STORY OF TWO MANUSCRIPTS

Ayse Demirkan

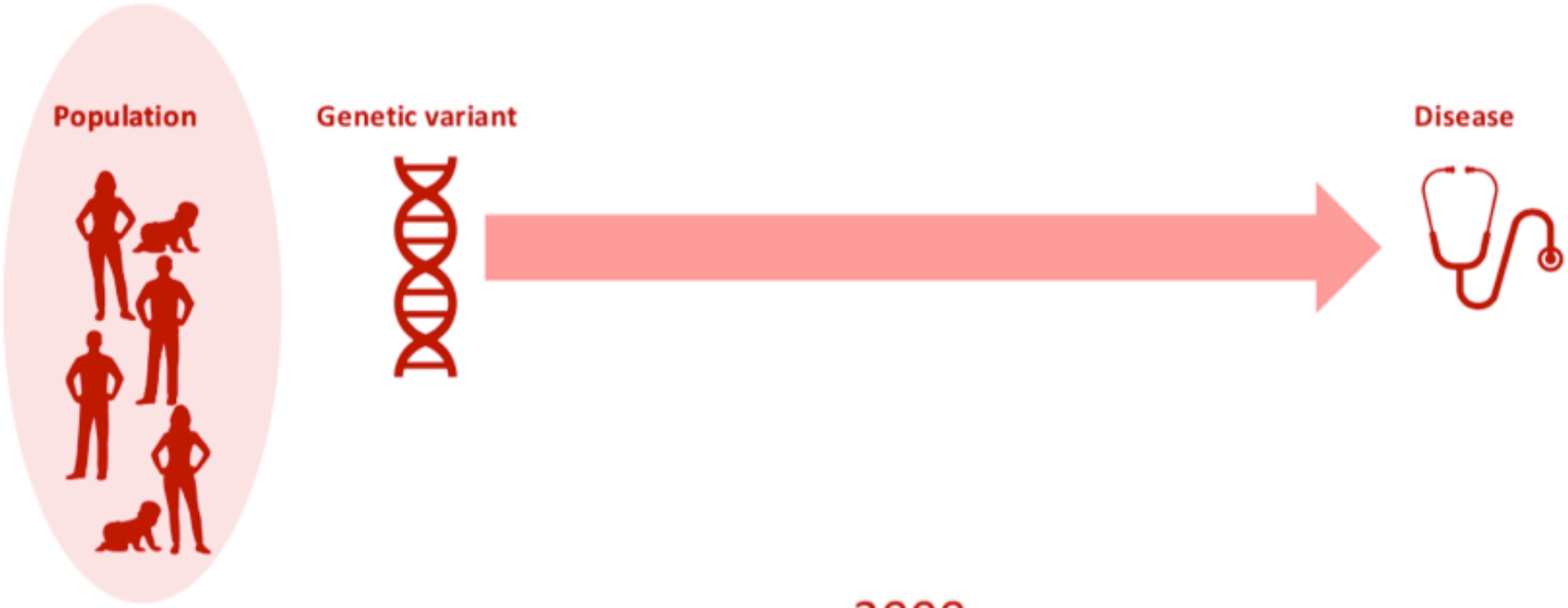
X-omics series

"Strategies to overcome your challenges in multi-omics data integration"

25 June 2020

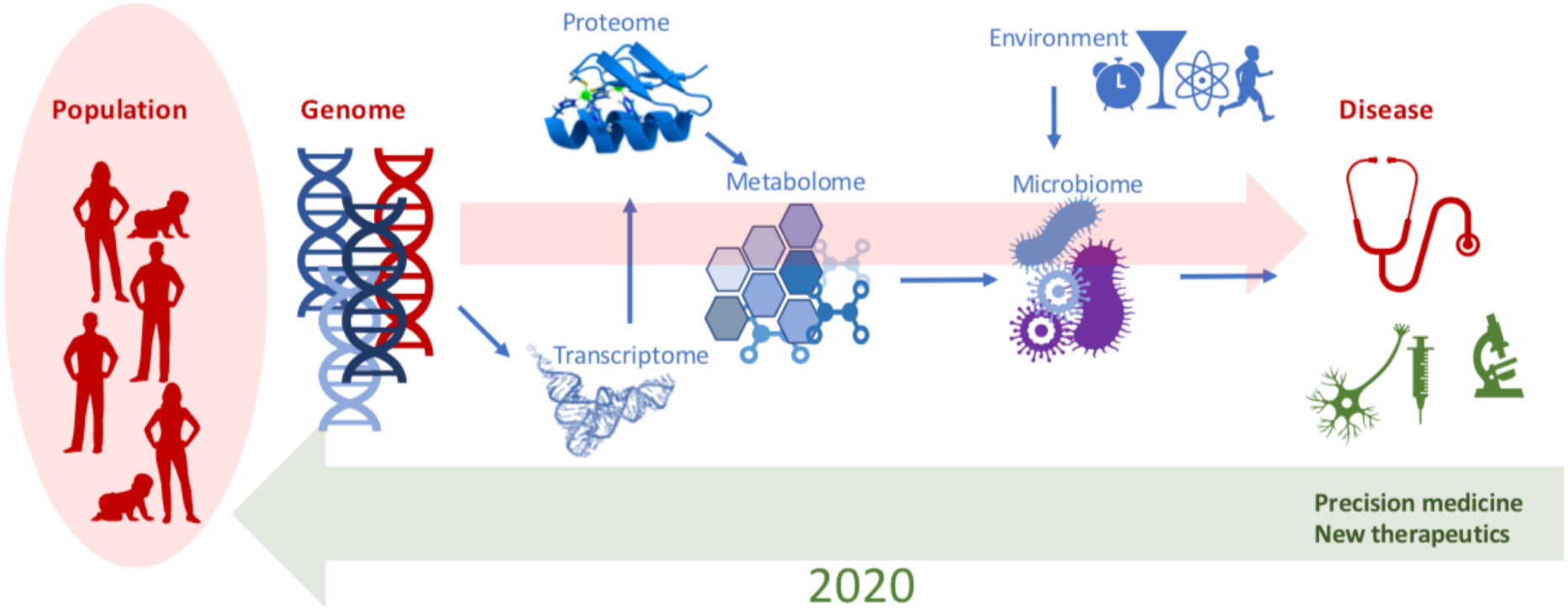
Department of Clinical & Experimental Medicine, School of Biosciences & Medicine, University of Surrey, Guildford, UK
Department of Genetics, University Medical Center Groningen, Groningen, the Netherlands
@demyrkana @StatMultiOmics

Genetic epidemiology

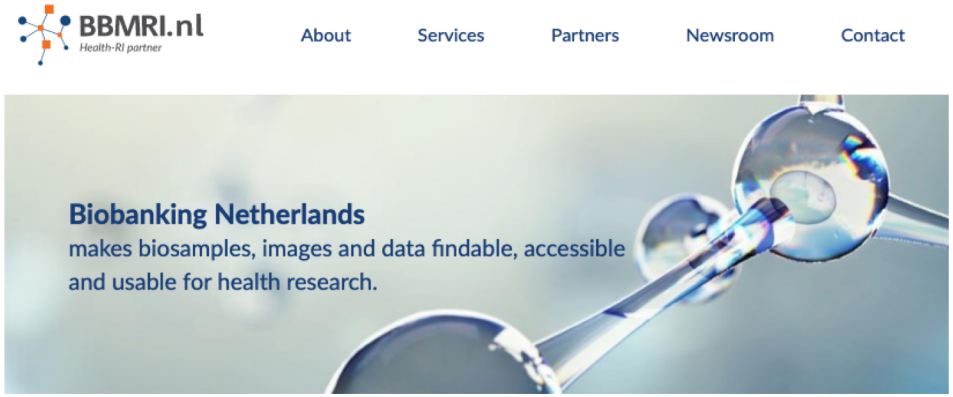


2000

contemporary genetic epidemiology

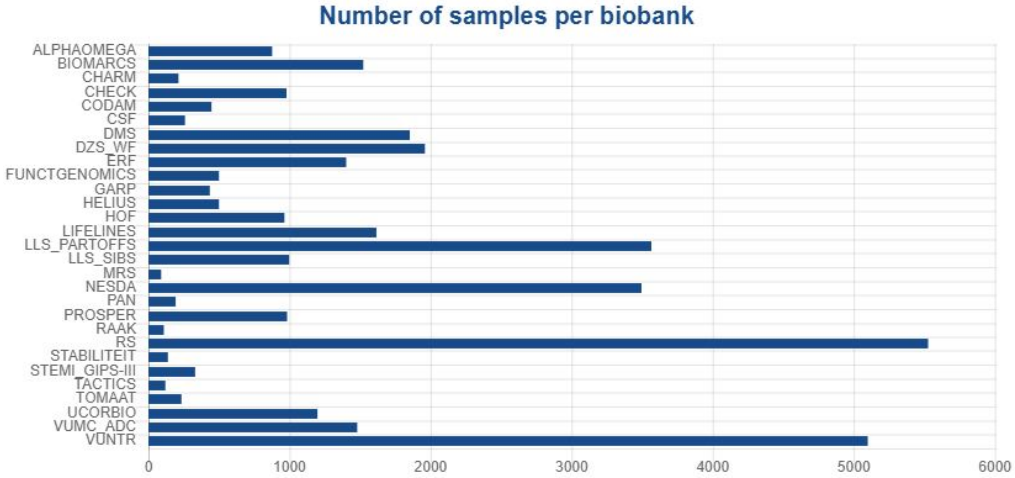


Biobanking for Biomedical Research Infrastructure



https://www.bbmri.nl/
https://www.bbmri.nl/samples-images-data

Integrated –omics dataset



Omics measure	Technique	Number of samples	
Whole Genome Sequencing	Illumina	750	GoNL
Imputed genotypes	HRC	~6000	
DNA Methylation	Illumina 450K	~4000	BIOS
Transcriptome	Illumina RNA seq	~4000	
Metabolome	Nightingale NMR	~32,000	

BBMRI CATALOGUE
https://catalogue.bbmri.nl/menu/main/app-molgenis-app-biobank-explorer/biobankexplorer

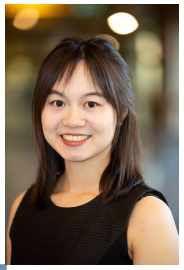
BBMRI-Omics

Collection types: Cohort, Cross-sectional, Population-based, Case-Control, Disease specific

Juridical person: BIOS Consortium

Collection	Type	Materials	Standards	#Samples
RNA sequencing (BIOS)	Cohort, Cross-sectional, Population-based	cDNA / mRNA, peripheral blood cells, Whole Blood		1000 - 10.000
BBMRI-Omics	Case-Control, Cohort, Cross-sectional, Disease specific	Not available		10.000 - 100.000
Genome of the Netherlands	Population-based	Not available		100 - 1000
Metabolomics	Case-Control, Cohort, Cross-sectional, Disease specific	Plasma		10.000 - 100.000
DNA Methylation (BIOS)	Cohort, Cross-sectional, Population-based	DNA, peripheral blood cells, Whole Blood		1000 - 10.000

integration



nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-10487-4 OPEN

An integrative cross-omics analysis of DNA methylation sites of glucose and insulin homeostasis

Jun Liu et al.#

Despite existing reports on differential DNA methylation in type 2 diabetes (T2D) and obesity, our understanding of its functional relevance remains limited. Here we show the effect of differential methylation in the early phases of T2D pathology by a blood-based epigenome-wide association study of 4808 non-diabetic Europeans in the discovery phase and 11,750 individuals in the replication. We identify CpGs in *LETM1, RBM20, IRS2, MAN2A2* and the 1q25.3 region associated with fasting insulin, and in *FCRL6, SLAMF1, APOBEC3H* and the 15q26.1 region with fasting glucose. In silico cross-omics analyses highlight the role of differential methylation in the crosstalk between the adaptive immune system and glucose homeostasis. The differential methylation explains at least 16.9% of the association between obesity and insulin. Our study sheds light on the biological interactions between genetic variants driving differential methylation and gene expression in the early pathogenesis of T2D.

