Molecular Epidemiology Biomarkers (III): Univariate approaches

Ioanna Tzoulaki

itzoulak@cc.uoi.gr

Reading list I

D Thomas. High-volume “-omics” technologies and the future of molecular epidemiology. *Epidemiology* 2006 17(5):490-1.

D Hunter. The future of molecular epidemiology. *Int J Epidemiol*. 1999 28(5): S1012-4.

C Wild, G Law and E Roman. Molecular epidemiology and cancer: promising areas of future research in the post-genomic era. Mutation Research 2002 499(1): 3-12.

Tzoulaki et al. Design and analysis of metabolomics studies in epidemiologic research: a primer on -omic technologies. Am J Epidemiol. 2014 Jul 15;180(2):129-39.

DeBord DG, Carreón T, Lentz TJ, Middendorf PJ, Hoover MD, Schulte PA. Use of the "Exposome" in the Practice of Epidemiology: A Primer on -Omic Technologies. Am J Epidemiol. 2016 Aug 15;184(4):302-14.

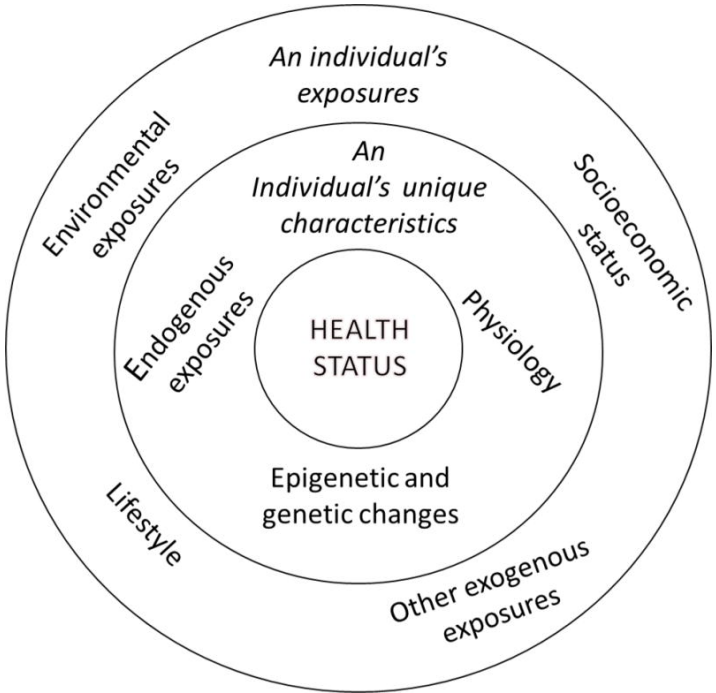
Yan H, Tian S, Slager SL, Sun Z, Ordog T. Genome-Wide Epigenetic Studies in Human Disease: A Primer on -Omic Technologies. Am J Epidemiol. 2016 Jan 15;183(2):96-109.

Mischak H, Critselis E, Hanash S, Gallagher WM, Vlahou A, Ioannidis JP. Epidemiologic design and analysis for proteomic studies: a primer on -omic technologies. Am J Epidemiol. 2015 May 1;181(9):635-47.

Coughlin SS. Toward a road map for global -omics: a primer on -omic technologies. Am J Epidemiol. 2014 Dec 15;180(12):1188-95.

The exposome

- Limited contribution of genetic factors to the risk pf chronic disease
- 79-90% attributed top exposures
  - Exposure measurement suboptimal
    - Which?
    - How?



Exposome measures the totality of environmental exposures from conception onwards

Omics data

High throughput biochemical measures of the abundance and/or structural features of molecules involved in main biological processes such as metabolism and its regulation.

	Supporting Structure	Platforms ($\log_{10}$ order of magnitude)	Features
Genome	DNA	Microarrays (6) Sequencing (9)	Categorical data Distance-driven correlation Extremely stable over time
Epigenome	DNA methylation Histone modifications Non-coding RNA	Microarrays (5) Bisulfite sequencing (1)	Continuous data Affected by time and exposures (with reduced plasticity)
Transcriptome	mRNA	Microarrays (5) RNA sequencing (9)	Continuous data Affected by time and exposures Strong measurement noise
Proteome	Proteins	Microarrays (5) Mass spectrometry (5)	Continuous data Affected by time and exposures
Metabolome	Small molecules	Mass spectrometry (5) NMR spectroscopy (4)	Continuous data Structured correlation Strongly affected by exposures

Technologies

- Genomics
  - Targeted (SNPs) or untargeted (WGS)
  - Binary or continuous (dosage data)
- Epigenetics (DNA methylation)
  - Methylation sites
  - Percentage of methylated cytosines at each CpG locus
  - Average over many cells, possibly of different types
- Transcriptomics
  - Targeted (micro-arrays) or untargeted (RNA sequence)
  - Intensities proportional to RNA abundances or sequence reads
- Metabolomics / Proteomics
  - Targeted or untargeted (MS and NMR)
  - Quantified proteins/metabolites or mass and retention times, or spectra

Challenges for Biomarker Studies in the 'Omics' Era

1. Precious and limited biobanked material, not easily accessed

2. Single (spot) biological samples

3. Usually blood, not urine (which may be better e.g. for metabolomics)

4. No cohorts allow life-course epidemiology

5. In-depth exposure assessment is limited by feasibility

6. Lab measurements and omics have the same limitations related to sample size and feasibility

7. Biostatistical approaches and causal interpretation

8. Ethical issues

Advantages of Omics data

- Agnostic view of cellular activity
- Measure the main biological processes involved in the regulation of cellular metabolism
- Use for the Exposome: Omics biomarkers have the potential to highlight internal responses to external stresses

Main data characteristics

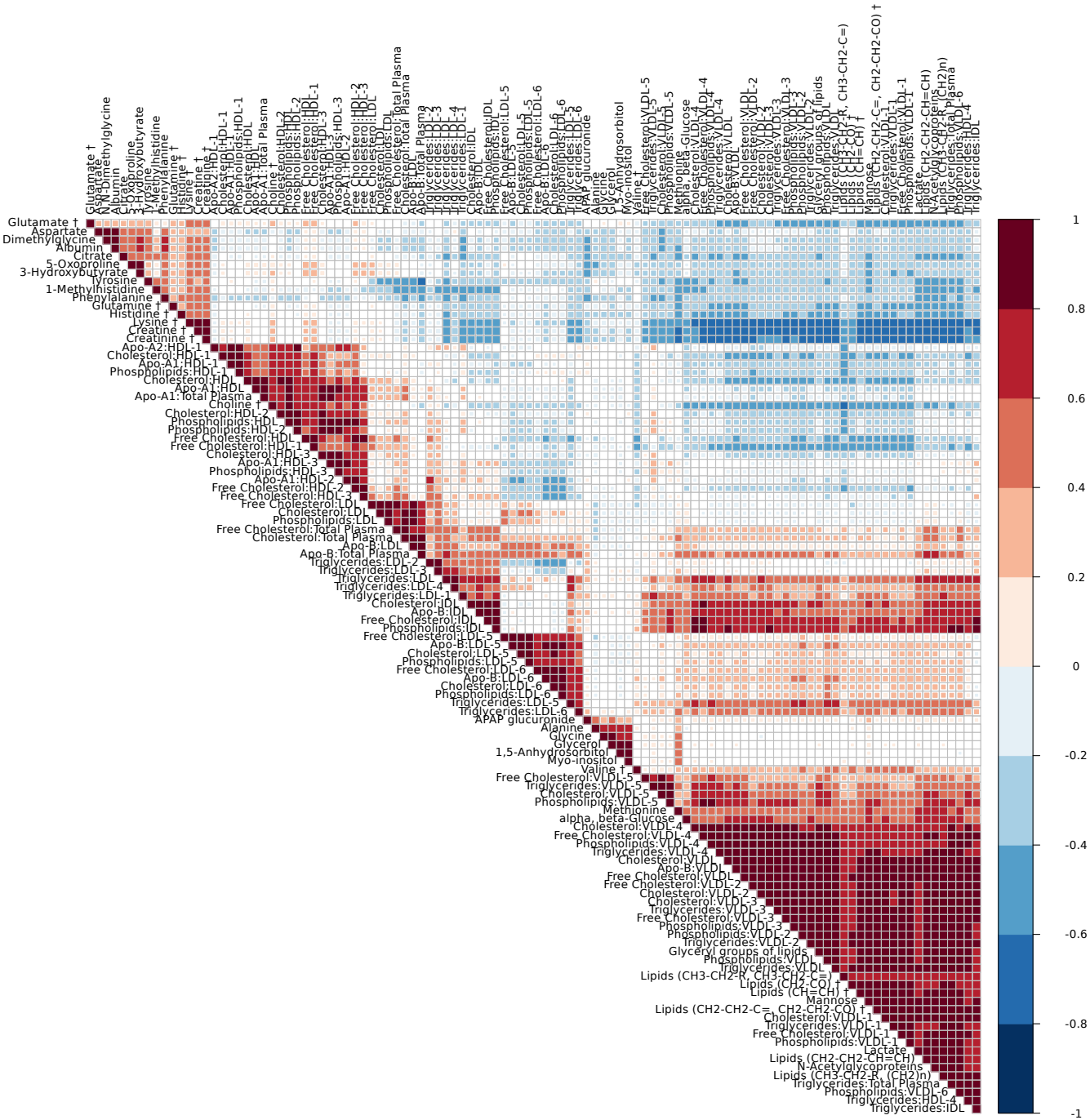
- High dimension
  - ranging from hundreds to millions
- Nature
  - continuous/binary/categorical/counts
- Noise/ Measurement error
  - sensitive to experimental conditions
- Stability

➢ need for flexible statistical framework to accommodate huge heterogeneity in data, response and dose-response relationships

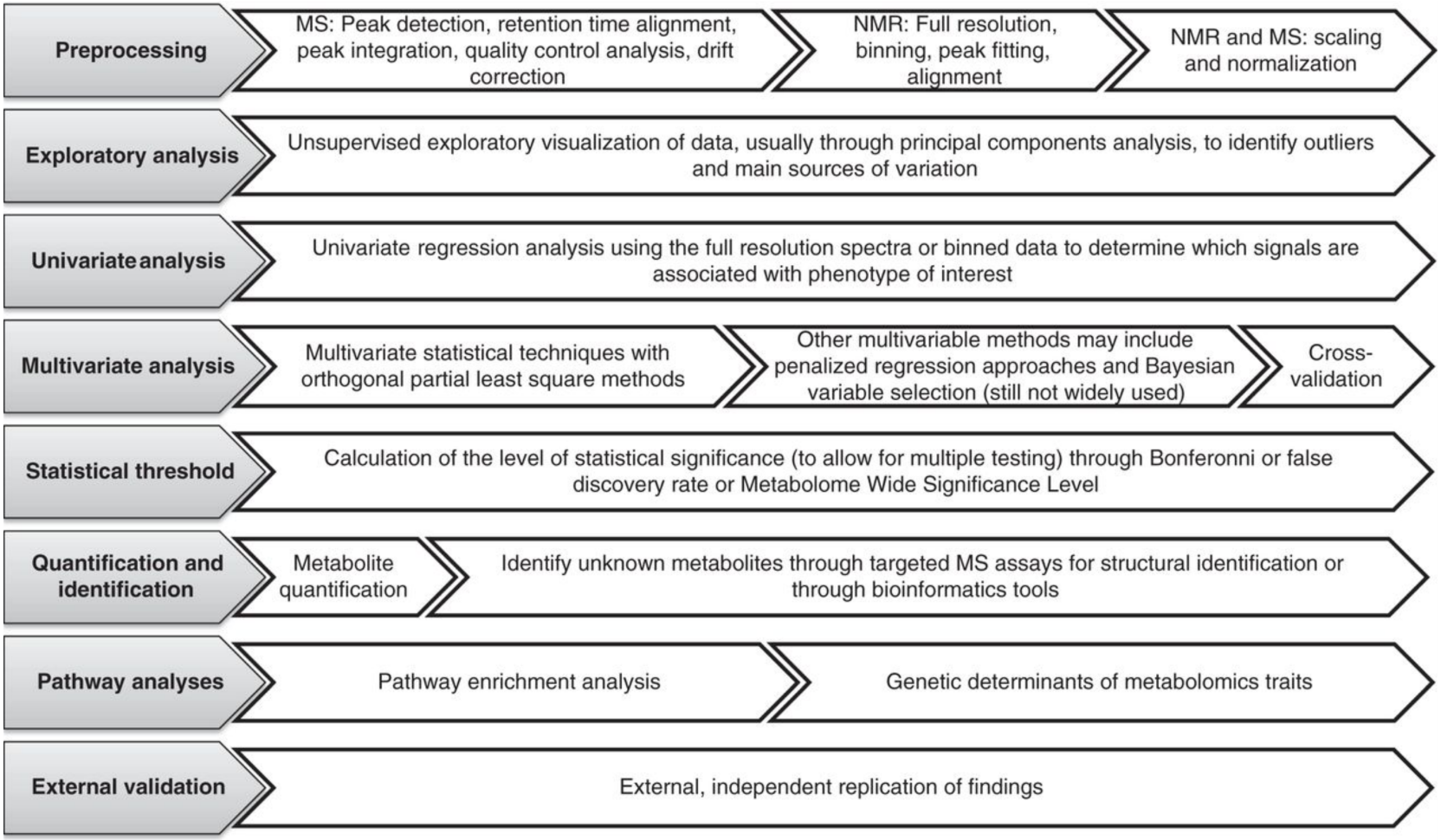
➢ (generalised) linear models

Heterogeneity

- Nature of the data
  - Binary variables (haplotype data)
  - Categorical variables (e.g. genotype data)
  - Continuous variables (e.g. % of methylation . . .)
- Dimension: wide range of scales
  - Hundreds of measurements (proteins levels)
  - Tens of thousands of variables: (NMR-MS spectral data)
  - Hundreds of thousands of variables (full genome scans)
- Correlated structure in the data:
  - Strength of the correlation varies
  - Correlation structure can either be ‘distance-driven’ (e.g LD genomics data) or more complex (e.g. NMR spectral data).


