



Enabled by **DNA**nexus®

Analyzing UK Biobank proteomics data on the UKB-RAP

JUNE 2023

Speakers & Agenda

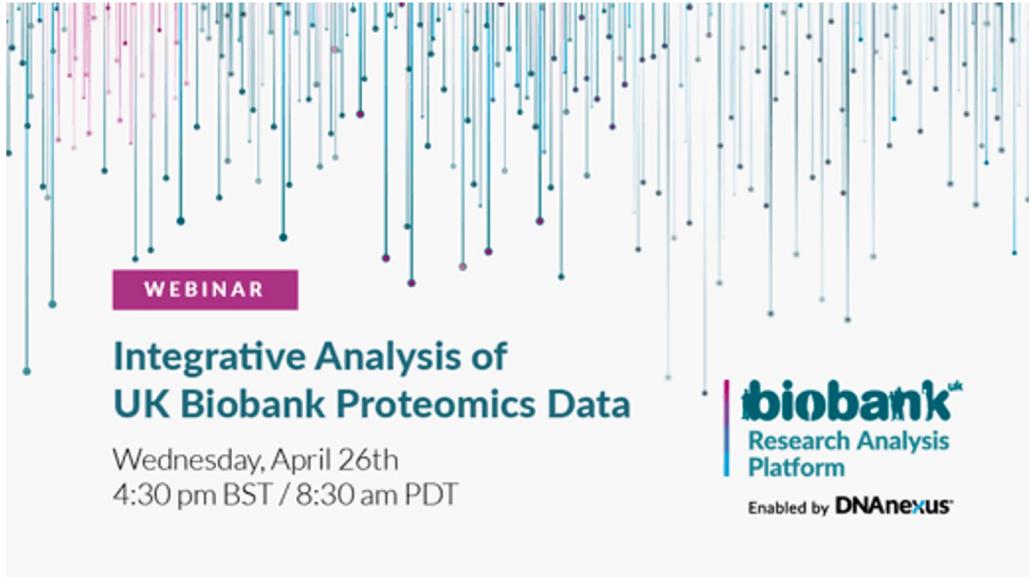


Alexandra Lee, PhD

Sr. Community Engagement/
Biomedical data scientist

1. New proteomics data on the UKB-RAP
2. How to access proteomics on UKB-RAP
3. Example: Differential expression analysis
4. Example: pQTL analysis

Previous proteomics webinar



<https://www.youtube.com/watch?v=btOYvmgwZGA>

Helpful resources

- ▶ [UKB Research analysis platform overview - webinar](#)
- ▶ [Introduction to JupyterLab notebooks on RAP - webinar](#)
- ▶ [End to end target discovery with GWAS and PheWAS on the UKB research analysis platform - webinar](#)

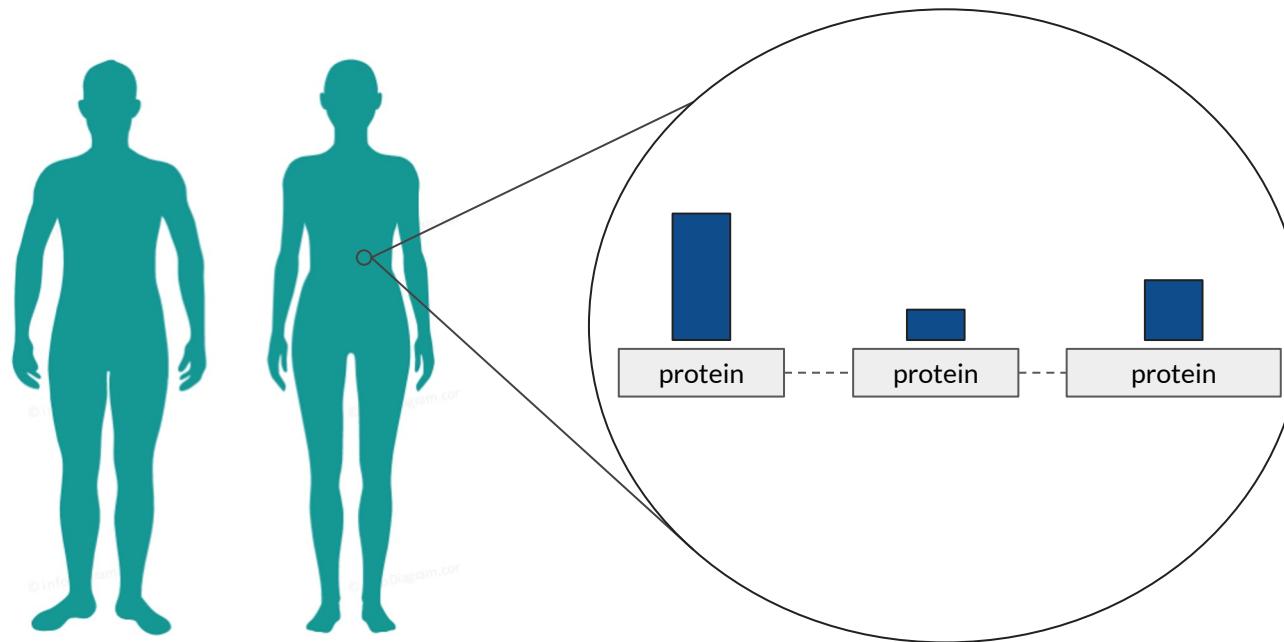
Learning Objectives

**By the end of this session,
you should be able to:**

- ▶ **Articulate** the proteomics data available on the UKB-RAP
- ▶ **Apply** steps how to extract and access proteomic data on the UKB-RAP
- ▶ **Apply** steps how to download and access analysis code on the UKB-RAP
- ▶ **Execute** various proteomic analyses with the tools available on the UKB-RAP
- ▶ **Access** the UKB-RAP community forum and additional courses

1. About Proteomics

Proteomics gives a snapshot of an organism's state



Proteomics data from UKB Pharma Proteomics Project (UKB-PPP)

biobank^{uk}
Enabling scientific discoveries that improve human health

The Pharma Proteomics Project

Proteins circulating in our blood may play a role in the development of many life-threatening diseases.

A greater understanding of such markers offers opportunities for more precise, targeted treatment.

53,000 UK Biobank participants

Analyse over 1,500 proteins

Measured by Olink

Genentech
Biogen

Bristol Myers Squibb[®]

AMGEN

AstraZeneca

REGENERON

gsk

Pfizer

Takeda

Janssen
INTERNATIONAL COMPANY OF
Johnson & Johnson

- Measured by Olink
- ~53,000 participants
- ~1,500 proteins

<https://www.biorxiv.org/content/10.1101/2022.06.17.496443v1.full.pdf>

Olink technology



	ACAN;Aggrecan core protein	ABHD14B;Protein ABHD14B	AARSD1;Alanyl-tRNA editing protein Aarsd1	ACTN4;Alpha-actin-4
sample 1	7.6	4.3	7.2	2.3
sample 2	6.9	6.7	6.8	1.7
sample 3	3.2	9.2	3.0	0.2

[White paper about Olink technology](#)

Olink technology

~53,000
samples

	ACAN;Aggrecan core protein	ABHD14B;Protein ABHD14B	AARSD1;Alanyl-tRNA editing protein Aarsd1	ACTN4;Alpha-actin-4
sample 1	7.6	4.3	7.2	2.3
sample 2	6.9	6.7	6.8	1.7
sample 3	3.2	9.2	3.0	0.2

[White paper about Olink technology](#)

Olink technology

Protein panels (~1,500 proteins):

- Inflammation
- Oncology
- Cardiometabolic
- Neurology

	ACAN;Aggrecan core protein	ABHD14B;Protein ABHD14B	AARSD1;Alanyl-tRNA editing protein Aarsd1	ACTN4;Alpha-actin-4
sample 1	7.6	4.3	7.2	2.3
sample 2	6.9	6.7	6.8	1.7
sample 3	3.2	9.2	3.0	0.2

[White paper about Olink technology](#)

Olink technology

	ACAN ;Aggrecan core protein	ABHD14B ;Protein ABHD14B	AARSD1 ;Alanyl-tRNA editing protein Aarsd1	ACTN4 ;Alpha-actin-4
sample 1	7.6	4.3	7.2	2.3
sample 2	6.9	6.7	6.8	1.7
sample 3	3.2	9.2	3.0	0.2

Normalized protein expression (NPX)

UKB-PPP preprint

2. How to access proteomics data on UKB RAP

Adding proteomics data to your project

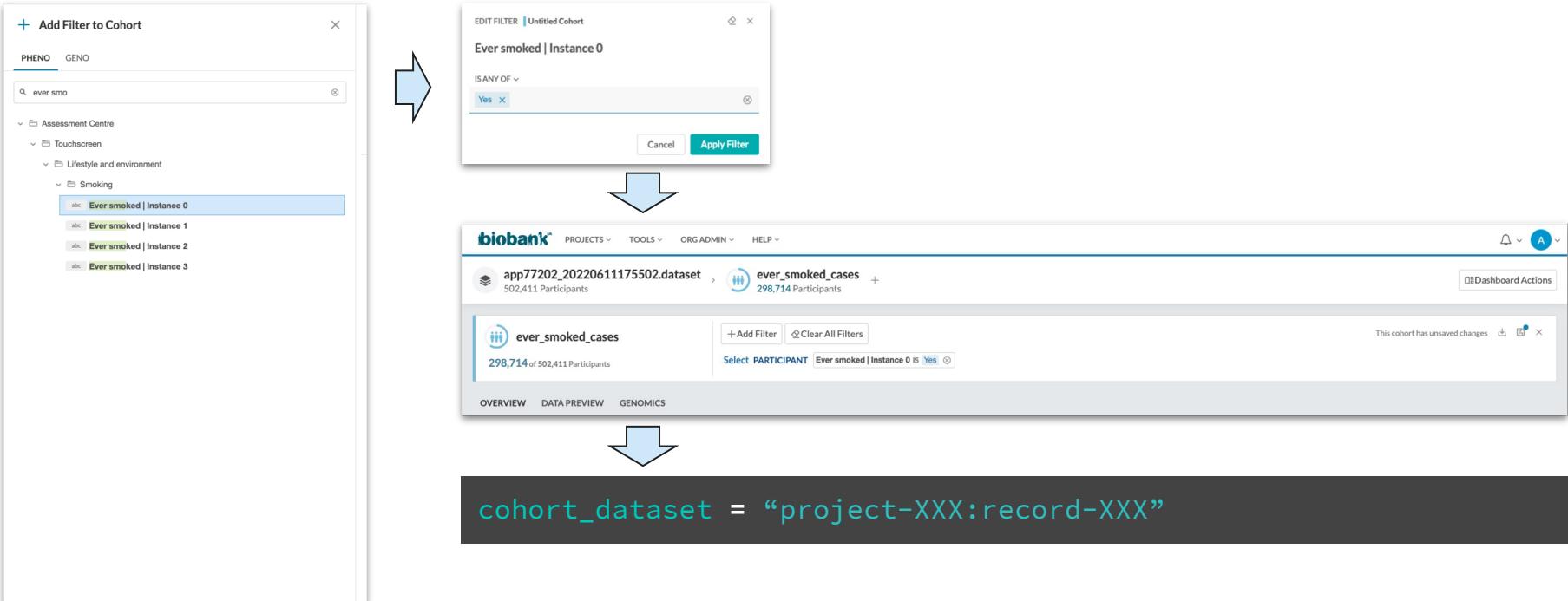
Note: Users will need to their UKB application be in Tier 2 or above in order to access this proteomics data

- ▶ Refresh data on existing project
- ▶ Create a new project and dispense proteomics data to this (*if refresh takes a while)

Collect data

1. Phenotype data
2. Protein expression data

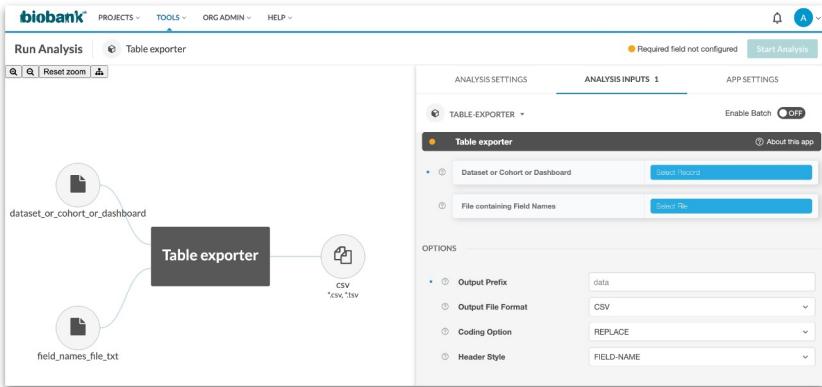
Get phenotype data from Cohort Browser



AD phenotype example
Ischemic disease example

Get protein expression data from Cohort Browser

Table Exporter App



dx extract_dataset

```
$ dx extract_dataset <dataset> --fields  
“entity1.field1, entity1.field2,  
entity2.field4”
```

Get protein expression data using Table Exporter app

The screenshot shows the 'Table exporter' configuration page. At the top right is an 'Enable Batch' toggle switch set to 'OFF'. Below it is an 'About this app' link.