ARTICLES | FOCUS
https://doi.org/10.1038/s41591-019-0722-x
nature medicine

Integration of epidemiologic, pharmacologic, genetic and gut microbiome data in a drug–metabolite atlas

Jun Liu[1,2*], Lies Lahousse[1,3], Michel G. Nivard[4,5], Mariska Bot[4,5], Lianmin Chen[6,7], Jan Bert van Klinken[8,9,10], Carisha S. Thesing[4,5], Marian Beekman[11], Erik Ben van den Akker[11,12,13], Roderick C. Slieker[5,14,15], Eveline Waterham[16], Carla J. H. van der Kallen[17,18], Irene de Boer[19], Ruifang Li-Gao[20], Dina Vojinovic[1], Najaf Amin[1], Djawad Radjabzadeh[21], Robert Kraaij[21], Louise J. M. Alferink[22], Sarwa Darwish Murad[22], André G. Uitterlinden[1,21], Gonneke Willemsen[4,5], Rene Pool[4,5], Yuri Milaneschi[4,5], Diana van Heemst[23], H. Eka D. Suchiman[11], Femke Rutters[5,14], Petra J. M. Elders[5,24], Joline W. J. Beulens[5,14], Amber A. W. A. van der Heijden[5,24], Marleen M. J. van Greevenbroek[17,18], Ilja C. W. Arts[18,25,26], Gerrit L. J. Onderwater[19], Arn M. J. M. van den Maagdenberg[8,19], Dennis O. Mook-Kanamori[20,27], Thomas Hankemeier[28,29], Gisela M. Terwindt[19], Coen D. A. Stehouwer[17,18], Johanna M. Geleijnse[16], Leen M. 't Hart[5,11,14,15], P. Eline Slagboom[11], Ko Willems van Dijk[8,9,30], Alexandra Zhernakova[6], Jingyuan Fu[6,7], Brenda W. J. H. Penninx[4,5], Dorret I. Boomsma[4,5], Ayşe Demirkan[1,6,31], Bruno H. C. Stricker[1,21,32] and Cornelia M. van Duijn[1,2,28*]

Progress in high-throughput metabolic profiling provides unprecedented opportunities to obtain insights into the effects of drugs on human metabolism. The Biobanking BioMolecular Research Infrastructure of the Netherlands has constructed an atlas of drug–metabolite associations for 87 commonly prescribed drugs and 150 clinically relevant plasma-based metabolites assessed by proton nuclear magnetic resonance. The atlas includes a meta-analysis of ten cohorts (18,873 persons) and uncovers 1,071 drug–metabolite associations after evaluation of confounders including co-treatment. We show that the effect estimates of statins on metabolites from the cross-sectional study are comparable to those from intervention and genetic observational studies. Further data integration links proton pump inhibitors to circulating metabolites, liver function, hepatic steatosis and the gut microbiome. Our atlas provides a tool for targeted experimental pharmaceutical research and clinical trials to improve drug efficacy, safety and repurposing. We provide a web-based resource for visualization of the atlas (http://bbmri.researchlumc.nl/atlas/).

Data *integration* in contemporary genetic epidemiology

"Bringing together different elements to make a whole unit"

Individual level data pooling → bringing together different populations/genotyping/ phenotyping

Analysis by correction of population structures of different ethnic background

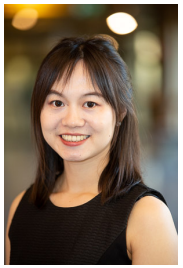
Summary level data pooling → bringing together different populations/genotyping/ phenotyping

Meta-analysis, post evaluation of heterogeneity

Integrating –omics data → different phenotypes, also different research questions

Currently no method to correct for heterogeneity and confounding exists, though, some tools may make life easier

Published in 2019, started in ? 2015



nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-10487-4 OPEN

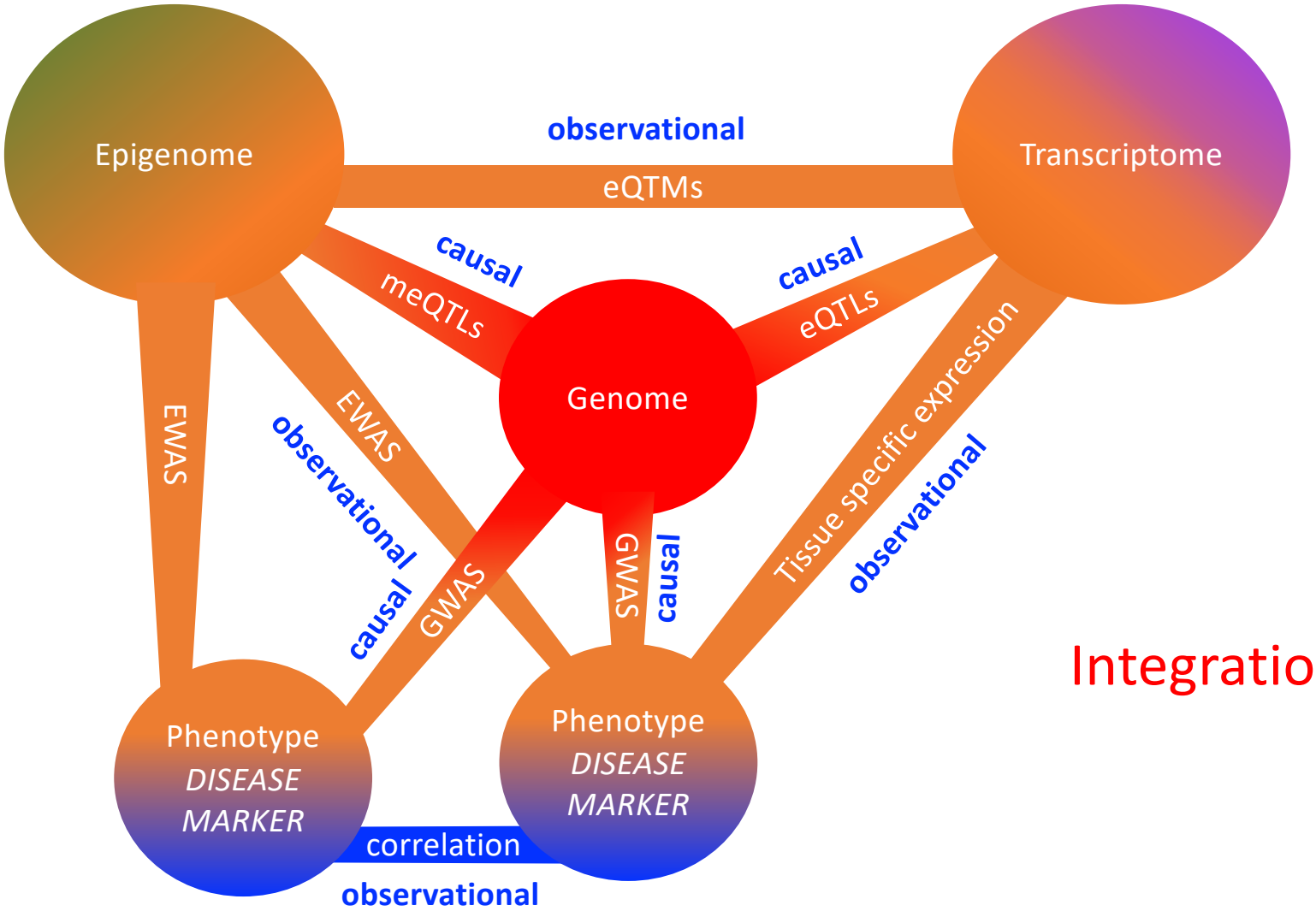
An integrative cross-omics analysis of DNA methylation sites of glucose and insulin homeostasis

Jun Liu et al.#

Despite existing reports on differential DNA methylation in type 2 diabetes (T2D) and obesity, our understanding of its functional relevance remains limited. Here we show the effect of differential methylation in the early phases of T2D pathology by a blood-based epigenome-wide association study of 4808 non-diabetic Europeans in the discovery phase and 11,750 individuals in the replication. We identify CpGs in *LETM1, RBM20, IRS2, MAN2A2* and the 1q25.3 region associated with fasting insulin, and in *FCRL6, SLAMF1, APOBEC3H* and the 15q26.1 region with fasting glucose. In silico cross-omics analyses highlight the role of differential methylation in the crosstalk between the adaptive immune system and glucose homeostasis. The differential methylation explains at least 16.9% of the association between obesity and insulin. Our study sheds light on the biological interactions between genetic variants driving differential methylation and gene expression in the early pathogenesis of T2D.

- Illumina 450 methylation array
- Fasting glucose, insulin, HOMA-IR
- Project initiated in 2015 by a visiting postdoc
- n=4808 sample size initially
- Submitted to Nature Communications in 2016, quick rejection
- Re-analysed in 2017:
  - Refined the statistical models
  - Added replication of n=11750
  - Added
  - Accepted with minor suggestions

Lots of data mining!

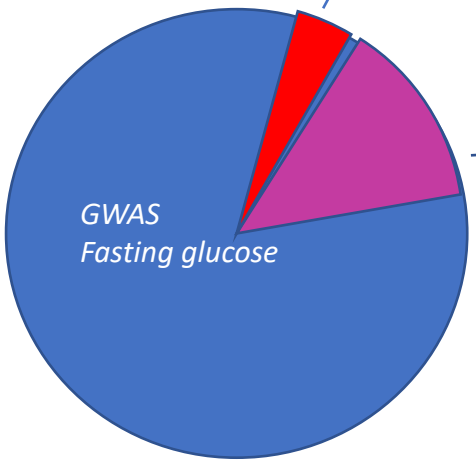


Integration space

Harvesting the maximum from a GWAS output

WG UTILIZING APPROACHES

- POLYGENIC RISK SCORES
- LD-SCORE BASED H2
- LD-SCORE BASED GENETIC CORRELATION
- H2 PARTITIONING (ENCODE)
- GReX / METAXcan



TOP SNP UTILIZING TOOLS

- PHEWAS/coloc
- Functional annotations
- Physical annotations
- MR

TOOLS USING SUGGESTIVE SNPs, REGIONS

- GENE SET / PATHWAY ENRICHMENT

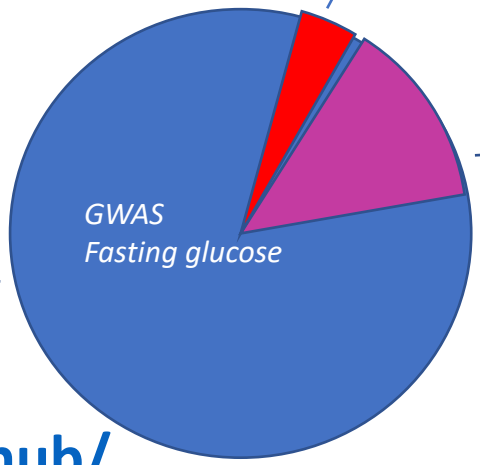
Harvesting the maximum from a GWAS output

WG UTILIZING APPROACHES

- POLYGENIC RISK SCORES
- **LD-SCORE BASED H2**
- **LD-SCORE BASED GENETIC CORRELATION**
- **H2 PARTITIONING (ENCODE)**
- GReX / METAXcan

TOP SNP UTILIZING TOOLS

- PHEWAS/coloc
- Functional annotations
- Physical annotations
- MR



TOOLS USING SUGGESTIVE SNPs, REGIONS

- GENE SET / PATHWAY ENRICHMENT

http://ldsc.broadinstitute.org/ldhub/

The intuition behind the approach is that if a trait is genetically influenced, then variants that tag more of the genome (i.e. have high LD scores) should have a greater opportunity to tag causal variants and therefore have higher test statistics on average than variants that have low LD scores. In this way genome-wide inflation of test statistics due to genuine polygenicity can be distinguished from biases such as population stratification and cryptic relatedness.

