



biobank^{uk}

Research Analysis
Platform

Enabled by **DNA**nexus[®]

Analyzing UK Biobank proteomics data on the UKB-RAP

JUNE 2023

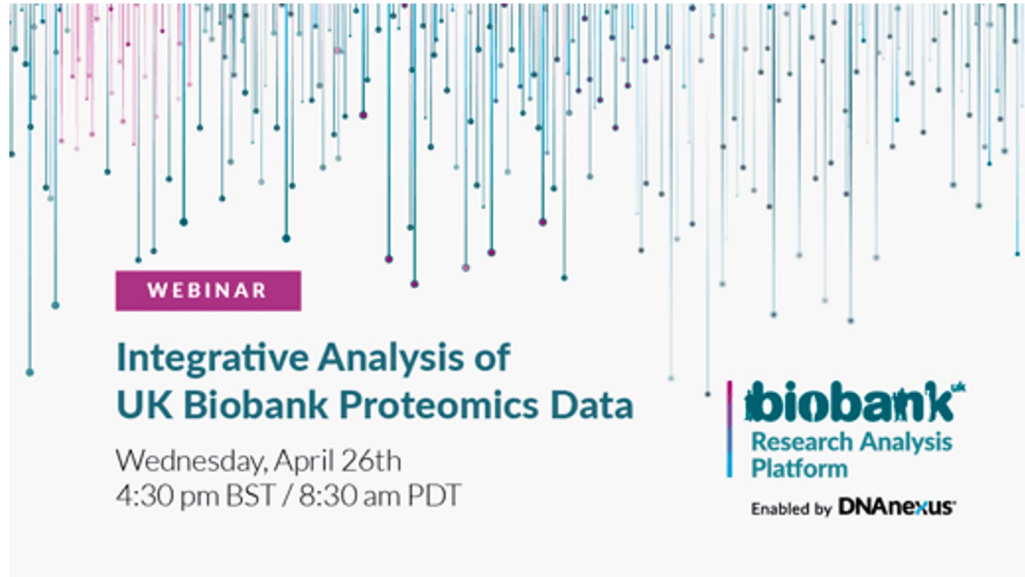
Speakers & Agenda



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



WEBINAR

Integrative Analysis of UK Biobank Proteomics Data

Wednesday, April 26th
4:30 pm BST / 8:30 am PDT

biobank^{uk}
Research Analysis
Platform

Enabled by **DNAxus**

<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](#)

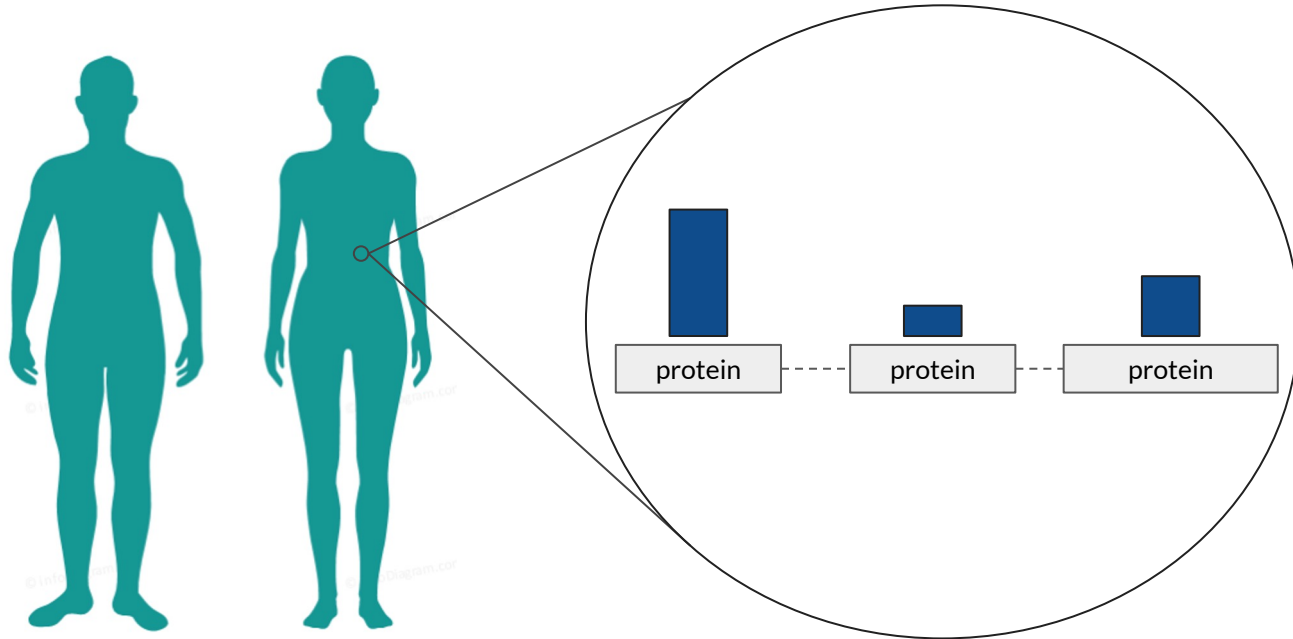
Learning Objectives

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
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

+ Add Filter to Cohort

PHENO GENO

ever smo

- Assessment Centre
 - Touchscreen
 - Lifestyle and environment
 - Smoking
 - abc Ever smoked | Instance 0
 - abc Ever smoked | Instance 1
 - abc Ever smoked | Instance 2
 - abc Ever smoked | Instance 3



EDIT FILTER | Untitled Cohort

Ever smoked | Instance 0

IS ANY OF ▾

Yes X

Cancel Apply Filter



biobank PROJECTS TOOLS ORG ADMIN HELP

app77202_20220611175502.dataset 502,411 Participants

ever_smoked_cases 298,714 Participants

Dashboard Actions

ever_smoked_cases 298,714 of 502,411 Participants

+ Add Filter Clear All Filters

Select PARTICIPANT Ever smoked | Instance 0 is Yes

OVERVIEW DATA PREVIEW GENOMICS

This cohort has unsaved changes



cohort_dataset = "project-XXX:record-XXX"

AD phenotype example
Ischemic disease example

Get protein expression data from Cohort Browser

Table Exporter App

The screenshot shows the 'Table exporter' app interface. On the left, a workflow diagram shows 'dataset_or_cohort_or_dashboard' and 'field_names_file_txt' as inputs to the 'Table exporter' process, which outputs a 'CSV' file. The main interface is divided into three sections: 'ANALYSIS SETTINGS', 'ANALYSIS INPUTS 1', and 'APP SETTINGS'. Under 'ANALYSIS INPUTS 1', there are two input fields: 'Dataset or Cohort or Dashboard' and 'File containing Field Names', both with 'Select Record' and 'Select File' buttons respectively. The 'APP SETTINGS' section includes an 'Enable Batch' toggle set to 'OFF' and an 'About this app' link. Below these are 'OPTIONS' for 'Output Prefix' (set to 'data'), 'Output File Format' (set to 'CSV'), 'Coding Option' (set to 'REPLACE'), and 'Header Style' (set to 'FIELD-NAME').

dx extract_dataset

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

Get protein expression data using Table Exporter app

The screenshot displays the 'Table Exporter' application interface. At the top, there is a header with 'TABLE-EXPORTER' and a toggle for 'Enable Batch' set to 'OFF'. Below the header, a dark bar contains 'Table exporter' and an 'About this app' link. The main configuration area is divided into several sections:

- Dataset or Cohort or Dashboard:** A dropdown menu showing 'ischaemic_cases'.
- File containing Field Names:** A dropdown menu showing 'field_names.txt'.
- OPTIONS:**
 - Output Prefix:** An empty text input field.
 - Output File Format:** A dropdown menu set to 'CSV'.
 - Coding Option:** A dropdown menu set to 'REPLACE'.
 - Header Style:** A dropdown menu set to 'FIELD-NAME'.
- ADVANCED OPTIONS:**
 - Entity:** A text input field containing 'olink_instance_0'.
 - Field Names:** An empty text input field.
 - Field Titles:** An empty text input field.
- COHORT/DASHBOARD OPTIONS:**
 - Cohort Table Entity Names:** An empty text input field.
 - Cohort Table Entity Titles:** An empty text input field.

