Funktionelle Genomanalysen 2023

Übung 1: Grundlagen der genetischen Statistik

Dr. Janne Pott

09.-11. Juni 2023

Vorstellung

- Name
- Fachrichtung
- Standort
- Erwartung an Übung

Vorstellung

- Name: Janne Pott
- Fachrichtung
  - Genetische Statistik im Allgemeinen
  - Entwicklung neuer kausalen Methoden unter Verwendung von Genetikdaten im Speziellen (Stichwort **Mendelische Randomisierung**)
- Standort: MRC BSU, University of Cambridge, UK
- Erwartung an Übung
  - Veranschaulichung einiger Konzepte
  - Hinweise auf praktische Anwendung
  - Kontakt zu Medizinern

Hinweise zu Moodle

- Alle relevanten Unterlagen stehen auf Moodle zur Verfügung
  - Übungsaufgaben
  - Folien
  - Video (vermutlich)
- Asynchrones Lernen: Multiple Choice Aufgaben zu
  - SNP-Clusterplots
  - GWAS-Plots
  - ...
- Forum: bitte Nutzen, andere haben evtl. die gleiche Frage!

Hinweise zur Übung

- Die Aufgaben werden in der Übung gemeinsam erarbeitet, daher bitte **Kamera an**.
- Zur Lösung von manchen Aufgaben wird ein Taschenrechner o.ä. benötigt.
- Am Ende des Moduls wird eine Musterlösung bereitgestellt.

Präsenzaufgabe 1

Definieren Sie folgende Begriffe

- SNP
- nicht-synonyme Mutation
- Frameshift-Mutation

Präsenzaufgabe 1 - Lösung

- SNP: single nucleotid polymorphism = Einzelnukleotid Polymorphismus = Punktmutation
  - Variation eines Basenpaares an einer Stelle im Genom
  - Bsp.: SNP in mcm6 führt zu Laktoseintoleranz
- nicht-synonyme Mutation: SNP im codierenden Bereich eines Gens, der dazu führt, dass eine andere Aminosäure eingebaut wird.
  - betrifft nur eine Aminosäure
- Frameshift-Mutation: SNP im codierenden Bereich eines Gens, der dazu führt, dass das Leseraster der RNA-Polymerase sich ändert.
  - betrifft alle folgenden Aminosäure bzw Länge des Proteins

Aufgabe 1: Crossing-over & LD (1)

1. Definieren Sie anhand der Abbildung 1 den Begriff **Crossing-over**.
2. Erläutern Sie den Zusammenhang zwischen der **Crossing-over** und **LD-Struktur** des Genoms.
3. Betrachten Sie Tabelle 1. Bestimmen Sie die **Randverteilungen** und berechnen Sie das **LD-Maß** r^2! Formel:

$$r^2 = \frac{(p_{00}p_{11} - p_{01}p_{10})^2}{p_{0.}p_{.0}p_{1.}p_{.1}}$$

4. Interpretieren Sie das Ergebnis! Was sind die **häufigen Haplotypen**? Was bedeutet dies für ein doppelt heterozygotes Individuum?
5. Würden Sie zwischen SNP 1 und SNP 3 ein höheres oder niedrigeres r^2 erwarten? Begründen Sie Ihre Entscheidung!

Aufgabe 1: Crossing-over & LD (1)

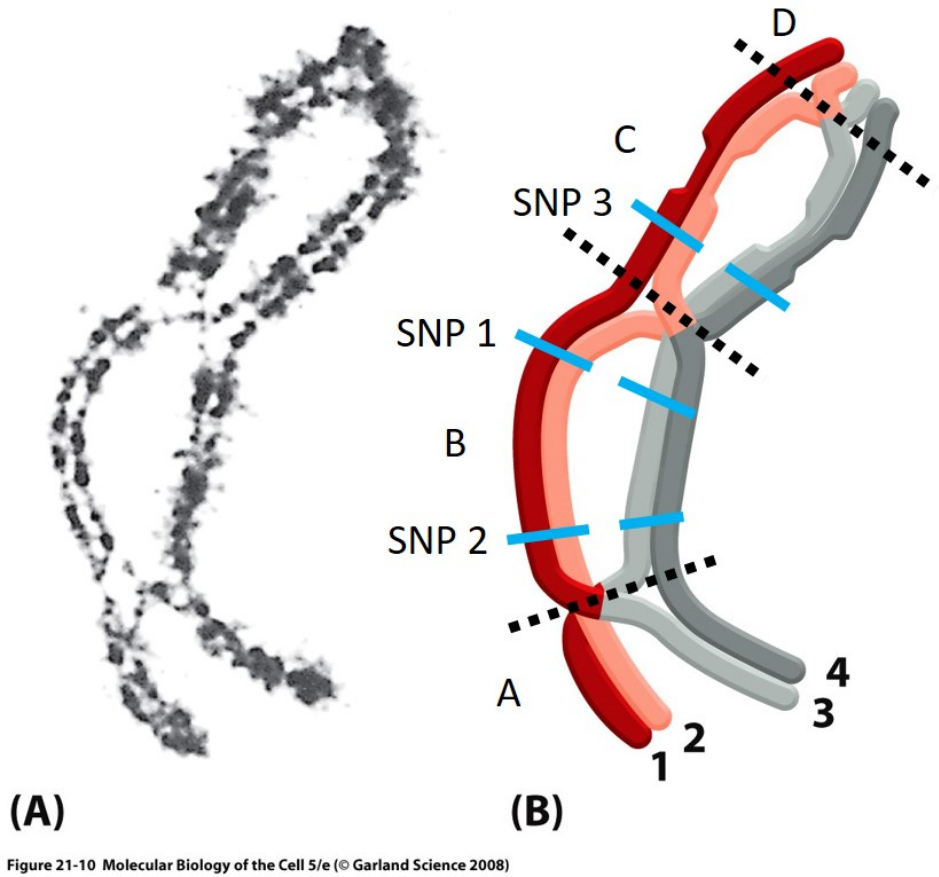

Figure 21-10 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Abbildung: Crossing-over eines Chromosoms. Modifiziert aus Alberts et al., Molecular Biology of the Cell, 2008

Aufgabe 1: Lösung (1)

Crossing-over: Austausch von Teilen homologer Chromosomen während der Meiose $\longrightarrow$ Rekombination, Mischung der Erb-Informationen der Eltern

- Step 1: DNA Replikation der väterlichen und mütterlichen DNA
- Step 2: Alignierung der duplizierten homologen Chromosomen
- z.T. Interaktion zwischen komplementären DNA-Sequenzen

Aufgabe 1: Lösung (2)

Haploblöcke wechseln sich mit **Rekombinations-Hotspots** ab. Dort finden die Crossing-overs mit erhöhter Wahrscheinlichkeit statt.

