Funktionelle Genomanalysen 2023

Übung 2: GWAS und Sekundäranalysen

Dr. Janne Pott

09.-11. Juni 2023

Übersicht Ablauf

- Fragen zur Vorlesung?
- GWAS Regressionsmodelle
- Mendelian Randomization + Kolokalisation

Aufgabe 1 - Ausgangslage

- autosomale SNPs rs123456 & rs127890
- normalverteilter Phänotyp X
- vier unabhängige Studien
- SNPs in Genregion ABC
- Ihre Rolle
  - Verantwortlich für Analysen in Studie 1
  - Interessiert an Zusammenhang zwischen ABC und X

Aufgabe 1a: Regressionsmodelle (2)

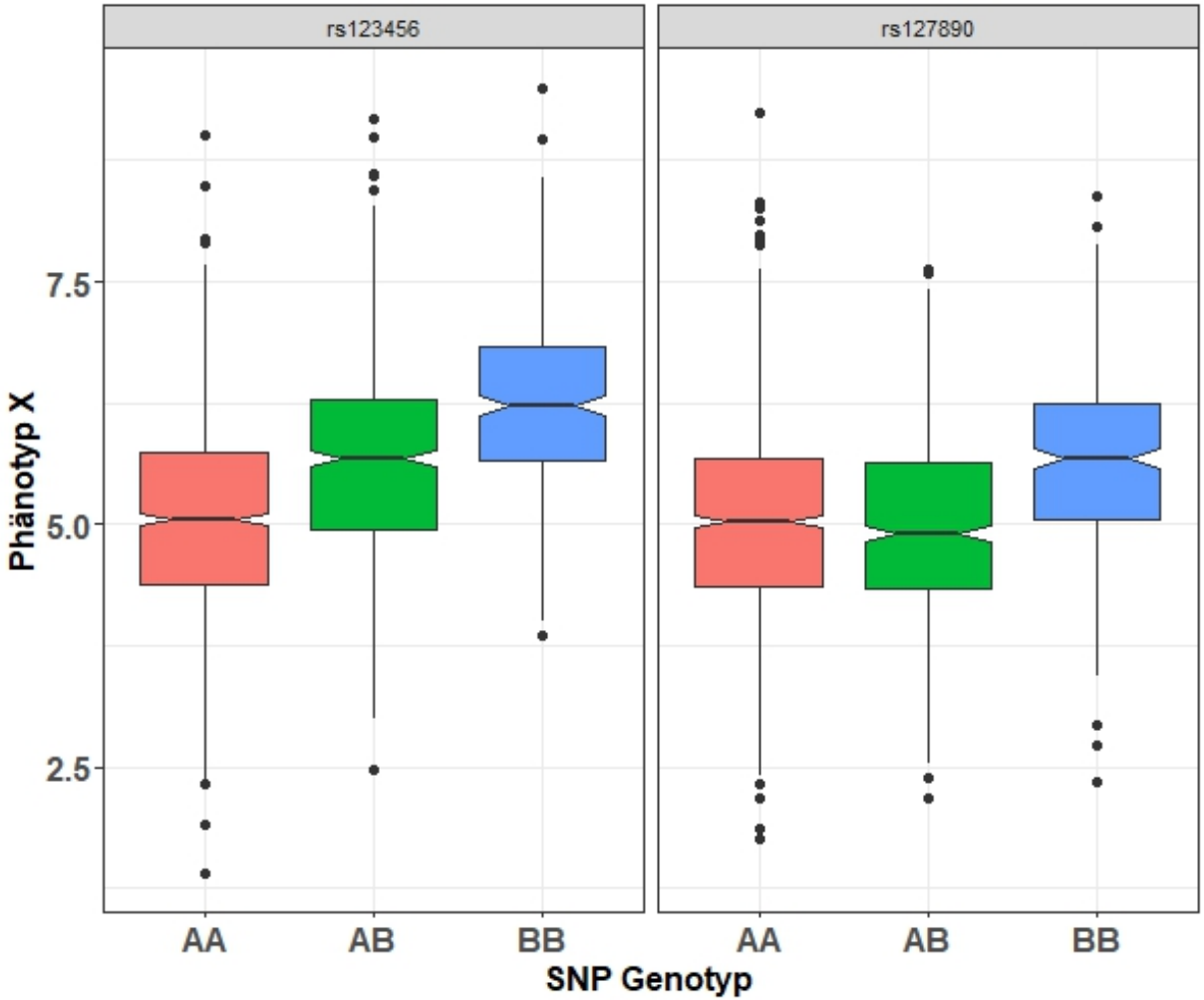


Abbildung: Boxplots

Aufgabe 1a: Regressionsmodelle (3)

- Welches Regressionsmodell wäre hier geeignet?
- Stellen Sie das Modell auf!
- Was ist die Nullhypothese?
- Kann es zwischen dominant & rezessiv unterscheiden?
- Wie kommt man von diesem Regressionsmodell zu einer GWAS?

Lösung 1a

- rs123456 sieht nach additivem Effekt aus: pro Allel B ein Effekt
- rs127890 sieht nach rezessivem Effekt aus: nur Unterschied zwischen AA/AB und BB

Option 1 (typische GWAS Regression, additives Modell):

$$x = \mu + \beta_1 \cdot G + \epsilon; \text{ mit AA=0, AB=1, und BB=2}$$

Nullhypothese: Der SNP G hat keinen Effekt auf X ($H_0 : \beta_1 = 0$)

Lösung 1a

- rs123456 sieht nach additivem Effekt aus: pro Allel B ein Effekt
- rs127890 sieht nach rezessivem Effekt aus: nur Unterschied zwischen AA/AB und BB

Option 1 (typische GWAS Regression, additives Modell):

$$x = \mu + \beta_1 \cdot G + \epsilon; \text{ mit AA=0, AB=1, und BB=2}$$

Option 2 (rezessives Modell):

$$x = \mu + \beta_1 \cdot G + \epsilon; \text{ mit AA=0, AB=0, und BB=1}$$

Nullhypothese: Der SNP G hat keinen Effekt auf X ($H_0 : \beta_1 = 0$)

