Value of proteomics at population scale

UK Biobank Annual meeting

20th June 2018

Adam Butterworth, Reader in Molecular Epidemiology
What is the potential value of molecular ‘omics eg, soluble proteomics?

To offer new insights into:

- Biology
- Disease aetiology / subclassification
- Risk prediction
- Therapeutic targeting
The INTERVAL study is a systems genomics study of 50,000 blood donors. The study includes the analysis of 350 proteins, 3,500 proteins, and 450 lipid species. Additionally, the study includes 230 lipoproteins & lipids, 90 cell parameters, >15,000 genes, 450 lipid species, and ~1000 metabolites. The data is generated using Affy array, WES (50X), WGS (15X), and e-health records.
Why are proteins a valuable domain to study?

• Fundamental ‘effectors’ of biology and often therapeutic targets

• Maturation of wide-angled protein assays \(\rightarrow\) relevant to a broad set of diseases and phenotypes

• Imperfect correlation with RNA/metabolomics \(\rightarrow\) an additional biological dimension

• “Proximal” endophenotypes \(\rightarrow\) good for causal inference
SomaScan: a broad high-throughput aptamer-based proteomics assay

~5000 proteins across 8 orders of magnitude of abundance

High-specificity aptamer-based approach

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

Lollo et al, Proteomics 2014
“Genetic validation” of the assay: ~2000 genotype-protein associations

Genes encoding target proteins

Genetic variants
- cis
- trans

Sun et al, Nature 2018
Predictive inference is enabled by population scale validation

A 9 protein score can predict future vascular events

e.g., people with higher levels of MMP-12 have higher risk of events

Survival free of vascular event (%) vs Time (years)

Decile 1 to 10

Ganz et al, JAMA 2016
Soluble proteins can help yield potential causal insights into disease

Integrating genetics and proteomics ("Mendelian randomisation") suggests a protective role for matrix metalloproteinase-12 in coronary disease

Risk of coronary heart disease (per-allele log odds ratio)

Effect on plasma MMP-12 levels (per-allele change in SD)

Sun et al, Nature 2018
Soluble proteins to identify gene-disease relationships

Complex genetic region associated with inflammatory bowel disease (IBD)

Liu et al, Nature Genetics 2015
Soluble proteins to identify gene-disease relationships.

IBD signal matches signal for MST1 protein.
Soluble proteins to identify gene-disease relationships

IBD signal also matches BLIMP-1 protein signal

Sun et al, Nature 2018
There is also an IBD signal in the *PRDM1* gene, which codes for the BLIMP-1 protein.
Some key challenges to enable robust inference from molecular ‘omics

Need to:

- Validate assays in population studies
- Understand complexities in interpreting assays (eg, epitope effects, free vs complexed proteins)
- Control for non-biological variation (eg, batch effects)
- Address statistical, computational and bioinformatics challenges
Conclusions

- Multiplex proteomic assays are becoming sufficiently high-throughput and reliable for use in large population studies.

- Such wide-angled assays, when performed at scale, can answer a diverse range of questions, including biological and aetiological insight, disease prediction etc.

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

Cambridge
Ben Sun
Jimmy Peters
John Danesh
Dirk Paul
David Stacey
James Staley
& many others

Merck
Joe Maranville
Heiko Runz
Caroline Fox
Robert Plenge
>100 genetic loci associated with coronary disease
Allelic series to inform therapeutic target validation

Why did darapladib Phase 3 trials fail?

- coronary disease outcomes
- soluble Lp-PLA₂ activity levels

Risk ratio for CHD (95% CI)

Val379Ala Per allele
Val279Phe (LoF) Heterozygotes vs. common homozygotes
Four LoF* variants Heterozygotes vs. common homozygotes
Darapladib 160mg vs. Placebo
Val279Phe (LoF) Homozygotes (complete LoF) vs. common homozygotes

Gregson et al, Eur J Card Prev, 2017
INTERVAL – a 50,000 participant bioresource

- 50,000 healthy blood donors nested within a randomised trial
- Aged 18-80 (mean ~43)
- ~50% male / ~50% female
- Recruited at 25 centres across England
- Stored serum, plasma, buffy coat, whole blood, RNA
Key features of the INTERVAL Bioresource

1. **Scale**
   - 50,000 blood donors aged 18-80 recruited across England, equal split of men/women

2. **Dense multi-omic phenotyping**
   - >25,000 molecular markers including lipids, lipoproteins, metabolites, RNA-sequencing, blood cell traits, autoantibodies etc

