Meta-analysis and replication

of Genome Wide Association Studies (GWAS)

Tinashe Chikowore, PhD

tinashedoc@gmail.com

Outline of this session

- When and why use meta-analysis in genetic studies
- Theoretical background of meta-analysis
- Fixed Effects
- Random Effects
- Bayesian and Trans-ethnic meta-analysis
- Tools
- Sample Replication approaches

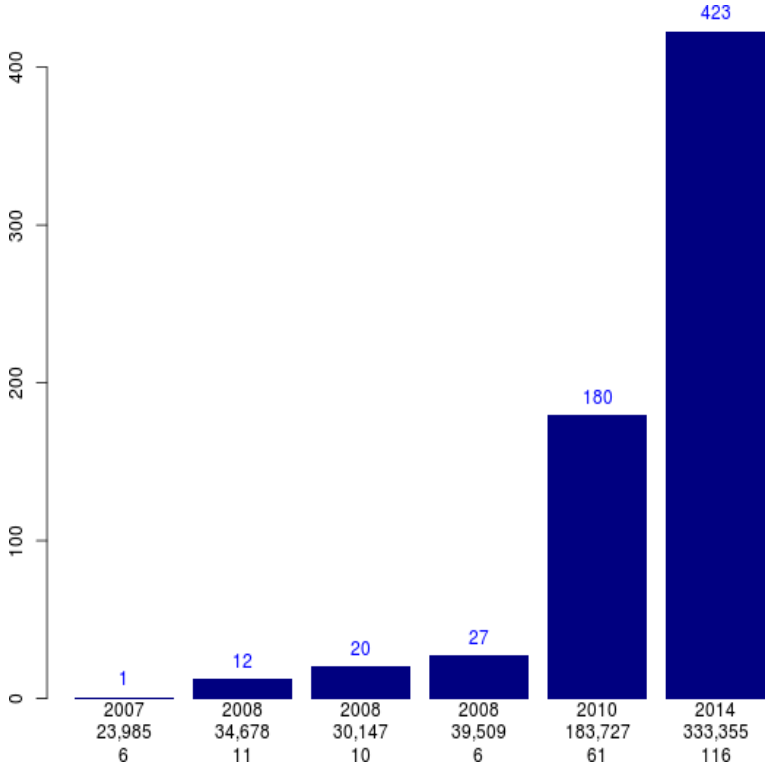
Genetic association studies

- Meta-analysis of candidate gene studies
  → Adhoc analysis of published results
  → Replication

- Meta-analysis of GWA studies
  → Replication of most significant hits from discovery sample
  → International consortia

