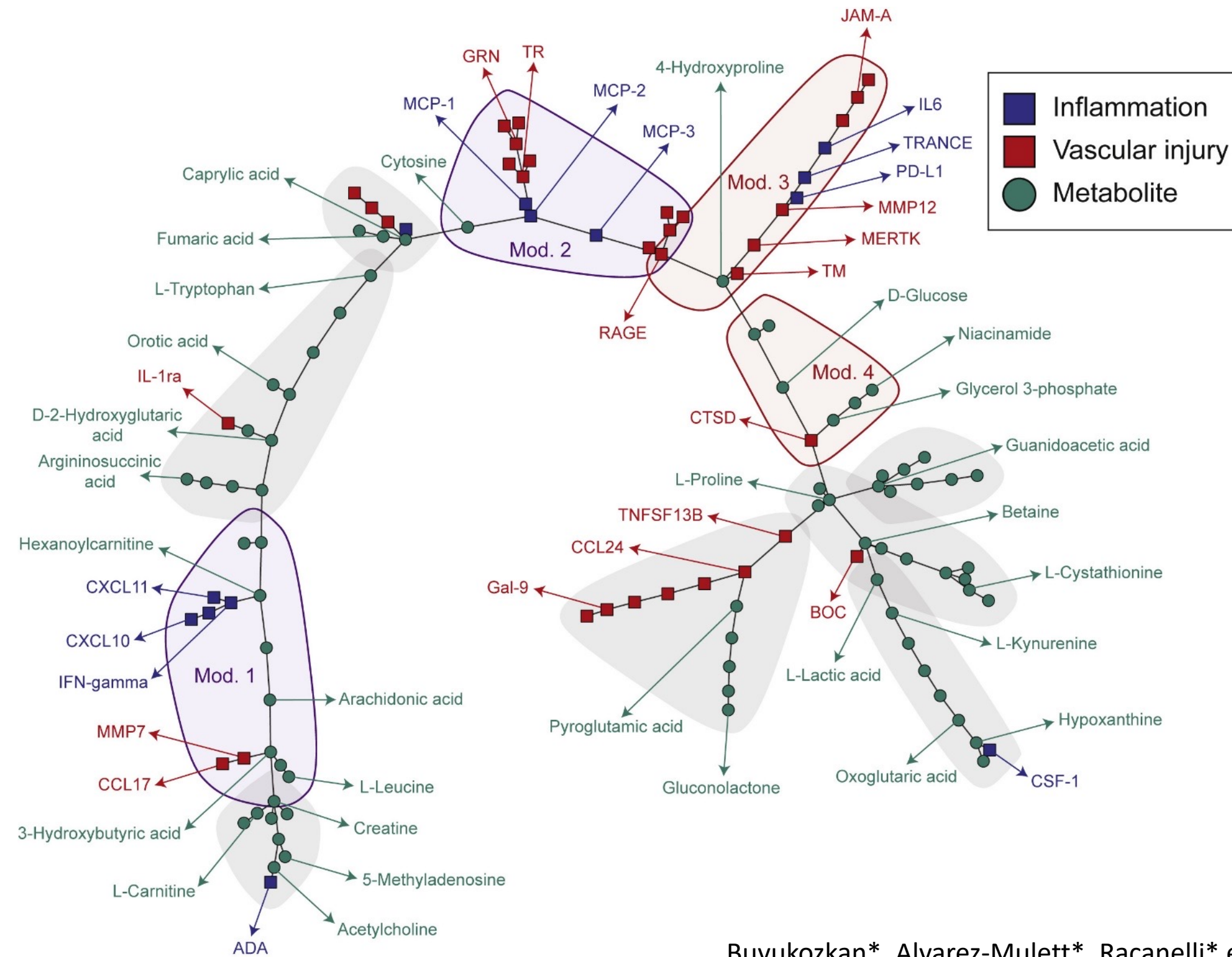


- Metabolon Urine
- Metabolon Plasma (HD4)
- Metabolon Plasma (HD2)
- Somalogic Proteomics
- Chenomix NMR Metabolomics
- Biocrates Metabolomics
- Brainshake Metabolomics
- Metabolon Saliva