Harvesting the maximum from a GWAS output

WG UTILIZING APPROACHES

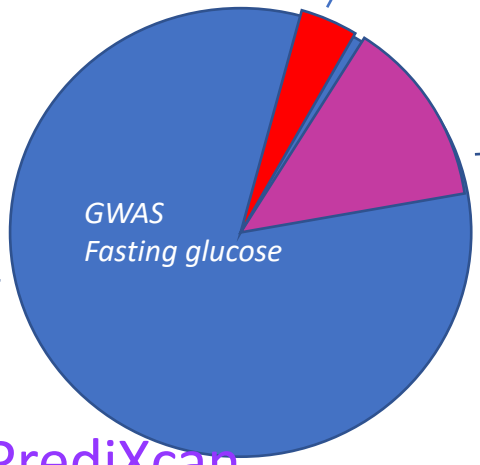
- POLYGENIC RISK SCORES
- LD-SCORE BASED H2
- LD-SCORE BASED GENETIC CORRELATION
- H2 PARTITIONING (ENCODE)
- **GReX / METAXcan**

TOP SNP UTILIZING TOOLS

- PHEWAS/coloc
- Functional annotations
- Physical annotations
- MR

TOOLS USING SUGGESTIVE SNPs, REGIONS

- GENE SET / PATHWAY ENRICHMENT

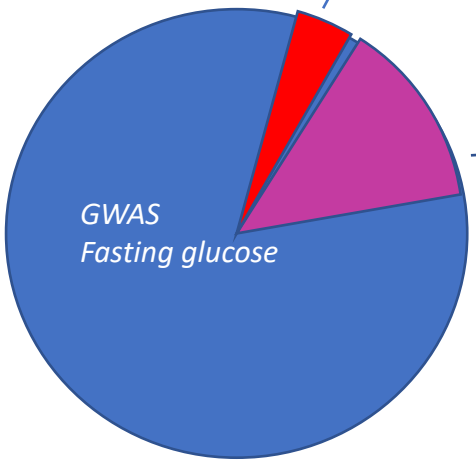


https://github.com/hakyimlab/PrediXcan
http://predictdb.org/
https://gtexportal.org/home/

Harvesting the maximum from a GWAS output

WG UTILIZING APPROACHES

- POLYGENIC RISK SCORES
- LD-SCORE BASED H2
- LD-SCORE BASED GENETIC CORRELATION
- H2 PARTITIONING (ENCODE)
- GReX / METAXcan



TOP SNP UTILIZING TOOLS

- **PHEWAS**/coloc
- **Functional annotations**
- Physical annotations
- MR

TOOLS USING SUGGESTIVE SNPs, REGIONS

- GENE SET / PATHWAY ENRICHMENT

https://atlas.ctglab.nl/
https://fuma.ctglab.nl/

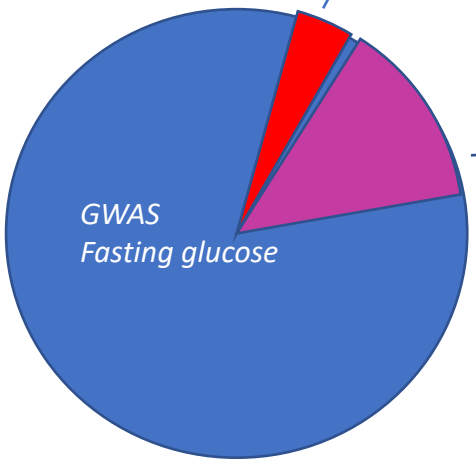
Harvesting the maximum from a GWAS output

WG UTILIZING APPROACHES

- POLYGENIC RISK SCORES
- LD-SCORE BASED H2
- LD-SCORE BASED GENETIC CORRELATION
- H2 PARTITIONING (ENCODE)
- GReX / METAXcan

TOP SNP UTILIZING TOOLS

- PHEWAS/coloc
- Functional annotations
- Physical annotations
- **MR**



TOOLS USING SUGGESTIVE SNPs, REGIONS

- GENE SET / PATHWAY ENRICHMENT

http://app.mrbase.org/

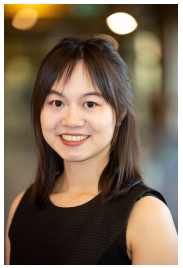
nature COMMUNICATIONS

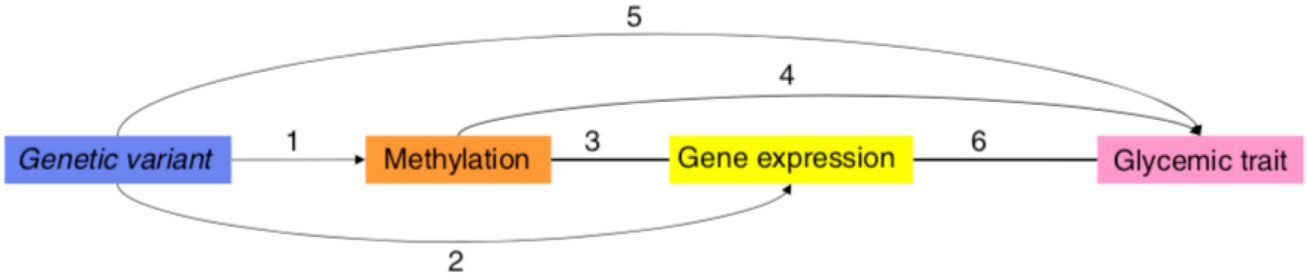
ARTICLE

https://doi.org/10.1038/s41467-019-10487-4 OPEN

An integrative cross-omics analysis of DNA methylation sites of glucose and insulin homeostasis

Jun Liu et al.





1 BIOS tables from here: https://genenetwork.nl/biosqtlbrowser/ file=Full list of primary cis-meQTLs FDR<0.05

2 BIOS tables from here: https://genenetwork.nl/biosqtlbrowser/ file=Cis-eQTLs Gene-level all primary effects, FDR<0.05

3 BIOS tables from here: https://genenetwork.nl/biosqtlbrowser/ file=Cis-eQTMs independent top effects, FDR<0.05

4 the actual tables from our EWAS here: https://figshare.com/s/1a1e8ac0fd9a49e2be30

5 Big N GWAS : https://www. magicinvestigators.org

6 In silico tissue specific expression prediction models by PrediXcan/GTeX: https://s3.amazonaws.com/imlab-open/Data/MetaXcan/results/ metaxcan_results_database_v0.1.tar.gz

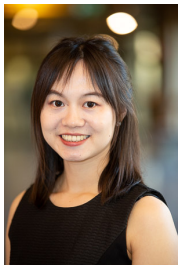
nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-10487-4 OPEN

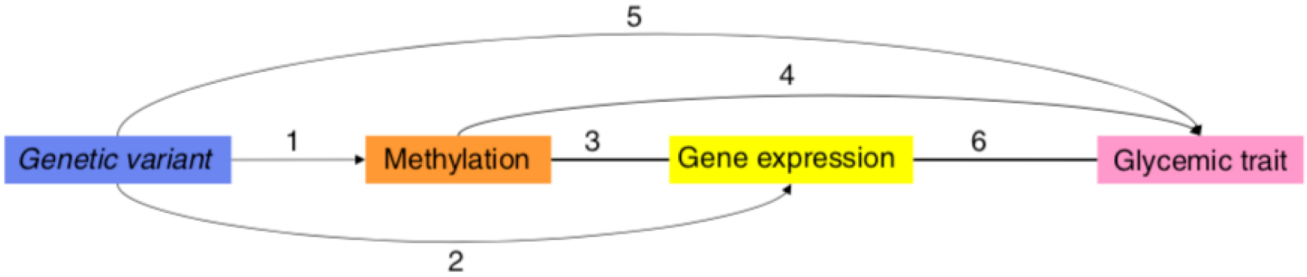
An integrative cross-omics analysis of DNA methylation sites of glucose and insulin homeostasis

Jun Liu et al.



Sample size therefore POWER is not comparable

5 >>> 4 > 1,2,3 > 6



1 BIOS tables from here: https://genenetwork.nl/biosqtlbrowser/ file=Full list of primary cis-meQTLs FDR<0.05

2 BIOS tables from here: https://genenetwork.nl/biosqtlbrowser/ file=Cis-eQTLs Gene-level all primary effects, FDR<0.05 ← Not clear if the SNPs are not tested or not significant

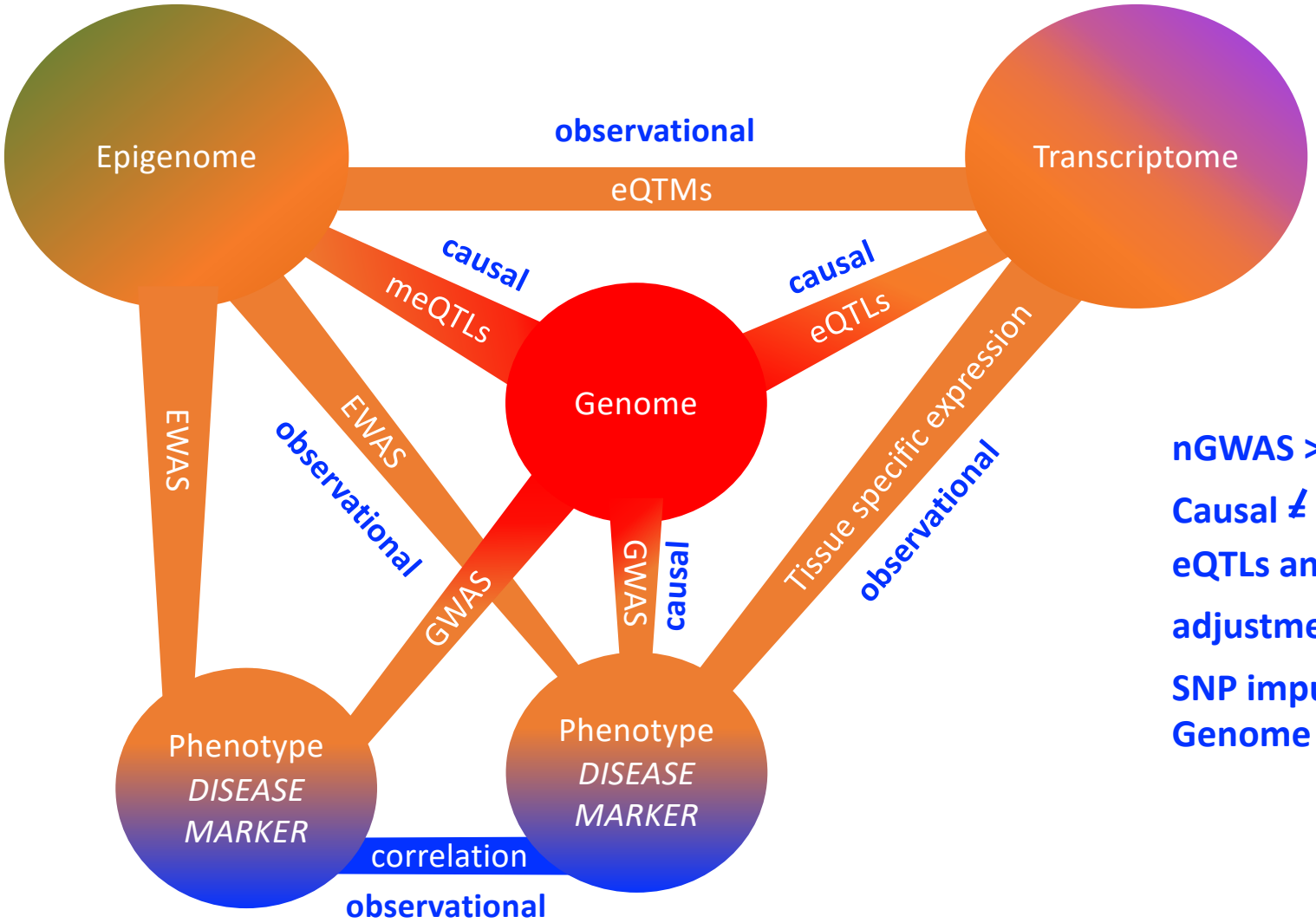
3 BIOS tables from here: https://genenetwork.nl/biosqtlbrowser/ file=Cis-eQTMs independent top effects, FDR<0.05