- SNPs im gleichen Haploblock haben tendenziell hohes paarweises LD (werden häufiger gemeinsam vererbt, statistisch **abhängig** voneinander)
- SNPs in unterschiedlichen Haploblöcken haben tendenziell niedriges LD (werden seltener gemeinsam vererbt, statistisch **unabhängig** voneinander)

Aufgabe 1: Lösung (3)

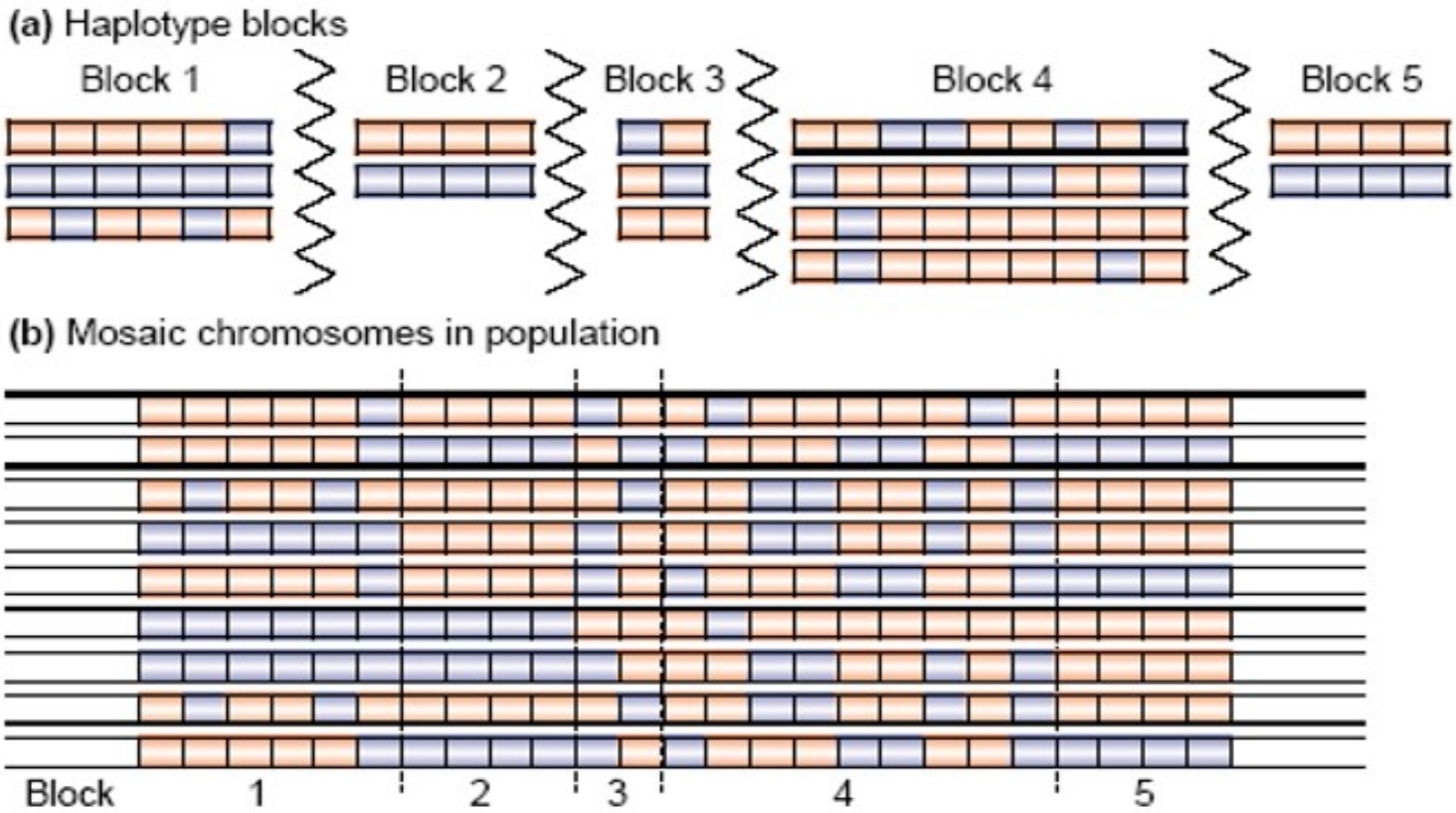

Abbildung: LD-Struktur des Genoms.

