HELMHOLTZ MUNICH

Meta-analysis strategies in genome-wide association studies

Ozvan Bocher
June 13, 2023

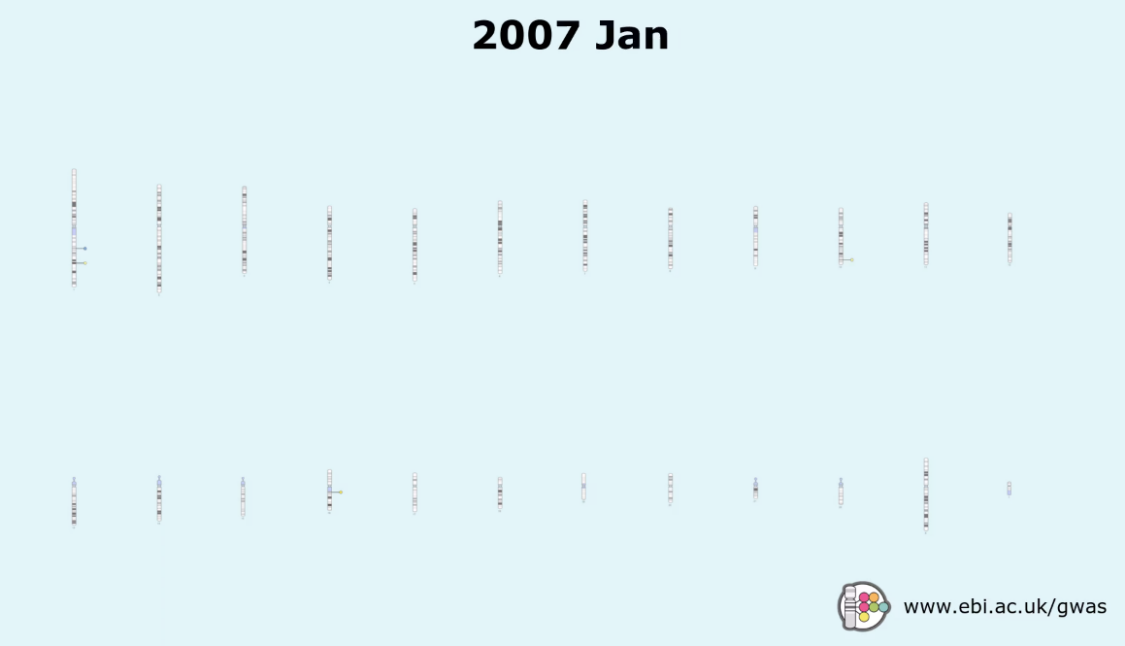
Agenda

1. Introduction
2. Principles of meta-analysis
3. Meta-analysis approaches
4. Considerations
5. Meta-analysis results

1

Introduction

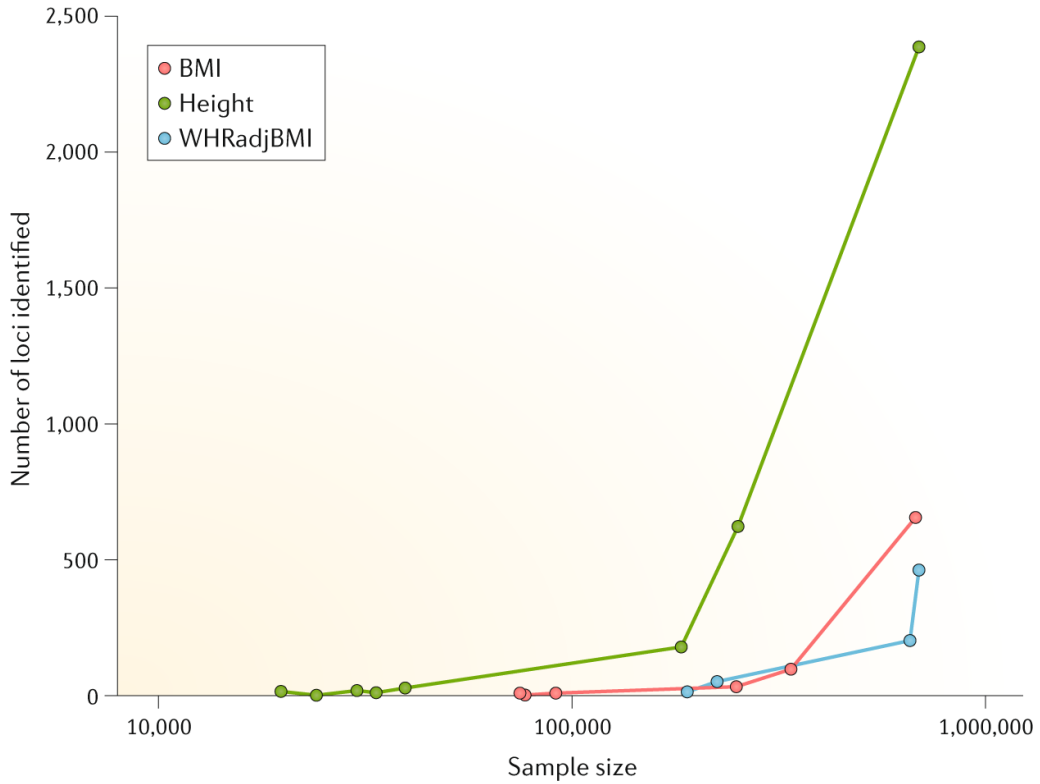
Genome-wide association studies

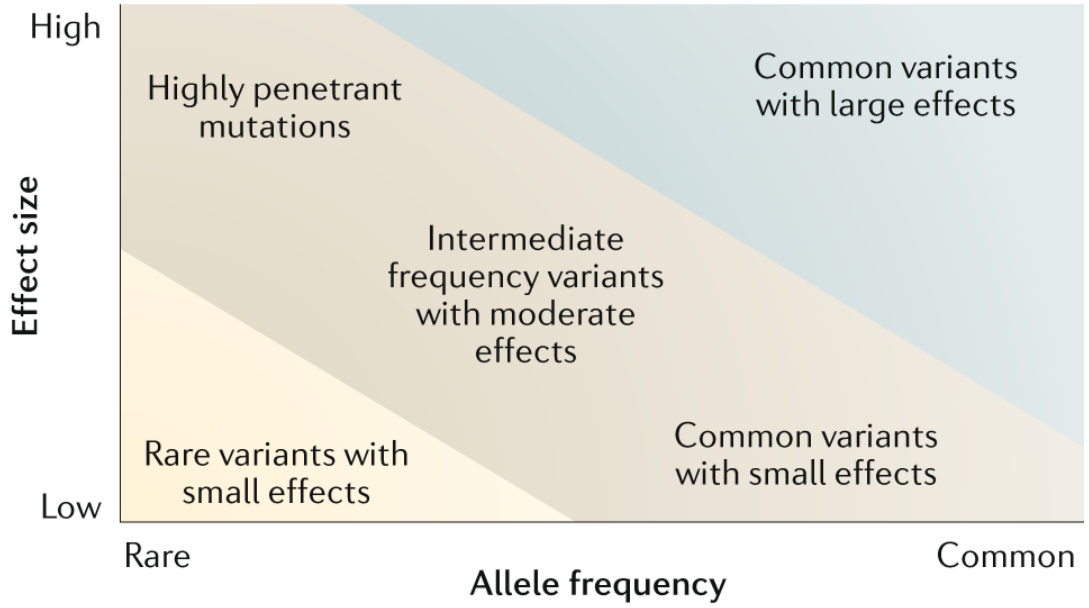


HELMHOLTZ MUNICH

https://www.ebi.ac.uk/gwas/

Motivation for meta-analysis





HELMHOLTZ MUNICH

Tam et al. 2019. Nat Rev Genet

Motivation for meta-analysis

