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

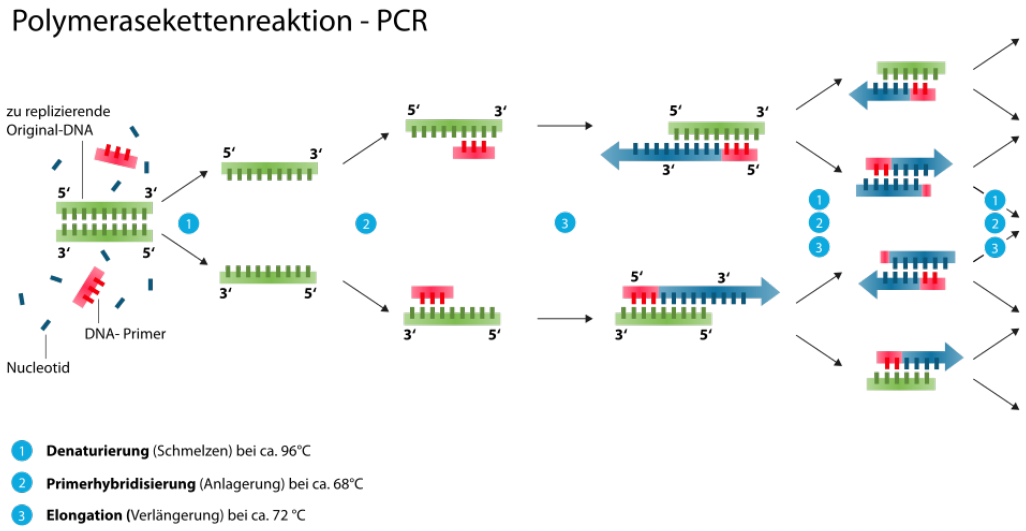
- SNP[2]:
  - Single Nucleotide Polymorphism - Einzelnukleotid-Polymorphismus
  - Variation eines einzelnen Basenpaares in einem DNA-Strang
  - SNPs sind geerbte und vererbbare genetische Varianten. Begrifflich davon abzugrenzen ist der Begriff der Mutation, der in der Regel eine neu aufgetretene Veränderung 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