Get protein expression data using Table Exporter app

The screenshot displays the 'Table Exporter' application interface. At the top, there is a header with 'TABLE-EXPORTER' and a toggle for 'Enable Batch' set to 'OFF'. Below the header, the main configuration area is titled 'Table exporter'. It features two input fields: 'Dataset or Cohort or Dashboard' with the value 'ischaemic_cases' and 'File containing Field Names' with the value 'field_names.txt'. A callout box points to the 'ischaemic_cases' value, containing the text 'Phenotype dataset e.g. *.dataset'. The interface is organized into sections: 'OPTIONS' and 'ADVANCED OPTIONS'. Under 'OPTIONS', there are four settings: 'Output Prefix' (empty text field), 'Output File Format' (dropdown menu set to 'CSV'), 'Coding Option' (dropdown menu set to 'REPLACE'), and 'Header Style' (dropdown menu set to 'FIELD-NAME'). Under 'ADVANCED OPTIONS', there are three settings: 'Entity' (text field with 'olink_instance_0'), 'Field Names' (empty text field), and 'Field Titles' (empty text field). At the bottom, there is a section for 'COHORT/DASHBOARD OPTIONS' with two settings: 'Cohort Table Entity Names' and 'Cohort Table Entity Titles', both with empty text fields.

Get protein expression data using Table Exporter app

The screenshot shows the 'Table Exporter' app interface. At the top, there is a header with 'TABLE-EXPORTER' and an 'Enable Batch' toggle set to 'OFF'. Below the header, the app title 'Table exporter' is displayed with an 'About this app' link. The main configuration area includes:

- Dataset or Cohort or Dashboard:** A dropdown menu showing 'ischaemic_cases'.
- File containing Field Names:** A dropdown menu showing 'field_names.txt'.
- OPTIONS:**
 - Output Prefix:** An empty text input field.
 - Output File Format:** A dropdown menu set to 'CSV'.
 - Coding Option:** A dropdown menu set to 'REPLACE'.
 - Header Style:** A dropdown menu set to 'FIELD-NAME'.
- ADVANCED OPTIONS:**
 - Entity:** A text input field containing 'olink_instance_0'.
 - Field Names:** An empty text input field.
 - Field Titles:** An empty text input field.
- COHORT/DASHBOARD OPTIONS:**
 - Cohort Table Entity Names:** An empty text input field.
 - Cohort Table Entity Titles:** An empty text input field.

List of field names
e.g. eid, aarsd1,...

List of field names

Get protein expression data using Table Exporter app

The screenshot displays the 'Table Exporter' application interface. At the top, there is a header with 'TABLE-EXPORTER' and an 'Enable Batch' toggle set to 'OFF'. Below the header, a dark bar contains 'Table exporter' and an 'About this app' link. The main configuration area is divided into several sections:

- Dataset or Cohort or Dashboard:** A dropdown menu showing 'ischaemic_cases'.
- File containing Field Names:** A dropdown menu showing 'field_names.txt'.
- OPTIONS:**
 - Output Prefix:** An empty text input field.
 - Output File Format:** A dropdown menu set to 'CSV'.
 - Coding Option:** A dropdown menu set to 'REPLACE'.
 - Header Style:** A dropdown menu set to 'FIELD-NAME'.
- ADVANCED OPTIONS:**
 - Entity:** A text input field containing 'olink_instance_0'. A callout box points to this field with the text 'entity table' and 'e.g. olink_instance_#'.
 - Field Names:** An empty text input field.
 - Field Titles:** An empty text input field.
- COHORT/DASHBOARD OPTIONS:**
 - Cohort Table Entity Names:** An empty text input field.
 - Cohort Table Entity Titles:** An empty text input field.

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	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	aca1	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

~1,500 proteins

~53,000
samples

	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

SETTINGS MANAGE MONITOR VISUALIZE Back to

All Projects > [redacted] > Bulk > Protein biomarkers > Olink > helper_files

Current Folder Only Any Name Any ID Any Type Any Class

<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 management interface with a sidebar on the left and a main content area. The sidebar contains a tree view with folders: .Notebook_archive, .table-exporter, alee_example, Bulk (expanded), QC_output, and Showcase metadata. The main content area has a breadcrumb path: All Projects > [redacted] > Bulk > Protein biomarkers > Olink > helper_files. Below the breadcrumb are filter buttons: Current Folder Only, Any Name, Any ID, Any Type, and Any Class. A table lists the files in the folder:

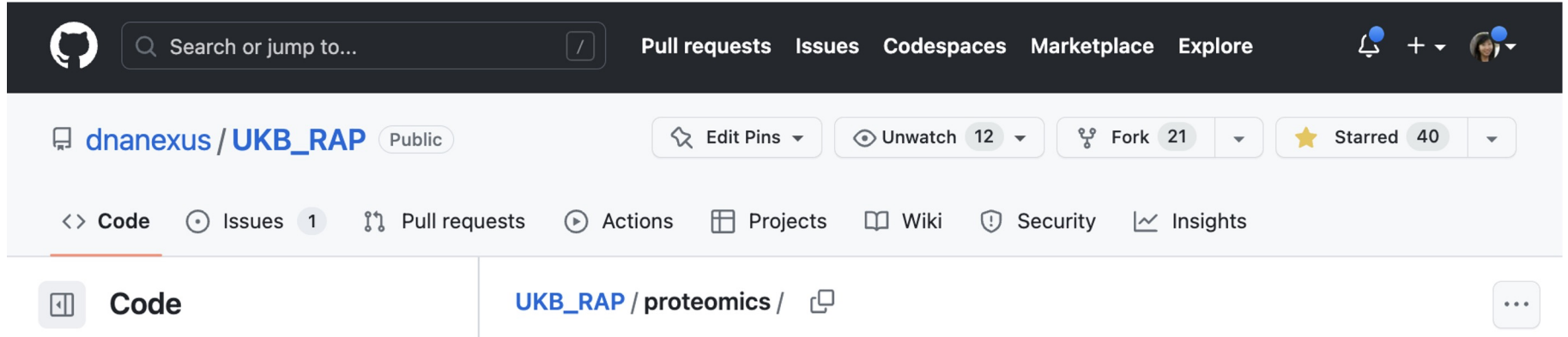
<input type="checkbox"/>	Name ↕	Type / Class	Created
<input type="checkbox"/>	olink_assay_version.dat	File	Mar 14 2023, 11:24
<input type="checkbox"/>	olink_assay_warning.dat	File	Mar 14 2023, 11:24
<input type="checkbox"/>	olink_batch_number.dat		Mar 14 2023, 11:24
<input type="checkbox"/>	olink_limit_of_detection.dat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

Callouts in the image:

- A callout box points to the file `olink_limit_of_detection.dat` with the text "Limit of detection".
- A callout box points to the file `Olink_proteomics_data.pdf` with the text "PDF of QC steps performed".

3. How to access analysis code on UKB-RAP

Code is available on github!



GitHub repository page for **dnanexus / UKB_RAP** (Public). The page shows navigation options: Code, Issues (1), Pull requests, Actions, Projects, Wiki, Security, and Insights. The current view is the Code tab, showing the path **UKB_RAP / proteomics /**.

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

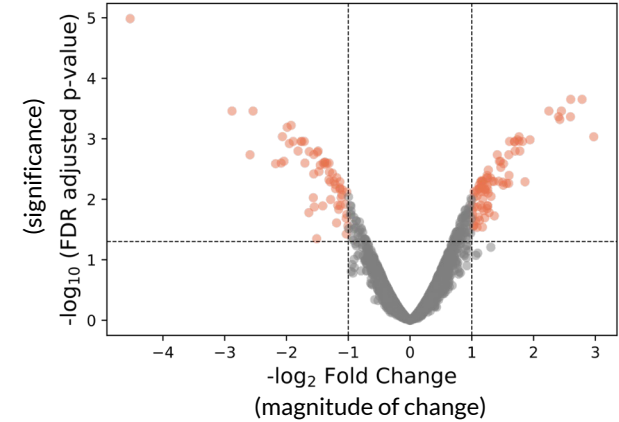
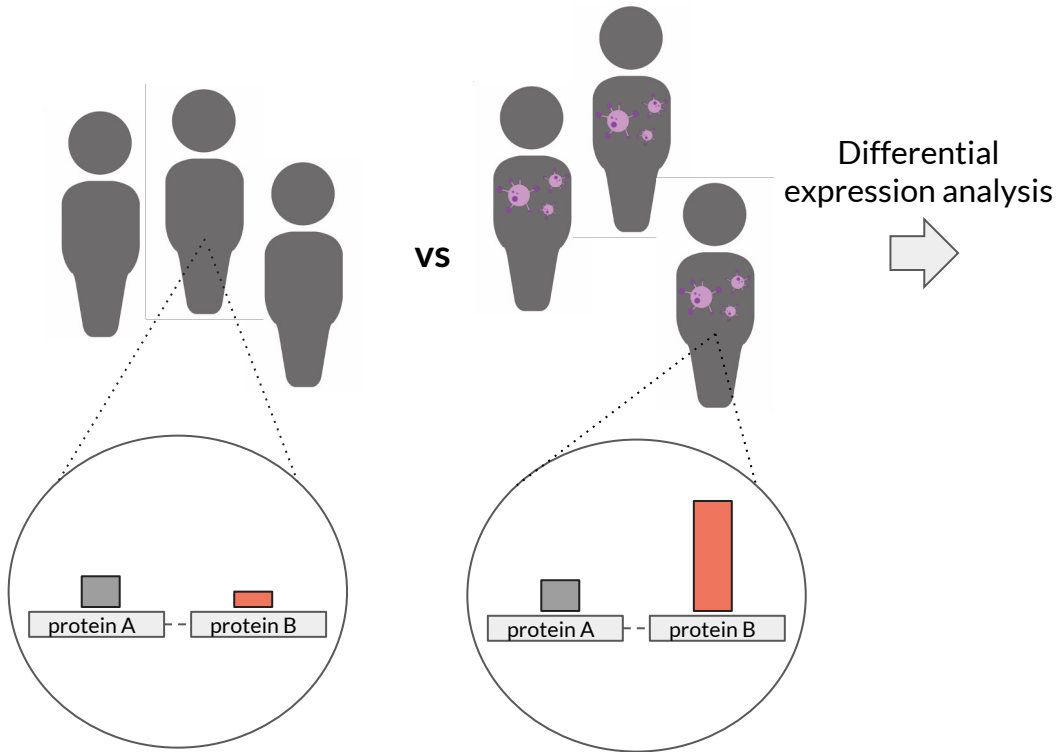
Steps to perform on your local machine:

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

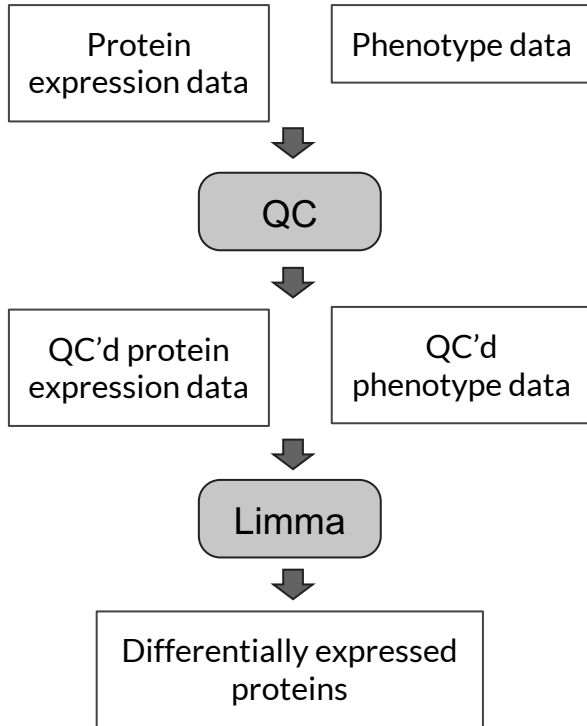
```
$git clone https://github.com/dnanexus/UKB\_RAP.git
$cd UKB_RAP
$dx login
$dx upload -r <proteomics> --destination <path on the UKB-RAP>
```


4. Example: Differential expression analysis

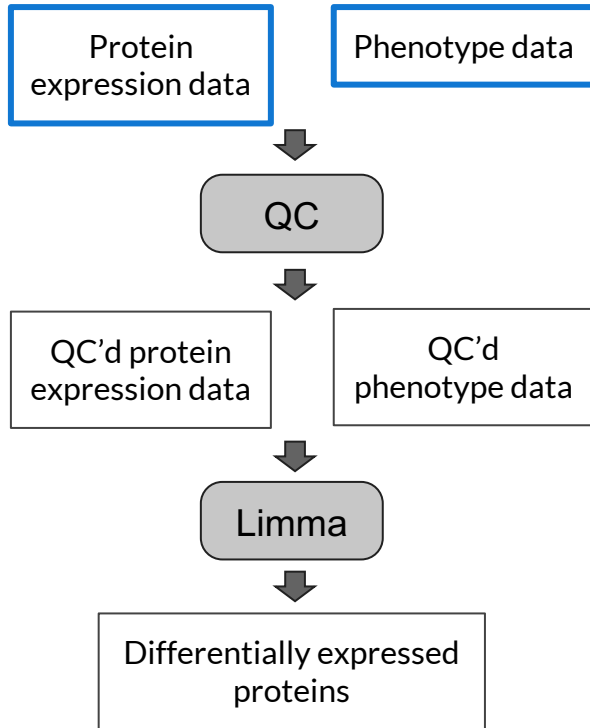
Differential expression analysis used to study mechanisms of disease



Approach

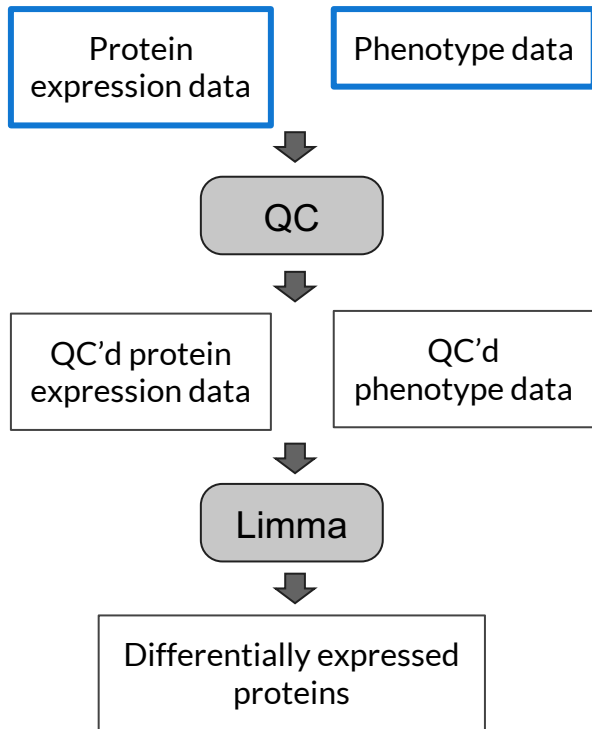


Approach



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

Collect input data

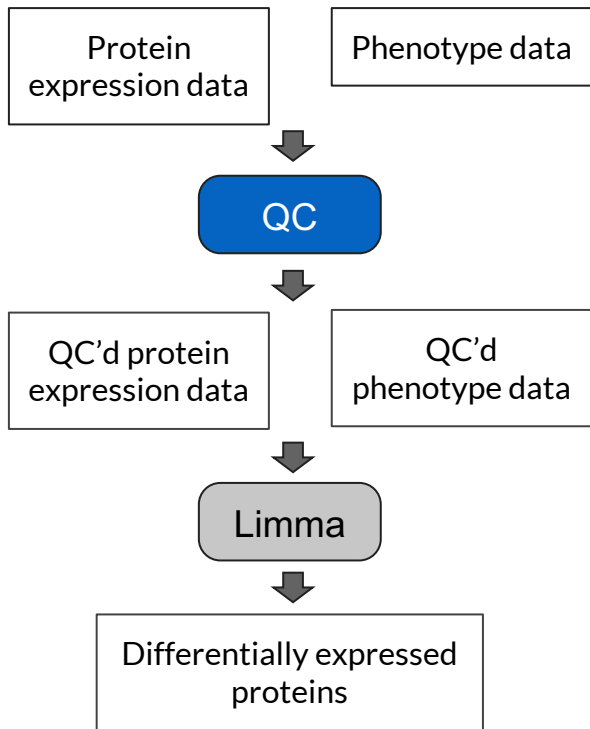


	CA1	ICAM1	CHL1	TGFBI	ENG
Plasma_Sample					
H0529.3	7.62107	6.79971	4.73174	9.33471	3.12445
H0441.1	6.96085	6.98459	4.31338	9.06819	3.31576
H0558.3	7.16983	7.04907	4.72713	8.92804	3.16308
H0499.2	7.59577	6.80282	4.51559	9.17979	3.19292
H0468.3	7.25945	6.91728	4.84307	9.91809	3.47692

	PIDN	Age_at_Baseline	Sex	Outcome
Plasma_Sample				
H0529.3	9677	90+	Male	MCI_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

Input data from [Kivisakk et al](#)