Lösung 1a

- rs123456 sieht nach additivem Effekt aus: pro Allel B ein Effekt
- rs127890 sieht nach rezessivem Effekt aus: nur Unterschied zwischen AA/AB und BB

Option 3 (komplexeres Modell):

$$x = \mu + \beta_1 \cdot AB + \beta_2 \cdot BB + \epsilon; \text{ mit AB, BB} \in \{0, 1\}$$

H_0: G hat keinen Effekt auf X ($\beta_1 = 0$ und $\beta_2 = 0$).
H_1: G hat einen dominaten Effekt auf X ($\beta_1 \neq 0$ und $\beta_1 \simeq \beta_2$)
H_2: G hat einen rezessiven Effekt auf X ($\beta_1 = 0$ und $\beta_2 \neq 0$)
H_3: G hat einen additiven Effekt auf X ($\beta_1 \neq 0$ und $\beta_1 \simeq 0.5 \cdot \beta_2$)

Lösung 1a

- rs123456 sieht nach additivem Effekt aus: pro Allel B ein Effekt
- rs127890 sieht nach rezessivem Effekt aus: nur Unterschied zwischen AA/AB und BB
- aktuell: nur Analyse von zwei SNPs (= nicht genomweit)
- GWAS: Teste **ALLE** SNPs auf Assoziation mit X (führe die Regressionsanalyse $\sim$ 1 Mio. mal aus)
- Multiples Testen $\rightarrow$ Korrektur der Signifikanzgrenze nötig ($\alpha = 0.05 \rightarrow \alpha_{Bonferroni} = \frac{\alpha}{k} = 5 \cdot 10^{-8}$, k=Anzahl getesteter SNPs)

Lösung 1a

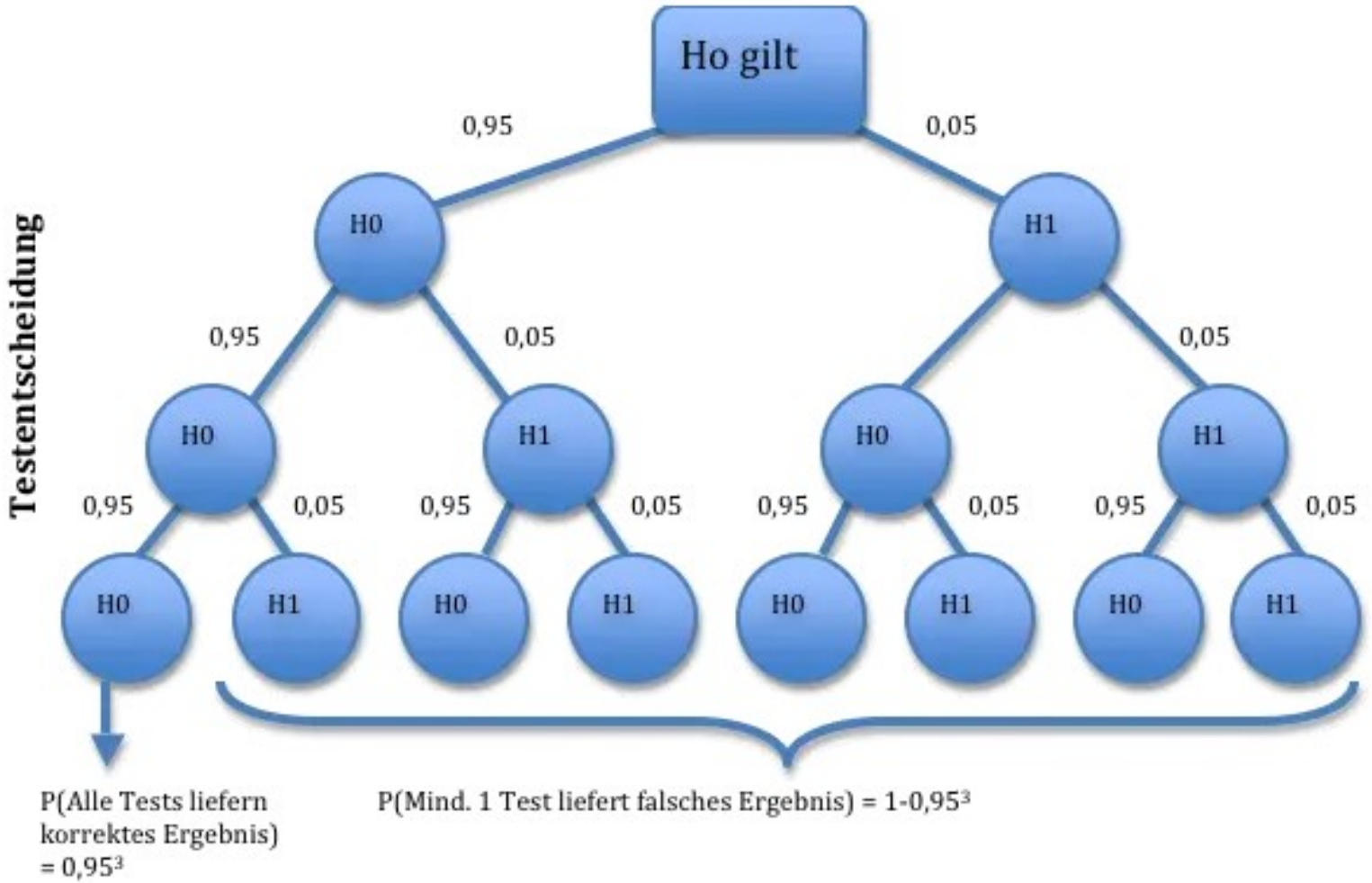


Abbildung: Alphafehler-Kumulierung

Zusatzfrage

Warum $\sim$ 1 Mio. SNPs?

- auf einem Array sind etwa 600.000 SNPs
- in Referenzgenom sind $>$ 80 Mio. SNPs

Zusatzfrage

Warum $\sim$ 1 Mio. SNPs?

