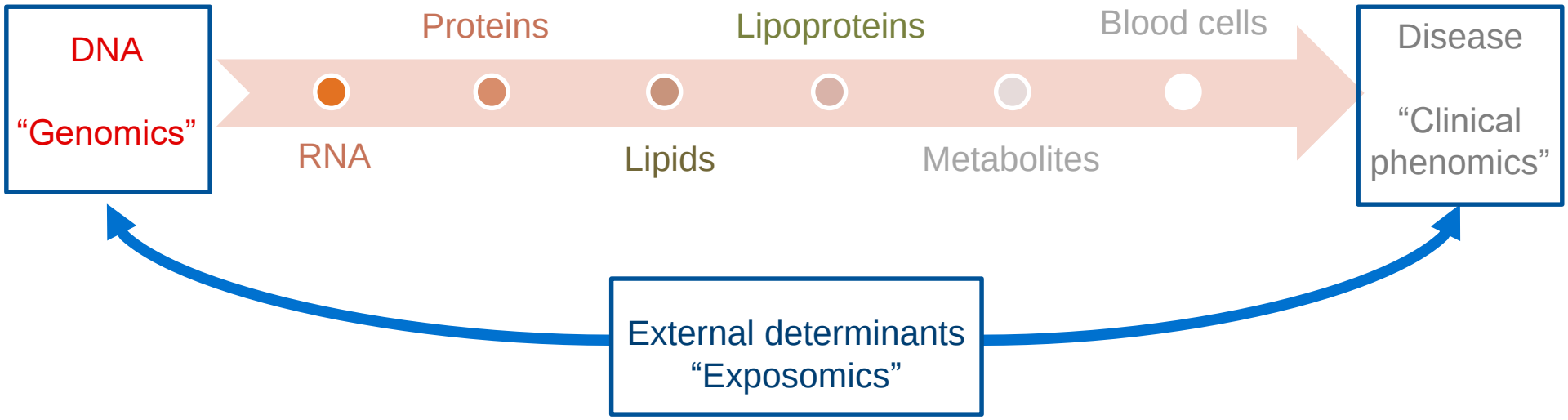
Multi-omic studies at population scale: opportunities and challenges

John Danesh
University of Cambridge

What are ‘omics?

Extensive (or even complete) measurement of a particular domain, eg:



What is the potential value of molecular ‘omics?

To offer new insights into:

- Biology
- Disease aetiology / subclassification
- Risk prediction
- Therapeutic targeting

Some key challenges to enable robust inference from ‘omics

Need to:

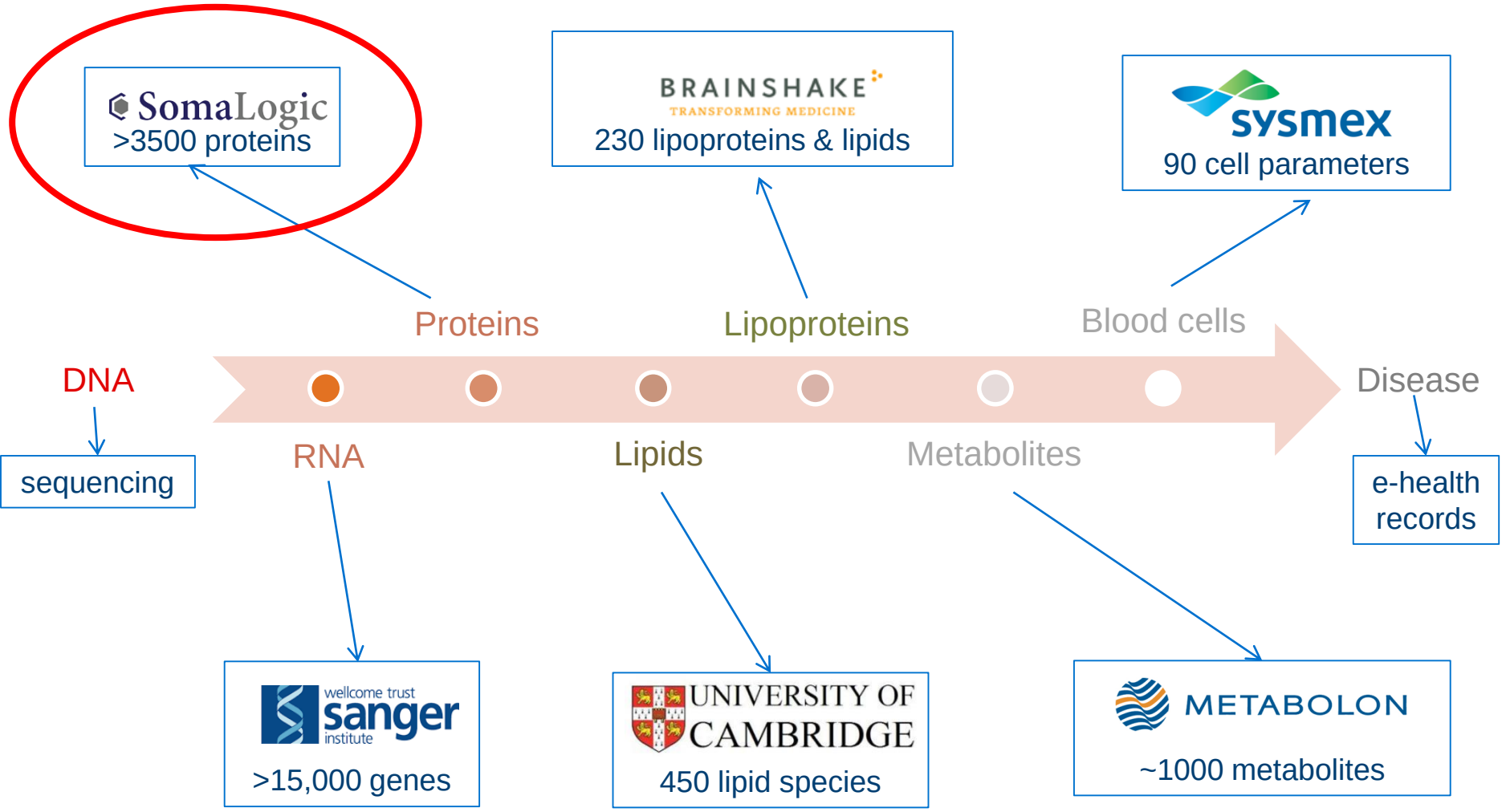
- Validate assays in population studies
- Understand complexities in interpreting assays
- Control for technical artefacts
- Address statistical and computational challenges

Some key challenges to enable robust inference from ‘omics

Need to:

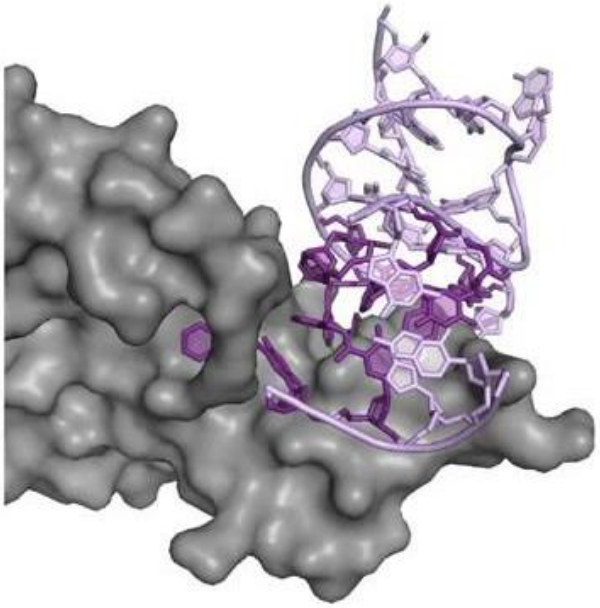
- Validate assays in population studies
- Understand complexities in interpreting assays
- Control for technical artefacts
- Address statistical and computational challenges

INTERVAL: a proof-of-concept study of 50,000 people

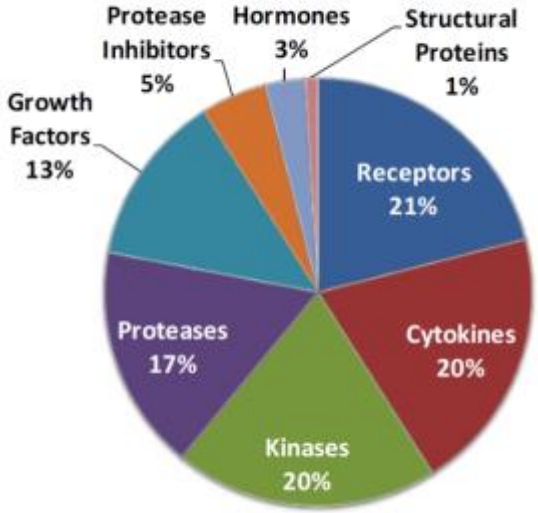


SomaScan plasma proteomics

~5000 proteins across 8 orders of magnitude of abundance



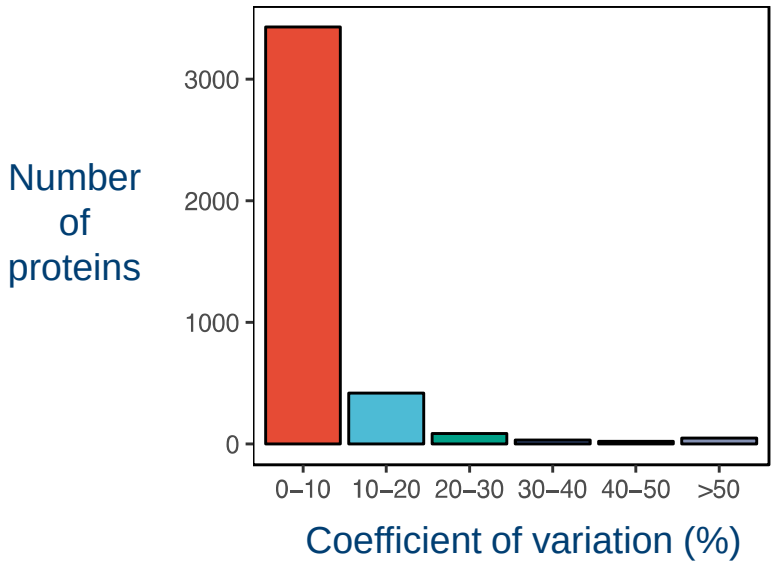
High-specificity aptamer-based approach



Diverse set of proteins, with a bias towards clinical and pharmaceutical relevance

Population-scale validation of the SomaScan assay

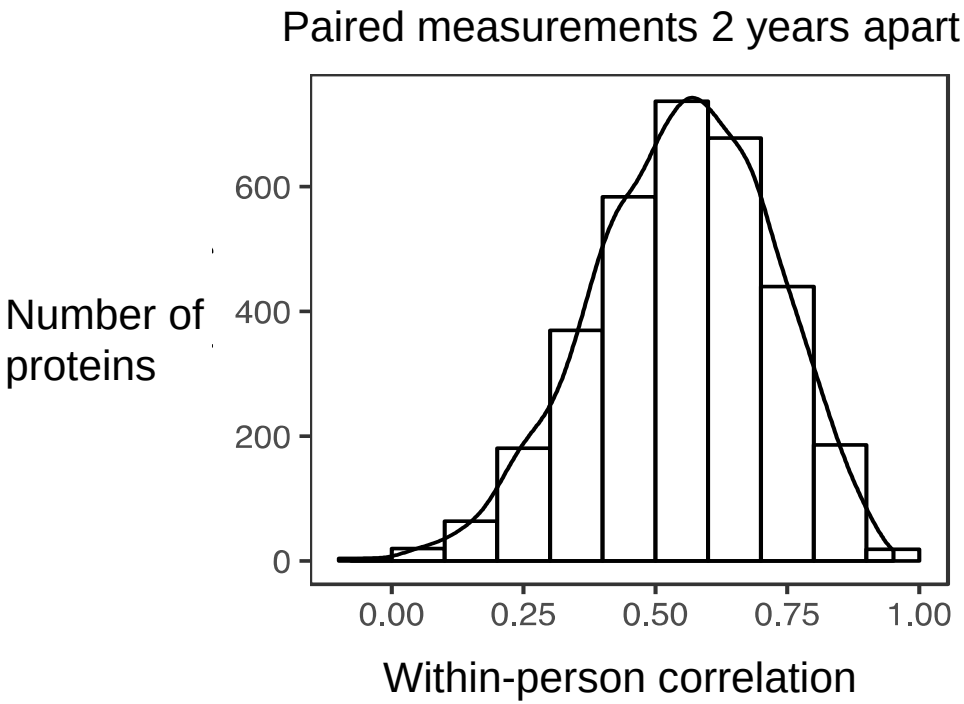
- What is the variability between assays run with the same samples?
  - More than 90% of proteins had CVs less than 20%



Sun *et al.*, *Nature* (in press)

Population-scale validation of the SomaScan assay

- What is the variability between assays run in the same samples?
- How stable are protein levels within participants over time?