- Amino acid
- No pathway information
- Carbohydrate
- Lipid
- Amino Acid
- Peptide
- Xenobiotics
- Energy
- Nucleotide

# Network integration



Buyukozkan\*, Alvarez-Mulett\*, Racanelli\* et al., *iScience*, 2022





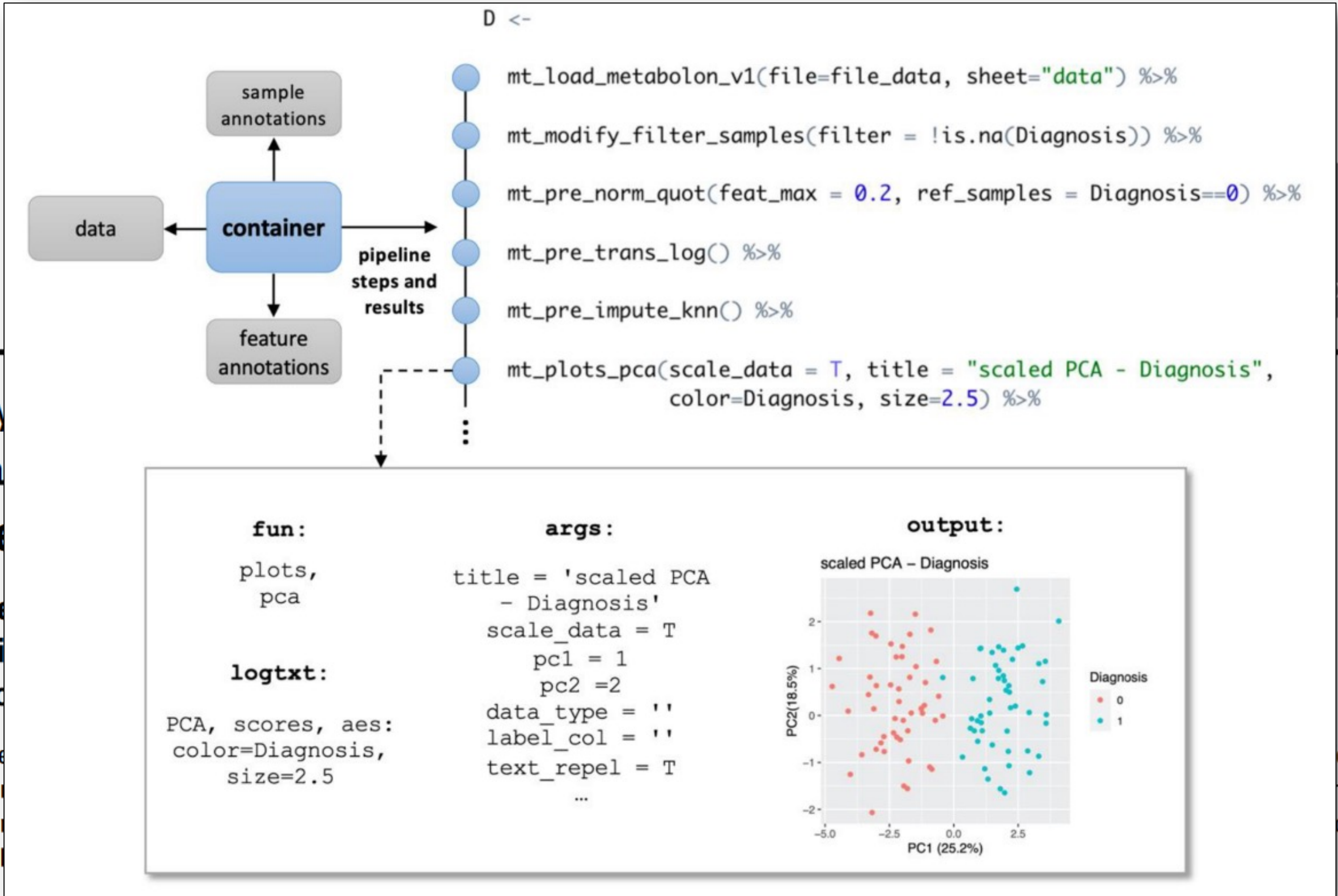
---

Systems biology

# **maplet: an extensible R toolbox for modular and reproducible metabolomics pipelines**

**Kelsey Chetnik<sup>1</sup>, Elisa Benedetti<sup>1</sup>, Daniel P. Gomari<sup>2</sup>, Annalise Schweickart<sup>1</sup>, Richa Batra<sup>1</sup>, Mustafa Buyukozkan<sup>1</sup>, Zeyu Wang<sup>1</sup>, Matthias Arnold <sup>2</sup>, Jonas Zierer<sup>1,†</sup>, Karsten Suhre <sup>3</sup> and Jan Krumsiek <sup>1,\*</sup>**

<sup>1</sup>Department of Physiology and Biophysics, Institute for Computational Biomedicine, Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY 10021, USA; <sup>2</sup>Institute of Computational Biology, Helmholtz Zentrum München—German Research Center for Environmental Health, Neuherberg, Germany and <sup>3</sup>Department of Physiology and Biophysics, Weill Cornell Medical College—Qatar Education City, Doha, Qatar



[Home](#) » [Bioconductor 3.16](#) » [Software Packages](#) » SummarizedExperiment

## SummarizedExperiment

platforms **all** rank **15 / 2183** support **3 / 4** in Bioc **7.5 years**  
build **ok** updated **before release** dependencies **24**