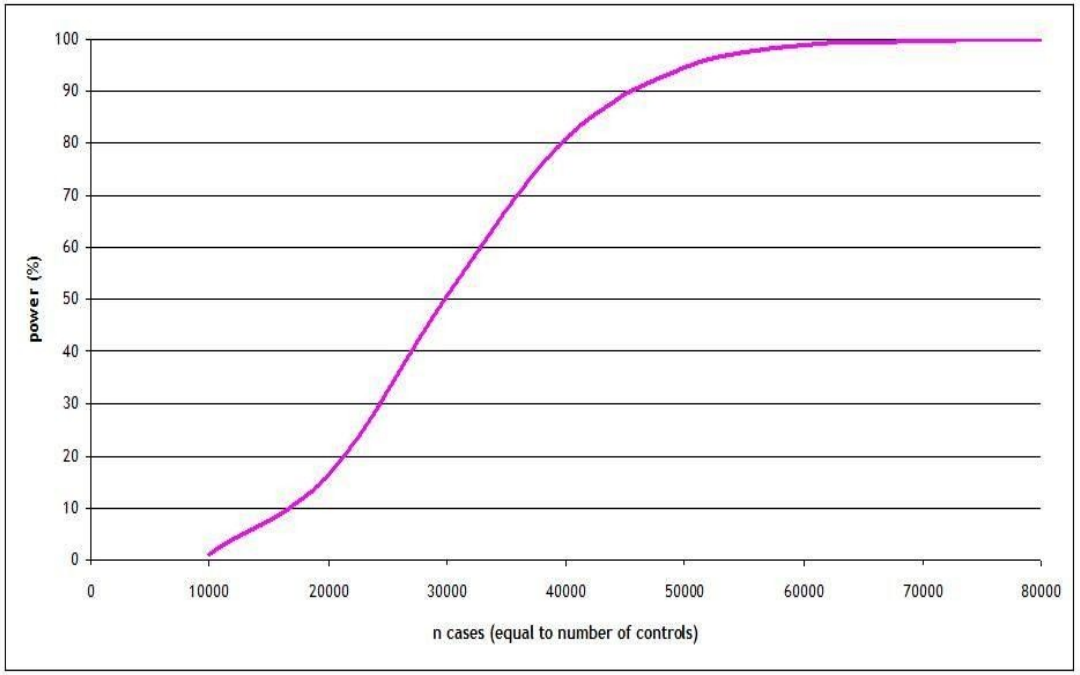
Height and SNPs discovered

- Number of studies, sample size and SNPs discovered



Motivation for GWAS

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



Why not combine samples for GWAS?

- Ethical Constraints
- Population stratification

- GWAS consortium have been formed

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

Goal of meta-analysis

- Quantitative synthesis of results from different samples/studies
- Larger N -> More power!
- Done by pooling:
  - Genetic effect of a SNP on a phenotype
  - P-value of the association test

Types of meta-analysis

- Pooling effect estimates

  - What is 'true' effect in population?

  ➡ Inverse variance weighted method

  ➡ Fixed vs. Random models

- Pooling p-values

  - Is association significant?

  ➡ Pooled z-score method

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

Inverse variance weighted method

Fixed models

Assumptions:

- There is one underlying ‘true’ effect
- All deviations of sample effects from the ‘true’ effect are due to chance

Prerequisites:

- Same scale must be used across samples!
- Same reference allele on same strand!

Inverse variance weighted method

Computing pooled effect:

$$\text{Pooled effect} = \frac{\text{Sum (weights * effect estimates)}}{\text{Sum (weights)}}$$

$$\beta_{pooled} = \frac{\sum_{i=1}^{N} (w_i * \beta_i)}{\sum_{i=1}^{N} (w_i)}$$

$$w_i = \frac{1}{\mathrm{var}(\beta_i)}$$

$$\mathrm{var}(\beta_i) = se\ (\beta_i)^2$$

$$se\ (\beta_i) = \frac{SD_i}{\sqrt{n_i}}$$

i=1…N samples

Inverse variance weighted method

Computing pooled standard error:

$$\text{Pooled standard error} = \text{Square root}\left(\frac{1}{\text{Sum (weights)}}\right)$$

$$se_{pooled} = \sqrt{\frac{1}{\sum_{i=1}^{N}(w_i)}}$$

$$w_i = \frac{1}{\mathrm{var}(\beta_i)}$$

$$\mathrm{var}(\beta_i) = se\ (\beta_i)^2$$

Inverse variance weighted method

Computing 95% confidence interval:

Pooled effect +/- 1.96 * pooled standard error

Inverse variance weighted method

Computing test statistic:

$$\chi^2_{df=1} = \frac{\beta_{pooled}^{\ 2}}{se_{pooled}^{\ 2}} = \frac{(\sum_{i=1}^{N} w_i * \beta_i)^2}{\sum_{i=1}^{N} w_i}$$

$$z = \frac{\beta_{pooled}}{se_{pooled}} = \frac{\sum_{i=1}^{N} w_i * \beta_i}{\sqrt{\sum_{i=1}^{N} w_i}}$$

Look up or compute the associated p-value

Inverse variance weighted method

Computing test statistic:

$$\chi^2_{df=1} = \frac{\beta_{pooled}^{\ 2}}{se_{pooled}^{\ 2}} = \frac{(\sum_{i=1}^{N} w_i * \beta_i)^2}{\sum_{i=1}^{N} w_i}$$

$$z = \frac{\beta_{pooled}}{se_{pooled}} = \frac{\sum_{i=1}^{N} w_i * \beta_i}{\sqrt{\sum_{i=1}^{N} w_i}}$$

P=0.05 → χ^2=3.84
Z=1.96

P=0.001 → χ^2=10.83
Z=3.29

Do assumptions of fixed model hold?