Sun *et al., Nature* (in press)

Population-scale validation of the SomaScan assay

- What is the variability between assays run in the same samples?
- How stable are protein levels within participants over time?
- Can well-known associations of proteins with phenotypes be reproduced?

For example,

- cystatin C and beta-2-microglobulin were associated with eGFR, a marker of renal function
- leptin, insulin and ghrelin were associated with body-mass index, a marker of obesity

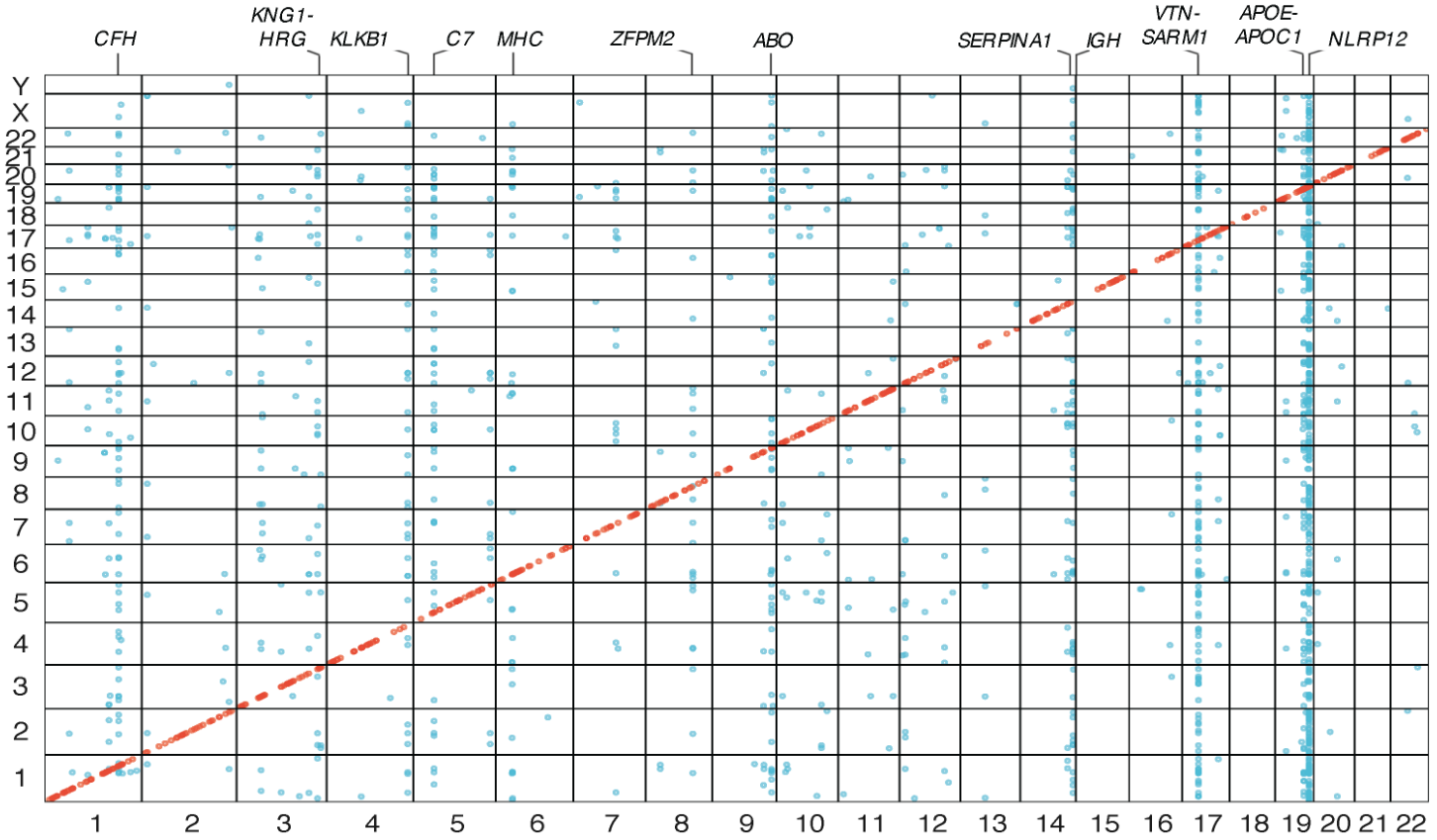
Sun *et al., Nature* (in press)

Population-scale validation of the SomaScan assay

- What is the variability between assays run in the same samples?
- How stable are protein levels within participants over time?
- Can well-known associations of proteins with phenotypes be reproduced?
- Can biologically plausible genetic associations be detected?