3. **Healthy participants**
   - Blood donors are largely free from the potential distorting effects of existing disease and medication use

4. **Linkage to electronic health records**
   - Donors are linked to mortality and hospital in-patient registries, with further linkages to out-patient registries, primary care records and disease-specific registries planned

5. **Recallability**
   - Participants can be brought back on the basis of protein level or genotype for detailed characterisation
Measurement of ~7000 soluble traits: ‘omics selected because of causal precedents

>7000 soluble traits

DNA → RNA → Proteins → Metabolites → Lipids → Other intermediate traits → Disease

50,000 healthy participants with ‘omics assays in subsets
Other phenotypes and consent for further research

**Physical activity profiling:** 7 day accelerometer recordings

**Cognitive function tests:** Attention (Stroop test), Memory (Pairs test), Executive function (Trail test), Intelligence (Reasoning test)

**Health records data:** linkage to multiple health records (eg, death registries, hospital episode statistics, morbidity registers) to follow-up incident health events

**Future studies:** consent to be contacted about future studies, including recall-by-phenotype/genotype studies
SOMAscan is overrepresented for secreted and membrane-bound proteins.
Within-person 2-year reproducibility of protein levels
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 2018
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 2018
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 2018
Genetic analysis of soluble proteomics at scale

3,600 proteins

10.1 million variants

3,300 blood donors

GWAS to discover the genetic determinants of soluble protein levels

765 genomic regions associated with 1528 proteins at $p<1.3\times10^{-11}$
1928 associations ("pQTLs") involving 1019 unique sentinel variants

Sun et al, Nature 2018
Genomic atlas of the human plasma proteome: ~2000 genetic associations

- Four-fold more associations
- Replicate almost all previous associations
- ~1/4 cis, ~3/4 trans
- 60 of the loci overlap with GWAS hits for disease
- 40% of cis signals overlap with cis RNA expression loci

Sun et al, Nature 2018
Genetic architecture of pQTLs

Sun et al, Nature 2018
*cis* signals cluster near to the start of the relevant genes

Sun et al, *Nature* 2018
Missense variants are strongly overrepresented amongst the sentinel variants
Genetic replication across proteomic assays

- 106 (65%) out of 163 pQTLs replicated using Olink (81% of cis, 52% of trans)

Sun et al, Nature 2018
Creation of a translational toolkit for MR from genetic discoveries

- For >600 proteins we explain >5% of variance → strong instruments

Sun et al, Nature 2018
Creation of a translational toolkit for MR from genetic discoveries

- More than 300 proteins have multiple associated regions of the genome
Creation of a translational toolkit for MR from genetic discoveries

- 435 associations (271 cis) have multiple independent variants
Cis pQTL for MMP12 levels

Sun et al, Nature 2018
Multiple independent signals at *MMP12* locus

Sun et al, Nature 2018
MR suggests inverse causal association with CHD

MMP12
- Univariate MR (conditionally significant variants only)  
  P-value: $3.7 \times 10^{-4}$
- Univariate MR  
  P-value: $2.8 \times 10^{-13}$
- Multivariable MR  
  P-value: $8.6 \times 10^{-6}$
- Univariate MR (Olink, conditionally significant variants only)  
  P-value: $4.2 \times 10^{-4}$
- Univariate MR (Olink)  
  P-value: $8.4 \times 10^{-11}$
- Multivariable MR (Olink)  
  P-value: $1.7 \times 10^{-4}$
- Univariate MR (Suhre, conditionally significant variants only)  
  P-value: $4.2 \times 10^{-4}$
- Univariate MR (Suhre)  
  P-value: $3.7 \times 10^{-13}$
- Multivariable MR (Suhre)  
  P-value: $1.0 \times 10^{-5}$
A potential protective role for MMP12 in CHD?

- MMP12 levels have been observationally associated with **increased** risk of CHD events.
- However, our allele score for MMP12 suggests that the causal relationship is **inverse**.
- Variants associated with higher MMP12 levels have been shown to:
  - ↓ large artery atherosclerotic stroke
  - ↑ **MMP12** promoter activity and expression
  - ↓ fibrinogen levels
  - ↓ pulmonary function &
  - ↑ risk of COPD
- Genetic evidence points to **opposing** effects of MMP12 on atherosclerotic disease and COPD: a potential safety concern for MMP12-inhibitors.

