Meta-analysis methods

For genome-wide association studies and beyond

What is meta-analysis?

- Meta-analysis είναι η μέθοδος που χρησιμοποιεί τον συνδυασμό και τη σύνοψη ερευνητικών μελετών που εχουν γίνει στο παρελθόν για το ίδιο θέμα ώστε να οδηγηθουμε σε ένα συνολικό συμπέρασμα
- Advantages:
  - Results can be generalized to a larger population
  - Can use summary data (no sharing individual-level data)
  - The precision and accuracy of estimates can be improved
- Disadvantages:
  - Sources of bias are not controlled by the method: a good meta-analysis of badly designed studies will still result in bad statistics
  - Publication bias: studies show negative or insignificant results are less likely to be published

Models for Data Synthesis

- Μερικά μοντέλα που χρησιμοποιούντε στη GWAS meta-analysis είναι :
- P values and Z scores
- Fixed effects
- Random effects
- Optimal weights
- Bayesian meta-analysis
- Q values (false discovery rate)

P values

Η πιο γνωστη μεθοδος που χρησιμοποιει τα P values είναι η μεθοδος του Fischer.

Σε αυτή τη προσέγγιση η αρχική υπόθεση λέει ότι το true effect είναι null σε όλα τα data sets .Ένα σημαντικο μειονεκτημα του P values είναι πως δεν μπορεί να μας παρέχει μια συνολική εκτίμηση του effect size .Επιπλέον ο συνδυασμός των P values μπορεί να είναι ψευδής όταν η κατευθυνση των effects των διαφορων ερευνών δεν συμφωνούν.

Z scores

Μία άλλη μέθοδος που σχετίζεται με τη μέθοδο του ficher είναι η μεθοδος των Z scores .Το Z score υπολογίζεται από τον τύπο $\frac{\sum_i w_i z_i}{\sqrt{\sum_i w_i^2}}$

Όπου το Wi είναι η τετραγωνική ριζα του μεγέθους του δείγματος της i- οστης ερευνας και το $Z\iota=\Phi^{-1}*(1-\frac{Pi}{2})*$(effect direction for study-i).

Η μεθοδος των Z scores βασιζεται στο μεσο ορο των Zi values.Ενα πλεονεκτημα της μεθόδου είναι ότι λαμβάνει υπόψιν το direction of effect

Fixed Effects

Η fixed effect προσέγγιση είναι η πιο γνωστή για τη σύνθεση GWAS data και η πιο ισχυρή για την ανακαλυψη των phenotype-associated SNPs .Η fixed effect υποθέτει ότι το true effect of each risk allele είναι το ιδιο σε κάθε σετ δεδομένων .Υπάρχουν πολλα μοντέλα που χρησιμοποιούντε για τα fixed effects . Αυτό που χρησιμοποιειται κυριως είναι το inverse variance weighting. Ένα άλλο μοντέλο που χρησιμοποιείται είναι το cochran-mantel-haenszel

- Inverse variance weighting : is a method of aggregating 2 or more random variables to minimize the variance of the weighted average
- The Cochran-Mantel-Haenszel : method is a technique that generates an estimate of an association between an exposure and an outcome after adjusting for or taking into account confounding. The method is used with a dichotomous outcome variable and a dichotomous risk factor. We stratify the data into two or more levels of the confounding factor (as we did in the example above). In essence, we create a series of two-by-two tables showing the association between the risk factor and outcome at two or more levels of the confounding factor, and we then compute a weighted average of the risk ratios or odds ratios across the strata

Random effects

Το μοντέλο random effects δεν χρησιμοποιείται όσο το fixed effects .

Παρόλα αυτά το random effect είναι πιο κατάλληλο όταν ο στόχος είναι να λάβει υπόψιν και το generalizability of the observed association .Το πιο δημοφιλές random effect model είναι το DeSimonian and Laird estimator ,αλλά υπαρχούν και μοντέλα όπως το Sidic-Jonkman ,Hedges-Vevea,Hunte-Schmidt και Schuster.