“Genetic validation” of the assay

Genes encoding target proteins

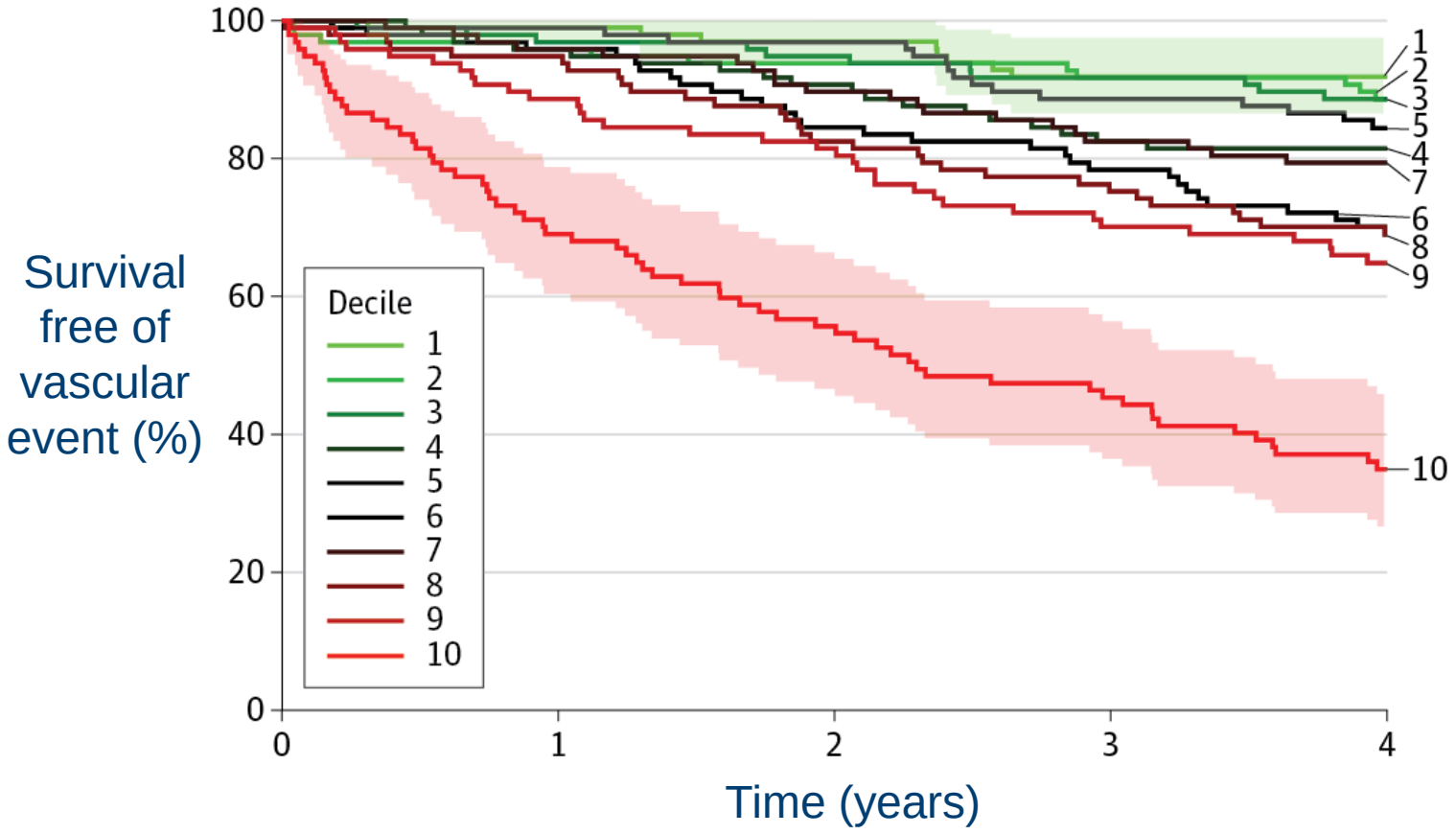


Genetic variants

Sun *et al.*, *Nature* (in press)

Predictive inference is enabled by such population scale validation

Adding 9 proteins from the SomaScan could enhance conventional vascular risk prediction



Ganz *et al.*, *JAMA* 2016

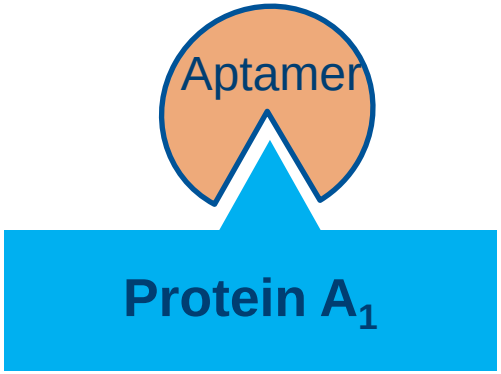
Some key challenges to enable robust inference from ‘omics

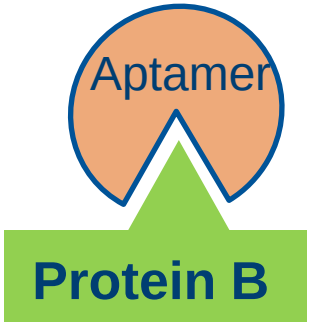
Need to:

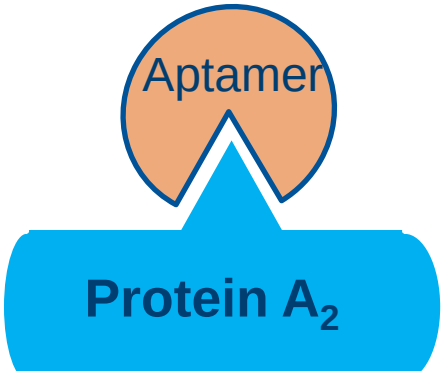
- Validate assays in population studies
- Understand complexities in interpreting assays
- Control for non-biological variation
- Address statistical and computational challenges

Aetiological inference for protein assays involves additional considerations

- Does the measurement reflect only one protein?





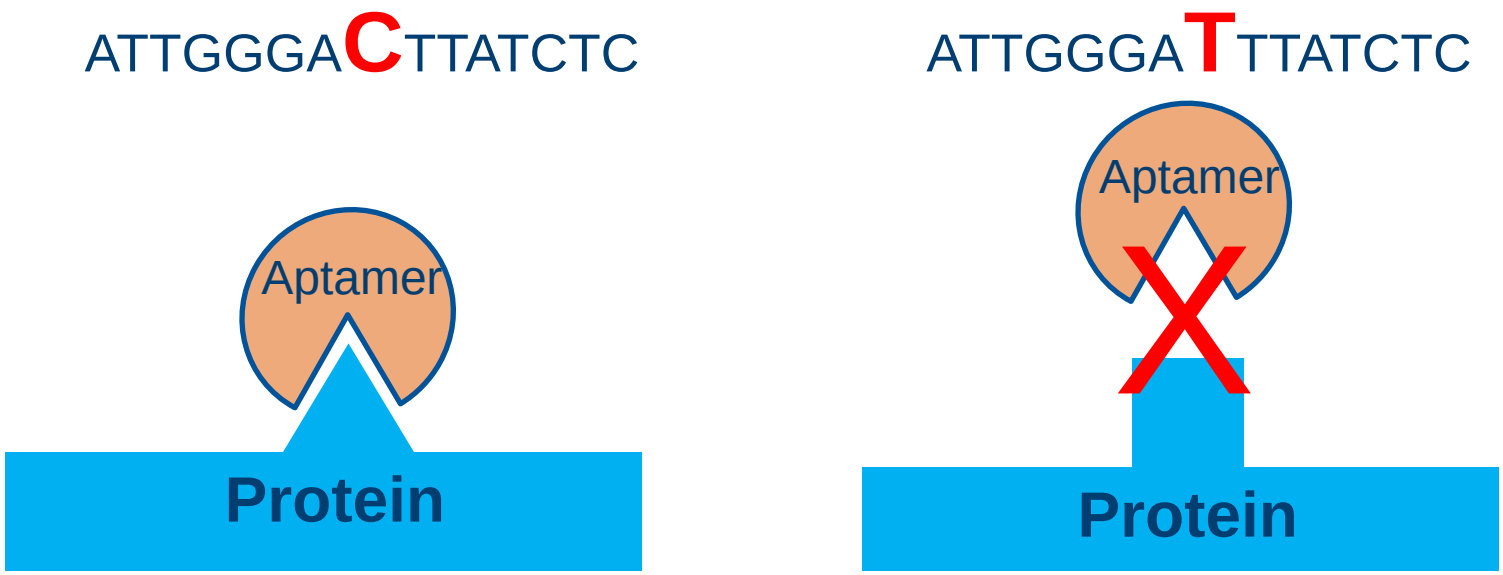


However, 87% of ~1000 SomaScan aptamers tested reflected only one protein

Sun *et al.*, *Nature* (in press)