Aufgabe 1: Crossing-over & LD (3)

Tabelle: 4-Felder-Tafel der beiden biallelischen SNPs: SNP 1 (Allele A1/B1) und SNP 2 (Allele A2/B2) aus Daten von 500 gemessenen diploiden Individuen

	SNP 1 - Allel A1	SNP 1 - Allel B1
SNP 2 - Allel A2	570	15
SNP 2 - Allel B2	25	390

$$r^2 = \frac{(p_{00}p_{11} - p_{01}p_{10})^2}{p_{0.}p_{.0}p_{1.}p_{.1}}$$

Aufgabe 1: Lösung (4)

Tabelle: 4-Felder-Tafel der beiden biallelischen SNPs: SNP 1 (Allele A1/B1) und SNP 2 (Allele A2/B2) aus Daten von 500 gemessenen diploiden Individuen

	Allel A1	Allel B1	Randverteilung
Allel A2	p_{00}=570	p_{01}=15	$p_{0.}$=585
Allel B2	p_{10}=25	p_{11}=390	$p_{1.}$=415
Randverteilung	$p_{.0}$=595	$p_{.1}$=405	1000

$$
\begin{aligned}
r^2 &= \frac{(p_{00}p_{11} - p_{01}p_{10})^2}{p_{0.}p_{.0}p_{1.}p_{.1}} \\
&= \frac{570 \cdot 390 - 25 \cdot 15)^2}{585 \cdot 595 \cdot 415 \cdot 405} \\
&= 0.842
\end{aligned}
$$

Aufgabe 1: Lösung (5)

- Hohes LD zwischen SNP1 und SNP2, nicht statistisch unabhängig!
- Häufige Haplotypen: A1A2 und B1B2
- Selten Haplotypen: A1B2 und B1A2
- Erwartung: niedrigeres LD zwischen SNP 1 und 3

Zusatz zu LD (1)

Was ist LD und warum ist es wichtig in der Funktionelle Genomanalysen?

Zusatz zu LD (2)

- Allele sind in LD wenn sie häufiger gemeinsam vorkommen als zufällig erwartet (Korrelation)
- r^2 ist ein häufig genutztes Maß für LD
- r^2 ist zwischen 0 (kein LD) and 1 (perfekte Korrelation).
- LD kann von verschiedenen Faktoren beeinflusst werden:
  - Mutationsrate
  - Genetischer Drift
  - Nicht-zufällige Paarung (Zucht)
  - Populationsstruktur
  - Selektion
  - Genkopplung
  - Rekombinationsrate

Zusatz zu LD (3)

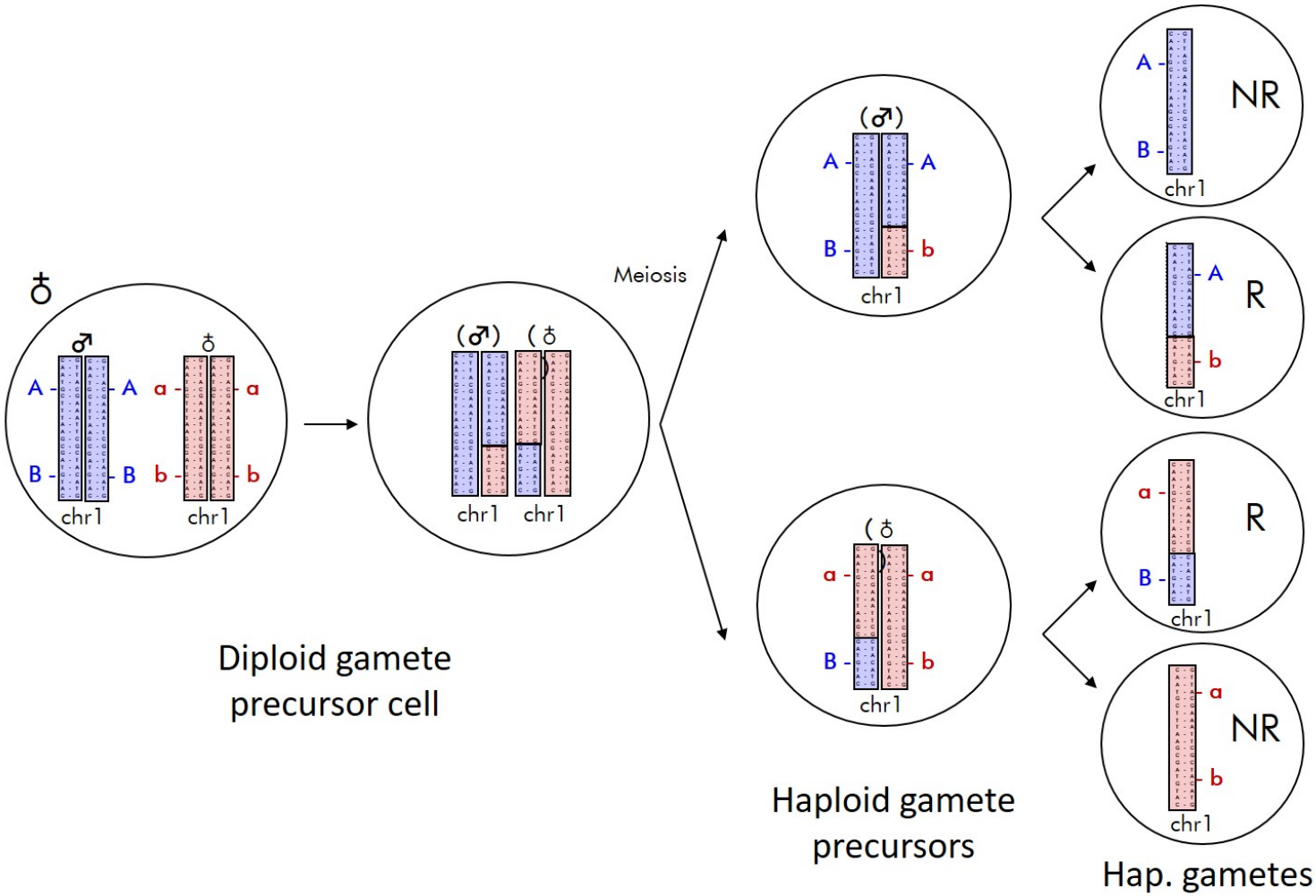

Abbildung: Recombination between unlinked loci (= not in LD).