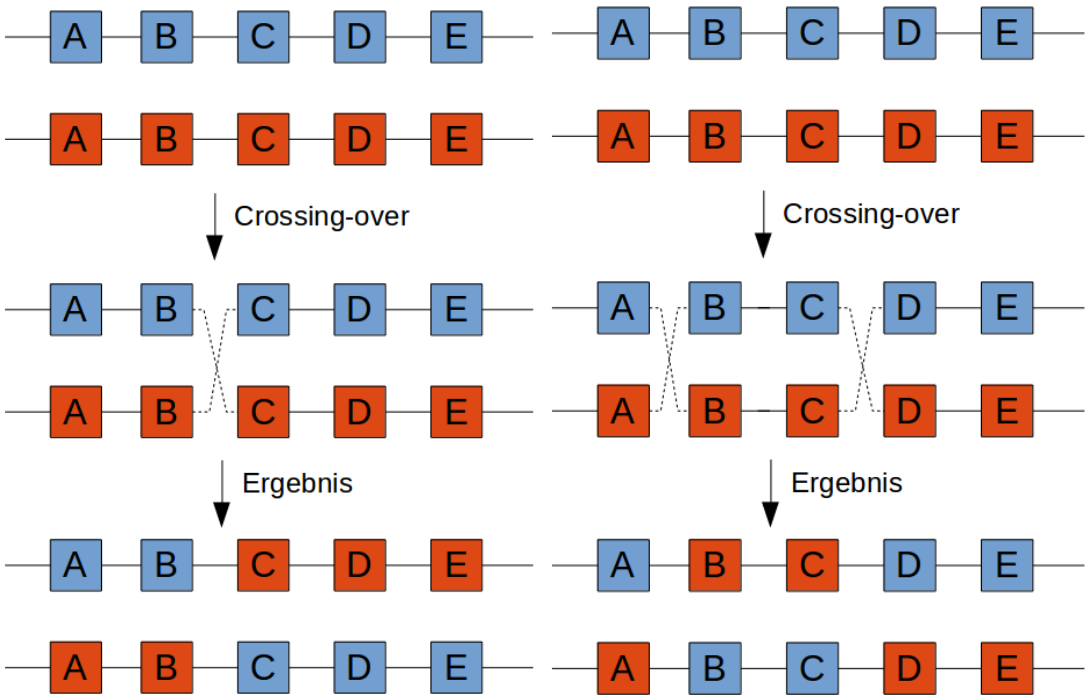
2.8.1 Aufgabe 1

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

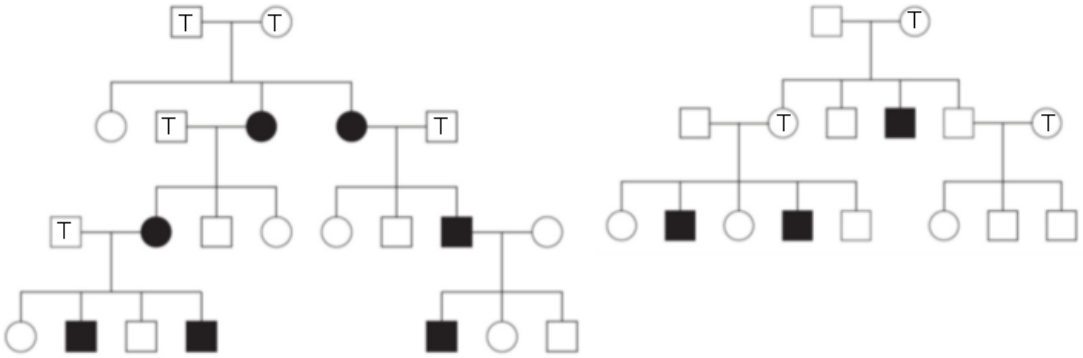
2.8.3 Aufgabe 3

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

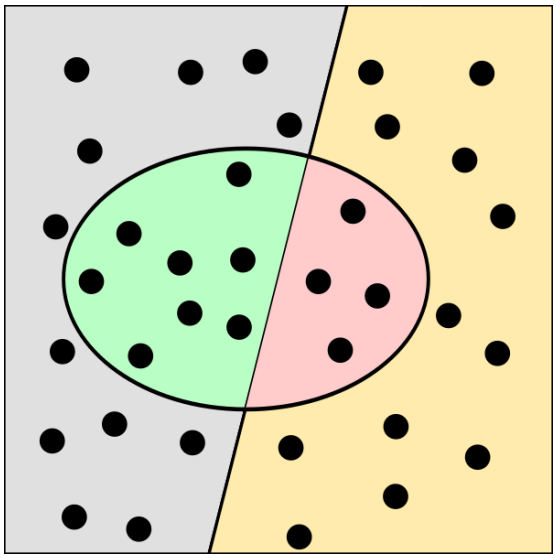
3.5.1 Aufgabe 1

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

5.2 Logik des statistischen Testens, Testdurchführung und Interpretation

5.3 Typ I und Typ II Fehler, Einfluß der Fallzahl

5.4 Problem des multiplen Testens und Korrekturmöglichkeiten

5.5 Faktoren für die Auswahl des richtigen Tests

5.6 Zusammenhangsmaße auf Vierfeldertafeln

5.7 Korrelation, Scheinkorrelation und Confounder

5.8 Aufgaben zur Übung 5

6 V6

6.1 Lineare Regression

6.1.1 Modellannahme

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

6.1.3 Varianzzerlegung und erklärte Varianz bei linearer Regression

6.1.4 Multivariate Regression

6.1.5 AIC

6.2 Multivariate Regression

6.2.1 Schätzen von Kontrasten

6.2.2 AIC

6.2.3 Interaktion

6.3 Auswahl einer passenden Regressionsmethode

6.4 Aufgaben zur Übung 6

7 V7

7.1 Motivation und Ansatz für gemischte Modelle

7.2 Feste und zufällige Effekte

7.3 Idee der Hauptkomponentenanalyse

7.4 Interpretation PCA-Plots und Eigenwerte

7.5 Aufgaben zur Übung 7

8 V8

8.1 Hardy-Weinberg Gleichgewicht incl. Test

8.2 Kinship-Koeffizient, Verwandtschaftsschätzung

8.3 Kopplungsungleichgewicht

8.3.1 Entstehung und Entwicklung

8.3.2 Bewertung (Maße)

8.3.3 Bedeutung (Interpretation, Tagging, LD-Heatmaps)

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}

9.2 Bootstrap, Jackknife als Schätzverfahren für Standardfehler

9.3 Hauptkomponentenanalyse in der Genetik (Interpretation)

9.4 ROH: Definition und Interpretation

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[16]

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.
- **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

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

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.[17]

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

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

10 V10

10.1 Heritabilität, Definition + Möglichkeiten zur Schätzung

10.2 Genetische Assoziation (Prinzip)

10.3 Stratifikationsbias bei genetischen Studien

10.4 Genetische Modelle und deren Schätzung

10.5 Spezifik gonosomaler Markeranalysen

10.6 Genomweite Assoziationsstudie

10.6.1 Ansatz

10.6.2 Replikation

10.6.3 Mehrstufendesign

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:[18]

- 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[19]
- grafische Darstellung mittels Bland-Altman Diagramm[20]

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"[21]
- 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

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

11.3.2 Kohortenstudien

- „Cohort study“[22], 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

[23]

- 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[24]

[22] `https://de.wikipedia.org/wiki/Kohortenstudie`
[23] `https://de.wikipedia.org/wiki/Fall-Kontroll-Studie`
[24] `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: $p < 10^{-7}$

12.4 Aufgaben zur Übung 12