Aetiological inference for protein assays involves additional considerations

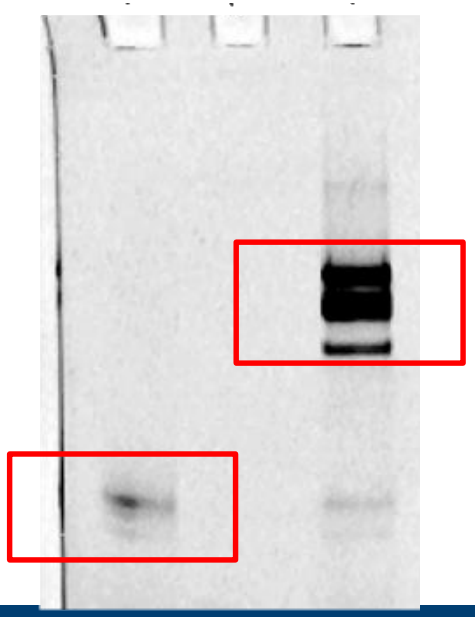
- Does the measurement reflect only one protein?
- Does the measurement reflect different versions of the same protein rather than differences in protein levels?



Aetiological inference for protein assays involves additional considerations

- Does the measurement reflect only one protein?
- Does the measurement reflect different versions of the same protein rather than differences in protein levels?
- Does the measurement reflect free protein or a “protein complex”?

Mass-spec ‘pull-down’ of proteinase-3 (PR3) aptamer



Strong binding to PR3:A1AT complex

Weak binding to free PR3

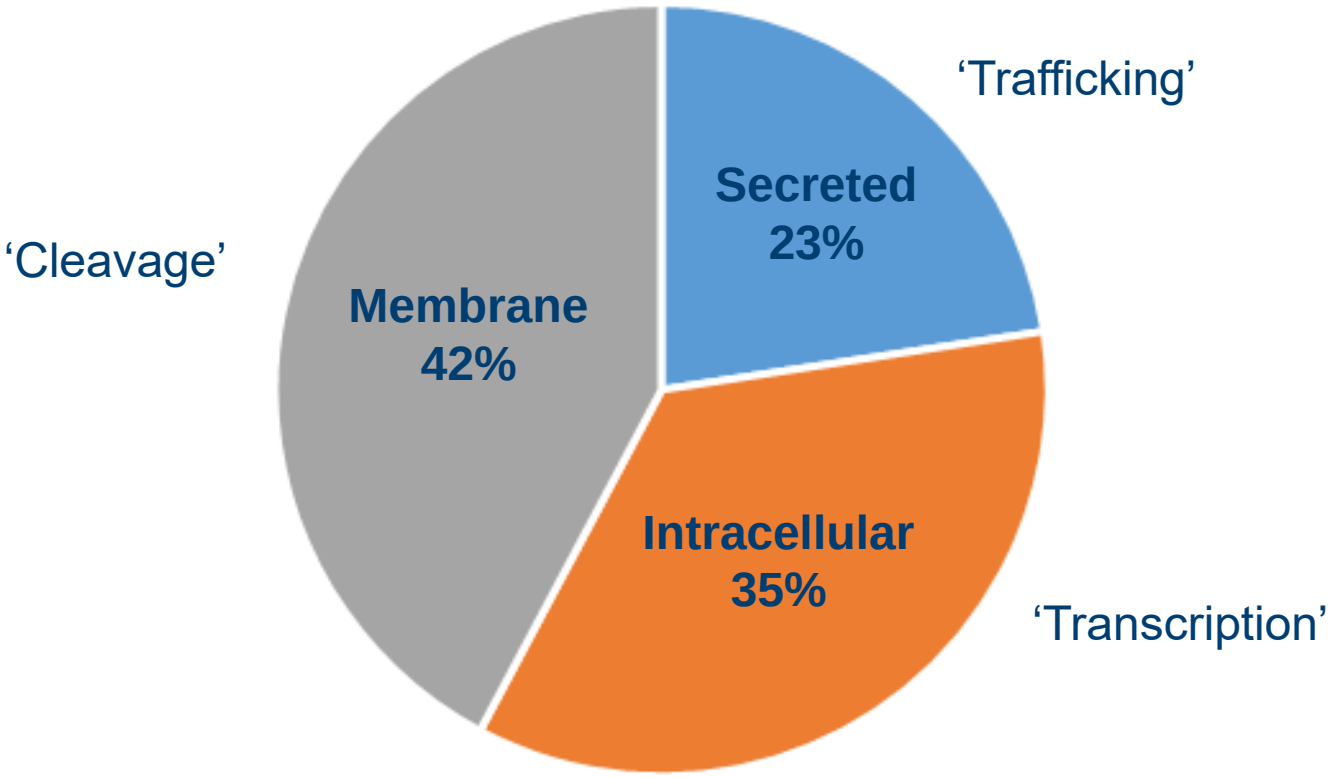
Sun *et al.*, *Nature* (in press)

Aetiological inference for protein assays involves additional considerations

- Does the measurement reflect only one protein?
- Does the measurement reflect different versions of the same protein rather than differences in protein levels?
- Does the measurement reflect free protein or a “protein complex”?
- What do protein levels in plasma reflect?

Protein concentrations in plasma may not reflect cell- or tissue-specific concentrations