- auf einem Array sind etwa 600.000 SNPs
- in Referenzgenom sind $>$ 80 Mio. SNPs

Aber

- *historisch*: man hat mit knapp 1 Mio die erste GWAS durchgeführt, der Grenzwert wurde beibehalten
- entspricht in etwa der Anzahl unabhängigen SNPs (paarweises LD $r^2 < 0.1$)

Aufgabe 1b: Meta-Analyse

Um die Power zu maximieren sollen die Daten der vier Studien in einer Meta-Analyse kombiniert werden.
Welche Annahme wird hier häufig getroffen und wie kann diese geprüft werden?

Lösung 1b

Fixed Effect Model (FEM)

- genetischer Effekt ist gleich (keine Heterogenität).
- in gemischten Modell: $x_{ij} = \mu + b_i + \beta_j \cdot G + \epsilon_{ij}$
- studien-spezifischer Intercept für Studie i (random) + fester Effekt für SNP j (fix = für alle Studien gleich)

Random Effect Model (REM)

- genetischer Effekt ist in allen Studien unterschiedlich (Berücksichtigung der Heterogenität).
- in gemischten Modell: $y_{ij} = \mu + b_{i1} + b_{ij2} \cdot G + \epsilon_{ij}$
- studien-spezifischer Intercept für Studie i (random) + studien-spezifischer Effekt für SNP j (random)

Test: Cochrans Q oder I^2 Statistik

Lösung 1b

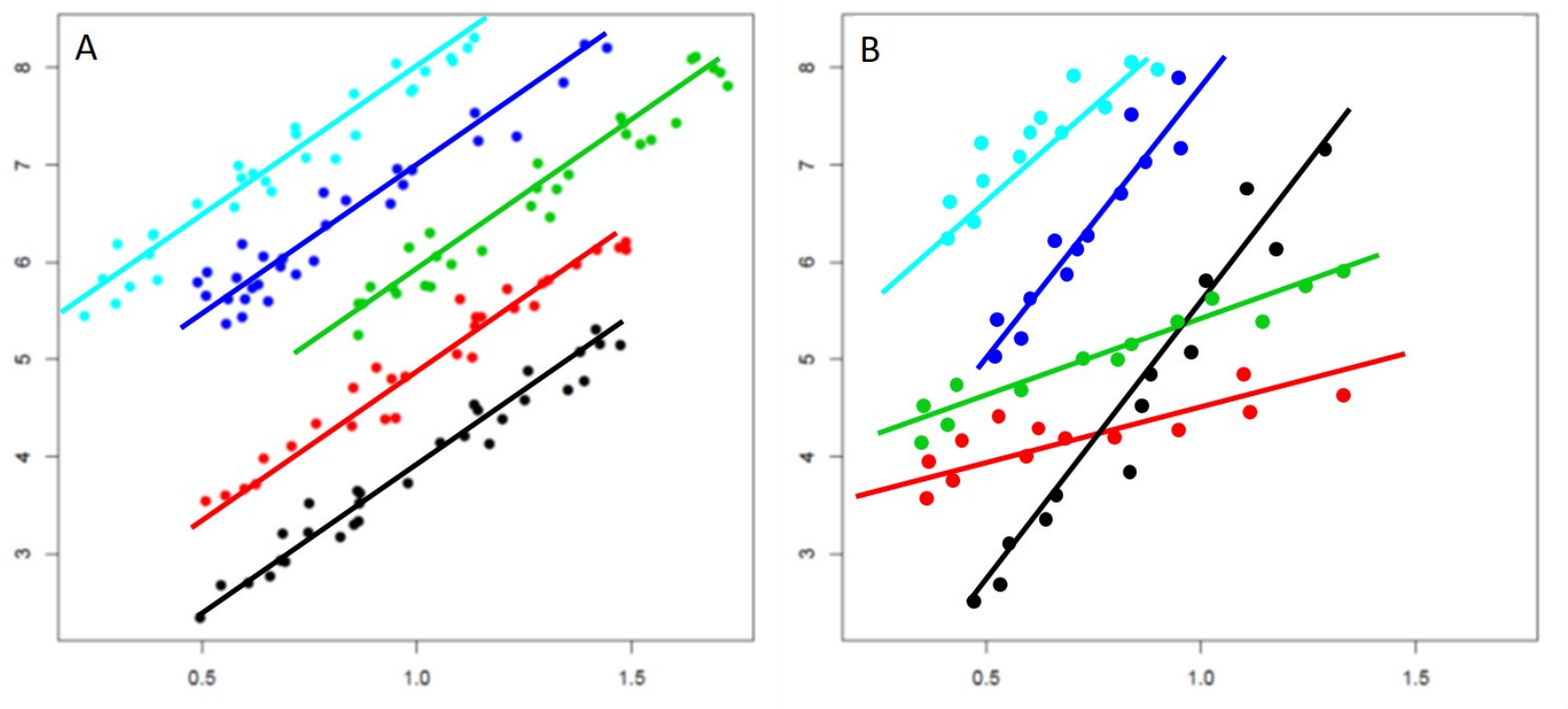
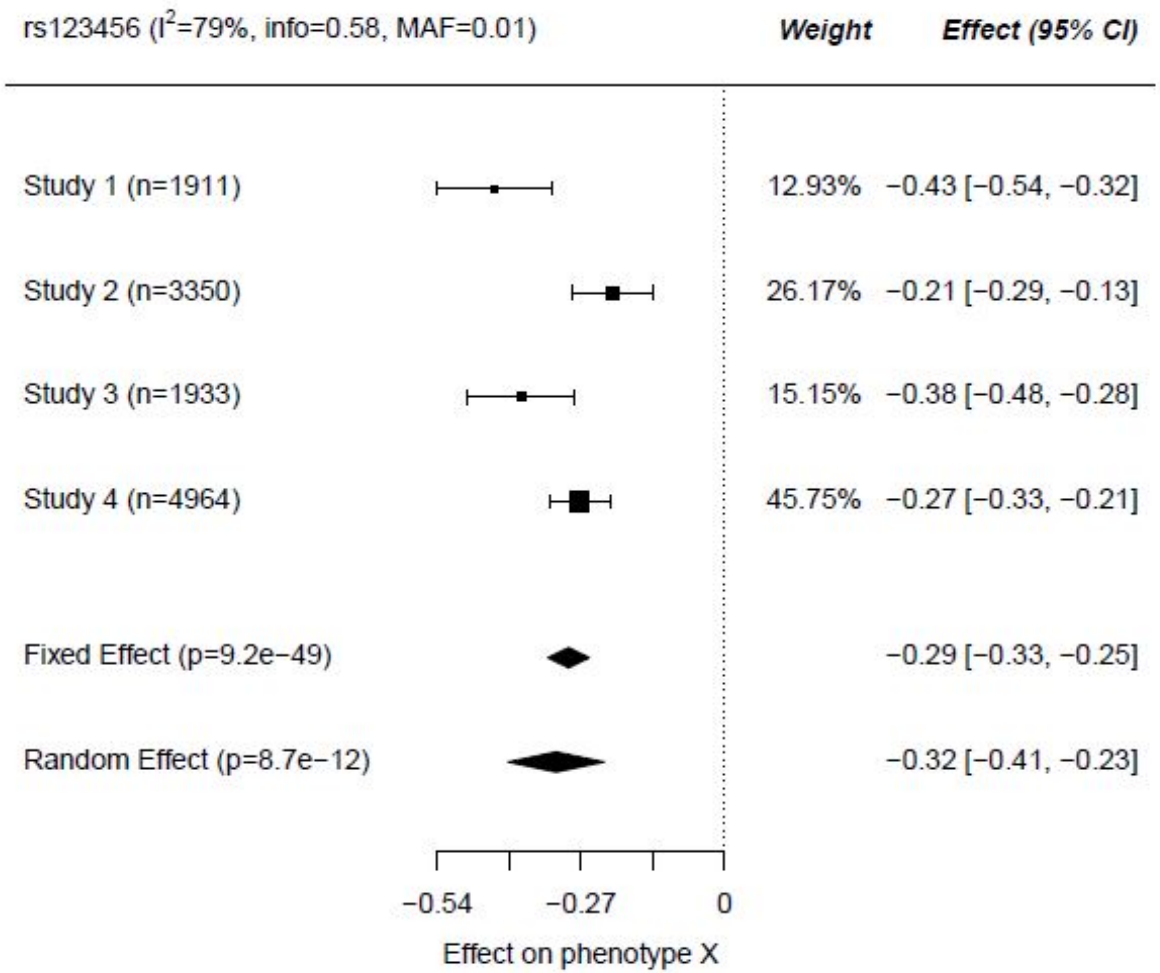