- **Random-effects (DerSimonian and Laird) method for meta-analysis**

A variation on the inverse-variance method is to incorporate an assumption that the different studies are estimating different, yet related, intervention effects. This produces a random-effects meta-analysis, and the simplest version is known as the DerSimonian and Laird method (DerSimonian)

Optimal weights

Το μοντέλο optimal weights είναι το inverse variance weighting . Οι ιδιότητες αυτου του μοντέλου είναι γνωστές για την αρχική αξιολόγηση και ανακάλυψη του GWAS meta-analysis

Bayesian meta-analysis

Η bayesian μέθοδος εχει αναπτυχθεί για να βρίσκει το καλύτερο ταίριασμα των διάφορων παραλλάγων από το GWAS meta analysis και να χαρακτηρίζει την αρχιτεκτονική σύνθετων χαρακτηριστικών στα οποία υπαρχουν χιλιαδες top SNPs.

Q values

Η μεθοδος Q values είναι γνωστη στους κλάδους της γενετικης που τελειώνουν σε ”omics”. Το Q value είναι το έλαχιστο false discovery rate που επιτυγχανεται .

Τα Q values για μια λιστα από P values μπορουν να υπολογιστουν χρησιμοποιώντας το διαθεσιμο λογισμικο QVALUE.

Fisher's method

- Combine p values from independent tests bearing upon the same overall hypotheses:

$$X_{2m}^2 = -2\sum_{i=1}^{m} \ln pi$$

- When the p values tend to be small, the test statistic will be large suggesting that the null hypotheses are not true for every test
- Under null (all null hypotheses are true) and when all p values are independent, it is a chi-squared distribution with 2m degrees of freedom.
- Extend to dependent tests
  - Scaled chi-squared distribution random variable
  - Brown's method: known covariance
  - Kost's method: unknown covariance

Heterogeneity

- Heterogeneity:
  - Sources of heterogeneity:
    - Some phenotypes are difficult to define and standardize, e.g., behavioral traits
    - Effect size might be higher in studies when individuals are older, or more educated or healthier
    - Genetic studies: different ethnicity groups, different genotyping platform or imputation software
- In this case, there may be different underlying true effect sizes for different studies

Method	Description	Advantages	Disadvantages	Main software used
P value meta-analysis	Simplest meta-analytical approach	Allows meta-analysis when effects are not available	Direction of effect is not always available; inability to provide effect sizes; difficulties in interpretation	METAL, GWAMA, R packages
Fixed effects	Synthesis of effect sizes. Between-study variance is assumed to be zero	Effects readily available through specialized software	Results may be biased if a large amount of heterogeneity exists	METAL, GWAMA, R packages
Random effects	Synthesis of effect sizes. Assumes that the individual studies estimate different effects	Generalizability of results	Power deserts in discovery efforts; may yield spuriously large summary effect estimates when there are selection biases	GWAMA, R packages
Bayesian approach	Incorporates prior assessment of the genetic effects	Most direct method for interpretation of results as posterior probabilities given the observed data	Methodologically challenging; GWAS-tailored routine software not available; subjective prior information used	R packages
Multivariate approaches	Incorporates the possible correlation between outcomes or genetic variants	Increased power can identify variants that conventional meta-analysis do not reveal using the same data sets	Computationally intensive; software not available for all analyses; some may require individual-level data	GCTA for multi-locus approaches
Other extensions	A set of different approaches that allows for the identification of multiple variants across different diseases	Summary results of previous meta-analyses can be used	May need additional exploratory analyses for the identification of variants; prone to systematic biases	Software developed by the authors of the proposed methodologies

