Übung "Statistische Aspekte der Analyse molekularbiologischer und genetischer Daten"

Inhaltsverzeichnis

1 Übung 1: Biologische Grundlagen – Teil 1 **1**
1.1 Aufgabe 1 . . . 1
1.2 Aufgabe 2 . . . 1
1.3 Aufgabe 3 . . . 1
1.4 Aufgabe 4 . . . 1
1.5 Aufgabe 5 . . . 2
1.6 Aufgabe 6 . . . 3

2 Übung 2 **4**
2.1 Aufgabe 1 . . . 4
2.2 Aufgabe 2 . . . 4
2.3 Aufgabe 3 . . . 5
2.4 Aufgabe 4 . . . 5

3 Übung 3 **7**
3.1 Aufgabe 1 . . . 7
3.2 Aufgabe 2 . . . 8
3.3 Aufgabe 3 . . . 8

4 Übung 8 **9**
4.1 Aufgabe 1 . . . 9
4.2 Aufgabe 2 . . . 9
4.3 Aufgabe 3 . . . 9
4.4 Aufgabe 4 . . . 9

5 Übung 9 **10**
5.1 Aufgabe 1 . . . 10
5.2 Aufgabe 3 . . . 10
5.3 Aufgabe 3 . . . 10

6 Übung 10 **12**
6.1 Aufgabe 1 . . . 12
6.2 Aufgabe 2 . . . 12
6.3 Aufgabe 3 . . . 12

1 Übung 1: Biologische Grundlagen – Teil 1

1.1 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

1.2 Aufgabe 2

1.3 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

1.4 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

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

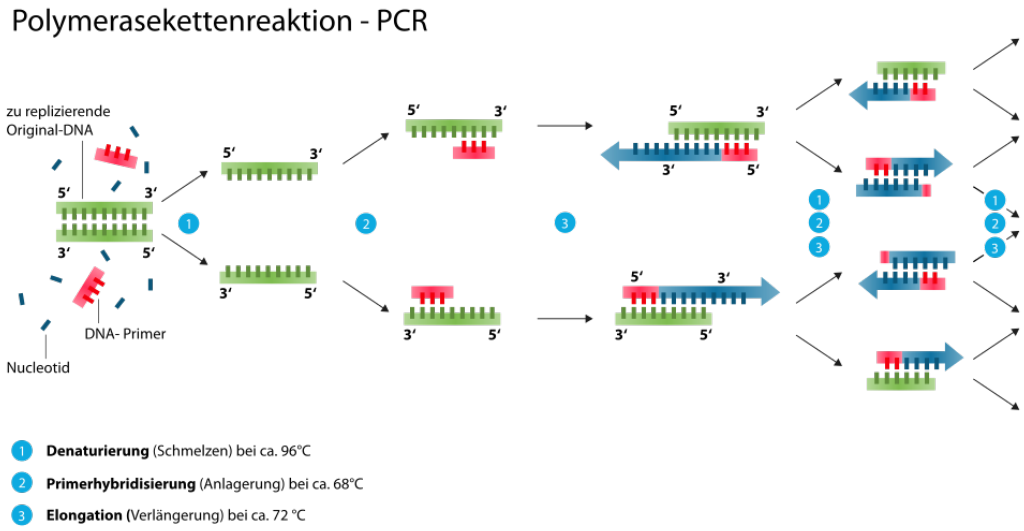
- Deletion: Ein Teilstück des Chromosoms (Endstück oder mittlerer Abschnitt) geht verloren
- Translokation: Chromosomen können auseinanderbrechen und dabei Teilstücke verlieren, welche in die Chromatide eines anderen Chromosoms angeheftet werden
- Duplikation: Ein Abschnitt des Chromosoms ist doppelt 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

1.5 Aufgabe 5

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

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.