DOI: [10.18129/B9.bioc.SummarizedExperiment](https://doi.org/10.18129/B9.bioc.SummarizedExperiment)

### SummarizedExperiment container

Bioconductor version: Release (3.16)

The SummarizedExperiment container contains one or more assays, each represented by a matrix-like object of numeric or other mode. The rows typically represent genomic ranges of interest and the columns represent samples.

Author: Martin Morgan [aut], Valerie Obenchain [aut], Jim Hester [aut], Hervé Pagès [aut, cre]

Maintainer: Hervé Pagès <hpages.on.github at gmail.com>

Citation (from within R, enter `citation("SummarizedExperiment")`):

Morgan M, Obenchain V, Hester J, Pagès H (2022). *SummarizedExperiment: SummarizedExperiment container*. R package version 1.28.0, <https://bioconductor.org/packages/SummarizedExperiment>.

### Documentation »

#### *Bioconductor*

- Package [vignettes](#) and manuals.
- [Workflows](#) for learning and use.
- Several [online books](#) for comprehensive coverage of a particular research field, biological question, or technology.
- [Course and conference](#) material.
- [Videos](#).
- Community [resources](#) and [tutorials](#).

R / [CRAN](#) packages and [documentation](#)

### Support »

Please read the [posting guide](#). Post questions about Bioconductor to one of the following locations:

- [Support site](#) - for questions about Bioconductor packages
- [Bioc-devel](#) mailing list - for package

