Genetische Statistik

Präsenzübung 9: GX-Analysen

Dr. Janne Pott (janne.pott@uni-leipzig.de)

January 18, 2022

Fragen

Gibt es Fragen zu

- Vorlesung?
- Übung?
- Seminar?

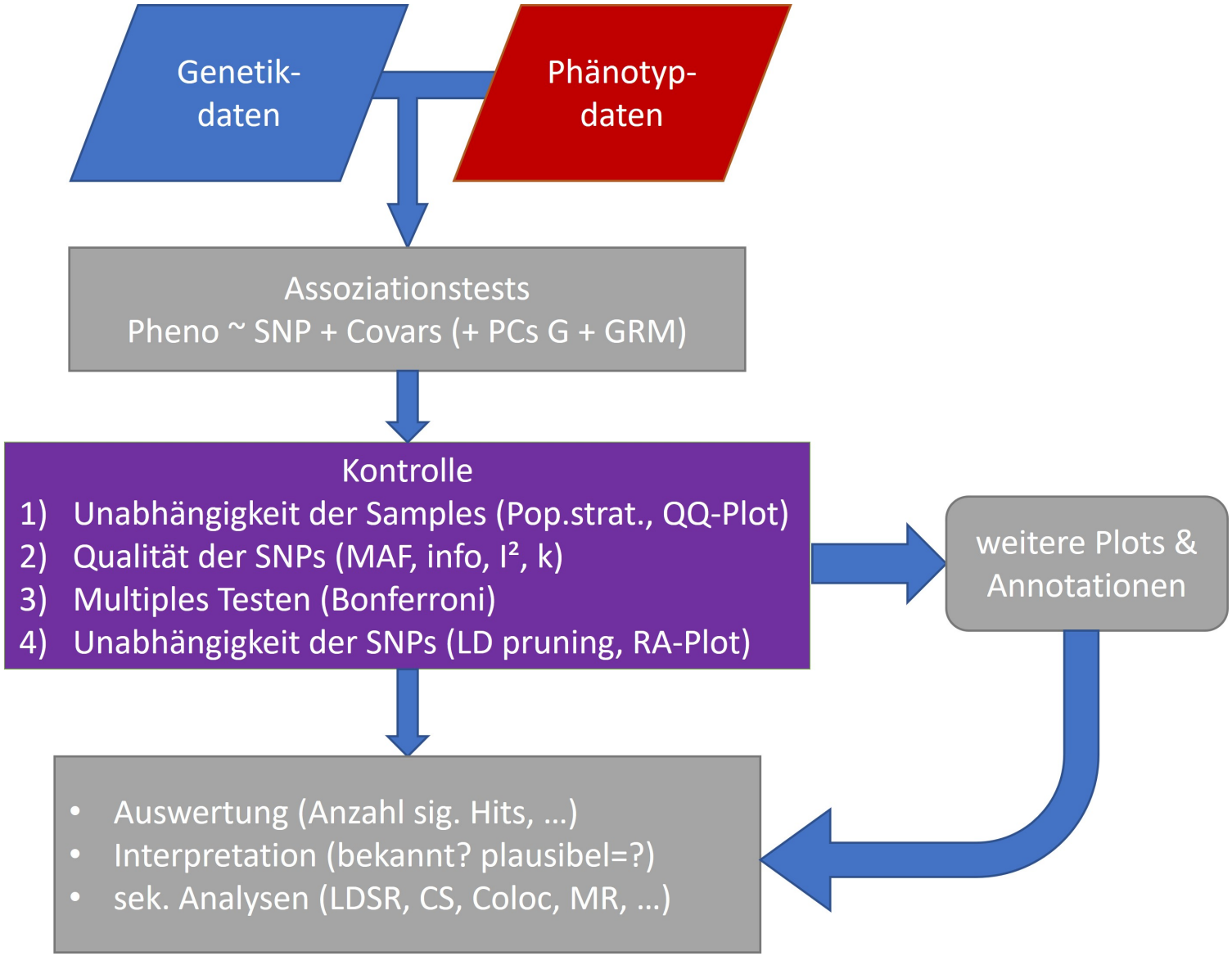
Plan heute

- Blatt 6 - A1: Genexpressionsanalysen
- Blatt 6 - A2: Pathway-Analysen
- Blatt 6 - A3: Hierarchische Testkorrektur
- Blatt 6 - A4: Interpretation von Plots

Abschnitt 1

Genexpressionsanalysen

Wiederholung GWAS-Workflow



Aufgabe 1

a) **eQTL** vs. **TWAS**
b) **cis** vs. **trans** eQTLs
c) Warum adjustiert man auf Lymphozyten und Monozyten?
d) Wie adjustiert man auf technische bzw. biologische Confounder?
e) Skizzieren Sie den Ablauf einer eQTL-Analyse!