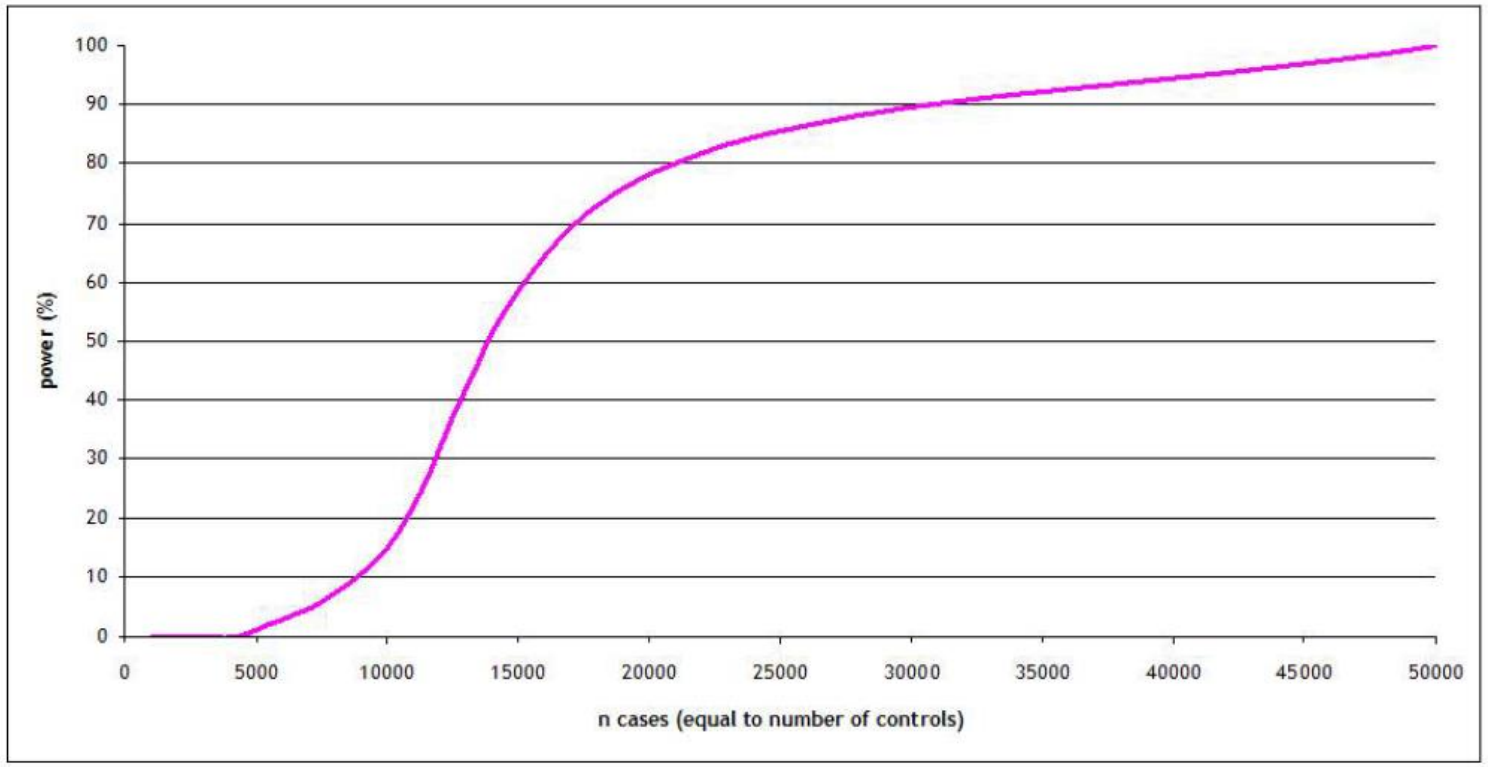
- Gather diverse studies on the same trait in a single framework
- Expected increase of power by:

  → Increased sample sizes

  → Imputation of untyped variants

Motivation for meta-analysis

Power to detect association ($p = 5 \times 10^{-8}$) at a variant with risk allele frequency 0.30 and allelic OR 1.10



Motivation for meta-analysis

- Increased sample size → access to low-frequency variants (1-5%) and rare variants (<1%)
- Issue of sample size and power more pronounced for rare variants