# Acknowledgements

Most of my present work is funded by



under the  
Biomedical Research Program (BMRP)  
and by multiple QNRF grants





**biobank**<sup>uk</sup>

Research Analysis  
Platform

Enabled by **DNAnexus**<sup>®</sup>

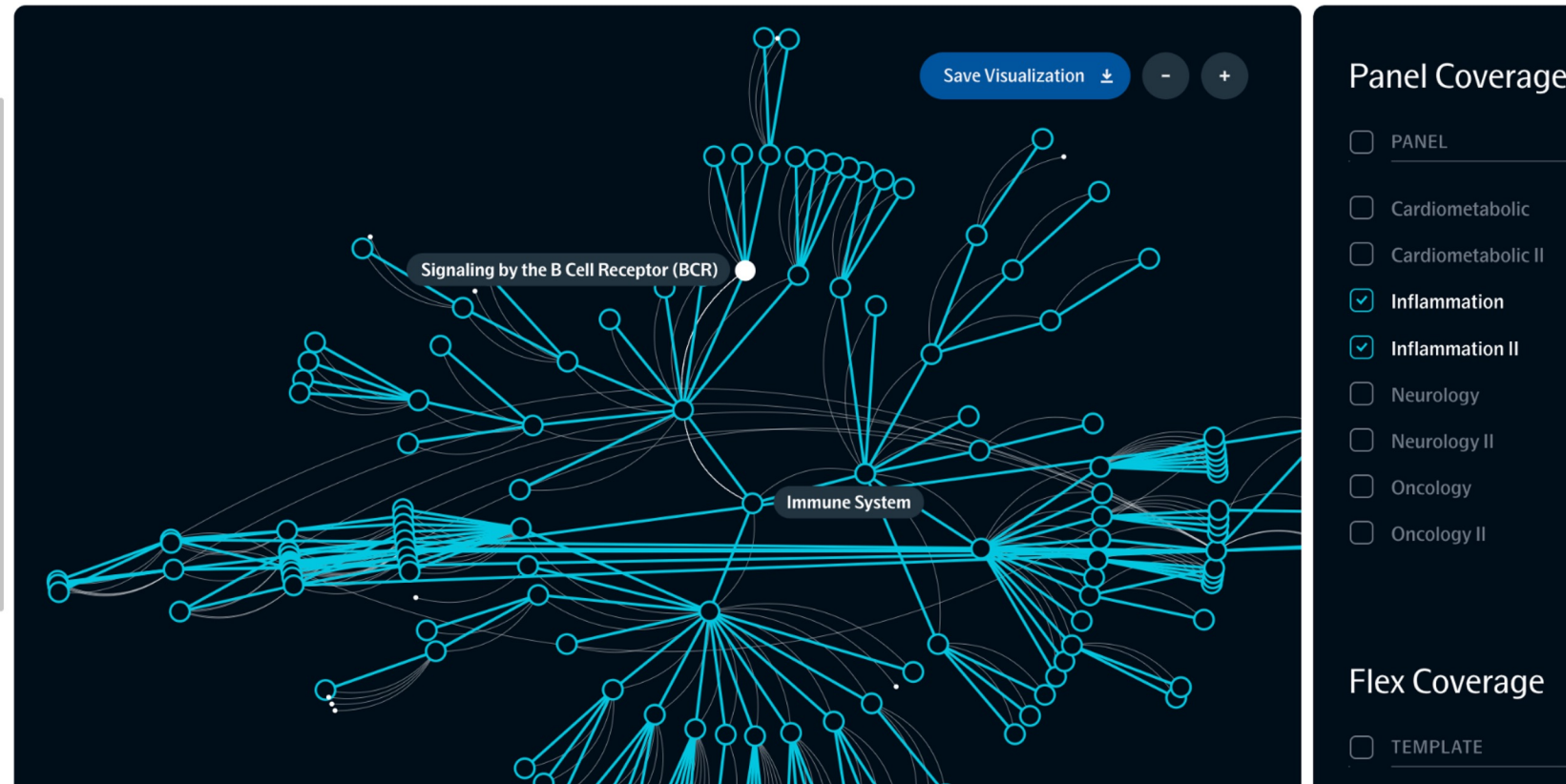
# Using Proteomics as Features for Disease Subtyping

