Statistische Aspekte der Analyse molekularbiologischer und genetischer Daten (WS 2016/17)

Quelle: Vorlesungsunterlagen

Inhaltsverzeichnis

1 V1

1.1 Aufbau und Struktur der DNA

1.2 Genetischer Code

1.3 Replikation / Transkription / Translation

1.4 Mitose

1.5 Nicht-kodierende RNAs

1.6 Aufgaben zur Übung 1

1.6.1 Aufgabe 1

- zu a: siehe Codonsonne[1]
  AUG (ATG) als Startcodon, UGA (TGA) als Stopcodon
  5' - ATG GTT AAA CAC GTG CAC GAG TGA - 3'
  3' - TAC CAA TTT GTG CAC GTG CTC ACT - 5'
- zu b:
  5' - AUG GUU AAA CAC GUG CAC GAG UGA - 3'
- zu c: tRNA für Valin, Lysin, Histidin, Valin, Glutamin, Glutaminsäure (das komplementäre der RNA)
- zu d: unpolar/neutral, positiv/basisch, positiv/basisch, unpolar/neutral, polar/neutral, negativ/sauer

1.6.2 Aufgabe 2

1.6.3 Aufgabe 3

- E. coli: $4,6 * 10^6$ Basen, 4500 Gene
- Bäckerhefe: $2 * 10^7$ Basen, 6000 Gene
- Ackerschmalwand: 10^8 Basen, 25500 Gene
- Fruchtfliege (Drosophila Melanogaster): $2 * 10^8$ Basen, 13500 Gene
- Menschen: $3,27 * 10^9$ Basen, 23000 Gene

[1] `https://de.wikipedia.org/wiki/Code-Sonne`

1.6.4 Aufgabe 4

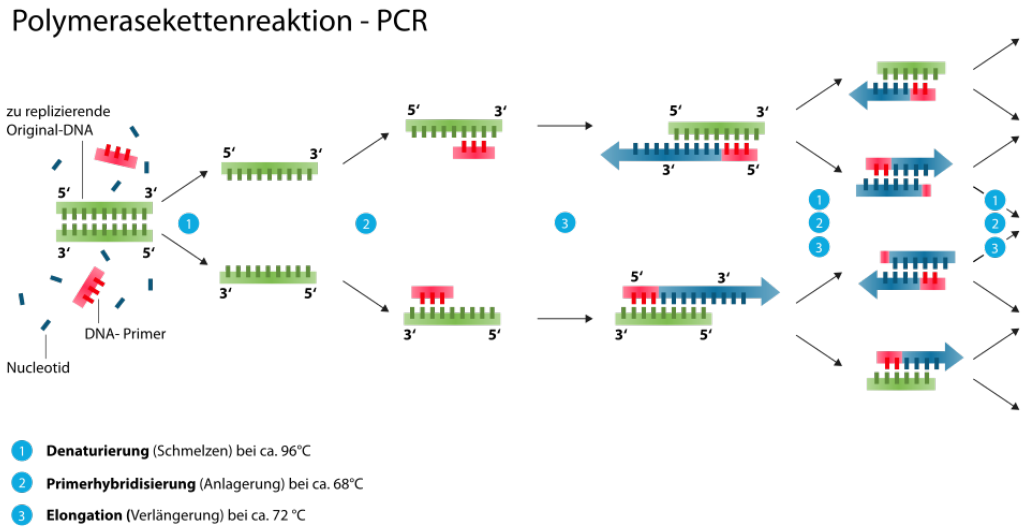
- SNP[2]:
  - Single Nucleotide Polymorphism - Einzelnukleotid-Polymorphismus
  - Variation eines einzelnen Basenpaares in einem DNA-Strang
  - SNPs sind geerbte und vererbbare genetische Varianten. Begrifflich davon abzugrenzen ist der Begriff der Mutation, der in der Regel eine neu aufgetretene Veränderung bezeichnet
  - Laktosetoleranz: durch einen SNP im Intron des Gens mcm6 entwickelt, welches 5' von LCT(Lactase) liegt
- CNV[3]:
  - Copy number variation - Kopienzahlvariation
  - struktureller Variation des Erbguts, die Abweichungen der Anzahl der Kopien eines bestimmten DNA-Abschnittes innerhalb eines Genoms erzeugt
- Chromosomen-Mutationen[4]:
  - strukturelle Veränderung eines Chromosoms, 5 Arten
  - Deletion: Ein Teilstück des Chromosoms (Endstück oder mittlerer Abschnitt) geht verloren
  - Translokation: Chromosomen können auseinanderbrechen und dabei Teilstücke verlieren, welche in die Chromatide eines anderen Chromosoms angeheftet werden
  - Duplikation: Ein Abschnitt des Chromosoms ist doppelt vorhanden, da ein auseinandergebrochenes Teilstück in die Schwesterchromatide eingegliedert wurde
  - Inversion: Innerhalb eines Chromosoms kann sich nach einem doppelten Bruch ein Stück wieder umgekehrt einfügen
  - Insertion (auch: Addition): Hier besitzt ein Chromosom ein zusätzliches Teilstück

1.6.5 Aufgabe 5

- PCR[5]: Polymerase-Kettenreaktion (polymerase chain reaction)
- Prozess besteht aus etwa 20–50 Zyklen, jeder Zyklus besteht aus drei Schritten

[2] https://de.wikipedia.org/wiki/Einzelnukleotid-Polymorphismus
[3] https://de.wikipedia.org/wiki/Gene_copy_number_variants
[4] https://de.wikipedia.org/wiki/Chromosomenmutation
[5] https://de.wikipedia.org/wiki/Polymerase-Kettenreaktion

1. Denaturierung (Melting, Schmelzen): Zunächst wird die doppelsträngige DNA auf 94–96 °C erhitzt, um die Stränge zu trennen. Die Wasserstoffbrückenbindungen, die die beiden DNA-Stränge zusammenhalten, werden aufgebrochen. Im ersten Zyklus wird die DNA oft für längere Zeit erhitzt (Initialisierung), um sicherzustellen, dass sich sowohl die Ausgangs-DNA als auch die Primer vollständig voneinander getrennt haben und nur noch Einzelstränge vorliegen. Manche (sogenannte Hot-Start-) Polymerasen müssen durch eine noch längere anfängliche Erhitzungsphase (bis zu 15 Minuten) aktiviert werden. Danach wird schnell auf 65 °C abgekühlt, um die Rückbildung der Doppelhelix zu verhindern.
2. Primerhybridisierung (primer annealing): Die Temperatur wird ca. 30 Sekunden lang auf einem Wert gehalten, der eine spezifische Anlagerung der Primer an die DNA erlaubt. Die genaue Temperatur wird hierbei durch die Länge und die Sequenz der Primer bestimmt (bzw. der passenden Nukleotide im Primer, wenn durch diesen Mutationen eingeführt werden sollen = site-directed mutagenesis). Wird die Temperatur zu niedrig gewählt, können sich die Primer unter Umständen auch an nicht hundertprozentig komplementären Sequenzen anlagern und so zu unspezifischen Produkten („Geisterbanden“) führen. Wird die Temperatur zu hoch gewählt, ist die thermische Bewegung der Primer u. U. so groß, dass sie sich nicht richtig anheften können, so dass es zu gar keiner oder nur ineffizienter Produktbildung kommt. Die Temperatur, welche die beiden oben genannten Effekte weitgehend ausschließt, liegt normalerweise 5–10 °C unter dem Schmelzpunkt der Primersequenzen; dies entspricht meist einer Temperatur von 55 bis 65 °C.
3. Elongation (Extending, Polymerisation, Verlängerung, Amplifikation): Schließlich füllt die DNA-Polymerase die fehlenden Stränge mit freien Nukleotiden auf. Sie beginnt am 3'-Ende des angelagerten Primers und folgt dann dem DNA-Strang. Der Primer wird nicht wieder abgelöst, er bildet den Anfang des neuen Einzelstrangs. Die Temperatur hängt vom Arbeitsoptimum der verwendeten DNA-Polymerase ab (68–72 °C). Dieser Schritt dauert etwa 30 Sekunden je 500 Basenpaare, variiert aber in Abhängigkeit von der verwendeten DNA-Polymerase. Übliche Thermocycler kühlen die Reaktionsansätze nach Vollendung aller Zyklen auf 4–8 °C, so dass eine PCR am Abend angesetzt werden kann und die Proben am Morgen darauf weiterverarbeitet werden können.



zu amplifizierende Sequenz:
5‘ACCGCGGCTT AGGAAAXXXX XXXXXXCCCG GGGCGTATGC TGACGG3‘
3’-CGAA TCCTTT-5’ 3’-GGGC CCCGCA-5’

1.6.6 Aufgabe 6

Didesoxymethode nach Sanger[6]:

- Didesoxynukleotide weil: wird als Stopp-Nukleotiden benutzt, an Ribose (Zucker) an Position 2’ und 3’ desoxidiert ist. Dadurch fehlt am 3’-Kohlenstoff-Atom die Hydroxygruppe, an der bei der Polymerisation das nächste Nukleotid angehängt wird.
- auch Desoxynukleotide weil: sonst funktioniert die Verlängerung nicht
- Ergebnis nur Didesoxynukleotide: es gibt keine Verländerung

nur Didesoxynukleotide

[6] https://de.wikipedia.org/wiki/DNA-Sequenzierung#Didesoxymethode_nach_Sanger

2 V2

2.1 Mechanismen der epigenetischen Modifikation

2.2 Mechanismen der DNA Reparatur

2.3 Typische Mutationen

2.4 PCR

2.5 Sanger Sequenzierung

2.6 TaqMan

2.7 SNP-Microarray

2.8 Aufgaben zur Übung 2

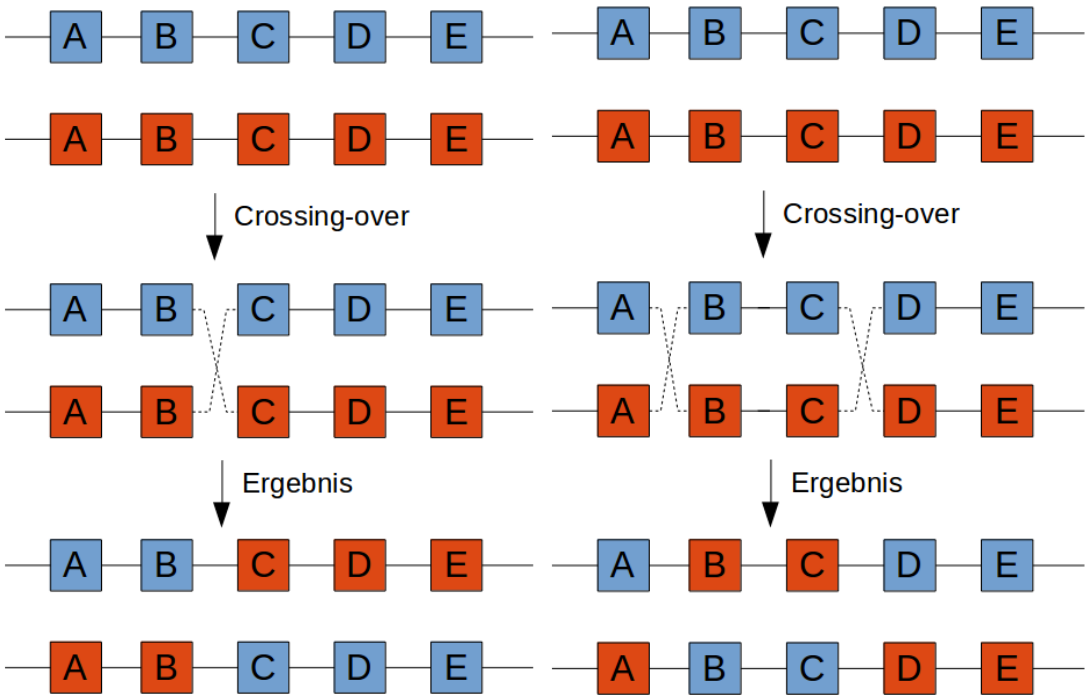
2.8.1 Aufgabe 1

a.)
Als Crossing-over[7] wird in der Genetik eine kreuzweise Überlagerung zweier Chromatiden mit nachfolgendem, gegenseitigem Austausch von Abschnitten bezeichnet, wie er zwischen väterlichen und mütterlichen homologen Chromosomen bei einer Meiose auftreten kann.

b.) A und B sind rekombiniert zu C,D,E
c.) A, D,E sind rekombiniert mit B,C

[7] `https://de.wikipedia.org/wiki/Crossing-over`

zu b.) **zu c.)**



2.8.2 Aufgabe 2

Gen: ABO[8] rs8176719[9]:

- (-;-): likely to be of blood type O
- (-;G): most likely to be of blood type A or B
- (G;G): most likely to be of blood type A, B or AB

rs8176747[10]:

- G führt zu Blutgruppe A, C zu Blutgruppe B

rs8176750[11]: definiert Untergruppe von A

- (-;C): A1
- (-;-): A2

Kombinationsmöglichkeiten:

- praktisch durch Allele vorgegeben: $3 \cdot 2 \cdot 2 = 12$[12]

[8] http://www.snpedia.com/index.php/ABO
[9] http://www.snpedia.com/index.php/rs8176747
[10] http://www.snpedia.com/index.php/rs8176747
[11] http://www.snpedia.com/index.php/rs8176750
[12] https://sites.google.com/site/abobloodgroup/14.aboalleles%28oalleles%29

- theoretisch: $5^3 = 125$
- Musterlösung: 3 SNPs auf einem Allel → 8 Kombinationen; 2 Allele: 36 Möglichkeiten

A und B kodominant, Faktor 0 rezessiv

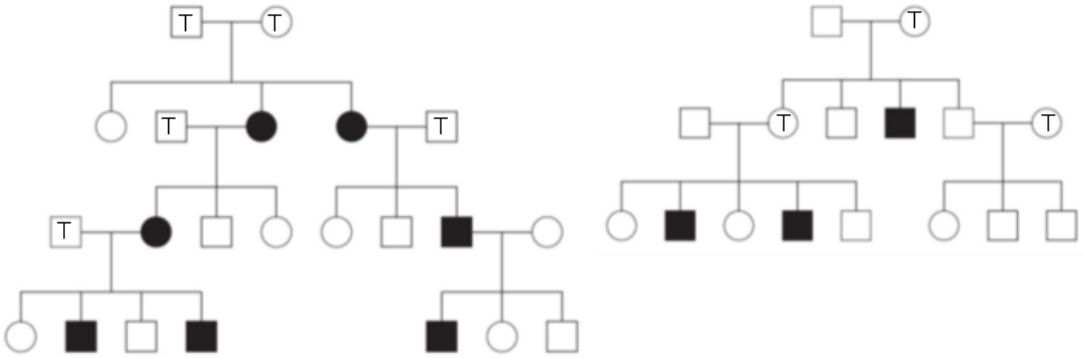
2.8.3 Aufgabe 3

a.)
b.)
c.)