Glutamate †
Aspartate
Dimethylglycine
Albumin
Citrate
5-Oxoproline
3-Hydroxybutyrate
Tyrosine
1-Methylhistidine
Phenylalanine
Glutamine †
Histidine †
Lysine †
Creatine †
Creatinine †
Apo-A2:HDL-1
Cholesterol:HDL-1
Apo-A1:HDL-1
Phospholipids:HDL-1
Cholesterol:HDL
Apo-A1:HDL
Apo-A1:Total Plasma
Choline †
Cholesterol:HDL-2
Phospholipids:HDL
Phospholipids:HDL-2
Free Cholesterol:HDL
Free Cholesterol:HDL-1
Cholesterol:HDL-3
Apo-A1:HDL-3
Phospholipids:HDL-3
Apo-A1:HDL-2
Free Cholesterol:HDL-2
Free Cholesterol:HDL-3
Free Cholesterol:LDL
Cholesterol:LDL
Phospholipids:LDL
Free Cholesterol:Total Plasma
Cholesterol:Total Plasma
Apo-B:LDL
Apo-B:Total Plasma
Triglycerides:LDL-2
Triglycerides:LDL-3
Triglycerides:LDL
Triglycerides:LDL-4
Triglycerides:LDL-1
Cholesterol:IDL
Apo-B:IDL
Free Cholesterol:IDL
Phospholipids:IDL
Free Cholesterol:LDL-5
Apo-B:LDL-5
Cholesterol:LDL-5
Phospholipids:LDL-5
Free Cholesterol:LDL-6
Apo-B:LDL-6
Cholesterol:LDL-6
Phospholipids:LDL-6
Triglycerides:LDL-5
Triglycerides:LDL-6
APAP glucuronide
Alanine
Glycine
Glycerol
1,5-Anhydrosorbitol
Myo-inositol
Valine †
Free Cholesterol:VLDL-5
Triglycerides:VLDL-5
Cholesterol:VLDL-5
Phospholipids:VLDL-5
Methionine
alpha, beta-Glucose
Cholesterol:VLDL-4
Free Cholesterol:VLDL-4
Phospholipids:VLDL-4
Triglycerides:VLDL-4
Cholesterol:VLDL
Apo-B:VLDL
Free Cholesterol:VLDL
Free Cholesterol:VLDL-2
Cholesterol:VLDL-2
Cholesterol:VLDL-3
Triglycerides:VLDL-3
Free Cholesterol:VLDL-3
Phospholipids:VLDL-3
Phospholipids:VLDL-2
Triglycerides:VLDL-2
Glycerol groups of lipids
Phospholipids:VLDL
Triglycerides:VLDL
Lipids (CH3-CH2-R, CH3-CH2-C=)
Lipids (CH2-CO) †
Lipids (CH=CH) †
Mannose
Lipids (CH2-CH2-C=, CH2-CH2-CO) †
Cholesterol:VLDL-1
Triglycerides:VLDL-1
Free Cholesterol:VLDL-1
Phospholipids:VLDL-1
Lactate
Lipids (CH2-CH2-CH=CH)
N-Acetylglycoproteins
Lipids (CH3-CH2-R, (CH2)n)
Triglycerides:Total Plasma
Phospholipids:VLDL-6
Triglycerides:HDL-4
Triglycerides:IDL
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Data analysis in *Omics*



Tzoulaki et al AJE 2014

Threats to the Validity of Molecular Epidemiology Studies

- Bias-systematic error
  - Information Bias (imprecision in measurements)
  - Confounding

- Statistical Issues
  - Over-emphasis on P-values
  - Multiple Comparisons
  - Association ≠ Causation

Advances in epigenome-wide association studies for common diseases

Dirk S. Paul and Stephan Beck

UCL Cancer Institute, University College London, London, WC1E 6BT, UK

Epigenome-wide association studies (EWASs) provide a systematic approach to uncovering epigenetic variants underlying common diseases. Discoveries have shed light on novel molecular mechanisms of disease and enabled the application of epigenetic variants as biomarkers. Here, we highlight the recent advances in this emerging line of research and discuss key challenges for current and future studies.

Many common diseases in humans are mediated by genetic disease. The authors probed DNA methylation marks in whole blood. Indeed, whole blood has proven to be the tissue of choice for most EWASs owing to its ease of accessibility. Importantly, they found that the proportions of the major circulating leukocytes differ between cases and controls. Statistical methods are capable of inferring and correcting for such cellular heterogeneity, either with [4] or without [5,6] the use of reference data sets. Following reference-based adjustment, Liu *et al.* achieved a substantial reduction of spurious association signals attributed to

nature COMMUNICATIONS

ARTICLE

Received 18 Sep 2015 | Accepted 20 Jul 2016 | Published 1 Sep 2016

DOI: 10.1038/ncomms12649 OPEN

Proteome-wide association studies identify biochemical modules associated with a wing-size phenotype in *Drosophila melanogaster*

Hirokazu Okada[1], H. Alexander Ebhardt[1], Sibylle Chantal Vonesch[1], Ruedi Aebersold[1,2] & Ernst Hafen[1,2]

Published in final edited form as:

J Proteome Res. 2010 September 3; 9(9): 4620–4627. doi:10.1021/pr1003449.

Metabolic Profiling And The Metabolome-Wide Association Study: Significance Level For Biomarker Identification

Marc Chadeau-Hyam[†,#], Timothy M D Ebbels[‡,#], Ian J Brown[†], Queenie Chan[†], Jeremiah Stamler[¶], Chiang Ching Huang[¶], Martha L Daviglus[¶], Hirotsugu Ueshima[§], Liancheng Zhao[‖], Elaine Holmes[‡,⊥], Jeremy K Nicholson[‡,⊥], Paul Elliott[*,†,⊥], and Maria De Iorio[*,†]

Department of Epidemiology and Biostatistics, School of Public Health, Imperial College, London W2 1PG, UK, Biomolecular Medicine, Department of Surgery and Cancer, Faculty of Medicine, Imperial College, London SW7 2AZ, UK, Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, Illinois 60611, US, Department of Health Science, Shiga University of Medical Science, Otsu, Japan, Department of Epidemiology, Fu Wai Hospital and Cardiovascular Institute, Chinese Academy of Medical Sciences, Beijing, People's Republic of China, and MRC-HPA Center for Environment and Health, Imperial College London UK

Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data

Joshua C Denny[1,2], Lisa Bastarache[2], Marylyn D Ritchie[3], Robert J Carroll[2], Raquel Zink[2], Jonathan D Mosley[1], Julie R Field[4], Jill M Pulley[4,5], Andrea H Ramirez[1], Erica Bowton[4], Melissa A Basford[4], David S Carrell[6], Peggy L Peissig[7], Abel N Kho[8], Jennifer A Pacheco[9], Luke V Rasmussen[10], David R Crosslin[11], Paul K Crane[12], Jyotishman Pathak[13], Suzette J Bielinski[14], Sarah A Pendergrass[3], Hua Xu[15], Lucia A Hindorff[16], Rongling Li[16], Teri A Manolio[16], Christopher G Chute[13], Rex L Chisholm[17], Eric B Larson[6], Gail P Jarvik[11,12], Murray H Brilliant[18], Catherine A McCarty[19], Iftikhar J Kullo[20], Jonathan L Haines[21], Dana C Crawford[21], Daniel R Masys[22] & Dan M Roden[1,23]

Candidate gene and genome-wide association studies (GWAS) have identified genetic variants that modulate risk for human disease; many of these associations require further study to replicate the results. Here we report the first

large number of single variant–phenotype associations has led to the serendipitous identification of single loci associated with multiple diseases, or pleiotropy. Notable examples include variants at 9p21.3, which were associated initially with early myocardial infarction[2] and

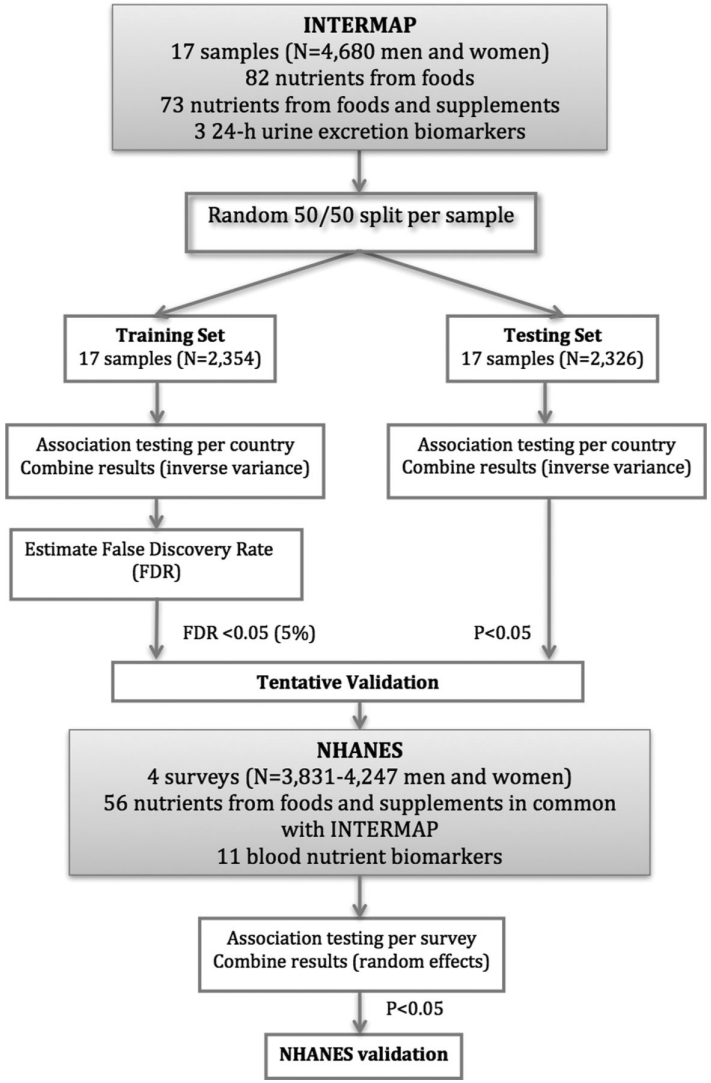
A Nutrient-Wide Association Study on Blood Pressure

Ioanna Tzoulaki, PhD;* Chirag J. Patel, PhD;* Tomonori Okamura, MD, PhD; Queenie Chan, PhD; Ian J. Brown, PhD; Katsuyuki Miura, MD, PhD; Hirotsugu Ueshima, MD, PhD; Liancheng Zhao, MD; Linda Van Horn, PhD; Martha L. Daviglus, MD, PhD; Jeremiah Stamler, MD; Atul J. Butte, MD, PhD; John P.A. Ioannidis, MD, DSc; Paul Elliott, MB BS, PhD

Background—A nutrient-wide approach may be useful to comprehensively test and validate associations between nutrients (derived from foods and supplements) and blood pressure (BP) in an unbiased manner.