Abbildung: Schema FEM (A) und REM (B). Jede Farbe stellt eine Studie dar. In Panel A hat jede Studie ihren eigenen Intercept, aber die gleiche Steigung (=SNP-Effekt). In Panel B hat jede Studie sowohl einen eigenen Intercept als auch eine eigene Steigung.

Zusatzfrage

Was ist das für ein Plot und wie ist dieser zu interpretieren?



Zusatzfrage

Forest-Plot: Zusammenfassung der Meta-Analyse (eines SNPs)

- Quadrate: Effekte der einzelnen Studien
- Striche: 95%-Konfidenzintervalle.
- Größe des Quadrats: Fallzahl/Gewicht
- Rauten: Meta-Effektschätzer (FEM, REM).
- SNP-Quali: I^2, info und MAF

Interpretation: genomweit signifikant, aber hohe Heterogenität und schlechte info

Zusatzfrage

Typischer SNP-Filterkriterien bei einer Meta-GWAS: (Filtern, wenn...)

- MAF (z.B. $mean(MAF) < 0.01$)
- MAC (minor allele count, Anzahl des Minorallels, z.B. $MAC < 6$)
- info-score (z.B. $min(info) < 0.8$)
- Heterogenität (z.B. $I^2 > 0.75$)
- Vollständigkeit zwischen den Studien bzw. mindestens zwei Studien (z.B. $k < 2$)
- Bonferroni-adjustierte P-Wert (z.B. $p > \frac{0.05}{1{,}000{,}000} = 5 \cdot 10^{-8}$)
- LD (abh. von Ethnie, z.B. $r^2 > 0.1$)

Aufgabe 1c: Stratifikationsbias

Nachdem alle Ihre Analysen abgeschlossen sind, meldet sich ein Kollege von Studie 2 bei Ihnen. Er teilt Ihnen mit, dass bei der Analyse leider vergessen wurde auf die Populationsstruktur zu korrigieren.
Was bedeutet das und welche Konsequenzen hat das für Ihre Analyse?

Lösung 1c

Stratifikationsbias: Durch die **gemeinsame Analyse** von Personen **unterschiedlicher genetischer Herkunft** bei gleichzeitigem Vorliegen **nichtgenetisch bedingter Unterschiede** zwischen den Personengruppen können sich **falsche Schätzer genetischer Effekte** ergeben.
Mögliche Maßnahmen:

- Analyse der Populationsstruktur (Structure, PCA, MDS)
- Korrektur auf Hauptkomponenten
- Berücksichtigung der Verwandtschaftsstruktur in genetischen Daten
- Genomic Control
- Genetische Outlier weglassen

$\Rightarrow$ Option 1: Kollege rechnet die GWAS in seiner Studie nochmal neu, unter Berücksichtigung der Populationsstruktur
$\Rightarrow$ Option 2: Sie führen Genomic Control für die Summary Statistics der Studie 2 durch, und wiederholen dann die Meta-Analyse

Aufgabe 1d: Heritabilität

In Ihrer finalen Analyse erklärt der SNP 4% der Varianz von X. Die Gesamt-Heritabilität von X liegt jedoch laut Literatur bei 40%. Definieren Sie den Begriff Heritabilität und erklären Sie den Unterschied zwischen den Werten!