Test of homogeneity

- Cochran's Q statistic evaluates if heterogeneity exists

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

χ^2-distributed with df=k-1

k=Number of samples

α=0.10

➡ With small sample size, low power!!

Quantify heterogeneity

I^2 statistic

$$I^2 = \frac{Q-(k-1)}{Q} * 100$$

Range 0-100%

$I^2 > 50\%$: Large heterogeneity

$I^2 > 75\%$: Very large heterogeneity

Causes of heterogeneity

Possible causes related to bias in samples:

- Differential selection of cases and controls
- Poor genotyping
- Poor imputation
- Poor genotype data cleaning
- Different SNP platforms
  (imputed vs. observed SNPs)
- Poor/differential phenotyping
- Population stratification

Causes of heterogeneity

Possible causes related to genuine differences across samples:

- Different LD structure across populations (truly associated SNP vs. tested SNP)
- 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.
- Winner's curse: The originally identified effect size is likely to be overestimated in comparison to its true value.

Solution to heterogeneity

Random effects model

Assumptions:

- Assume that there is one underlying *distribution* of effects
- Normal distribution of effects

Random effects models

Used if:

- Large differences across samples
  (expected or observed)
- Same scale is used across samples

But:

- Number of samples should be sufficiently large

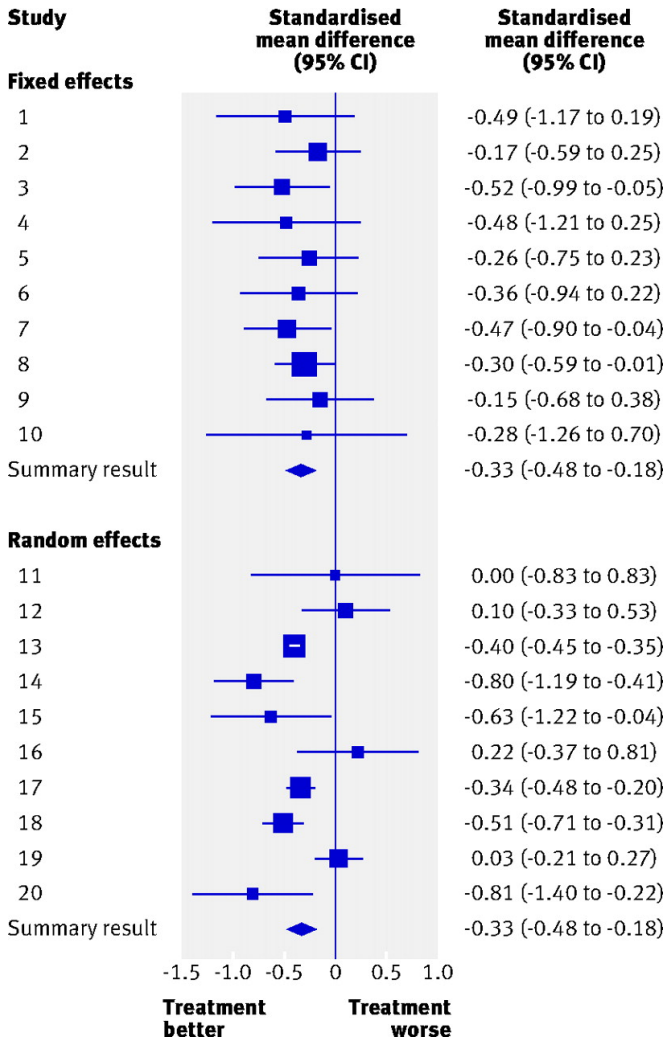
Random effects models

Estimate between study variance
(DerSimonian Laird estimator)

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

- τ^2 is incorporated in the weights
- Random effects model are more conservative (larger se)

Fig 1 Forest plots of two distinct hypothetical meta-analyses that give the same summary estimate (centre of diamond) and its 95% confidence interval (width of diamond).