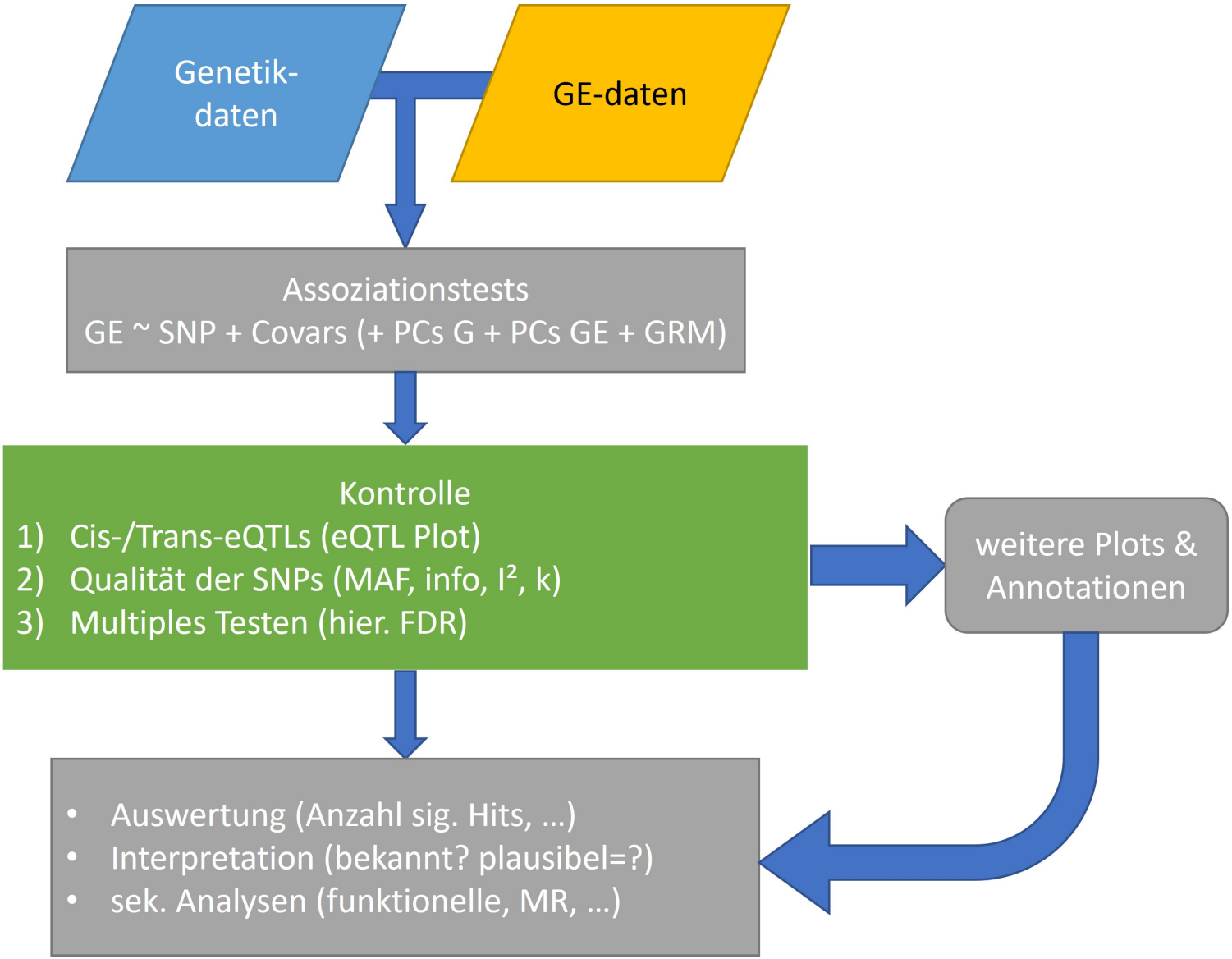
Aufgabe 1 - Lösung a) & b)