Lösung 1d

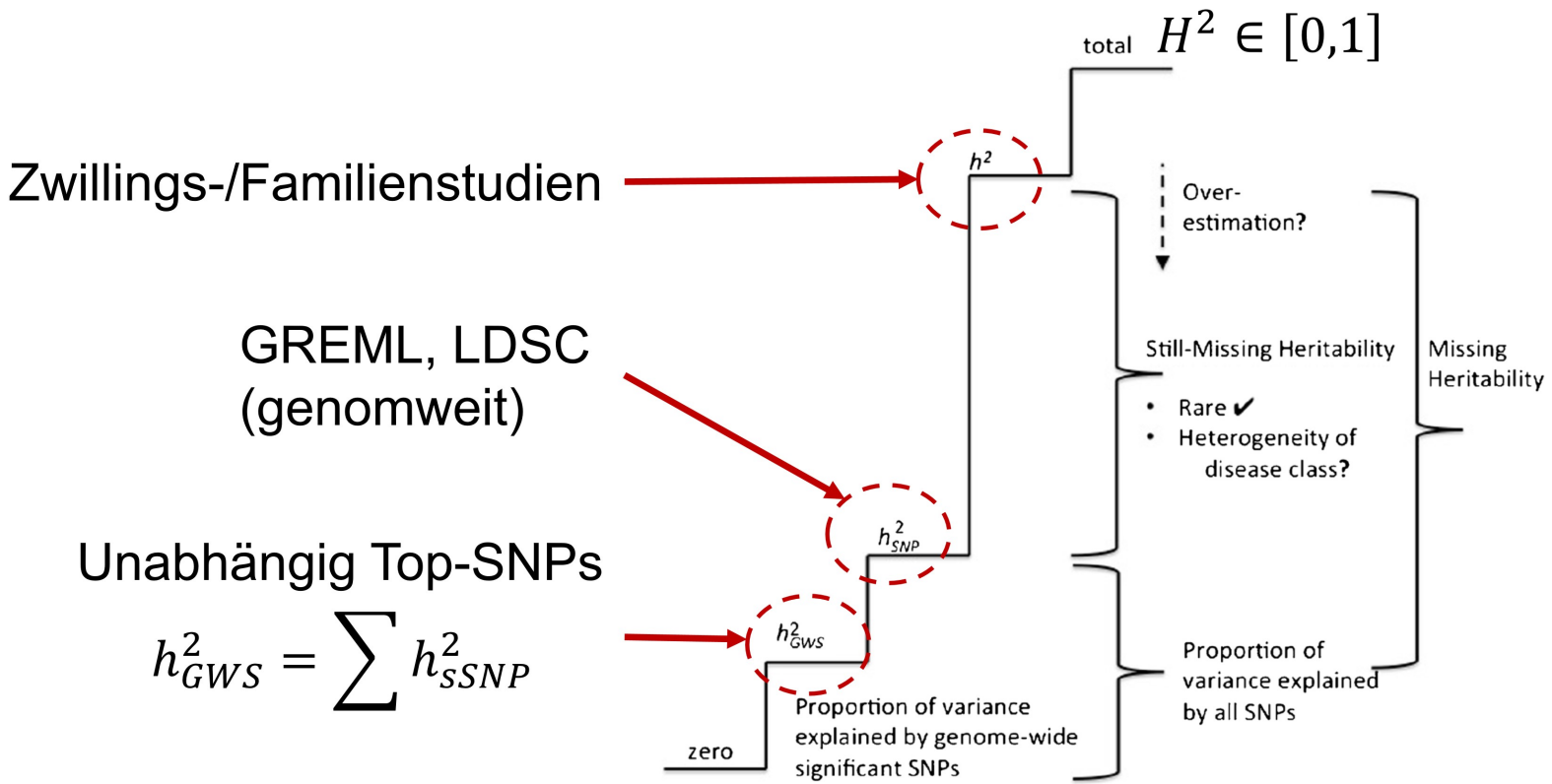
Heritabilität: Anteil der Varianz eines Merkmals, der durch die Genetik erklärt wird.
Beantwortet in wie fern Gene den Unterschied (Varianz) einer Eigenschaft erklären, **NICHT** welche Gene die Eigenschaft beeinflussen.

$$h^2 = \frac{Var(G)}{Var(Merkmal)} = \frac{Var(G)}{Var(G) + Var(U) + 2 \cdot Cov(G, U)}$$

Methoden

- Verwandtschaftstudien bzw Zwillingsstudien:
  - Falconers Formel $h^2 = 2 \cdot (r(MZ) - r(DZ))$
  - Vergleich der Merkmalskonkordanz zwischen monozygoten (MZ) und dizygoten (DZ) Zwilligen
- Querschnittsstudien von unverwandten Personen
  - Genetik-Daten vorhanden: GREML (z.B. in GCTA implementiert)
  - Nur Summary Statistics vorhanden: LD Score Regression (python-basiert, bislang nur für weiße Europäer/Amerikaner etabliert)

Lösung 1d



Die 40% laut Literatur kommen vermutlich aus einer Zwillingsstudie (genomweit, h^2).
Die 4% kommen von einem einzelnen SNP (rs123456, h^2_{sSNP})

Aufgabe 2 - Ausgangslage

- Summary Statistics für X (s. Aufgabe 1)
- Hits in Genregion ABC
- Kollege
  - Genomweite Analyse für Krankheit Y (in unabhängige Studien)
  - ebenfalls Hits in Genregion ABC
- Ihre Rolle
  - Verantwortlich für Analysen in Studie 1
  - Interessiert an Zusammenhang zwischen ABC und X
  - Wollen auf kausale Beziehung zwischen Risikofaktor X und Krankheit Y testen

Aufgabe 2a: Grundlagen Mendelischer Randomisierung

Was ist die Idee der MR und welche drei Bedingungen müssen dafür gelten?