[5] `https://de.wikipedia.org/wiki/Polymerase-Kettenreaktion`



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

1.6 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

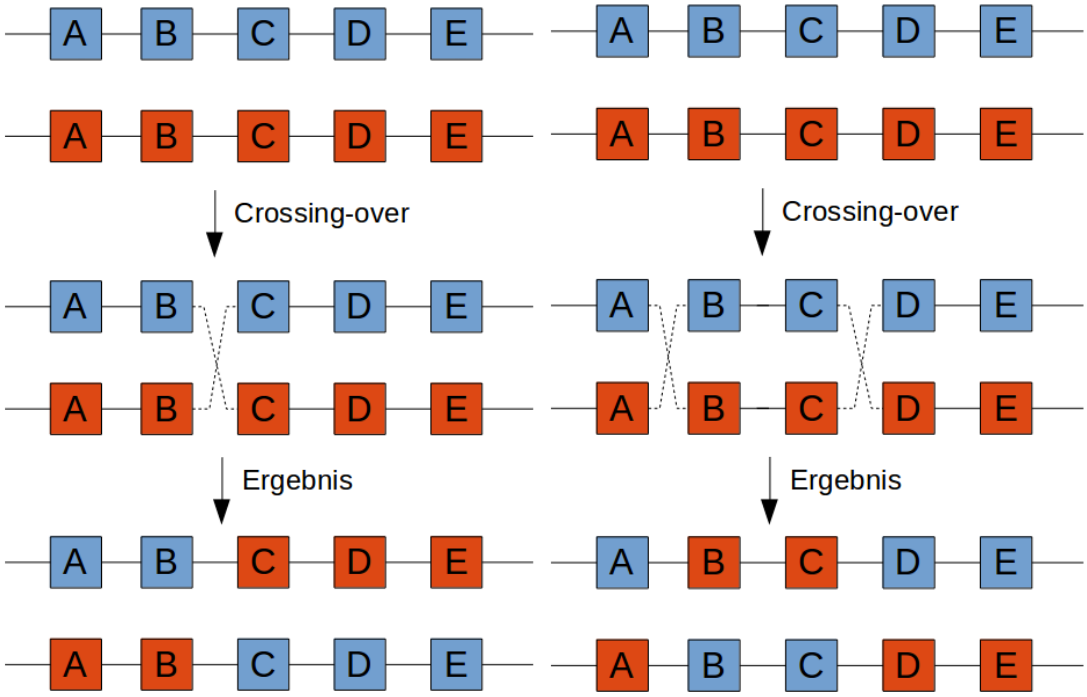
2 Übung 2

2.1 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

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



2.2 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

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

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

Kombinationsmöglichkeiten:

- praktisch durch Allele vorgegeben: $3 \cdot 2 \cdot 2 = 12$[12]
- theoretisch: $5^3 = 125$
- Musterlösung: 3 SNPs auf einem Allel $\rightarrow$ 8 Kombinationen; 2 Allele: 36 Möglichkeiten

A und B kodominant, Faktor 0 rezessiv

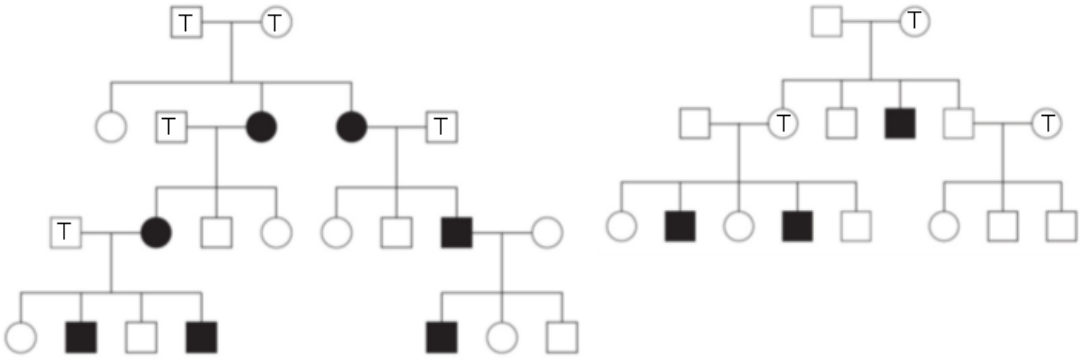
2.3 Aufgabe 3

a.)
b.)
c.)

2.4 Aufgabe 4

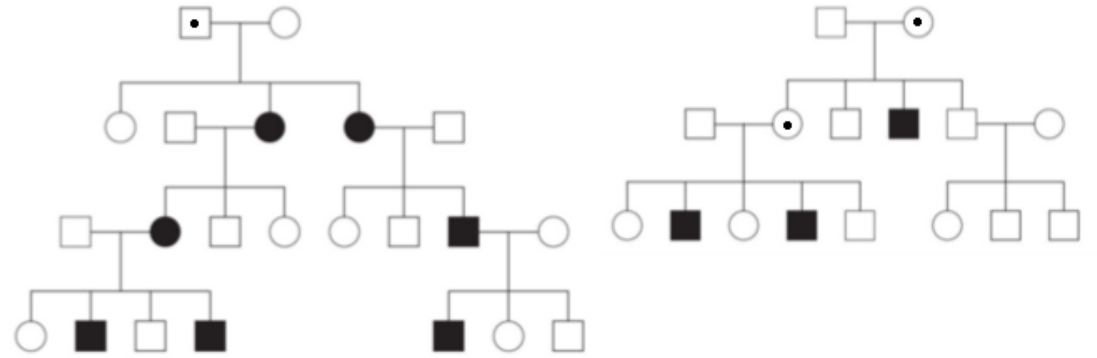
a.)
rezessiv:[13] bedeutet in der Genetik „zurücktretend“ oder auch „nicht in Erscheinung tretend“
dominant:[14] ein dominantes Allel setzt sich in der Merkmalsausprägung gegenüber einem rezessiven Allel durch
Penetranz:[15] prozentuale Wahrscheinlichkeit, mit der ein bestimmter Genotyp zur Ausbildung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

b.)



