Populationsgenetik 3: Kopplungsungleichgewicht (LD)

Peter N. Robinson

Institut für medizinische Genetik
Charité Universitätsmedizin Berlin

1. Dezember 2014

Outline

Hardy-Weinberg-Gesetz

- Letztes Mal ...
- Eigenschaften eines *einzelnen* Genlocus: Allelfrequenzen, Genotypfrequenzen
- Hardy-Weinberg-Gesetz: Beziehung zwischen Allel- und Genotypfrequenzen

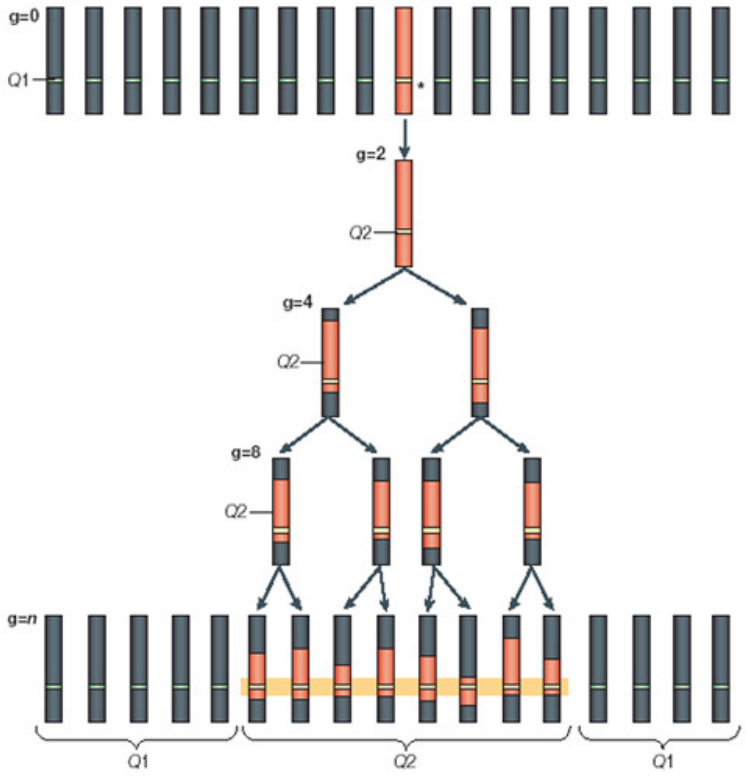
$$p^2 + 2pq + q^2$$

Kopplungsungleichgewicht

- Dieses Mal ...
- Eigenschaften von Gruppen von Genorten
- Haplotypen
- Kopplungsgleichgewicht
- Kopplungsungleichgewicht (englisch: Linkage Disequilibrium, LD)

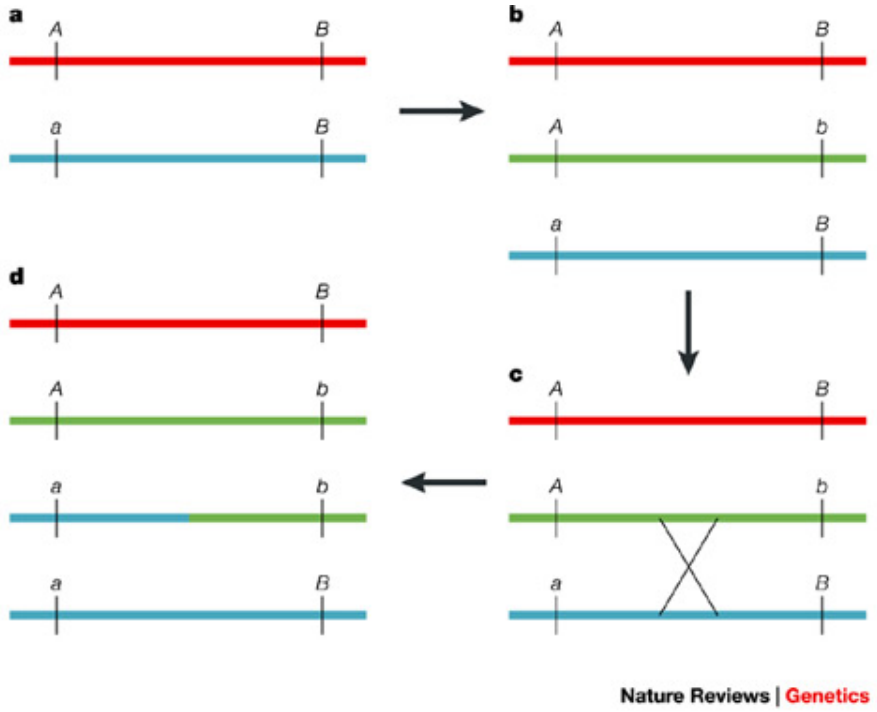
Eine Geschichte zweier Mutationen

- Heute existierende Haplotypen sind durch weit zurückliegende Mutationsereignisse entstanden