4 the actual tables from our EWAS here: https://figshare.com/s/1a1e8ac0fd9a49e2be30

5 Big N GWAS : https://www. magicinvestigators.org <-- overlap between SNP sets <100%

6 In silico tissue specific expression prediction models by PrediXcan/GTeX: https://s3.amazonaws.com/imlab-open/Data/MetaXcan/results/ metaxcan_results_database_v0.1.tar.gz <-- overlap between SNP sets <100%

Integration space



nGWAS >> nOTHERS

Causal ≠ direct

eQTLs and meQTLs are tissue specific

adjustments differ

SNP imputation

Genome build

Data *integration* in contemporary genetic epidemiology

"Bringing together different elements to make a whole unit"

Integrating –omics data → different phenotypes, also different research questions

Currently no method to correct for heterogeneity and confounding exists, though, some tools may make life easier

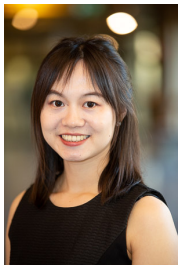
Summary level data pooling → bringing together different populations/genotyping/ phenotyping

Meta-analysis, post evaluation of heterogeneity

Individual level data pooling → bringing together different populations/genotyping/ phenotyping

Analysis by correction of population structures of different ethnic background

The BBMRI drug metabolite study



18873 persons from 10 BBMRI-NL cohorts

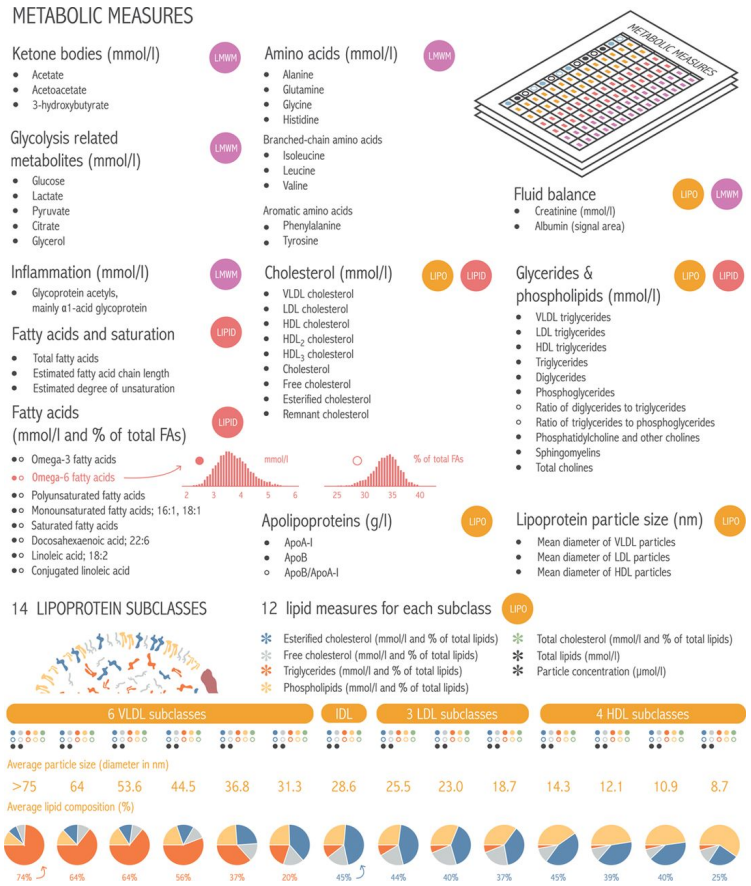
89 drug categories, 150 metabolites

ARTICLES | FOCUS
https://doi.org/10.1038/s41591-019-0722-x
nature medicine

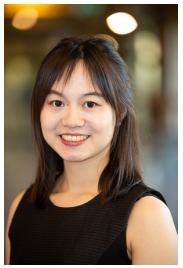
Integration of epidemiologic, pharmacologic, genetic and gut microbiome data in a drug–metabolite atlas

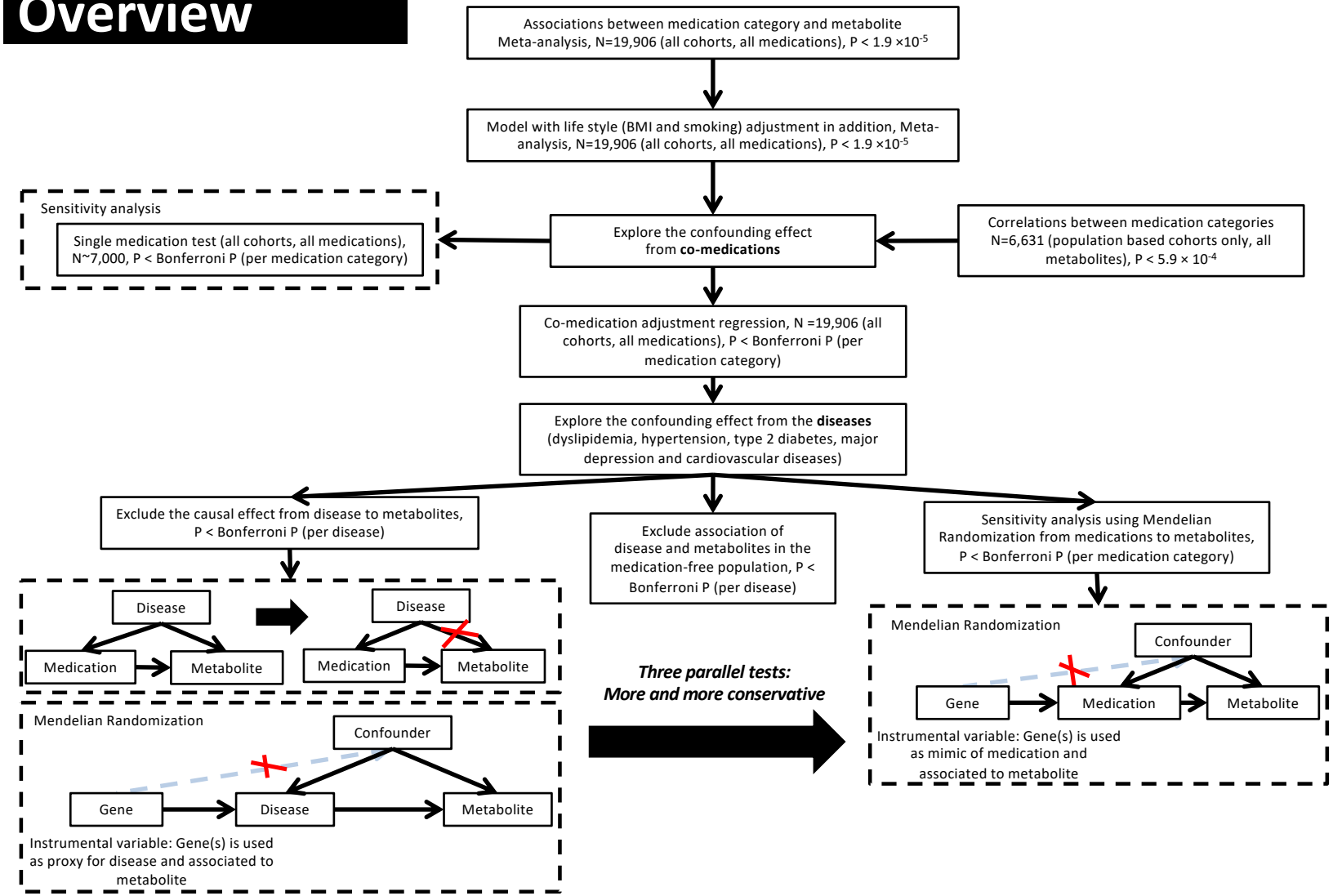
Jun Liu[1,2*], Lies Lahousse[1,3], Michel G. Nivard[4,5], Mariska Bot[4,5], Lianmin Chen[6,7], Jan Bert van Klinken[8,9,10], Carisha S. Thesing[4,5], Marian Beekman[11], Erik Ben van den Akker[11,12,13], Roderick C. Slieker[5,14,15], Eveline Waterham[16], Carla J. H. van der Kallen[17,18], Irene de Boer[19], Ruifang Li-Gao[20], Dina Vojinovic[1], Najaf Amin[1], Djawad Radjabzadeh[21], Robert Kraaij[21], Louise J. M. Alferink[22], Sarwa Darwish Murad[22], André G. Uitterlinden[1,21], Gonneke Willemsen[4,5], Rene Pool[4,5], Yuri Milaneschi[4,5], Diana van Heemst[23], H. Eka D. Suchiman[11], Femke Rutters[5,14], Petra J. M. Elders[5,24], Joline W. J. Beulens[5,14], Amber A. W. A. van der Heijden[5,24], Marleen M. J. van Greevenbroek[17,18], Ilja C. W. Arts[18,25,26], Gerrit L. J. Onderwater[19], Arn M. J. M. van den Maagdenberg[8,19], Dennis O. Mook-Kanamori[20,27], Thomas Hankemeier[28,29], Gisela M. Terwindt[19], Coen D. A. Stehouwer[17,18], Johanna M. Geleijnse[16], Leen M. 't Hart[5,11,14,15], P. Eline Slagboom[11], Ko Willems van Dijk[8,9,30], Alexandra Zhernakova[6], Jingyuan Fu[6,7], Brenda W. J. H. Penninx[4,5], Dorret I. Boomsma[4,5], Ayşe Demirkan[1,6,31], Bruno H. C. Stricker[1,21,32] and Cornelia M. van Duijn[1,2,28*]

Progress in high-throughput metabolic profiling provides unprecedented opportunities to obtain insights into the effects of drugs on human metabolism. The Biobanking BioMolecular Research Infrastructure of the Netherlands has constructed an atlas of drug–metabolite associations for 87 commonly prescribed drugs and 150 clinically relevant plasma-based metabolites assessed by proton nuclear magnetic resonance. The atlas includes a meta-analysis of ten cohorts (18,873 persons) and uncovers 1,071 drug–metabolite associations after evaluation of confounders including co-treatment. We show that the effect estimates of statins on metabolites from the cross-sectional study are comparable to those from intervention and genetic observational studies. Further data integration links proton pump inhibitors to circulating metabolites, liver function, hepatic steatosis and the gut microbiome. Our atlas provides a tool for targeted experimental pharmaceutical research and clinical trials to improve drug efficacy, safety and repurposing. We provide a web-based resource for visualization of the atlas (http://bbmri.researchlumc.nl/atlas/).