---

Sun et al, Nature 2018
External replication in other studies using Olink

Integrative studies implicate matrix metalloproteinase-12 as a culprit gene for large-artery atherosclerotic stroke

H. Mahdessian1,*, L. Perisic Matic1,*, M. Lengquist2, K. Gertow1, B. Sennblad1, D. Baldassarre3, F. Veglia4, S. E. Humphries5, R. Rauramaa6, U. de Faire7,8, A. J. Smit9, P. Giral10,11, S. Kurl12, E. Mannarino13, E. Tremoli13,4, A. Hamsten1, P. Eriksson1, U. Hedin2,# & A. Mälarstig1,14,# IMPROVE study group

From the 1Cardiovascular Medicine, Department of Medicine Solna; 2Vascular Surgery, Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden; 3Department of Medical Biotechnology and Translational Medicine, Università degli Studi di Milano & Centro Cardiologico Monzino I.R.C.C.S.; 4Centro Cardiologico Monzino, IRCCS, Milan, Italy; 5Department of Medicine, British Heart Foundation Laboratories, University College of London, London, UK; 6Foundation for Research in Health Exercise and Nutrition, Kuopio Research Institute of Exercise Medicine, Kuopio, Finland; 7Division of Cardiovascular Epidemiology, Institute of Environmental Medicine, Karolinska Institutet; 8Department of Cardiology, Karolinska University Hospital, Solna, Stockholm, Sweden; 9Department of Medicine, University Medical Center Groningen, Groningen, The Netherlands; 10Assistance Publique-Hopitaux de Paris; 11Service Endocrinologie-Metabolisme, Unités de Prévention Cardiovaskulaire, Groupe Hôpitalier Pitie-Salpetriere, Paris, France; 12Institute of Public Health and Clinical Nutrition, University of Eastern Finland, Kuopio, Finland; 13Internal Medicine, Angiology and Arteriosclerosis Diseases, Department of Clinical and Experimental Medicine, University of Perugia, Perugia, Italy; and 14Pfizer Worldwide Research and Development, Stockholm, Sweden

Mahdessian et al, J Intern Med 2017
Multivariable MR to resolve complex GWAS loci

IL-1/IL-18 locus for eczema

Multivariable MR to tease out relevant proteins

Protein | Effect on AD risk per unit change in protein levels | P-Value
--- | --- | ---
IL18R1 | Lower levels decrease risk | 9.3e−72
| Multivariable MR | 1.5e−28
IL1RL1 | Multivariable MR | 5.7e−27
| Univariable MR | 0.01
| Multivariable MR | 0.013
IL1RL2 | Multivariable MR | 1.1e−69
| Univariable MR | 0.7
| Multivariable MR | 0.0094
IL1R2 | Multivariable MR | 0.49
| Univariable MR | 0.97
IL1R1 | Multivariable MR | 0.97
IL18RAP | Multivariable MR |

Sun et al, Nature 2018
**MST1** is the only signal associated with BLIMP1 levels.
Soluble proteins to identify causal pathways

- TNS2
- TMPRSS11D
- CRBB2
- DOCK9

Sun et al, Nature 2018
Protein target mis-specification

- Manhattan plot from GWAS of ALPL protein

Sun et al, Nature 2018
Binding assays have potential for distortion due to the effects of protein-altering variants

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

ATTGGGA\text{C}TTATCTC

ATTGGGA\text{T}TTATCTC

\begin{align*}
\text{Aptamer} & \quad \text{Protein} \\
\text{Aptamer} & \quad \text{Protein}
\end{align*}
‘Distortion’ or ‘detection’?!
Proteinase-3: implicated in vasculitis

Sun et al, Nature 2018
Wet-lab follow-up to assess affinity for free vs complexed PR3

Prefers PR3:A1AT complex
\(\rightarrow\) SERPINA1 trans pQTL

Prefers free PR3
\(\rightarrow\) No SERPINA1 trans pQTL

Sun et al, Nature 2018
pQTLs help to build causal models for disease

- Protective alleles at $PRTN3$ & $SERPINA1$
- Hepatocyte
- Blood
- PR3
- A1AT
- Antibody
- Neutrophil
- $PRTN3$

Sun et al, Nature 2018
Pleiotropic *trans* pQTLs may shed light on disease processes

Disease-associated missense variant (rs28929474 – ‘Z allele’) in *SERPINA1* is a *trans* pQTL hotspot

*Sun et al, Nature 2018*
Single locus allele scores created from multiple independent variants: IL-1Ra example

No. of IL-1Ra-raising ("C") alleles carried

[Graph showing log(IL-1Ra) levels in pg/mL (95% CI)]
Is having a causal therapeutic target enough? CETPi