# The UKB-RAP and Proteomics – Hypothesis Confirmation

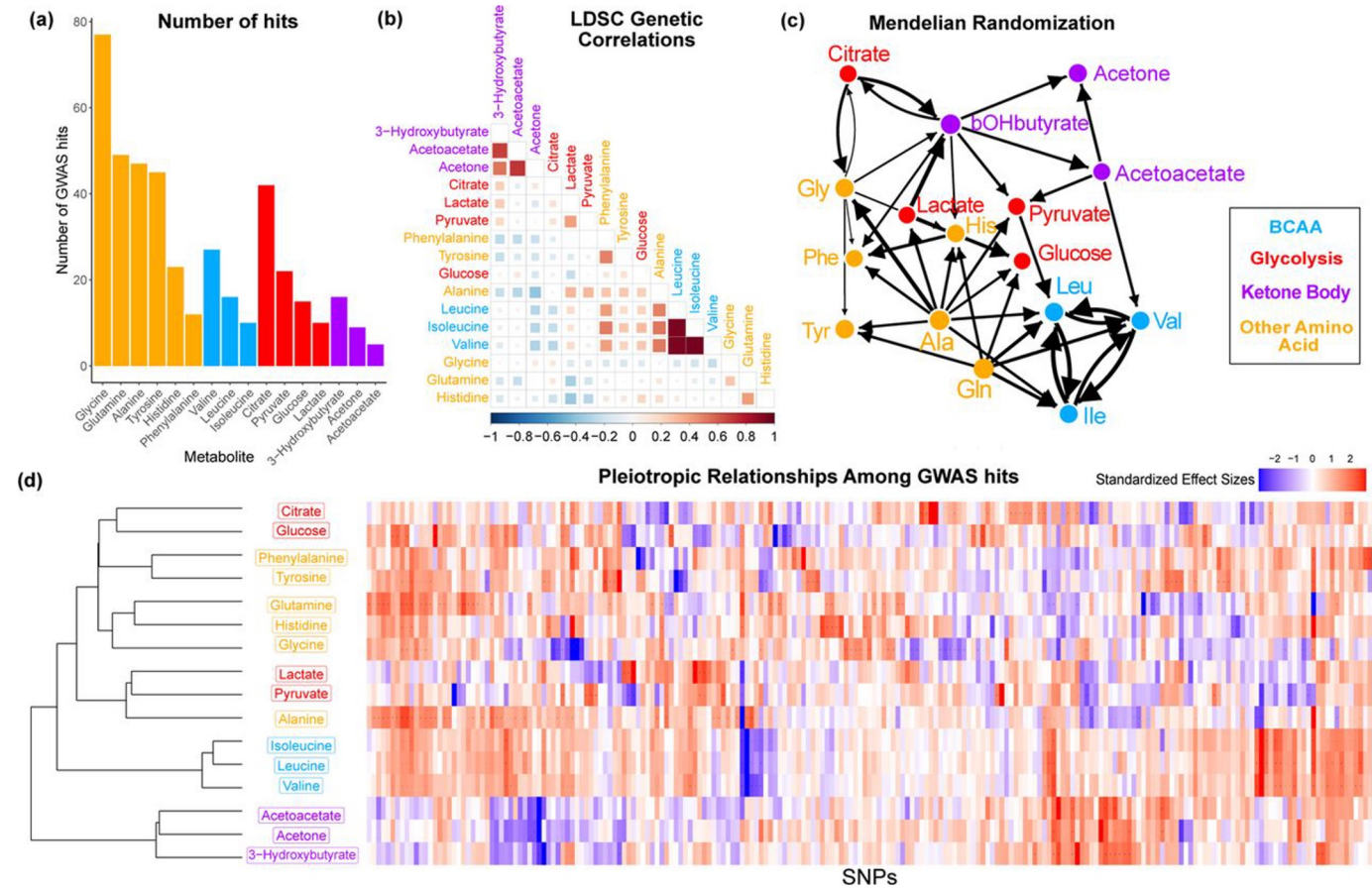
Search UniProt ID, Gene, Pathway or List

Pathway Browser

- Autophagy
- Cell Cycle
- Cell-Cell communication
- Cellular responses to stim...
- Chromatin organization
- Circadian Clock
- DNA Repair
- DNA Replication
- Developmental Biology
- Digestion and absorption
- Disease
- Drug ADME
- Extracellular matrix orga...
- Gene expression (Transcri...
- Hemostasis
- Immune System
- Metabolism
- Metabolism of RNA
- Metabolism of proteins
- Muscle contraction