Zusatz zu LD (4)

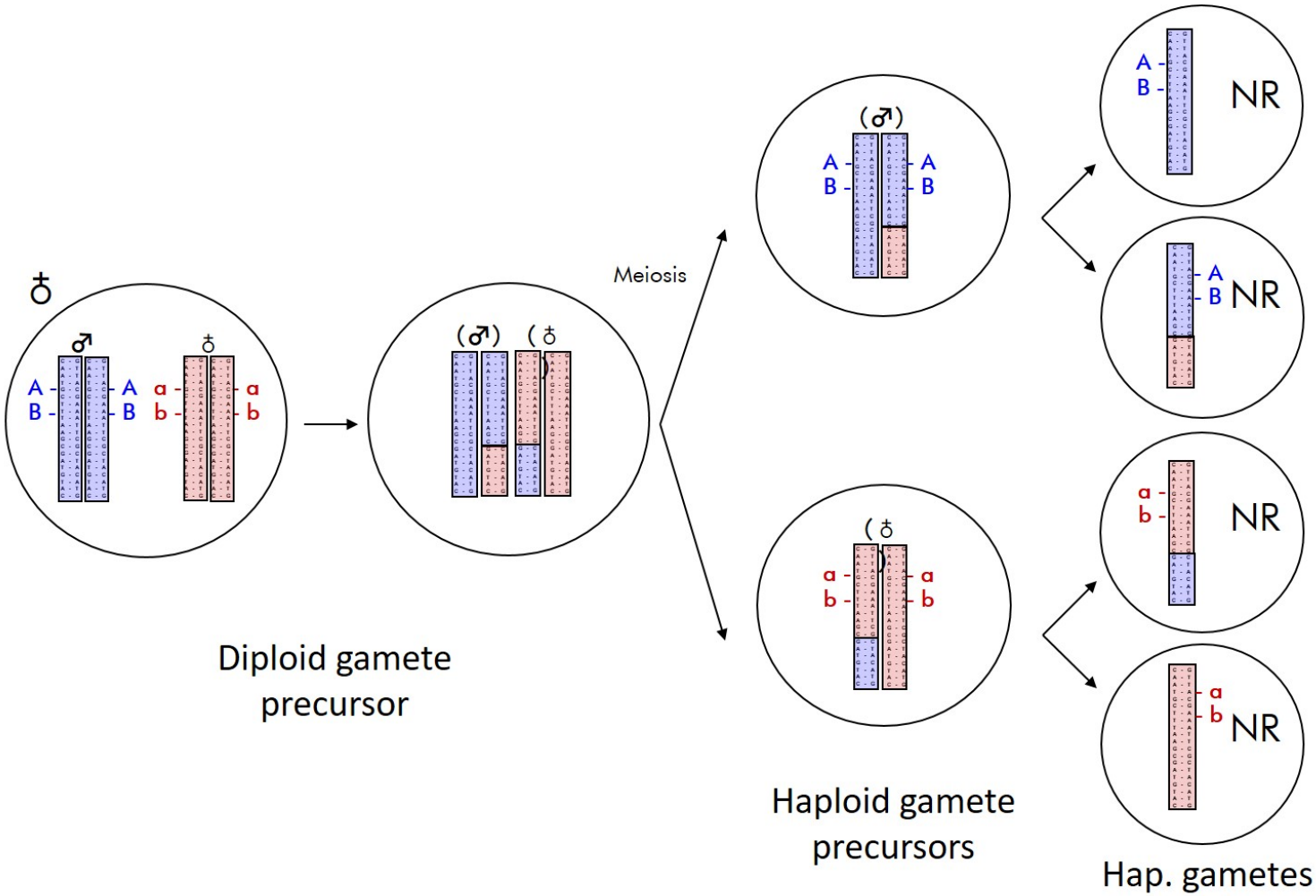

Abbildung: Recombination between linked loci (= in LD).

Aufgabe 2: HWE (1)

Für den biallelischen SNP 1 mit Allelen A/B wird folgende Genotypverteilung beobachtet:

Genotyp	AA	AB	BB	Missing
Häufigkeit	824	1326	463	87

1. Welche Modellannahmen werden Hardy-Weinberg-Gleichgewicht (HWE) getroffen (Stichwort **ideale Population**)?
2. Bestimmen Sie auf drei Nachkommastellen genau die
   1. **Callrate** des SNPs,
   2. **Allelfrequenzen** für A und B, und
   3. **erwartete Genotypverteilung** im HWE!

Aufgabe 2: Lösung (1)

Ideale Population:

- Diploide Organismen
- Nur geschlechtliche Vermehrung
- Keine Überlappung der Generationen
- Zufällige Paarungen
- Unendliche Populationsgröße
- Allelfrequenzen sind in beiden Geschlechtern gleich
- Keine Migration, Gendrift, Mutation oder Selektion

Aufgabe 2: Lösung (2)

Callrate:

$$CR = \frac{N_1}{N_0} = \frac{824 + 1326 + 463}{824 + 1326 + 463 + 87} = \frac{2613}{2700} = 0.968$$

Allelfrequenzen:

$$p = AF_A = \frac{2 \cdot AA + AB}{2 \cdot N_1} = \frac{2 \cdot 824 + 1326}{2 \cdot 2613} = 0.569$$

$$q = AF_B = \frac{2 \cdot BB + AB}{2 \cdot N_1} = \frac{2 \cdot 463 + 1326}{2 \cdot 2613} = 0.431$$

Aufgabe 2: Lösung (3)

Erwartete Genotypverteilung: Im HWE gilt:

$$1 = p + q = (p + q)^2 = p^2 + 2pq + q^2$$

Genotyp	AA	AB	BB	Missing
Häufigkeit	824	1326	463	87
p_{obs}	0.315	0.507	0.177	
p_{exp}	p^2=0.324	$2pq$=0.490	q^2=0.186	

Aufgabe 2: HWE (2)

Zusatz: Testen Sie auf HWE mit 5% Irrtumswahrscheinlichkeit. Stellen Sie dazu die **Nullhypothese** auf. Berechnen Sie die **Teststatistik** für diese und interpretieren Sie das Ergebnis.
Formel:

$$\sum_i \frac{(O_i - E_i)^2}{E_i}, i \in AA, AB, BB$$

Aufgabe 2: Lösung (4)

Nullhypothese: Die beobachteten Häufigkeiten der Genotypen befinden sich im Hardy-Weinberg-Gleichgewicht.
Um die Hypothese zu testen, muss das χ^2 bestimmt werden:

$$\chi^2 = N_1 \sum \frac{(p_{obs} - p_{exp})^2}{p_{exp}} = 3.1416$$

Da $m = 2$ Allele betrachten werden, haben wir einen Freiheitsgrad:

$$df = \frac{m(m-1)}{2} = \frac{2 \cdot 1}{2} = 1 \rightarrow \chi^2_1 = 3.841$$

Das berechnete χ^2 ist kleiner als die Schranke χ^2_1, daher kann die Nullhypothese nicht verworfen werden.