Dataset or Cohort or Dashboard: ischaemic_cases

File containing Field Names: field_names.txt

OPTIONS

- Output Prefix:** (empty input field)
- Output File Format:** CSV
- Coding Option:** REPLACE
- Header Style:** FIELD-NAME

ADVANCED OPTIONS

- Entity:** olink_instance_0
- Field Names:** (empty input field)
- Field Titles:** (empty input field)

COHORT/DASHBOARD OPTIONS

- Cohort Table Entity Names:** (empty input field)
- Cohort Table Entity Titles:** (empty input field)

Get protein expression data using Table Exporter app

The screenshot shows the 'Table exporter' configuration page. At the top, there are two dropdown menus: 'Dataset or Cohort or Dashboard' set to 'ischaemic_cases' and 'File containing Field Names' set to 'field_names.txt'. A callout bubble highlights these fields with the text 'Phenotype dataset e.g. *.dataset'. Below these are sections for 'OPTIONS' and 'ADVANCED OPTIONS'. In 'OPTIONS', 'Output Prefix' is empty, 'Output File Format' is set to 'CSV', 'Coding Option' is 'REPLACE', and 'Header Style' is 'FIELD-NAME'. In 'ADVANCED OPTIONS', 'Entity' is 'olink_instance_0', 'Field Names' is empty, and 'Field Titles' is empty. Under 'COHORT/DASHBOARD OPTIONS', 'Cohort Table Entity Names' and 'Cohort Table Entity Titles' are both empty. At the top right, there is an 'Enable Batch' switch set to 'OFF'.

Enable Batch OFF

Phenotype dataset
e.g. *.dataset

Dataset or Cohort or Dashboard: ischaemic_cases

File containing Field Names: field_names.txt

OPTIONS

Output Prefix:

Output File Format: CSV

Coding Option: REPLACE

Header Style: FIELD-NAME

ADVANCED OPTIONS

Entity: olink_instance_0

Field Names:

Field Titles:

COHORT/DASHBOARD OPTIONS

Cohort Table Entity Names:

Cohort Table Entity Titles:

Get protein expression data using Table Exporter app

The screenshot shows the Table Exporter app interface. At the top, there's a dropdown menu labeled "TABLE-EXPORTER" and an "Enable Batch" toggle switch set to "OFF". Below that is a title bar with "Table exporter" and a "About this app" link. The main area has two sections: "Dataset or Cohort or Dashboard" (selected) and "File containing Field Names". Under "Dataset or Cohort or Dashboard", the value is "ischaemic_cases". Under "File containing Field Names", the value is "field_names.txt". A callout box points to this section with the text "List of field names e.g. eid, aarsd1,...". The "OPTIONS" section includes "Output Prefix" (empty), "Output File Format" (set to "CSV"), "Coding Option" (set to "REPLACE"), and "Header Style" (set to "FIELD-NAME"). The "ADVANCED OPTIONS" section includes "Entity" (set to "olink_instance_0"), "Field Names" (empty), and "Field Titles" (empty). The "COHORT/DASHBOARD OPTIONS" section includes "Cohort Table Entity Names" (empty) and "Cohort Table Entity Titles" (empty). A second callout box at the bottom right points to the "Field Names" and "Field Titles" fields with the text "List of field names".

Get protein expression data using Table Exporter app

The screenshot shows the 'Table exporter' app interface. At the top, there are two input fields: 'Dataset or Cohort or Dashboard' set to 'ischaemic_cases' and 'File containing Field Names' set to 'field_names.txt'. An 'Enable Batch' switch is turned off. Below these are sections for 'OPTIONS' and 'ADVANCED OPTIONS'. In the 'OPTIONS' section, there are four dropdowns: 'Output Prefix' (empty), 'Output File Format' (CSV), 'Coding Option' (REPLACE), and 'Header Style' (FIELD-NAME). In the 'ADVANCED OPTIONS' section, there are three dropdowns: 'Entity' (set to 'olink_instance_0'), 'Field Names' (empty), and 'Field Titles' (empty). A callout box highlights the 'Entity' field with the text 'entity table e.g. olink_instance_#'. At the bottom, there are sections for 'COHORT/DASHBOARD OPTIONS' with dropdowns for 'Cohort Table Entity Names' and 'Cohort Table Entity Titles'.

Get protein expression data using dx extract_dataset

```
cmd = ['dx',
       'extract_dataset',
       cohort_dataset,
       '--fields',
       'olink_instance_0.eid, olink_instance_0.aarsd1,...',
       '--delimiter',
       ',',
       '--output',
       'filename.csv',
     ]
subprocess.check_call(cmd)
```

Get protein expression data using dx extract_dataset

```
cmd = ['dx',
       'extract_dataset',
       cohort_dataset,
       '--fields',
       'olink_instance_0.eid, olink_instance_0.aarsd1,...',
       '--delimiter',
       ',',
       '--output',
       'filename.csv',
     ]
subprocess.check_call(cmd)
```

Specify phenotype dataset identifier
cohort_dataset = “project-XXX:record-XXX”

Get protein expression data using dx extract_dataset

```
cmd = ['dx',
       'extract_dataset',
       cohort_dataset,
       '--fields',
       'olink_instance_0.eid, olink_instance_0.aarsd1,...',
       '--delimiter',
       ',',
       '--output',
       'filename.csv',
     ]
subprocess.check_call(cmd)
```

List of field names to extract
entity_table.field_name

Get list of field names for all proteins

```
cmd = ["dx", "extract_dataset", dataset, "-ddd", "--delimiter", ","]
subprocess.check_call(cmd)

data_dict_df = pd.read_csv("data_dictionary.csv")
```

data_dictionary

	entity	name		title	units	coding_name
26	participant	p31		Sex	NaN	data_coding_9
27	participant	p34		Year of birth	years	NaN
23823	participant	p20107_i0	Illnesses of father Instance 0	0	NaN	data_coding_1010



entity_dictionary

	entity	entity_title	entity_description	entity_label_plural	entity_label_singular
0	participant	Participant	NaN	Participants	Participant
1	death	Death Record	NaN	Death Records	Death Record
2	death_cause	Death Cause Record	NaN	Death Cause Records	Death Cause Record
3	hesin	Hospitalization Record	NaN	Hospitalization Records	Hospitalization Record

coding_dictionary

	coding_name	code	meaning	concept	display_order	parent_code
87816	data_coding_1010	13	Prostate cancer	NaN	1	NaN
87817	data_coding_1010	12	Severe depression	NaN	2	NaN
87818	data_coding_1010	11	Parkinson's disease	NaN	3	NaN

Get list of field names for all proteins

```
field_names = list(data_dict_df.loc[data_dict_df["entity"] == "olink_instance_0", "name"].values)
```

	entity	name	type
28093	gp_scripts	drug_name	string
28094	gp_scripts	quantity	string
28095	olink_instance_0	eid	string
28096	olink_instance_0	aarsd1	float
28097	olink_instance_0	abhd14b	float
28098	olink_instance_0	abl1	float
28099	olink_instance_0	acaa1	float
28100	olink_instance_0	acan	float
28101	olink_instance_0	ace2	float
28102	olink_instance_0	acox1	float

Get list of field names for all proteins

```
field_names_str = [f"olink_instance_0.{f}" for f in field_names]  
field_names_query = ",".join(field_names_str)
```

Output

```
'olink_instance_0.eid,olink_instance_0.aarsd1,olink_instance_0.abhd14b,olink_instance_0.abl1,...
```

Get protein expression data using dx extract_dataset

```
cmd = ['dx',
       'extract_dataset',
       cohort_dataset,
       '--fields',
       'olink_instance_0.eid, olink_instance_0.aarsd1,...',
       '--delimiter',
       ',',
       '--output',
       'filename.csv',
     ]
subprocess.check_call(cmd)
```

Write result to [filename.csv](#)

[Notebook to extract proteomics data](#)

Sample protein expression data

~53,000
samples

~1,500 proteins

	ACAN	ABHD14B	AARSD1	ACTN4
sample 1	7.6	4.3	7.2	2.3
sample 2	6.9	6.7	6.8	1.7
sample 3	3.2	9.2	3.0	

npx values

Three protein expression datasets

Protein dataset	No. samples	Description	Entity Table
1	~53,000	Randomly selected from 500K UKB participants, pre-selected	olink_instance_0
2	~1,000	COVID-19 imaging first visit	olink_instance_2
3	~1,000	COVID-19 imaging second visit	olink_instance_3

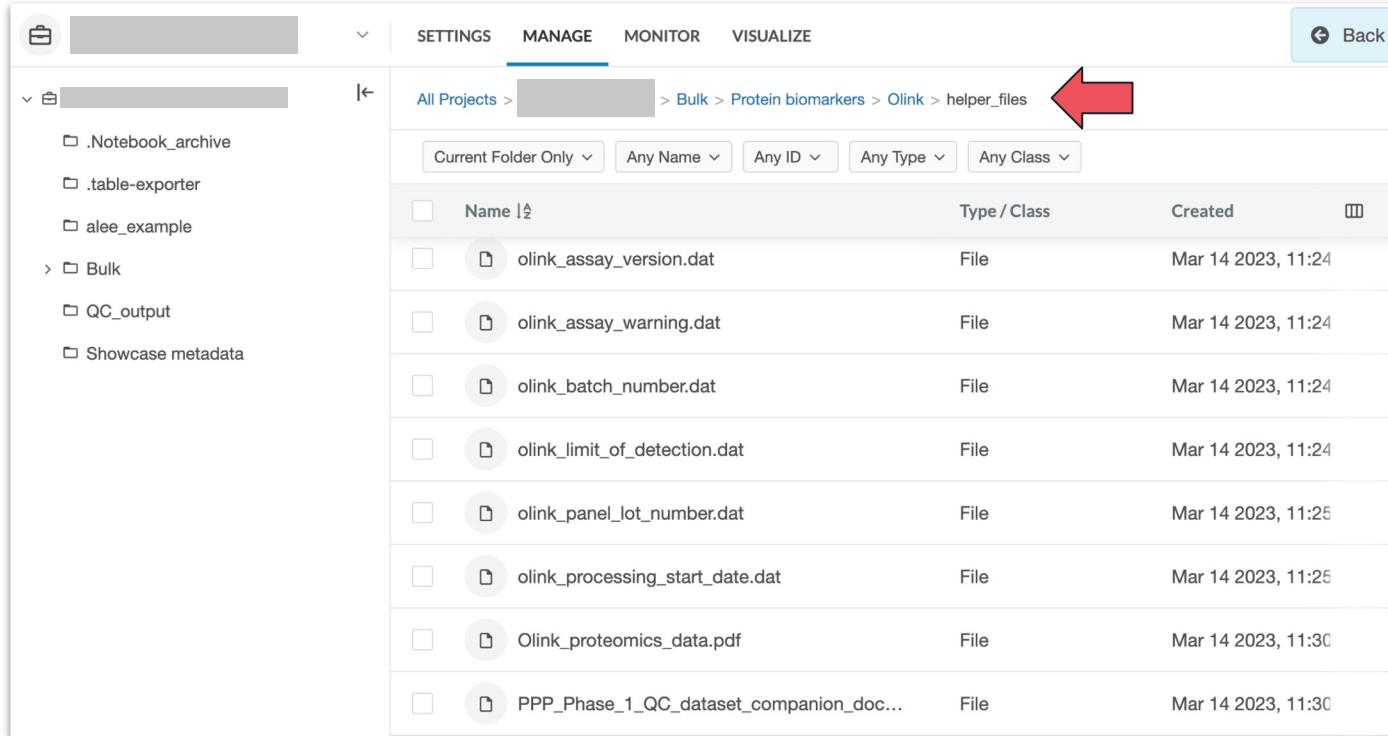
Three protein expression datasets

Protein dataset	No. samples	Description	Entity Table
1	~53,000	Randomly selected from 500K UKB participants, pre-selected	olink_instance_0
2	~1,000	COVID-19 imaging first visit	olink_instance_2
3	~1,000	COVID-19 imaging second visit	olink_instance_3