2.8.4 Aufgabe 4

a.)
rezessiv:[13] bedeutet in der Genetik „zurücktretend“ oder auch „nicht in Erscheinung tretend“
dominant:[14] ein dominantes Allel setzt sich in der Merkmalsausprägung gegenüber einem rezessiven Allel durch
Penetranz:[15] prozentuale Wahrscheinlichkeit, mit der ein bestimmter Genotyp zur Ausbildung des zugehörigen Phänotyps führt

b.)



[13] https://de.wikipedia.org/wiki/Rezessiv
[14] https://de.wikipedia.org/wiki/Dominanz_(Genetik)
[15] https://de.wikipedia.org/wiki/Penetranz_(Genetik)

aus Musterlösung:

c.)
links: autosomal rezessiv, aus Musterlösung: autosomal dominat mit reduzierter Penetranz, weil:

- beide Geschlechter betroffen
- in jeder Generation
- etwa die Hälfte der Kinder betroffen

rechts: genosomal rezessiv, auf einem X-Chromosom der Mutter

3 V3

3.1 Meiose

3.2 Mendelsche Gesetze

3.3 Erbgänge / Stammbäume

3.4 Gründe für Abweichungen von Mendelschen Erbgängen

3.5 Aufgaben zur Übung 3

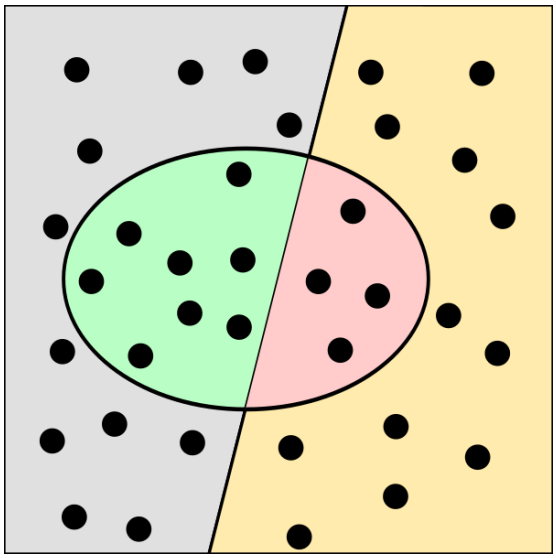
3.5.1 Aufgabe 1

a.)

- Sensitivität: gibt den Anteil der korrekt als positiv klassifizierten Objekte an der Gesamtheit der tatsächlich positiven Objekte an ($\mathbb{P}(P|K)$)
- Spezifität: gibt den Anteil der korrekt als negativ klassifizierten Objekte an der Gesamtheit der in Wirklichkeit negativen Objekte an ($\mathbb{P}(\overline{P}|\overline{K})$)
- Prävalenz: welcher Anteil der Menschen einer bestimmten Gruppe (Population) definierter Größe zu einem bestimmten Zeitpunkt an einer bestimmten Krankheit erkrankt ist
  Prävalenz=Anzahl der zum Untersuchungszeitpunkt Kranken / Anzahl der in die Untersuchung einbezogenen Individuen

Vierfeldertafel

	Person ist krank (r_p+f_n)	**Person ist gesund (f_p+r_n)**
Test positiv (r_p+f_p)	richtig positiv (r_p)	falsch positiv (f_p)
Test negativ (f_n+r_n)	falsch negativ (f_n)	richtig negativ (r_n)



b.)
gegeben:

- K={Patient ist krank}
- P={Test ist positiv}
- Sensitivität: $\mathbb{P}(P|K) = 0,95$
- Spezifität: $\mathbb{P}(\overline{P}|\overline{K}) = 0,90$
- Prävalenz: $\mathbb{P}(K) = 0,1$

gesucht:

- positiv prädiktiver Wert (PPW):
  $\mathbb{P}(K|P) = \underbrace{\frac{\mathbb{P}(P|K) \cdot \mathbb{P}(K)}{\mathbb{P}(P)}}_{Satz\ von\ Bayes} = \underbrace{\frac{\mathbb{P}(P|K) \cdot \mathbb{P}(K)}{\underbrace{\mathbb{P}(P|\overline{K})}_{=1-\mathbb{P}(\overline{P}|\overline{K})} \cdot \mathbb{P}(\overline{K}) + \mathbb{P}(P|K) \cdot \mathbb{P}(K)}}_{totale\ Wahrscheinlichkeit}$
  $\mathbb{P}(K|P) = \frac{0,95 \cdot 0,1}{0,1 \cdot 0,9 + 0,95 \cdot 0,1} = \underline{\underline{0,513513514}}$
- negativ prädiktiver Wert (NPW):
  $\mathbb{P}(\overline{K}|\overline{P}) = \frac{\mathbb{P}(\overline{P}|\overline{K}) \cdot \mathbb{P}(\overline{K})}{\underbrace{\mathbb{P}(\overline{P})}_{1-\mathbb{P}(P)}} = \frac{\mathbb{P}(\overline{P}|\overline{K}) \cdot \mathbb{P}(\overline{K})}{1-(\mathbb{P}(P|\overline{K}) \cdot \mathbb{P}(\overline{K}) + \mathbb{P}(P|K) \cdot \mathbb{P}(K))}$
  $\mathbb{P}(\overline{K}|\overline{P}) = \frac{0,9 \cdot 0,9}{1-(0,1 \cdot 0,9 + 0,95 \cdot 0,1)} = \underline{\underline{0,993865031}}$

c.)
gegeben:

- Sensitivität: $\mathbb{P}(P|K) = 0,95$
- Spezifität: $\mathbb{P}(\overline{P}|\overline{K}) = 0,90$
- Prävalenz: $\mathbb{P}(K) = 0,05$

gesucht:

- positiv prädiktiver Wert (PPW)= $\underline{\underline{0,\overline{33}}}$
- negativ prädiktiver Wert (NPW)= $\underline{\underline{0,997084548104956}}$

d.) siehe R-Script

3.5.2 Aufgabe 2

3.5.3 Aufgabe 3

siehe R-Script

4 V4

4.1 Bias und Präzision

4.2 Frequentistischer und Bayesianischer Wahrscheinlichkeitsbegriff

4.3 Zufallsvariablen (Erwartungswert, Varianz, Standardabweichung, Covarianz, Unabhängigkeit, Randverteilung)

4.4 Bedingte Wahrscheinlichkeit, Bayessche Lernformel

4.5 Einige wichtige Verteilungsfunktionen

4.6 Aufgaben zur Übung 4

5 V5

5.1 Konfidenzintervall

$P(E(x)) \in [\bar{x_n} - 1.96\frac{\hat{\sigma}}{\sqrt{N}}, \bar{x_n} + 1.96\frac{\hat{\sigma}}{\sqrt{N}}] = 0.95$
Interpretation des 95% Konfidenzintervalls

- Interpretation (frequentistisch): In 95% aller Stichproben vom Umfang N liegt der wahre Erwartungswert in dem oben angegebenen Intervall.
- Alternativ (frequentistisch): 95% der so konstruierten Konfidenzintervalle enthalten den wahren Wert
- Interpretation (Bayesianisch): Mit 95% Sicherheit (nicht Wahrscheinlichkeit!) liegt der wahre Erwartungswert in diesem Intervall.
- Für eine spezielle gegebene Stichprobe können wir jedoch nicht sagen, ob der wahre Wert in diesem Intervall liegt oder nicht. In der Regel gibt es auch keine Möglichkeit, dies anderweitig zu zeigen.

5.2 Logik des statistischen Testens, Testdurchführung und Interpretation

Beobachtung: Unterschied zwischen zwei Gruppen (z.B. zwei Therapievarianten, zwei genetischen Varianten A und B (bei Variante A mehr Ereignisse))

Skeptiker: Dieses Ergebnis ist zufällig!

Annahme: Der Skeptiker hat recht. Es gibt keinen Unterschied (Nullhypothese).

Frage: Wie gut können Zufallseffekte die Daten erklären? Wie wahrscheinlich ist meine Beobachtung, wenn die Nullhypothese war ist (p-Wert)?

Maßzahl: Wahrscheinlichkeit p (p-Wert) den beobachteten Unterschied (oder noch extremere Abweichungen von der Nullhypothese) rein zufällig zu erhalten.
Statistische Schlussweise: Wenn p klein ist...

- Etwas sehr Unwahrscheinliches ist geschehen
- Unsere Annahme (Nullhypothese) ist falsch

Je kleiner p, desto unplausibler ist die Nullhypothese
Konvention: $p < \alpha = 0.05$ („signifikant“) $\rightarrow$ Nullhypothese ablehnen

Allgemeine Konstruktion eines statistischen Tests:

1. Auswahl eines Effektmaßes (z.B. (standardisierte) Mittel- wertsdifferenz)
2. Bestimme zugehörige Verteilung unter der Nullhypothese
3. Lege extreme Abweichungen von der Nullhypothese fest (z.B. zweiseitig oder einseitig)
4. Berechne die beobachtete Teststatistik z
5. Berechne den p-Wert aus z und der Verteilung unter der Nullhypothese

5.3 Typ I und Typ II Fehler, Einfluß der Fallzahl

Konfidenzniveau

- Bemerkung 1: α wird Typ 1 Fehler genannt.
- Bemerkung 2: Statistische Tests korrespondieren fast immer zu Berechnungen von 95% Konfidenzintervallen. Die Nullhypothese kann verworfen werden, wenn deren Effekt außerhalb des Konfidenzintervalls des beobachteten Effekts liegt.
- Bemerkung 3: Wenn p groß ist, kann man gar nichts schlussfolgern. Die Annahme, dass dann die Nullhypothese gilt bzw. gar bewiesen wurde, ist ein (leider sehr weit verbreiteter) Irrtum.
- Bemerkung 4: Statistische Signifikanz hat nichts mit Relevanz zu tun. Selbst die kleinsten Effekte können bei großen Fallzahlen signifikant werden.

Power

- Neben dem Typ-1 Fehler α gibt es auch einen Typ-2 Fehler β. Er beschreibt die Wahrscheinlichkeit, eine falsche Nullhypothese nicht aufdecken zu können.
- 1-β heißt Power und sollte möglichst groß sein (Studienplanung, für klinische Studien $\sim$80%)
- Die Power hängt ab von der Wahl des Signifikanzniveaus α, von der Stärke der Abweichung von der Nullhypothese (Effektstärke) und der Fallzahl.
- Kleine Signifikanzniveaus und Effektstärken erfordern eine große Fallzahl um eine akzeptable Power zu erreichen (vergleiche später)

5.4 Problem des multiplen Testens und Korrekturmöglichkeiten

- Bei einem Test beträgt die Wahrscheinlichkeit, dass man die Nullhypothese fälschlicherweise ablehnt α
- typischer Wert für α ist 0,05
- beim mehrmaligen Testen von Hypothesen innerhalb einer Stichprobe kommt es zu einer sogenannten α-Fehler-Kumulierung
- Testet man zum Beispiel 1 Mio. genetische Varianten erhält man schon durch Zufall 50.000 signifikante Ergebnisse

Korrekturmöglichkeiten

- Kontrolle der family-wise error rate (FWER): Bonferroni, Bonferroni-Holm
- False discovery rate (FDR): Benjamini-Hochberg, Schätzen des Anteils wahrer Nullhypothesen

5.5 Faktoren für die Auswahl des richtigen Tests

Statistische Tests müssen passend zum Variablentyp gewählt werden.
Gepaarte versus ungepaarte Vergleiche

- Ungepaart: Vergleich zwischen verschiedenen Gruppen
- Beispiele:
  - Geschlechter
  - Fälle versus Kontrollen
  - verschieden behandelte Individuen
- Gepaart: Vergleiche innerhalb eines Individuums oder Entität
- Beispiele:
  - rechtes gegen linkes Auge/Hand
  - wiederholte Messungen innerhalb von Probanden
  - gematchte Fälle und Kontrollen

Überblick (Vergleich stetiger Variablen)

	Zwei Gruppen		Mehr als zwei Gruppen	
Daten	Unabhängige Gruppen	Gepaarte Gruppen	Unabhängige Gruppen	Gepaarte Gruppen
Normal-verteilt	t-Test	Gepaarter t-Test	ANOVA	ANOVA mit Messwieder-holung
Nicht Normal-verteilt	U test	Wilcoxon-Test	Kruskal-Wallis Test	Friedman Test

Überblick (Vergleich binärer Variablen)

	ungepaart	gepaart
Fallzahl groß	Chiquadrat Test	McNemar Test
Fallzahl klein	Fishers exakter Test	Exakter Binomialtest

5.6 Zusammenhangsmaße auf Vierfeldertafeln

	Risikofaktor liegt vor	Risikofaktor liegt nicht vor	
Erkrankt	A	b	a+b
Gesunf	C	d	c+d
	a+c	b+d	N

Maße:

- Absolute Risikodifferenz: $ARD = \frac{a}{a+c} - \frac{b}{b+d}$
- Relatives Risiko: $RR = \frac{\frac{a}{a+c}}{\frac{b}{b+d}}$
- Odds-ratio: $OR = \frac{ad}{bc}$

Warum sollte man die Odds-ratio verwenden?