Richard D Riley et al. BMJ 2011;342:bmj.d549

Trans ethnic GWAS Meta-analysis

Genetic Epidemiology 35:809–822 (2011)

Transethnic Meta-Analysis of Genomewide Association Studies

Andrew P. Morris*

Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom

- Fixed effects assumes the allele has the same effect in all populations
- Random effects assumes that each population has an underlying effect.
- Problem: It is assumed populations from the same ethnic group should have a homogenous effect compared to those they are distantly related to.
- Using a *Bayesian partition model*, in MANTRA, the population is separated into clustered due to relatedness. Inside the cluster fixed effect is assumed and among the clustered random effects are assumed

Z-score pooling method

Good to use if:

- Large differences across samples
- Number of samples is small
- Same scale is NOT used across samples

Z-score pooling method

Computing pooled z-score:

$$\text{Pooled z-score} = \frac{\text{Sum (weights * z-scores)}}{\text{Sum (weights)}}$$

Individual z-scores computed by:

- Converting individual p-values into z-scores
- Taking the sign of the effects into account

Z-score pooling method

Computing pooled z-score:

$$\text{Pooled z-score} = \frac{\text{Sum (weights * z-scores)}}{\text{Sum (weights)}}$$

$$z_{pooled} = \frac{\sum_{i=1}^{N} (w_i * z_i)}{\sqrt{\sum_{i=1}^{N} (w_i^2)}}$$

$$w_i = \sqrt{n_i}$$

nature genetics

Genome-wide association study identifies 112 new loci for body mass index in the Japanese population

Masato Akiyama[1], Yukinori Okada[1–3], Masahiro Kanai[1], Atsushi Takahashi[1,4], Yukihide Momozawa[5], Masashi Ikeda[6], Nakao Iwata[6], Shiro Ikegawa[7], Makoto Hirata[8], Koichi Matsuda[9], Motoki Iwasaki[10], Taiki Yamaji[10], Norie Sawada[10], Tsuyoshi Hachiya[11], Kozo Tanno[11,12], Atsushi Shimizu[11], Atsushi Hozawa[13,14], Naoko Minegishi[13,14], Shoichiro Tsugane[15], Masayuki Yamamoto[13,14], Michiaki Kubo[16] & Yoichiro Kamatani[1,17]