Three protein expression datasets

Protein dataset	No. samples	Description	Entity Table
1	~53,000	Randomly selected from 500K UKB participants, pre-selected	olink_instance_0
2	~1,000	COVID-19 imaging first visit	olink_instance_2
3	~1,000	COVID-19 imaging second visit	olink_instance_3

Metadata available in Bulk folder



The screenshot shows the DAnexus web interface with the 'MANAGE' tab selected. The left sidebar shows a project structure with 'Bulk' expanded. The main area displays a list of files in the 'helper_files' subfolder of 'Bulk'. The breadcrumb navigation bar at the top right shows the path: All Projects > [redacted] > Bulk > Protein biomarkers > Olink > helper_files. A large red arrow points to this breadcrumb bar.

<input type="checkbox"/>	Name	Type / Class	Created
<input type="checkbox"/>	olink_assay_version.dat	File	Mar 14 2023, 11:24
<input type="checkbox"/>	olink_assay_warning.dat	File	Mar 14 2023, 11:24
<input type="checkbox"/>	olink_batch_number.dat	File	Mar 14 2023, 11:24
<input type="checkbox"/>	olink_limit_of_detection.dat	File	Mar 14 2023, 11:24
<input type="checkbox"/>	olink_panel_lot_number.dat	File	Mar 14 2023, 11:25
<input type="checkbox"/>	olink_processing_start_date.dat	File	Mar 14 2023, 11:25
<input type="checkbox"/>	Olink_proteomics_data.pdf	File	Mar 14 2023, 11:30
<input type="checkbox"/>	PPP_Phase_1_QC_dataset_companion_doc...	File	Mar 14 2023, 11:30

Metadata available in Bulk folder

The screenshot shows a file browser interface with the following details:

- Left Sidebar:** Shows a tree view of project files:
 - .Notebook_archive
 - .table-exporter
 - alee_example
 - Bulk** (selected)
 - QC_output
 - Showcase metadata
- Top Navigation:** SETTINGS, MANAGE (selected), MONITOR, VISUALIZE, Back to [Project]
- Breadcrumb:** All Projects > [Project] > Bulk > Protein biomarkers > Olink > helper_files
- Filter Bar:** Current Folder Only, Any Name, Any ID, Any Type, Any Class
- Table View:** Displays the following files:

	Name	Type / Class	Created
<input type="checkbox"/>	olink_assay_version.dat	File	Mar 14 2023, 11:24
<input type="checkbox"/>	olink_assay_warning.dat	File	Mar 14 2023, 11:24
<input type="checkbox"/>	olink_batch_number.dat	File	Mar 14 2023, 11:24
<input type="checkbox"/>	olink_limit_of_detection.dat	File	Mar 14 2023, 11:24
<input type="checkbox"/>	olink_panel_lot_number.dat	File	Mar 14 2023, 11:25
<input type="checkbox"/>	olink_processing_start_date.dat	File	Mar 14 2023, 11:25
<input type="checkbox"/>	Olink_proteomics_data.pdf	File	
<input type="checkbox"/>	PPP_Phase_1_QC_dataset_companion_doc...	File	

A callout box labeled "Limit of detection" points to the "olink_limit_of_detection.dat" file. Another callout box labeled "PDF of QC steps performed" points to the "Olink_proteomics_data.pdf" file.

3. How to access analysis code on UKB-RAP

Code is available on github!

A screenshot of a GitHub repository page. The repository is named `dnanexus/UKB_RAP` and is public. The main navigation bar includes links for Pull requests, Issues, Codespaces, Marketplace, and Explore. Below the repository name, there are buttons for Edit Pins, Unwatch (12), Fork (21), and Starred (40). The navigation tabs at the top of the page are Code (selected), Issues (1), Pull requests, Actions, Projects, Wiki, Security, and Insights. The URL `https://github.com/dnanexus/UKB_RAP/tree/main/proteomics` is visible in the browser's address bar.

https://github.com/dnanexus/UKB_RAP/tree/main/proteomics

Add the analysis scripts to the UKB-RAP

Steps to perform on your local machine:

1. Clone the repository:
2. Navigate into the cloned repository:
3. Login to UKB-RAP:
4. Upload analysis scripts to the platform:

```
$git clone https://github.com/dnanexus/UKB_RAP.git  
$cd UKB_RAP  
$dx login  
$dx upload -r <proteomics> --destination <path on the UKB-RAP>
```

Add the analysis scripts to the UKB-RAP

Steps to perform on your local machine:

1. Clone the repository:
2. Navigate into the cloned repository:
3. Login to UKB-RAP:
4. Upload analysis scripts to the platform:

Clone repository to local machine

```
$git clone https://github.com/dnanexus/UKB\_RAP.git
```

Add the analysis scripts to the UKB-RAP

Steps to perform on your local machine:

1. Clone the repository:
2. Navigate into the cloned repository:
3. Login to UKB-RAP:
4. Upload analysis scripts to the platform:

```
$git clone https://github.com/dnanexus/UKB_RAP.git  
$cd UKB_RAP
```

Add the analysis scripts to the UKB-RAP

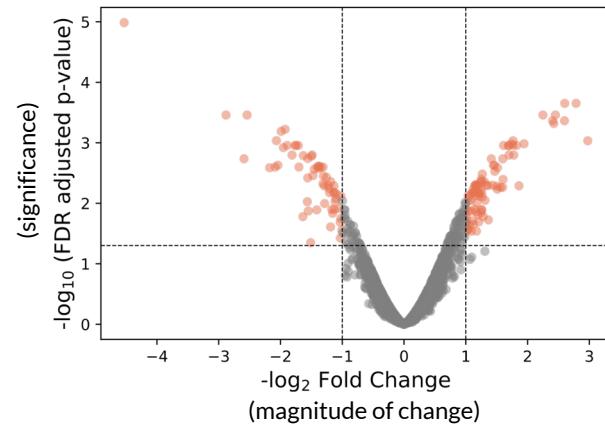
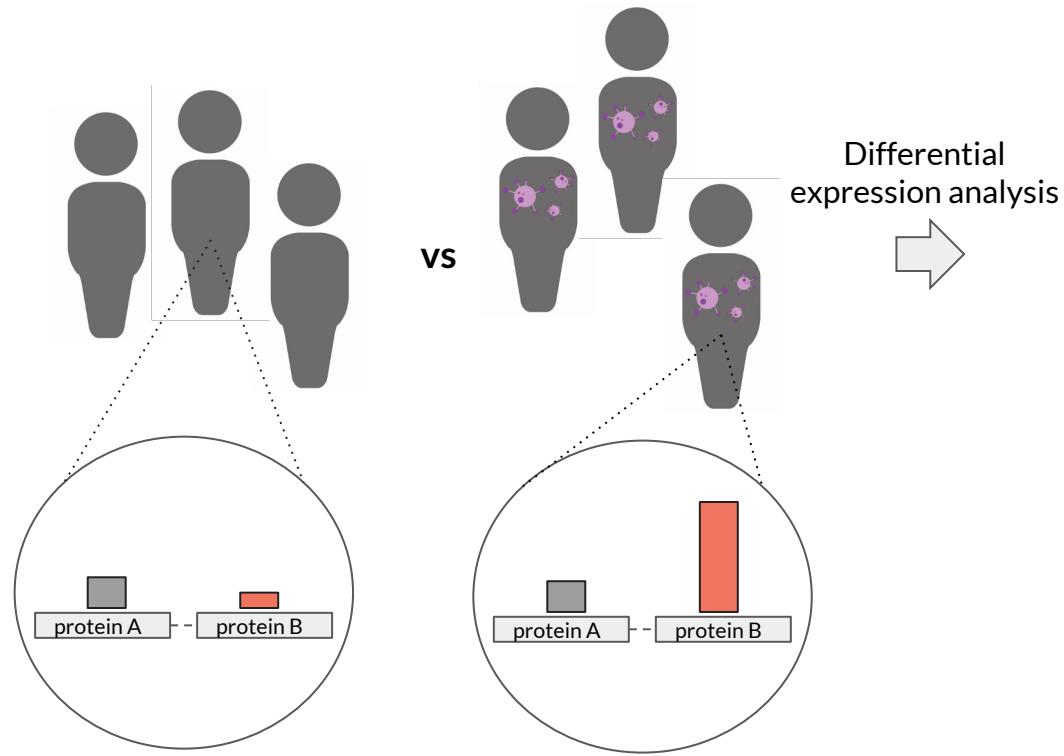
Steps to perform on your local machine:

1. Clone the repository:
2. Navigate into the cloned repository:
3. Login to UKB-RAP:
4. Upload analysis scripts to the platform:

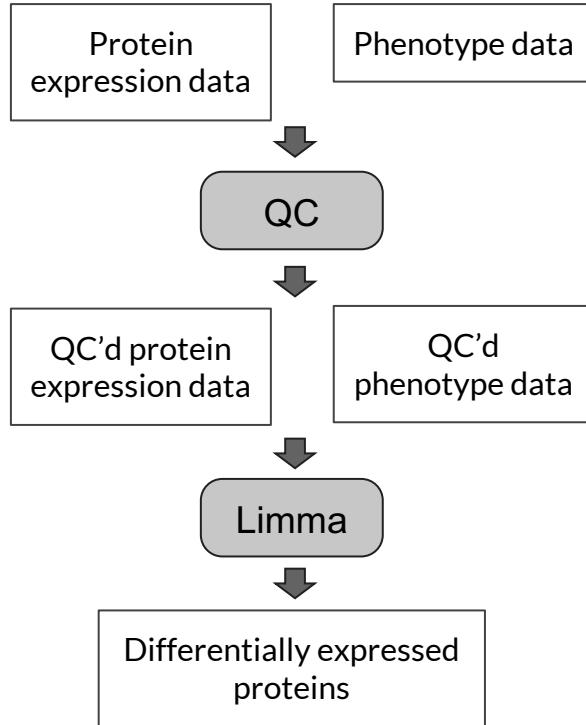
```
$git clone https://github.com/dnanexus/UKB_RAP.git  
$cd UKB_RAP  
$dx login  
$dx upload -r <proteomics> --destination <path on the UKB-RAP>
```