QC input data

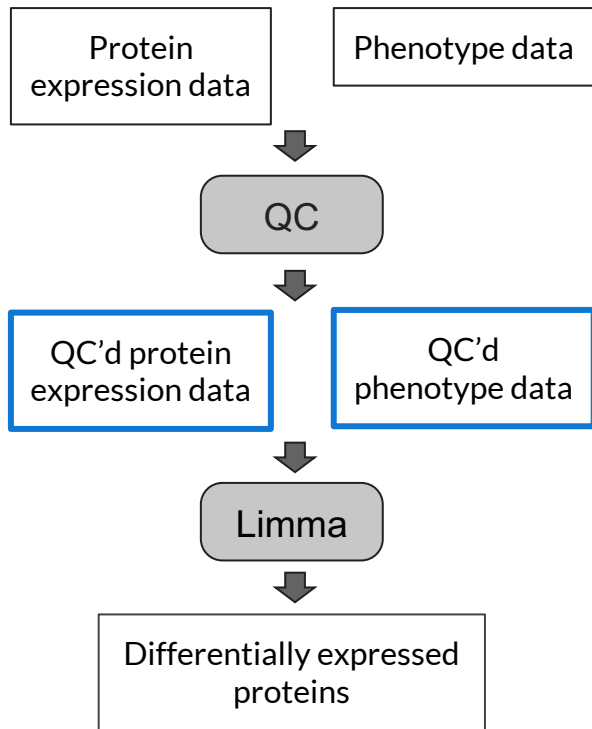


	CA1	ICAM1	CHL1	TGFBI	ENG
Plasma_Sample					
H0529.3	7.62107	6.79971	4.73174	9.33471	3.12445
H0441.1	6.96085	6.98459	4.31338	9.06819	3.31576
H0558.3	7.16983	7.04907	4.72713	8.92804	3.16308
H0499.2	7.59577	6.80282	4.51559	9.17979	3.19292
H0468.3	7.59577	6.80282	4.51559	9.17979	3.19292

- Removing missing, outlier data
- Normalize/scale data

	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

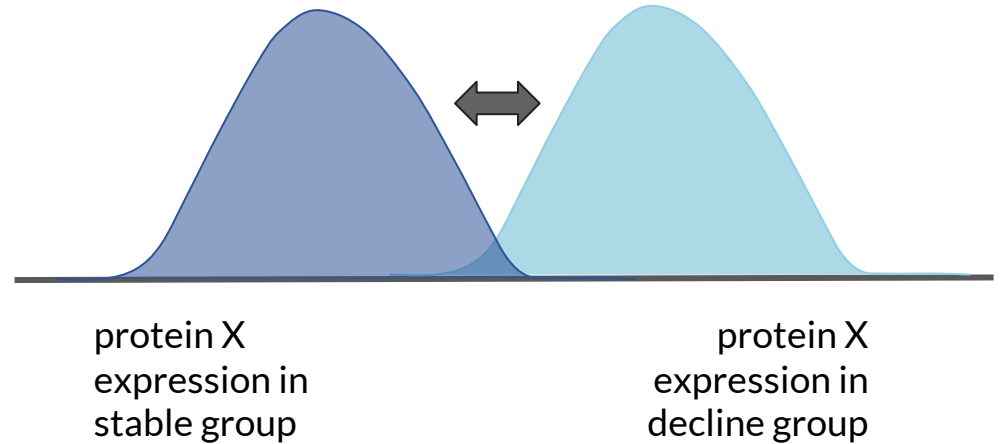
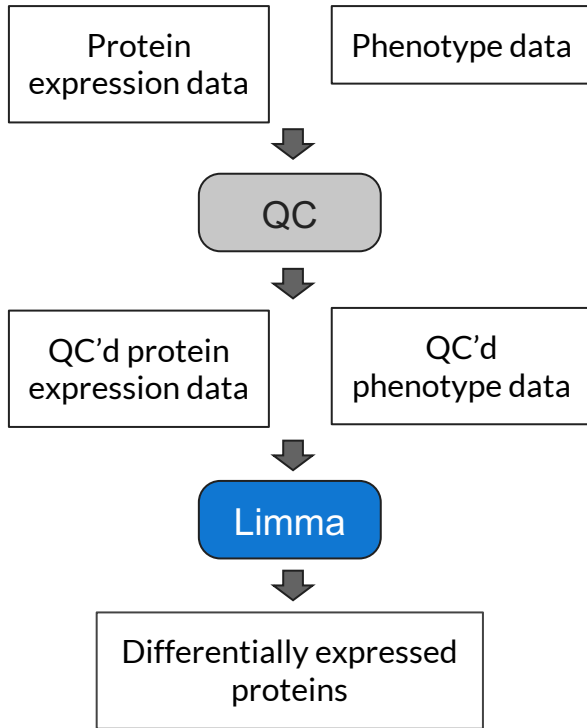
Get QC'd input data



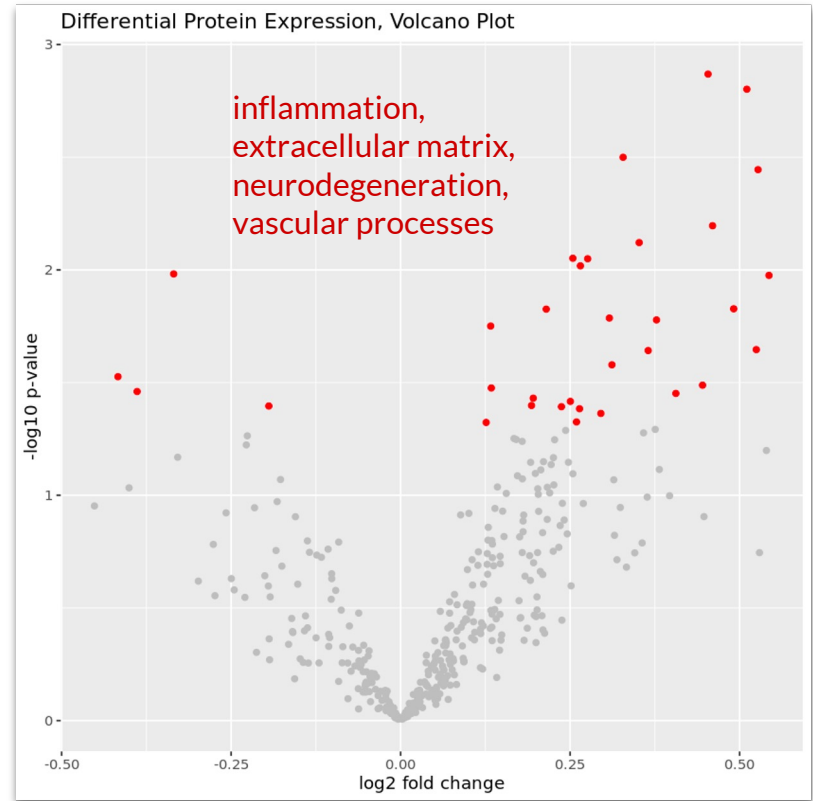
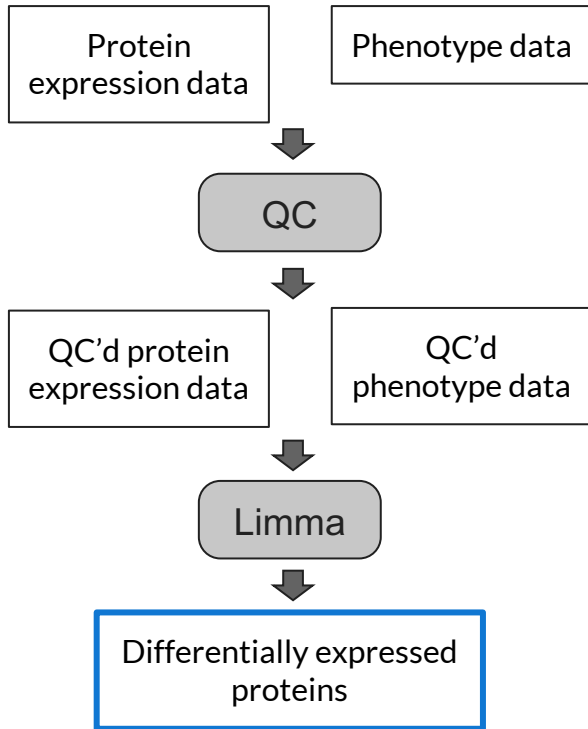
	CA1	ICAM1	CHL1	TGFBI	ENG
Plasma_Sample					
H0529.3	7.62107	6.79971	4.73174	9.33471	3.12445
H0441.1	6.96085	6.98459	4.31338	9.06819	3.31576
H0558.3	7.16983	7.04907	4.72713	8.92804	3.16308
H0499.2	7.59577	6.80282	4.51559	9.17979	3.19292
H0468.3	7.25945	6.91728	4.84307	9.91809	3.47692