Dealing with heterogeneity

- ***Phenotype-based heterogeneity.***
  - GWASs have been less successful for diseases in which phenotypes have been more difficult to define and to standardize, such as cognitive traits and mental-health-related diseases55, behavioural traits56 and osteoarthritis57,58. Evidence from other fields, such as obesity, also suggests that the establishment of associations may be dependent on phenotype definition59 and that variability in defi- nitions may cause heterogeneity in effect size or even spurious associations60. In some cases, harmoniza- tion of different phenotype definitions61,62 is possible, whereas in other situations this may not be feasible: for example, if phenotypes have already been collected and it is not possible to go back and remeasure them. The process must balance the need to augment the sample size (to increase power for gene discovery) with the likelihood of increased heterogeneity, which dilutes the average genetic effect and thus leads to loss of

Dealing with heterogeneity

- ***Ancestry-based heterogeneity.***
  - Synthesizing data from populations of different ancestry may increase the observed heterogeneity. The agnostic GWAS approach usually captures common markers that are likely to be in linkage only with the functional or causative cul- prits; the most strongly associated SNP may not be the functional or the causative variant. An assessment of GWAS-discovered variants shows modest correlation in MAF of the variants between ancestries and differ- ent genetic effects in different ancestries65. However, consistency across different ancestries may be higher for some common diseases66. A proposed transethnic meta-analysis approach takes into account the similarity in allelic effects between the most closely related popu- lations while allowing for heterogeneity between more diverse ethnic groups67; this approach may occasionally improve power to detect a novel association and localize causal variants67.

Dealing with heterogeneity

- ***Other sources of heterogeneity.***
  - There are several other sources that can introduce heterogeneity in meta-analysis of genetic data. Population stratification may exist even in populations that are assumed to be fairly homogeneous. Gene–gene interactions and gene–environment interactions with differential non-genetic environmental exposures across different populations may also introduce heterogeneity
  - Finally, sex differences may induce heterogeneity, and some studies suggest differential genetic effects with respect to sex for many common variants, although this has not been borne out in other large-scale empiri- cal studies.

Μετρικές ανομοιογένειας

- Οι μετρικές που χρησιμοποιούντε για την ανομοιογένεια είναι η Cochran's Q statistic I^2.
  - Cochran's *Q* statistic follows a X^2 distribution with *k*–1 degrees of freedom, where *k* is the number of studies and is typically considered to be significant at $a = 0.10$ (where *a* is the type I error)

$$Q = \sum_i w_i(\theta_i - \theta_F)$$

  - I^2 quantifies the heterogeneity by measuring the amount of heterogeneity that is not due to chance. It ranges from 0–100% and is considered low, moderate, large and very large for values 0–25%, 25–50%, 50–75% and >75%, respectively.