Overview




Associations between medication category and metabolite
Meta-analysis, N=19,906 (all cohorts, all medications), P < 1.9 ×10-5
Model with life style (BMI and smoking) adjustment in addition, Meta-analysis, N=19,906 (all cohorts, all medications), P < 1.9 ×10-5
Sensitivity analysis
Single medication test (all cohorts, all medications), N~7,000, P < Bonferroni P (per medication category)
Explore the confounding effect from co-medications
Correlations between medication categories N=6,631 (population based cohorts only, all metabolites), P < 5.9 × 10-4
Co-medication adjustment regression, N =19,906 (all cohorts, all medications), P < Bonferroni P (per medication category)
Explore the confounding effect from the diseases (dyslipidemia, hypertension, type 2 diabetes, major depression and cardiovascular diseases)
Exclude the causal effect from disease to metabolites, P < Bonferroni P (per disease)
Exclude association of disease and metabolites in the medication-free population, P < Bonferroni P (per disease)
Sensitivity analysis using Mendelian Randomization from medications to metabolites, P < Bonferroni P (per medication category)
Disease
Medication
Metabolite
Disease
Medication
Metabolite
Mendelian Randomization
Confounder
Gene
Disease
Metabolite
Instrumental variable: Gene(s) is used as proxy for disease and associated to metabolite
Three parallel tests:
More and more conservative
Mendelian Randomization
Confounder
Gene
Medication
Metabolite
Instrumental variable: Gene(s) is used as mimic of medication and associated to metabolite

Overview

Baseline model and lifestyle adjusted model

Associations between medication category and metabolite Meta-analysis, N=19,906 (all cohorts, all medications), $P < 1.9 \times 10^{-5}$

2,087 significant associations

Model with life style (BMI and smoking) adjustment in addition, Meta-analysis, N=19,906 (all cohorts, all medications), $P < 1.9 \times 10^{-5}$

1,640 significant associations

Sensitivity analysis

Single medication test (all cohorts, all medications), N~7,000, P < Bonferroni P (per medication category)

Explore the confounding effect from **co-medications**

Correlations between medication categories N=6,631 (population based cohorts only, all metabolites), $P < 5.9 \times 10^{-4}$

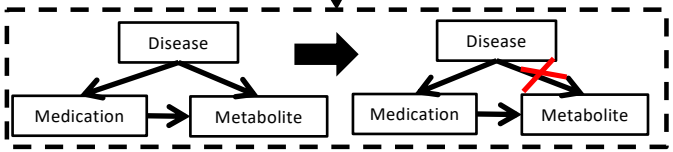
Co-medication adjustment regression, N =19,906 (all cohorts, all medications), P < Bonferroni P (per medication category)

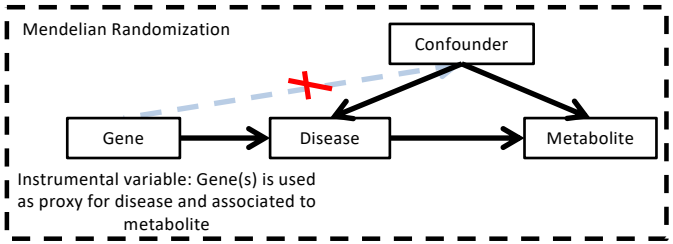
Explore the confounding effect from the **diseases** (dyslipidemia, hypertension, type 2 diabetes, major depression and cardiovascular diseases)

Exclude the causal effect from disease to metabolites, P < Bonferroni P (per disease)

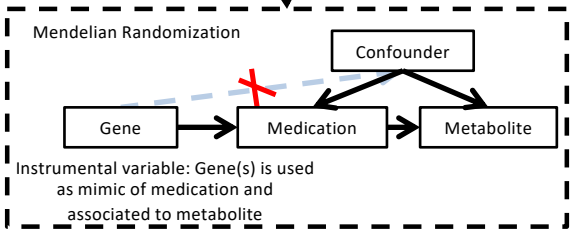
Exclude association of disease and metabolites in the medication-free population, P < Bonferroni P (per disease)

Sensitivity analysis using Mendelian Randomization from medications to metabolites, P < Bonferroni P (per medication category)

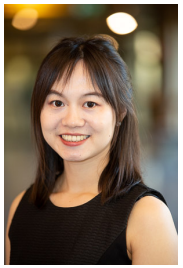




Three parallel tests: More and more conservative



Overview



Lifestyle adjusted model and co-medication adjusted model

Correlation matrix: Rotterdam Study and LifeLines DEEP, *n* = 6,631) and meta-analyzed.

Associations between medication category and metabolite Meta-analysis, N=19,906 (all cohorts, all medications), P < 1.9 ×10^{-5}

Model with life style (BMI and smoking) adjustment in addition, Meta-analysis, N=19,906 (all cohorts, all medications), P < 1.9 ×10^{-5}

Sensitivity analysis

Single medication test (all cohorts, all medications), N~7,000, P < Bonferroni P (per medication category)

Explore the confounding effect from **co-medications**

Correlations between medication categories N=6,631 (general based cohorts only, all metabolites), P < 5.9 × 10^{-4}

Co-medication adjustment regression, N =19,906 (all cohorts, all medications), P < Bonferroni P (per medication category)

Explore the confounding effect from the **diseases** (dyslipidemia, hypertension, type 2 diabetes, major depression and cardiovascular diseases)

1,071 significant associations

Exclude the causal effect from disease to metabolites, P < Bonferroni P (per disease)

Exclude association of disease and metabolites in the medication-free population, P < Bonferroni P (per disease)

Sensitivity analysis using Mendelian Randomization from medications to metabolites, P < Bonferroni P (per medication category)

Disease

Medication

Metabolite

Disease

Medication

Metabolite

Mendelian Randomization

Confounder

Gene

Disease

Metabolite

Instrumental variable: Gene(s) is used as proxy for disease and associated to metabolite

Three parallel tests:
More and more conservative

Mendelian Randomization

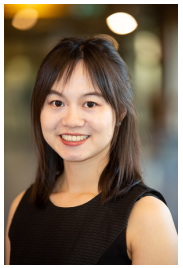
Confounder

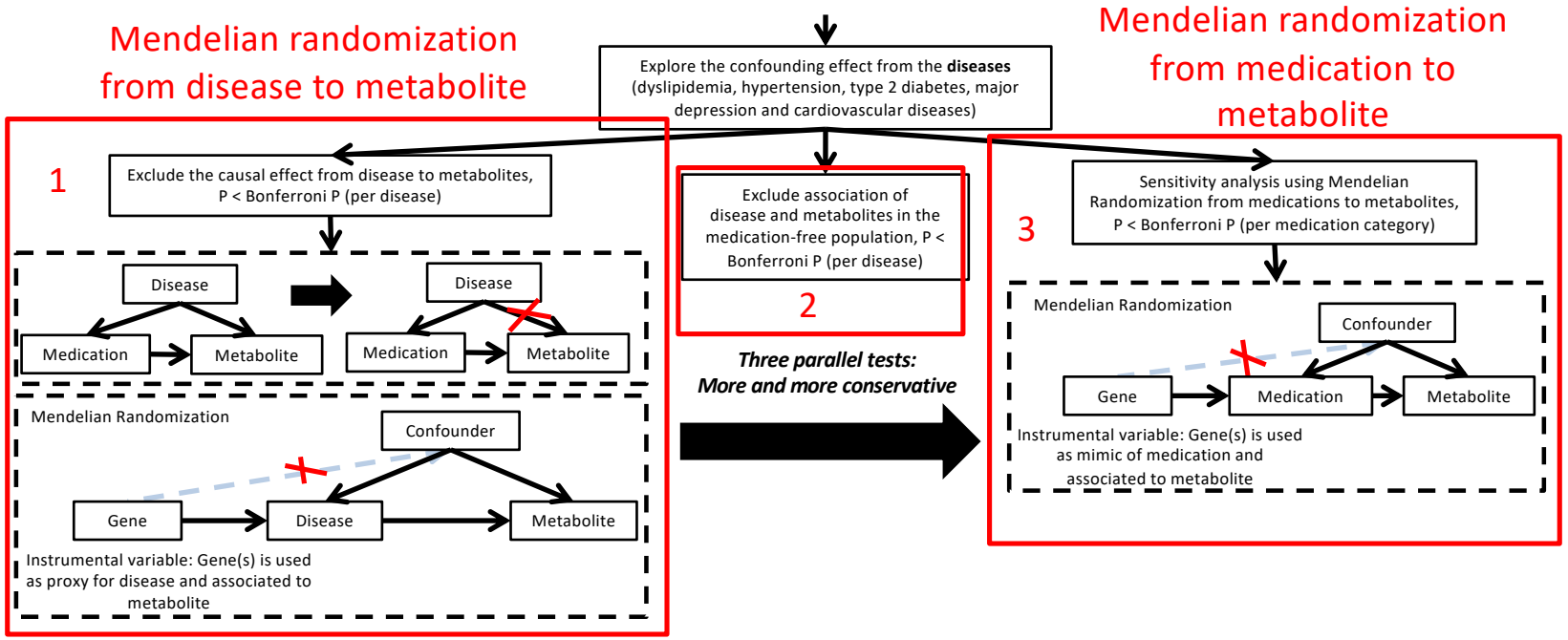
Gene

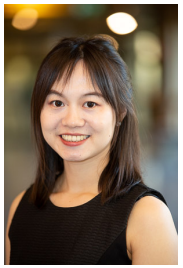
Medication

Metabolite

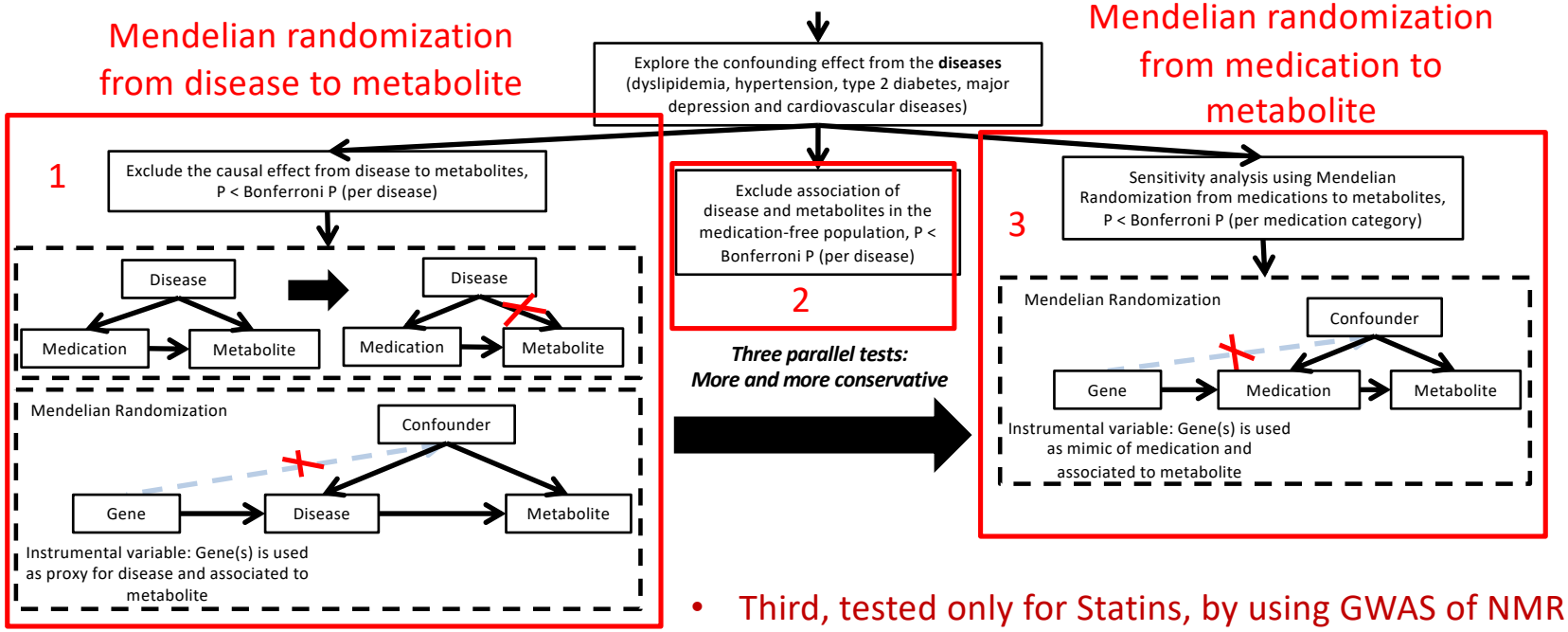
Instrumental variable: Gene(s) is used as mimic of medication and associated to metabolite




Mendelian randomization from disease to metabolite
Explore the confounding effect from the diseases (dyslipidemia, hypertension, type 2 diabetes, major depression and cardiovascular diseases)
Mendelian randomization from medication to metabolite
1
Exclude the causal effect from disease to metabolites, P < Bonferroni P (per disease)
Disease
Medication
Metabolite
Disease
Medication
Metabolite
Mendelian Randomization
Confounder
Gene
Disease
Metabolite
Instrumental variable: Gene(s) is used as proxy for disease and associated to metabolite
Exclude association of disease and metabolites in the medication-free population, P < Bonferroni P (per disease)
2
Three parallel tests:
More and more conservative
3
Sensitivity analysis using Mendelian Randomization from medications to metabolites, P < Bonferroni P (per medication category)
Mendelian Randomization
Confounder
Gene
Medication
Metabolite
Instrumental variable: Gene(s) is used as mimic of medication and associated to metabolite