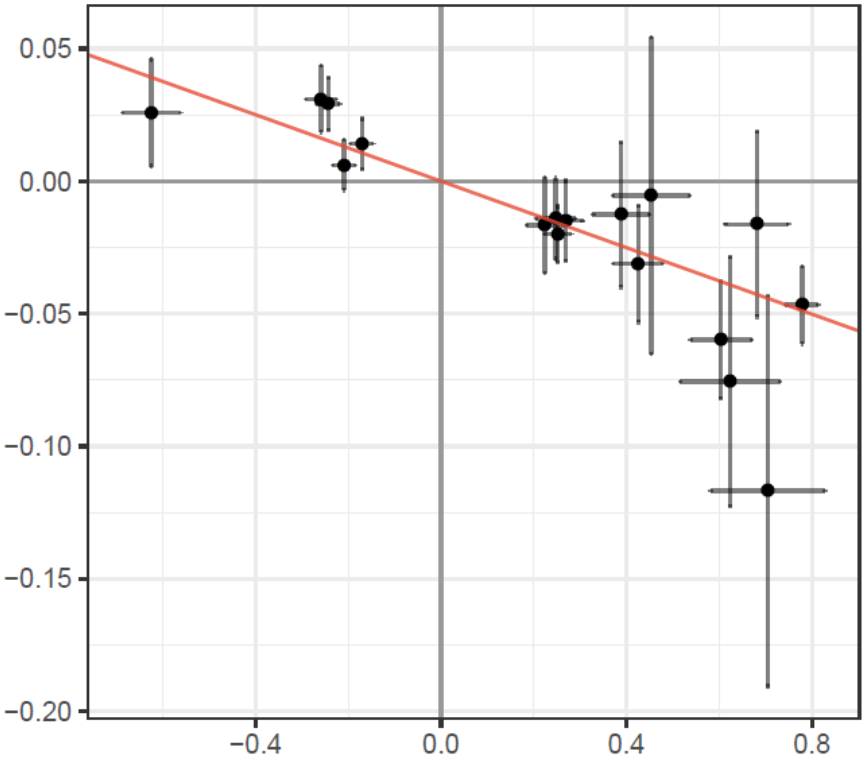
Proteins measured by the SomaScan assay
are influenced by different processes



Sun *et al.*, *Nature* (in press)

Plasma proteomics can help yield potential causal insight into disease

Mendelian randomisation suggests a causal role for matrix metalloproteinase-12 in coronary disease



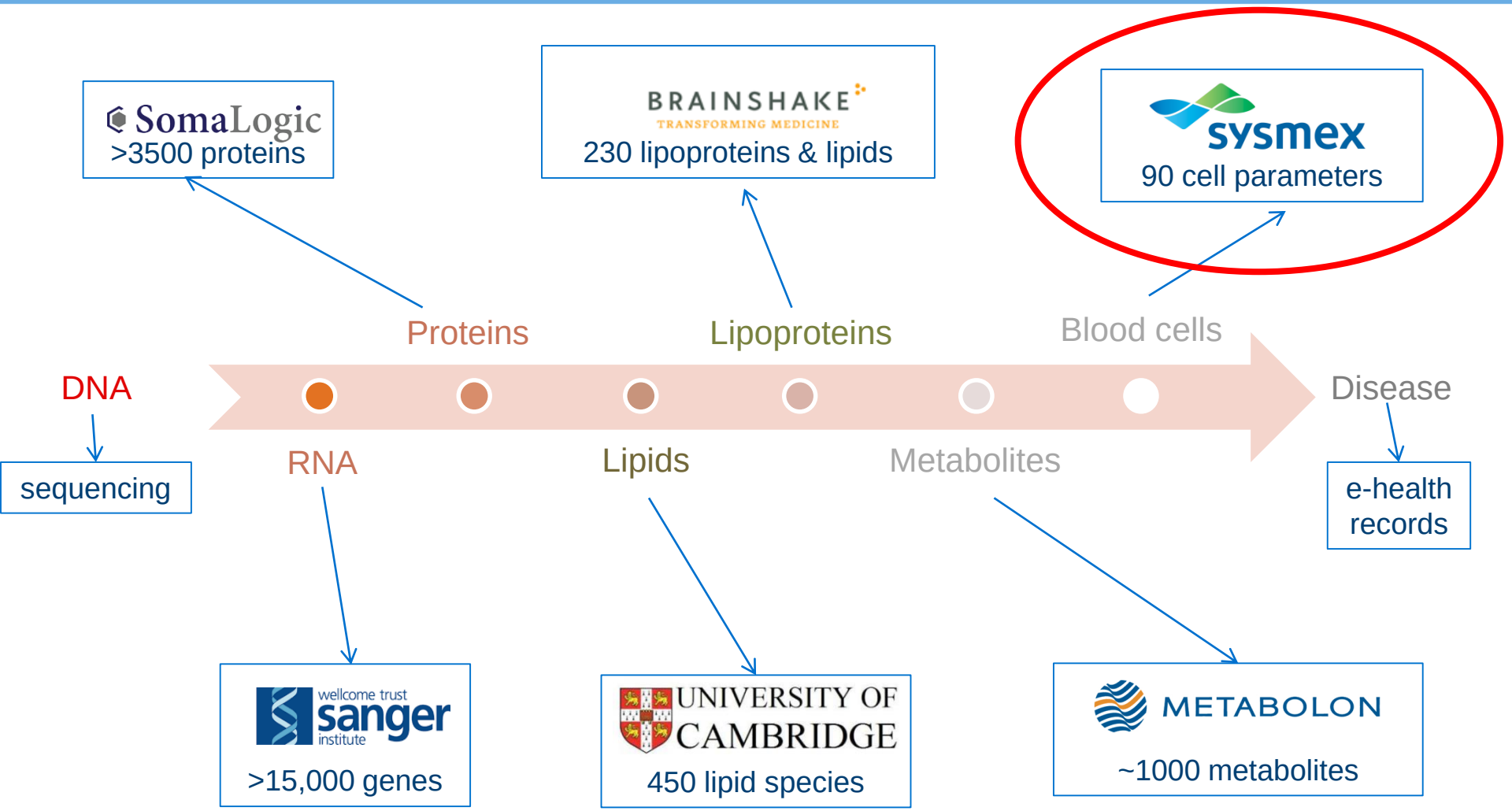
Sun *et al.*, *Nature*, in press

Some key challenges to enable robust inference from ‘omics

Need to:

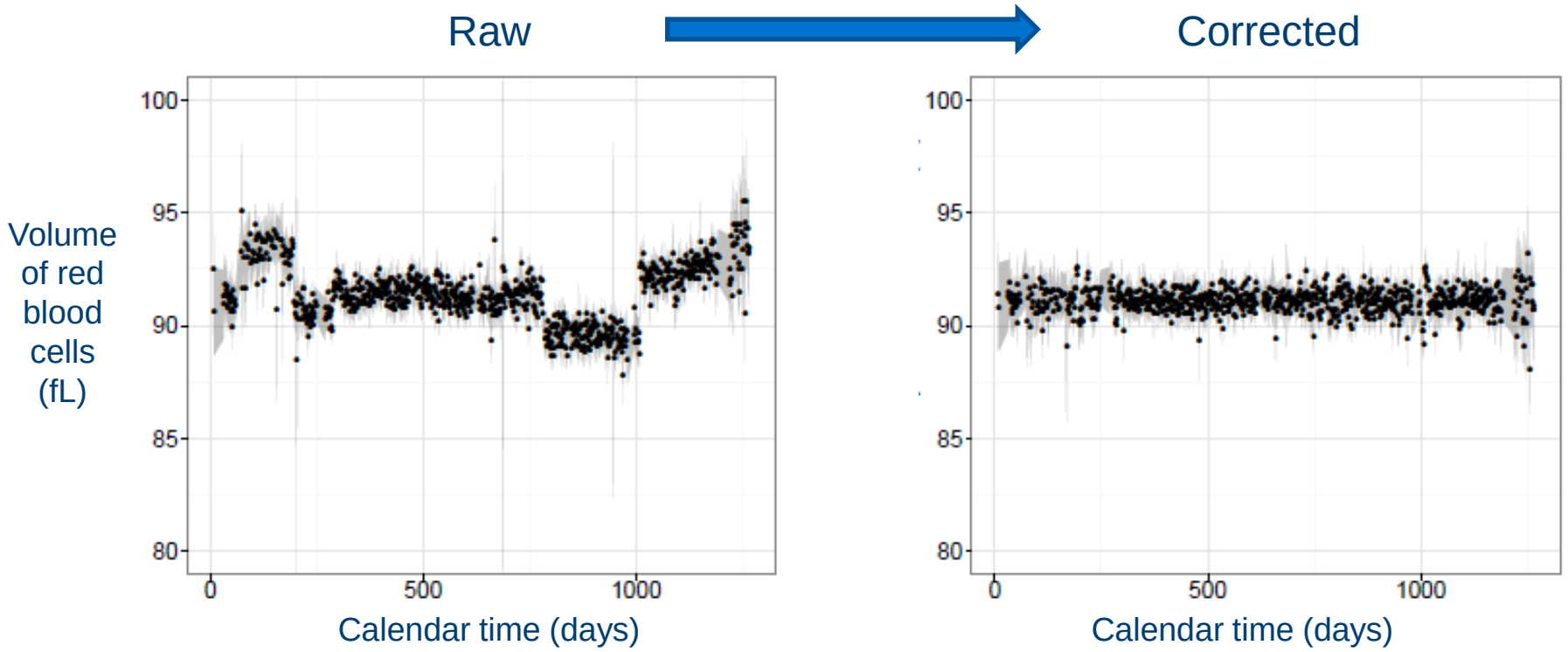
- Validate assays in population studies
- Understand complexities in interpreting assays
- Control for non-biological variation
- Address statistical and computational challenges

INTERVAL: a proof-of-concept study of 50,000 people



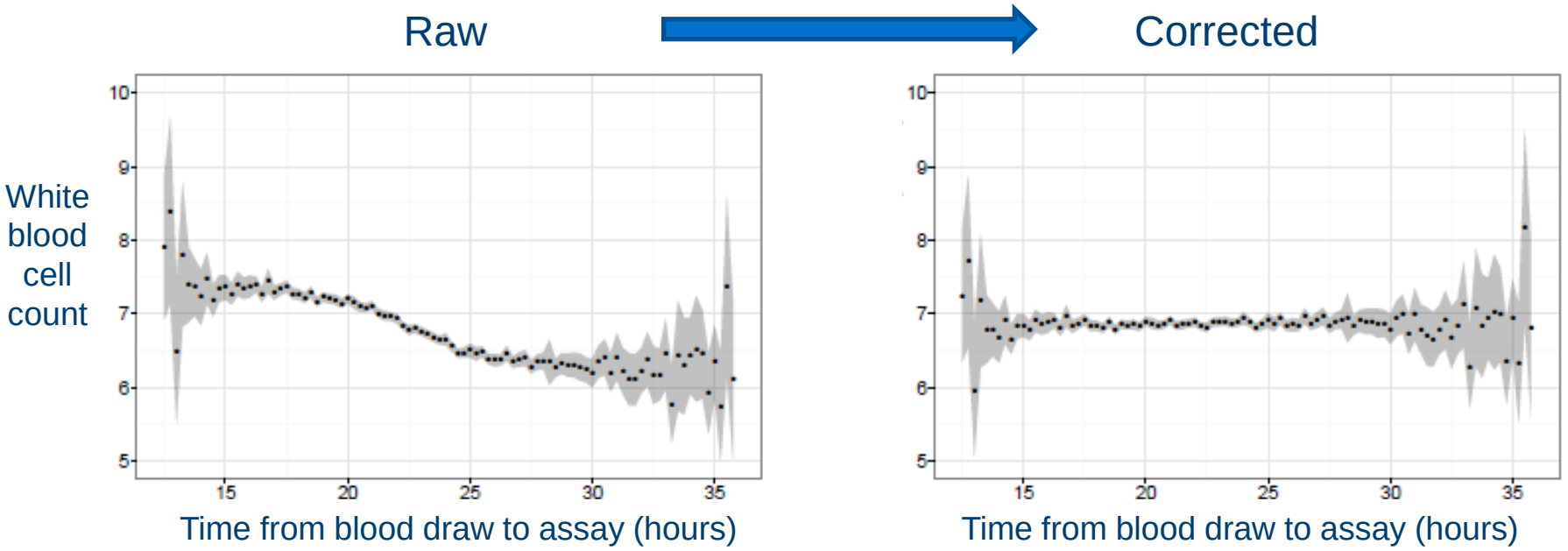
Non-biological variation is not always identified but is potentially correctable

- Batch effects due to lengthy periods of assay



Technical artefacts are important but correctable sources of variability

- Batch effects due to lengthy periods of assay
- Subtle distortions due to operational effects