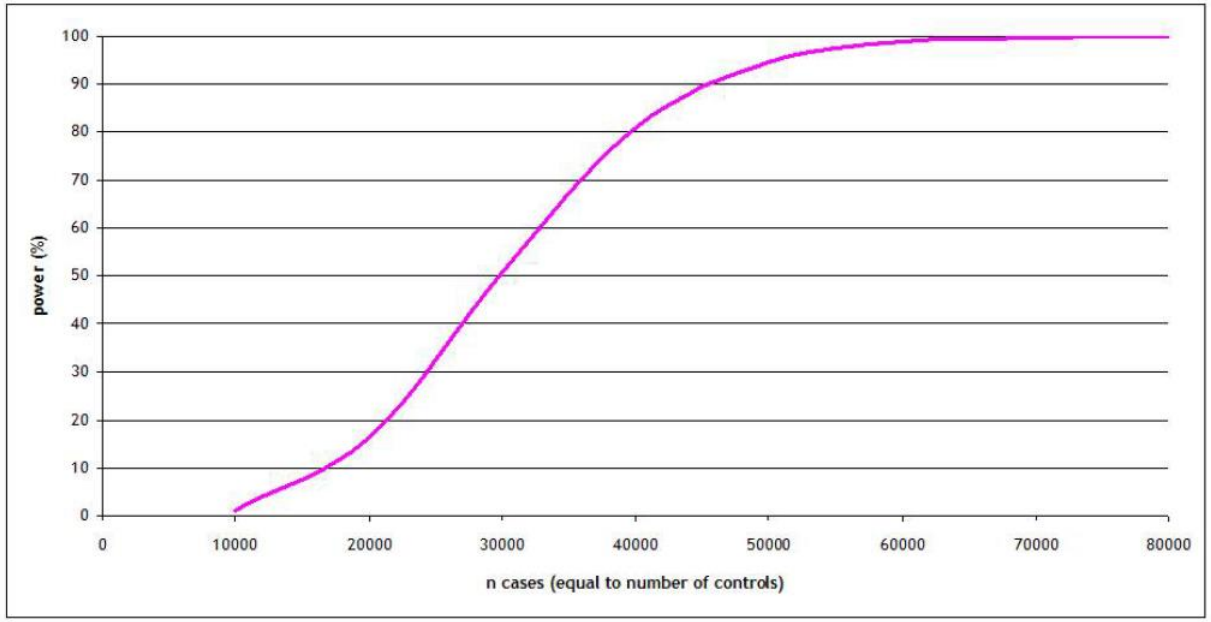
  → Don’t expect to have high-effect variants in polygenic disorders
  → Rare variant association tests

Motivation for meta-analysis

- Increased sample size → access to low-frequency variants (1-5%) and rare variants (<5%)
- Issue of sample size and power more pronounced for rare variants

  → Don't expect to have high-effect variants in polygenic disorders
  → Rare variant association tests

Power to detect association ($p = 5 \times 10^{-8}$) at a variant with risk allele frequency 0.005 and allelic OR 1.50



2

Principles of meta-analysis

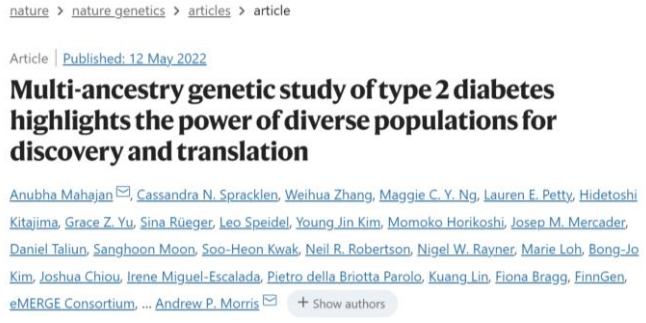
Motivation for meta-analysis

- Summary based on evidence from a combined dataset
- Detect variants with moderate and small effect sizes
- Can be carried out sequentially and updated when new GWAS from the same trait emerge

Motivation for meta-analysis

- Summary based on evidence from a combined dataset
- Detect variants with moderate and small effect sizes
- Can be carried out sequentially and updated when new GWAS from the same trait emerge

32 GWAS
74,124 T2D cases,
824,006 controls

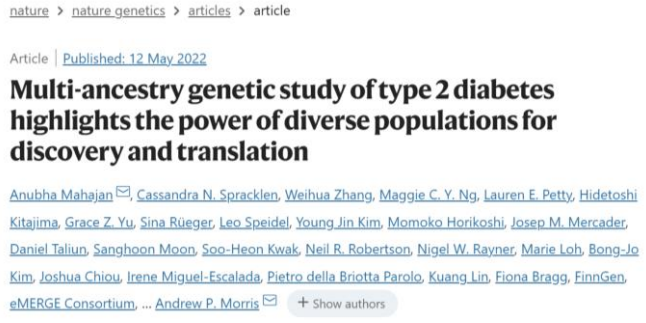


121 GWAS
180,834 T2D cases,
1,159,055 controls

Motivation for meta-analysis

- Summary based on evidence from a combined dataset
- Detect variants with moderate and small effect sizes
- Can be carried out sequentially and updated when new GWAS from the same trait emerge

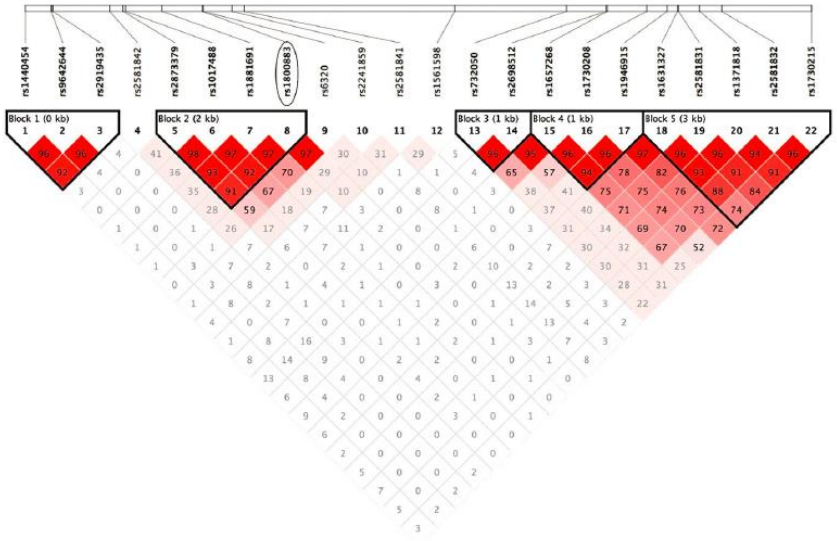
32 GWAS
74,124 T2D cases,
824,006 controls