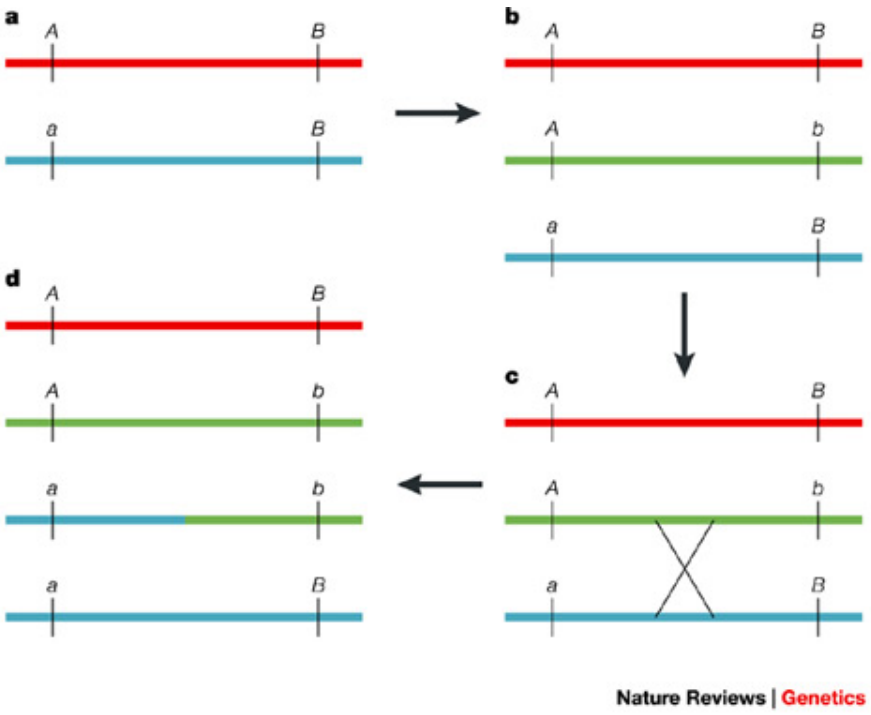
- Das Allel **Q2** entsteht in der Generation 0 durch Mutation von Allele **Q1**.
- Zunächst besteht ein komplettes LD zwischen Q2 und den Allelen aller anderen Loci im ersten Gameten, welcher Q2 trägt
- Dieses LD verringert sich allmählich im Laufe der Generationen auf Grund von Rekombination
- LD bleibt für eng benachbarte Loci bestehen
- In Generation *n* kann noch ein minimaler Haplotyp um Q2 bei mehreren Individuen identifiziert werden (gelber Balken)

Eine Geschichte zweier Mutationen (2)



a Anfangs besteht ein polymorpher Locus mit zwei Allelen, **A** und **a**.