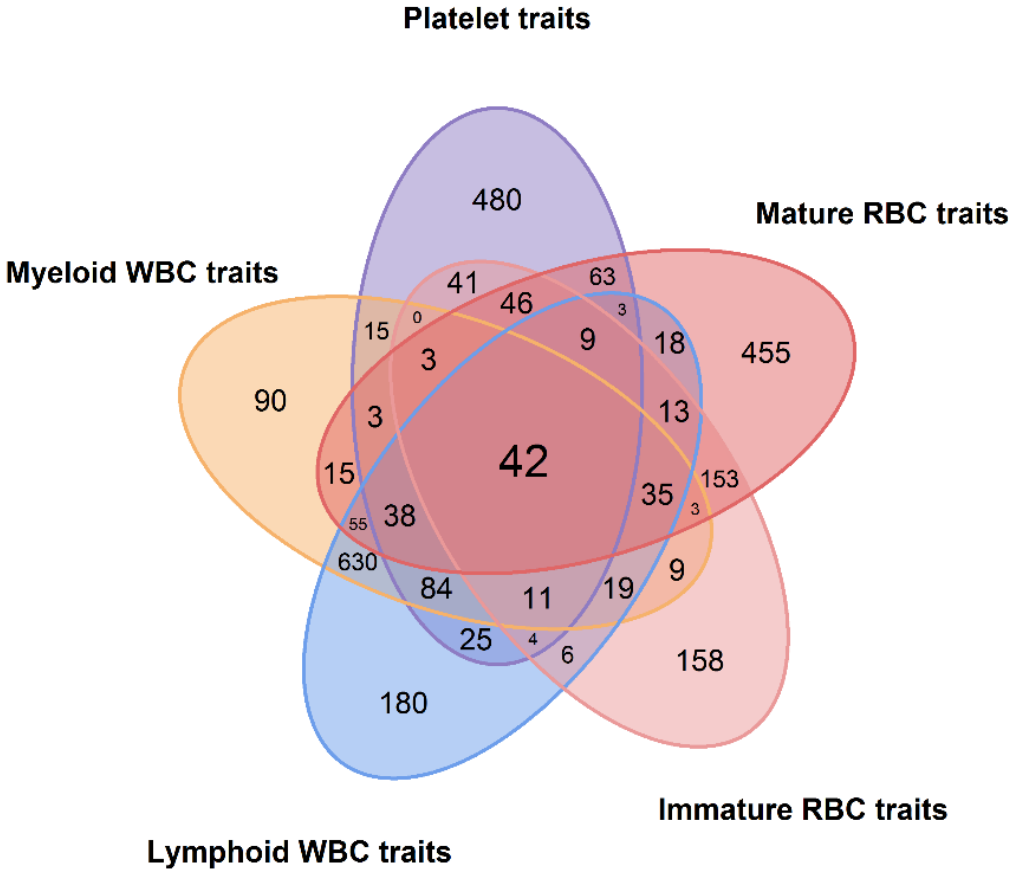
Discovery power can be substantially enhanced after correction for non-biological variation

	Uncorrected	→	Corrected
Number of RBC loci discovered with n=50,000	10		36

RBC = red blood cell count

Blood cell ‘omics can help yield biological insight

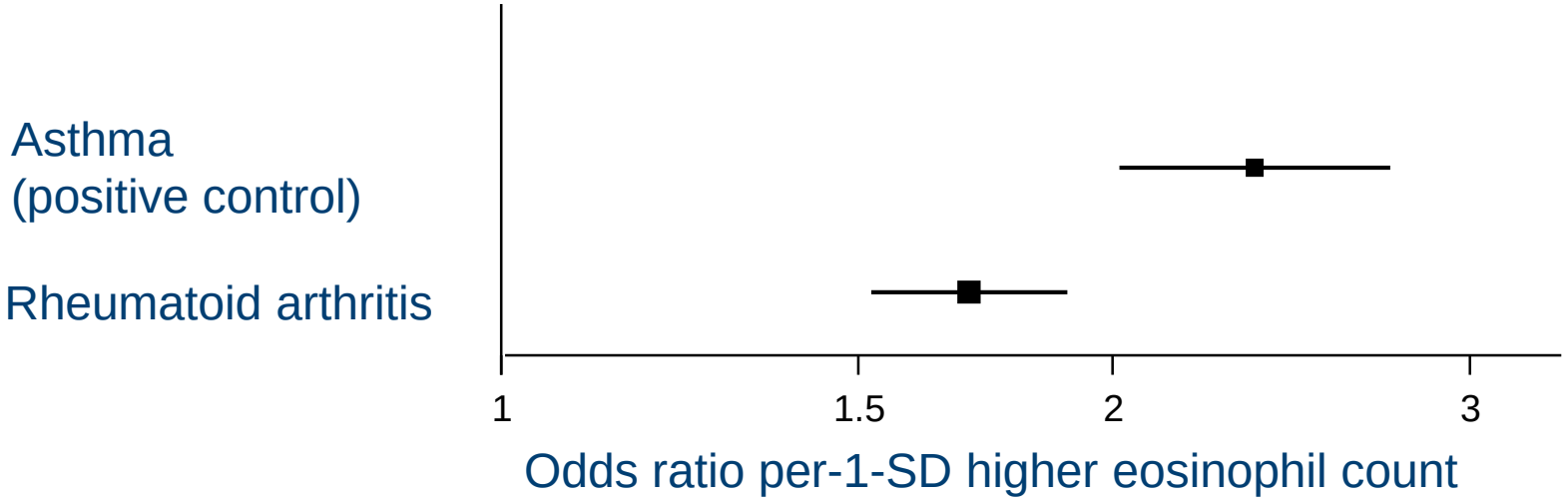
Discovery of the genetic architecture of blood cell traits: 2700 variants



Astle *et al.*, *Cell* 2016

Blood cell ‘omics can help yield potential causal insights into disease

Mendelian randomisation suggests a causal role for eosinophils in rheumatoid arthritis



Astle *et al., Cell* 2016

Some key challenges to enable robust inference from ‘omics

Need to:

- Validate assays in population studies
- Understand complexities in interpreting assays
- Control for non-biological variation
- Address statistical and computational challenges

Non-trivial computational challenges of combining multi-omic data

~ 50,000 participants

~ 1600 distinct lipids / metabolites (plus ratios/combinations)

~ 80 million genotypes

➔ ~ 130 billion calculations

High-performance computing is essential

	Time to achieve ~130 billion calculations
Conventional computing	~1300 days
High-performance computing*	~4 days

* Cambridge University Peta 5 HPC

Conclusions

- Multi-omics can help address the post-GWAS grand challenge of bridging molecular gaps from genotype to disease

- ‘Omic assays have common and assay-specific technical and interpretive challenges

- Several assays are being used at population scale, with results being pooled across cohorts

“Progress in science depends on new techniques, new discoveries and new ideas, probably in that order.”

- Sydney Brenner

Acknowledgements

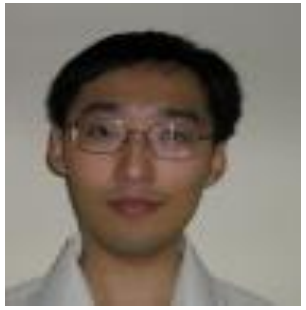
Ben Sun

Adam Butterworth

Jimmy Peters

Will Astle

Nicole Soranzo



Tao Jiang

Funders


British Heart
Foundation


MRC
Medical
Research
Council


NHS
National Institute for
Health Research


erc
European
Research
Council


NHS
Blood and Transplant


SEVENTH FRAMEWORK
PROGRAMME


wellcome


MERCK
INVENTING FOR LIFE

Examples of Cambridge-led 100K+ cohorts

UK blood donor cohorts (consented for recall)

Today: 100K participants

Recruiting 250K further participants

South Asian cohorts (enriched for autozygosity)

Today: 95K participants

Recruiting 150K further participants