4. Example: Differential expression analysis

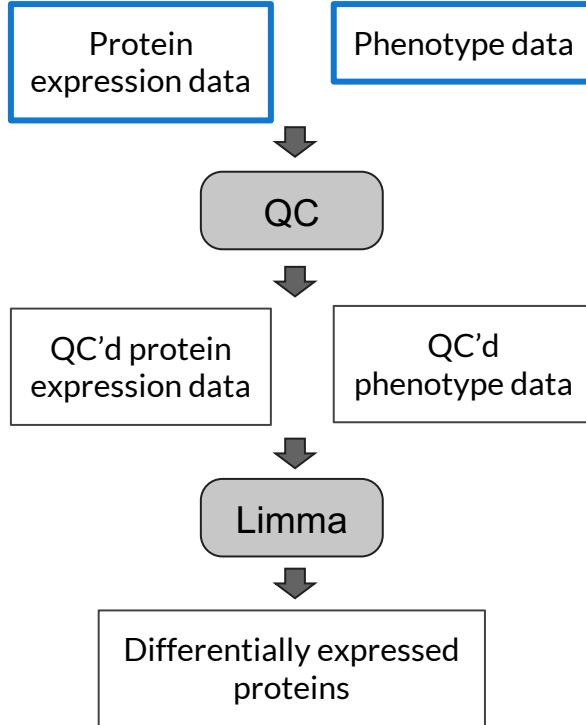
Differential expression analysis used to study mechanisms of disease



Approach

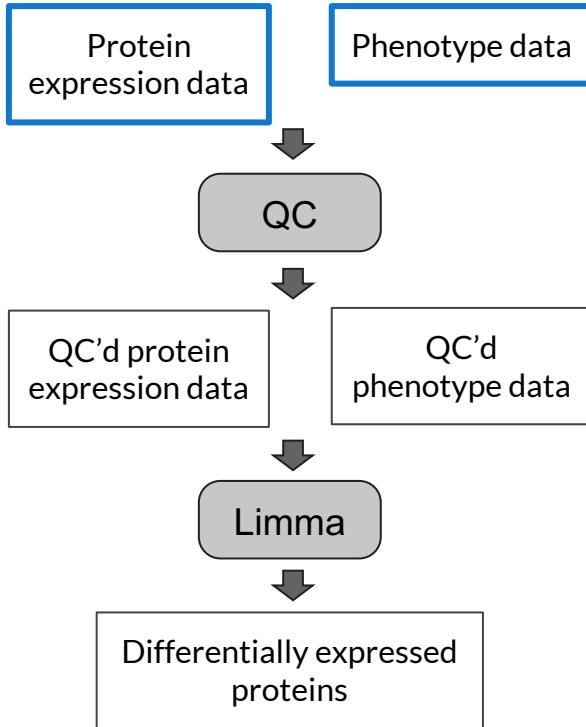


Approach



NOTE: We are using public proteomic data, not UKB data, for demonstration purposes

Collect input data

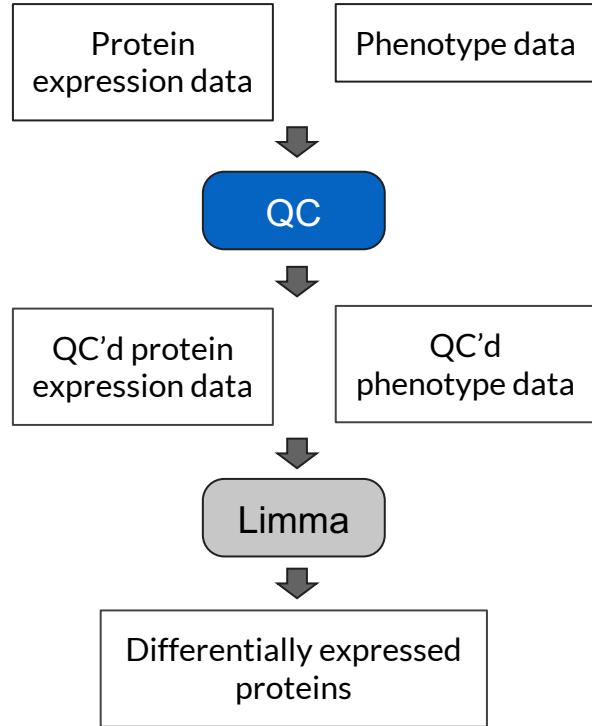


	CA1	ICAM1	CHL1	TGFBI	ENG
Plasma_Sample					
H0529.3	7.62107	6.79971	4.73174	9.33471	3.12445
H0441.1	6.96085	6.98459	4.31338	9.06819	3.31576
H0558.3	7.16983	7.04907	4.72713	8.92804	3.16308
H0499.2	7.59577	6.80282	4.51559	9.17979	3.19292
H0468.3	7.25945	6.91728	4.84307	9.91809	3.47692

	PIDN	Age_at_Baseline	Sex	Outcome
Plasma_Sample				
H0529.3	9677	90+	Male	MCI_Dcline_AD
H0441.1	9974	90+	Female	MCI_Stable_AD
H0558.3	9681	90+	Female	MCI_Dcline_AD
H0499.2	9502	88	Male	MCI_Stable_AD
H0468.3	9635	87	Female	MCI_Stable_AD

Input data from [Kivisakk et al](#)

QC input data



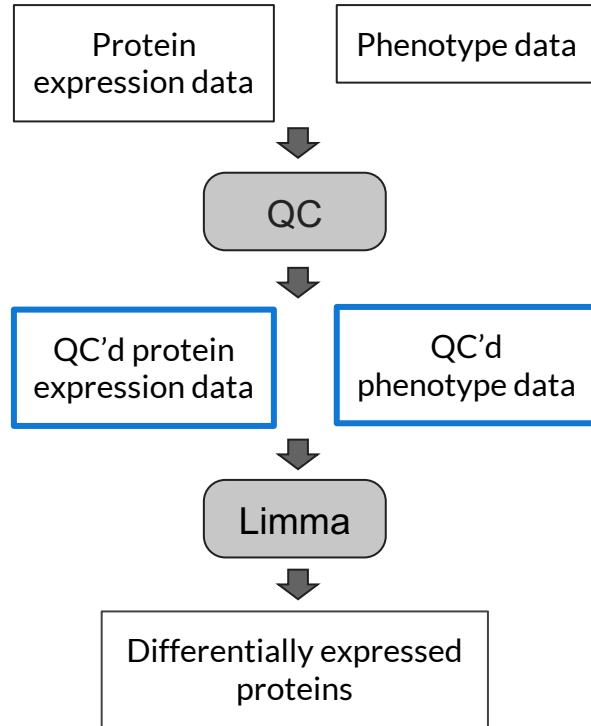
A funnel icon is positioned above two tables of data. The top table shows numerical values for CA1, ICAM1, CHL1, TGFBI, and ENG across five samples (H0529.3, H0441.1, H0558.3, H0499.2, H0468.3). The bottom table shows categorical information for PIDN, Age_at_Baseline, Sex, and Outcome for the same samples.

Plasma_Sample	CA1	ICAM1	CHL1	TGFBI	ENG
H0529.3	7.62107	6.79971	4.73174	9.33471	3.12445
H0441.1	6.96085	6.98459	4.31338	9.06819	3.31576
H0558.3	7.16983	7.04907	4.72713	8.92804	3.16308
H0499.2	7.59577	6.80282	4.51559	9.17979	3.19292
H0468.3	5.045	6.01729	8.107	0.0112	3.17612

- Removing missing, outlier data
- Normalize/scale data

Plasma_Sample	PIDN	Age_at_Baseline	Sex	Outcome
H0529.3	9677	90+	Male	MCI_Decline_AD
H0441.1	9974	90+	Female	MCI_Stable_AD
H0558.3	9681	90+	Female	MCI_Decline_AD
H0499.2	9502	88	Male	MCI_Stable_AD
H0468.3	9635	87	Female	MCI_Stable_AD

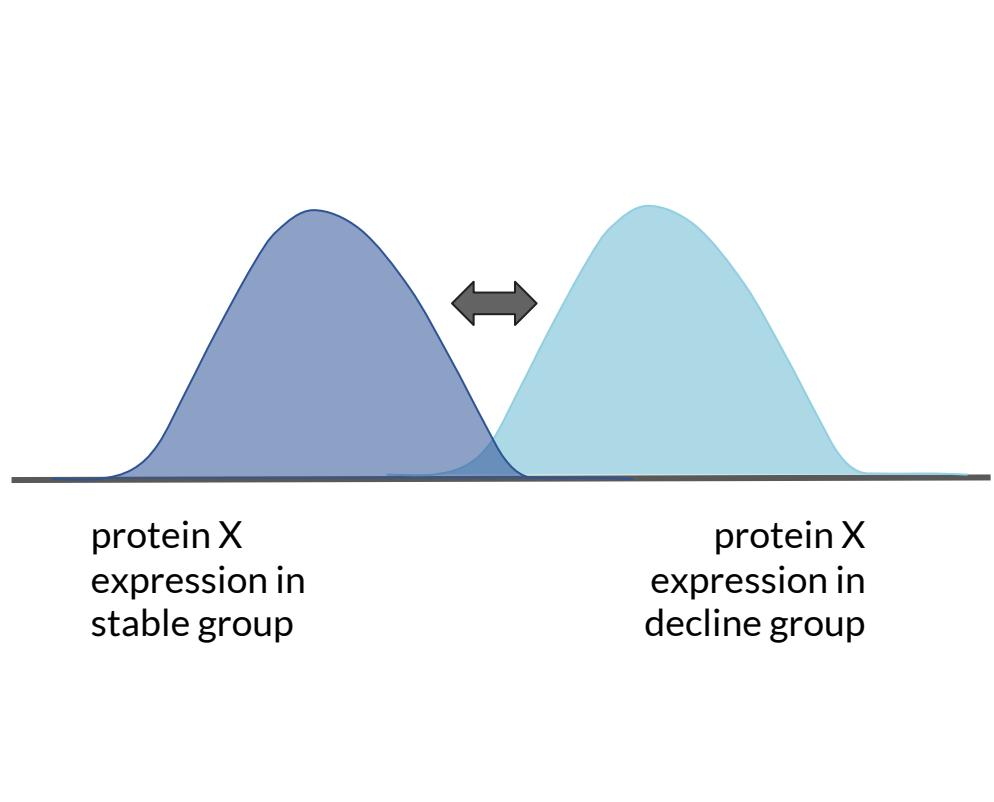
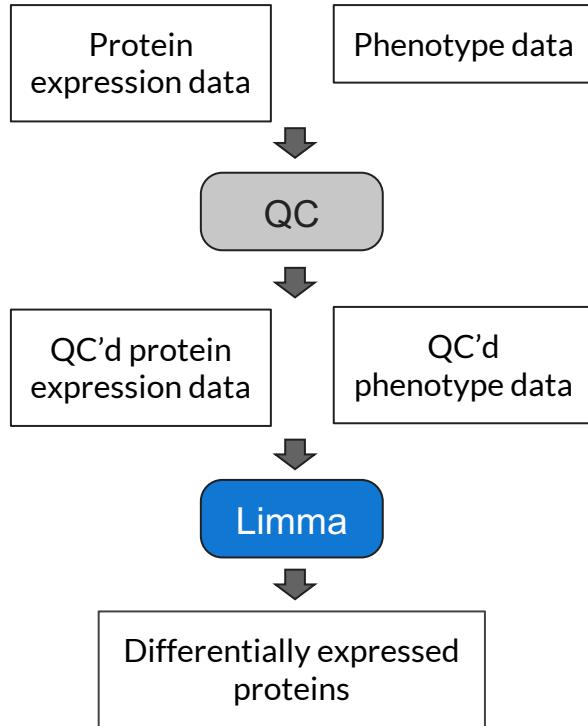
Get QC'd input data