	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

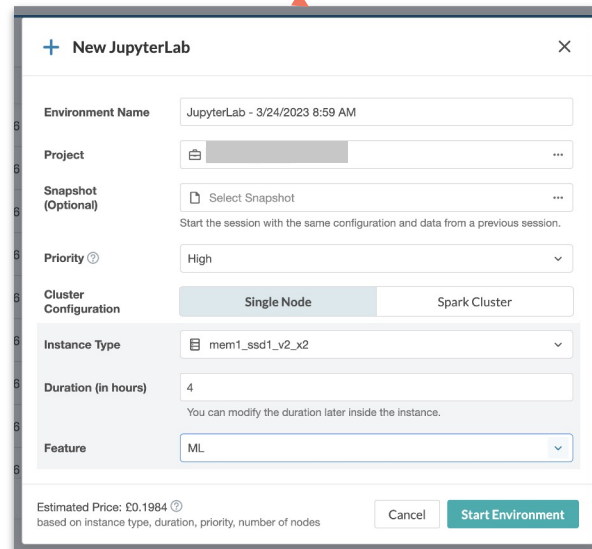
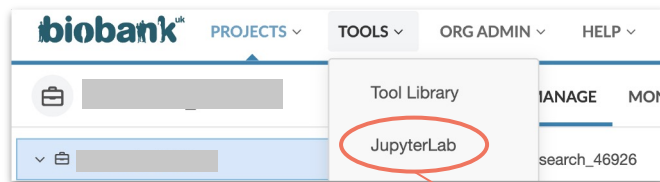
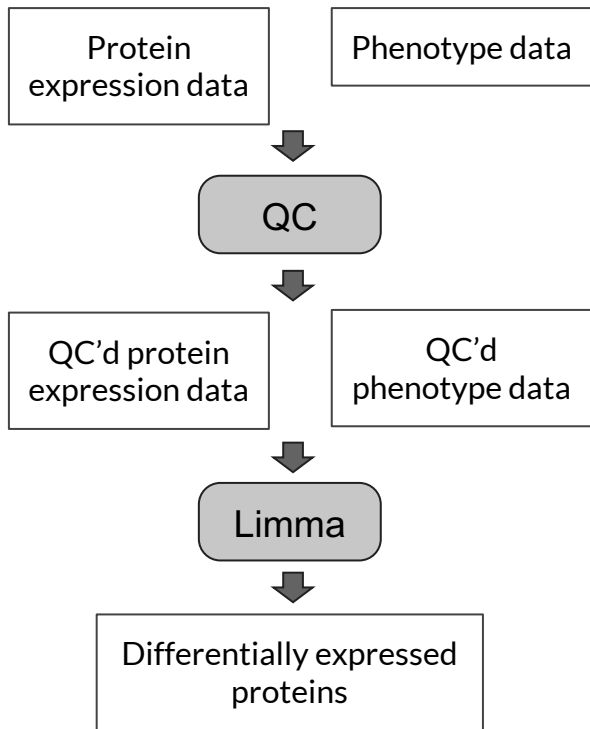
Perform differential expression analysis using Limma



Found differentially expressed proteins

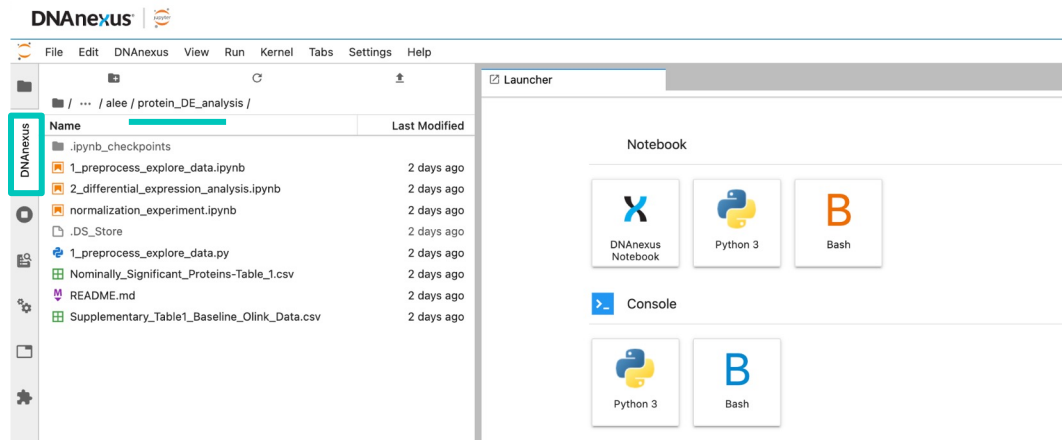
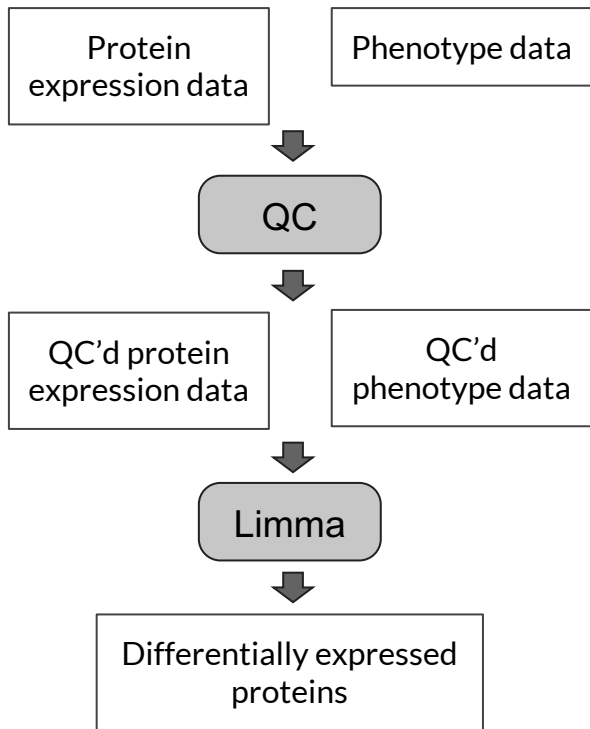


Run analysis using JupyterLab



JupyterLab webinar

Run analysis using JupyterLab



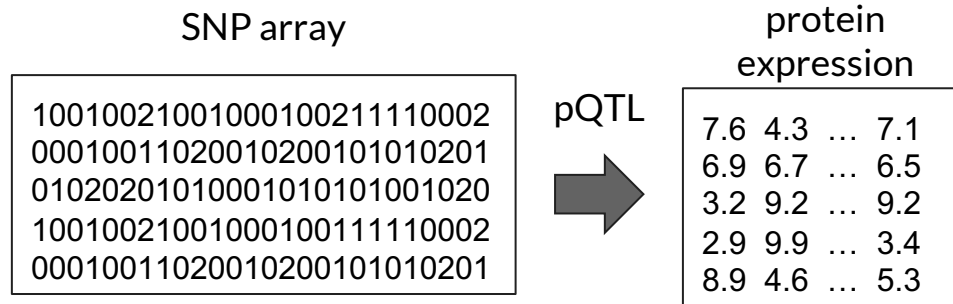
Resources

	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#366642284		

5 minute break Course evaluation

5. Example: pQTL analysis

pQTL identify SNPs that influence changes in protein expression



GWAS identify SNPs that influence trait

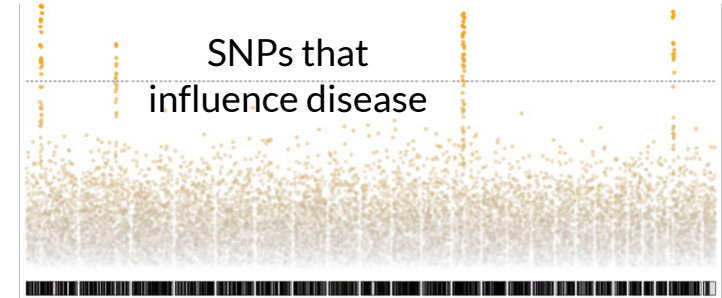
SNP array

```
10010021001000100211110002
00010011020010200101010201
01020201010001010101001020
10010021001000100111110002
00010011020010200101010201
```

disease

```
1
1
1
0
0
```

GWAS
➔



Input genotype

Phenotype

Result

GWAS identify SNPs that influence trait

SNP array

disease

```
10010021001000100211110002
00010011020010200101010201
01020201010001010101001020
10010021001000100111110002
00010011020010200101010201
```

```
1
1
1
0
0
```

GWAS



SNP array

protein
expression

```
10010021001000100211110002
00010011020010200101010201
01020201010001010101001020
10010021001000100111110002
00010011020010200101010201
```

```
7.6 4.3 ... 7.1
6.9 6.7 ... 6.5
3.2 9.2 ... 9.2
2.9 9.9 ... 3.4
8.9 4.6 ... 5.3
```

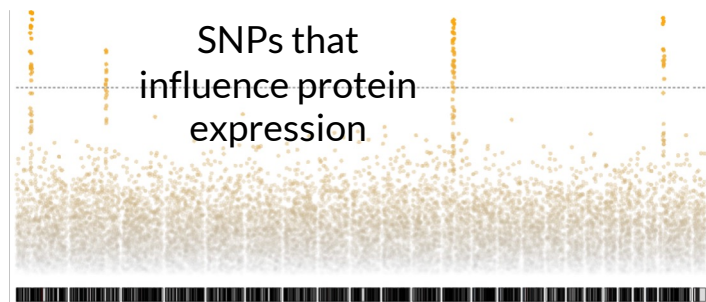
pQTL



SNPs that
influence disease



SNPs that
influence protein
expression



Input genotype

Phenotype

Result

GWAS identify SNPs that influence trait

SNP array

```
10010021001000100211110002
00010011020010200101010201
01020201010001010101001020
10010021001000100111110002
00010011020010200101010201
```