Zusatz zu HWE (1)

Warum ist HWE wichtig in der Funktionelle Genomanalysen?

Zusatz zu HWE (2)

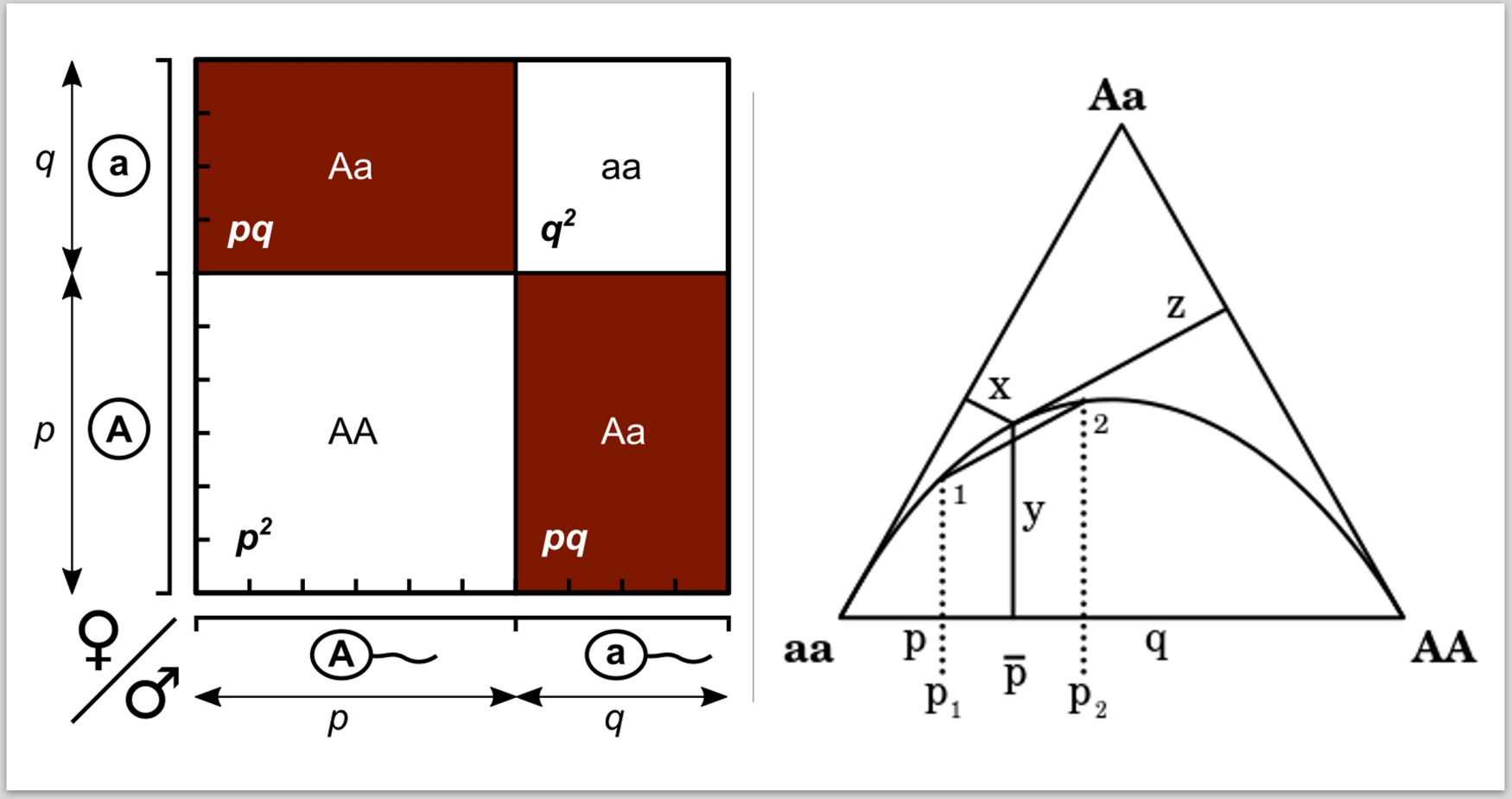

Abbildung: Punnett square and De Finetti diagram.

Aufgabe 3: Stammbäume (1)

1. Definieren Sie die Begriffe **dominant**, **rezessiv** und **Penetranz**.
2. Betrachten Sie die drei Stammbäume und geben Sie folgendes an:
   - eine Legende,
   - die Träger/in,
   - das wahrscheinlichstes Segregationsmuster mit Begründung

Aufgabe 3: Lösung (1)

- **dominant**: eine Kopie des Risiko-Allels reicht aus, um den Phänotyp zu erzeugen
- **rezessiv**: nur wenn das Risiko-Allel homozygot vorliegt, kommt es zum Phänotypen
- **Penetranz**: WSK, mit der ein bestimmter Genotyp zur Ausbildung des zugehörigen Phänotyps führt
- Legende:
  - Form = Geschlecht (Kreis: Frau; Quadrat: Mann)
  - Füllung = Phänotyp/Krankheit (Keine Füllung: gesund; rote Füllung: erkrankt; ein Punkt im Kreis oder Quadrat: Träger/in)

Aufgabe 3: Plot A

Bestimmen Sie den Erbgang des vorliegenden Stammbaums und den Genotyp aller Mitglieder!