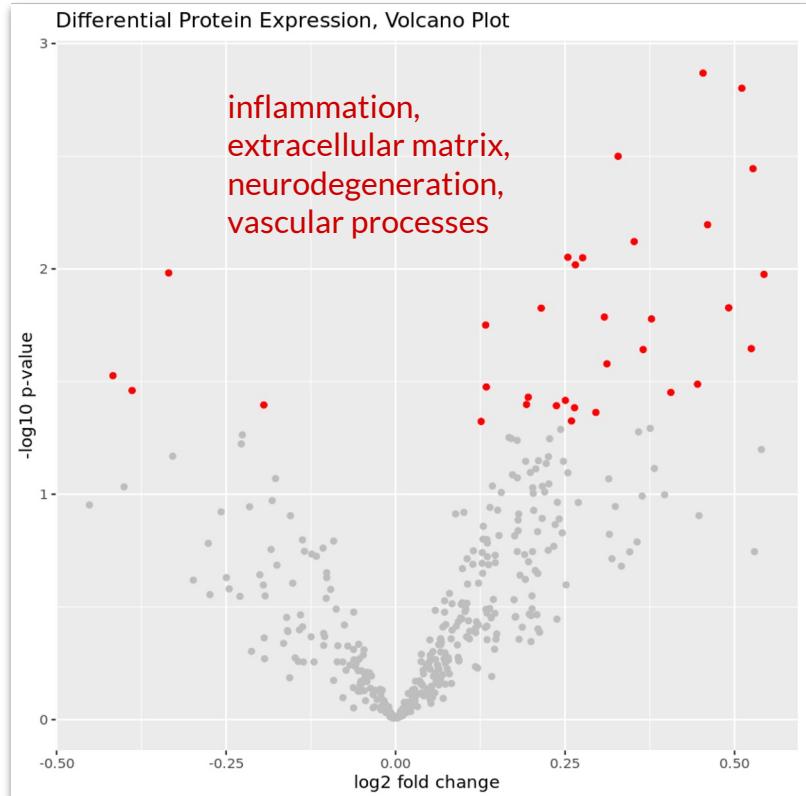
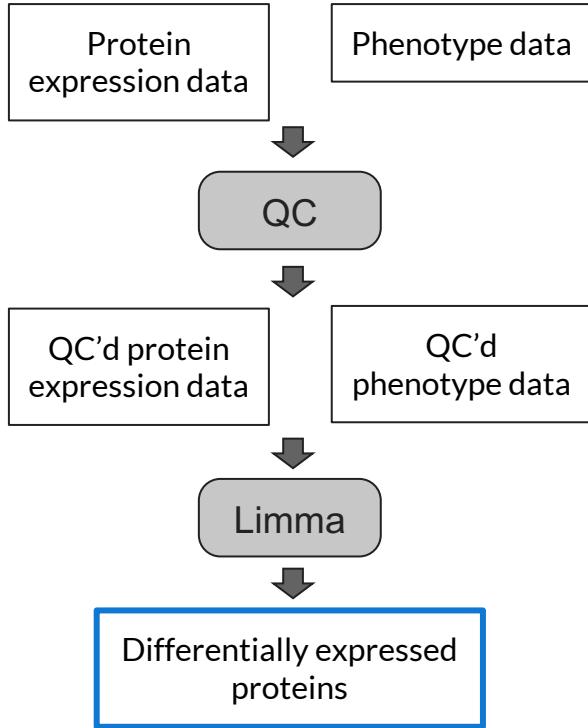
	CA1	ICAM1	CHL1	TGFBI	ENG
Plasma_Sample					
H0529.3	7.62107	6.79971	4.73174	9.33471	3.12445
H0441.1	6.96085	6.98459	4.31338	9.06819	3.31576
H0558.3	7.16983	7.04907	4.72713	8.92804	3.16308
H0499.2	7.59577	6.80282	4.51559	9.17979	3.19292
H0468.3	7.25945	6.91728	4.84307	9.91809	3.47692

	PIDN	Age_at_Baseline	Sex	Outcome
Plasma_Sample				
H0529.3	9677	90+	Male	MCI_Decline_AD
H0441.1	9974	90+	Female	MCI_Stable_AD
H0558.3	9681	90+	Female	MCI_Decline_AD
H0499.2	9502	88	Male	MCI_Stable_AD
H0468.3	9635	87	Female	MCI_Stable_AD

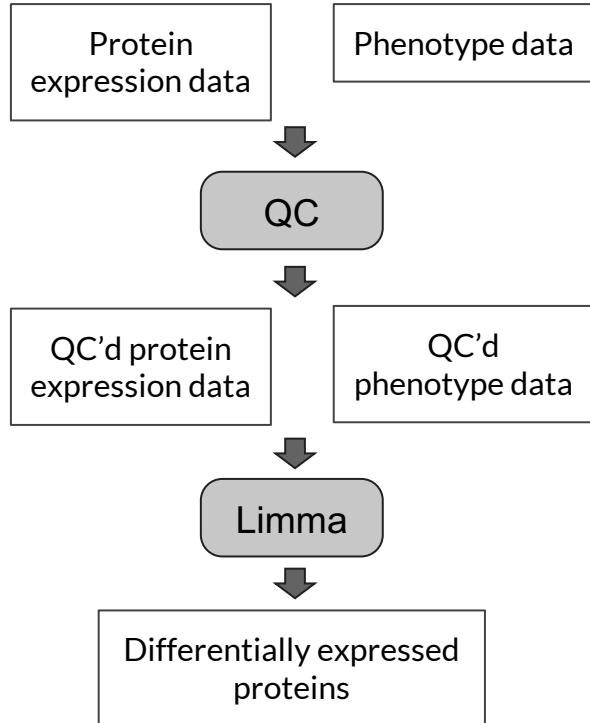
Perform differential expression analysis using Limma



Found differentially expressed proteins



Run analysis using JupyterLab



The screenshot shows the biobank.uk web interface. At the top, there are navigation links: PROJECTS, TOOLS, ORG ADMIN, and HELP. Below these are sections for Tool Library, IANAGE, and MON. A red circle highlights the 'JupyterLab' button in the Tool Library section. An arrow points from this circle to a larger window titled '+ New JupyterLab'. This window contains the following fields:

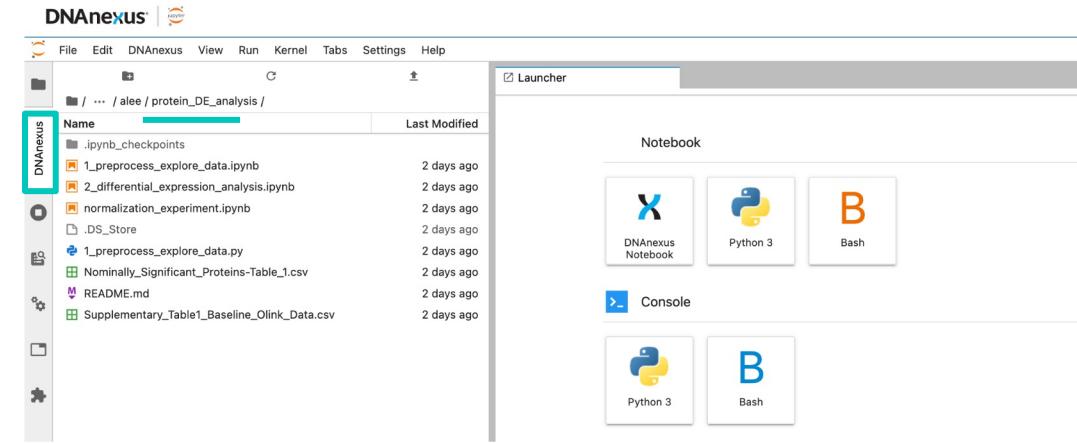
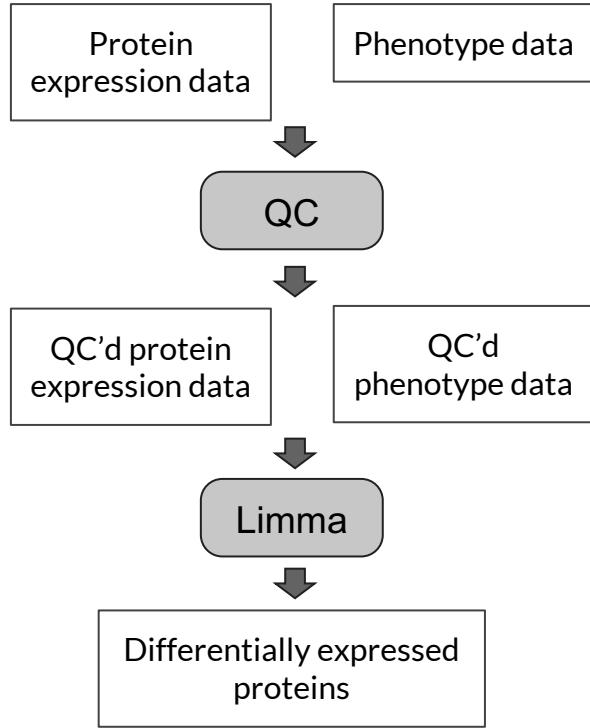
Environment Name	JupyterLab - 3/24/2023 8:59 AM
Project	White
Snapshot (Optional)	Select Snapshot
Priority	High
Cluster Configuration	Single Node
Instance Type	mem1_ssd1_v2_x2
Duration (in hours)	4
Feature	ML

At the bottom of the window, it says 'Estimated Price: £0.1984 based on instance type, duration, priority, number of nodes'.



JupyterLab webinar

Run analysis using JupyterLab



Resources

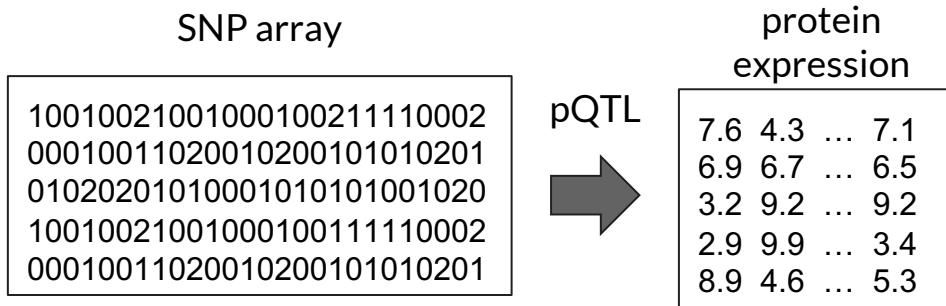
	Link	Configuration	Runtime & cost
Code to explore input data	https://github.com/dnanexus/UKB_RAP/blob/main/proteomics/protein DE analysis/1 preprocess_explore_data.ipynb	Kernel: ML Priority: normal Recommended instance: mem1(ssd1_v2_x2)	Runtime: ~ 1min Cost: ~£ 0.082
Code to perform differential expression	https://github.com/dnanexus/UKB_RAP/blob/main/proteomics/protein DE analysis/2 differential_expression_analysis.ipynb	Kernel: PYTHON_R Priority: normal Recommended instance: mem1(ssd1_v2_x2)	Runtime: ~ 5 min Cost: ~£0.015
Input data publication	https://academic.oup.com/braincomms/article/4/4/fcac155/6608340?login=false#36642284		