Lösung 2a

Ziel: Detektion eines kausalen Effekts von X auf Y
Randomisiert: elterliche Allele zufällig bei Meiose + zufällige Kombination von paternalen und maternalen Allelen
Bedingungen:
(IVs = Instrumentale Variablen, die in der MR genutzt werden)

- IVs sind mit X assoziiert
  - Das haben Sie in Ihrer GWAS gezeigt
- IVs sind unabhängig von möglichen Confoundern U
  - In der Regel nur plausibilisierbar
  - Abgleich mit Datenbanken wie dem GWAS Katalog (welche anderen Phänotypen sind für diese SNPs)
- IVs sind unabhängig von Y bis auf seinen Effekt auf X
  - In der Regel nur plausibilisierbar
  - Der Effekt, den Ihr Kollege beobachtet, sollte also von dem Effekt kommen, den Sie schon in Ihrer GWAS gesehen haben.

Lösung 2a

Idee MR: der durch die *IVs* erklärte Effekt von X auf Y ist ein kausaler Schätzer. Modell:

$$Y \sim \beta_{IV} \cdot X = \beta_{IV}(\beta_X \cdot G) = \beta_Y \cdot G$$

$$\hat{\beta}_{IV} = \frac{\beta_Y}{\beta_X}$$

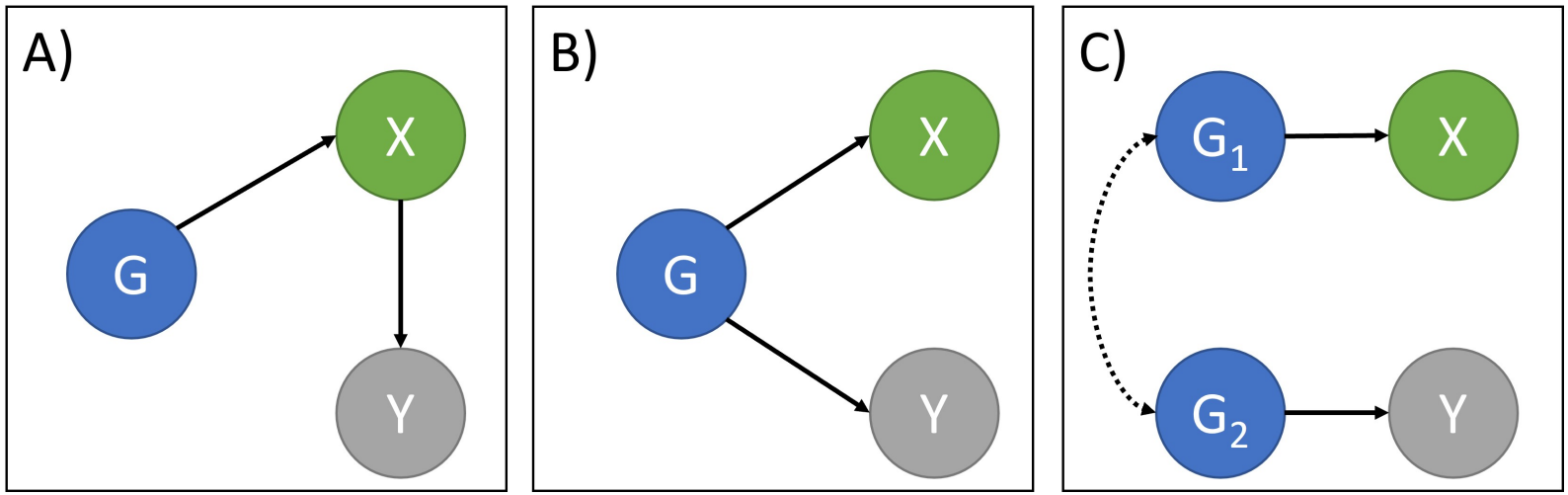
Den Standardfehler kann man mittels Jackknife oder Delta-Methode abschätzen.

Unterschied zu Randomisierter Studie:

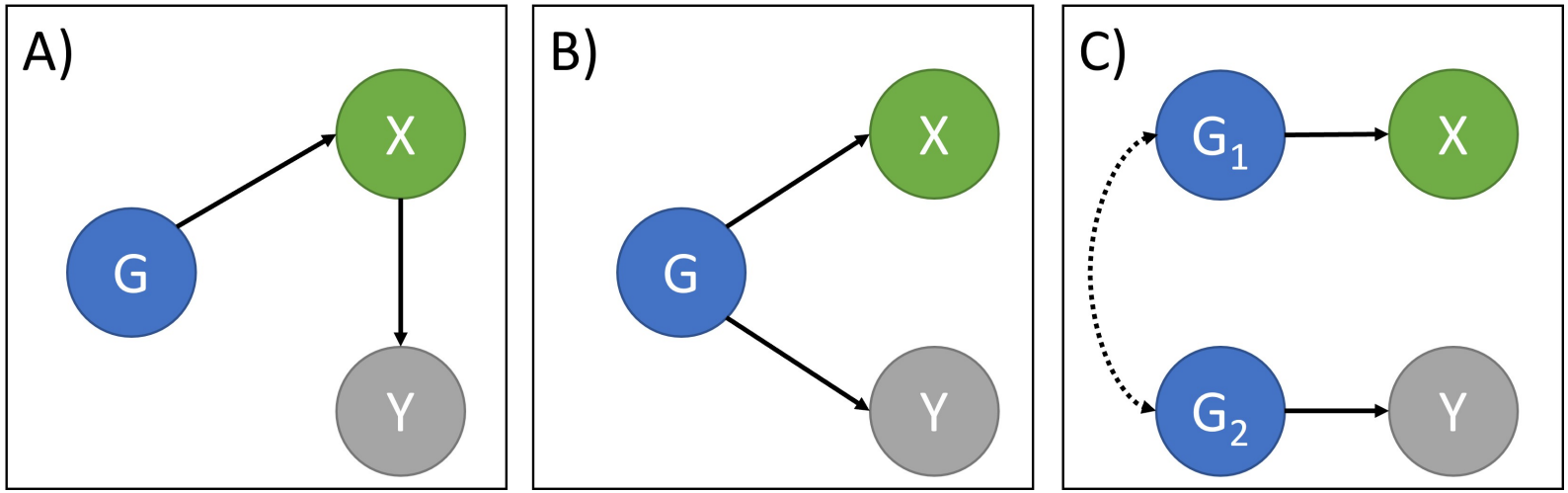
- Statt Einteilung in Medikament vs Placebo Einteilung anhand der Risiko-Allele.
- Lebenslange Wirkung in der Genetik, temporäre Wirkung eines Medikaments

Aufgabe 2b: DAGs

Erläutern Sie das jeweilige Szenario!