- **eQTL**:= Expression Qantitative Trait Locus $\Rightarrow GX \sim G$
- **TWAS**: Transkriptionsweite Assoziationsstudie $\Rightarrow$ *Phänotyp* $\sim GX$
- Ziel eQTL:
  - Funktionelle Relevanz, Validierung
  - Identifizierung neuer genetischer Risikofaktoren
  - Aufklärung grundlegender biologischer Zusammenhänge
- Ziel TWAS:
  - Identifizierung von Gen - Phänotyp - Beziehungen
  - Pathwayanalysen
- **Cis**: SNP „in Nähe“ der Expressionssonde (1 MB Fenster)
- **Trans**: SNP „weit weg“ von Expressionssonde (mehr als 1 MB, auch andere Chromosomen)

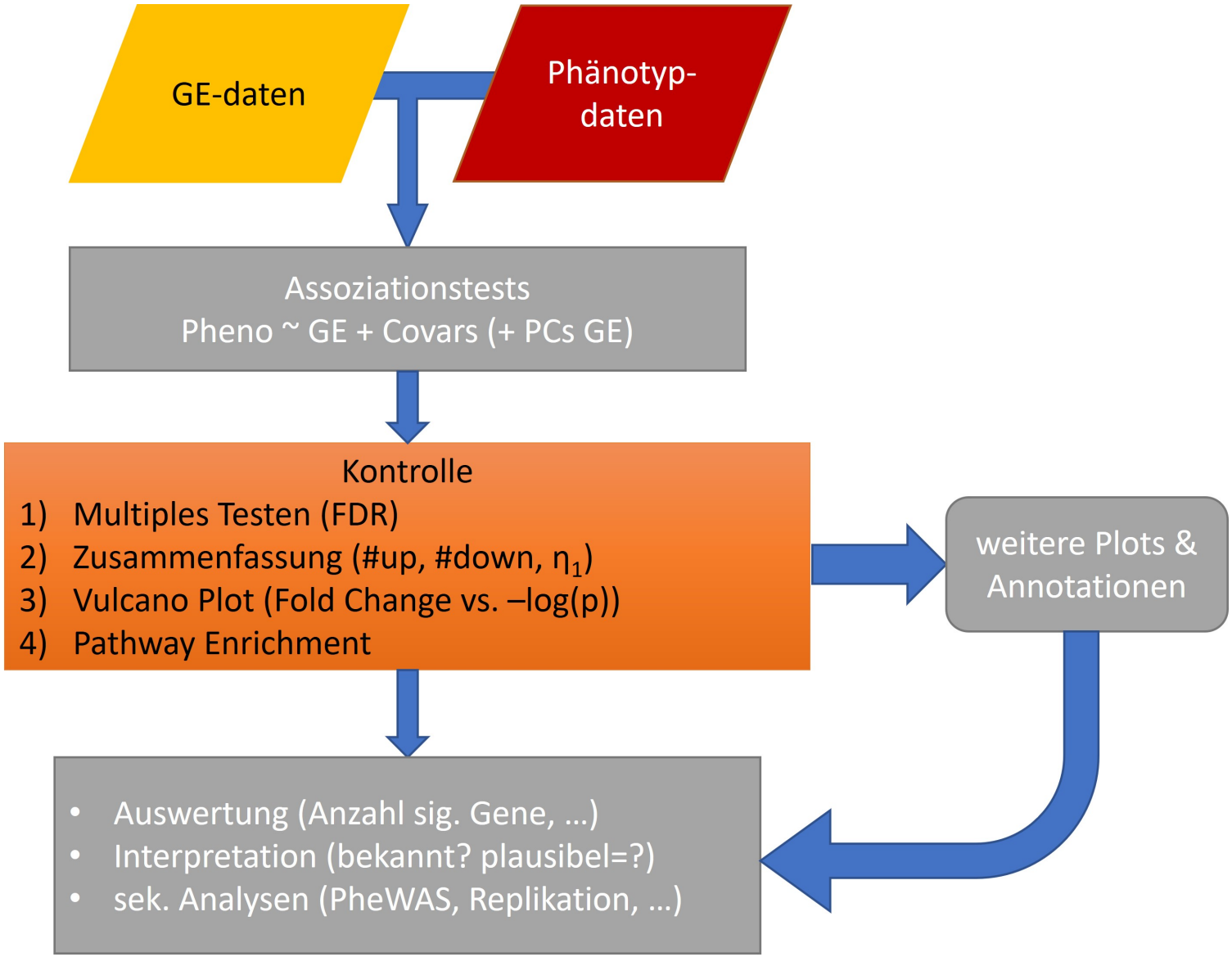
Aufgabe 1 - Lösung c) & d)