5 minute break

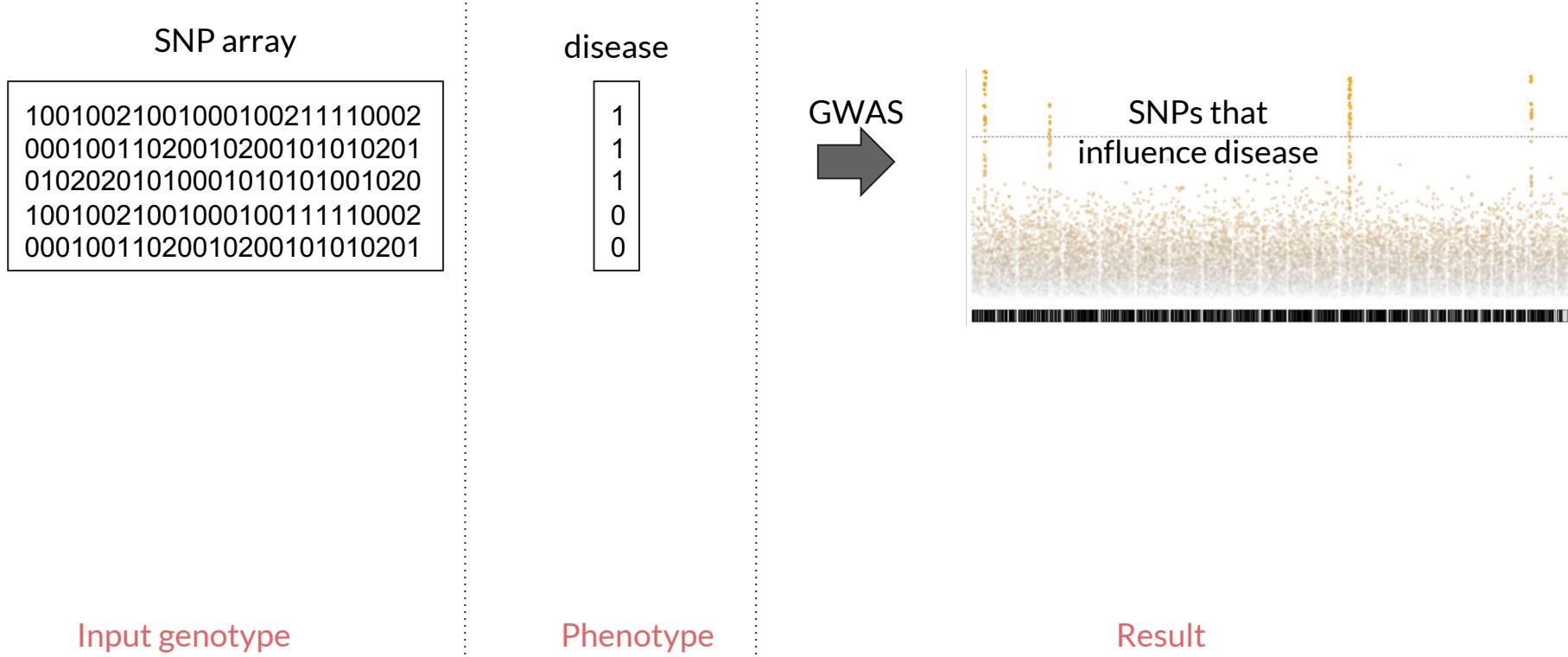
Course evaluation

5. Example: pQTL analysis

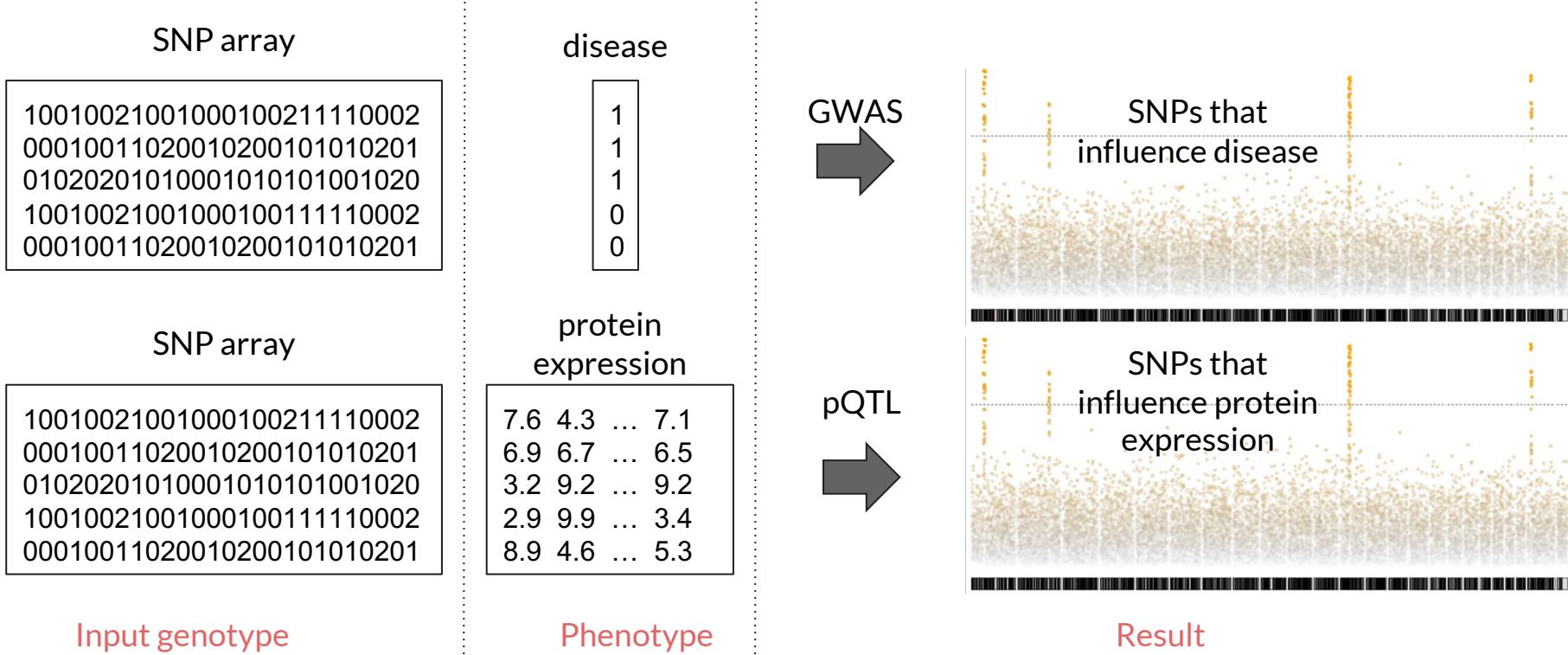
pQTL identify SNPs that influence changes in protein expression



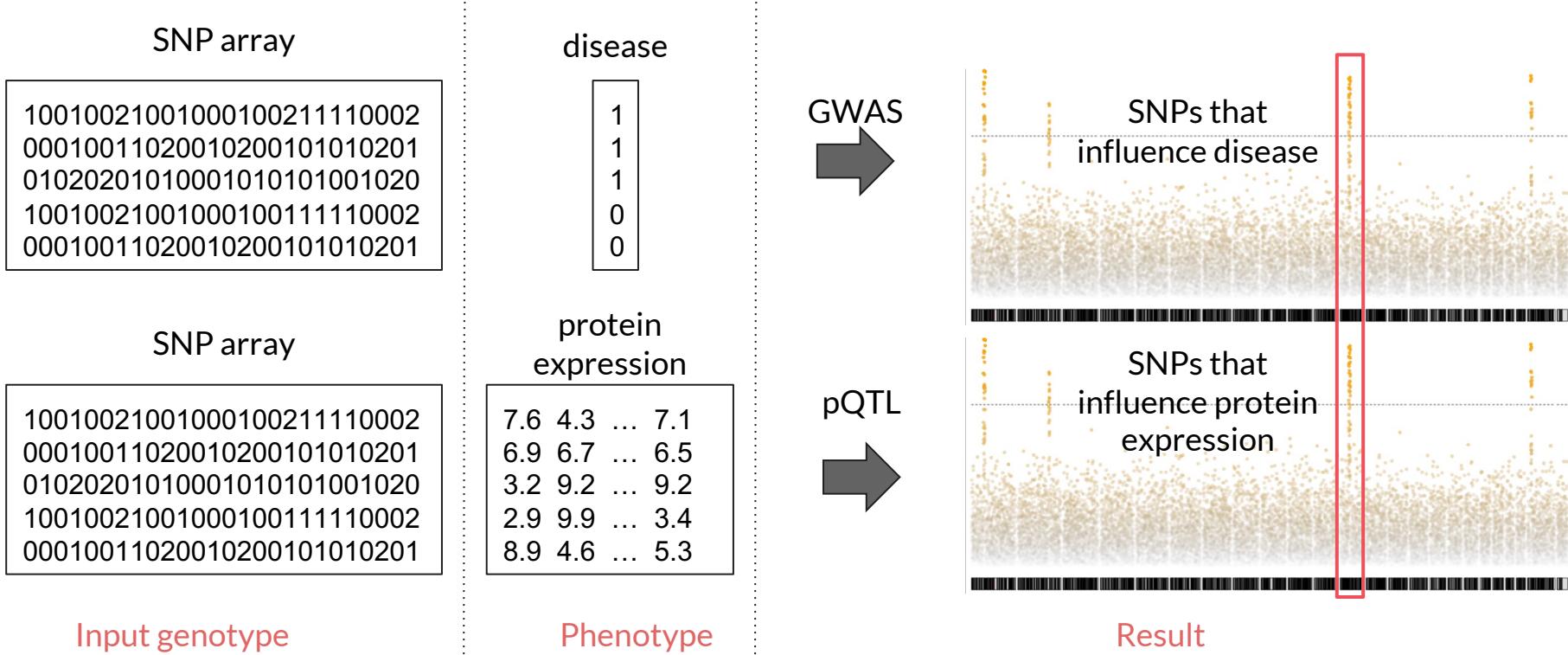
GWAS identify SNPs that influence trait



GWAS identify SNPs that influence trait



GWAS identify SNPs that influence trait



About RGENIE

First step - calculate Polygenic Risk Score (PRS)
for background association correction



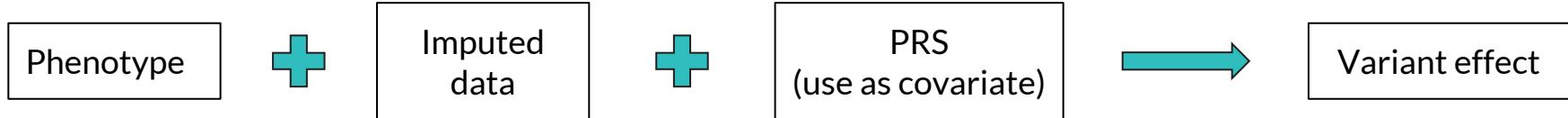
[REGENIE paper](#)
[End-to-end Target Discovery webinar](#)

About RGENIE

First step - calculate Polygenic Risk Score (PRS)
for background association correction