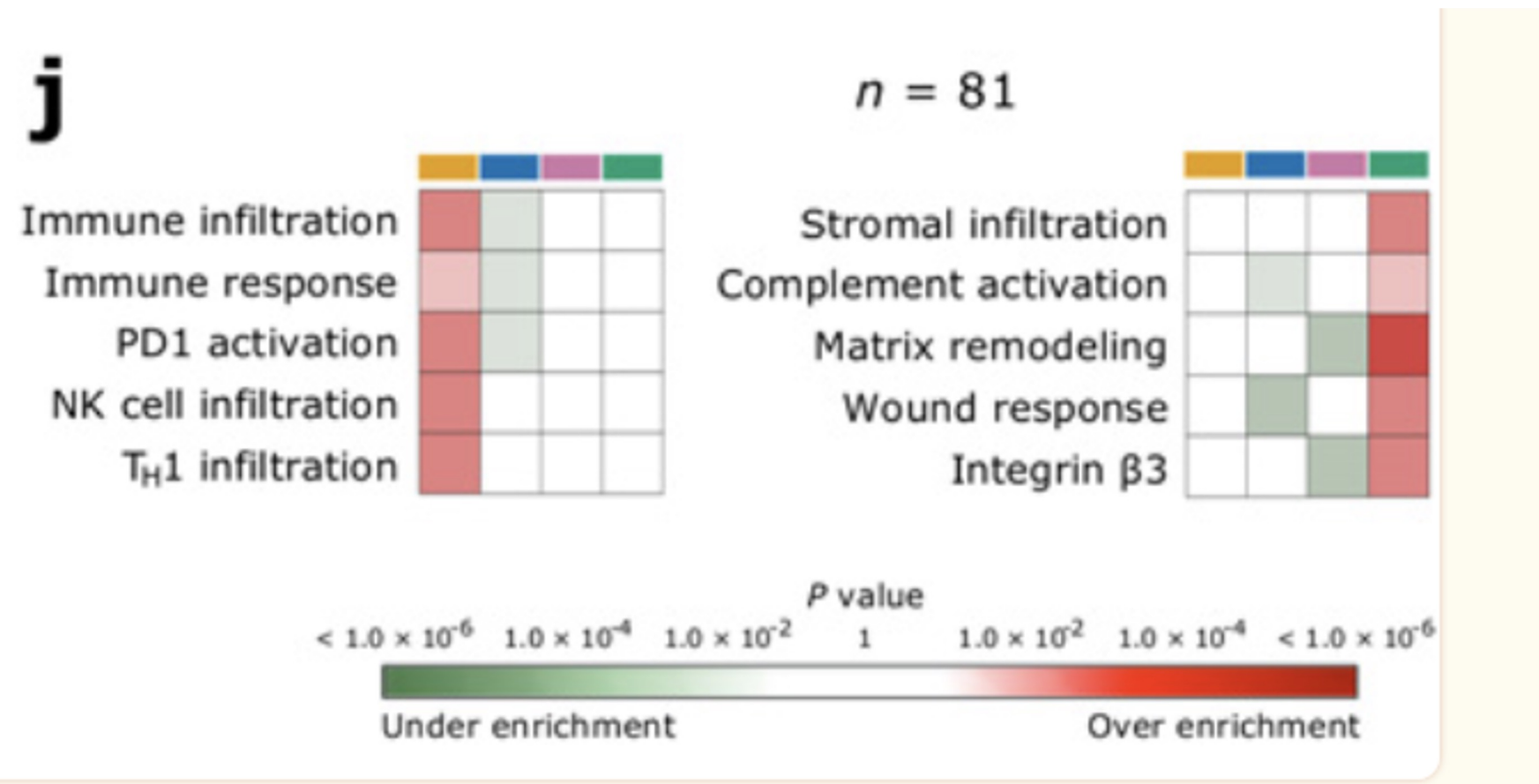
<https://insight.olink.com/pathway-browser>