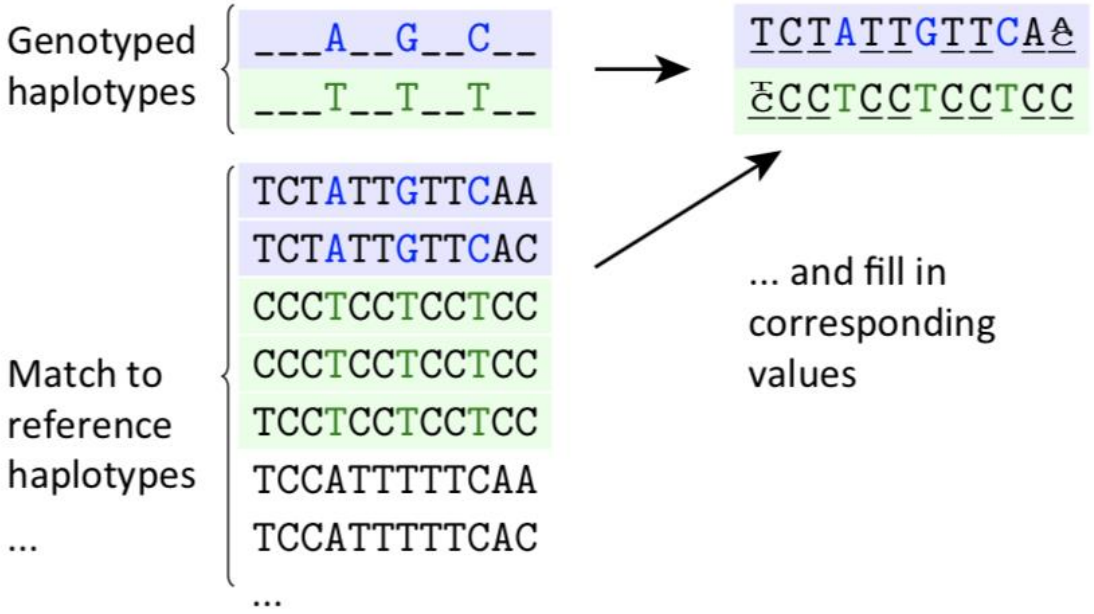
121 GWAS
180,834 T2D cases,
1,159,055 controls

- Facilitated by imputation which enables the combination of data across different genotyping platforms

Imputation



- Goal = get genotypes at untyped positions in the target dataset
- Based on the linkage disequilibrium (LD) between variants
  - “Haplotypes” = group of alleles inherited together
- Reference datasets such as www.1000genomes.org, www.uk10k.org and www.haplotype-reference-consortium.org



Consortia

- Consortia to study specific traits of interests and gather multiple GWAS

ENGAGE
European Network of Genomic and Genetic Epidemiology

MAGIC

GIANT CONSORTIUM

DIAGRAM consortium

DIAbetes Genetics
Replication And Meta-analysis

UKRAG
UK Rheumatoid Arthritis Genetics Consortium

nature › articles › article

Article | Open Access | Published: 12 October 2022

A saturated map of common genetic variants associated with human height

Loïc Yengo, Sailaja Vedantam, Eirini Marouli, Julia Sidorenko, Eric Bartell, Saori Sakaue, Marielisa Graff, Anders U. Eliasen, Yunxuan Jiang, Sridharan Raghavan, Jenkai Miao, Joshua D. Arias, Sarah E. Graham, Ronen E. Mukamel, Cassandra N. Spracklen, Xianyong Yin, Shyh-Huei Chen, Teresa Ferreira, Heather H. Highland, Yingjie Ji, Tugce Karaderi, Kuang Lin, Kreete Lüll, Deborah E. Malden, 23andMe Research Team, VA Million Veteran Program, DiscovEHR (DiscovEHR and MyCode Community Health Initiative), eMERGE (Electronic Medical Records and Genomics Network), Lifelines Cohort Study, The PRACTICAL Consortium, Understanding Society Scientific Group, … Joel N. Hirschhorn + Show authors

Nature **610**, 704–712 (2022) | Cite this article

45k Accesses | **2** Citations | **1350** Altmetric | Metrics

>12,000 independent SNPs

Meta-analysis

- Requires a robust predefined protocol with information on
  - Genetic model examined
  - Definition of inclusion and exclusion criteria
  - Strategies for covariate adjustment
  - Size of the study
  - Independence of samples
  - Strand and build of the human genome
  - Allele coding
- Majority of meta-analysis combine data **retrospectively**

  → Harmonization of study design is difficult
- Requires summary statistics for each variant