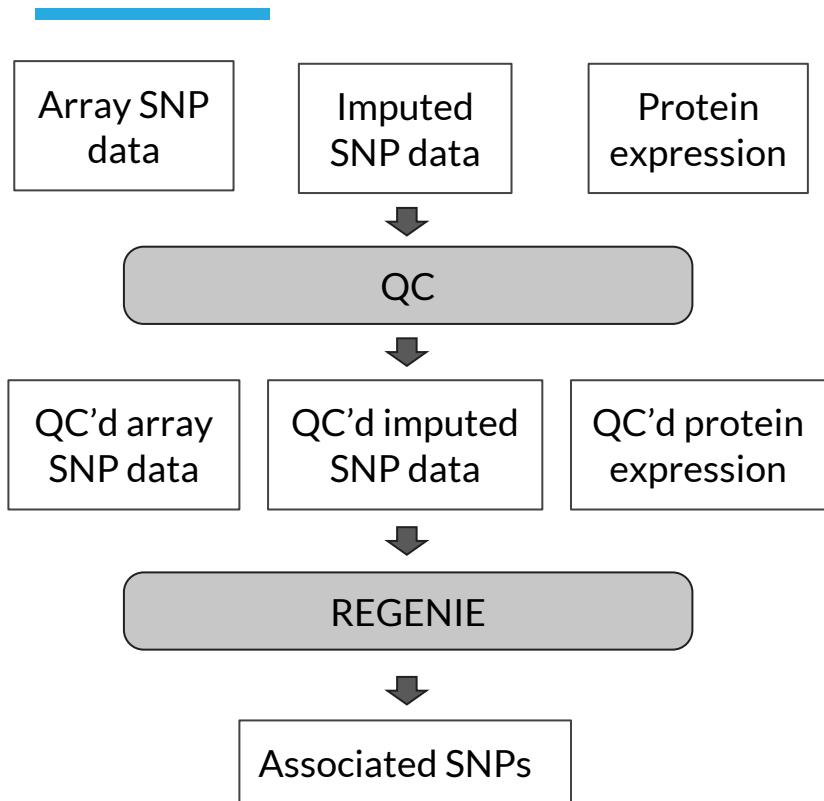


Second step - test variant-phenotype association

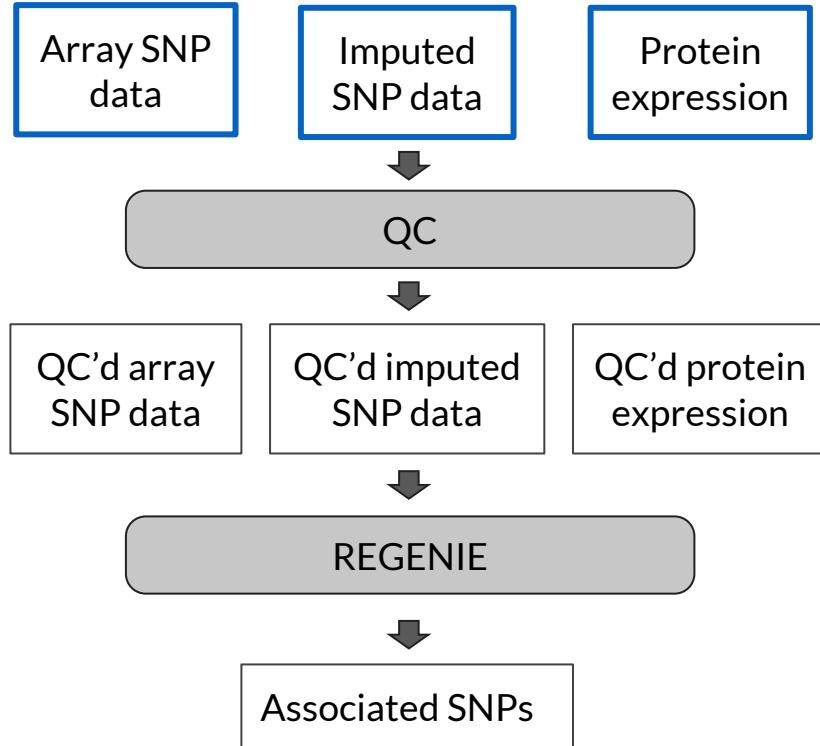


[REGENIE paper](#)
[End-to-end Target Discovery webinar](#)

Approach

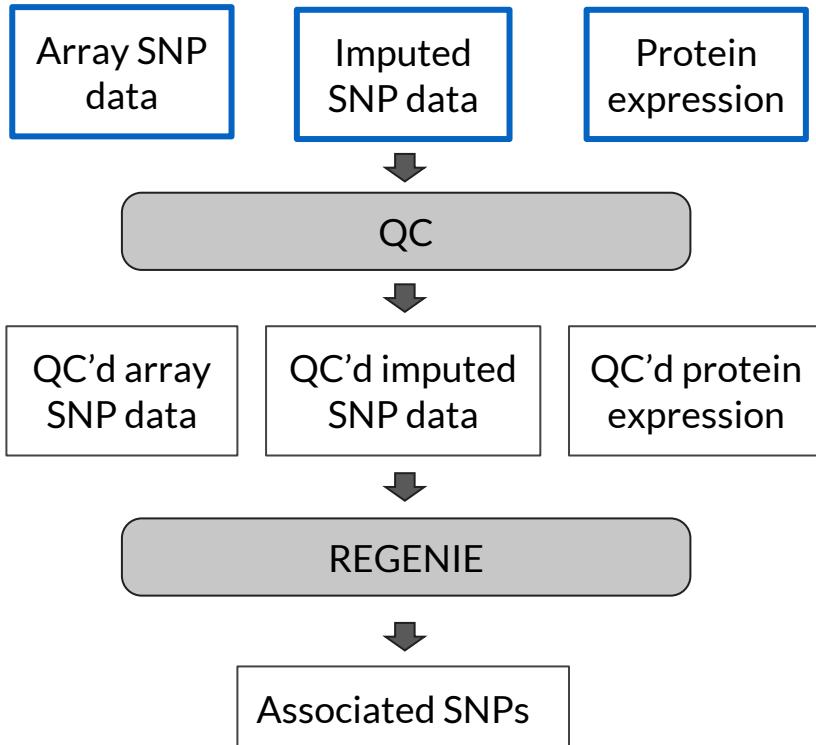


Simulate input data

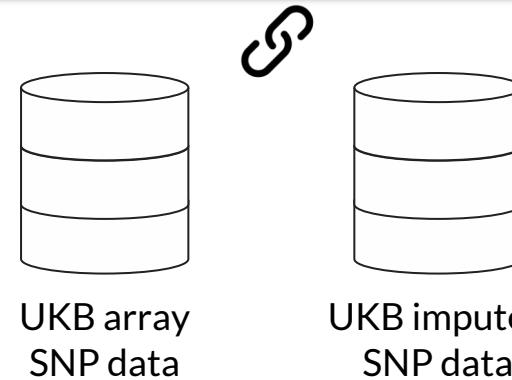


NOTE: We are using public proteomic data, not UKB data, for demonstration purposes

Matched genotype and protein expression data

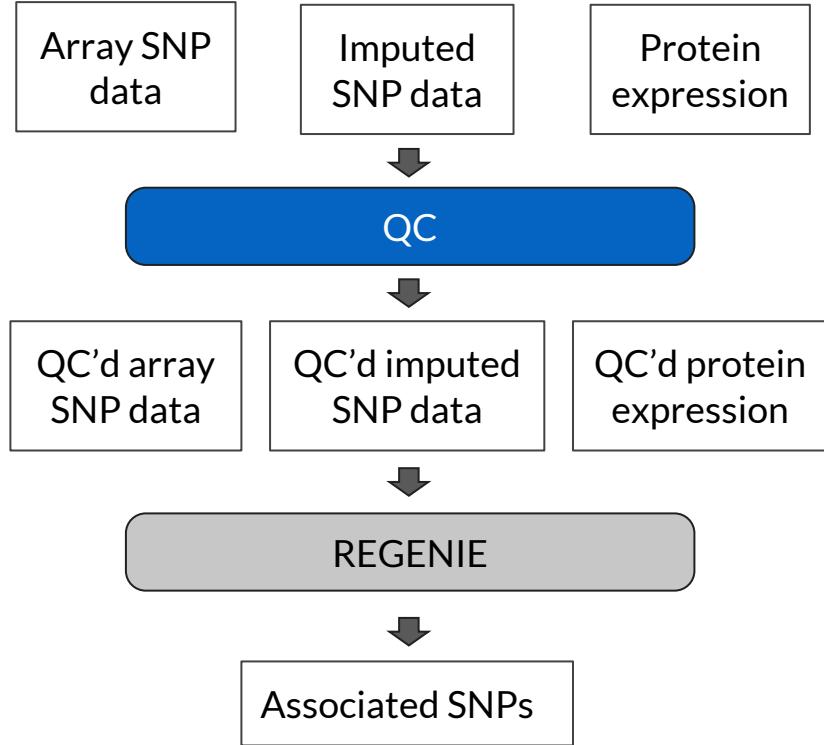


FID	IID	CA1	ICAM1	CHL1	TGFBI	ENG
2894753	2894753	7.62107	6.79971	4.73174	9.33471	3.12445
2352368	2352368	6.96085	6.98459	4.31338	9.06819	3.31576
1483346	1483346	7.16983	7.04907	4.72713	8.92804	3.16308
2352196	2352196	7.45724	6.89523	4.57029	9.27165	3.06199
4886500	4886500	7.81354	6.71708	4.93904	9.51350	3.66898



[Notebook to simulate data](#)

QC input data



FID	IID	CA1	ICAM1	CHL1	TGFBI	ENG
2894753	2894753	7.62107	6.79971	4.73174	9.33471	3.12445
2352368	2352368	6.96085	6.98459	4.31338	9.06819	3.31576
1483346	1483346	7.16983	7.04907	4.72713	8.92804	3.16308
2352196	2352196	7.45724	6.89523	4.57029	9.27165	3.06199
4886500	4886500	7.81354	6.71708	4.93904	9.51350	3.66898



- Filter samples to remove possible confounders
- Removing missing or low quality variants and proteins

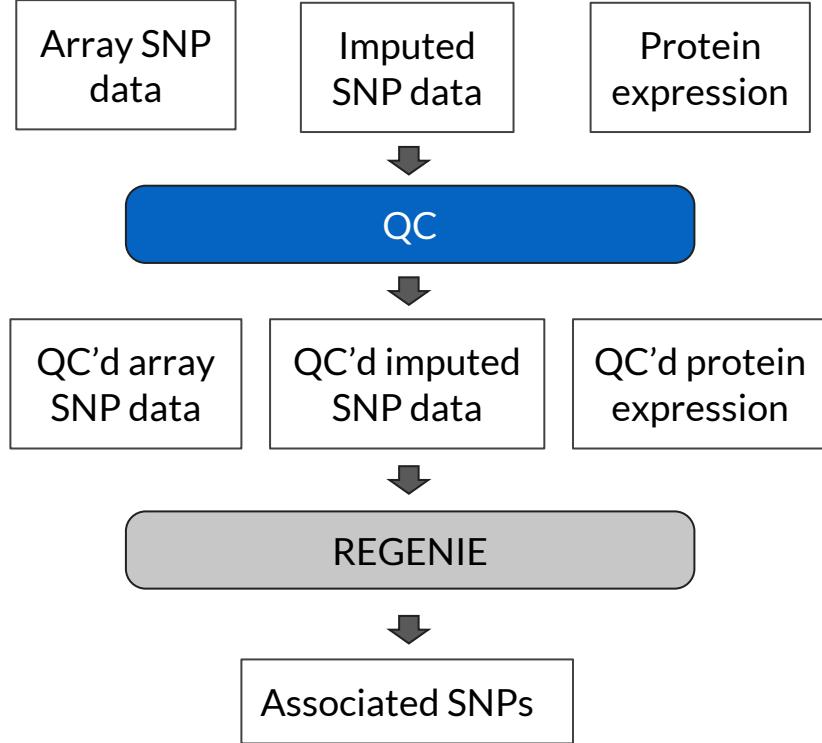


UKB array
SNP data



UKB imputed
SNP data

QC input data



	FID	IID	CA1	ICAM1	CHL1	TGFBI	ENG
2894753	2894753	7.62107	6.79971	4.73174	9.33471	3.12445	
2352368	2352368	6.96085	6.98459	4.31338	9.06819	3.31576	
1483346	1483346	7.16983	7.04907	4.72713	8.92804	3.16308	
2352196	2352196	7.45724	6.89523	4.57029	9.27165	3.06199	
4886500	4886500	7.81354	6.71208	4.93904	9.51350	3.66898	



- [Sample QC steps](#)
- [Array variant QC steps](#)
- [Imputed variant QC steps](#)