- ARD und RR hängen vom Basisrisiko ab → Vergleich zwischen Gruppen kann irreführend sein (wie in unserem Emesis-Beispiel)
- OR ist unabhängig vom Basisrisiko → kann zwischen Gruppen verglichen werden
- Die Nullhypothese des Chiquadrat-Tests ist OR=1 → OR ist das zum Chiquadrat-Test passende Effektmaß

5.7 Korrelation, Scheinkorrelation und Confounder

Pearsons Korrelationskoeffizient $r(X,Y) = \frac{var(X+Y) - var(X) - var(Y)}{2\sqrt{var(X)}\sqrt{var(Y)}}$

Eigenschaften:

1. Misst den Grad der linearen Abhängigkeit zwischen X und Y
2. $-1 \leq r(X,Y) \leq 1$
3. r(X,Y) ist invariant unter linearer Transformation von X und Y
4. Falls Y=aX+b dann r(X,Y)=sign(a)
5. Misst nicht die Übereinstimmung (Konkordanz) von X und Y

Grobe Einteilung der Stärke des linearen Zusammenhangs:

- $|r| <$ 0,3 schwach
- 0,3 $\leq |r| <$ 0,6 mittel
- 0,6 $\leq |r| <$ 0,8 stark
- 0,8 $\leq |r| <$ 1,0 sehr stark
- $|r| =$ 1,0 perfekter Zusammenhang

Kombination nicht vergleichbarer Stichproben führt zu **Scheinkorrelation**. Größen die Scheinkorrelationen erzeugen heißen **Confounder**.

5.8 Aufgaben zur Übung 5

6 V6

6.1 Lineare Regression

6.1.1 Modellannahme

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ mit $i \in \{1, ..., n\}$

Schätzen mit Kleinste Quadrate Methode: Regressionsgerade minimiert die Summe der quadrierten Abstände zwischen den y-Werten der Regressionsgerade und den y-Werten der Datenpunkte (Residuen)

6.1.2 Schätzen der Betas („Intercept“ und „Slope“)

Ableiten der Residuenquadratsumme nach den Betas und Nullsetzen.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}$$

$$\hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \cdot x_n$$

Mean square error (MSE) $= \frac{1}{n} \cdot RSS$

residual sum of squares (RSS) $= \sum_{i=1}^{n}(y_i - f(x_i))^2$

6.1.3 Varianzzerlegung und erklärte Varianz bei linearer Regression

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{Gesamtvarianz}} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{Varianz der Residuen}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{Varianz der Regresswerte}}$$

Erklärte Varianz (= Bestimmtheitsmaß R^2)

$$R^2 = 1 - \frac{\text{Varianz der Residuen}}{\text{Gesamzvarianz}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$R^2 = 1 - \frac{\text{Varianz der Regresswerte}}{\text{Gesamzvarianz}} = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Problem: Steigt die Anzahl der Parameter m im Modell, so steigt R2 unabhängig vom Einfluss der Variablen, deshalb Korrektur

korrigiertes Bestimmtheitsmaß:

$$\bar{R}^2 = 1 - \frac{\frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

mit p=Anzahl der unabhängigen Variablen

6.1.4 Multivariate Regression

Wie lineare, aber mehrere betas

6.1.5 AIC

AIC (Akaike Informations Kriterium): je kleiner desto besser
$AIC = -2L + 2m$

Bayesianisches Informationskriterium (BIC): je kleiner desto besser
$BIC = -2L + mlog(n)$

Log-Likelihood für Beide (bei linearer Regression):
$-L = \frac{n}{2}ln(2\pi\sigma^2) + \sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{2\sigma^2}$

6.2 Multivariate Regression

6.2.1 Schätzen von Kontrasten

Was soll das sein???

6.2.2 AIC

6.2.3 Interaktion

- möglich innerhalb multivariate Regression
- Berücksichtigung nicht-additiver Effekte (Modulation einer Einflußgröße durch eine andere)

z. B. multivariate lineare Regression:
$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \beta_3 (x_i^{(1)} \cdot x_i^{(2)}) + \epsilon_i$

$x_i^{(1)} \cdot x_i^{(2)}$ - Interaktionsterm:
$Y_i = \beta_0 + \underbrace{(\beta_1 + \beta_3 x_i^{(2)})}_{=:\tilde{\beta_1}=\tilde{\beta_2}(x_i^{(2)})} x_i^{(1)} + \beta_2 x_i^{(2)} + \epsilon_i$

6.3 Auswahl einer passenden Regressionsmethode

Best-Subset-Selection

- betrachte alle $\binom{m}{k}$ Modelle
  - m: Anzahl der Einflussvariablen
  - k $\leq$ m: Anzahl der ausgewählten Einflussvariablen
- bestimme für jedes k das beste Modell anhand RSS, MSE oder R^2
- wähle aus den k Modellen das beste Modell anhand AIC, BIC, Adjusted R^2
- Problematisch für große m
- Die Regressionskoeffizienten sind bei diesem Vorgehen gebiast (vergleiche später Sequentialtest)

Vorwärts-Selektion

- Beginn mit Null-Modell (keine Einflußvariablen)
- füge diejenige Einflußsvariable hinzu, bei der RSS, MSE oder R 2 am besten ist
- wähle aus den k Modellen das beste Modell anhand AIC, BIC, Adjusted R^2 oder MSE

Rückwärts-Eliminierung

- Beginn mit vollständigem Modell (alle Einflußvariablen)
- entferne diejenige Ausgangsvariable, bei der RSS, MSE oder R^2 am besten ist

- wähle aus den k Modellen das beste Modell anhand AIC, BIC, Adjusted R^2 oder MSE

Idealerweise kommen hierbei die selben Modelle heraus (ist in der Praxis jedoch selten der Fall), auch hier sind die Schätzer der β_j des besten Modells gebiast

6.4 Aufgaben zur Übung 6

7 V7

7.1 Motivation und Ansatz für gemischte Modelle

Das lineare Modell stößt bei machen Analysen an seine Grenzen, z.B wegen:

- Fehlen der Linearität in den realen Daten → nichtlineare Regression
- Verletzung der Verteilungsannahmen an die zufälligen Fehler
- tichprobe besteht aus Subgruppen, für die jeweils „eigene“ β's geschätzt werden müssten. D.h. die Grundannahme der Unabhängigkeit der Beobachtungen ist verletzt.

Typische Beispielsituationen

- Longitudinale Daten (Im Rahmen von Längsschnittstudien werden in zeitlicher Abfolge wiederholte Beobachtungen an denselben Subjekten -unter möglicherweise verschiedenen Bedingungen- erhoben)
- Clusterdaten (Aus Primäreinheiten (Cluster) werden mehrere Individuen ausgewählt und die interessierenden Variablen erhoben: Kinder aus Familien, Schüler auf Schulen)
- Hierarchische Daten (z.B. Kontinent → Land → Stadt)
- In der Genetik: Verwandtschaft

Warum hier nicht lineare Regression? Hauptproblem ist hier die Korrelation in den Daten:

- Fehlerterme in linearer Regression sind nicht mehr unabhängig
- beta-Schätzer in linearer Regression können die Struktur der Daten nicht angemessen berücksichtigen

Longitudinaldaten: Im Rahmen von Längsschnittstudien werden in zeitlicher Abfolge wiederholte Beobachtungen an denselben Subjekten (unter möglicherweise verschiedenen Bedingungen) erhoben.

Clusterdaten: Aus Primäreinheiten (Cluster) werden mehrere Individuen ausgewählt und die interessierenden Variablen erhoben.

Bei Longitudinal- und Clusterdaten lässt sich ein typische Datensatz (Stichprobe) wie folgt notieren: $(y_{i1}, ..., y_{in_i}, x_{i1}, ..., x_{in_i})$, i=1,...,m

- m die Anzahl der Individuen bzw. Cluster
- $n_i$die Anzahl der Messzeitpunkte für Individuum i bzw. die Anzahl der Elemente im Cluster i

- Longitud.Daten: y_{ij} ist der Wert der Zielgröße von Subjekt i zum Zeitpunkt t_{ij}
- Cluster: y_{ij} ist der Wert der Zielgröße vom j-ten Element in Cluster i

Beachte: Die Daten vom gleichen Subjekt/Cluster neigen oft dazu ähnlicher zu sein als die Daten verschiedener Subjekte/Cluster.

Ziele:

- Schätzung der Subjekt- /Clusterspezifischen Effekte
- Schätzung der Populationsspezifischen Effekte
- Schätzung der Korrelationsstruktur

7.2 Feste und zufällige Effekte

Variabilitätsquellen

- Zwischen den Subjekten/Clustern, d.h. Abweichungen eines Subjekts/Clusters vom Populationsmittel.
- Innerhalb des Subjekts/Clusters, d.h. Abweichungen einer Messung vom Mittelwert des entsprechenden Subjekts/Clusters.

Lösung: Zwei - Stufen - Verfahren oder lineares gemischtes Modell

Zwei-Stufen-Verfahren:

1. Stufe: Das subjektspezifische Modell $Y_i = Z_i\beta_i + \epsilon_i$

   - i aus {1,...,m} das Individium i
   - n_i Anzahl der Messwiederholungen bei Individuum i
   - $Y_i = (Y_{i1}, ..., Y_{in_i})^T$ der n_i-dimensionale Vektor der Messungen von Individuum i
   - Z_i ist die $(n_i \times q)$-Matrix der (evtl. zeitlich abhängigen q Kovariablen
   - β_i (später random effects)
   - β_i ist der (unbekannte) q-dimensionale subjektspezifische Vektor der Effekte der Kovariablen
   - $\epsilon_i = (\epsilon_{i1}, ..., \epsilon_{in_i})^T$ Fehlerterm mit $\epsilon_i \sim N(0, \sigma^2 I_{n_i})$

2. Stufe: Das Modell $\beta_i = K_i\beta + b_i$ welches die Variabilität zwischen den Individuen beschreibt

   - i aus {1,...,m} das Individuum i
   - n_i Anzahl der Messwerte bei Individuum i
   - K_i eine $(q \times p)$-Matrix bekannter p Kovariaten auf Populationsebene (später fixed effects)
   - β ist der p-dimensionale Vektor der unbekannten Regressionskoeffizienten dieser Kovariaten
   - $b_i \sim N(0, D)$ Fehlerterm mit Kovarianzmatrix D (also hier im Unterschied zur linearen Regression Fehler nicht mehr i.i.d.[independent and identically distributed])

Man will Modelle aus beiden Stufen gleichzeitig fitten → **Idee:** einsetzen der 2. Stufe in die 1. Stufe.

Stufe 1: $Y_i = Z_i\beta_i + \epsilon_i$
Stufe 2: $\beta_i = K_i\beta + b_i$
Zusammen: $Y_i = Z_iK_i\beta + Z_ib_i + \epsilon_i$

$\Rightarrow$ Umwandlung des hierachischen Modells in ein
lineares gemischtes Modell (linear random effects model):

$$y_i = \underbrace{X_i\beta}_{\text{Feste Effekte}} + \underbrace{U_i b_i}_{\text{Zufällige Effekte}} + \underbrace{\epsilon_i}_{\text{Residuale Fehler}} \quad , \text{i=1,...,m}$$

Hierbei sind:

- Die zufälligen Effekte b_i unabhängig identisch normalverteilt (i.i.d.) mit Erwartung 0: $b_i \overset{i.i.d}{\sim} N(0, D)$
- D ist eine positiv semi-definite Matrix (Kovarianzmatrix) und beschreibt ggf. Abhängigkeiten zwischen den zufälligen Effekten
- ϵ_i unabhängig (aber nicht identisch!) normalverteilt: $\epsilon_i \sim N(0, \Sigma_i)$
- $\Sigma_1, ..., \Sigma_m$ individuelle Kovarianzmatrizen der Zeitreihen
- $b_1, ..., b_m, \epsilon_1, ..., \epsilon_m$ unabhängig

Bemerkungen:

- Ein lineares gemischtes Modell (LGM) muss nicht aus einem 2-Stufen-Verfahren abgeleitet werden: Jedes 2 Stufen Verfahren führt zu einem LGM, aber nicht jedes LGM lässt sich aus einem 2-Stufen-Verfahren herleiten.
- Bei Longitudinaldaten können Kovariablen zeitlich variieren;
- Auch bei Clusterdaten können Variablen im gleichen Cluster variieren oder konstant sein
- Bis auf die Definitheit gibt es keine weiteren Annahmen an die Kovarianzmatrizen der zufälligen Effekte D und der Fehlersumme. Zusätzliche Annahmen hinsichtlich der Struktur der Matrizen ergeben Modelle verschiedener Komplexität und Flexibilität

Spezialfälle:

- **Conditional Independence:** ???
- **Varianzkomponenten-Modell:** ???
- **Random Intercept Modell:** Hier wird für jedes Individuum ein eigenes Intercept zugewiesen
- **Random Intercept – Random Slope Modell:** ???

Abschließende Bemerkungen zum LGM: Warum also LGM's?

- Stichproben mit einer Substruktur verletzen typischerweise die für viele statistische Verfahren erforderliche Unabhängigkeit der Beobachtungen. Nichtbeachten dieser Substruktur führt zu fehlerhafter Ergebnissen (Tests, Konfidenzintervalle, ...)
- Abhängigkeiten zwischen Beobachtungen reduzieren den Informationsgehalt im Vergleich mit unabhängigen Beobachtungen.
- Beachtung individuenspezifischer Informationen kann Schätzgenauigkeit erhöhen
- Schätzer für zufällige Effekte ermöglichen individuelle Prognosen
- Ermöglichen Modelle mit fehlenden Werten
- Bei vielen Meßverfahren können mit Hilfe der LGMs Batch-Effekte berücksichtigt werden.
- In der Genetik sind diese Verfahren notwendig um Verwandtschaftsstrukturen zu berücksichtigen.
- Wichtig für die Kombination von Studien (Metaanalyse)

7.3 Idee der Hauptkomponentenanalyse

Ziele:

- Exploratives Untersuchen von (hochdimensionalen) Daten
- Visualisierung (hochdimensionaler) Daten
- Dimensionsreduktion
- Denoising (Rauschen der Daten mindern)

Zwischenziel: Finde im m-dimensionalen Datenraum die Richtung größter Varianz

Pearsons Ansatz: Welche Gerade passt am besten zu den Daten?

- Die Gerade w, die das Rauschen minimiert und das Signal maximiert.
- Oder anders gesagt: Die Gerade, die die Varianz in den Daten maximiert

Was hier noch???

7.4 Interpretation PCA-Plots und Eigenwerte

Und was hier so???