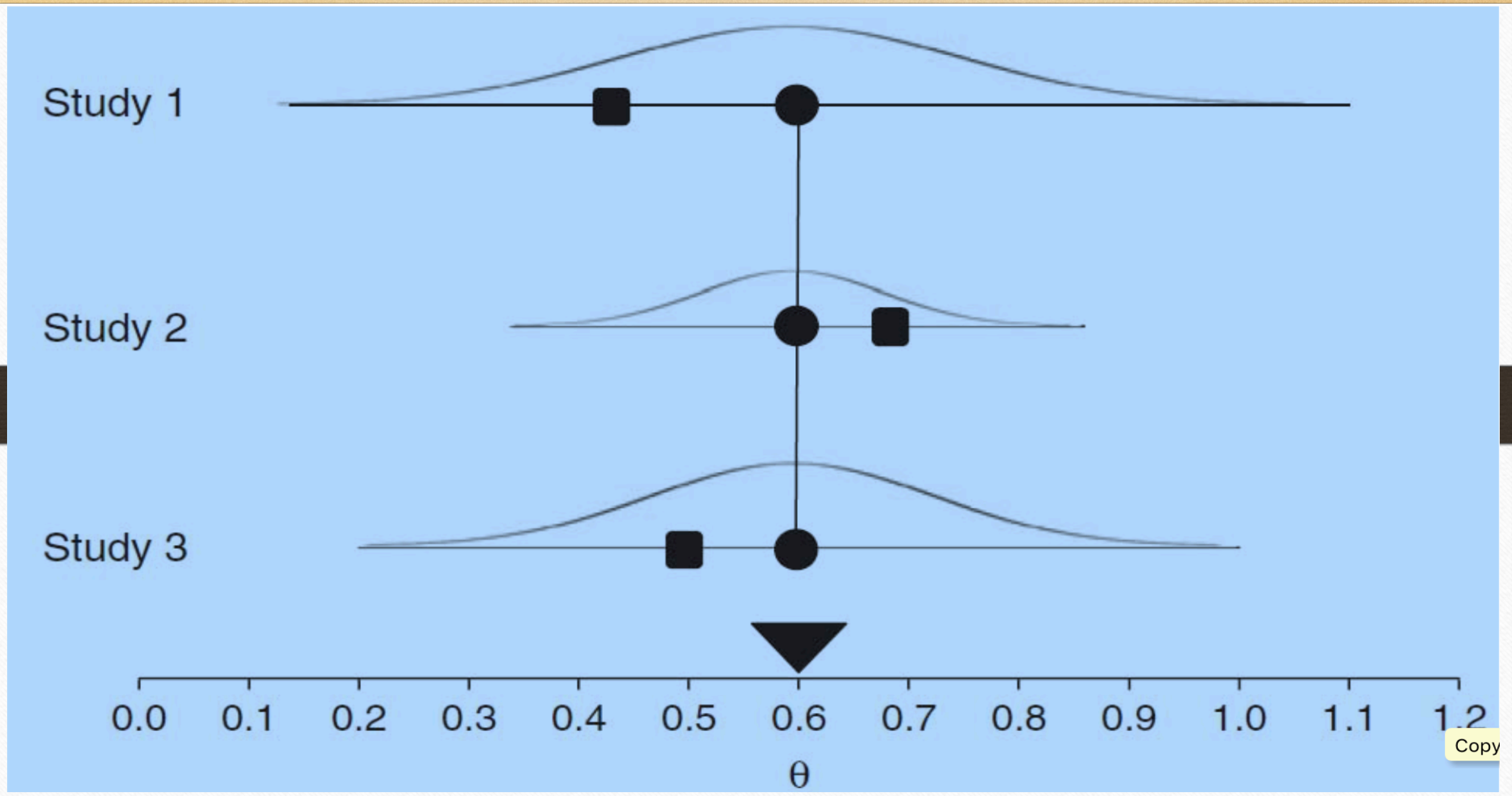
$$I^2 = 100 * (Q - (k - 1))$$

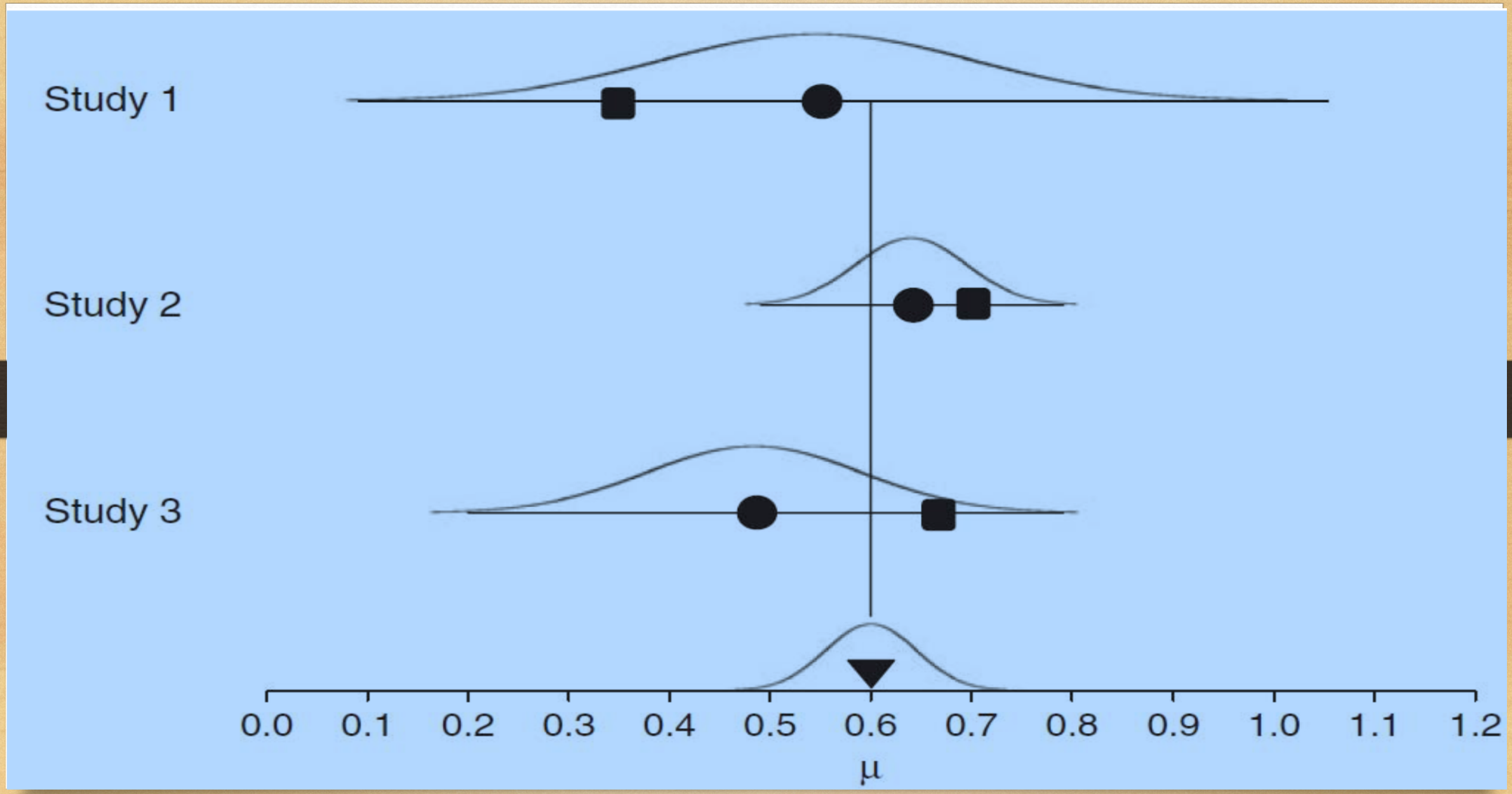
Data Storage

Data storage is an important aspect of meta-analysis as the individual-level data collected by each partner and also single participants' genotypes should be kept secured and unidentifiable. Most collaborative meta-analyses use online storage options to deposit summary data, giving access to members of the analysis team. This enables the partners to retain control of individual-level primary data. In most settings, summary data are statistically as efficient for meta-analysis as individual-level data16. The major drawback of working with summary data comes when more detailed investigations are required, such as conditional analyses, gene–gene interactions or adjusted analyses.

Meta-analyses of meta-analyses

Για κάποιους κρίσιμους φενότυπους μπορεί να υπάρχουν πολλές ατομικές έρευνες καθώς και πολλές συλλογικές. Αυτά τα αποτελέσματα μπορεί να έχουν προέρθει είτε από διαφορετικούς μηχανισμούς είτε από διαφορετικές έρευνες. Η επιτυχής συνένωση των αποτελεσμάτων των διαφορετικών ερευνών θα επιτρέψει στις mega-analyses να έχει αποτελέσματα πολλών παλιών ερευνών καθώς και πολλών meta-analyses. Αυτό θα έχει ως αποτέλεσμα τη ραγδαία αύξηση στον εντοπισμό σπάνιων φενότυπων. Ειδικά για φενότυπους με σπάνια εμφάνιση.


Study 1
Study 2
Study 3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
θ


Study 1
Study 2
Study 3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
μ