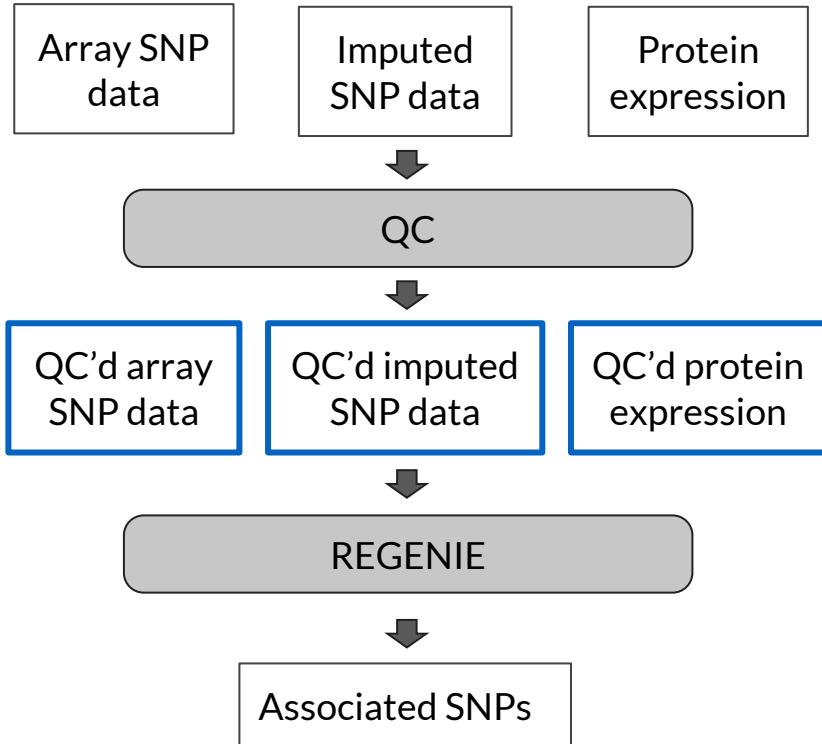


UKB array
SNP data

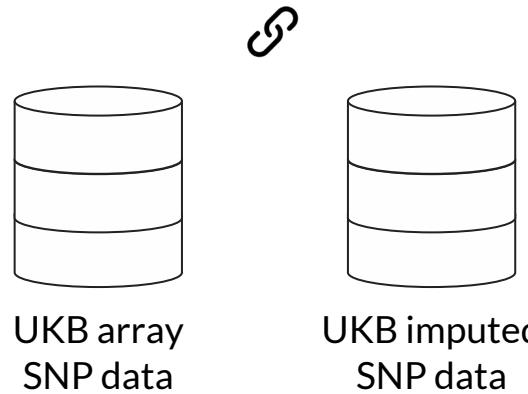


UKB imputed
SNP data

Get QC'd input data

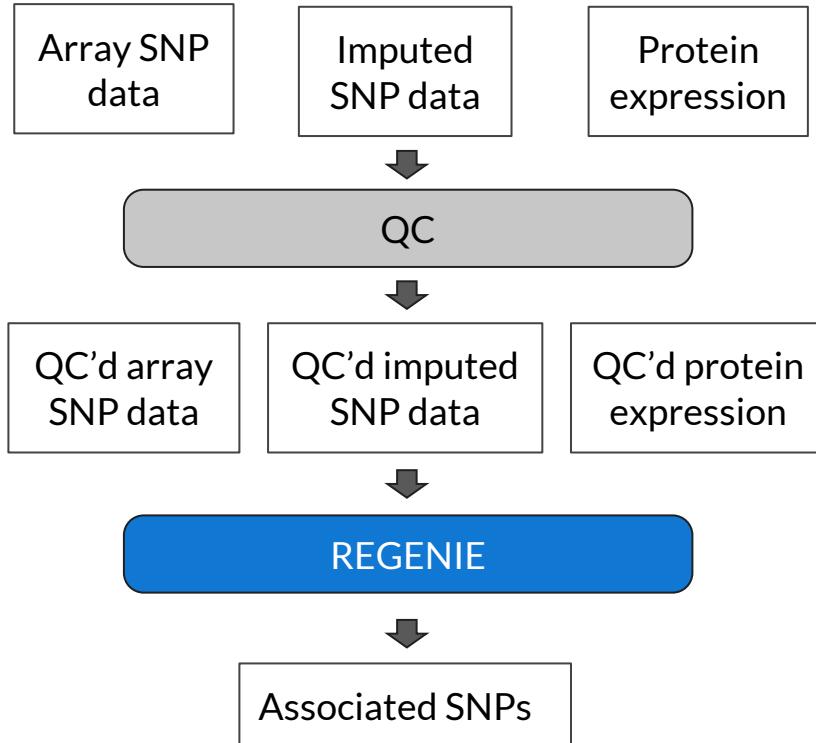


FID	IID	CA1	ICAM1	CHL1	TGFBI	ENG
2894753	2894753	7.62107	6.79971	4.73174	9.33471	3.12445
2352368	2352368	6.96085	6.98459	4.31338	9.06819	3.31576
1483346	1483346	7.16983	7.04907	4.72713	8.92804	3.16308
2352196	2352196	7.45724	6.89523	4.57029	9.27165	3.06199
4886500	4886500	7.81354	6.71708	4.93904	9.51350	3.66898



- 100,100 samples
- ~500,000 variants
- 200 proteins

Run GWAS



The screenshot shows the biobank platform's Tools Library interface. The REGENIE tool is selected. The tool's description is: "REGENIE: Genome-wide association analysis of large cohorts for quantitative and binary phenotypes using regenie". Below this, the "Run Analysis" interface for the REGENIE tool is shown. The "ANALYSIS SETTINGS" section is set to "REGENIE". The "ANALYSIS INPUTS" section lists several required files:

- Genotype BED for Step 1: Select File
- Genotype BIM for Step 1: Select File
- Genotype FAM for Step 1: Select File
- Genotype BGEN files for Step 2: Select File (Bgen)
- Genotype BGI index files for Step 2: Select File (Bped)
- Sample files for Step 2: Select File (Bped)
- Phenotype file: Select File
- Individuals to remove (Step 1): Select File (Bped)

Other sections include "APP SETTINGS" (Enable Batch: OFF), "Required fields not configured", and "Start Analysis".

REGENIE GWAS analysis settings

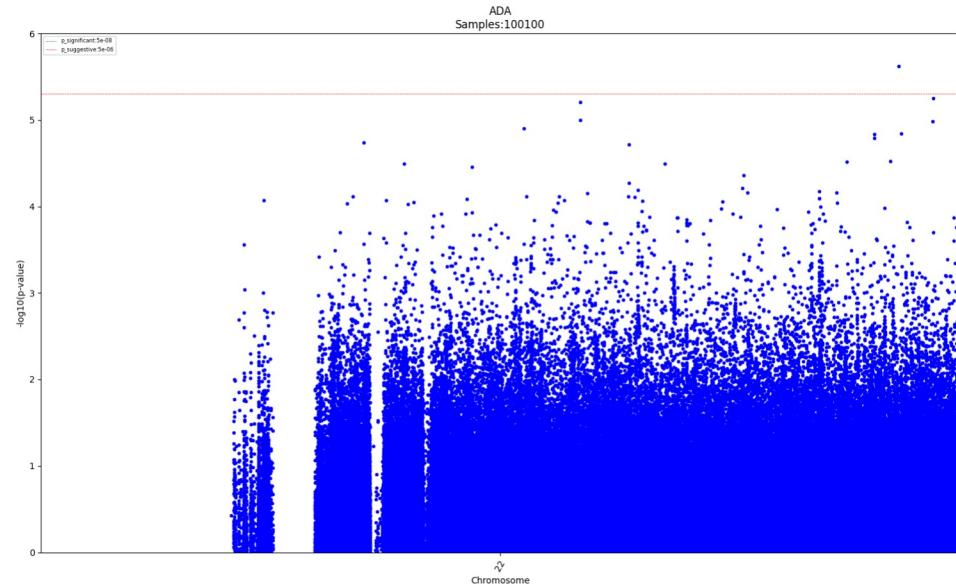
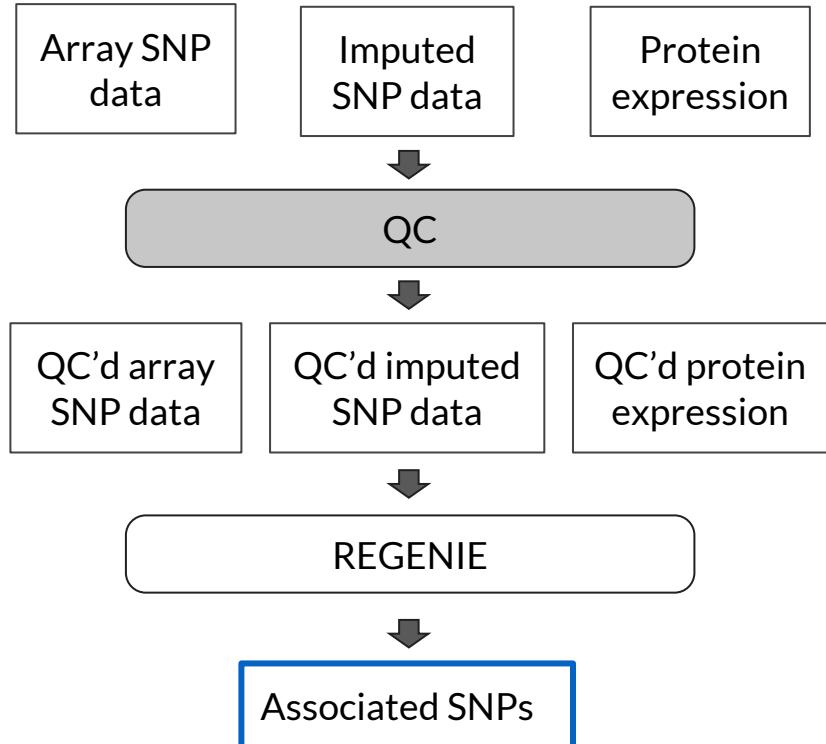
Step 1

- ▶ **Phenotype File:** <.phe file containing protein expression>
- ▶ **QC'D genotype Files (SNP array):** <. [bed, bim, fam] QC'd genotype files after liftover>
- ▶ **Variant IDs to extract:** <.snplist file containing list of QC'd array SNPs>

Step 2

- ▶ **Phenotype File:** <.phe file containing protein expression>
- ▶ **Sample ID File:** <.phe file containing protein expression>
- ▶ **BGEN, BGI, SAMPLE genotype Files (Imputed SNP):**
/Bulk/Imputation/ukb21008_c22_b0_v1.[bgen, bgi, sample]
- ▶ **Variant IDs to extract:** <.snplist file containing list of QC'd imputed SNPs>

Found significantly associated SNPs



Resources

	Link	Configuration	Runtime & cost
Code to create simulated protein expression data	https://github.com/dnanexus/UKB_RAP/blob/main/proteomics/protein_pQTL/1_simulate_input_data.ipynb	Kernel: PYTHON_R Priority: normal Recommended instance: mem1_ss1_v2_x2	Runtime: ~ 1min Cost: ~£ 0.0069
QC steps from end-to-end webinar	https://github.com/dnanexus/UKB_RAP/tree/main/end_to_end_gwas_phewas		
Steps to run REGENIE	https://github.com/dnanexus/UKB_RAP/blob/main/proteomics/protein_pQTL/REA_DME.md		Runtime: ~ 19 hours Cost: ~£1.04
REGENIE publication	https://www.nature.com/articles/s41588-021-00870-7		

Conclusion

- ▶ Researchers can use [UKB-RAP](#) to analyze proteomic data
- ▶ Proteomic data can be extracted via the [Cohort Browser](#)
- ▶ Differential expression analysis can be done using custom code in [JupyterLab](#)
- ▶ pQTL analysis can be done using [REGENIE](#) app following [end-to-end tutorial steps](#)

Helpful resources

- ▶ [Integrative analysis of UKB proteomics data - webinar](#)
- ▶ [UKB Research analysis platform overview - webinar](#)
- ▶ [Introduction to JupyterLab notebooks on RAP - webinar](#)
- ▶ [End to end target discovery with GWAS and PheWAS on the UKB research analysis platform - webinar](#)

Upcoming events

- ▶ **Webinar: Dementia and Multimorbidity in Late-Life disease: Longitudinal and Multimodal Data Science Approaches**
 - ▶ When: Late June TBA
 - ▶ Registration TBA
- ▶ [Subscribe](#)
- ▶ [All webinar recordings](#)

Join the conversation to:



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

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

Click Here to Join 

OR



Acknowledgements



Ondrej Klempir, PhD
Sr. Community Engagement
Scientist



Arkarachai Fungtammasan, PhD
Scientific Community Manager/
Principal Scientist



Anastazie Sedlakova, PhD
Community Engagement/
Principal Scientist

UKB-RAP team
Ben Busby, PhD
Ted Laderas, PhD
SciProd team

Thank you! Questions?