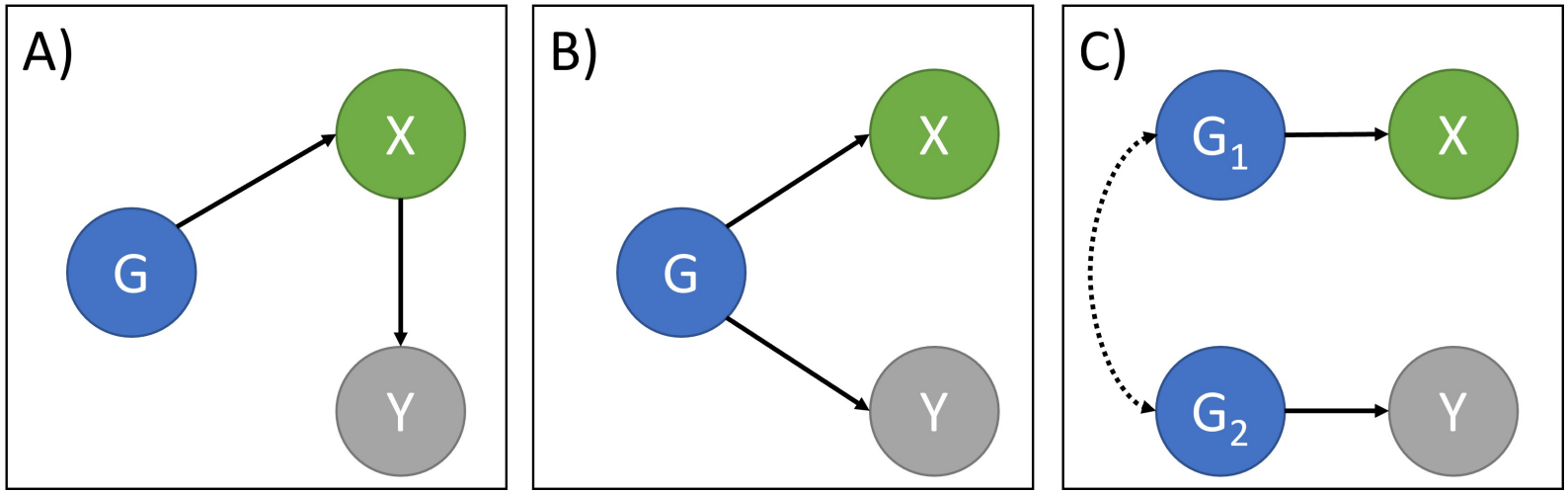
Lösung 2b - Panel A



G hat einen Effekt auf X, und X hat einen kausalen Effekt auf Y. Der SNP-Effekt auf Y kommt nur über die kausale Struktur zustande.

- Dies ist ein typischer, **valider** DAG für eine MR
- Ideale Situation
- Die MR würde zu einem wahr-postiven Ergebnis führen!

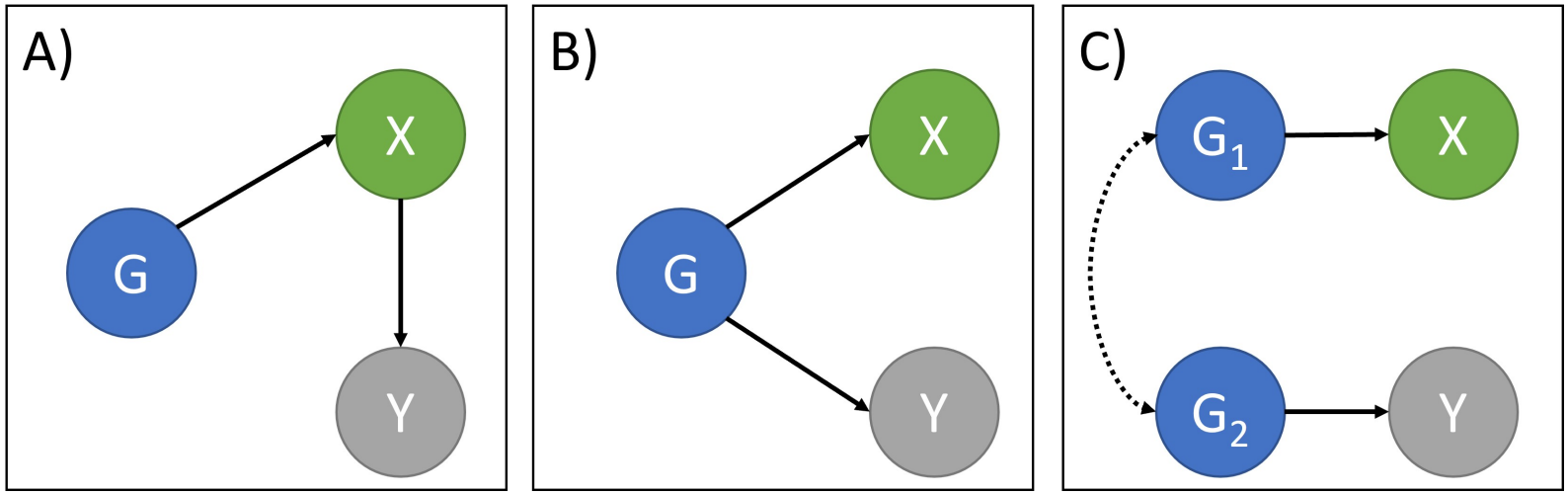
Lösung 2b - Panel B



Der SNP G hat einen pleiotropen Effekt, d.h. er beeinflusst sowohl X als auch Y unabhängig voneinander; es besteht keine kausale Beziehung zwischen X und Y.