New biological insights using trans-ethnic meta-analysis of BMI

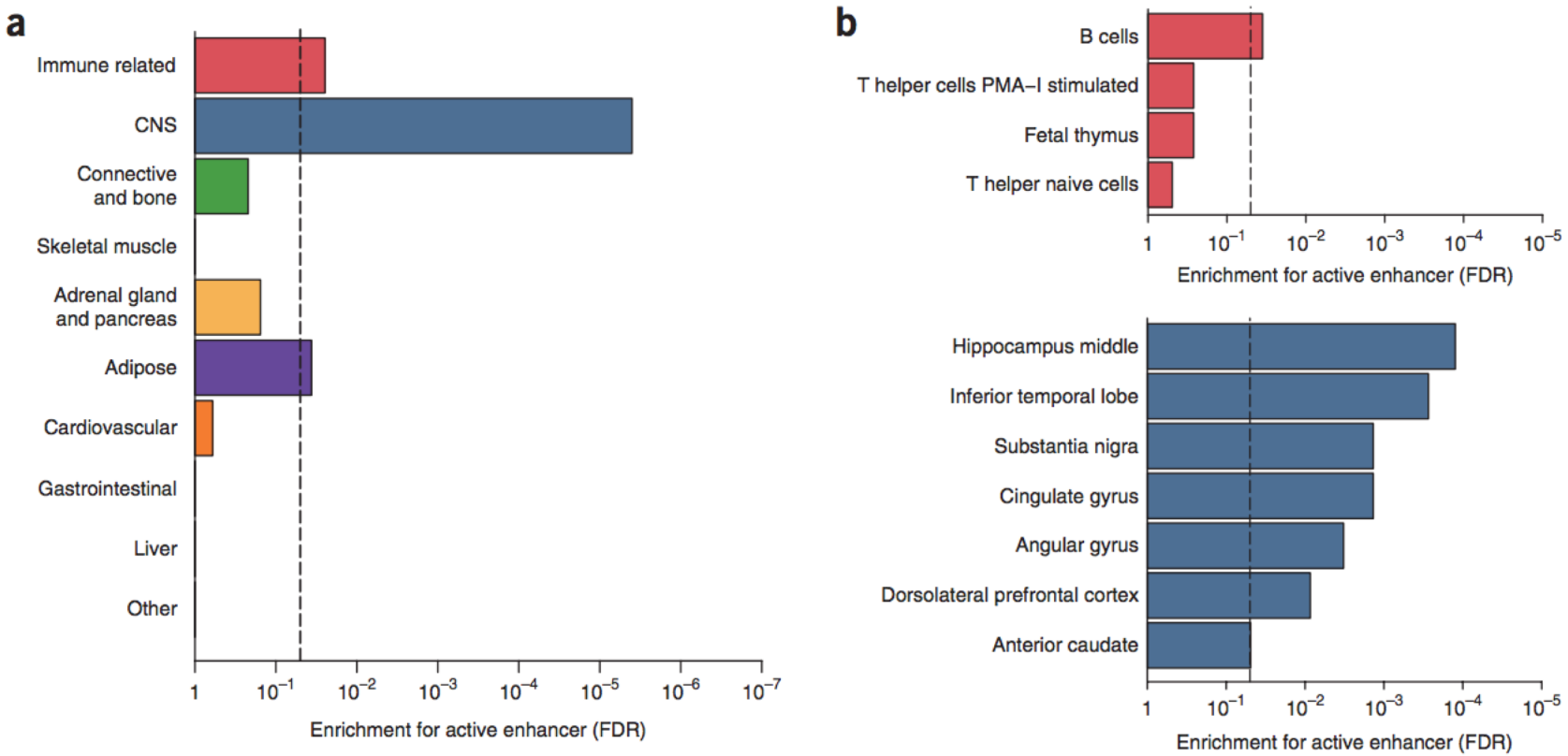


Figure 1 Enrichment of identified variants in active enhancers. (**a**) Enrichment of the variants included in the 99% credible sets for active enhancer in 10 cell groups (**a**) and immune-related cell and CNS groups (**b**). Shown are cell types with $P < 0.05$ in **b**. P values were calculated by 1×10^7 permutations. FDR was estimated using Benjamini–Hochberg method. Vertical dashed lines denote FDR = 0.05.

Meta-analysis in GWA studies

Example study: Frayling et al.