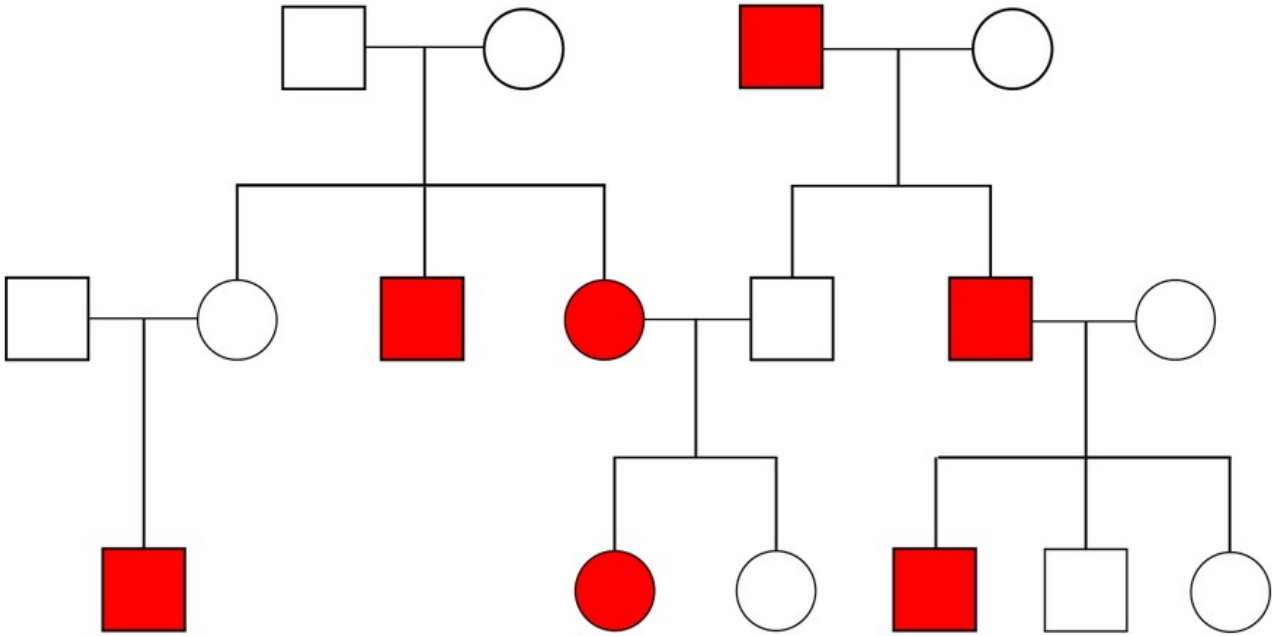

Abbildung: Stammbaum A.

Aufgabe 3: Plot A - Lösung

Lösung: autosomal-rezessiv, weil

- Beide Geschlechter betroffen
- Generationen können übersprungen werden

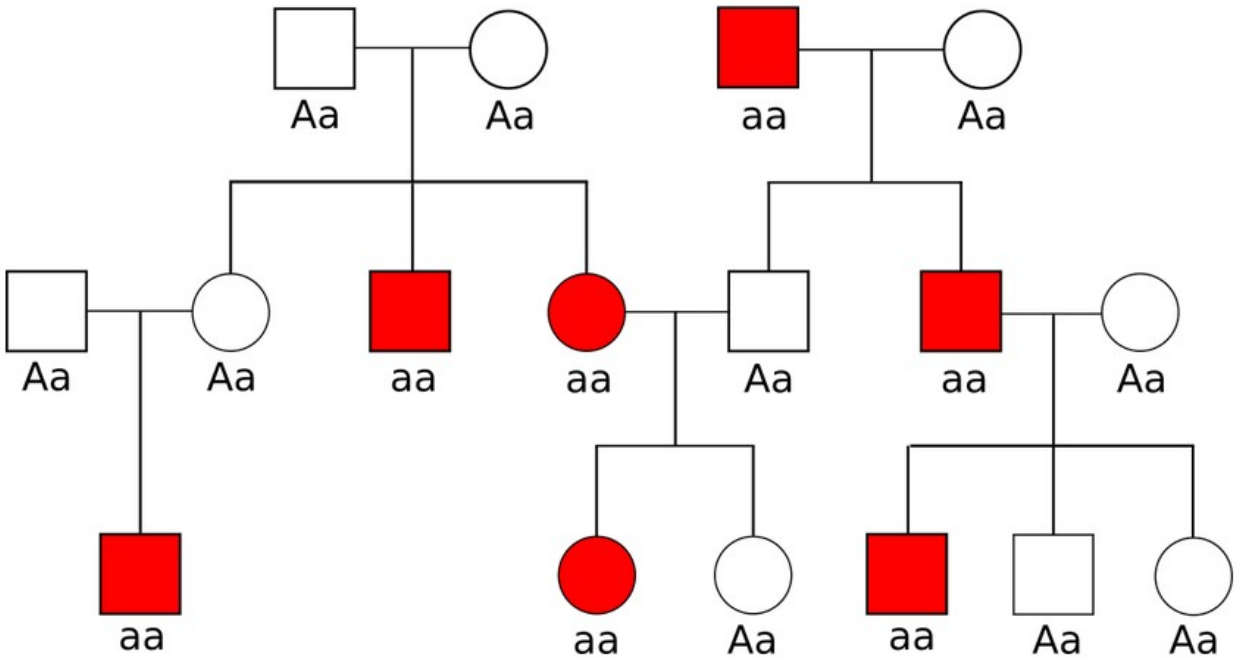

Abbildung: Stammbaum A - Lösung.

Aufgabe 3: Plot B

Bestimmen Sie den Erbgang des vorliegenden Stammbaums und den Genotyp aller Mitglieder!