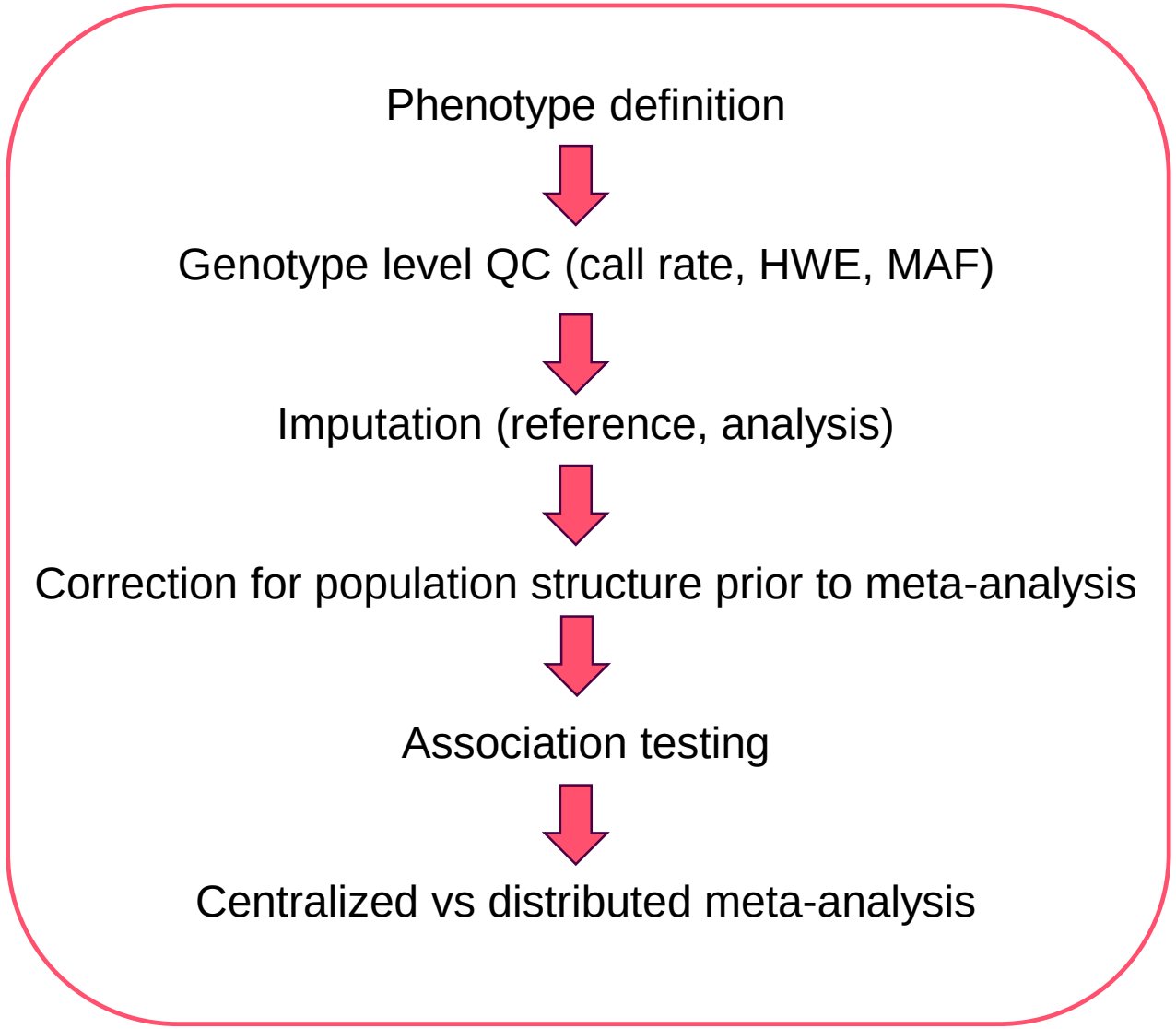
Typical data sharing table format

STUDY TITLE	
General information	
Name of study	
Name of analyst	
Email of analyst	
Study design	population-based, family-based –please give details
Sample information	
Number of cases (females)	
Number of controls (males)	
Ethnic composition	
Possible relatedness issues	are individuals related (how?)
Possible structure issues	mixed population?
Genotyping and imputation information	
Genotyping platform	
Summary of key QC metrics	
# SNPs passed QC	
Imputation method	
Imputation settings	
Reference data used for imputation	including build
Analytical information	
Association analysis method for imputed genotypes	accounting for uncertainty using SNPTEST or other (which?) program, using only genotypes with P(call)>X (which threshold?) as hard calls, using best guess genotypes
Calculated GC lambda (typed SNPs)	
Calculated GC lambda (imputed SNPs)	
Covariates included	PCA, GC, none
Genetic model	

Typical data sharing table format

Column header	Description
SNP	SNP rs number (if unknown, e.g. with some Affymetrix SNPs, report Affy SNP ID)
build	e.g. "36", human genome build used
strand	e.g. "+", human genome strand used
chromosome	chromosome on which SNP resides
position	position of SNP on chromosome in base pairs, based on human genome build used
imputed	"1" for imputed, "0" for directly-typed SNP passing QC
major_allele	e.g. "G", major allele at that SNP, based on control frequency
minor_allele	e.g. "A", minor allele at that SNP, based on control frequency
MAF_controls	e.g. "0.246", minor allele frequency in controls -provide 3 digits to the right of the decimal
OR_allele	e.g. "A", allele to which the OR has been estimated
call_rate	e.g. "0.985", call rate for this SNP across cases and controls -provide 3 digits to the right of the decimal
exact_HWE_cases	exact HWE p value in cases
exact_HWE_controls	exact HWE p value in controls
OR	e.g. "1.097", allelic odds ratio –provide 3 digits to the right of the decimal
lower_95%CI	e.g. "0.874", lower 95% confidence interval of the OR –provide 3 digits to the right of the decimal
upper_95%CI	e.g. "1.267", upper 95% confidence interval of the OR –provide 3 digits to the right of the decimal
additive_p_uncorr	additive model p value, uncorrected for genomic control
additive_p_corr	additive model p value, corrected for genomic control
impute_acc	e.g. "0.98", metric for imputation accuracy (i.e. value for r^2hat or proper_info measures, depending on imputation programme used; if some other measure used, please specify)

Meta-analysis pipeline



3

Meta-analysis methods

Meta-analysis methods

- Fixed-effects
- Random-effects
- Bayesian
- Trans-ancestry
- Based on estimate of effect size (β or OR) or on p-value
    - Must have independent set of effect sizes
    - Larger studies should carry more weight

Pooling effect estimates

Phenotype	Analysis	Effect estimate
Case-control	Chi-square test	$OR=e^{\beta}$ $\beta=\ln(OR)$
Case-control	Logistic regression	$OR=e^{\beta}$ $\beta=\ln(OR)$
Quantitative trait	Linear regression	β

3

Meta-analysis methods

3.1

Fixed-effect

Fixed-effect meta-analysis

→ Fixed-effect model assumption: $\beta_1=\beta_2$ ($\pm\varepsilon$)

→ Higher weights to the largest studies

Fixed-effect meta-analysis

Inverse-variance based

- β_i: effect size estimate for study i
- se_i: standard error for study i

$$\boldsymbol{w_i = \frac{1}{se_i^2}}$$

$$SE = \sqrt{\frac{1}{\sum_i w_i}}$$

$$\beta = \frac{\sum_i (\beta_i \cdot w_i)}{\sum_i w_i}$$

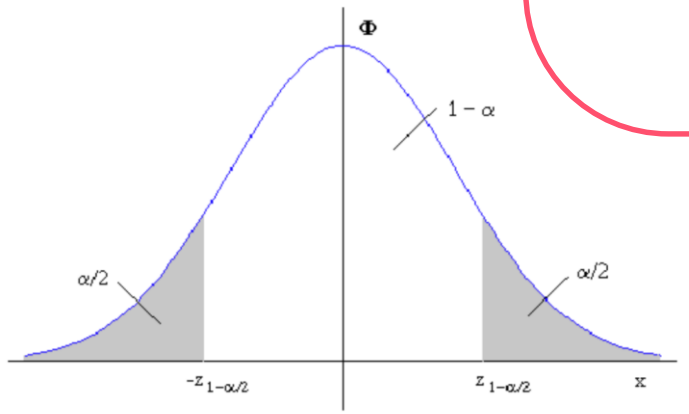
$$Z = \frac{\beta}{SE}$$

$$P = 2 * (1 - \phi(|Z|))$$

Sample size based

- N_i: Sample size for study i
- P_i: p-value for study i
- Δ_i: direction of effect for study i