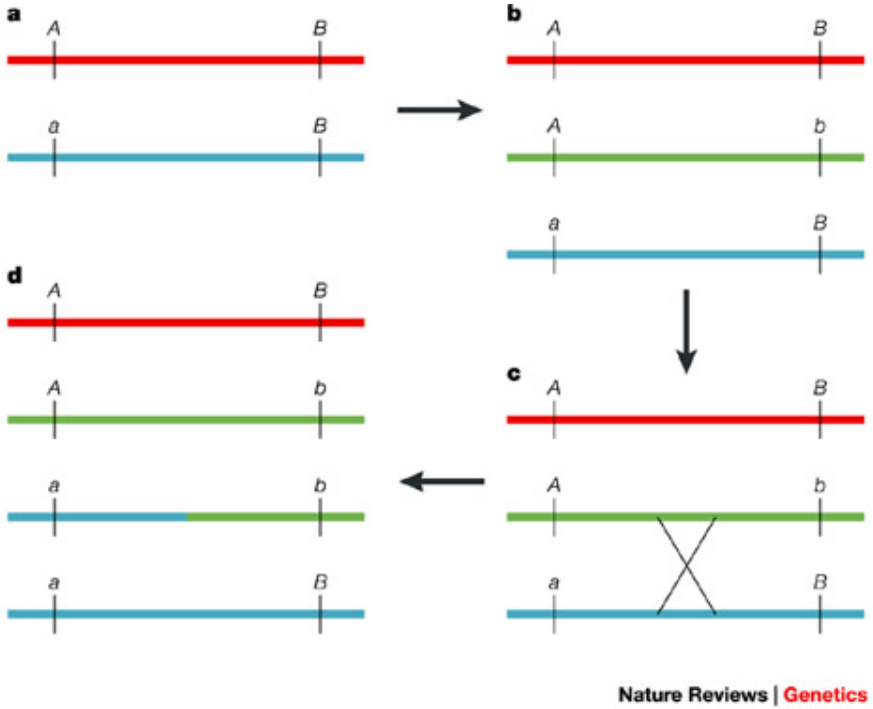
Eine Geschichte zweier Mutationen (3)



b Tritt eine Mutation an einem benachbarten Locus auf (**B** $\rightarrow$ **b**), betrifft die Mutation ein einzelnes Chromosom mit entweder dem **A**- oder dem **a**-Allel.

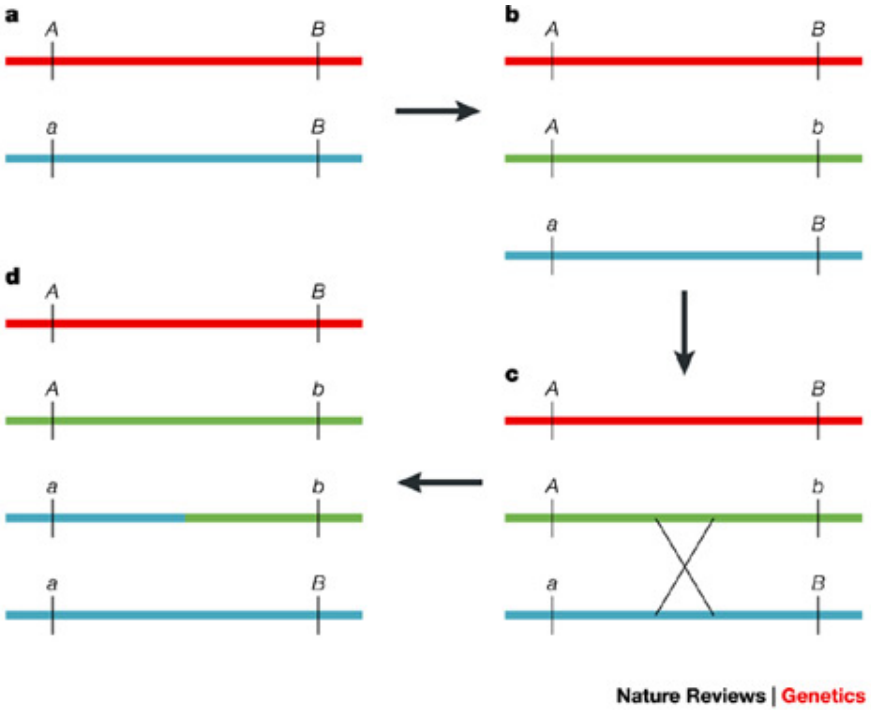
Daher werden früh in der Geschichte der Mutation nur drei der möglichen vier Haplotypen in der Population beobachtet (In diesen Beispiel: **AB**, **aB** und **Ab** aber nicht **ab**).

Eine Geschichte zweier Mutationen (4)



- Die anfänglich absolute Assoziation zwischen den Allelen an den beiden Loci wird im Laufe der Generationen durch Rekombination zwischen den Loci verringert.