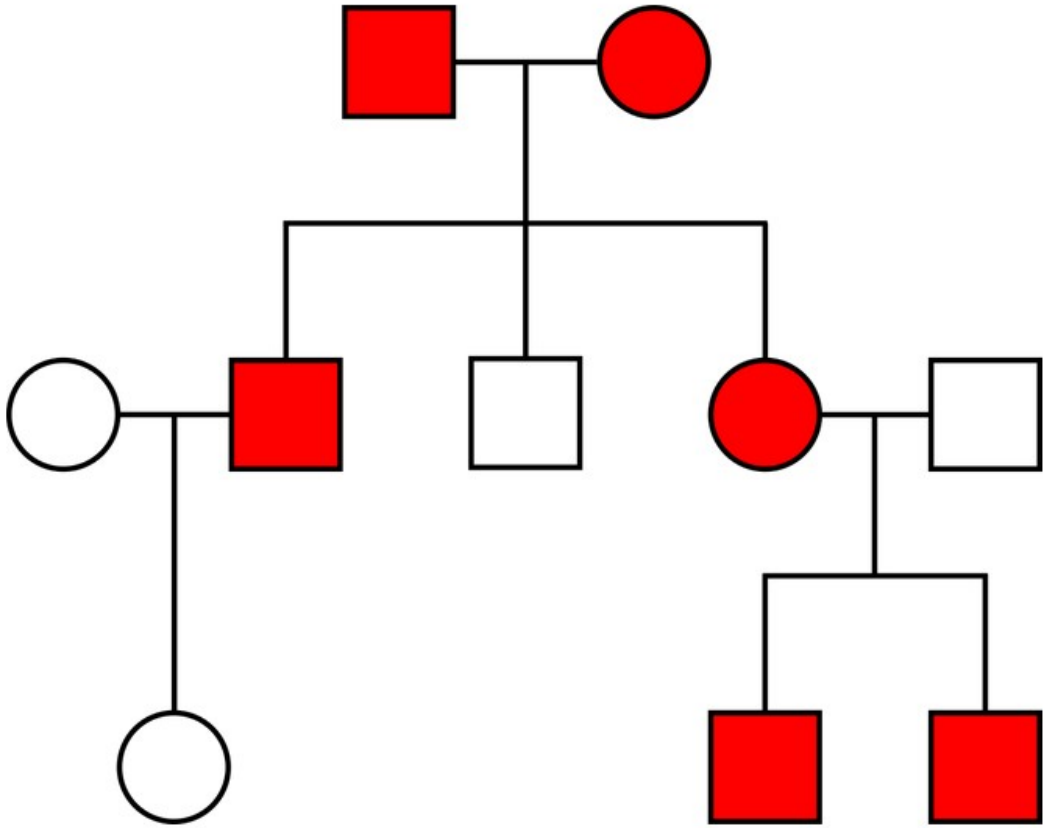

Abbildung: Stammbaum B.

Aufgabe 3: Plot B - Lösung

Lösung: autosomal-dominant, weil

- Beide Geschlechter betroffen
- jede Generation betroffen

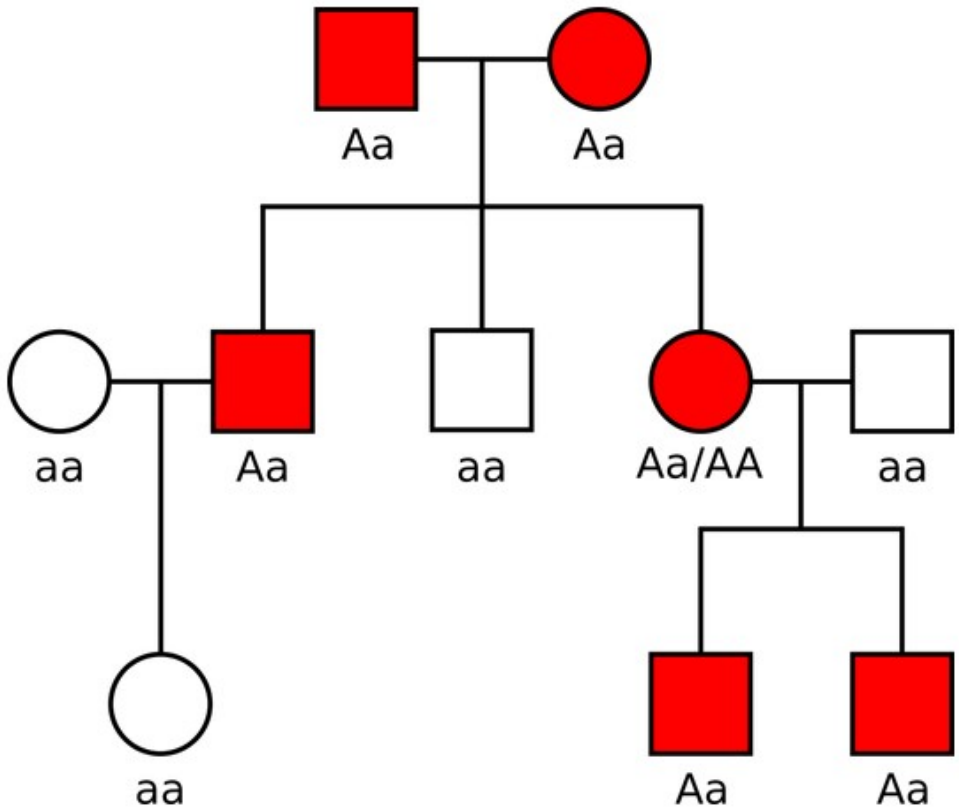

Abbildung: Stammbaum B - Lösung.

Aufgabe 3: Plot C

Kinderwunsch in Generation III. Bestimmen Sie den Erbgang und den Genotypen der Mutter. Mit welcher Wahrscheinlichkeit werden die Kinder dieses Paares erkranken?