Methods and Results—Data from 4680 participants aged 40 to 59 years in the cross-sectional International Study of Macro Micronutrients and Blood Pressure (INTERMAP) were stratified randomly into training and testing sets. US National Health and Nutrition Examination Survey (NHANES) four cross-sectional cohorts (1999–2000, 2001–2002, 2003–2004, 2005–2006) were used for external validation. We performed multiple linear regression analyses associating each of 82 nutrients and 3 urine electrolytes with systolic and diastolic BP in the INTERMAP training set. Significant findings were validated in the INTERMAP testing set and further in the NHANES cohorts (false discovery rate <5% in training, $P<0.05$ for internal and external validation). Among the validated nutrients, alcohol and urinary sodium-to-potassium ratio were directly associated with systolic BP, and dietary phosphorus, magnesium, iron, thiamin, folacin, and riboflavin were inversely associated with systolic BP. In addition, dietary folacin and riboflavin were inversely associated with diastolic BP. The absolute effect sizes in the validation data (NHANES) ranged from 0.97 mm Hg lower systolic BP (phosphorus) to 0.39 mm Hg lower systolic BP (thiamin) per 1-SD difference in nutrient variable. Inclusion of nutrient intake from supplements in addition to foods gave similar results for some nutrients, though it attenuated the associations of folacin, thiamin, and riboflavin intake with BP.

Conclusions—We identified significant inverse associations between B vitamins and BP, relationships hitherto poorly investigated. Our analyses represent a systematic unbiased approach to the evaluation and validation of nutrient-BP associations. **(*Circulation*. 2012;126:2456-2464.)**


INTERMAP
17 samples (N=4,680 men and women)
82 nutrients from foods
73 nutrients from foods and supplements
3 24-h urine excretion biomarkers
Random 50/50 split per sample
Training Set
17 samples (N=2,354)
Testing Set
17 samples (N=2,326)
Association testing per country
Combine results (inverse variance)
Association testing per country
Combine results (inverse variance)
Estimate False Discovery Rate
(FDR)
FDR <0.05 (5%)
P<0.05
Tentative Validation
NHANES
4 surveys (N=3,831-4,247 men and women)
56 nutrients from foods and supplements in common
with INTERMAP
11 blood nutrient biomarkers
Association testing per survey
Combine results (random effects)
P<0.05
NHANES validation

OPEN ACCESS Freely available online

PLoS one

An Environment-Wide Association Study (EWAS) on Type 2 Diabetes Mellitus

Chirag J. Patel[1,2,3], Jayanta Bhattacharya[4], Atul J. Butte[1,2,3]*

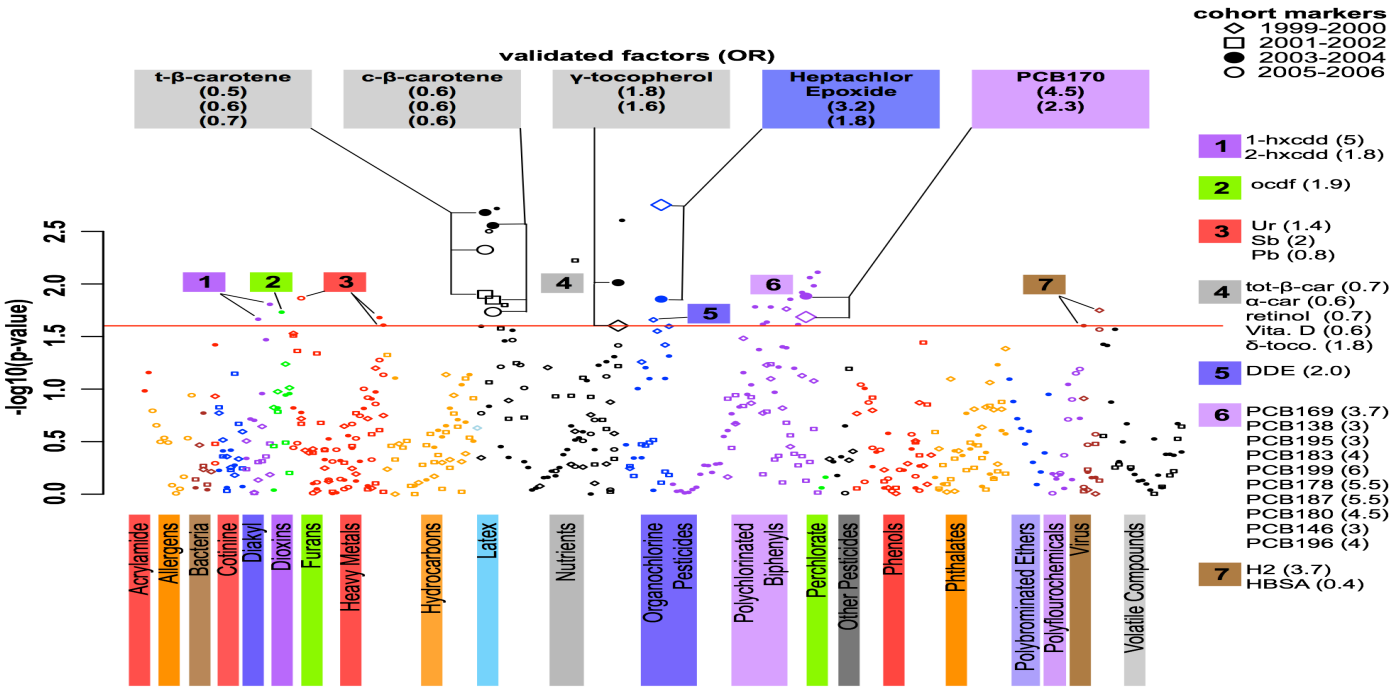
1 Department of Pediatrics and Medicine, Stanford University School of Medicine, Stanford, California, United States of America, **2** Stanford Center for Biomedical Informatics Research, Stanford University School of Medicine, Stanford, California, United States of America, **3** Lucile Packard Children's Hospital, Palo Alto, California, United States of America, **4** Center For Primary Care and Outcomes Research, Stanford University School of Medicine, Stanford, California, United States of America

Abstract

Background: Type 2 Diabetes (T2D) and other chronic diseases are caused by a complex combination of many genetic and environmental factors. Few methods are available to comprehensively associate specific physical environmental factors with disease. We conducted a pilot Environmental-Wide Association Study (EWAS), in which epidemiological data are comprehensively and systematically interpreted in a manner analogous to a Genome Wide Association Study (GWAS).

Methods and Findings: We performed multiple cross-sectional analyses associating 266 unique environmental factors with clinical status for T2D defined by fasting blood sugar (FBG) concentration $\geq$126 mg/dL. We utilized available Centers for Disease Control (CDC) National Health and Nutrition Examination Survey (NHANES) cohorts from years 1999 to 2006. Within cohort sample numbers ranged from 503 to 3,318. Logistic regression models were adjusted for age, sex, body mass index (BMI), ethnicity, and an estimate of socioeconomic status (SES). As in GWAS, multiple comparisons were controlled and significant findings were validated with other cohorts. We discovered significant associations for the pesticide-derivative heptachlor epoxide (adjusted OR in three combined cohorts of 1.7 for a 1 SD change in exposure amount; $p<0.001$), and the vitamin γ-tocopherol (adjusted OR 1.5; $p<0.001$). Higher concentrations of polychlorinated biphenyls (PCBs) such as PCB170 (adjusted OR 2.2; $p<0.001$) were also found. Protective factors associated with T2D included β-carotenes (adjusted OR 0.6; $p<0.001$).

Environment wide association study on type 2 diabetes



Multiple Comparison Problem in 'Omics' studies

	Normal					Disease					
Gene 1	0.701365258	0.847689154	0.945472154	0.644555958	0.86880259	0.553831918	0.216928593	0.973412306	0.999717081	0.030686471	0.258952072
Gene 2	0.019693544	0.998953774	0.79541506	0.784368111	0.786279804	0.488011858	0.109621914	0.370060164	0.699715047	0.906833389	0.477616141
Gene 3	0.823234225	0.009390884	0.173507875	0.86814406	0.781284479	0.084611403	0.697088945	0.592397243	0.158629413	0.387556786	0.517460405
Gene 4	0.831201089	0.672332684	0.709812715	0.614309625	0.058084282	0.057314605	0.036616132	0.515439251	0.824838113	0.902083252	0.641959022
Gene 5	0.618048089	0.493722217	0.582979716	0.909020223	0.089930431	0.435987475	0.300954006	0.401800668	0.36287023	0.721856109	0.550259337
Gene 6	0.314244277	0.693208332	0.507662222	0.910433429	0.642351972	0.650730411	0.694156972	0.952770501	0.165252532	0.503087392	0.903471832
Gene 7	0.834701125	0.975953907	0.538782775	0.544151697	0.431703426	0.40012594	0.090574576	0.778406246	0.099311443	0.59307239	0.14690471
Gene 8	0.632542712	0.320787292	0.573479184	0.600636977	0.280344436	0.840668539	0.953859038	0.93067047	0.183795382	0.638818057	0.194666534
Gene 9	0.613812632	0.943127333	0.789148665	0.740696336	0.756161519	0.225290514	0.998161929	0.192950694	0.152709112	0.672583819	0.104214494
Gene 10	0.326036635	0.138067146	0.613095022	0.782722541	0.055087176	0.105971326	0.89495784	0.619088186	0.798195475	0.416937562	0.379330623
Gene 11	0.634973714	0.556111533	0.843606126	0.770987963	0.243204132	0.625448193	0.774528794	0.350605578	0.36276179	0.835054279	0.893488236
Gene 12	0.965398561	0.057168922	0.567125297	0.763013231	0.413766749	0.327217012	0.311494135	0.134875146	0.517469133	0.95852006	0.634666711
Gene 13	0.12216374	0.433638925	0.669994608	0.929084475	0.946953019	0.204031316	0.656656377	0.009321932	0.637010051	0.141680378	0.194537816
Gene 14	0.414223175	0.383942752	0.682146127	0.918495607	0.382467827	0.782112064	0.333122917	0.143586717	0.898119274	0.557894875	0.941420469
Gene 15	0.285974499	0.155930996	0.330072963	0.383671395	0.716907409	0.864141357	0.490873804	0.781127292	0.92330326	0.021729016	0.240506468
Gene 16	0.672888773	0.772635752	0.674517227	0.765489034	0.713345501	0.317341191	0.415206224	0.385831293	0.378462402	0.730507282	0.00693229
Gene 17	0.016216298	0.008760328	0.122856594	0.911411537	0.054231562	0.094487454	0.345526591	0.057715898	0.016620408	0.8738592	0.821530697
Gene 18	0.551922437	0.097837061	0.6162674	0.410259157	0.913703161	0.789701193	0.026344507	0.093459699	0.292196191	0.590586608	0.44261104
Gene 19	0.88922594	0.629840151	0.642071927	0.437341731	0.349580595	0.717605676	0.253664017	0.681060437	0.682633708	0.585084141	0.965814376
Gene 20	0.679047253	0.610385651	0.984636956	0.522444904	0.983714469	0.008354579	0.54121905	0.910983448	0.862391892	0.104260295	0.23427917

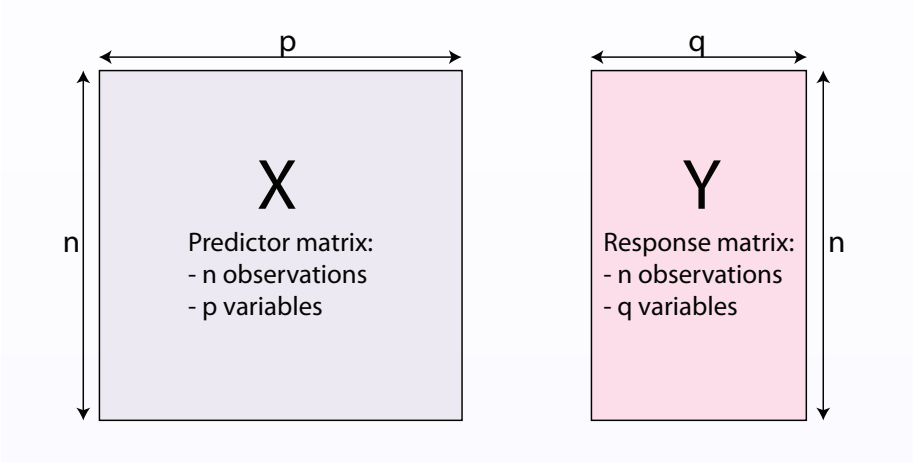
Gene by gene 2-tailed t-test; $P<0.05$ significant