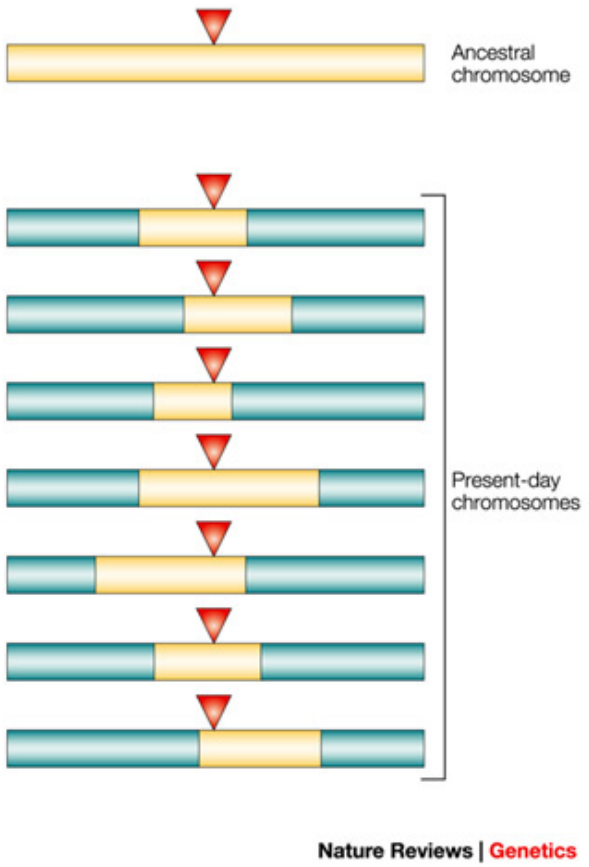
Eine Geschichte zweier Mutationen (5)



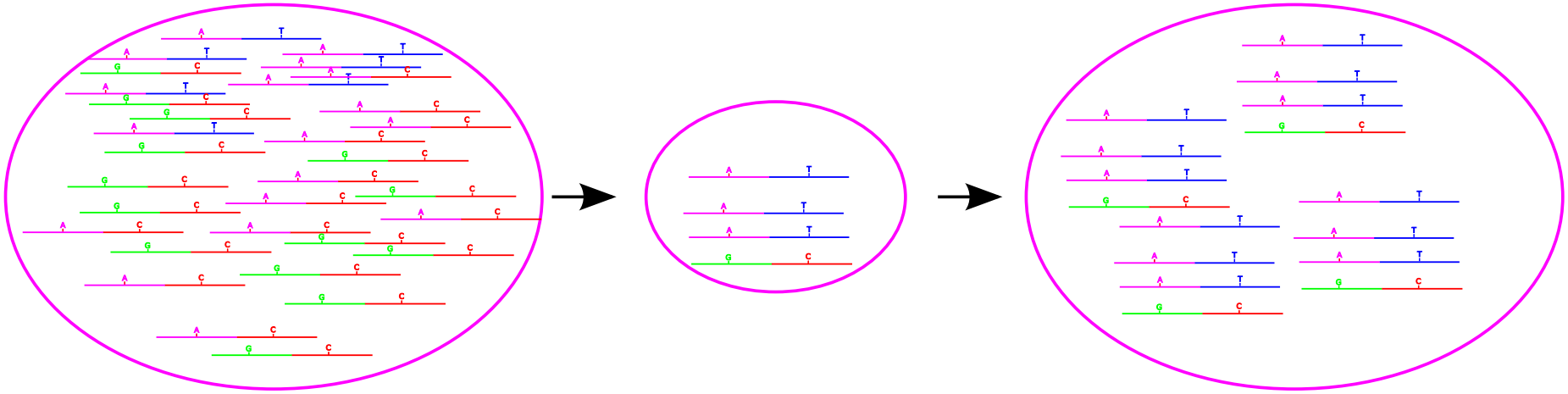
d Das Ergebnis ist ein vierter Haplotyp und eine Abnahme des LD durch eine Zunahme der Frequenz des rekombinanten Haplotypes **ab** in der Population.

Kopplungsungleichgewicht

- Linkage Disequilibrium: **LD**
- Chromosomen sind Mosaike
  - Rekombination
  - Mutation
  - Genetische Drift
  - Natürliche Auslese (Selektion)
- Kombinationen von Allelen in nah beieinander liegenden Loci: weit zurückliegende (”ancestral”) Haplotypen



Gründereffekt und LD



- Eine in der Stammpopulation bestehende, große genetische Variabilität reduziert sich bei der Gründung einer Kolonie durch wenig Individuen
- Häufige Ursache von LD in menschlichen Populationen