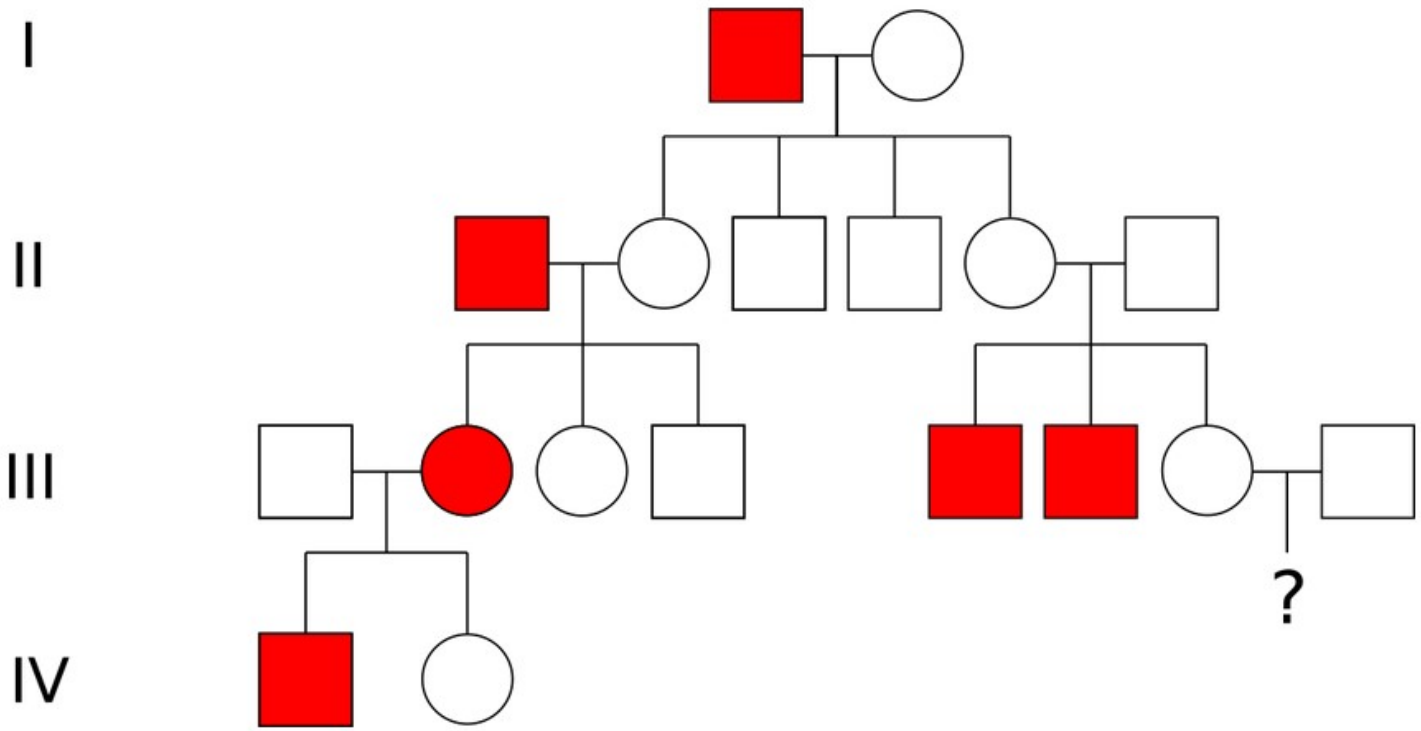

Abbildung: Stammbaum C.

Aufgabe 3: Plot C - Lösung

Lösung: X-chromosomal rezessiv, weil

- Deutlich mehr Männer betroffen
- Generationen können übersprungen werden

Mutter hat 50% Chance Trägerin zu sein

- Töchter werden alle gesund sein (können Trägerinnen sein)
- Söhne werden zu 25% erkranken (WSK(Mutter Trägerin) * WSK(rezessives Allel wird weitergegeben) = 0.5 * 0.5 = 0.25)

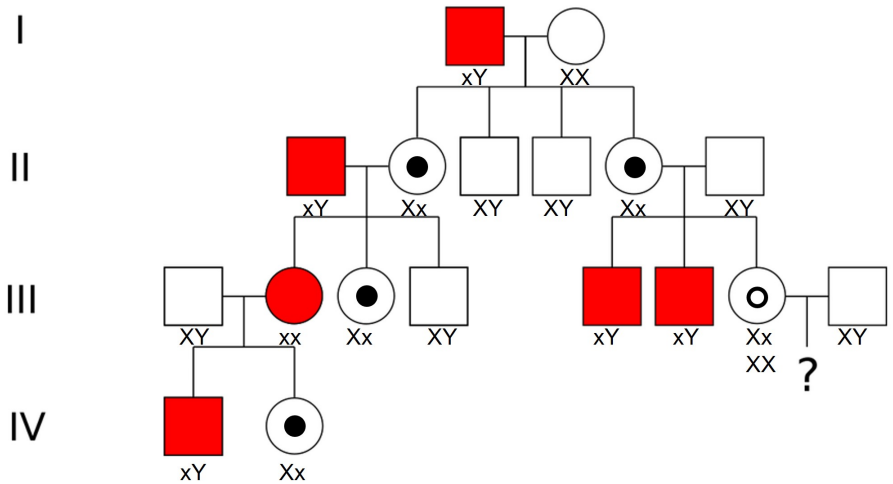

Abbildung: Stammbaum C - Lösung.

Zusammenfassung

- Warum ist LD wichtig in der Funktionelle Genomanalysen?
  - (Un-)Abhängigkeit von genetischen Markern in statistischen Analysen!
- Warum ist HWE wichtig in der Funktionelle Genomanalysen?
  - Annahme einer best. Verteilung in stat. Analysen
  - Kenntnis von best. Eigenschaften (z.B. Varianz)
- Warum sind Segregationsmuster wichtig in der Funktionelle Genomanalysen?
  - Relevant in Festlegung des Regressionsmodelles