SNP array

```
10010021001000100211110002
00010011020010200101010201
01020201010001010101001020
10010021001000100111110002
00010011020010200101010201
```

disease

```
1
1
1
0
0
```

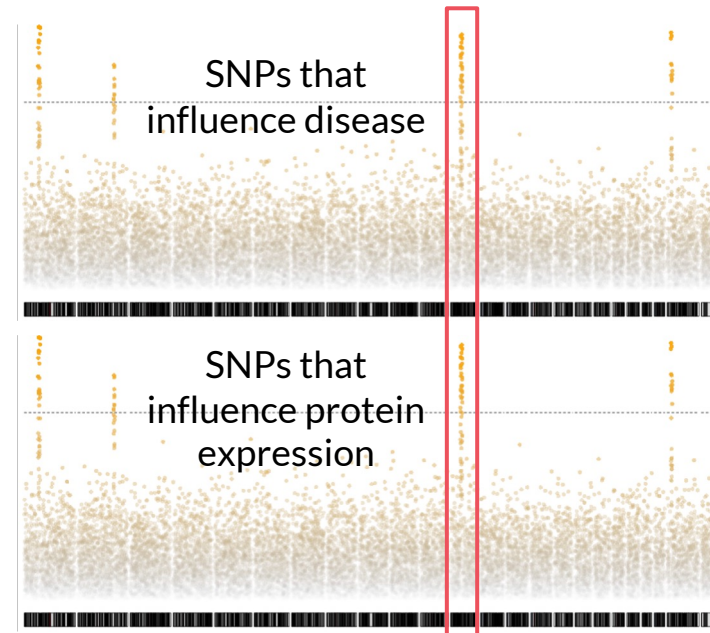
protein
expression

```
7.6 4.3 ... 7.1
6.9 6.7 ... 6.5
3.2 9.2 ... 9.2
2.9 9.9 ... 3.4
8.9 4.6 ... 5.3
```

GWAS



pQTL



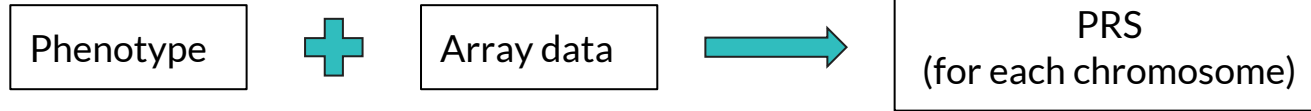
Input genotype

Phenotype

Result

About REGENIE

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



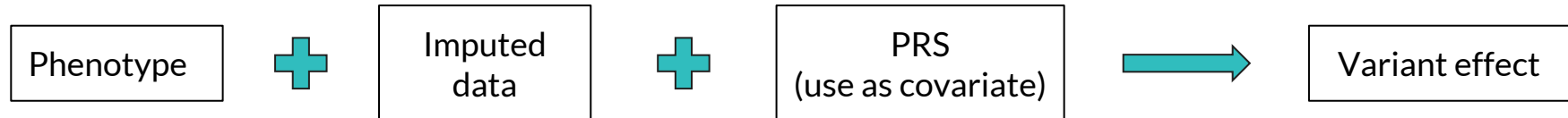
REGENIE paper
End-to-end Target Discovery webinar