- **Technische Confounder**: z.B. Batcheffekte, Hintergrundrauschen, sollte in der Präprozessierung der GX-Arrays bereits korrigiert werden.
- **Biologische Confounder**: können meist als Modellparameter berücksichtigt werden:
  - **Blutwerte**: nötig bei Blutgewebe da verschiedene Blutkörperfraktionen mit unterschiedliche GE
  - Alter, Geschlecht, Medikamente: bekannter Einfluss (bsp. Sexualhormone sind Transkriptionsfaktoren)
  - Weitere, unbekannte Confounder: PCA der GE

Aufgabe 1 - Lösung e)



Aufgabe 1 - Zusatz



Abschnitt 2

Pathwayanalysen

Aufgabe 2

Frage: $GX \sim Medikament?$, $N = 15,397$ GEs ($k = 1,042$ sig. assoziiert, $m = 2,587$ im Lipidpathway)

Wie viele signifikante Gene müssten im Lipidstoffwechsel liegen, um von einer signifikanten Anreicherung ausgehen zu können? Gehen Sie dazu von einer hypergeometrischen Verteilung aus.

Hintergrund (1)

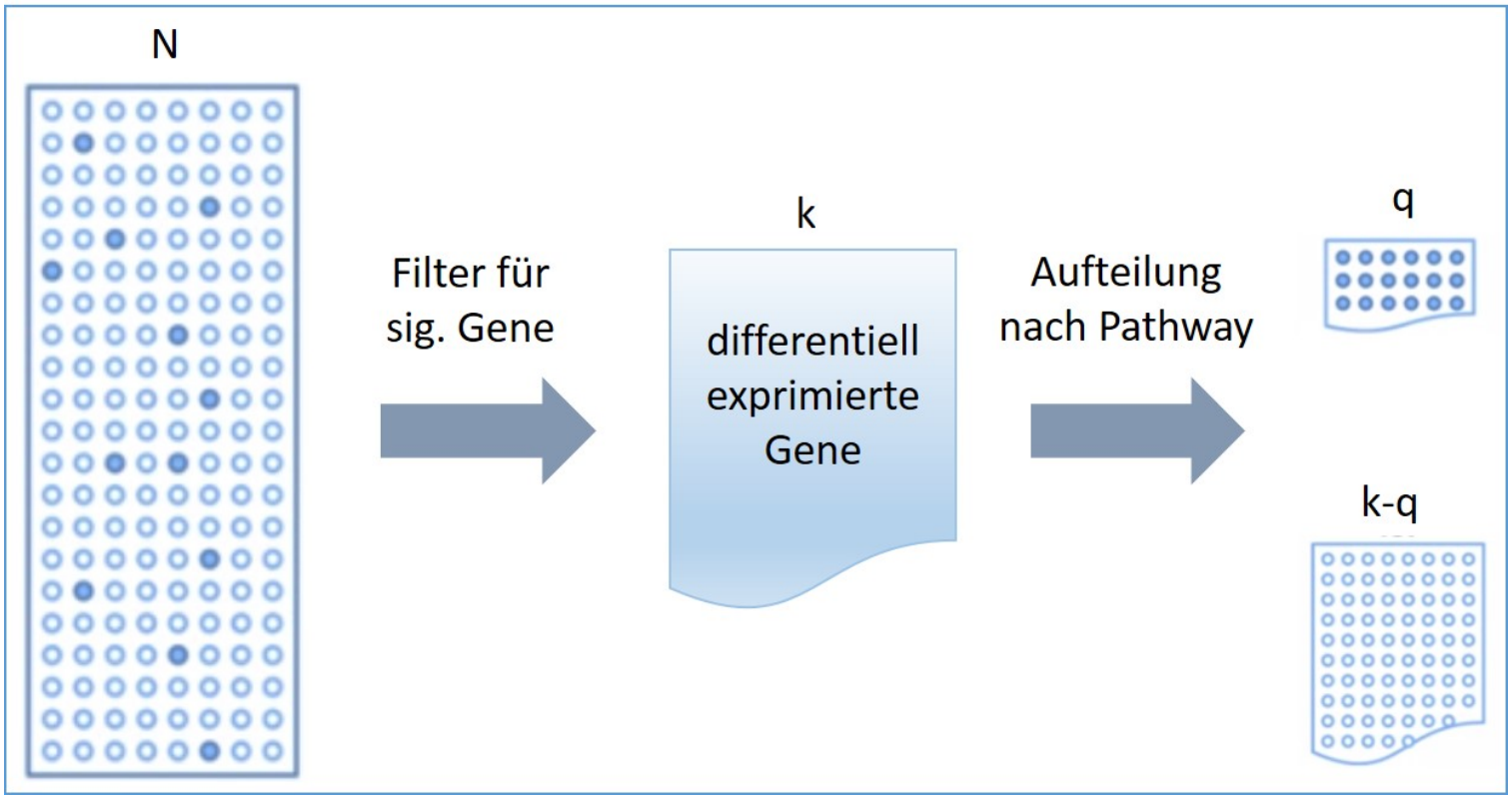


Abbildung 1: Schema einer Überrepräsentationsanalyse.

Hintergrund (2)

Idee: Urne mit $N - m$ schwarzen und m weißen Kugeln, aus der k mal ohne zurücklegen gezogen wird. Wie hoch ist die Wahrscheinlichkeit, dass q Kugeln weiß sind?

- $N = 15,397 =$ Größe der Grundmenge (alle Gene),
- $k =$ Größe der Stichprobe, $k \leq N$ (differenziell exprimierte Gene), und
- $m =$ Größe einer spezifischen Teilmenge von N, $m \leq N$ (Gene eines Pathway),

so gilt für die Wahrscheinlichkeit, dass die Stichprobe k Elemente von der spezifischen Teilmenge enthält

$$P(q) = \frac{\binom{m}{q}\binom{N-m}{k-q}}{\binom{N}{k}}$$