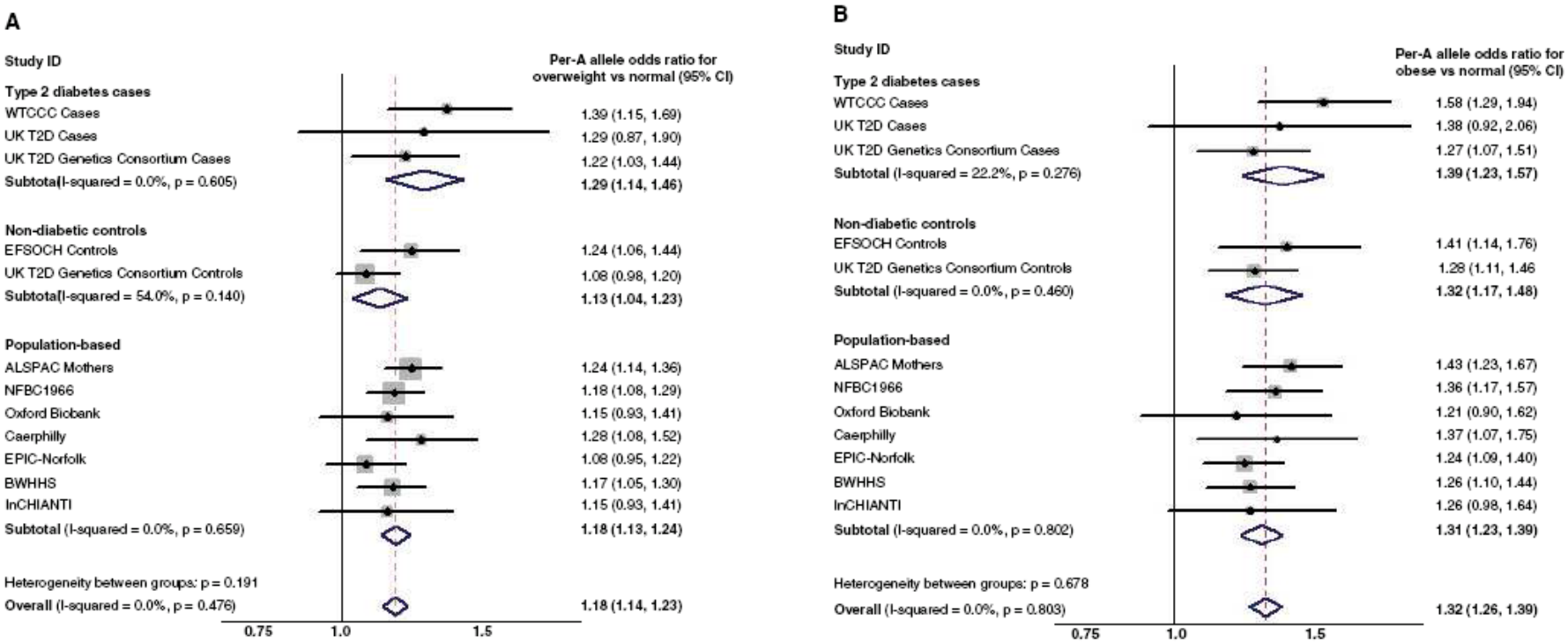