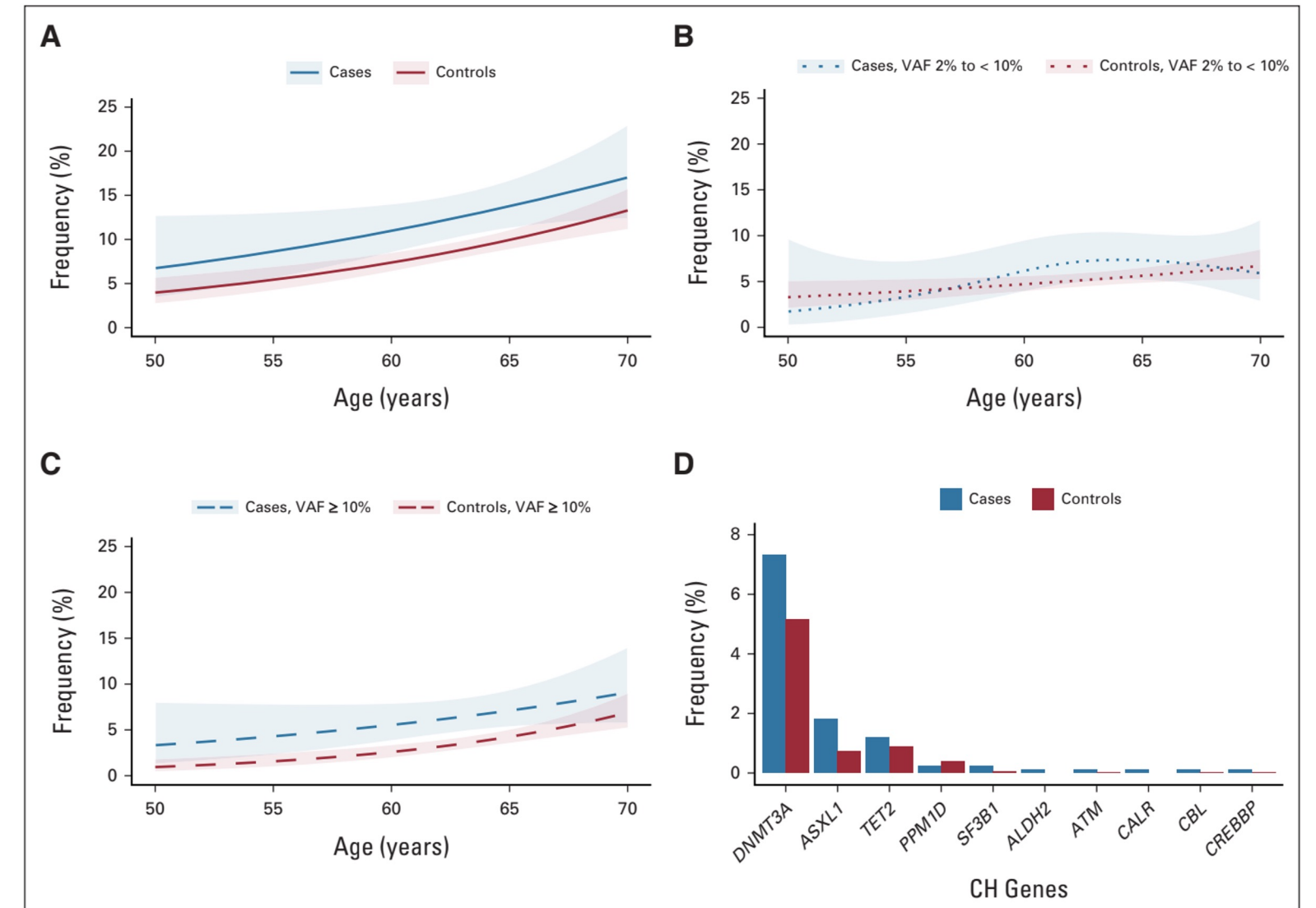
Smith, et al. <https://doi.org/10.1101/2022.04.02.486791>

# The UKB-RAP and Proteomics – Disease Subtyping

**j**



<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4636487/>



<https://ascopubs.org/doi/abs/10.1200/JCO.22.00857>



# Quick Tech Tips for Proteomics Data Analysis on the UKB-RAP

---

- Dispense to a new project instead of refreshing
- Work on the data alongside genomic data on the RAP
- Watch this webinar!