CANTOS trial: 10,000 patients, 4 years of follow-up

REVEAL trial: 30,000 patients, 4 years of follow-up

“The modest absolute clinical benefit of canakinumab cannot justify its routine use …unless a price restructuring and formal cost-effectiveness evaluation supports it,” Dr. Robert Harrington, Stanford University
Going beyond qualitative causal conclusions: \( Lp(a)i \)

Allelic spectrum of 43 independent \( LPA \)-region variants

- **Per-allele change in \( Lp(a) \) [mg/dL]**

- **Absolute change in mean of lipoprotein(a) per allele (mg/dL, 95% CI)**

- **Odds ratio for coronary heart disease per allele (95% CI)**

- **Arithmetic mean of genetically-predicted lipoprotein(a) (mg/dL)**

- **Odds ratio for coronary heart disease (95% CI)**

- **Allele score of genetically-predicted \( Lp(a) \)**

- **Genetically-predicted \( Lp(a) \)**

Burgess et al, *JAMA Cardiol* 2018
Therapeutic target quantification

**LDL-c**
- Genetic estimate
- Observational estimate
- Trial estimate

**Lp(a)**
- Genetic estimate
- Observational estimate
- Trial estimate (predicted)

Burgess et al, *JAMA Cardiol* 2018
Lp(a) phase 3 trials demand stratified approaches

Lp(a) ASO lowers plasma Lp(a) by ~80-90%, but.....

only ~10% of Europeans have Lp(a)>10mg/dL

Burgess et al, *JAMA Cardiol* 2018
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:

- **Membrane**: 42%
- **Intracellular**: 35%
- **Secreted**: 23%
- **Cleavage**
- **Trafficking**
- **Transcription**

*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

Risk of coronary heart disease (per-allele log odds ratio)

Effect on plasma MMP-12 levels (per-allele change in SD)

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

DNA sequencing

- >15,000 genes
- ~1000 metabolites

RNA

- >3500 proteins
- 450 lipid species

Proteins

- 230 lipoproteins & lipids

Lipoproteins

- 230 lipoproteins & lipids

Blood cells

- 90 cell parameters

Disease

- e-health records

SomaLogic

- >3500 proteins

BRAINSHAKE

- 230 lipoproteins & lipids

Metabolon

- ~1000 metabolites
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.

Number of RBC loci discovered with n=50,000

<table>
<thead>
<tr>
<th>Uncorrected</th>
<th>Corrected</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>36</td>
</tr>
</tbody>
</table>

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

Asthma (positive control)
Rheumatoid arthritis

Odds ratio per 1-SD higher eosinophil count

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
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>~130 billion calculations</td>
<td>~1300 days</td>
</tr>
</tbody>
</table>

| Conventional computing    | ~1300 days       |
| High-performance computing*| ~4 days          |

* Cambridge University Peta 5 HPC
Current challenges of plasma proteomics

- Scalability
- Coverage
- Between assay differences
- Target mis-specification
- Cross-reactivity
- Epitope effects of protein-altering variants
- How does plasma relate to cell/tissue?
- Free vs complexed
Conclusions

• Robust ‘omics technologies, such as soluble proteomics, can greatly enhance our biological and aetiological understanding when combined with genetics at scale

• These results, which are all available for download (see www.phpc.cam.ac.uk/ceu/proteins) provide a toolkit for the community for 2-sample Mendelian randomisation studies of hundreds of pathways

• Scaling up measurements and pooling results from international consortia will enhance these findings in future, as will interrogating other matrices/technologies at scale
Future work in causal evaluation using proteomics

- Scale-up of protein GWAS in INTERVAL & through consortia
- Extensions to sequencing for rare variant analysis
- Validation using non-binding protein assays
- Assess specific therapeutic targets of relevance to pharma
- Develop pipelines for agnostic target prioritisation
- Improve approaches to MR to deal with complexity (eg, correlated variants, multivariable approaches etc)
Enhancing insights through international collaboration

**Olink**
- 350 proteins
- 1 study, 5k participants
- 15 studies, >~20k participants

**SomaLogic**
- 3500 proteins
- 1 study, 3k participants
- 12 studies, >20k participants

**Nightingale**
- 230 lipoproteins & lipids
- 1 study, 40k participants
- 15 studies, ~150k participants

**Metabolon**
- ~1000 metabolites
- 1 study, 9k participants
- 3 studies, ~20k participants

**Sysmex**
- 90 cell parameters
- 2 studies, 170k participants
- 24 studies, ~1M participants