Aufgabe 2 - Lösung

- $N = 15,397, m = 2,587, k = 1,042$,
  $m/N = 16.8\% \rightarrow 16.8\% \cdot k = 175$,
- Bei $q = 175$ Treffern ist die Ratio gleich (= Erwartung, keine Anreicherung).
- Suche nach $q > 175$, sodass $P(q > x) > 0.95$ (Anreicherung, mehr Treffer als zufällig erwartet).

```
qhyper(0.95,2587,15397-2587,1042)
```

```
## [1] 194
```

- $P(x \leq 194) = 0.95 \rightarrow$ ab $q \geq 195$ signifikanten Anreicherung

Abschnitt 3

Hierarchische Testkorrektur

Aufgabe 3

Hierarchische Korrektur mittels Bonferroni für folgende Daten:

Gen 1 |x || Gen 2 |x || Gen 3 |x —— | ——— || —— | ——— || —— | ——— SNP | p-Wert || SNP | p-Wert || SNP | p-Wert rs1001 | 0.05 || rs2001 | 0.1 || rs1004 | 0.0124 rs1002 | 0.04 || rs2002 | 0.2 || rs1005 | 0.2 rs1003 | 0.005 || rs2003 | 0.04 || rs1006 | 0.0025 rs1004 | 0.4 || rs2004 | 0.0001 || rs2001 | 0.5 rs1005 | 0.3 || rs2005 | 0.004 || rs2002 | 0.00001 rs1006 | 0.8 || rs2006 | 0.02 || rs2003 | 0.054 | || rs2007 | 0.00005 || —— | ——— || —— | ——— || —— | ———

Hintergrund

- **Multiples Testproblem**: Alphafehler-Kumulierung (globale Erhöhung des Fehler 1. Art, mehr falsch positive). *Hierarchisches FDR:
  1) Adjustierung Genebene: n_i Tests pro Transkript
  2) min. adj. p-Wert pro Transkript bestimmen
  3) Adjustierung global: n Gene
  4) Bestimmung der Anzahl k der sig. assoziierten Transkripte mittels Step 3
  5) Bestimmung des globalen Sig.niveaus $\alpha_1 = 0.05 \cdot k/n$
  6) Anwendung von α_1 auf die adjustierten p-Werte von Step 1