[12] https://sites.google.com/site/abobloodgroup/14.aboalleles%28oalleles%29
[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 Übung 3

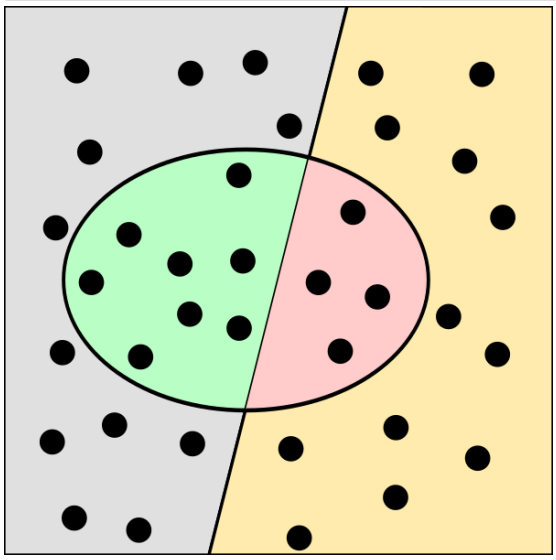
3.1 Aufgabe 1

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

3.2 Aufgabe 2

3.3 Aufgabe 3

siehe R-Script

4 Übung 8

4.1 Aufgabe 1

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

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!

4.4 Aufgabe 4

5 Übung 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 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

5.2 Aufgabe 3

siehe R-Script (prüfungsrelevant: Plots und Inhalte bewerten)

5.3 Aufgabe 3

a.)
siehe vorherige Aufgabe, Outlier außerhalb von Clustern

b.)
Callrate ausrechnen, wenn Callrate $< 97\% \rightarrow$ Missings entfernen

[16] `http://www.molmine.com/magma/global_analysis/batch_effect.html`
[17] `http://www-stud.rbi.informatik.uni-frankfurt.de/~linhi/SeminarSS04/Ausarbeitungen/03ausarbeitung_evgenji_yusuf.pdf`

c.)
Imputation:

1. Referenzabgleich: Welche Array-SNPs sind in der Referenz und sind sie gleich codiert? (ID, Chromosom, Position, Strang, Allel A, Allel B)

2. Phasierung: Schätzung der Haplotypen

6 Übung 10

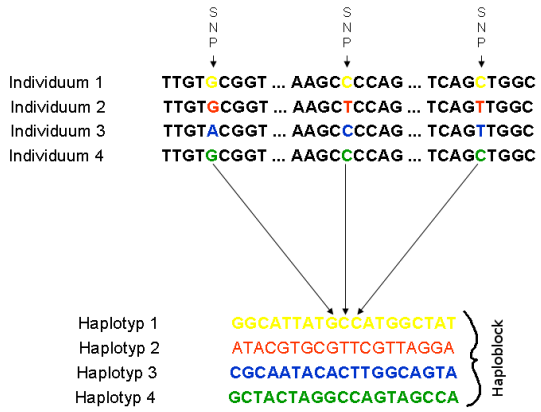
6.1 Aufgabe 1

siehe R-Skript (Prüfungsrelevant: QQ-Plot bewerten (Grenze beachten), Manhatten-Plot (Signifikanzgrenzen beachten))
signifikante Assoziation zwischen betrachteten SNP und phenotypischen Merkmal
so viele grüne übereinander da in einem Block (ohne Korrektur durch LD)

6.2 Aufgabe 2

a.)

- Haplotyp[18]: Variante einer Nukleotidsequenz auf ein und demselben Chromosom im Genom eines Lebewesens
- Haploblock: Haplotypen verschiedener Individuen als Block geschrieben sind ein Haploblock
- Rekombinations-Hotspot[19]: Bereich im Genom der eine erhöhte Rekombinationsrate im Vergleich zur neutralen Erwartung aufweist



Haplotypen aus SNPs von Chromosomenabschnitten des gleichen Chromosoms von vier haploiden Individuen

b.)
862 weitere SNPs auf Gen ABO[20]
Aus snpedia: 36
c.)
Durch die drei bekannten SNPs sind die relevanten Haploblöcke vollständig beschrieben
d.)
wenige SNPs reichen aus, um den Genotyp der restlichen SNPs im Block zu bestimmen

6.3 Aufgabe 3

a.)
Mischen von fixed- und random-Effects
b.)
$y_{ijk} = \mu + \beta_j + b_j + \epsilon_{ijk}$

[18] https://de.wikipedia.org/wiki/Haplotyp
[19] https://en.wikipedia.org/wiki/Recombination_hotspot
[20] Vorgehen siehe https://www.ncbi.nlm.nih.gov/guide/howto/view-all-snps/

Proband: i=1-6
Ergometer: j=1-3 Wiederholung: k=1-3
β_j: Effekt des Ergometers j (fixer Effekt)
b_i: Effekt des Probanden i (zufälliger Effekt, Variation innerhalb der Probanden)
jeder Proband hat seine eigenen Intercept, aber alle haben die gleiche Steigung
Beispiel oben: Random Intercept Modell == fixed effect meta analyse, denn Studien haben eigenen Intercept

c.)
d.)