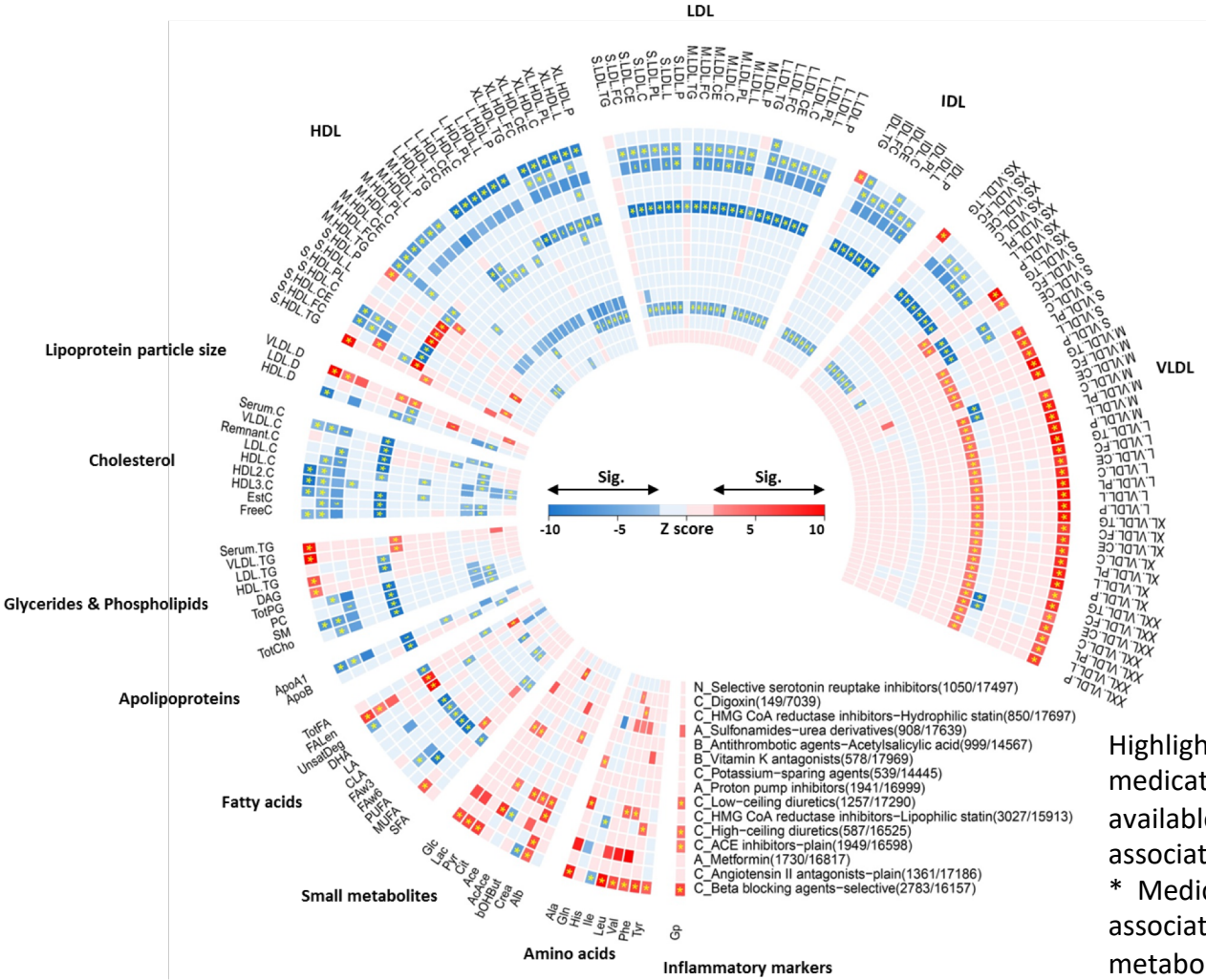


- First, we tested whether indicated diseases causally related to drug-related metabolites using the genetic risk score of the disease as an instrumental variable in Mendelian randomization (MR) →HT, T2D , MD
- Second, we associated drug-related metabolites with the indicated disease in individuals who were not receiving treatment → HT, T2D , MD, Dyslipidemia
  - T2D analyses were performed based on Rotterdam Study and NEO
  - Dyslipidemia and HT were tested in ERF and Rotterdam Study
  - MD were tested in a parallel BBMRI paper by Mariska Bot



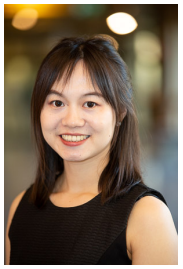
- Third, tested only for Statins, by using GWAS of NMR metabolites

Top 15 medications effecting Nightingale metabolites that are NOT explained by co-medication or disease



Highlight red/blue: Sig. in co-medication adjusted model and available to do disease-metabolite association analysis

* Medication-metabolite association not explain by disease-metabolite association



Challenges

- Organisation
- Dealing with heterogeneity among the cohorts,
- Data cleaning
- Getting the data from the cohorts
- Dealing with the cohorts which has no analysists with R/computing experience
- Too many significant findings, how to find the highlights?
- How to interpret the results
- Varying power in each analysis

ANGLIA - An online platform for GWAS/EWAS summary statistics and imputation

Ayse Demirkan[1,2], Konstantin Rudometkin[1], Liudmila Zudina[1], Zhanna Balkhiyarova[1,3], Marika Kaakinen[1,3], Inga Prokopenko[1]

1 Department of Clinical & Experimental Medicine, School of Biosciences & Medicine, University of Surrey, Guildford, UK

2 Department of Genetics, University Medical Center Groningen, Groningen, the Netherlands

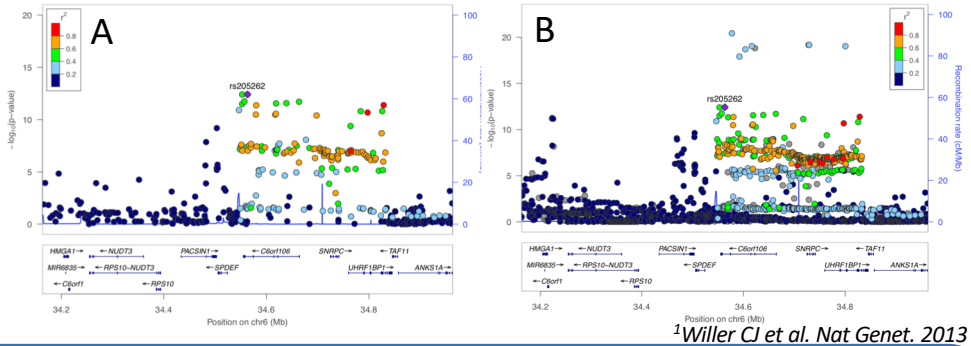
3 Department of Genomics of Common Disease, Imperial College London, London, UK

@demyrkana @StatMultiOmics

Background

- Genotype imputation is a necessity for data harmonisation to perform Genome-Wide Associations Study (GWAS) meta-analyses, replication and Phenome –wide association studies.
- Less than 30% of summary statistics are imputed to the 1000 Genomes/Haplotype Reference Consortium variant density, with the majority being imputed to the HapMap reference panel.
- Genotype imputation at summary statistics level can be used to upcycle out-of-date GWAS data and can increase power and resolution of the association signals.

Example Locus zoom plots below show information gained by summary statistics imputation, for the HDL-C locus on chr 6[1] . A is based on HapMap density, while B is based on SS-imp imputation to 1000 Genomes density.



[1]*Willer CJ et al. Nat Genet. 2013*

AIM: to develop an online platform for data harmonisation, including imputation of summary statistic level data

ANGLIA: An online platform for imputing And aNalyzinG epidemioLogical -omIcs dAta

UNIVERSITY OF SURREY

ANGLIA - An online platform for GWAS/EWAS summary statistics and imputation

Tools for upcycling the data

- **Summary statistics imputation SS-imp**[1,2] adds additional lines to the GWAS data table for the additional SNPs found in the given reference panel
- **LiftOver**[3] updates the positional mapping of the SNPs according to the new human genome build

Tools for running multi-phenotype analysis

- **SCOPA**[5] performs multi-phenotype GWAS using individual level genetic data
- **MARV**[6,7] performs multi-phenotype GWAS for rare variants using individual level genetic data
- **methylSCOPA**[8] performs multi-phenotype epigenome wide association study using individual level genetic data

Tools for generating new data

- **Genome-Wide Inferred Statistics (GWIS)** [4] calculates association statistics for composite measurements, from existing summary statistic data
- **sumSCOPA (being prepared)** performs multi-phenotype GWAS from existing summary statistics data

Tool for multi-phenotype imputation

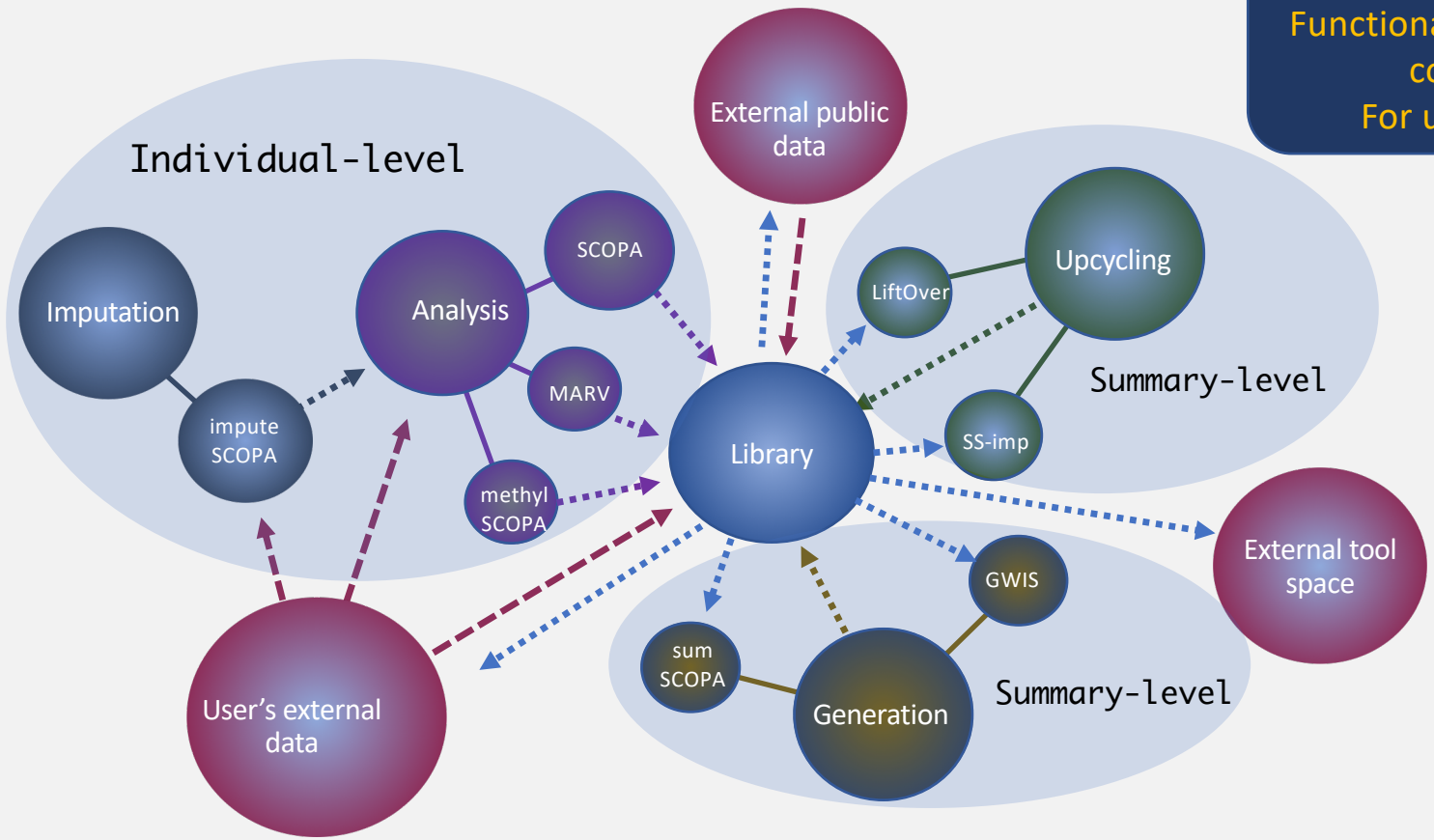
- **ImputeSCOPA** (being prepared) performs multi-phenotype imputations using the random forest method