7.5 Aufgaben zur Übung 7

8 V8

Ideale Population

- Modellannahmen für genetische Statistik:
- unendlich große (sehr große Individuenzahl)
- keine Selektion
- keine Mutation
- keine Migration
- zufällige Partnerwahl (random mating)
- getrennte Generationen (keine Verpaarungen zwischen z.B. Elterngeneration und Kindergeneration)

8.1 Hardy-Weinberg Gleichgewicht incl. Test

Seien an einem genetischen Lokus die Allele A und B vorhanden. Betrachte Genotyphäufigkeiten:
$P(AA) = p_1$, $P(AB) = p_2$, $P(BB) = 1 - p_1 - p_2$

Die Allelhäufigkeit läßt sich daraus berechnen:
$P(A) = p_1 + \frac{p_2}{2}$, $P(B) = 1 - P(A)$

Unter Annahme zufälliger Partnerwahl erhält man: $P(AA) = P(A)^2$, $P(AB) = 2P(A)P(B)$, $P(BB) = P(B)^2$

Also eine charakteristische Verteilung der Genotypen in Abhängigkeit von der Allelfrequenz das so genannte Hardy-Weinberg-Gleichgewicht (HWE).

Test auf HWE
Chi^2-Test des Hardy-Weinberg-Gleichgewichts (2 Allele, n=n_{11} +n_{12} +n_{22} Beobachtungen):

$$\chi_1 \sim \sum \frac{(O-E)^2}{E} = \frac{(n_{11} - n \cdot \hat{P}(A)^2)^2}{n \cdot \hat{P}(A)^2} + \frac{(n_{12} - n \cdot 2\hat{P}(A)\hat{P}(B))^2}{n \cdot 2\hat{P}(A)\hat{P}(B)} + \frac{(n_{22} - n \cdot \hat{P}(B)^2)^2}{n \cdot \hat{P}(B)^2}$$

Erweiterungen:

- Für kleine Allelfrequenzen oder kleine Fallzahlen gibt es exakte Tests in Analogie zu Fishers exaktem Test.
- Für gemischte Populationen gibt es stratifizierte Tests.

Hypothesentest für das Hardy-Weinberg-Gleichgewicht (HWE) mit H_0: Die beobachteten Häufigkeiten der Genotypen sind im HWE.

Verwerfe H_0 , wenn bei einem Test zum Signifikanzniveau α die Prüfgröße $\sum \frac{(O-E)^2}{E}$ größer ist als der kritische Wert $\chi^2_{df;1-\alpha}$.

Beachte: Wenn H_0 nicht abgelehnt wird, ist das kein Beweis für HWE. Wenn man HWE beweisen will, benötigt man Äquivalenztests für diese Situation.

Beispielrechnung:

Phenotype	White-spotted (AA)	Intermediate (Aa)	Little spotting (aa)	**Total**
Number	1469	138	5	1612

From this, allele frequencies can be calculated:

$$p = \frac{2 \times \text{obs(AA)} + \text{obs(Aa)}}{2 \times (\text{obs(AA)} + \text{obs(Aa)} + \text{obs(aa)})}$$

$$= \frac{1469 \times 2 + 138}{2 \times (1469 + 138 + 5)}$$

$$= \frac{3076}{3224}$$

$$= 0.954$$

and

$$q = 1 - p$$
$$= 1 - 0.954$$
$$= 0.046$$

So the Hardy–Weinberg expectation is:

$$\text{Exp(AA)} = p^2 n = 0.954^2 \times 1612 = 1467.4$$
$$\text{Exp(Aa)} = 2pqn = 2 \times 0.954 \times 0.046 \times 1612 = 141.2$$
$$\text{Exp(aa)} = q^2 n = 0.046^2 \times 1612 = 3.4$$

Pearson's chi-squared test states:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$
$$= \frac{(1469 - 1467.4)^2}{1467.4} + \frac{(138 - 141.2)^2}{141.2} + \frac{(5 - 3.4)^2}{3.4}$$
$$= 0.001 + 0.073 + 0.756$$
$$= 0.83$$

8.2 Kinship-Koeffizient, Verwandtschaftsschätzung

Die Unabhängigkeit von Elementen einer Stichprobe ist eine grundlegende Annahme in der Statistik. Durch Verwandtschaft ist diese Annahme verletzt.

IBD: Ein Genlokus in den Individuen X und Y heißt „identical by descent“, wenn er von einem gemeinsamen Vorfahren ererbt wurde.

IBS: Ein Genlokus in den Individuen X und Y heißt „identical by state“, wenn er sich nicht zwischen den Individuen X und Y unterscheidet.

IBS ist leicht festzustellen. Liegen keine Familienstammbäume vor, muß IBD geschätzt werden.

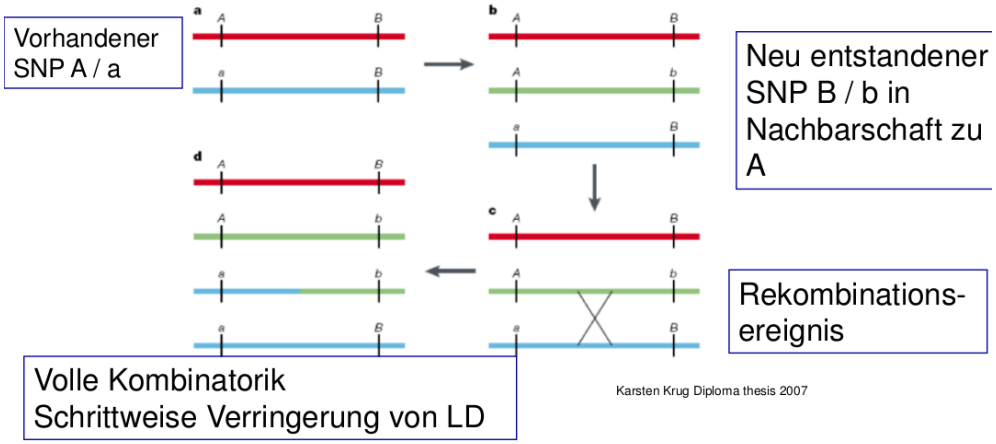
Kinship-Koeffizient k_{ij} zweier Personen i und j
Ziehe an einem Genlokus von jeder Person ein Allel k_{ij}=P(die Allele sind IBD) $\in$ [0,$\frac{1}{2}$] Schätzung nach Astle & Balding: $\hat{k_{ij}} = \frac{1}{L}\sum_{l=1}^{L} \frac{(g_{l,i} - 2p_{l,1})(g_{l,j} - 2p_{l,1})}{4p_{l,1}(1-p_{l,1})}$ mit
L #Loci
g kodiert Anzahl eines bestimmten Allels (0,1,2)
p Allelfrequenz dieses Allels

Was hier noch???

8.3 Kopplungsungleichgewicht

- Auf ein- und demselben Chromosom benachbarte Marker werden gemeinsam vererbt, wenn zwischen ihnen keine Rekombination stattfindet.
- Der Verteilung benachbarter Marker ist deshalb i.d.R. nicht stochastisch unabhängig. Es besteht eine Assoziation zwischen benachbarten Markern.
- Für die Wahrscheinlichkeitsverteilung der Genotypen der Marker X und Y gilt dann: $F_{X,Y}(x, y) \neq F_X(x)F_Y(y)$

8.3.1 Entstehung und Entwicklung



Gemessen / beobachtbar sind i.d.R. nur Genotypen. LD beschreibt aber den Zusammenhang auf Chromosomenebene (Haplotypen).

Beobachtbare **Genotypfrequenzen**:

		Lokus 2		
		BB	Bb	bb
Lokus 1	AA	P_{AABB}	P_{AABb}	P_{AAbb}
	Aa	P_{AaBB}	P_{AaBb}	P_{Aabb}
	aa	P_{aaBB}	P_{aaBb}	P_{aabb}

Dilemma: Individuen mit dieser Genotypkonstellation lassen sich nicht eindeutig einer Haplotypkonstellation zuordnen

Zugrundeliegende (versteckte) **Haplotypfrequenzen**:

		Lokus 2	
		B	b
Lokus 1	A	P_{AB}	P_{Ab}
	a	P_{aB}	P_{ab}

8.3.2 Bewertung (Maße)

Alle LD-Maße können als Zusammenhangsmaße auf Vierfeldertafeln beschrieben werden. Formale Tests auf Unabhängigkeit sind i.d.R. uninteressant. Interessant ist vielmehr die Stärke des Zusammenhangs.

p_{00}	p_{01}	$p_{0.}$
p_{10}	p_{11}	$p_{1.}$
$p_{.0}$	$p_{.1}$	1

Abweichung von Unabhängigkeit
$D = p_{00} - p_{0.}p_{.0}$
D standardisiert auf [-1,1]
$$D' = \frac{D}{D_{max}} \text{ where } D_{max} \begin{cases} min\{p_{0.}p_{.1}, p_{.0}p_{1.}\} & \text{if } D \geq 0 \\ min\{p_{0.}p_{.0}, p_{1.}p_{.1}\} & \text{if } D < 0 \end{cases}$$

Dieses Maß hängt von den Allelfrequenzen ab und ist ± 1 für Tafeln mit einer Null.

Korrelationskoeffizient r: $r = \frac{D}{sqrtp_{0.}p_{.0}p_{1.}p_{.1}}$

Mutual information: $M = \sum_{i,j} p_{ij} log_2 p_{ij} - \sum_{i,j} p_{i.} log_2 p_{i.} - \sum_{i,j} p_{.j} log_2 p_{.j}$

Odds ratio: $\lambda = \frac{p_{00}p_{11}}{p_{01}p_{10}}$

Odds ratio standardisiert auf [-1,1]:
Yule's Q: $Q = \frac{\lambda - 1}{\lambda + 1}$
Yule's Y: $Y = \frac{\sqrt{\lambda} - 1}{\sqrt{\lambda} + 1}$

Diese Maße sind unabhängig von der Allelfrequenzen und sind extrem für Tafeln mit einer Null. Eine interessante Eigenschaft ist, dass D', r und Y auf Diagonaltafeln übereinstimmen.

Faustregel zur Verwendung unterschiedlicher LD Maße

- D' ist ein Maß für stattgehabte Rekombinationen zwischen zwei Markern
- r, M messen den Grad der Übereinstimmung von markerbasierten Teststatistiken z.B. in Assoziationsstudien
- Odds-ratio basierte Maße sind Maße für stattgehabte Rekombinationen zwischen zwei Markern und eignen sich zum Vergleichen von Populationen (da unabhängig von Randverteilung = Allelfrequenz)
- D' wird dennoch sehr oft verwendet trotz der schlechten Eigenschaften

8.3.3 Bedeutung (Interpretation, Tagging, LD-Heatmaps)

Das mittlere LD unterscheidet sich zwischen verschiedenen genomischen Bereichen → unterschiedliche Rekombinationswahrscheinlichkeiten. Die Bereiche sind für unterschiedliche Populationen auffällig identisch.
Die Faustregel 1cM = 1MB (gilt nur ungefähr)
Es gibt so genannte **„recombination hot spots"**, d.h. Regionen mit hoher

Rekombinationswahrscheinlichkeit.Die Rekombinationswahrscheinlichkeit ist ein alternatives Maß zur physikalischen Positionsangabe und definiert die sogenannte Genetische Karte (genetic map) bzw. entsprechende Abstände (genetic map distance)

8.4 Aufgaben zur Übung 8

8.4.1 Aufgabe 1

8.4.2 Aufgabe 2

	LIFE-Adult (N=10000)	LIFE-Heart (N=7000)
Design	Zunächst Querschnittstudie	Kohortenstudie
Frage (konkret)	Identifizierung molekulargenetischer und umweltbedingter Faktoren für komplexer Erkankungen → Volkskrankheit	Identifizierung von Lebensstil- und molekulargenetischer Modifikatorebn des Atherosklerose-Risiko und verwandter Phänotypen (z.B. Lipidmetabolismus)
Frage (generell)	Wie gesund oder Krank ist die Bevölkerung?	Was haben die Kranken gemeinsam, sodass sich krankheiten entwickeln?
Vorteil	Billig, einfach durchführbar	Erfassung der Inzidenz eines Endpunktes und zeitlichen Zusammenhang zwischen Risikofaktor und Endpunkt
Nachteil	Ursache-Wirkung schlecht abbildbar	Teuer, seltene Endpunkte können nicht erfasst werden, selection bias

8.4.3 Aufgabe 3

Sie haben in der Vorlesung den Begriff Coverage kennengelernt.

1. Von was hängt die Coverage einer Microarrays ab?
   - „Qualität meines Arrays“, wie viel Prozent des Array-SNPs sind in hinreichend hohem LD mit den Refernz-SNPs.
   - Nimm Array-SNP und prüfe, ob dieser in der Referenz vorkommt bzw. in LD mit der Referenz-SNPs ist. Coverage ist der Anteil der in der Referenz vorkommenden SNPs
   - Abhängig von Referenz, Ethnien, LD-Niveau, cutt-off für seltene Varianten
2. Was sind die üblichen Referenz-Panels und wie unterscheiden diese sich? international HapMap Project, 1000 Genomes Project

3. Beschreiben Sie stichpunktartig den Workflow der Affymetrix Axiom Plattform!

8.4.4 Aufgabe 4

9 V9

9.1 Interpretation der Fixationsindices F_{st} und F_{is}

Fixationsindizes (F-Statistiken) sind Maße der genetischen Variabilität (Varianzkomponenten).