- $Z_i = \phi^{-1}\left(\frac{P_i}{2}\right) \cdot sign(\Delta_i)$
- $\boldsymbol{w_i = \sqrt{N_i}}$
- $Z = \frac{\sum_i (Z_i \cdot w_i)}{\sum_i w_i^2}$
- $P = 2 * (1 - \phi(|Z|))$



Correcting for population structure

- Within study variation
  - Correct test statistics $\chi_i^2 = \left(\frac{\beta_i}{se_i}\right)^2$ by the genomic inflation factor λ $\left(\lambda = \frac{median(\chi_i^2)}{0.456}\right)$
  - Calculate λ separately for directly genotyped and imputed SNPs, λ_{Di} and λ_{D*i} for study i
  - Adjust weights to $w_i^{adj} = \lambda_{Ki} * w_i$, where K is replaced by D or D* as appropriate

- Between studies variation
  - $X^2 = Z^2 = \frac{\beta^2}{(\lambda * se^2)}$, where λ is the genomic control inflation factor over all meta-analyzed association test statistics

Assessment of heterogeneity

- Important step in meta-analysis = assessing heterogeneity across the combined studies

- Statistics:
  - Cochran's Q: is there heterogeneity ?

$$Q = \sum_{i=1}^{N} w_i(\beta_i - \beta_{pooled})^2$$

Q_j for $SNP\ j$ depends on the number of studies for which an allelic effect is reported
Small number of studies: low power

Assessment of heterogeneity

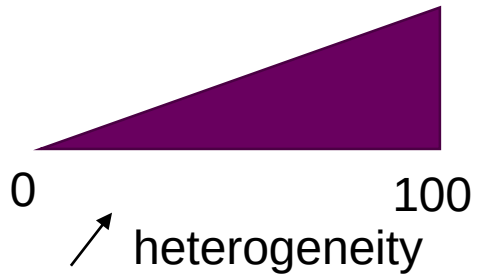
- Important step in meta-analysis = assessing heterogeneity across the combined studies

- Statistics:
  - Cochran's Q: is there heterogeneity ?

$$Q = \sum_{i=1}^{N} w_i(\beta_i - \beta_{pooled})^2$$

Q_j for $SNP\ j$ depends on the number of studies for which an allelic effect is reported
Small number of studies: low power

  - I^2: how much heterogeneity is there ?
    - I_j^2 is more robust to variability in the number of studies included in the meta-analysis for $SNP\ j$
    - $I_j^2 > 50\%$: Large heterogeneity
    - $I_j^2 > 75\%$: Very large heterogeneity



Potential causes for heterogeneity (bias)

- Study-specific bias:
  - Genotyping errors
  - Poor genotype data cleaning
  - Poor imputation
  - Different SNP platforms
  - Population stratification
  - Phenotype misclassification
- Winner’s curse: the originally identified effect size is likely to be overestimated in comparison to its true value

Potential causes for heterogeneity

- Variable LD patterns across studies

  The identified marker is not the causal polymorphism, but has a different LD pattern with the causal polymorphism across different studies.

- Gene-environment interactions with different environmental exposures across populations
- Genuine genetic heterogeneity in effect sizes across different ethnic backgrounds and population-specific effects

Interpreting heterogeneity

- Heterogeneity may represent genuine differences in genetic effects across different populations and different biological settings (**truly informative heterogeneity**)
- Informative heterogeneity may reveal interesting facts about biology, e.g. the mechanism through which the variant is acting on disease risk.
- Recognizing the potential for heterogeneity can rescue associations from being discarded as replication failures.

Interpreting heterogeneity

- Heterogeneity may represent genuine differences in genetic effects across different populations and different biological settings (**truly informative heterogeneity**)
- Informative heterogeneity may reveal interesting facts about biology, e.g. the mechanism through which the variant is acting on disease risk.
- Recognizing the potential for heterogeneity can rescue associations from being discarded as replication failures.
- Example of informative heterogeneity: relationship between the obesity associated (FTO) gene and T2D.
    - Overall the evidence suggests that FTO affects weight, which in turn increases risk of T2D.
    - As a consequence of being on the same pathway, FTO showed an association with T2D in population-based studies but failed to replicate in studies that controlled for weight by only recruiting lean subjects.

3

Meta-analysis methods

3.2

Random-effect

Random-effect models

→ No assumption on the relation between β_1 and β_2
→ Assumption that there is an underlying distribution of effects

A variance component τ^2 is used to inflate the variance of the estimated allelic effect in each study

Random-effect models

- Used if large differences across samples
- Same scale is used across samples
- Number of samples should be sufficiently large

$$\tau^2 = \frac{\boldsymbol{Q} - (k-1)}{\sum_{i-1}^{N} w_i - \frac{\sum w_i^2}{\sum w_i}}$$

$$w_i = \frac{1}{se_i^2} \rightarrow w_i^* = \frac{1}{(\tau^2 + se^2)}$$

- Random effect models are more conservative (larger se)

Specialized software for GWAS meta-analysis

GWAMA
Genome-Wide Association Meta-Analysis
www.well.ox.ac.uk/gwama/contact.shtml

METAL
Meta-Analysis Helper
www.sph.umich.edu/csg/abecasis/metal

META
www.stats.ox.ac.uk/ jsliu/meta.html

MetABEL
R package part of GenABEL, an R library for GWAS analysis
www.genabel.org/packages/MetABEL

METASOFT
Han and Eskin's Random Effects model, Binary Effects model
http://genetics.cs.ucla.edu/meta
Ioanna Tachmazidou

Comparison of software

Software package	METAL	MetABEL	META	GWAMA
Pre-processing of GWA files	No	ABEL	SNPTEST	SNPTEST,PLINK
Strand flipping	Yes	Yes	Yes	Yes
Fixed effect analysis	Yes	Yes	Yes	Yes
Random effect analysis	No	No	Yes	Yes
Heterogeneity statistics	Q	No	Q, I^2	Q, I^2
Genomic control	Yes	Yes	Yes	Yes
Graphical visualization	No	Forest plot	No	QQ and Manhattan plots