References [1]Rueger *et al.* PLOS Genetics 2018 **Evaluation and Application of Summary Statistic Imputation to Discover New Height-Associated Loci** [2]Pasaniuc *et al.* Bioinformatics 2014 **Fast and Accurate Imputation of Summary Statistics Enhances Evidence of Functional Enrichment** [3]Kuhn *et al.* Brief Bioinform. 2013 **The UCSC Genome Browser and Associated Tools** [4]Nieuwboer *et al.* AJHG 2016 **GWIS: Genome-Wide Inferred Statistics for Functions of Multiple Phenotypes** [5]Magi *et al.* BMC Bioinformatics 2017 **SCOPA and META-SCOPA: Software for the Analysis and Aggregation of Genome-Wide Association Studies of Multiple Correlated Phenotypes**. [6]Kaakinen *at al.* BMC Bioinformatics 2017 **MARV: A Tool for Genome-Wide Multi-Phenotype Analysis of Rare Variants** . [7]Kaakinen *et al.* EJHG 2017 **A Rare-Variant Test for High-Dimensional Data**. [8]Draisma e*t al.* bioRxiv 2019 **methylSCOPA and META-methylSCOPA: software for the analysis and aggregation of epigenome-wide association studies of multiple correlated phenotypes.**

ANGLIA

Home | Library | Impute | Analyze | Generate | Upcycle | My jobs | Upload your data

Demo website
http://anglia.prokopenkogroup.org/
Functional website is under construction
For updates follow

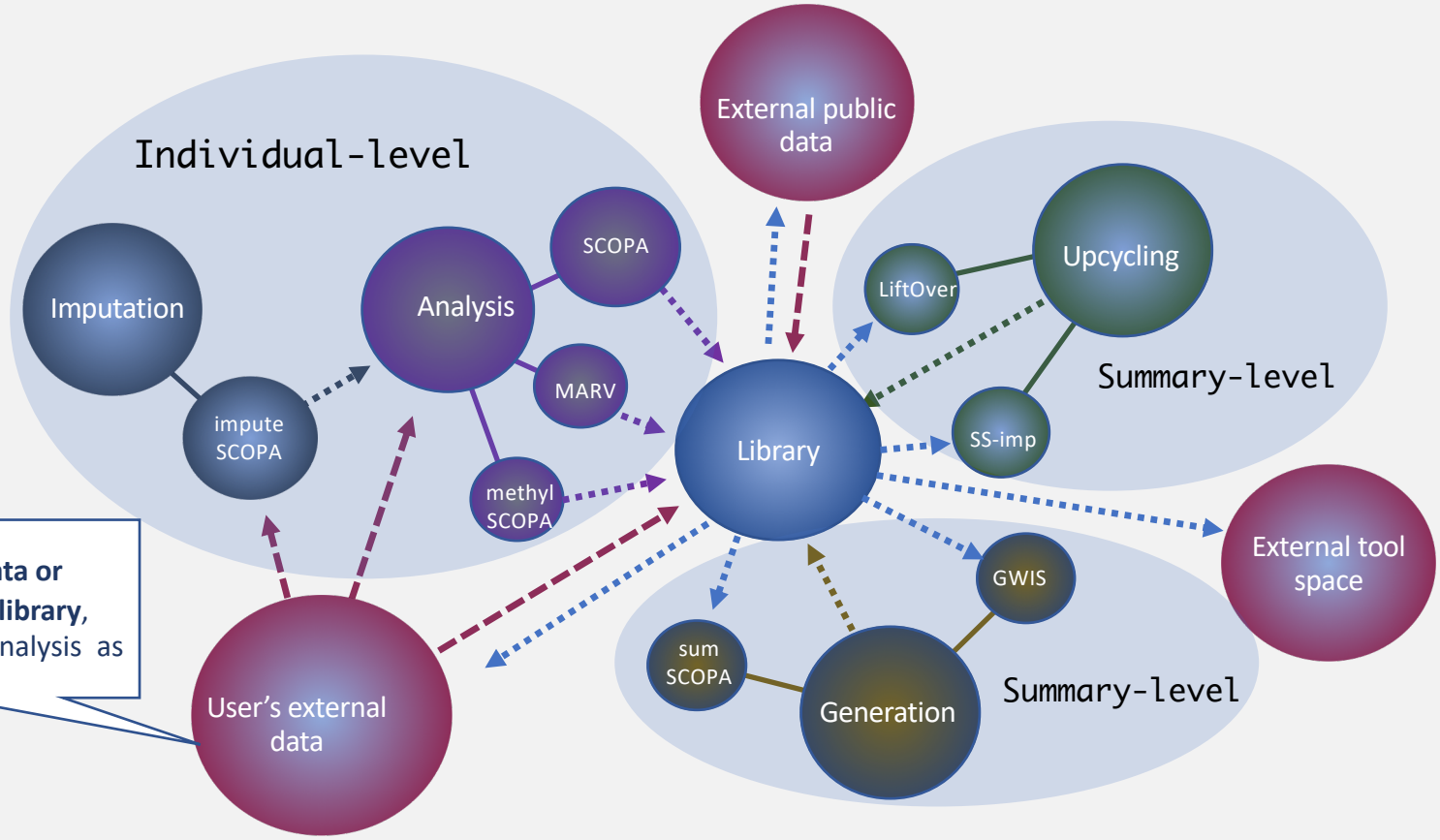


Contact:
anglia@prokopenkogroup.org

Prokopenko group
Section of Statistical multi-omics
University of Surrey

ANGLIA

Home | Library | Impute | Analyze | Generate | Upcycle | My jobs | Upload your data



The user can
- **upload their own data or**
- **download from the library,**
- and use them in analysis as desired.

Contact:
anglia@prokopenkogroup.org

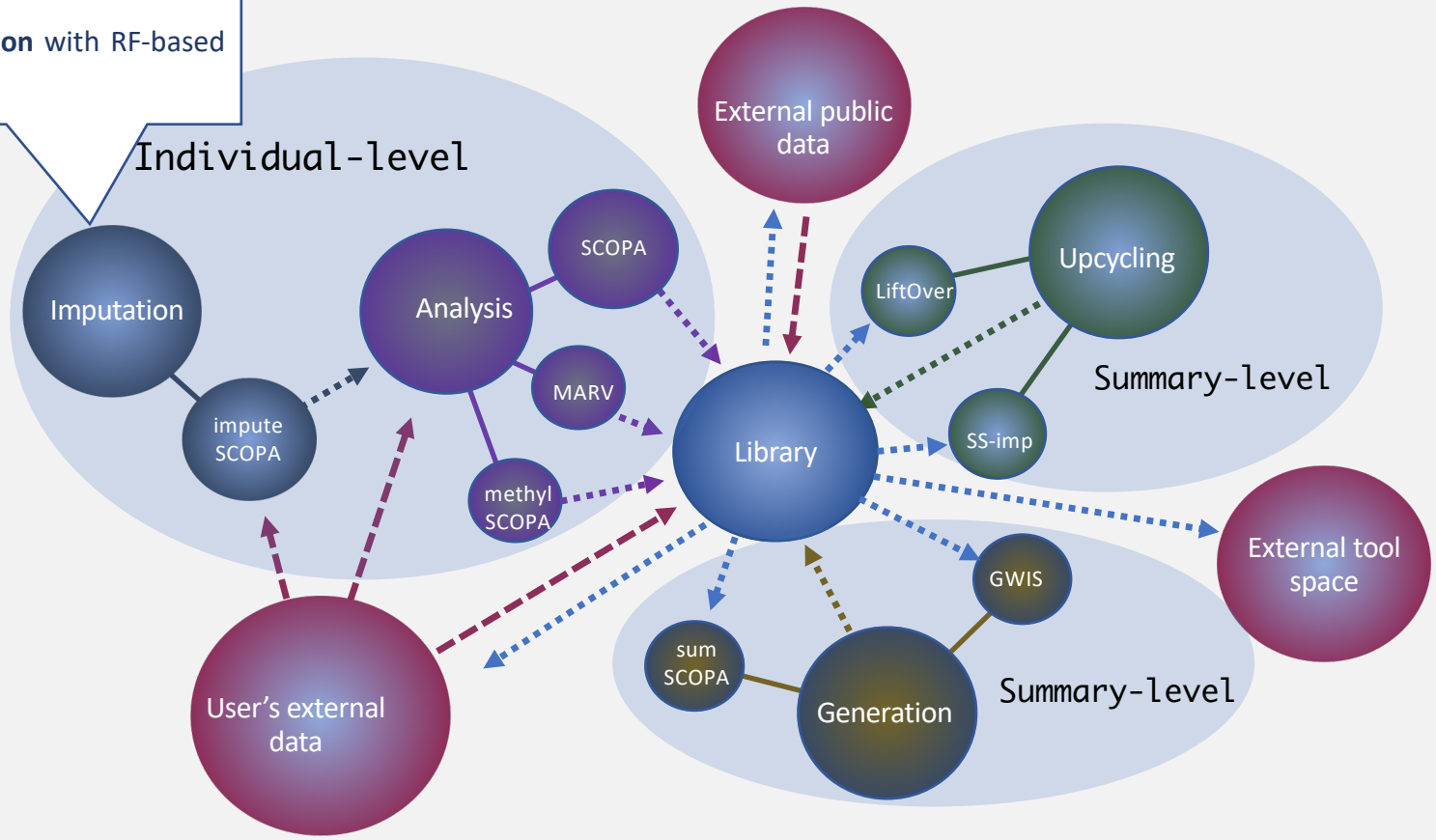
Prokopenko group
Section of Statistical multi-omics
University of Surrey

ANGLIA

Home | Library | Impute | Analyze | Generate | Upcycle | My jobs | Upload your data