- Das ist ein **invalider** DAG für eine MR!
- Horizontale Pleiotropie
- Die MR würde zu einem falsch-postiven Ergebnis führen!

Lösung 2b - Panel C



Zwei unterschiedliche SNPs, G_1 und G_2, haben einen Effekt auf X bzw Y, wobei keine kausale Beziehung zwischen X und Y besteht. Die beiden SNPs sind jedoch in LD miteinander.

- Das ist ein **invalider** DAG für eine MR!
- Counfounding durch LD
- Die MR würde zu einem falsch-postiven Ergebnis führen!

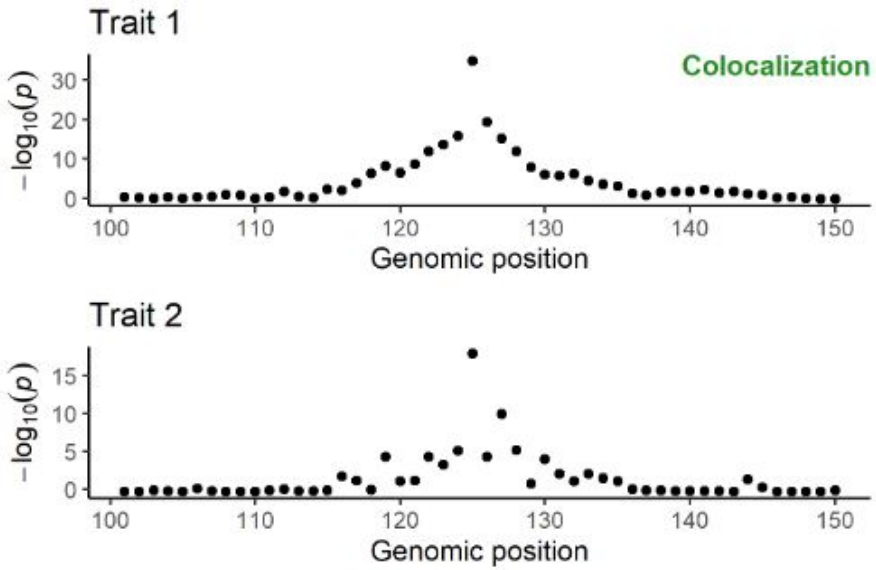
$\Rightarrow$ es reicht also nicht aus den besten SNP auf Pleiotropie zu prüfen, auch alle Varianten in LD müssen berücksichtigt werden!

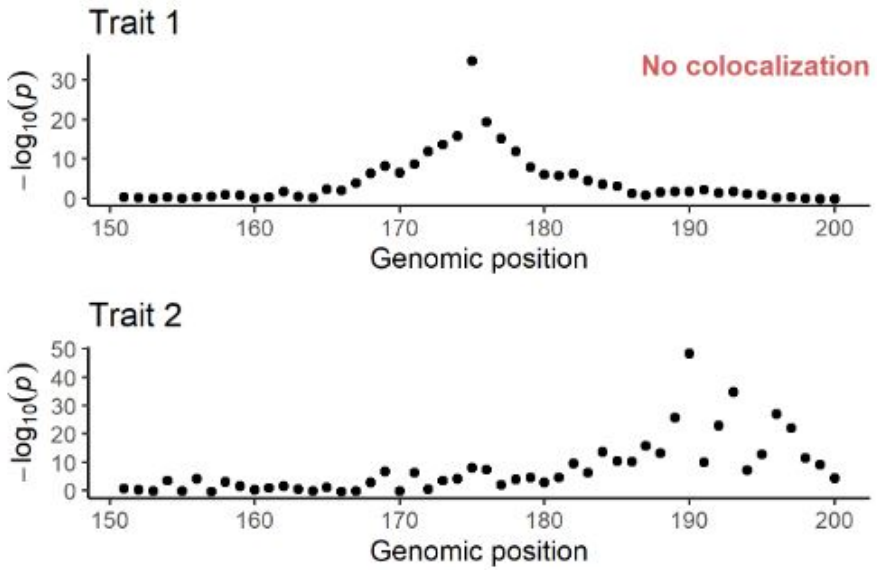
Aufgabe 2c: Grundlagen Colokalisierung

Ihr Kollege schlägt als Sensitivitätsanalyse der MR einen Test auf Colokalisierung vor.
Was ist unterdiesem Begriff zu verstehen und wie unterscheidet sich die Analyse von der MR?

Lösung 2c

Colokalisierung: Vergleich der genetischen Assoziationen zweier Phänotypen am gleichen genetischen Lokus





Lösung 2c

- unabhängig von MR
- agnostisch zu Effektrichtung bzw. kausale Beziehung
- in drug-targeted Analysen: Auswahl der Genregion anhand eines Risikofaktors, der von dem Medikament beeinflusst wird. Wir geben der Analyse daher eine Richtung vor
- $\rightarrow$ Colokalisierung kann helfen die **invaliden DAGs** zu erkennen!
  - Panel A: hohe PP für gemeinsames Signal
  - Panel B: hohe PP für gemeinsames Signal (falsch-positiv, aber im Kontext der Medikamenten-Entwicklung selten!)
  - Panel C: hohe PP für zwei unabhängige Signale

Zusammenfassung

- Typische GWAS Regression
  - Annahme additiver SNP-Effekt
- Typische Meta-Analyse
  - Fixed Effect Model (FEM) unter der Annahme, dass der SNP-Effekt in allen Studien gleich ist (keine Heterogenität)
- Stratifikationsbias
  - Inflations der Teststatistiken aufgrund fehlender Korrektur auf Populationsstruktur
- Heritabilität
  - Varianz die durch die Genetik erklärt wird
- Mendelische Randomisierung
  - Detektion eines kausalen Effekts anhand der genetisch-vorhergesagten Werte von X und Y
- Colokalisierung
  - Vergleich der genetischen Assoziationen zweier Phänotypen X und Y am gleichen genetischen Lokus