# Upcoming Webinars - Links in Related Content Section

- ▶ Oncology Researcher Roundtable: **May 25th**
- ▶ Analyzing the UK Biobank Proteomics Data on the UK Biobank Research Analysis Platform: **June 1st**
- ▶ [Find all Event announcements on the Community Forum](#)

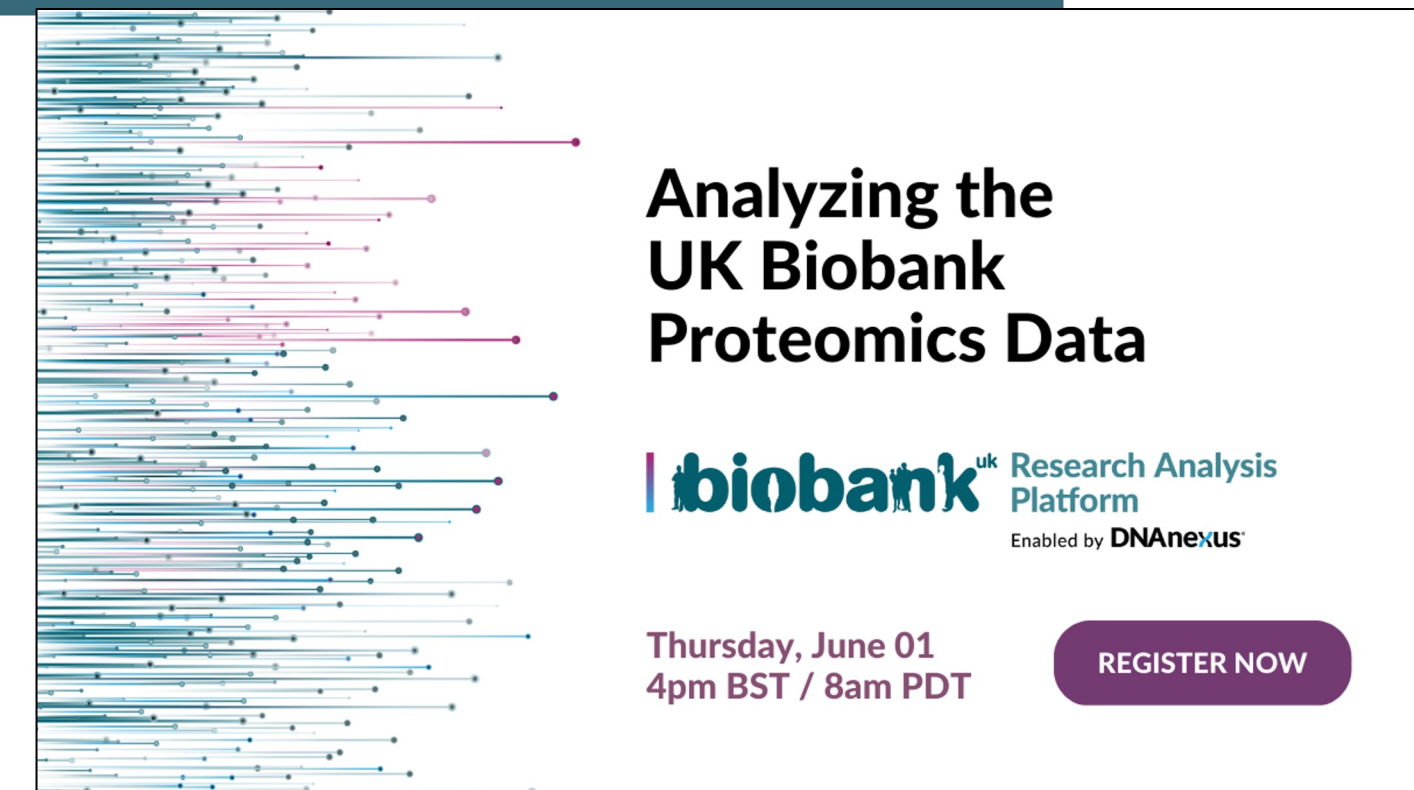


WEBINAR

**Oncology Researcher Roundtable:  
Working with Large-Scale Datasets  
to Enable Discovery**

Thursday, May 25th  
4:00 pm BST/8:00 am PDT

**biobank<sup>uk</sup>**  
Research Analysis  
Platform  
Enabled by **DNAexus**



**Analyzing the  
UK Biobank  
Proteomics Data**

**biobank<sup>uk</sup>** Research Analysis  
Platform  
Enabled by **DNAexus**

Thursday, June 01  
4pm BST / 8am PDT

**REGISTER NOW**

# Q&A