In the **IMPUTATION module**, the user can perform

- **phenotype imputation** with RF-based **imputeSCOPA tool**



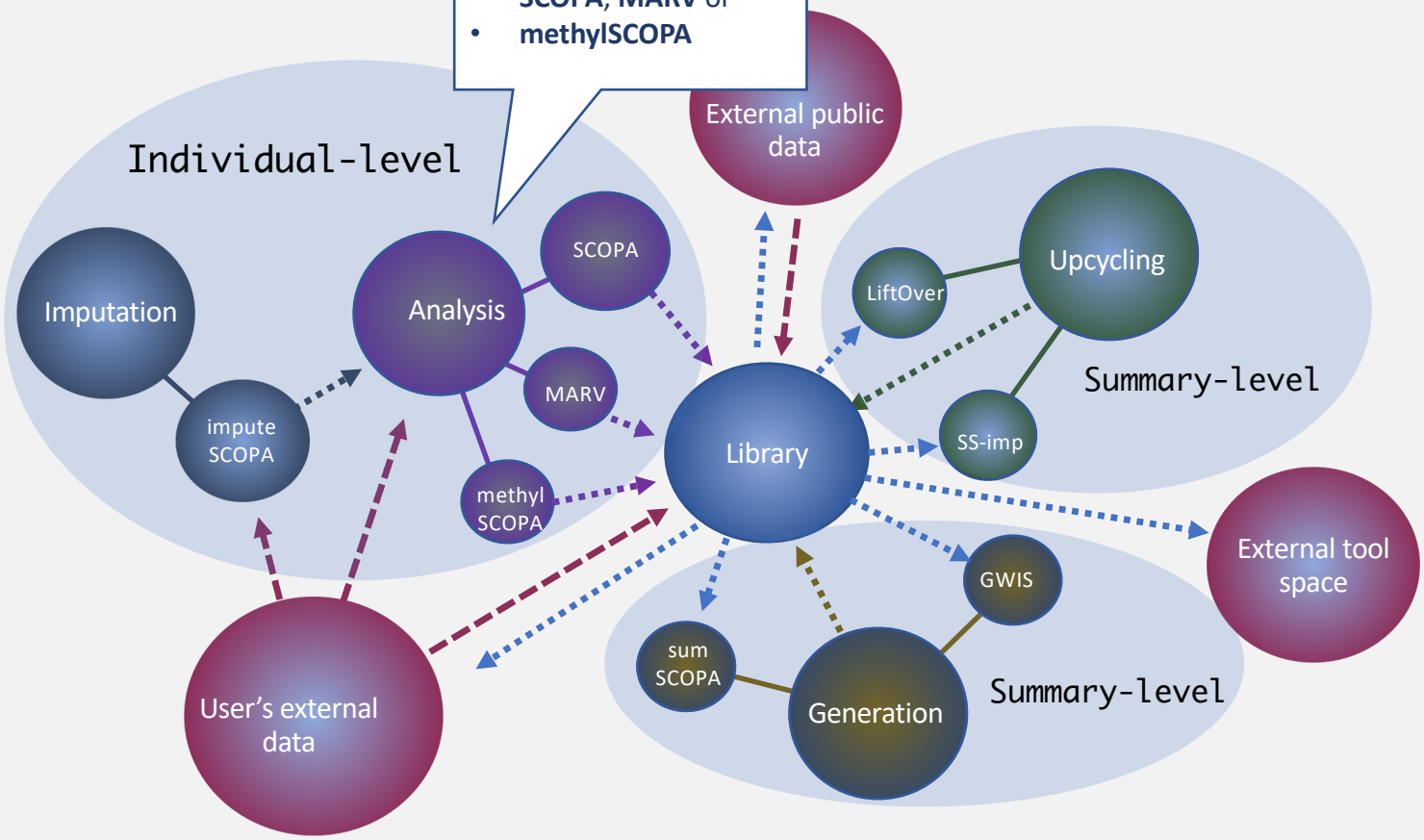
Contact:
anglia@prokopenkogroup.org

Prokopenko group
Section of Statistical multi-omics
University of Surrey

ANGLIA

Home | Library | Impute | ...rate | Upcycle | My jobs | Upload your data

The user can select multi-phenotype analysis methods among
- **SCOPA**, **MARV** or
- **methylSCOPA**



Contact:
anglia@prokopenkogroup.org

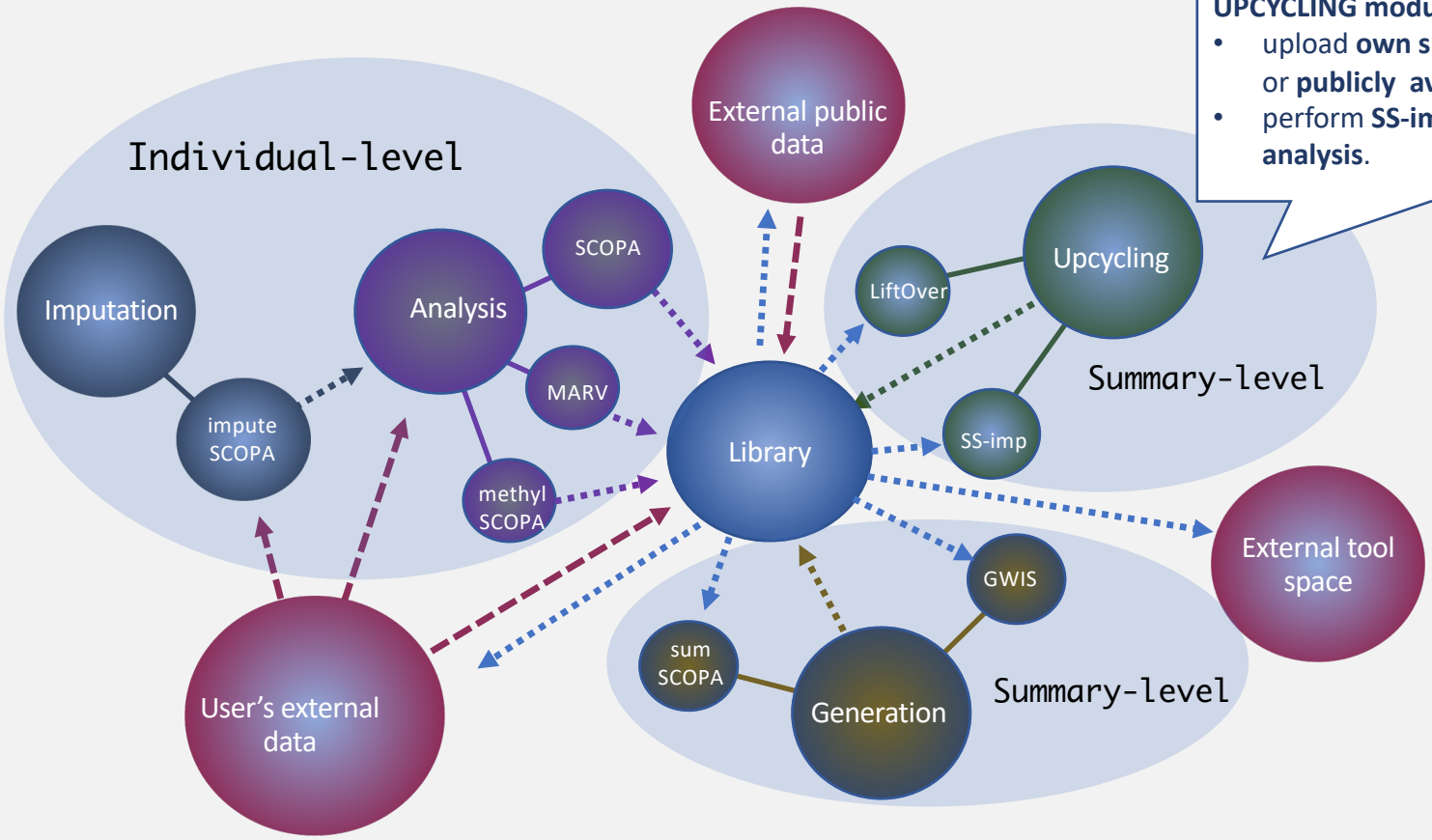
Prokopenko group
Section of Statistical multi-omics
University of Surrey

ANGLIA

Home | Library | Impute | Analyze | Generate | Upcycle | My jobs | Upload your data

UPCYCLING module helps the use to

- upload **own summary statistics data** or **publicly available data**
- perform **SS-imp** and/or **LiftOver analysis**.

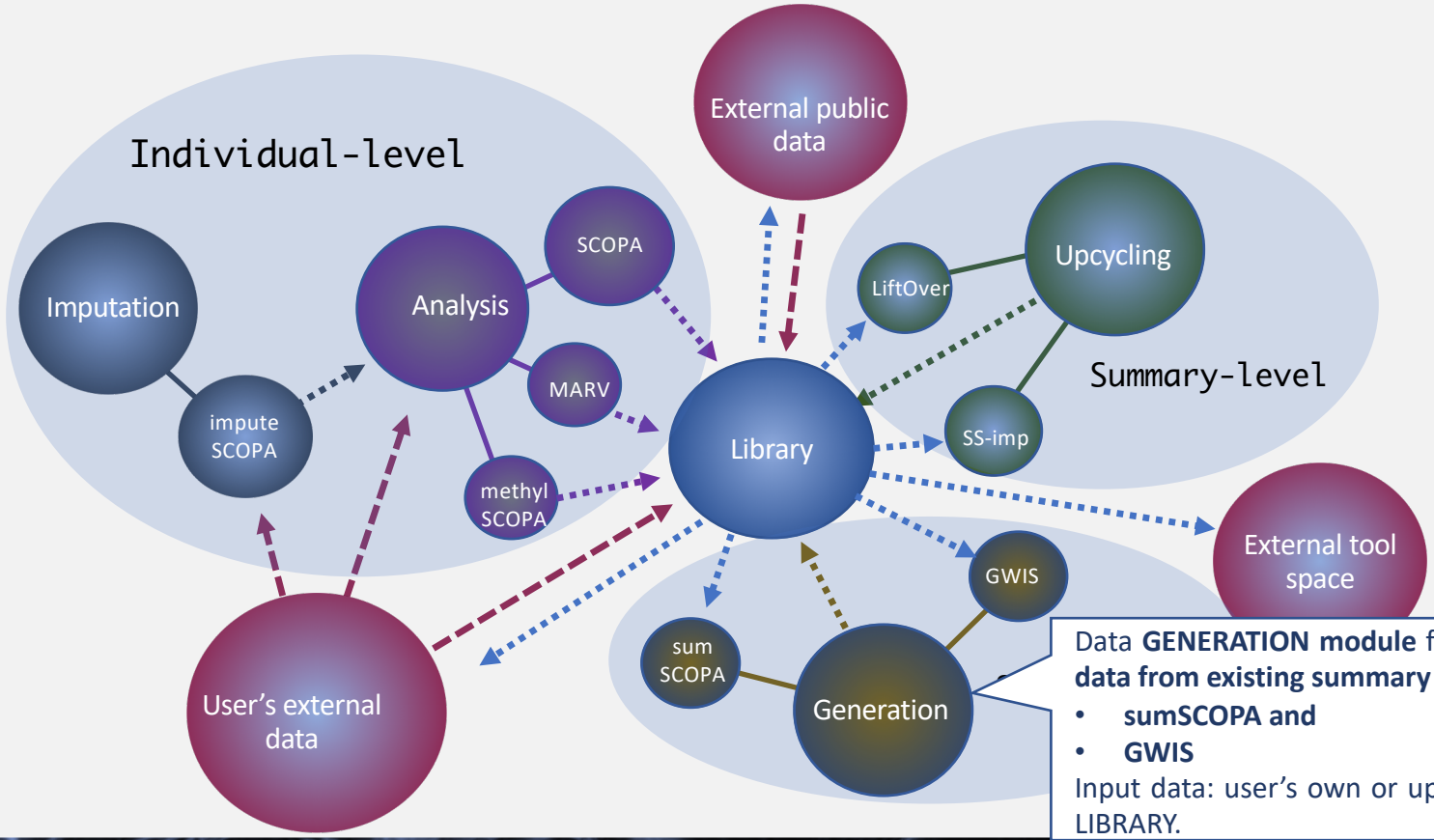


Contact:
anglia@prokopenkogroup.org

Prokopenko group
Section of Statistical multi-omics
University of Surrey

ANGLIA

Home | Library | Impute | Analyze | Generate | Upcycle | My jobs | Upload your data



Data **GENERATION module** facilitates generating **new data from existing summary statistics**, **by**

- **sumSCOPA and**
- **GWIS**

Input data: user's own or upcycled data file from the LIBRARY.

Contact:
anglia@prokopenkogroup.org

Prokopenko group
Section of Statistical multi-omics
University of Surrey

STORY OF TWO MANUSCRIPTS

Ayse Demirkan

X-omics series

"Strategies to hug your challenges in multi-omics data integration"

25 June 2020

Department of Clinical & Experimental Medicine, School of Biosciences & Medicine, University of Surrey, Guildford, UK
Department of Genetics, University Medical Center Groningen, Groningen, the Netherlands
@demyrkana @StatMultiOmics

Acknowledgements-ANGLIA

Prokopenko Group
Inga Prokopenko
Marika Kaakinen
Ayse Demirkan
Konstantin Rudometkin
Liudmila Zudina
Harmen Draisma
Zhanna Balkhiyarova
Mila Anasanti
Jared Maina
Igor Pupko
Anna Ulrich

Special thanks to:
SS-imp
Zoltan Kutalik
Sina Rueger

GWIS
Dorret Boomsma
Michel Nivard
Iryna Fedko

Funding:

Project LONGITOOLS

Contact:
a.demirkan@surrey.ac.uk

Spread the word:

@demyrkana
@StatMultiOmics