Multiple Comparison Problem in 'Omics' studies

	Normal					Disease					
Gene 1	0.701365258	0.847689154	0.945472154	0.644555958	0.86880259	0.553831918	0.216928593	0.973412306	0.999717081	0.030686471	0.258952072
Gene 2	0.019693544	0.998953774	0.79541506	0.784368111	0.786279804	0.488011858	0.109621914	0.370060164	0.699715047	0.906833389	0.477616141
Gene 3	0.823234225	0.009390884	0.173507875	0.86814406	0.781284479	0.084611403	0.697088945	0.592397243	0.158629413	0.387556786	0.517460405
Gene 4	0.831201089	0.672332684	0.709812715	0.614309625	0.058084282	0.057314605	0.036616132	0.515439251	0.824838113	0.902083252	0.641959022
Gene 5	0.618048089	0.493722217	0.582979716	0.909020223	0.089930431	0.435987475	0.300954006	0.401800668	0.36287023	0.721856109	0.550259337
Gene 6	0.314244277	0.693208332	0.507662222	0.910433429	0.642351972	0.650730411	0.694156972	0.952770501	0.165252532	0.503087392	0.903471832
Gene 7	0.834701125	0.975953907	0.538782775	0.544151697	0.431703426	0.40012594	0.090574576	0.778406246	0.099311443	0.59307239	0.14690471
Gene 8	0.632542712	0.320787292	0.573479184	0.600636977	0.280344436	0.840668539	0.953859038	0.93067047	0.183795382	0.638818057	0.194666534
Gene 9	0.613812632	0.943127333	0.789148665	0.740696336	0.756161519	0.225290514	0.998161929	0.192950694	0.152709112	0.672583819	0.104214494
Gene 10	0.326036635	0.138067146	0.613095022	0.782722541	0.055087176	0.105971326	0.89495784	0.619088186	0.798195475	0.416937562	0.379330623
Gene 11	0.634973714	0.556111533	0.843606126	0.770987963	0.243204132	0.625448193	0.774528794	0.350605578	0.36276179	0.835054279	0.893488236
Gene 12	0.965398561	0.057168922	0.567125297	0.763013231	0.413766749	0.327217012	0.311494135	0.134875146	0.517469133	0.95852006	0.634666711
Gene 13	0.12216374	0.433638925	0.669994608	0.929084475	0.946953019	0.204031316	0.656656377	0.009321932	0.637010051	0.141680378	0.194537816
Gene 14	0.414223175	0.383942752	0.682146127	0.918495607	0.382467827	0.782112064	0.333122917	0.143586717	0.898119274	0.557894875	0.941420469
Gene 15	0.265974499	0.155930996	0.330072963	0.363671395	0.716907409	0.864141357	0.490673604	0.781127292	0.92330326	0.021729016	0.240506466
Gene 16	0.672888773	0.772635752	0.674517227	0.765489034	0.713345501	0.317341191	0.415206224	0.385831293	0.378462402	0.730507282	0.00693229
Gene 17	0.016216296	0.006760326	0.122656594	0.911411537	0.054231562	0.094487454	0.345526591	0.057715696	0.016620406	0.8736592	0.621530697
Gene 18	0.551922437	0.097837061	0.6162674	0.410259157	0.913703161	0.789701193	0.026344507	0.093459699	0.292196191	0.590586608	0.44261104
Gene 19	0.88922594	0.629840151	0.642071927	0.437341731	0.349580595	0.717605676	0.253664017	0.681060437	0.682633708	0.585084141	0.965814376
Gene 20	0.679047253	0.610385651	0.984636956	0.522444904	0.983714469	0.008354579	0.54121905	0.910983448	0.862391892	0.104260295	0.23427917

Conclude: Gene 16 associated with disease

Limitations of x-WAS studies



The n < p situation:

◦ More predictors than observations

→ numerically intractable statistical inferences

◦ n > p

→ univariate approaches

→ dimension reductions techniques

→ variable selection methods

Univariate methods

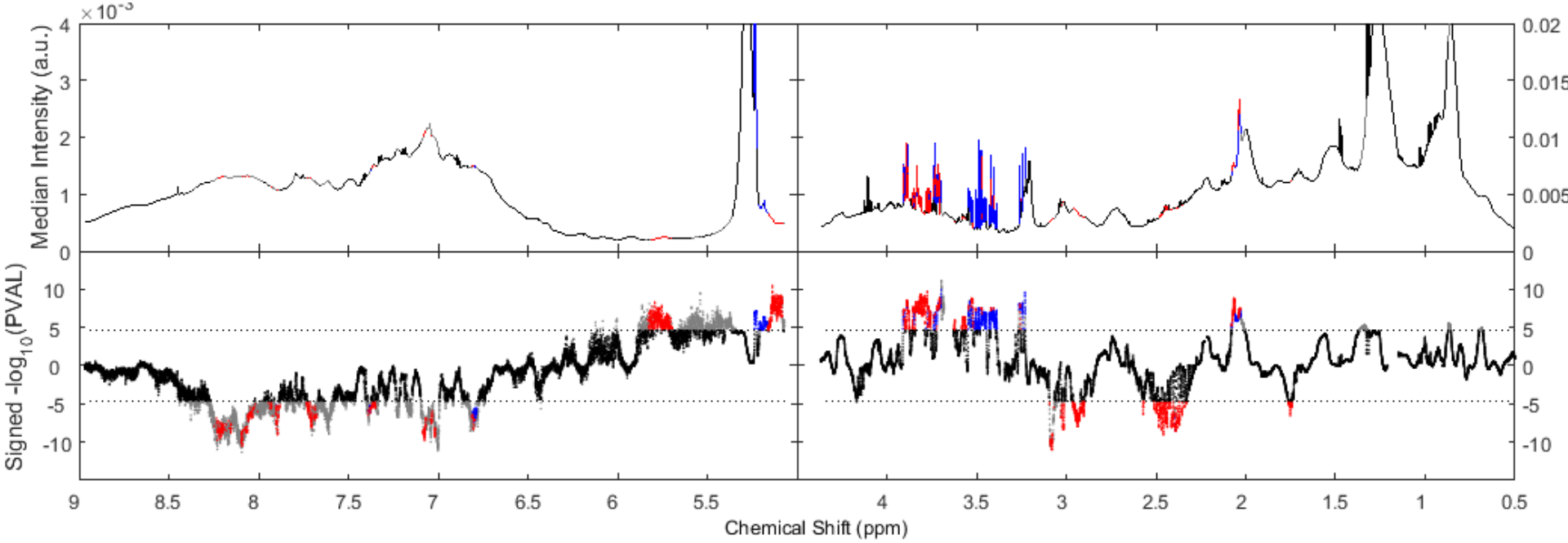
- Each X with each Y
- Each measurement with the outcome
- Common in GWAS

$$Y_i = \alpha + \beta X_{ij} + \epsilon_{ij},$$

where:

- Y_i is the measured outcome (possibly multivariate)
- X_{ij} is the observed value for j^{th} predictor
- α is the intercept
- β is the regression coefficient
- ϵ_{ij} is the residual error measuring the random deviation from the linear relationship

$\Rightarrow p$ models are estimated (one per predictor)


×10⁻³
Median Intensity (a.u.)
Signed -log10(PVAL)
Chemical Shift (ppm)
9
8.5
8
7.5
7
6.5
6
5.5
4
3.5
3
2.5
2
1.5
1
0.5

Multiple Comparisons

- However….need to consider the number of tests performed-$P<0.05$ means we accept the risk of erroneously rejecting H_0 in 5% of the cases (i.e. willing to accept 5% false positives)

- Each comparisons carries a 5% error probability so if we perform 20 tests, likely to detect 1 false positive

- The association between gene 16 and disease may be real but we do not have sufficient data to make that claim

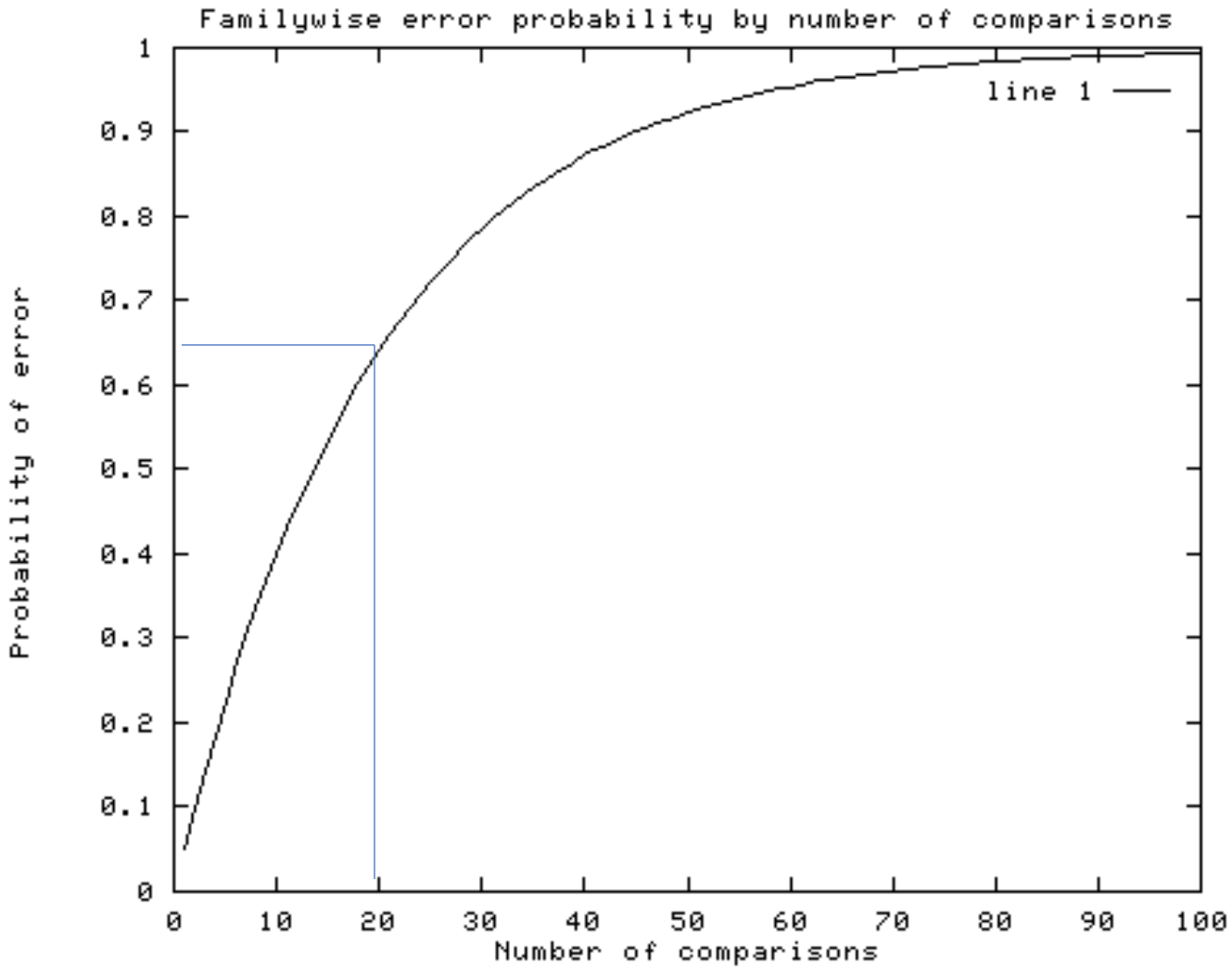
Probability of ≥1 False Positives by Chance