About REGENIE

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

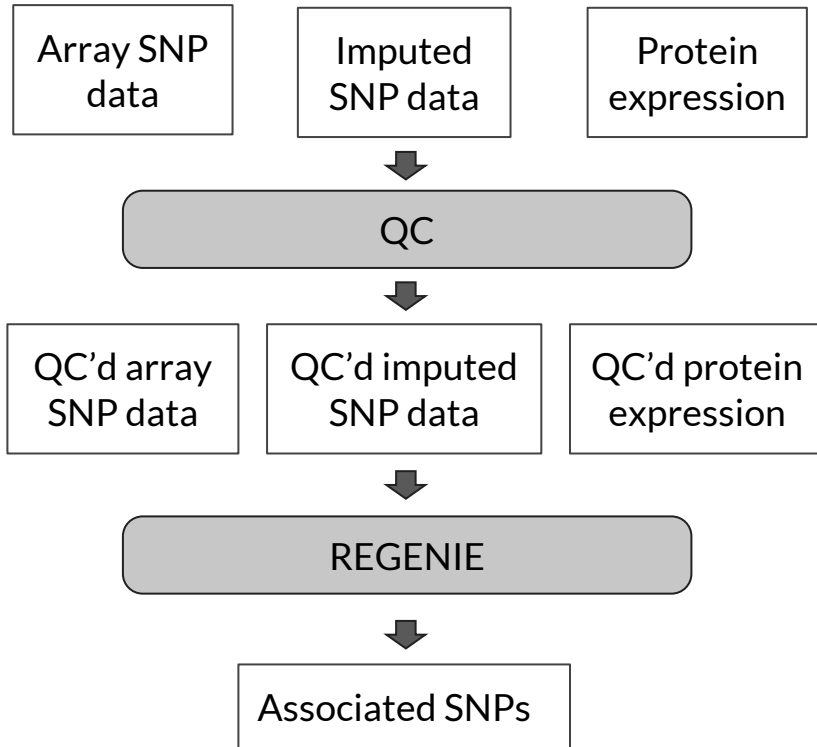


Second step - test variant-phenotype association

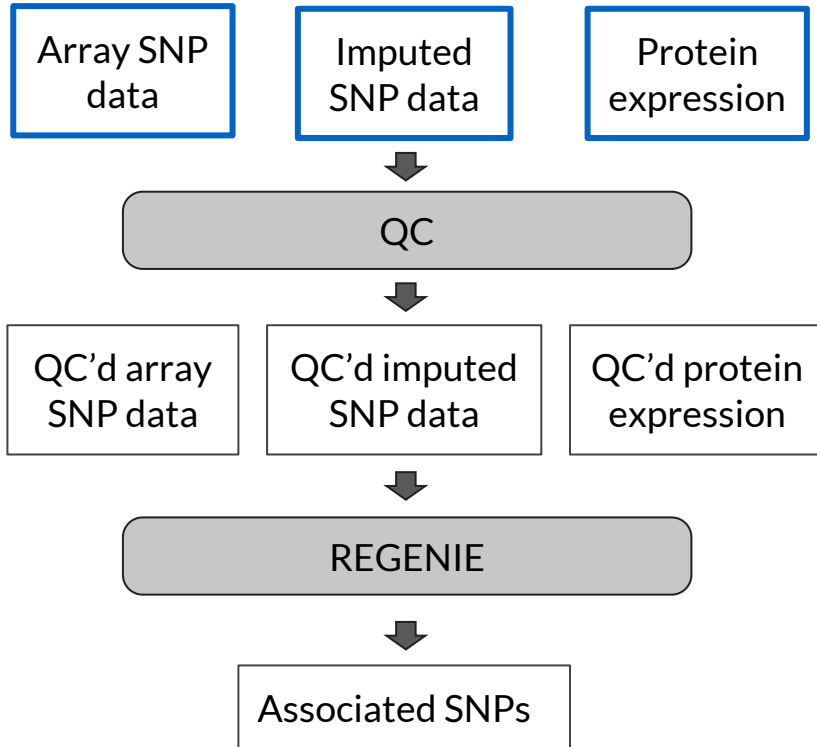


REGENIE paper
End-to-end Target Discovery webinar

Approach

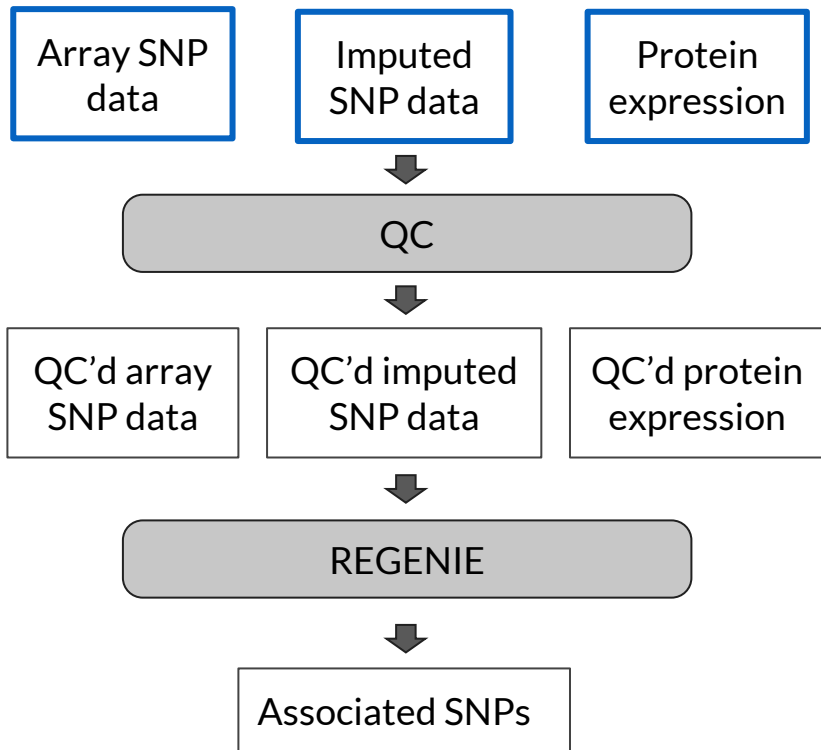


Simulate input data

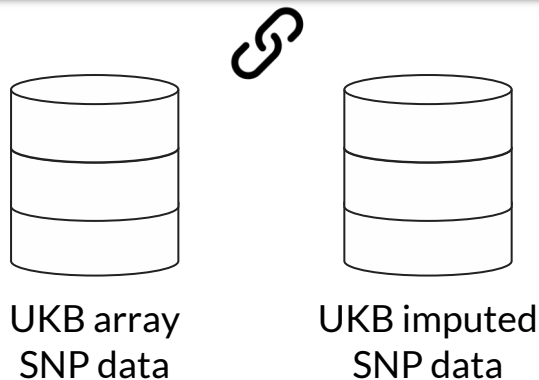


NOTE: We are using public proteomic data, not UKB data, for demonstration purposes

Matched genotype and protein expression data

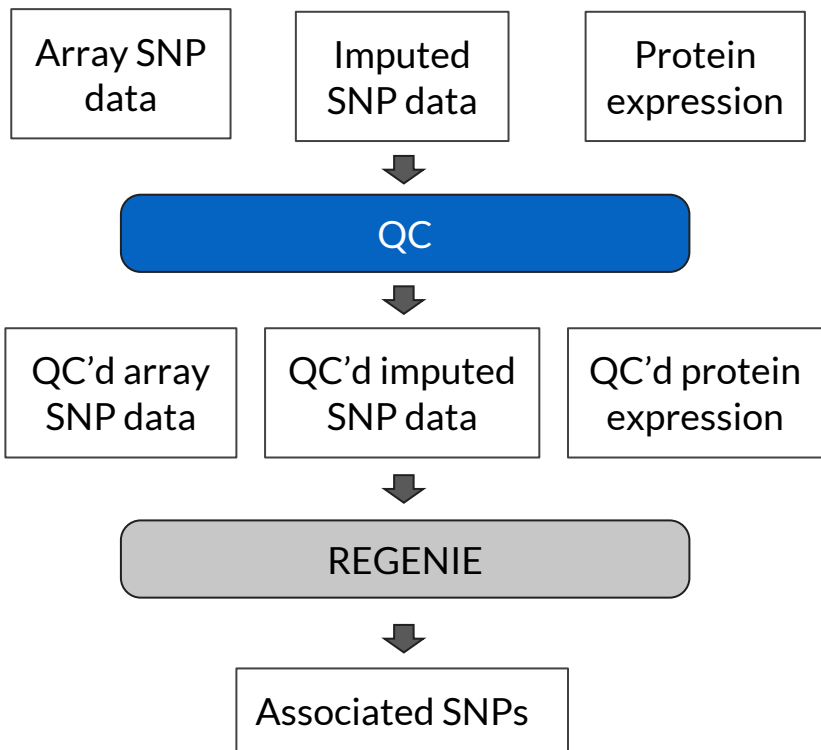


	IID	CA1	ICAM1	CHL1	TGFBI	ENG
FID						
2894753	2894753	7.62107	6.79971	4.73174	9.33471	3.12445
2352368	2352368	6.96085	6.98459	4.31338	9.06819	3.31576
1483346	1483346	7.16983	7.04907	4.72713	8.92804	3.16308
2352196	2352196	7.45724	6.89523	4.57029	9.27165	3.06199
4886500	4886500	7.81354	6.71708	4.93904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



[Notebook to simulate data](#)

QC input data



	IID	CA1	ICAM1	CHL1	TGFBI	ENG
FID						
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

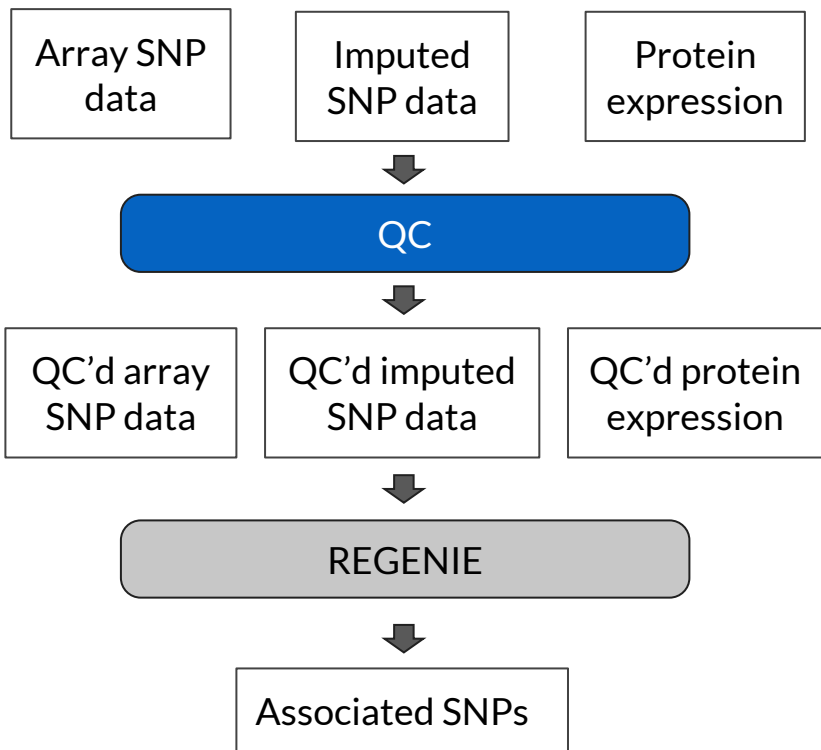


UKB array
SNP data



UKB imputed
SNP data

QC input data



	IID	CA1	ICAM1	CHL1	TGFBI	ENG
FID						
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	6.71758	4.93904	9.51350	3.66898



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

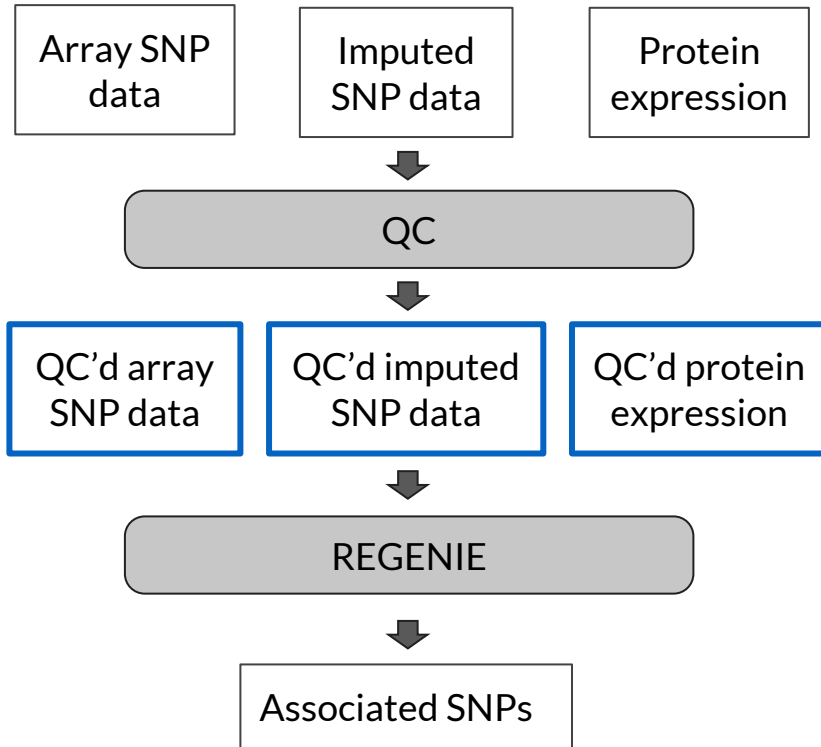


UKB array SNP data

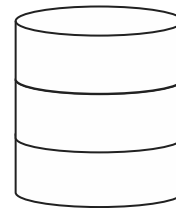


UKB imputed SNP data

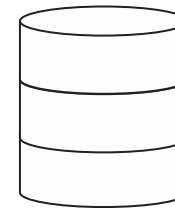
Get QC'd input data



	IID	CA1	ICAM1	CHL1	TGFBI	ENG
FID						
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