Verursacht durch evolutionäre Prozesse:

- Inzucht
- Migration
- Mutation
- Selektion
- Wahlund Effekt[16]

Fortpflanzung zwischen Verwandten Individuen führt zu einer Reduktion der Heterozygosität (im Vergleich zu HWE) und wird gemessen durch den Inzuchtskoeffizienten F
Zusammenhang zwischen F und Genotyp-Häufigkeiten für zwei Allele:
$p(A) = p(AA) + \frac{p(AB)}{2}$, $p(B) = 1 - p(A)$
$p(AA) = p(A)^2(1 - F) + p(A)F$
$p(AB) = 2p(A)p(B)(1 - f)$
$p(BB) = p(B)^2(1 - F) + p(B)F$

$$\Rightarrow F = 1 - \frac{O(AB)}{E(AB)} = 1 - \frac{p(AB)}{2p(A)p(B)}$$

F hängt offenbar von der Allelfrequenz einer Bezugs-Population ab

F_{IS} = Inzuchtkoeffizient eines Individuums I relativ zu Subpopulation S
F_{ST} = Maß für den genetischen Abstand einer Subpopulation S zur Gesamtpopulation T
F_{IT} = Inzuchtkoeffizient eines Individuums I relativ zur Gesamtpopulation T

[16]Die Vereinigung von zwei Populationen mit unterschiedlicher Allelfrequenz verringert die Heterozygosität im Vergleich zur HWE Erwartung.`https://de.wikipedia.org/wiki/Wahlund-Effekt`

Man kann die Fixationsindizes auch als Varianzkomponenten modellieren:

- F_{IS}:
  - entspricht der Varianzkomponente des Innersubjektfaktors innerhalb einer Subpopulation
  - $F_{IS} = 1 - \frac{\text{observed heterocygosity}}{heterocygosityunderHWE}$
- F_{ST}
  - entspricht der Varianzkomponente der Populationssubstruktur innerhalb der Gesamtpopulation
  - $F_{ST} = 1 - \frac{\sigma^2_{\text{within populations}}}{\sigma^2_{\text{between populations}}}$
- $F_{IT} = F_{IS} + F_{ST} - F_{IS}F_{ST}$

9.2 Bootstrap, Jackknife als Schätzverfahren für Standardfehler

Bootstrap

- Sei X eine Zufallsstichprobe, S eine interessierende Statistik auf X
- Betrachte die empirische Verteilung auf X
- Ziehe n=#X Zufallsstichproben aus X (mit Zurücklegen)
- Berechne S für diese Zufallsstichproben
- Wiederhole Ziehung der Zufallsstichproben und Berechnung von S
- Bestimme die Standardabweichung für diese Ergebnisse
- Diese Standardabweichung ist ein Schätzer des Standardfehlers von S auf X

Jackknife

- Sei X eine Zufallsstichprobe, S eine interessierende Statistik auf X
- Betrachte die n=#X Teilstichproben X_i die man erhält, wenn man einzelne Elemente wegläßt
- Berechne S für diese Zufallsstichproben $\rightarrow S_i$
- Berechne $\hat{\mu} = \frac{\sum_i^n S_i}{n}$ und $\hat{\sigma} = \sqrt{\frac{n-1}{n}\sum_i^n (S_i - \hat{\mu})^2}$
- Dies ist ein Schätzer des Standardfehlers von S auf X

9.3 Hauptkomponentenanalyse in der Genetik (Interpretation)

- "Structure" erkennt nur grobe genetische Unterschiede zwischen Populationen
- Hauptkomponentenanalyse vieler SNPs ist (viel) sensitiver
- Hauptkomponenten erlauben es, Genetische Assoziationen auf Populationsstruktur zu korrigieren

Vorgang:

1. Markermatrix normalisieren: $M(i,j) = \frac{C(i,j)-\mu(j)}{\sqrt{p(j)(1-p(,)}}$???
2. Wishart-Matrix aufstellen: $X = \frac{1}{n}MM'$
3. Eigenwerte berechnen und ordnen: $\lambda_1 > \lambda_2 > ... > \lambda_m$

Die hohe Empfindlichkeit der Methode erfordert eine sorgfältige Analyse:

- Angleichung der Fallzahlen der Cluster
- Sorgfältiger Umgang mit Ausreißern
- Strenge SNP-Qualitätskriterien
- SNPs vor Analyse entkorrelieren (pruning)
- Bekannte, kritische genomische Bereiche (konservierte Bereiche oder stark mit Abstammung assoziierte Bereiche) eliminieren
- Mit PCs assoziierte SNPs + Regionen eliminieren
- Funktioniert nicht ohne weiteres mit verwandten Individuen (Lösung: „Drop-one-in" Prozedur, Alternative: „swap-one-in")

9.4 ROH: Definition und Interpretation

Idee: Längere homozygote Bereiche im Genom weisen auf (ent- fernte) Verwandtschaft zwischen den Eltern hin (inbreeding).

Tritt das Phänomen in einer Population gehäuft auf, so spricht dies für einen kleinen Genpool (kleine Gründerpopulation = founder), bzw. geringen Austausch mit der Umgebung = Isolation

Praktische Probleme mit der Definition (SNP-Arrays):

- Genotypisierungsprobleme: einzelne, möglicherweise falsche heterozygote oder fehlende Genotypen in einem Bereich
- Wo beginnt, wo endet ein homozygoter Bereich?
- Die Chance auf Homozygotie hängt von der Anzahl der SNPs in einem Bereich bzw. der SNP-Dichte und den Allelfrequenzen ab
- Ab welcher Entfernung „teilt“ man homozygote Bereiche?
- Lösung: Hidden Markov Modelle

9.5 Aufgaben zur Übung 9

9.5.1 Aufgabe 1

a.) Was sind Batch-Effekte?
eine technische Quelle für Variation in den Daten durch die Verarbeitung[17]

b.) Durch was können sie entstehen, wie kann man sie vermeiden?
mögliche Quellen:

- **Spotting:** Die Menge der Probe in den Nadeln des Roboters, der damit das Array behandelt, kann leicht variieren.
- **PCR Amplikation:** Proben, die durch die Polymerase-Kettenreaktion(PCR) erzeugt werden, enthalten oft nicht die gleichen Vielfachen einer Sequenz, da die Amplikation der unterschiedlichen Nukleotidstränge mit unterschiedlicher Geschwindikeit verlaufen kann.
- **Probenaufbereitung:** bei der Vorbereitung der Proben ist eine Vielzahl komplexer biochemischer Reaktionen, wie zum Beispiel die reverse Transkription, durchzuführen. Diese können von Labor zu Labor und innerhalb eines Experiments Unterschiede aufweisen.

[17] `http://www.molmine.com/magma/global_analysis/batch_effect.html`

- **RNA-Abbau:** Unterschiedliche RNA-Stränge haben aufgrund ihrer Sekundärstruktur eine unterschiedliche Halbwertszeit. Um sie zu stabilisieren, werden eine Vielzahl von Gegenmaßnahmen angewendet, die auch Nebeneffekte nach sich ziehen können.

- **Array-Beschichtung:** Sowohl die Effizienz der Probenfixierung auf dem Array, als auch die Intensität des Hintergrundrauschens hängt stark von der Array-Beschichtung mit der Probe ab.

Diese Probleme sollten beim Design eines Mircoarray-Experiments beachtet werden. Kann man trotz allem einen Fehler nicht verhindern, so sollten die experimentellen Bedingungen so gewählt werden, dass die biologische Fragestellung nicht beeinflusst wird. Falls zum Beispiel ein Vergleich zwischen zwei Tumorprob en durchgeführt werden soll, so ist es ratsam, beide Prob en nicht in verschiedenen Labors aufbereiten zu lassen.[18]

c.) Erinnern Sie sich an Aufgabe 4 von Blatt 6. Statt verschiedener Populationen nehmen wir nun an, dass der SNP auf verschiedenen Platten gemessen wurde. Führen Sie einen Chi-Quadrat-Test durch, ob sich die Allelhäufigkeiten zwischen den Platten signifikant unterscheidet!

Ergebnisse siehe R-Skript

[18] `http://www-stud.rbi.informatik.uni-frankfurt.de/~linhi/SeminarSS04/Ausarbeitungen/03ausarbeitung_evgenji_yusuf.pdf`

10 V10

10.1 Heritabilität, Definition + Möglichkeiten zur Schätzung

Erklärte Varianz eines Merkmals durch Genetik.
Schätzung der Vererbbarkeit:

1. Zwillingsstudien: Vergleich der Merkmalskonkordanz zwischen ein- (MZ) und zweieiigen Zwillingen (DZ)
   Falconers Gleichung: $H^2 = 2r(MZ) - 2r(DZ)$

2. Familienstudien: Gemischte Modelle, Verwandtschaft = Kovarianz

Schätzen der Heritabilität aus genomweiten Daten gemischtes Modell:
$y_i = \mu + G_i + e_i$ mit
y_i: Phänoptyp Induviduum i
μ: Mittelwert
$G_i \sim \sigma^2_{Genetik} N(0, \Phi)$ (Polygenetischer Effekt)
Φ: Verwandschaftsmatrix
$e_i \sim \sigma^2_{Umwelt} N(0, 1)$ (Umwelteffekt)

Heritabilitätsschätzung: $H^2 = \frac{\sigma^2_{Genetik}}{\sigma^2_{Genetik} + \sigma^2_{Umwelt}}$

10.2 Genetische Assoziation (Prinzip)

SNP dient als Stellvertreter (proxy) eines Krankheitslokus. SNP muß nicht selbst kausal sein, sondern es genügt LD mit kausaler Variante.

Variante A: direkte Kausalität
Variante B: indirekte Kausalität über LD und proxy

10.3 Stratifikationsbias bei genetischen Studien

Gedankenexperiment:

- Phänotyp: Fähigkeit mit Stäbchen zu essen
- Studienpopulation: Gemisch aus Europäern und Asiaten
- Alle SNPs mit unterschiedlicher Allelfrequenz zwischen Europäern und Asiaten (viele!) sind mit dem Phänotyp assoziiert → Inflation der Teststatistiken → **Stratifikationsbias**

Andere Möglichkeit der Entstehung: Verwandtschaft zwischen den Individuen

Möglichkeiten der Bekämpfung von Stratifikationsbias (später genauer)

- Stratifizierte Analyse: Vorteil: funktioniert am besten, Nachteil: erfordert Kenntnis der Strata
- Adjustierung auf Hauptkomponenten: Vorteil: einfach durchführbar, Nachteil: Erhöht Freiheitsgrade der Assoziationsmodelle, beseitigt Stratifikation nicht immer vollständig, erfordert genomweite Daten (später)
- Local ancestry: Vorteil: Kann mit „Mischlingen“ umgehen, Nachteile wie bei „Adjustierung auf Hauptkomponenten“, erfordert genomweite Daten (später)
- Gemischte Modelle (Assoziation mit Korrelationsstruktur): Vorteil: geeignet bei Verwandtschaft, Nachteil: hohe Rechenbelastung
- Genomic Control: Einfache, phänomenologische Korrektur bei geringer Inflation, erfordert genomweite Daten (später)

10.4 Genetische Modelle und deren Schätzung

Allgemeines genetisches Modell ohne Kovariablen:
$y_i = \mu + \beta_{AB}\chi^i_{AB} + \beta_{BB}\chi^i_{BB} + e_i$ mit

y_i: Phänoptyp Induviduum i
μ: Mittelwert
$\chi^i_{XX} = 1$ falls Indiviuum i Genotyp XX hat, 0 sonst
β: Regressionskoeffizienten
$e_i \sim \sigma^2_{Umwelt} N(0,1)$ (Residualvarianz)

Modell	Vergleich	Tests
Additiv	AA vs. AB vs. BB	$\beta_{AB} \neq 0, \beta_{BB} = 2\beta_{AB}$
Dominant B	AA vs. AB, BB	$\beta_{AB} \neq 0, \beta_{BB} = \beta_{AB}$
Rezessiv B	AA, AB vs. BB	$\beta_{AB} = 0, \beta_{BB} \neq 0$
Heterozygotenvorteil	AA, BB vs. AB	$\beta_{AB} \neq 0, \beta_{BB} = 0$

- Additives Modell am flexibelsten
- Dominat B = Rezessiv A
- Heterozygotenvorteil selten (Beispiel Sichelzellanämie / Malaria)
- Problem: Genetisches Modell i.d.R. unbekannt

Strategien

1. • Man rechnet (trotz Unsicherheit über das tatsächliche Modell) mit nur einem (meist dem additiven) Modell
   • Rationale:
     – Modelle sind korreliert (Ausnahme Heterozygotenmodell)
     – Bsp: liegt tatsächlich eine Assoziation unter dem dominanten/rezessiven Modell vor, wird dies höchstwahrscheinlich auch im additiven Modell sichtbar
   • Diese Strategie wird überwiegend angewendet
   • Problem: Powerverlust bei falschem Modell, additives Modell erkennt Heterozygotenvorteil nicht
2. • Man vergleicht die Paßförmigkeit der Modelle für jede Situation (z.B. AIC), sucht sich das beste heraus und nimmt dessen p-Wert
   • Problem: Dieses Verfahren inflationiert den Typ 1 Fehler → erfordert Permutationstests, diese sind aufwendig
3. • Man testet mehrere Modelle und korrigiert nach Bonferroni
   • Problem: Überkonservativ, da Korrelation zwischen Modellen nicht berücksichtigt
4. Max-Test: ...

10.5 Spezifik gonosomaler Markeranalysen

- Die nPAR-Region ist bei Frauen auf einem Chromosom häufig inaktiviert.
- Die Inaktivierung ist wahrscheinlich(?) zufällig.
- Man modelliert die Inaktivierung indem man die Codierung der Genotypen der Männer entsprechend anpaßt:

$$\text{Codierung (Frau)} = \begin{cases} 0 & \text{für AA} \\ 1 & \text{für AB} \\ 2 & \text{für BB} \end{cases}$$

$$\text{Codierung (Mann)} = \begin{cases} 0 & \text{für A} \\ c & \text{für B} \end{cases}$$

mit c=1: Keine Inaktivierung; c=2: Vollständige Inaktivierung; 1<c<2: Unvollständige Inaktivierung

- Der Grad der Inaktivierung kann mit geschätzt werden (kompliziert)
- Man sollte unbedingt den Haupteffekt des Geschlechts in die Regressionsmodelle des X-Chromosoms einbauen

- Speziell bei Annahme keiner Inaktivierung, sollte die IA von Marker und Geschlecht mit ins Modell gesteckt werden.
- Obwohl Geschlecht ein unabhängiger Risikofaktor vieler Erkrankungen ist, heißt es nicht, dass auf Chr. X besonders viele Assoziationen zu finden sind (Beispiel: KHK)

10.6 Genomweite Assoziationsstudie

- Ziel: Identifikation genetischer Modifikatoren beobachtbarer Phänotypen
- Hinweise für Vererbbarkeit aus z.B. Zwillingsstudien
- „Komplexe Erkrankungen“ → Polygenetische Effekte: Häufige Varianten mit geringer Penetranz, seltene Varianten mit höherer (?) Penetranz
- Kandidatengenansätze häufig nicht replizierbar → hypothesenfreie Ansätze → Screening des Genomes mittels Marker (SNPs)
- Aktuell ca. 2240 publizierte GWAS, mehrere hundert verschiedenen Phänotypen

10.6.1 Ansatz

10.6.2 Replikation vs. kombinierte Analyse und Power

Replikation:

- Einzelanalyse der Top-Marker aus der ersten Stufe (GWAS)
- Verwendet nicht die Evidenz aus der ersten Stufe

Kombinierte Analyse:

- Kombiniert die Information aus erster und zweiter Stufe
- Achtung! Effektgrößen aus der erster Stufe sind inflationiert („winners curse“) ⇒ Korrektur (kompliziert)

Die Power der kombinierten Analyse ist immer höher als die der Replikation, aber Kombinierte Analyse hat nur dann (deutlich) höhere Power wenn:

- Mehrzahl der Samples in Stufe 1 und
- Viele Marker in der Replikation

10.6.3 Mehrstufendesign

1. GWAS: Genotype full set of SNP's in relatively small population at liberal p value
2. Replikationsstufe: Screen second, larger population at more stringent p value
3. optional third stage for increased stringency

10.6.4 Power

10.7 Aufgaben zur Übung 10

11 V11

11.1 Phänotyp, Genotyp-Phänotyp-Beziehung

Phänotyp: Erscheinungsbild/Merkmale eines Organismus (Morphologisch, Physiologisch, Psychisch)

- ererbt (Genotyp)
- erworben (Umwelt)
- akut (auf einen äußeren Reiz)

Phänotyp - Bestimmung:

- Messbarkeit
- Reliabilität
- Validität
- Vergleichbarkeit zwischen Studien

Intermediärer Phänotyp: Liegt in der Kausalbeziehung zwischen Genetik und Zielphänotyp; Beispiel: Cholesterin = intermediärer Phänotyp für Arteriosklerose

Genotyp-Phänotyp-Beziehung
Definition:

- Eine genetische Veränderung (Mutation) ist ursächlich für den Phänotyp (Kausalität)
- Grad der Abhängigkeit des Phänotyps vom Genotyp wird gemessen durch Heritabilität

Mögliche Ursachen für Abweichungen von einer strengen Genotyp-Phänotyp Beziehung:

- Phänokopie: Merkmalsausprägung aus anderer Ursache
- Phänotypische Plastizität: Modulierbarkeit durch Umwelteinflüsse
- Unvollständige Penetranz: Nichtausprägung trotz vorhandener Mutation, Kompensation eines Mechanismus
- Dramatyp: Phänokopie als Reaktion auf akutes Geschehen

11.2 Reliabilität, Validität

Reliabilität:[19]

- Anteil der Varianz von Messwerten, der durch tatsächliche Unterschiede des Merkmals begründet ist
- Hängt eng mit der Reproduzierbarkeit von Messungen zusammen
- Intra-Rater (observer) Reliabilität vs. Inter-Rater (observer) Reliabilität[20]
- grafische Darstellung mittels Bland-Altman Diagramm[21]

Konkordanz-Korrelations-Koeffizient (CCC)
geeignetes Maß zur Bewertung der Übereinstimmung zweier **quantitativer** Merkmale
$CCC(X,Y) = \frac{2cov(X,Y)}{var(X)+var(Y)+(E(X)-E(Y))^2}$

Cohen's Kappa
geeignetes Maß zur Bewertung der Übereinstimmung zweier **binärer** Merkmale
$\kappa = \frac{p_{00}+p_{11}-p_{0.}p_{.0}-p_{1.}p_{.1}}{1-p_{0.}p_{.0}-p_{1.}p_{.1}}$

Validität:

- Aussage zur Belastbarkeit einer Messmethode oder Operationalisierung. Wird tatsächlich das gemessen, was gemessen werden soll?
- Vergleich mit Goldstandard

11.3 (Genetische) Studiendesigns

11.3.1 Querschnittstudien

- „Cross-Sectional Study"[22]
- Untersucht gesamte Population oder repräsentative Zufallsstichprobe
- Momentaufnahme zu gegebenem Zeitpunkt: Analysiert Prävalenzen (kann nicht zwischen Effekt auf Inzidenz oder Dauer unterscheiden)
- Ungünstig für seltene Phänotypen
- Besonderheiten der Stichprobenziehung in Analysen berücksichtigen
- „Selective Survival Bias"
- Einfache Durchführbarkeit, sehr häufiger Studientyp

[19] https://de.wikipedia.org/wiki/Reliabilit%C3%A4t
[20] https://de.wikipedia.org/wiki/Interrater-Reliabilit%C3%A4t
[21] https://de.wikipedia.org/wiki/Bland-Altman-Diagramm
[22] https://de.wikipedia.org/wiki/Querschnitt_(empirische_Forschung)

11.3.2 Kohortenstudien

- „Cohort study“[23], Längsschnittliche Studie (longitudinal study“)
- Beobachtung des Auftretens eines Zielmerkmals in einer Population über einen gewissen Zeitraum
- Population ist initial frei vom Zielmerkmal
- Exposition wird anfänglich gemessen
- Anreicherung seltener Expositionen möglich
- Analysiert Risikofaktoren für Inzidenzen
- Aufwändig (Zeit und Geld)
- Probleme mit drop-out beim follow-up

11.3.3 Fall-Kontroll-Studien

[24]

- Definierte Fälle (mit Merkmal) und Kontrollen (ohne Merkmal)
- Definition einheitlicher Ein- und Ausschlusskriterien sehr wichtig
- Analysiert Expositionseffekte
- Vorsicht mit Confounding und Stratifizierung!
- Selective survival bias, differential recall bias
- Keine direkte Bestimmung des Relativen Risikos
- Fall-Kontroll-Matching kann schwierig sein

11.4 GxE Interaktion

Gene–environment interaction[25]

[23] `https://de.wikipedia.org/wiki/Kohortenstudie`
[24] `https://de.wikipedia.org/wiki/Fall-Kontroll-Studie`
[25] `https://en.wikipedia.org/wiki/Gene%E2%80%93environment_interaction`

11.5 Coverage von Microarrays

Maßzahl für die „Qualität“ des Inhalts eines Microarray-Produkts
Anteil der Referenz, die in hinreichend hohem LD (r^2) mit SNPs auf dem Microarray sind. Hängt ab von:

- Referenz (meist HapMap, 1000Genomes, verschiedene Panels)
- Ethnie (z.B. für afrikanische Populationen Coverage i.d.R. viel schlechter)
- gewünschtem LD-Niveau
- cut-off für seltene Varianten

11.6 Aufgaben zur Übung 11

12 V12

12.1 Calling von SNP-Daten

Intensitäten lassen sich mittels bioinformatischer Methoden übersetzen in „Genotyp einer Person an einem SNP“ → „Calling“

12.1.1 Calling-Algorithmen

Bei der Genotypisierung mittels Micro-Arrays werden Hybridisierungsintensitäten gemessen, die i.d.R. mittels Clusteranalysen in Genotypen umgerechnet werden Clusterplots→Genotypen
Wichtige Algorithmen:

- DM (dynamic model)
  - Calling-Algorithmus auf Basis einzelner Proben/Messungen
- BRLMM (Bayesian robust linear model)
  - benötigt die Information mehrerer Proben/Messungen

Genotypisierung ist immer fehlerbehaftet
Ziel: Eliminierung/Verringerung der Fehler(quellen) durch geeignete Filter

12.2 Clusterplots + Interpretation

- nach Genotypisierung erhält man Intensitätswerte (A und B) für die beiden Allele eines SNPs (bezeichnet mit a und b)
- man plottet nun für festen SNP und jede Person
  - auf der x-Achse: log 2 (A / B)
  - auf der y-Achse: (log 2 (A * B)) / 2
- Ergebnis: Clusterplot
- für qualitativ hochwertige SNPs sollten sich die Punktwolken gut trennen (in die Genotypen aa, ab und bb)

12.3 Maße zur Bewertung der Clusterplotirregularität

12.3.1 Typische SNP-QC Maße

- Fishers Linear Discriminant (FLD)
  - Problem:
  - Man bildet für die drei Gruppen (aa, ab und bb) jeweils die Mittelwerte der Einträge auf der x-Achse

  - Filterkriterium: FLD < 3.6

- Homozygote Ratio Offset (HomRO)
  - Problem: Homozygotencluster sollte ungefähr symmetrisch liegen
  - Filterkriterium:
    * 3 Cluster: HomRO < -0.9
    * 2 Cluster: HomRO < 0.3
    * 1 Cluster: HomRO < 0.6

- Heterozygous Cluster Strength Offset (HetSO)
  - Problem: Der AB-Cluster sollte höhere Intensität haben als von den AA / BB-Intensitäten zu erwarten wäre
  - HetSO ist der vertikale Abstand vom Mittelpunkt des AB-Clusters zur Verbindungslinie zwischen den Mittelpunkten des AA- und BB-Clusters
  - Filterkriterium: HetSO < -0.1

12.3.2 Typische Sample-QC Maße

- Callrate
  - Problem:
  - SNP-Call-Rate $= \frac{\#Calls\ für\ SNP}{\#Individuen}$
  - Filterkriterium: SNP-Call-Rate $< 97\%$

- Hardy-Weinberg Gleichgewicht
  - Problem: Verletzung des Hardy-Weinberg Gleichgewicht
  - Filterkriterium: $p < 10^{-6}$

- Minor allele frequency (MAF)
  - Problem: SNPs mit sehr geringer MAF sind aufgrund kleiner Cluster schlecht zu callen und haben außerdem nur geringen Informationsgehalt für Einzel-SNP-Assoziationen
  - Filterkriterium: MAF < 2

- Platten-Assoziation
  - Problem: Batcheffekte
  - mit Chi-Quadrat-Tests kann überprüft werden ob sich die Allelfrequenzen zwischen Platten unterscheiden
  - Filterkriterium: $p < 10^{-7}$

13 V13

Weitere Sample-Filter beruhen auf:

- Geschlechts-Analyse
- Verwandtschafts-Analyse
- Hauptkomponenten-Analyse (PCA nach den englischen „principal component analysis“)

13.1 Interpretation X-Y Intensitätsplots

Geschlechts-Analyse

- notwendig da zwei Quellen für das Geschlecht eines Probanden zur Verfügung stehen:
  - Probandendatenbank
  - Geschlechtsbestimmung beim Calling (computed gender)
  - Problem: computed gender wird (bei Affymetrix) nur über Heterozygosität des X-Chromosoms bestimmt
- Regel: DB-Geschlecht $\neq$ calling-Geschlecht $\rightarrow$ Proband filtern (obwohl mitunter auch DB-Eintrag falsch sein kann)
- weiteres Problem: beim Calling gibt es drei Geschlechtseinstufungen: female, male und unknown (für unknown Entscheidung anhand eines X-Y-Intensity-Plots & DB-Geschlecht)

Unregelmäßigkeiten in X-Y-Intensity-Plots

- YYX-Männer
- Frauen mit höherer Y-Intensität
- Monosomie X
- ungewöhliche X-Heterozygosität (Poly-X-Frauen)
- Geschlechtswidersprüche (gemessen vs. Datenbank)

13.2 Interpretation PCA

- vergleiche Grundlagen
- kann zur Plausibilisierung der Daten genutzt werden (Vergleich mit Referenz-Populationen)
- Identifikation von Ausreißern (ethnisch, schlechte Genotypisierung)
- Plausibilisierung von ethnischen Angaben (DB)
- Achtung - PCA interagiert mit Verwandtschaft!
  - verwandte Individuen sind i.d.R. PCA – Outlier
  - Lösung: „Drop one in procedure“ – sehr aufwendig

kann folgendes darstellen:

- Ethnische Ausreißer
- Batch-Effekte
- Plattform-Effekte
- Substrukturen in der Kohorte

13.3 CNV Detektion mit SNP-Array und Interpretation von R-Ratio und B-Allelfrequency plots

Problem

- CNV-Bestimmung aus Microarrays arbeitet direkt mit den Allel-Intensitäten
- kurze CNVs nur über spezielle CNV-Sonden detektierbar (z.B. Affymetrix SNP 6.0)
- mittellange CNVs i.d.R. auch schlecht detektierbar, Ergebnisse stark von Plattform abhängig
- nur sehr lange CNVs sind gut detektierbar (z.B. Tumor-DNA)
- SNP-Arrays nur bedingt für Keimbahn-CNV-Analyse geeignet (besser Sequenzierung)

R Ratio: Beobachtete Intensität / Referenzintensität, bei höherer Intensität Duplikation, bei geringerer Deletion

B-Allelhäufigkeit: Anzahl der Banden ergeben Ordnung der CNV (z.B. Anzahl Duplikationen)

14 V14

14.1 Prinzip der Genotyp-Imputation

Hintergrund: bei nichteindeutiger Clusterzuordnung der Intensitäten beider Allele wird Genotyp auf „fehlend“ gesetzt → „Löcher“ im Datensatz

→ durch Imputation können diese fehlenden Genotypen geschätzt und ersetzt werden

Mögliches Vorgehen:

- Ausfüllen der Löcher im Datensatz ohne Referenz
- Ausfüllen der Löcher im Datensatz mit Referenz
- Schätzen von nichtgemessenen Genotypen mittels Referenz

Nutzen:

- Vervollständigen der Daten für Analysezwecke
- Konstruktion einer gemeinsamen Marker-Menge für genetische Metaanalysen von Studien mit unterschiedlichen Genotypisierungsplattformen
- Erhöhung der Power
- Korrektur von Genotypisierungsfehlern (in geringem Maße)

Problem:

- Imputation ist ein Schätzverfahren
- Unsicherheit der resultierenden Genotypen muss bei der nachfolgenden Analyse geeignet berücksichtigt werden

14.2 Aufbau eines HMMs

siehe Vorlesung

14.3 Probleme des Referenzabgleichs

Problem: Die Daten müssen zur Referenz insbesondere deren Annotationen passen

Wichtige Informationen zu SNP-Daten werden mit zwei Versionsangaben versehen:

- **NCBI build (reference genome)**
  - Angaben zur Lage der SNPs im humanem Genom
  - betrifft Chromosom und Basenposition
  - beides kann sich zwischen Versionen verändern
- **dbSNP build**
  - Database of single nucleotide polymorphisms (SNPs) and multiple small-scale variations
  - gelistete SNPs werden auf Referenzgenom gemappt
  - alle SNPs erhalten eine offizielle "rsID" als SNP identifier (ändert sich zwischen Versionen!)