Magi and Morris, BMC Bioinformatics, 2010

3

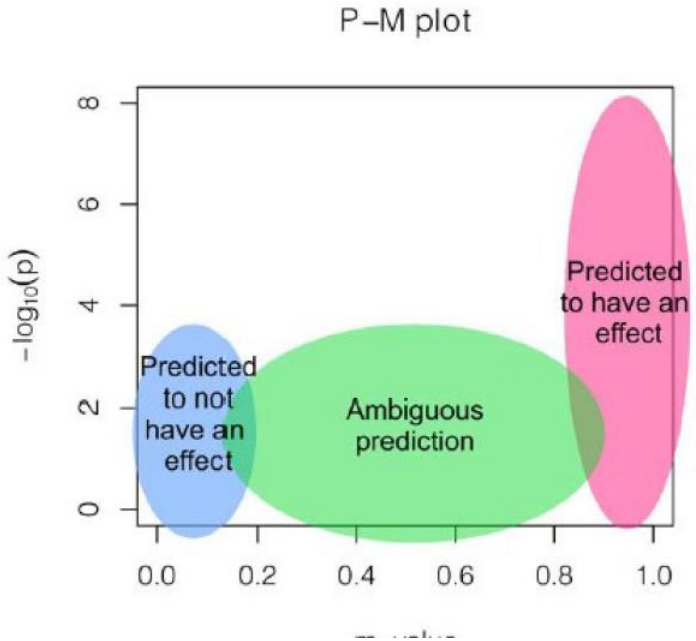
Meta-analysis methods

3.3

Bayesian

Bayesian meta-analysis: Binary effect model

- Based on the posterior probability that the effect exists in each study (m-value)
- Estimated using cross-study information via MCMC
- Segregates the studies predicted to have an effect, the studies predicted to not have an effect, and the underpowered ones



- If m-value > 0.9, the study is predicted to have an effect
- If m-value < 0.1, the study is predicted not to have an effect

- Binary effect test statistics

$$S_{BE} = \frac{\sum m_i \sqrt{W_i Z_i}}{\sqrt{\sum m_i^2 W_i}}$$

Where $Z_i = \frac{\beta_i}{se_i}$ and $\sqrt{W_i} = \sqrt{Ni}$

HELMHOLTZ MUNICH

Han and Eskin, PLoS Genetics 2012

3

Meta-analysis methods

3.4

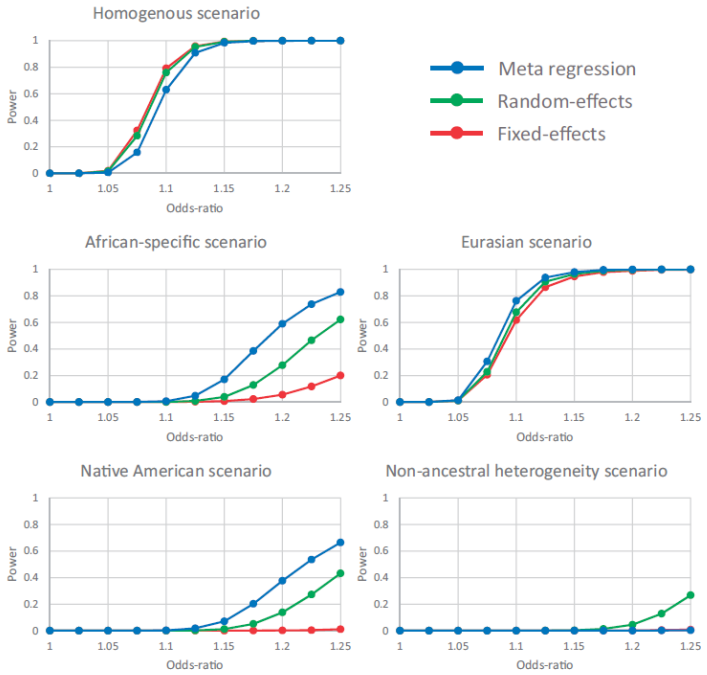
Trans-ancestry meta-analyses

Meta-regression

- Magi et al. 2017

  → Take into account heterogeneity related to ancestry
  → Integrates axes of genetic variations (PCs from PCA) in a linear regression framework

$$E(b_{kj}) = \alpha_j + \sum_{t=1}^{T} \beta_{tj} x_{kt}$$



4

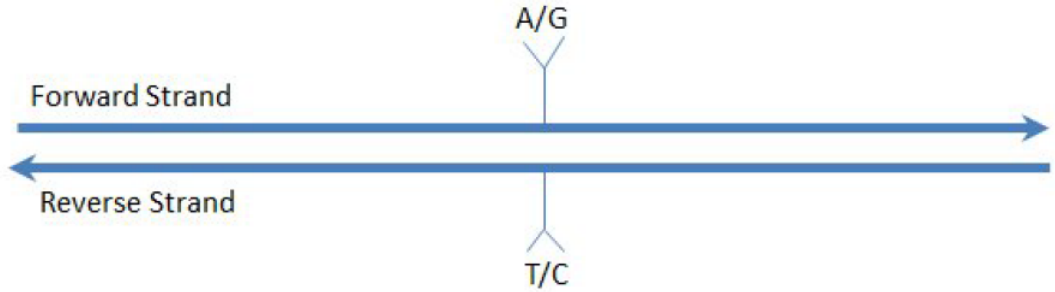
Considerations

Remapping genome build and strand flipping

- Ensuring all data are aligned to the same strand on the same genome build

  → Ensure the same risk allele across studies (effect sizes are associated to an allele)
  → Usually the forward or positive (+) strand

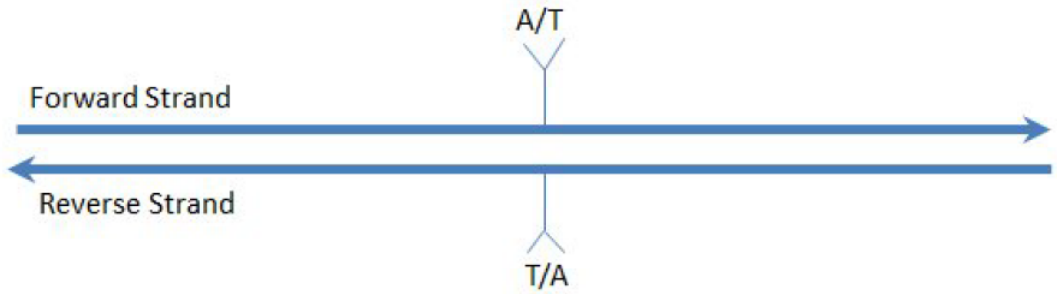
- For most SNPs, the strand can be identified by the alleles



- SNP is A/G on the forward strand and the T/C on the reverse, thereby uniquely identifying the strand

Alleles are not the full story

- For some SNPs, the strand cannot be determined using the alleles



- In this case, the alleles on either strand are A/T
- This is the same for G/C SNPs

Using allelic frequency to determine the strand

- Assume SNP is A/T, the Minor Allele is A with a frequency (MAF) of 30%.
- A second study with the SNP listed as Minor Allele T with a frequency of 32% is likely on the opposite strand.
- This is not conclusive as SNPs vary in frequency between populations.
- A/T or G/C SNPs with a frequency near 50% are particularly difficult to determine using frequency.