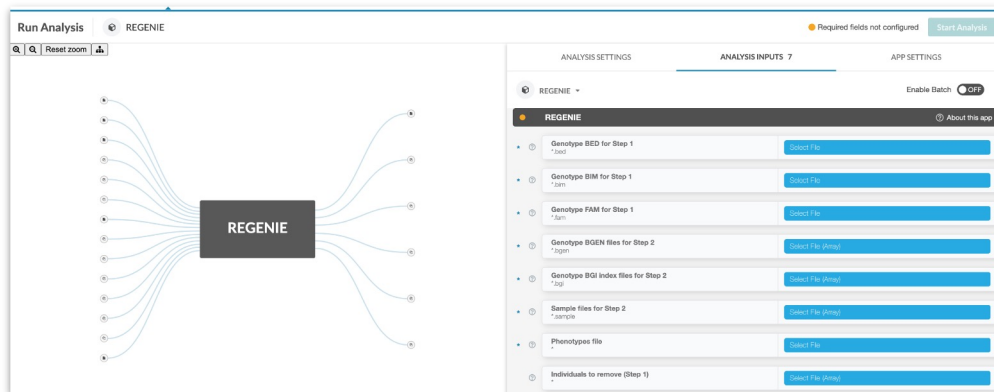
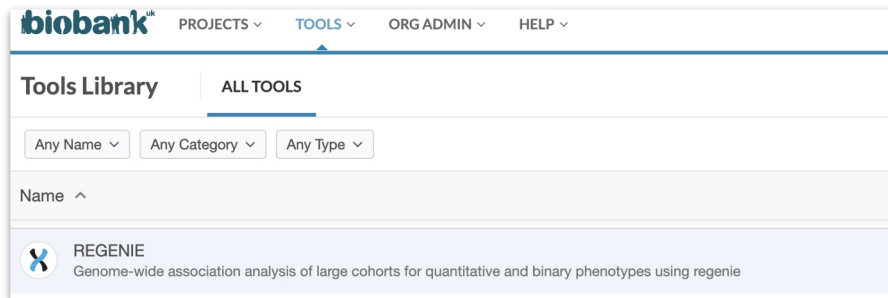
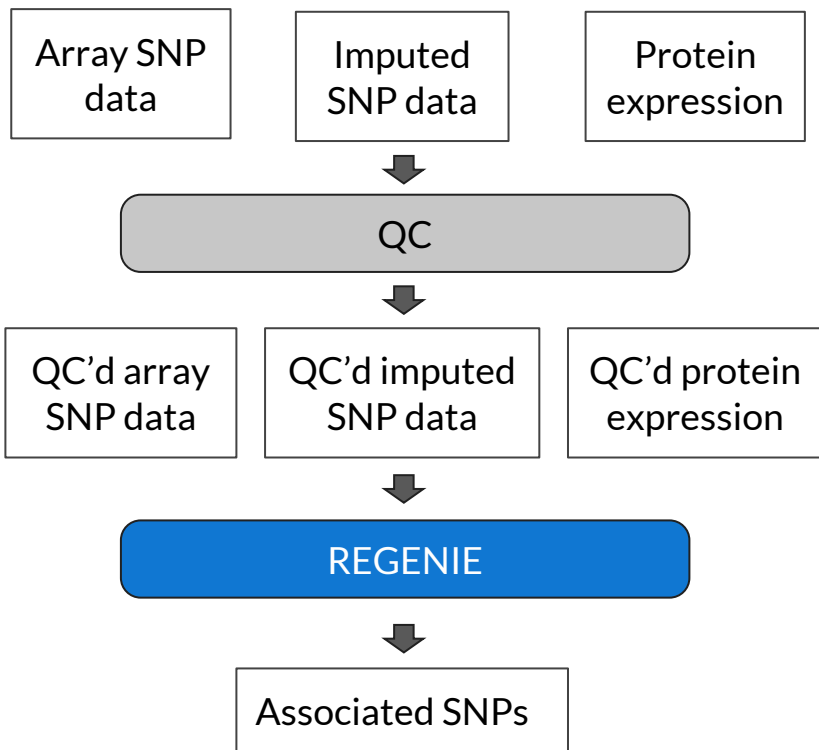
UKB array SNP data



UKB imputed SNP data

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

Run GWAS



REGENIE GWAS analysis settings

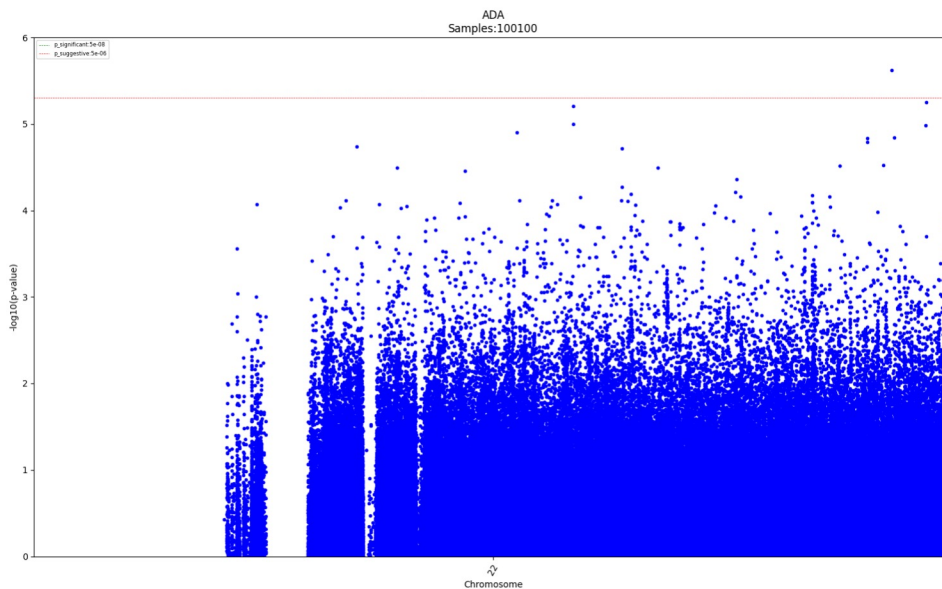
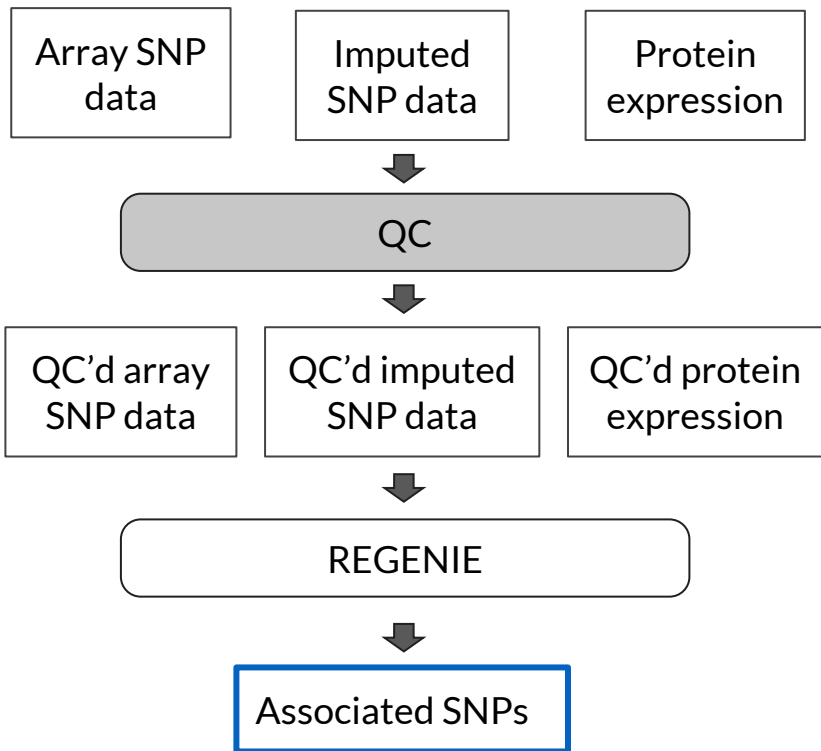
Step 1

- ▶ **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_ssd1_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 DNAxus

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 DNAxus webinars, trainings, and roundtable discussions

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

Click [Here](#) to Join 

OR



SCAN ME

Acknowledgements



Ondrej Klempir, PhD
Sr. Community Engagement
Scientist



Anastazie Sedlakova, PhD
Community Engagement/
Principal Scientist



Arkarachai Fungtammasan, PhD
Scientific Community Manager/
Principal Scientist

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

Thank you! Questions?