Vorgang

1. Eigene SNPs brauchen eine rsID: Falls keine gefunden, muss SNP gefiltert werden oder über Position auf Referenz abgebildet werden
2. NCBI build der eigenen Daten muss mit dem des Referenzdatensatzes übereinstimmen: Falls nicht gegebenen, muss einer der Datensätze geliftet werden (vorzugsweise der eigene Datensatz)
   - Hilfreiches Tool: LiftOver der UCSC (später)
   - Liftet SNPs anhand der gegebenen Positionen von einer Version des Human Genomes (HG-Version) zur nächsten
   - ändert nur Positionsangaben, die richtigen rsIDs für neue Positionen müssen selbst gesucht werden
   - dabei können SNPs verloren gehen
3. dbSNP build Identifier der eigenen Daten muss mit dem des Referenzdatensatzes übereinstimmen: Falls nicht gegeben, muss über die Positionsangaben aus dem gewünschten dbSNP build die korrekte rsID ermittelt werden
   - Achtung: für einige Positionen gibt es mehrere rsIDs
   - wenn möglich, rsID des Referenzdatensatzes übernehmen
4. SNP-Informationen sollten zwischen Referenz und eigenem Datensatz übereinstimmen

- identische Positionsangaben
- identische rsIDs
- identische Basenkombination (bzgl. des richtigen Strangs)
- Letzteres erfordert erfahrungsgemäß die meisten Probleme insbesondere bei A/T, C/G SNPs und Frequenzen nahe 50%
- Alle SNPs mit ungeklärten Widersprüchen sollten aus den eigenen Daten gefiltert werden, werden „mit etwas Glück“ durch Imputation wieder eingebracht

14.4 Messen der Imputationsqualität

Imputation liefert zwei relevante Wahrscheinlichkeiten:

- p_1 = Wahrscheinlichkeit für Heterozygot
- p_2 = Wahrscheinlichkeit für Homozygot BB

Wenn die wahren Genotypen bekannt sind, kann man Maße der Imputationsgüte über Abstände von Verteilungen definieren: Was ist das alles???

Die wahren Genotypen sind i.d.R. jedoch nicht bekannt. Die Imputations- güte wird dann aus der Abweichung der geschätzten Genotypverteilung von der zufällig erwarteten (durch Raten) ermittelt:
Alleldodid B des i-ten Individuums des j-ten SNPs: $e_{ij} = p_{ij1} + 2p_{ij2}$
???: $f_{ij} = p_{ij1} + 4p_{ij2}$
Empirische Allelfrequenz des j-ten SNPs: $\hat{\theta} = \frac{\sum_{i=1}^{N} e_{ij}}{2N}$

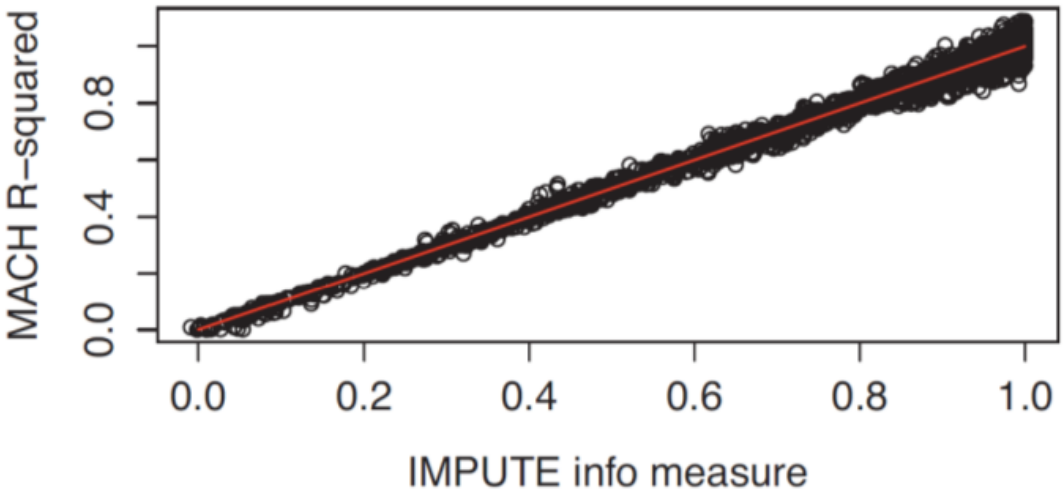
MaCH r^2
empirische Varianz der Gendosen geteilt durch die unter HWE erwartete Varianz

$$\hat{r}_j^2 = \begin{cases} \frac{\frac{\sum_{i=1}^{N} e_{ij}^2}{N} - (\frac{\sum_{i=1}^{N} e_{ij}}{N})^2}{2\hat{\theta}(1-\hat{\theta})} & \text{when } \hat{\theta} \in (0,1) \\ 1 & \text{when } \hat{\theta} = 0, \hat{\theta} = 1 \end{cases}$$

IMPUTE info score
1-Varianz der geschätzten Genotypen / Varianz der Genotypen bei zufälliger Stichprobenziehung mit θ

$$I_A = \begin{cases} 1 - \frac{\sum_{i=1}^{N}(f_{ij} - e_{ij}^2)}{2N\hat{\theta}(1-\hat{\theta})} & \text{when } \hat{\theta} \in (0,1) \\ 1 & \text{when } \hat{\theta} = 0, \hat{\theta} = 1 \end{cases}$$

führt zu Plot:



14.5 Einflußfaktoren auf die Imputationsqualität

- Referenzgenom
- Software (auf Ethnien)

14.6 Problematik der Assoziationsanalyse mit imputierten Genotypen

Die Unsicherheit bei der Imputation von Genotypen muß bei der Analyse berücksichtigt werden.

Variante 1 (best-guess genotype): Man nimmt den Genotyp mit der höchsten posterior-Wahrscheinlichkeit und rechnet mit diesem weiter. Mitunter werden dabei zu geringe posterior-Wahrscheinlichkeiten auf „missing“ gesetzt (=neue Löcher).

Variante 2 (Alleldosis): Man bestimmt die Erwartung der posterior-Verteilung (=Alleldosis) und rechnet mit dieser weiter.

Variante 3 (Mischmodell): Man fittet ein Mischmodell an (nicht zu verwechseln mit „mixed model“!).

z.B. Additives Modell:

$$f_k(\mu, \beta, \epsilon_i) = \begin{cases} \mu + \epsilon_i, & k = 0 \\ \mu + \beta + \epsilon_i, & k = 1 \\ \mu + 2\beta + \epsilon_i, & k = 2 \end{cases}$$

$y_i = \sum_{k=0}^{2} p_{ki} f_k(\mu, \beta, \epsilon_i)$
Erfordert numerische Optimierung der Likelihood → Höherer Rechenaufwand

Variante 4 (Score-Test): Wird als „state of the art“ angesehen
Score-Funktion (L=Likelihood, θ=genetischer Effekt): $U(\theta) = \frac{\delta log L(x,\theta)}{\delta \theta}$

Fisher-Information: $I(\theta) = -E[\frac{\delta^2}{\delta \theta^2} log L(x, \theta)]$

Verteilung unter H_0: $S(\theta_0) = \frac{U(\theta_0)^2}{I(\theta_0)} \sim \chi_1^2$

Problematisch ist die mangelnde Stabilität dieses Ansatzes, bei kleinen Allelfrequenzen!

15 VL15

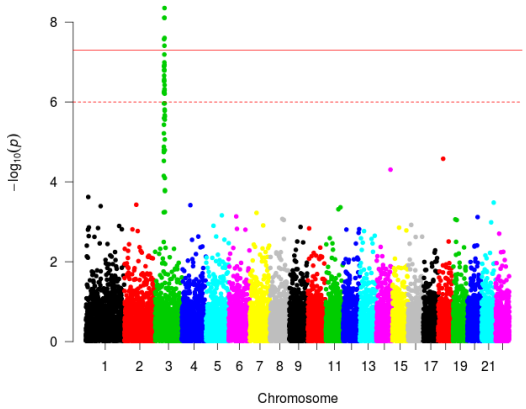
GWAS Aufgaben

- Assoziationstests
  - Einzel-SNP, verschiedene genetische Modelle
  - Scoring-Tests
  - Haplotypbasierte Analysen, Fine-Mapping
  - Metaanalyse
  - SNPxSNP, SNPxKovariablen Interaktionen
  - Subgruppenanalysen (Power-Problem)
- Post-Analyse QC
- Graphische Aufbereitung
- Extraktion von Kandidaten
- Replikation
- Vergleich mit Online Ressourcen

15.1 Pedigree Format

- Jede Zeile ein Individuum
- Vier IDs: Familie, Individuum, Vater, Mutter
- Vater, Mutter – ID muß als Individuum-ID auftreten (0 = fehlend)
- Pro SNP zwei Spalten (Allele)
- Datenformat erforderlich für viele Softwarepakete
- (Gewisse) Familienstruktur ist mit gegeben

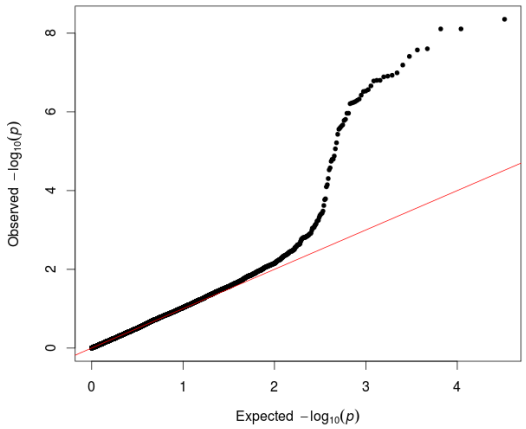
15.2 Manhattan Plots



Suggestive-Grenze=10^{-6}
Genomweite Grenze=$5 \cdot 10^{-8}$

15.3 QQ-Plots

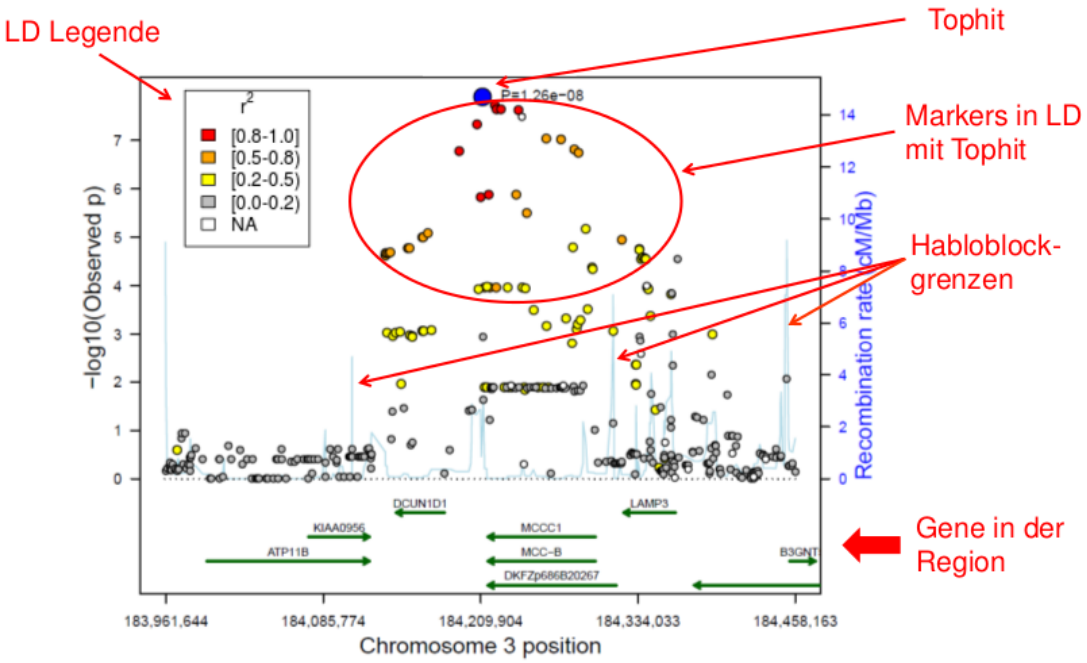
Idee: Die Mehrzahl der SNPs sollte nicht assoziiert sein, d.h. einer Nullverteilung entsprechen



Wenn Punkte der Mittellinie bis zur Mitte folgen → Inflation der Teststatistiken → erfordert Korrektur

<u>Genomic Control:</u>
Inflationsfaktor λ berechnen, wenn $\lambda < 1.05$ → keine Korrektur notwendig ⇒ siehe **Genomic Control**

15.4 Regional Association plot (RA)



Was kann man hier so ablesen?

15.5 Genomic Control

Seien $Y_1^2,\ldots,Y_m^2$ unabhängige, Chi-Quadrat-verteilte Zufallsvariablen (z.B. aus Fall-Kontrolltests, m = Anzahl Marker)

Unter H0 (keine Assoziationen) gilt: median$(\tilde{Y}_1^2,\ldots,\tilde{Y}_m^2)$=0,456
Definiere Inflationsfaktor: $\lambda = \frac{\text{median}(Y_1^2,\ldots,Y_m^2)}{0{,}456}$
Korrigiere Teststatistiken: $(\bar{Y}_1^2,\ldots,\bar{Y}_m^2) = \frac{Y_1^2,\ldots,Y_m^2}{\lambda}$
Wie kommt man von p-Werten zu Chi-Quadrat-verteilte Zufallsvariablen?

16 V16

17 V17

18 V18

18.1 Fixed-effects Modell

Schätzung (Fixed-effect Modell):
Angenommen die wahren Effekte sind zwischen den Studien homogen:
$\theta_1 = \theta_2 = ... = \theta_k$

Dann ist $\hat{\theta} = \frac{\sum_{i=1}^{k} \hat{\theta}_i \cdot w_i}{\sum_{i=1}^{k} w_i}$ ein Schätzer des gemeinsamen Effekts θ. Ein 95%-Konfidenzintervall kann abgeleitet werden: $\hat{\theta} \pm 1.96 \cdot \sqrt{\frac{1}{\sum_{i=1}^{k} w_i}}$

18.2 Random-effects Modell

Schätzung (Random-effects Modell):
Angenommen die Effekte sind zwischen den Studien heterogen. Nehmen hierarchisches Modell an, um die Effekte zu kombinieren. Hierzu wird angenommen, dass die wahren Effekte der Studien θ_i normalverteilt sind mit Mittelwert θ:

Das heisst $\theta_i \sim N(\theta, \tau^2)$ mit Varianz τ^2 welche die Heterogenität der Studien quantifiziert. Warum τ^2???

In jeder einzelnen Studie hat man $\theta_i \sim N(\theta, w_i^{-1})$ mit den Gewichten aus dem Fixed-effect Modell.