If set P-value at <0.05

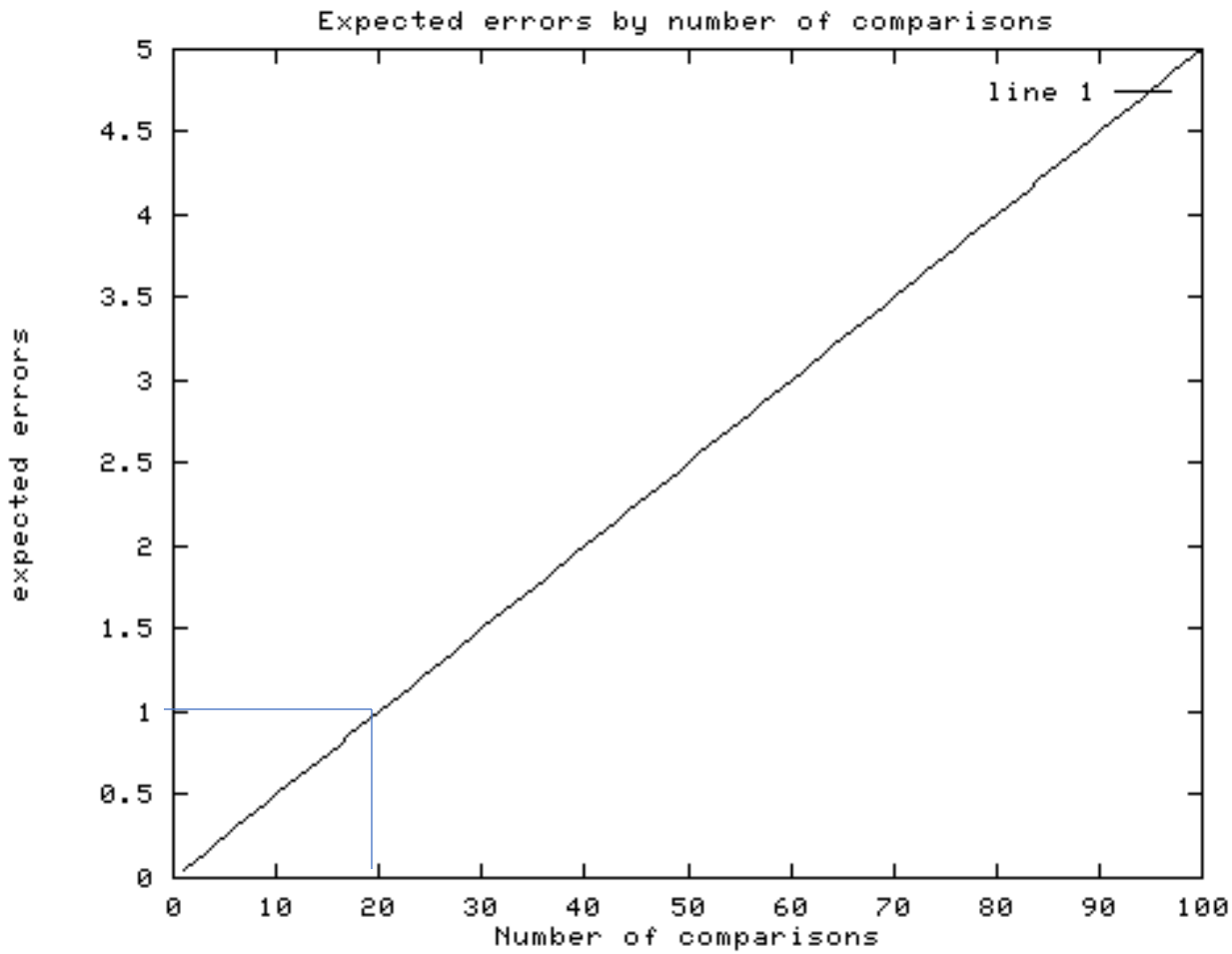
# Genes tested (N)	Incidence False Positives	Probability of detecting ≥ false +ves
1	1/20	5%
2	1/10	10%
20	1	64%
100	5	99.4%

$$100(1-0.95^N)$$

Probability of Error and Number of Comparisons



Expected Number of Errors and Number of Comparisons



Multiple Comparison Problem in 'Omics' studies

- On the 'omics' scale problem is magnified..
- ~10,000 genes on an array
  - Each gene-disease association has 5% chance of being false positive (Type I error)

- So by chance alone, we should detect 500 significant associations…..
- For α=0.05 and 2.5×10^6 SNPs: 125,000 FP

Sources of Multiple Comparisons

Source	Example
Multiple outcomes	a cohort study looking at the incidence of breast cancer, colon cancer, and lung cancer
Multiple predictors	an observational study with 40 dietary predictors or a trial with 4 randomization groups
Subgroup analyses	a randomized trial that tests the efficacy of an intervention in 20 subgroups based on prognostic factors
Multiple definitions for the exposures and outcomes	an observational study where the data analyst tests multiple different definitions for "moderate drinking" (e.g., 5 drinks per week, 1 drink per day, 1-2 drinks per day, etc.)
Multiple time points for the outcome (repeated measures)	a study where a walking test is administered at 1 months, 3 months, 6 months, and 1 year
Multiple looks at the data during sequential interim monitoring	a 2-year randomized trial where the efficacy of the treatment is evaluated by a Data Safety and Monitoring Board at 6 months, 1 year, and 18 months

Correction for Multiple Comparisons

- Major research issue for biostatisticians…debate as to the best approach

- Two ways to control for multiple testing:

- Controlling the Family-Wise Error Rate (FWER)
  - Traditional methods for controlling for multiple testing such as Bonferroni correction (α/n)
    - may be too conservative (↑ false negatives)

- Controlling the False Discovery Rate (FDR)
  - False discovery rate (FDR; Benjamin-Hochberg Test) now more commonly applied to 'omics' data sets

→multiple testing correction is achieved by either adjusting the p-value, or by altering the cut-off value

Correction for Multiple Comparisons

	H_0 true	H_0 false	Total
H_0 rejected	V	S	R
H_0 accepted	U	T	p-R
Total	p_0	p-p_0	p

- What is the probability of at least one type I error? α
  - Family-wise error rate (FWER)= α = *p*(V>=1)
- Single step FWER α'=α/p → FWER <=α
- Stepwise approaches: sequentially compare the sorted P-values to a threshold that depends on their rank
  - Too stringent

Correction for Multiple Comparisons

	H_0 true	H_0 false	Total
H_0 rejected	V	S	R
H_0 accepted	U	T	p-R
Total	p_0	p-p_0	p

- Correlated predictors: if correlated X same features are partially tested many times
- p models but less than p independent tests
- Resample techniques
- Effective Number of Tests (ENT)
  - the number of independent tests that would be required to obtain the same significance level using Bonferroni

J Proteome Res. Author manuscript; available in PMC 2011 Sep 3.
Published in final edited form as:
J Proteome Res. 2010 Sep 3; 9(9): 4620–4627.
doi: 10.1021/pr1003449

PMCID: PMC2941198
NIHMSID: NIHMS225485
PMID: 20701291

Metabolic Profiling And The Metabolome-Wide Association Study: Significance Level For Biomarker Identification

Marc Chadeau-Hyam,[†#] Timothy M D Ebbels,[‡#] Ian J Brown,[†] Queenie Chan,[†] Jeremiah Stamler,[¶] Chiang Ching Huang,[¶] Martha L Daviglus,[¶] Hirotsugu Ueshima,[§] Liancheng Zhao,[||] Elaine Holmes,[‡⊥] Jeremy K Nicholson,[‡⊥] Paul Elliott,[*†⊥] and Maria De Iorio[*†]

Author information ▶ Copyright and License information ▶ Disclaimer

Abstract

Go to:

High throughput metabolic profiling via the metabolome-wide association study (MWAS) is a powerful new approach to identify biomarkers of disease risk, but there are methodological challenges: high

J Proteome Res. Author manuscript; available in PMC 2011 Sep 3.
Published in final edited form as:
J Proteome Res. 2010 Sep 3; 9(9): 4620–4627.
doi: 10.1021/pr1003449

PMCID: PMC2941198
NIHMSID: NIHMS225485
PMID: 20701291

Metabolic Profiling And The Metabolome-Wide Association Study: Significance Level For Biomarker Identification

Marc Chadeau-Hyam,[†#] Timothy M D Ebbels,[‡#] Ian J Brown,[†] Queenie Chan,[†] Jeremiah Stamler,[¶] Chiang Ching Huang,[¶] Martha L Daviglus,[¶] Hirotsugu Ueshima,[§] Liancheng Zhao,[||] Elaine Holmes,[‡⊥] Jeremy K Nicholson,[‡⊥] Paul Elliott,[*†⊥] and Maria De Iorio[*†]

Author information ▶ Copyright and License information ▶ Disclaimer

Abstract

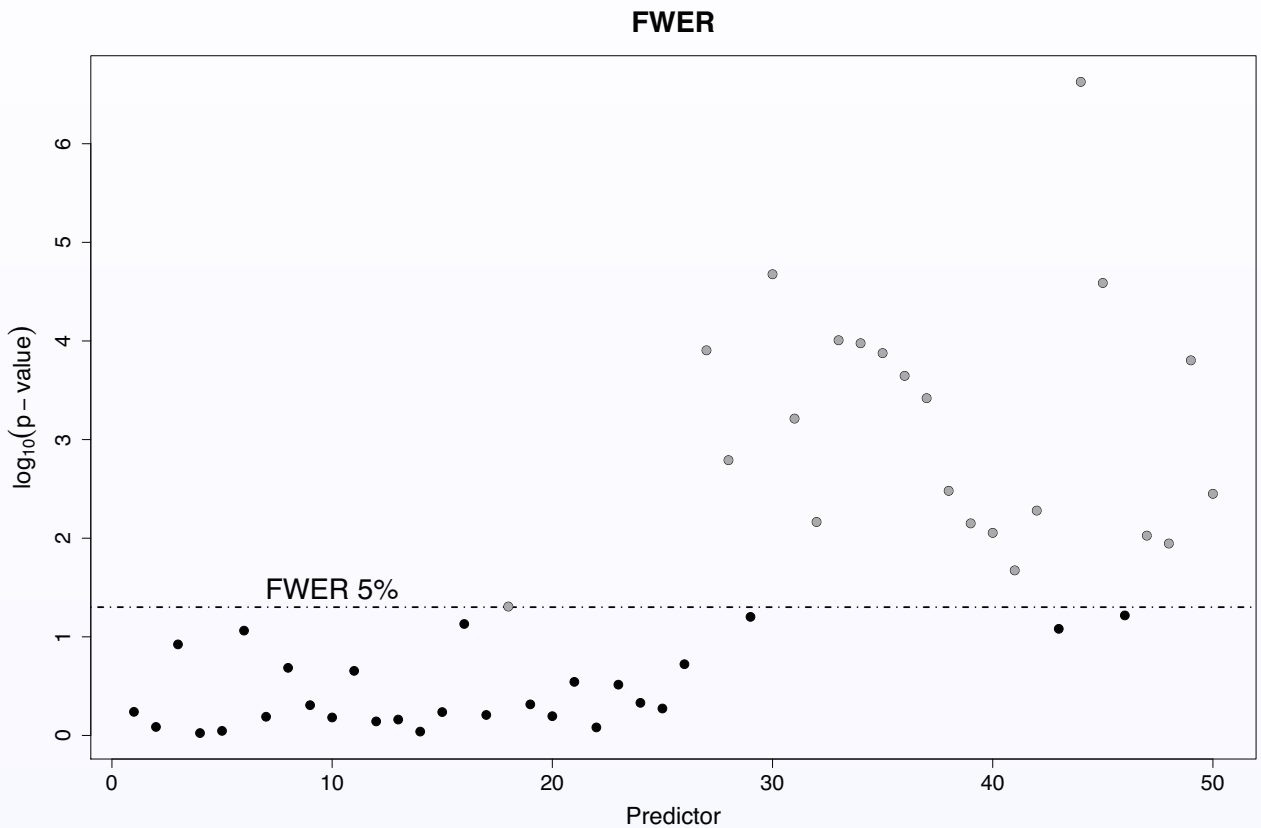
Go to:

High throughput metabolic profiling via the metabolome-wide association study (MWAS) is a powerful new approach to identify biomarkers of disease risk, but there are methodological challenges: high

500 cases and equal number of controls, assuming 7,100 spectral variables, the metabolome-wide significance level was estimated at $P = 2 \times 10^{-5}$ ($\alpha = 5\%$), resulting in a 60% reduction in the effective number of tests compared with Bonferonni correction