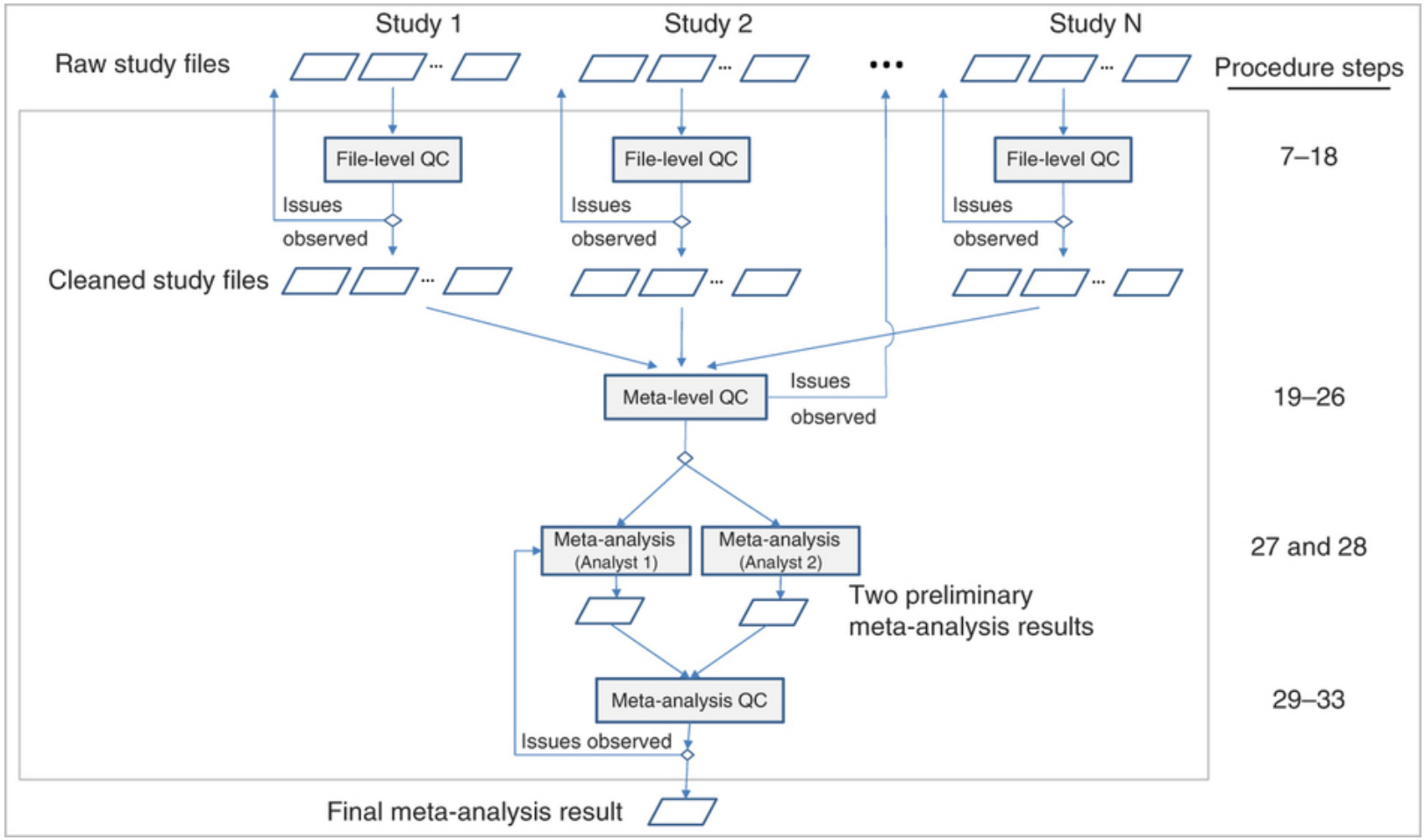