Für die Randverteilung von θ_i gilt: $\theta_i \sim N(\theta, w_i^{-1} + \tau^2)$

Deshalb hat unter einem Random-effects Modell der Fixed-effect Schätzer $\hat{\theta}$ immer noch die Erwartung θ aber eine größere Varianz

$$var(\hat{\theta}) = \frac{\sum w_i^2 \cdot var(\hat{\theta}_i)}{(\sum w_i)^2} = \frac{\sum w_i \cdot (w_i^{-1} + \tau^2)}{(\sum w_i)^2} = \frac{1}{(\sum w_i)} + \frac{\tau^2 \sum w_i^2}{(\sum w_i)^2}$$

entsprechend hat das Random-effects Modell den selben Schätzer des Meta-Effekts aber eine größere Varianz (und folglich größeres Konfidenzintervall, größere p-Werte usw.).

18.3 Fixed effect versus Random effects Modell

Fixed effect

- Konsistent mit der globalen Nullhypothese
- Homogenitätsannahme könnte unrealistisch sein
- Vorliegende Heterogenität muss erklärt werden
- Äquivalent zu einem „random intercept“- Regressionsmodell

Random effects

- Inkonsistent mit der globalen Nullhypothese
- Formale Berücksichtigung der Heterogenität
- Annahme Normalverteilung der Studieneffekte könnte auch unrealistisch sein
- Äquivalent zu einem „random slope-random-intercept“ Regressionsmodell

18.4 Forest plot

zeigt:

- Effekte der einzelnen Studien + Konfidenzintervall
- Gewichte der Studien
- Meta-Effekte und qualitative Bewertung der Heterogenität

18.5 Permutationstest

- Dilemma:
  - Fixed Effect Modell berücksichtigt mögliche Heterogenität nicht
  - Random Effects Modell ist genomweit überkonservativ (λ typischerweise deutlich kleiner 1)
- Idee: Man testet auf Heterogenität und entscheidet sich dann für das "richtige" Modell
- Problem: Dieser Sequentialtest hält das Irrtumsniveau nicht ein, ein Aspekt der selbst in einigen hochrangigen Publikationen nicht ausreichend berücksichtigt wurde
- **Lösung: Permutationstest**

Permutationstests gehören zu den so genannten nichtparametrischen Verfahren und beruhen auf "Resampling" (Analogie zu Jackknife/Bootstrap)

- Idee:
  - Bei Tests werden i.d.R. zwei (oder mehr) Größen miteinander in Beziehung gesetzt
  - Unter der Nullhypothese "kein Zusammenhang" sollte die Verbindung zwischen den Größen egal sein
  - Betrachte alle möglichen Verbindungen (oder eine Zufallsziehung daraus) und überprüfe, ob der vorliegende Zusammenhang extrem ist (z.B. in weniger als 5% der möglichen Fälle auftritt)

Dieses Prinzip läßt sich auf viele komplexe Situationen anwenden, ist aber auch sehr rechenintensiv

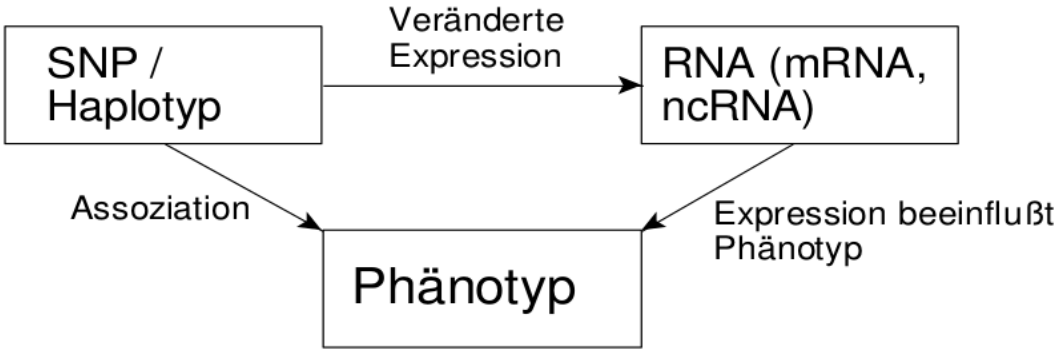
- Beispiel: Mittelwertsvergleich zwischen zwei Gruppen (Varianzgleichheit erforderlich!)
- Gruppen: 0, 1, Fallzahlen: n_0, n_1 , Werte: x_{0i}, x_{1j} mit i=1,...,n_0, j=1,...,n_1
- Berechne die Differenz der Mittelwerte μ_0 und μ_1 der zwei Gruppen: $d = \mu_0 - \mu_1$
- Lege eine Reihenfolge der Messwerte fest und permutiere die zugehörigen Gruppenlabels
- Berechne erneut die o.g. Differenz
- Wiederhole das Verfahren für alle Permutationen (exakter Test) oder für eine zufällige hinreichend große Auswahl an Permutationen

- Die so erhaltenen Differenzen liefern die Nullverteilung
- Die ursprüngliche Differenz wird mit dieser Nullverteilung verglichen (Man kann z.B. bestimmen, wieviel Prozent der Differenzen betragsmäßig größer sind als d. Dies liefert einen (empirischen) zweiseitigen Permutationstest p-Wert bezüglich der Hypothese d=0)

19 V19

19.1 Typische Konzepte der Genexpressions-Präprozessierung

Motivation



Genotyp-Phänotyp Assoziationen ergänzt durch funktionelle Analysen (z.B. Genotyp-Expression oder Expression-Phänotyp Korrelationen) ergeben Hinweise auf Pathomechanismen.

Prozessierung:

- RNA Extraktion
- RNA Aufreinigung / Fällung
- Sondensynthese / Hybridisierung / Scan
- Datenpräprozessierung (Normalisierung, Varianzstabilisierung, Filterungen, Adjustierungen) → Ziel: vertrauenswürdige Daten
- Datenanalyse

Typische Designfalle

- Wenn bei einer Fall-Kontroll-Studie die Fälle und Kontrollen in verschiedenen Batches laufen, hat man hinterher kaum eine Chance, Batcheffekte von den interessanten Kontrasten zu unterscheiden
- Ideal ist eine Balancierung der Fälle und Kontrollen
- Eine gleichmäßige Verteilung ist auch ok
- Deshalb ist es wichtig, den Meßprozeß bereits aus Sicht einer späteren Auswertung zu beeinflussen

Vergleich verschiedener Normalisierungs- und Varianzstabilisierungsmethoden

- **Hintergrundkorrektur:** Korrektur der Kontraste innerhalb eines Arrays
- **Normalisierung:** Angleichung der Intensitäten unterschiedlicher Arrays
- **Varianzstabilisierung:** Angleichung der Streuung unterschiedlicher Transkripte

Bestes Vorgehen hängt von Technologie und Frage ab

Präprozessierung Illumina HT12v4 (Beispiel)

1. **Ausgangspunkt:** Genexpressionsintensitäten von 47.000 Sonden
2. **Individuenfilter 1:** Individuen mit geringer genomweiter Genexpression
3. **Transkriptfilter 1:** Signal-to-noise Filter
4. **Reduktion technischer Artefakte 1:** Quantilnormalisierung, Varianzstabilisierung (log2-Transformation)
5. **Individuenfilter 2:** Individuen mit atypischen technischen Qualitätsparametern
6. **Reduktion technischer Artefakte 2:** Reduktion von Batcheffekten mittels Empirical Bayes Methoden
7. **Individuenfilter 3:** Ausreißer bezüglich Genexpression
8. **Reduktion biologischer Artefakte 3:** Residualisierung bezüglich Alter, Geschlecht, Blutwerte (Mo, Ly, Gra, Ret), Rauchen, ggf. unbekannter Batch-Effekte

Resultat der Präprozessierung: 28.000 Sonden

Was heißt „exprimiert“?

- Man prüft Intensität des Transkripts im Vergleich mit Leersonden (Sequenzen die im Genom keine Entsprechung haben, d.h. nicht hybridisieren sollten)
- Detection p-value = Anteil Leersonden mit höherer Expression (kein eigentlicher p-Wert)
- 5% (andere verwenden 1%) als cut-off zur Entscheidung „exprimiert“ vs. „nichtexprimiert“
- Stark gewebsabhängig
- Über alle Gene und Individuen hinweg sind ca. 60% der Gene / Transkripte in Blut (PBMCs) exprimiert

19.2 Typische Filter / Probleme

Individuenfilter 1

- Individuen mit atypischer Zahl exprimierter Gene

Transkriptfilter 1

- Transkript wird gefiltert, wenn z.B. in weniger als 5% der Individuen exprimiert
- Z.B. im Blut: Von den Genen sind ca. 60% in mehr als 5% der Individuen transkribiert, also gültig bzgl. Transkriptfilter 1

Quantilnormalisierung

- Idee: Bilde Quantile der Intensitäten aufeinander ab um Intensitätsprofile anzugleichen

Reduktion technischer Artefakte 1

- QL = Quantilnormalisierung + log2-Transformation(Varianzstabilisierung)
- Für was steht QL?

Individuenfilter 2

- Idee: Exkludiere Individuen mit atypischen Ergebnissen in speziellen Kontrollsonden
  - Filtere Individuen mit Zu großem Mahalanobis-Abstand zum Mittelwert der internen Qualitätsparameter = atypische (schlechte) Qualität
  - **Mahalanobis-Abstand:** Seien x,y Realisierungen einer multinominalen Normalverteilung mit Kovarianzmatrix S (in unserem Falle stellen die Qualitätsparameter für ein Sample eine solche Realisierung dar), so berechnet sich die Mahalanobisdistanz wie folgt:
    $d(x, y) = \sqrt{(x-y)^T S^{-1}(x-y)}$ Was ist T???

Reduktion technischer Artefakte 2

- Batch-Effekte verursacht durch Meßprozeß (hauptsächlich Hybridisierung)
- Problem: Viele Batches geringer Größe $\rightarrow$ keine klassische Adjustierung möglich (Regression)
- Lösung: Empirical Bayes – Verfahren

Individuenfilter 3

- Idee: Betrachte Genexpression als multivariaten Phänotyp, Filtere Individuen mit zu großem Abstand vom Zentrum = atypisches Expressionsprofil

Reduktion technischer Artefakte 3

- Adjustierung auf biologische Faktoren

19.3 Konzepte für Anreicherungsanalysen

Anreicherungsanalysen = Überrepräsentationsanalyse

Verwendet z.B. hypergeometrische Verteilung, um die Wahrscheinlichkeit zu bestimmen, in einer Anzahl von Versuchen (Vordergrund) eine bestimmte Anzahl Treffer in der Gesamtmenge (Hintergrund) zu erzielen

Wie groß ist die Wahrscheinlichkeit (p-Wert) dass genau 260 Gene (von 747) auf die Erkrankung abbildbar sind, gegeben dass 3005 dieser Gene im Hintergrund sind (13,101 genes)? → **Hypergeometrischer Test / Fisher's exakter Test**
Limitationen

- Reihenfolge der Assoziationsstärke der Top-Hits wird nicht beachtet
- Richtung der Effekte wird nicht beachtet
- Richtung von Gen-Gen-Interaktionen wird nicht beachtet
- Verschiedene Cut-offs für Genlisten erzeugen unterschiedliche Ergebnisse
- Überlappungen zwischen Pathways werden nicht berücksichtigt (Unabhängigkeit der Tests ist nicht gegeben)
- Signifikanz wird tendenziell überschätzt, wenn ein Gen in vielen Pathways liegt
- P-Werte aus diesen Analysen sollten eher als qualitative Orientierung dienen und keinesfalls überbewertet werden

Functional Class Scoring: GSEA

- GSEA: gene set enrichment analysis
- Ordnung der Gene nach Signifikanz. Liegen Gene eines Subsets tendenziell weiter vorn in der Liste?
- Signifikanztest per Komolgornov Smirnov/Permutation
- Vorteile:
  - Benötigen keinen Cutoff der Topliste
  - Sensitiv zu kleineren Effekten, wenn Sie alle im gleichen Pathway liegen
- Limitationen
  - Überlappungen zwischen Pathways werden nicht beachtet
  - Richtung des Zusammenhangs wird nicht beachtet

- Richtung von Gen-Gen-Interaktionen wird nicht beachtet

Pathway Topologie
Bewertet nicht nur Vorkommen des Gens im Pathway, sondern auch Rolle des Gens: Beeinflusst es zentral mehrere Gene (a) oder weniger (b) erzeugen verschiedene Ergebnisse Wie muss ich das lesen???
Limitationen

- Überlappungen zwischen Pathways werden nicht beachtet
- Topologische Annotationen oft nur eingeschränkt verfügbar
- Verlässlichkeit des eingepreisten Wissens nicht berücksichtigt
- Verschiedene Cut-offs erzeugen verschiedene Ergebnisse

20 V20

20.1 eQTL

- **E**xpression **Q**uantitativ **T**rait **L**ocus
- Genetik des Transkriptoms“ = GWAS für Genexpressionen
- Welche genetischen Faktoren ($\sim 10^7$) korrelieren mit welchen RNA-Transkripten ($\sim 10^4$)?
- eQTL = Schnittmenge zwischen Phänpty, genetischen Faktoren und Genexpression

Anwendung

- Funktioneller Relevanz / Validierung genetischer Hits (z.B. GWAS – Hits)
- Identifizierung neuer genetischer Risikofaktoren
- Aufklärung grundlegender biologischer Zusammenhänge bzw. Krankheitsmechanismen

20.2 Cis/trans Effekte

- cis–eQTL:
  - SNP korreliert mit benachbarter Expressionssonde
  - Stärkere Effekte
  - Massenphänomen aktuell bei mehr als der Hälfte aller Gene beobachtet
  - Weniger Falschpositive Unmittelbare Hypothesen bzgl. kausalem Gen $\rightarrow$ besonders relevant für Follow up
- trans–eQTL:
  - SNP korreliert mit distaler Expressionssonde
  - Teilweise bedeutend geringere Effektstärke
  - Erfordern größere Studien
  - Liefern Einblicke in die genetische Regulation (nc-RNA, Epigenetik)
  - Könnten GWAS-Hits in „Genwüsten“ erklären

Trennung ist schwammig und uneinheitlich, da „benachbart“ schlecht definierbar ist. Wir verwenden meist 1MB.

20.3 Bedeutung von eQTLs

20.3.1 Für Verständnis der Regulation der Genexpression

???

20.3.2 Zur Erklärung genetischer Assoziationen

???