Outcome of p=50 tests: a list of 50 p-values

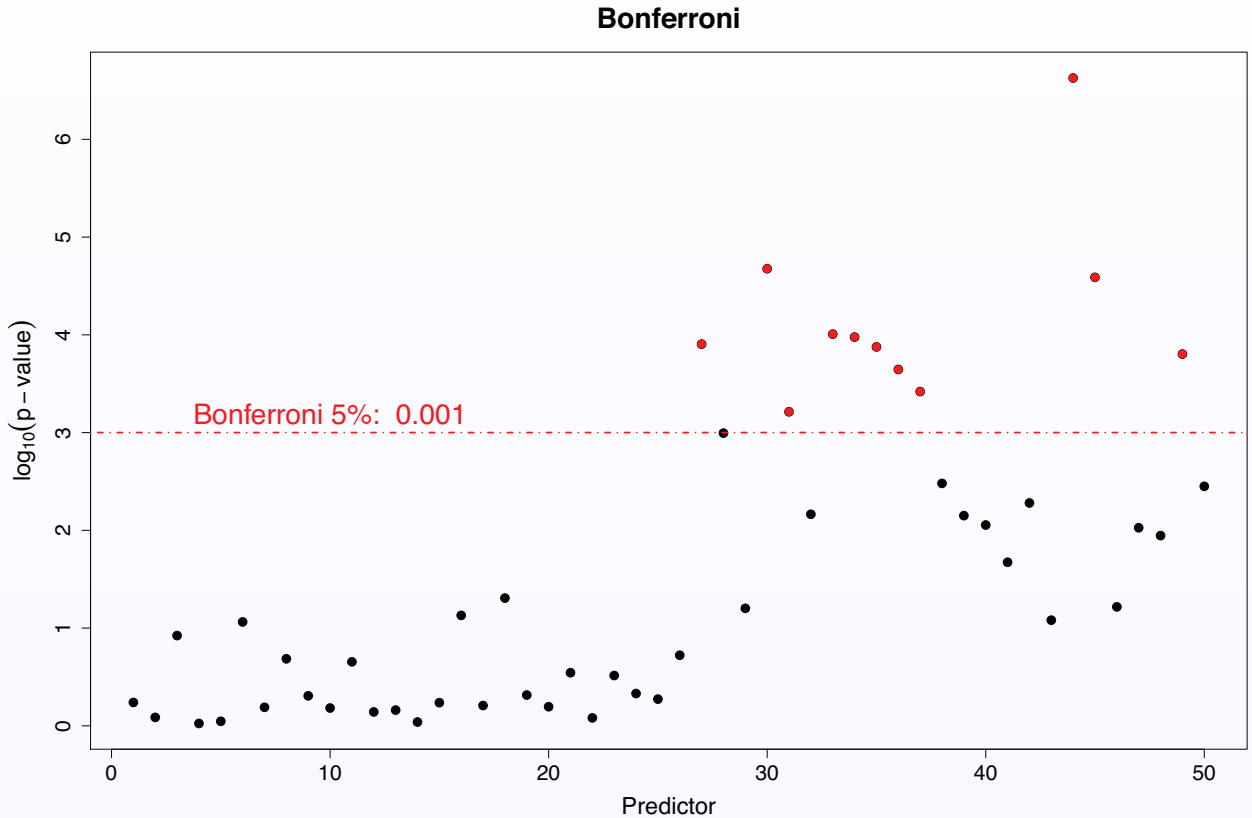
- 22 predictors have a p-value<0.05 (expected average #FP=2.5)



$\Rightarrow$ which of these 22 are likely TP?

Outcome of p=50 tests: a list of 50 p-values

- Bonferroni correction: $\alpha' = 0.05/50 = 0.001$



$\Rightarrow$ 11 predictors are associated (FWER 5%)

False Discovery Rate (FDR)

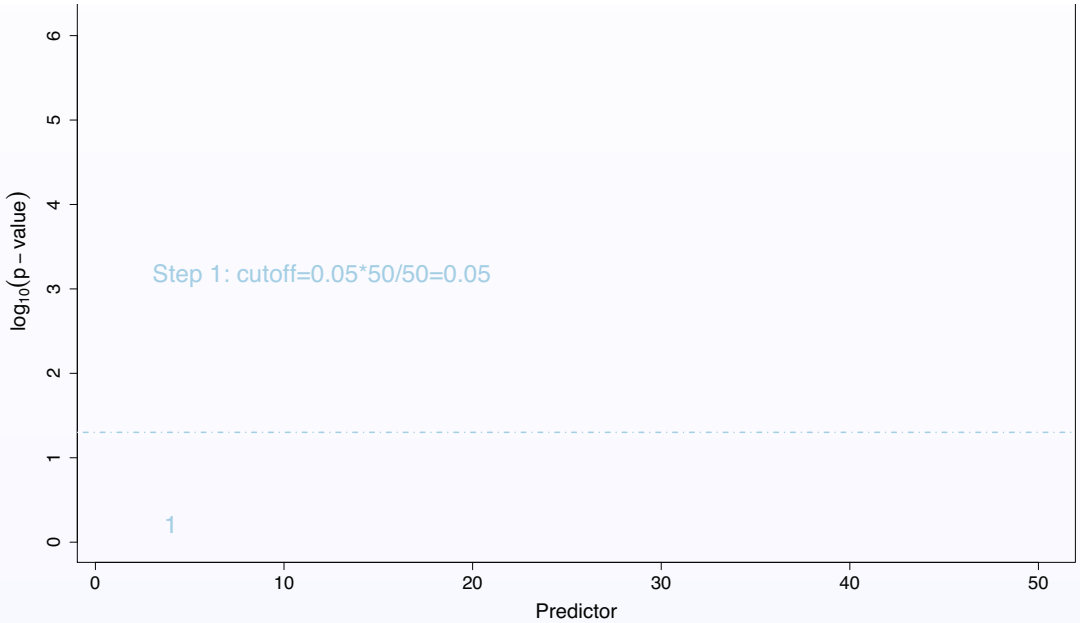
	H_0 true	H_0 false	Total
H_0 rejected	V	S	R
H_0 accepted	U	T	p-R
Total	p_0	p-p_0	p

$$Q = \frac{\#\text{ of false discoveries}}{\#\text{ of discoveries}} = \frac{V}{R}$$

- Q is set to be 0 when R=0
- FDR = expectation of Q = E(V/R; R>0)
- Benjamin-Hochberg Test (rank all P-values)
- FDR is less stringent than FWER
  - FWER control at 5% ensures that over 100 experiments <5 contain one FP
  - FDR control: over the 100 experiments the average #FP ≤ 5
- ⇒ FDR control may be preferred in an exploratory context

Benjamin-Hochberg FDR

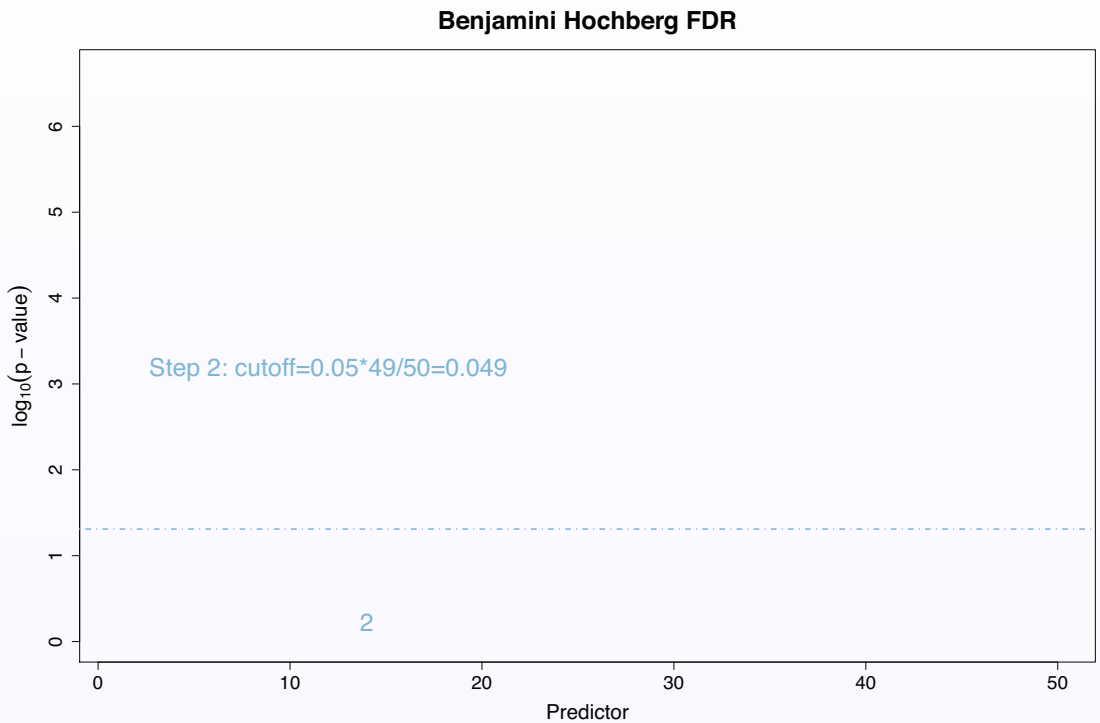
Order p values – start with max
Calculate critical value α(k/p)
Find largest p value that is smaller than critical value



$\Rightarrow$ k=1 largest p-value is not significant at 0.05, update the cut-off

Benjamin-Hochberg FDR

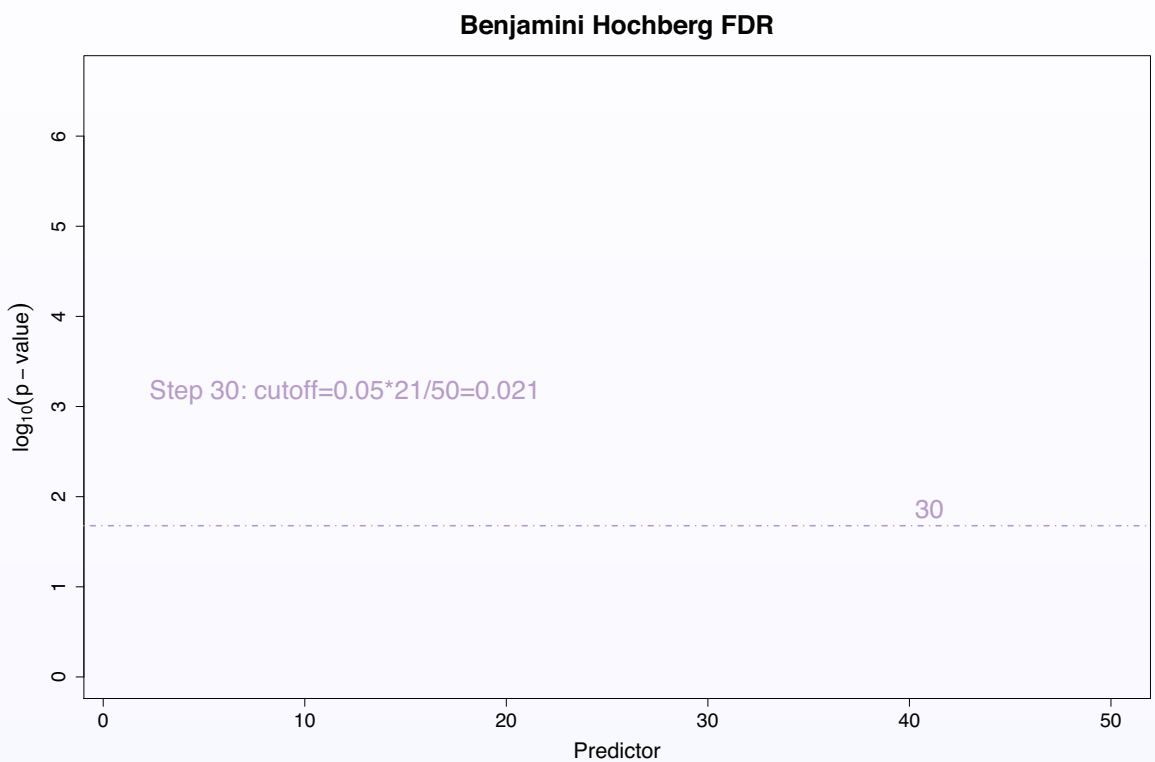
Order p values – start with max
Calculate critical value α(k/p)
Find largest p value that is smaller than critical value