Study alignment and error trapping

Example of alignment of allelic effects and error trapping or a single SNP in a meta-analysis of five studies of a dichotomous phenotype

Study	Reported strand	Effect allele[1]	Other allele	RAF	Odds ratio (95% confidence interval)	Aligned allelic effect (standard error)	Comment
1	+	A	G	0.12	1.12 (1.07-1.16)	0.11 (0.02)	Allele A taken as reference effect allele.
2	+	G	A	0.85	0.92 (0.87-0.98)	0.08 (0.03)	Effect aligned to allele A.
3	-	T	C	0.12	1.06 (1.02-1.10)	0.06 (0.02)	Effect aligned to allele A on + strand.
4	+	T	C	0.13	1.07 (0.99-1.16)	0.07 (0.04)	Effect aligned to allele A on + strand. Strand error reported to log file.
5	+	A	G	0.87	0.95 (0.90-1.01)	-0.05 (0.03)	Large discrepancy in EAF reported to log file.

[1] Effects are aligned to the reference allele in the first study. Errors in the reported strand are recorded in the log file together with warnings regarding potential discrepancies in reported data between studies, for example the aligned reference allele frequency (RAF).

HELMHOLTZ MUNICH

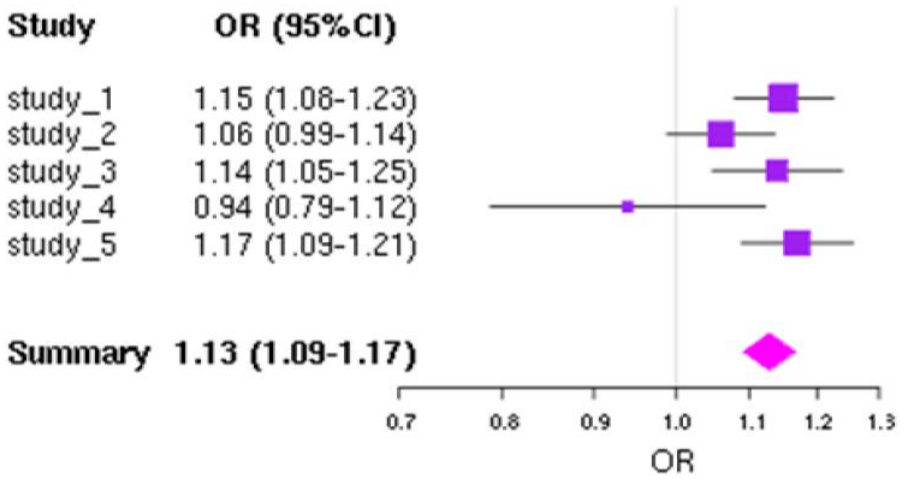
Magi and Morris, BMC Bioinformatics, 2010

5

Meta-analysis results

Visualization

- Classical tools: QQ-plots, Manhattan plots
- Forest plots: study-specific and overall effect



Interpretation

- No between-study heterogeneity:

  → Fixed effect = random effect calculations
  → Identical point estimates and confidence intervals

- Increasing between-study heterogeneity

  Random effects summary estimates have larger variance (wider confidence intervals)
  → Usually less prominent statistical significance

- Most meta-analysts would typically run both models

- In all situations: statistically significant associations need **replication**

Interpretation

- Effect sizes should be estimated with precision (if further used for predictive risk modelling - Lecture 5)
  - The analytical model used, e.g. genetic model specification, may affect the magnitude of the effect size
  - Winner's curse effect: significant associations are likely to have observed effect sizes that are inflated compared with the true effect size

- Is a newly discovered genetic variant the causal variant or simply linked to it?
  - Hard to answer
  - Even large-scale GWA meta-analyses require extensive **fine-mapping** and **targeted resequencing** experiments before the truly causal variants can be identified.

- Discovery of a genetic locus has important implications on its own

  → May highlight some interesting biological pathway and may give some insights into developing new therapeutics.

Replication

- Prioritize interesting signals for follow-up and replication.
- Follow-up sample sets should be adequately powered to detect the association.

  → The replication stage could be a large meta-analysis itself.
- Issues of set-up, information aggregation, estimation of heterogeneity and summary effects in a replication effort are similar to those described for discovery meta-analysis.
- Meta-analyze the discovery with the replication data to capture the totality of the evidence.
- Review the literature and bioinformatic databases to identify candidate variants both for the particular trait under study and for associated traits.
- Replication of previously published hits confirms previous publications and gives validity to the meta-analysis.

Stages of a meta-analysis

- Sensitivity analysis
  - Does the effect extend over a chromosomal region or is it confined to one variant ?
  - Do the results depend critically on a single study ? Why ?
  - Is the effect stronger under a recessive or dominant genetic model ?
  - Is there large heterogeneity at a locus ?
- Secondary analysis
  - Assess the importance of some variants adjusted for others, in order to see if the two sets act independently
  - Adjust for phenotypes that lie on potential causal pathways
  - Is the signal driven or stronger in males or females ?
  - Is the signal driven by some specific covariates ?

Summary

- Genetic effects of **common variants** associated with complex diseases are mostly **moderate/small** and require very large sample sizes to identify with certainty.
- Meta-analysis of genome-wide data can improve the power for detecting and validating such associations.
- Meta-analysis of data from studies using different platforms is enhanced by the use of **imputed** genotype data.
- Careful **collection and quality-checking** of information is essential to avoid errors.
- A wide array of methods may be used, including fixed effects, random effects, and Bayesian meta-analysis and they have particular advantages and disadvantages.
- Application of the meta-analysis methodology in genome-wide data has been successful in identifying more disease-related genes for various conditions.

Thank you.