Fig. 2. (**A** and **B**) Meta-analysis plots for odds of (A) overweight and (B) obesity, compared with normal weight in adults for each copy of the A allele of rs9939609 carried. (**C** and **D**) Bar charts showing (C) DEXA-measured fat mass in 9-year-old children and (D) DEXA-measured lean mass in 9-year-old children, both from the ALSPAC study. Error bars represent 95% confidence intervals

Interpreting GWAS findings

- In the absence of between-study heterogeneity, fixed and random effects calculations yield identical point estimates and confidence intervals.
- With increasing between-study heterogeneity, the random effects summary estimates have larger variance (wider confidence intervals) and usually less prominent statistical significance.
- Most meta-analysts would typically run both models, but prefer placing emphasis on random effects.
- Statistically significant associations in fixed or random effects calculations need replication.

QC in GWAS Meta-analyis



EasyQC

- File name errors -> sounds simple doesn’t it, but with 167 files it is essentiel that all files can be traced back to a specific cohort
- Incorrect specification of the Phenotype
- Flipped alleles
- Duplicated SNPs
- Bad imputation quality
- Association issues from incorrect analysis models
  - Population stratification
  - Improper model adjustments
  - Unaccounted relatedness of individuals

Principles of replication

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > features > article

Published: 06 June 2007

Replicating genotype–phenotype associations

NCI-NHGRI Working Group on Replication in Association Studies

Nature **447**, 655–660 (2007) | Cite this article

12k Accesses | **1085** Citations | **34** Altmetric | Metrics

COMMENT

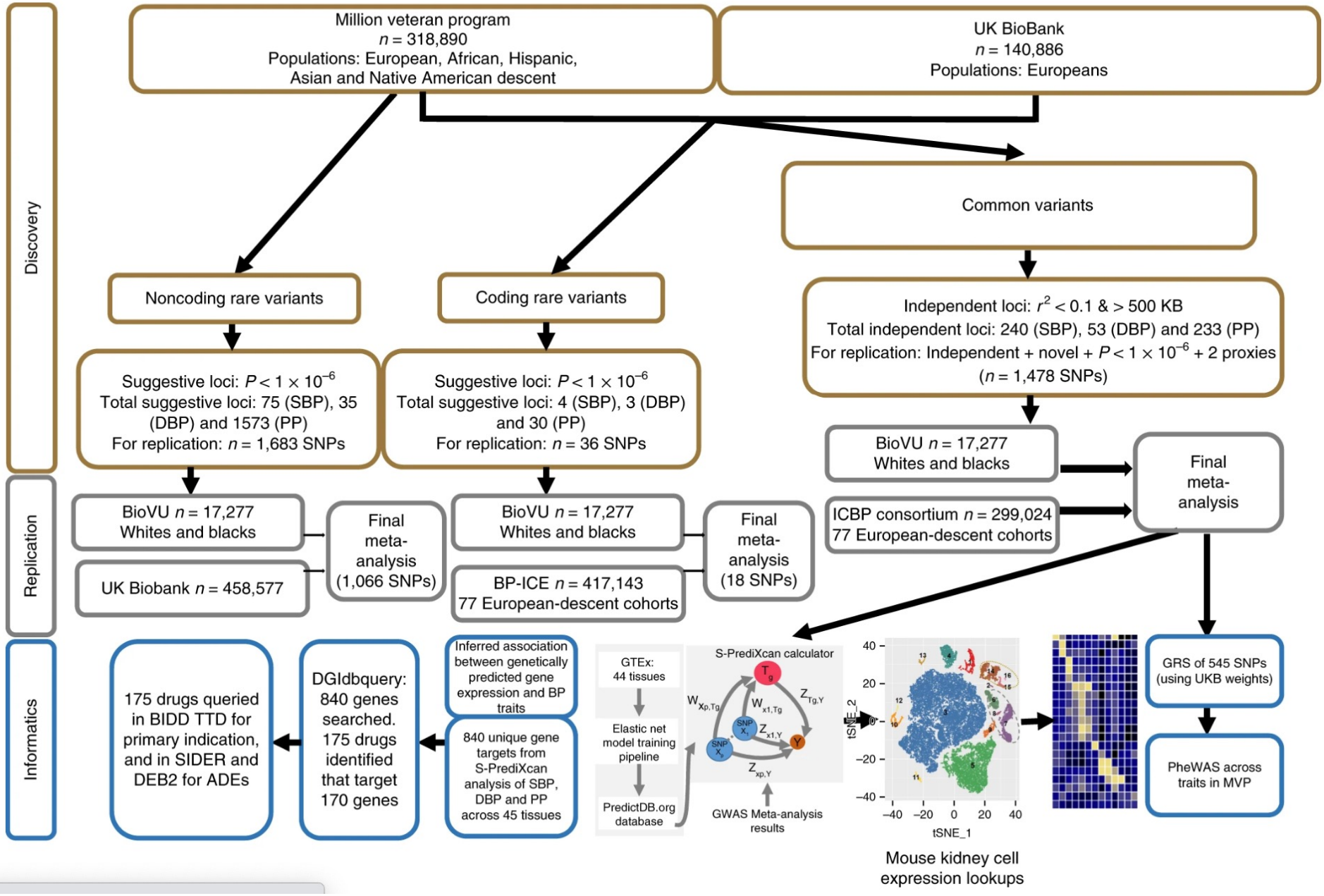
DOI: 10.1038/s41467-018-07348-x OPEN

Examining the current standards for genetic discovery and replication in the era of mega-biobanks

J.E. Huffman [1]

Fig. 1: Study design schematic.

From: Trans-ethnic association study of blood pressure determinants in over 750,000 individuals



- Thank You