$\Rightarrow$ k=2 is not significant set k=3, update the cut-off

Benjamin-Hochberg FDR

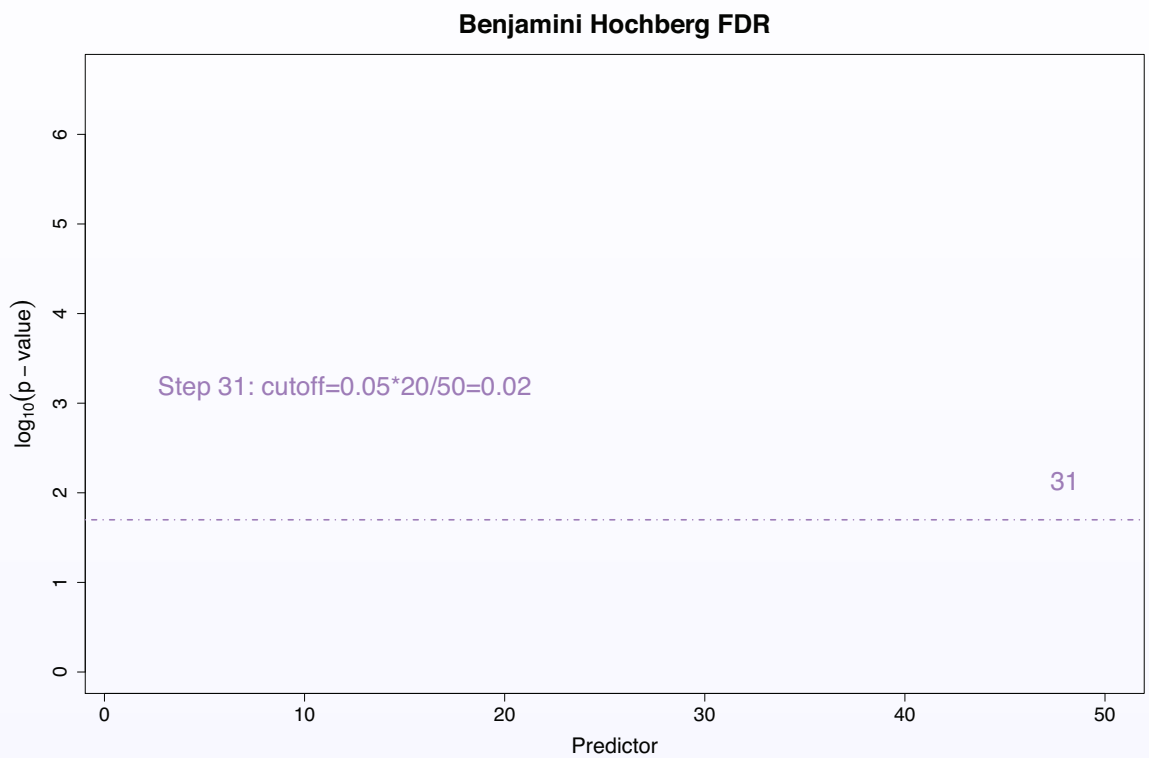
Order p values – start with max
Calculate critical value α(k/p)
Find largest p value that is smaller than critical value



$\Rightarrow k$ getting close to significance

Benjamin-Hochberg FDR

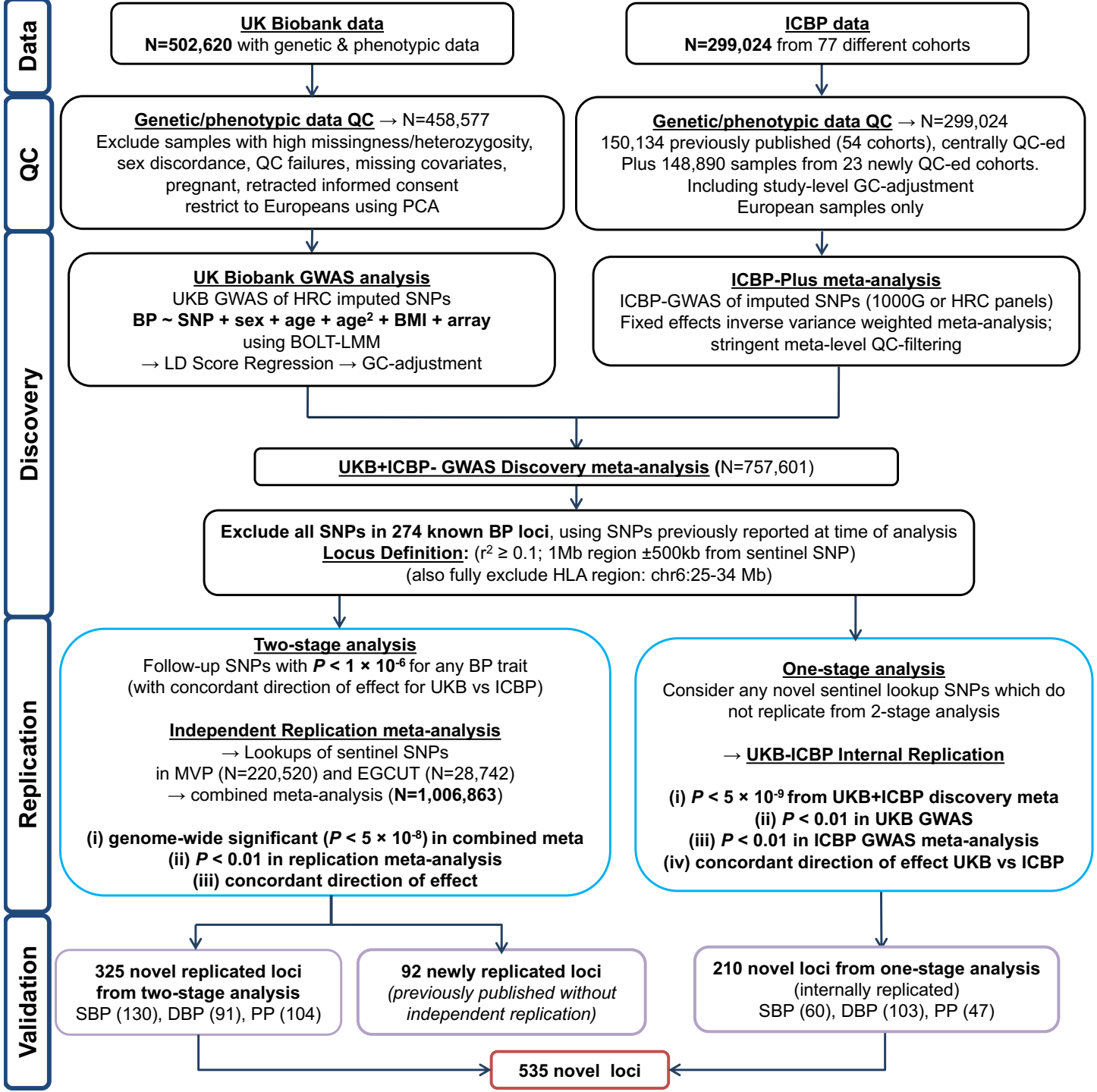
Order p values – start with max
Calculate critical value α(k/p)
Find largest p value that is smaller than critical value

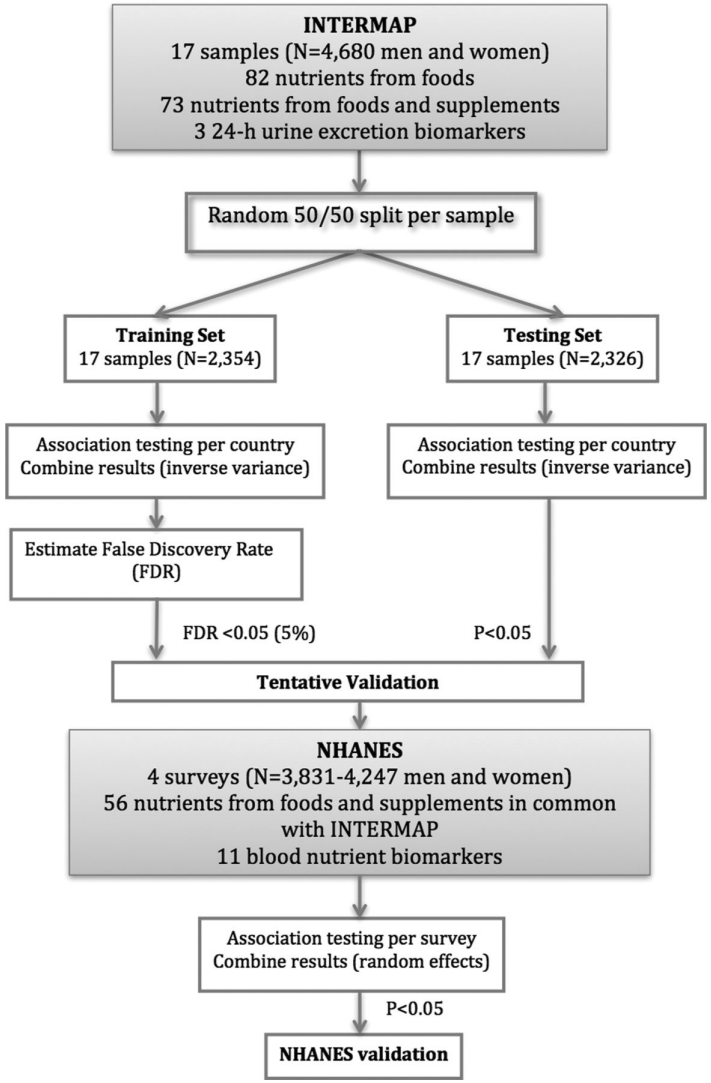


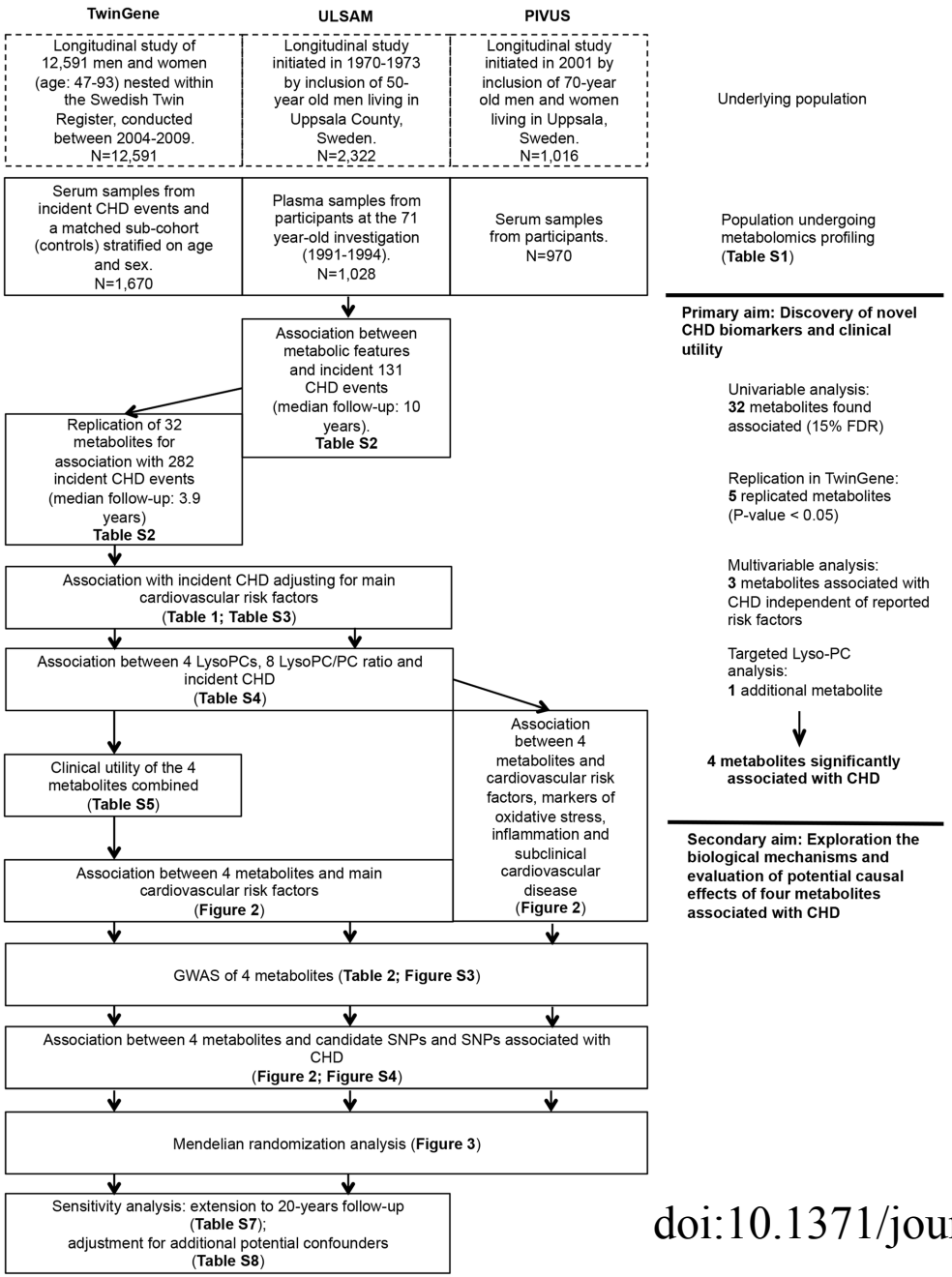
$\Rightarrow$ k=31 significant; the 30 1st assoc. are NOT signif. with FDR<5%

Two-Stage Study Designs

- Most widely accepted designs for ‘omics’ studies now….
- Discovery (training set): large sample size, identify discoveries (FDR)
- Validation (test set): independent from discovery set


Data
QC
Discovery
Replication
Validation
UK Biobank data
N=502,620 with genetic & phenotypic data
ICBP data
N=299,024 from 77 different cohorts
Genetic/phenotypic data QC → N=458,577
Exclude samples with high missingness/heterozygosity, sex discordance, QC failures, missing covariates, pregnant, retracted informed consent
restrict to Europeans using PCA
Genetic/phenotypic data QC → N=299,024
150,134 previously published (54 cohorts), centrally QC-ed
Plus 148,890 samples from 23 newly QC-ed cohorts.
Including study-level GC-adjustment
European samples only
UK Biobank GWAS analysis
UKB GWAS of HRC imputed SNPs
BP ~ SNP + sex + age + age² + BMI + array
using BOLT-LMM
→ LD Score Regression → GC-adjustment
ICBP-Plus meta-analysis
ICBP-GWAS of imputed SNPs (1000G or HRC panels)
Fixed effects inverse variance weighted meta-analysis;
stringent meta-level QC-filtering
UKB+ICBP- GWAS Discovery meta-analysis (N=757,601)
Exclude all SNPs in 274 known BP loci, using SNPs previously reported at time of analysis
Locus Definition: (r² ≥ 0.1; 1Mb region ±500kb from sentinel SNP)
(also fully exclude HLA region: chr6:25-34 Mb)
Two-stage analysis
Follow-up SNPs with P < 1 × 10⁻⁶ for any BP trait
(with concordant direction of effect for UKB vs ICBP)
Independent Replication meta-analysis
→ Lookups of sentinel SNPs
in MVP (N=220,520) and EGCUT (N=28,742)
→ combined meta-analysis (N=1,006,863)
(i) genome-wide significant (P < 5 × 10⁻⁸) in combined meta
(ii) P < 0.01 in replication meta-analysis
(iii) concordant direction of effect
One-stage analysis
Consider any novel sentinel lookup SNPs which do not replicate from 2-stage analysis
→ UKB-ICBP Internal Replication
(i) P < 5 × 10⁻⁹ from UKB+ICBP discovery meta
(ii) P < 0.01 in UKB GWAS
(iii) P < 0.01 in ICBP GWAS meta-analysis
(iv) concordant direction of effect UKB vs ICBP
325 novel replicated loci from two-stage analysis
SBP (130), DBP (91), PP (104)
92 newly replicated loci
(previously published without independent replication)
210 novel loci from one-stage analysis
(internally replicated)
SBP (60), DBP (103), PP (47)
535 novel loci


INTERMAP
17 samples (N=4,680 men and women)
82 nutrients from foods
73 nutrients from foods and supplements
3 24-h urine excretion biomarkers
Random 50/50 split per sample
Training Set
17 samples (N=2,354)
Testing Set
17 samples (N=2,326)
Association testing per country
Combine results (inverse variance)
Association testing per country
Combine results (inverse variance)
Estimate False Discovery Rate
(FDR)
FDR <0.05 (5%)
P<0.05
Tentative Validation
NHANES
4 surveys (N=3,831-4,247 men and women)
56 nutrients from foods and supplements in common
with INTERMAP
11 blood nutrient biomarkers
Association testing per survey
Combine results (random effects)
P<0.05
NHANES validation



doi:10.1371/journal.pgen.1004801.g001

Working Solutions

- Data analysis in ‘omics’ studies is challenging….
- Control for multiple testing is a necessity
- The Gold Standard is biological replication
- Training Sets and test sets should have no members in common
- Set up design as rigorously as possible (in advance)
- Training sets are proof of principle
- Test sets are, theoretically, validation

Factors to consider in evaluating (molecular) epidemiologic data

1. Analyses are exploratory.	The authors have mined the data for associations rather than testing a limited number of *a priori* hypotheses.
2. Many tests have been performed, but only a few p-values are "significant".	If there are no associations present, .05*k significant p-values (p<.05) are expected to arise just by chance, where k is the number of tests run.
3. The "significant" p-values are modest in size.	The closer a p-value is to .05, the more likely it is a chance finding. According to one estimate*, about 1 in 2 p-values <.05 is a false positive, 1 in 6 p-values <.01 is a false positive, and 1 in 56 p-values <.0001 is a false positive.
4. The pattern of effect sizes is inconsistent.	If the same association has been evaluated in multiple ways, an inconsistent pattern of effect sizes (e.g., risk ratios both above and below 1) is indicative of chance.
5. The p-values are not adjusted for multiple comparisons	Adjustment for multiple comparisons can help control the study-wide false positive rate.

Types of Validation for Biomarkers

- Analytical validation
  - When there is a gold standard
    - Sensitivity, specificity
  - No gold standard
    - Reproducibility and robustness
- Clinical validation
  - Does the biomarker predict what it's supposed to predict for independent data
- Clinical utility
  - Does use of the biomarker result in patient benefit
  - Depends on available treatments and practice standards

Univariate approaches

- Advantages
- Computational efficiency
- Model Flexibility
  - Generalized linear models
  - No need to model correlation structure in x
  - Adjustment for confounders easy
- Limitations
- Restricted to parametric marker outcomes relationship
- Models do not account for potential combined effects of X factors
  - Multivariate approaches

Multivariate approaches

- Dimension Reduction techniques:
  - Aim: Identify summary covariates (components) constructed as linear combinations of original variables which accurately reconstruct in a lower dimension the structure of the original data
  - Main approaches: unsupervised (e.g. PCA) and supervised (e.g. PLS-based approaches)
  - Main limitation: results may not guarantee easy interpretability $\Rightarrow$need to ensure sparsity of the results
- Variable selection techniques:
  - Aim: identify a sparse set of predictors that jointly predicts Y
  - Two main approaches: penalised regression (e.g. lasso approaches), and Bayesian Variable Selection approaches (BVS)
  - $\Rightarrow$variable selection approaches implicitly correct for multiple testing

M. Chadeau-Hyam et al. Deciphering the Complex: Methodological Overview of Statistical Models to Derive OMICS-Based Biomarkers. Environ Mol Mutagen, 2013 Aug;9(8).

Method Family	*Model*	*Outcome type*	*Available Implementation*	*Comment*
Univariate Approaches	Linear regression	*Continuous*	lm[1]	All linear models are special cases of generalised linear models. When running any of the GLM on multiple outcomes, results are equivalent to those obtained on each outcome independently
	Logistic regression	*Categorical/Binary*	glm[1]	
	Poisson regression	*Count data*		
	Linear mixed models	*Continuous*	lme4, nlme[2]	
	Generalized linear mixed models (GLM)	*Any kind of outcome (incl. survival data)*		
	Generalized additive models (GAM)	*Continuous/Categorical/Binary/Count data*	mgcv[2]	When running any of the GAM on multiple outcomes, results are equivalent to those obtained on each outcome independently. The package mgcv includes an L^2 (ridge) penalization capacity
	Generalized additive mixed models (GAMM)	*Continuous/Categorical /Binary/Count data*		
Dimension Reduction techniques	Principal Components Analysis (PCA)	*Continuous*	prcomp[2]	All dimension reduction techniques can accommodate multivariate outcomes. PCA and DAPC are unsupervised (i.e. do not account for the Y in constructing latent variables), CCA and all PLS-based approaches are supervised. StarPLS will soon be submitted to CRAN
	Discriminant Analysis (DA)	*Categorical/Binary*	lda[2]	
	Discriminant Analysis of Principal Components (DAPC)	*Continuous*	adegenet[2]	
	Partial Least Square (PLS)	*Continuous/ Categorical (and binary) for DA variants of the algorithms*	pls[2]	
	Canonical Correlation Analysis (CCA)	*Continuous*	CCA[2]	
	OPLS/O2PLS/OnPLS	*Continuous/ Categorical (and binary) for DA variants of the algorithms*	StarPLS[2]	
	Penalized (sparse) dimension regression methods (sPCA, sPLS, sPLS-DA)	*Continuous/ Categorical (and binary) for DA variants of the algorithms*	mixOmics[2]	These implementations include non-penalized versions as a special case
Regularization and Variable Selection	Ridge regression	*Continuous/Categorical /Binary*	lm.ridge[3], ridge[2]	ridge package adds logistic regression, as well as automatic selection of the penalty parameter
	Lasso/Elastic net regression	*Any kind of outcome*	glmnet[2]	Latest implementation of lasso methods accommodate multivariate outcomes. Variants of the lasso approach (e.g. bolasso, fused lasso, …) are implemented in separate packages
	Shotgun Stochastic search (SSS)	Continuous/Categorical/Binary	C++ stand-alone application	SSS and piMASS can accommodate any quantitative outcome. They differ in their prior specifications and search algorithms. Neither can handle multivariate outcomes
	piMASS	Continuous/Categorical/Binary	C++ stand-alone application	
	Evolutionary Stochastic Search (ESS/GUESS)	Continuous outcome	R2GUESS[2]	Accommodate multivariate outcomes
(Differential) Network models	(Shrinkage) Correlation Network	Continuous/Qualitative	GeneNet[2]	Can accommodate continuous outcomes

[1]Function included in the stats package; [2]R package available on CRAN; [3]Function included in the MASS package;

OPEN ACCESS Freely available online

PLOS | GENETICS

Large-scale Metabolomic Profiling Identifies Novel Biomarkers for Incident Coronary Heart Disease

Andrea Ganna[1], Samira Salihovic[2], Johan Sundström[2], Corey D. Broeckling[3], Åsa K. Hedman[4], Patrik K. E. Magnusson[1], Nancy L. Pedersen[1], Anders Larsson[5], Agneta Siegbahn[6], Mihkel Zilmer[7], Jessica Prenni[3,8], Johan Ärnlöv[4,9], Lars Lind[2], Tove Fall[4], Erik Ingelsson[4,10]*

1 Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden, **2** Department of Medical Sciences, Cardiovascular Epidemiology, Uppsala University, Uppsala, Sweden, **3** Proteomics and Metabolomics Facility, Colorado State University, Fort Collins, Colorado, United States of America, **4** Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden, **5** Department of Medical Sciences, Biochemial structure and function, Uppsala University, Uppsala, Sweden, **6** Department of Medical Sciences, Coagulation and inflammation science, Uppsala University, Uppsala, Sweden, **7** Department of Biochemistry, Centre of Excellence for Translational Medicine, University of Tartu, Tartu, Estonia, **8** Department of Biochemistry and Molecular Biology, Colorado State University, Fort Collins, Colorado, United States of America, **9** School of Health and Social Studies, Dalarna University, Falun, Sweden, **10** Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom

Abstract

Analyses of circulating metabolites in large prospective epidemiological studies could lead to improved prediction and better biological understanding of coronary heart disease (CHD). We performed a mass spectrometry-based non-targeted metabolomics study for association with incident CHD events in 1,028 individuals (131 events; 10 y. median follow-up) with validation in 1,670 individuals (282 events; 3.9 y. median follow-up). Four metabolites were replicated and independent of main cardiovascular risk factors [lysophosphatidylcholine 18:1 (hazard ratio [HR] per standard deviation [SD] increment = 0.77, P-value $<$ 0.001), lysophosphatidylcholine 18:2 (HR = 0.81, P-value $<$ 0.001), monoglyceride 18:2 (MG 18:2; HR = 1.18, P-value = 0.011) and sphingomyelin 28:1 (HR = 0.85, P-value = 0.015)]. Together they contributed to moderate improvements in discrimination and re-classification in addition to traditional risk factors (C-statistic: 0.76 *vs.* 0.75; NRI: 9.2%). MG 18:2 was associated with CHD independently of triglycerides. Lysophosphatidylcholines were negatively associated with body mass index, C-reactive protein and with less evidence of subclinical cardiovascular disease in additional 970 participants; a reverse pattern was observed for MG 18:2. MG 18:2 showed an enrichment (P-value = 0.002) of significant associations with CHD-associated SNPs (P-value = 1.2×10^{-7} for association with rs964184 in the *ZNF259/APOA5* region) and a weak, but positive causal effect (odds ratio = 1.05 per SD increment in MG 18:2, P-value = 0.05) on CHD, as suggested by Mendelian randomization analysis. In conclusion, we identified four lipid-related metabolites with evidence for clinical utility, as well as a causal role in CHD development.