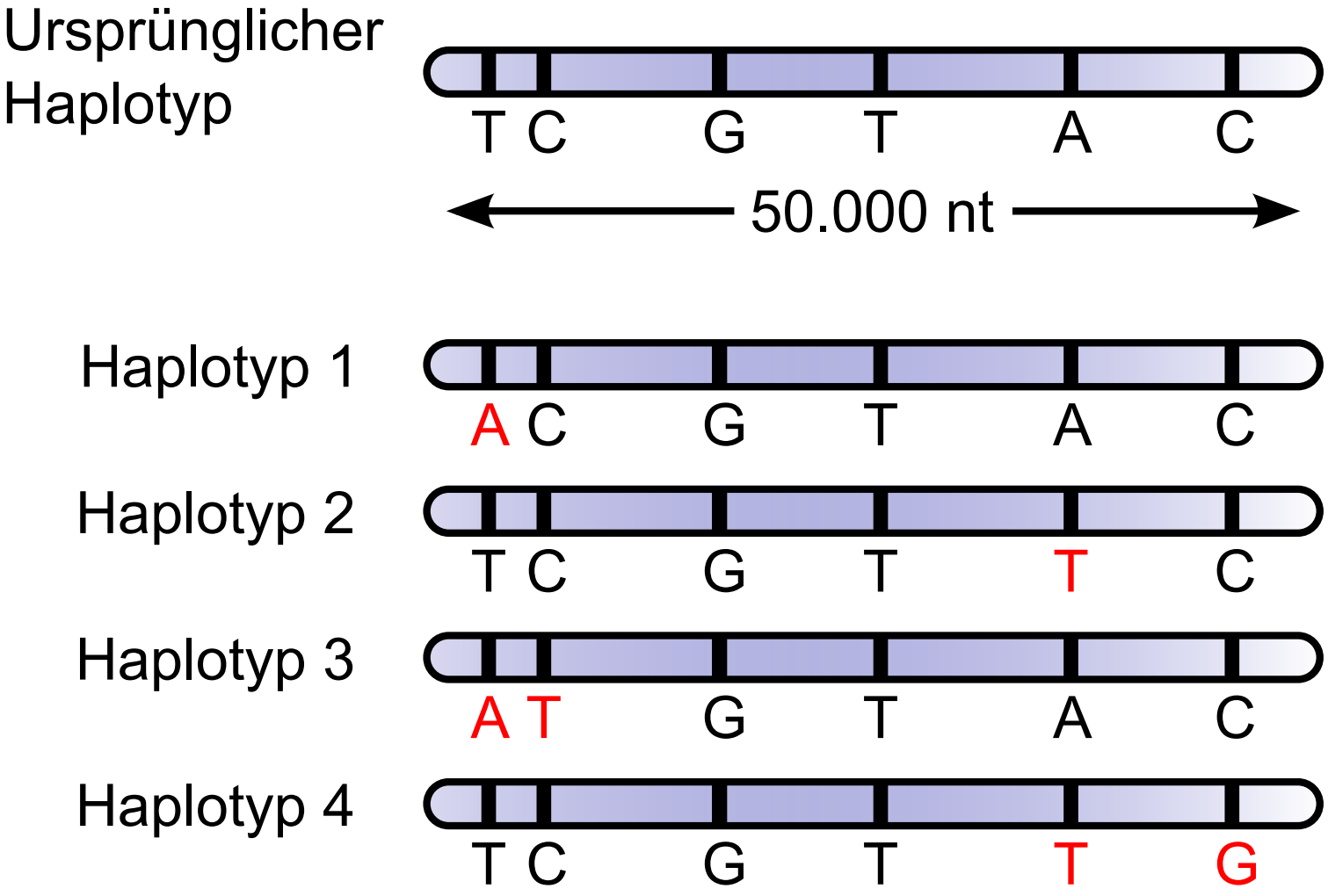
LD und Kartierung (Vorschau)

- In einer späteren Vorlesung werden wir die Bedeutung des LD für die Entdeckung von genetischen Varianten, die mit einer erhöhten Anfälligkeit für häufige Krankheiten wie Diabetes, Herzinfarkt, Schlaganfall, Allergie, ..., eingehen
- Dieses Mal wollen wir die mathematischen Hintergründe und die biologischen Grundlagen erklären
- *Wichtig*: Unterscheide zwischen Linkage und Linkage Disequilibrium (LD):
  - **Linkage**: gemeinsame Vererbung zweier Loci in Familien
  - **LD**: Beziehung zwischen zwei Allelen an 2 Loci in einer Population

Genotyp vs. Haplotyp

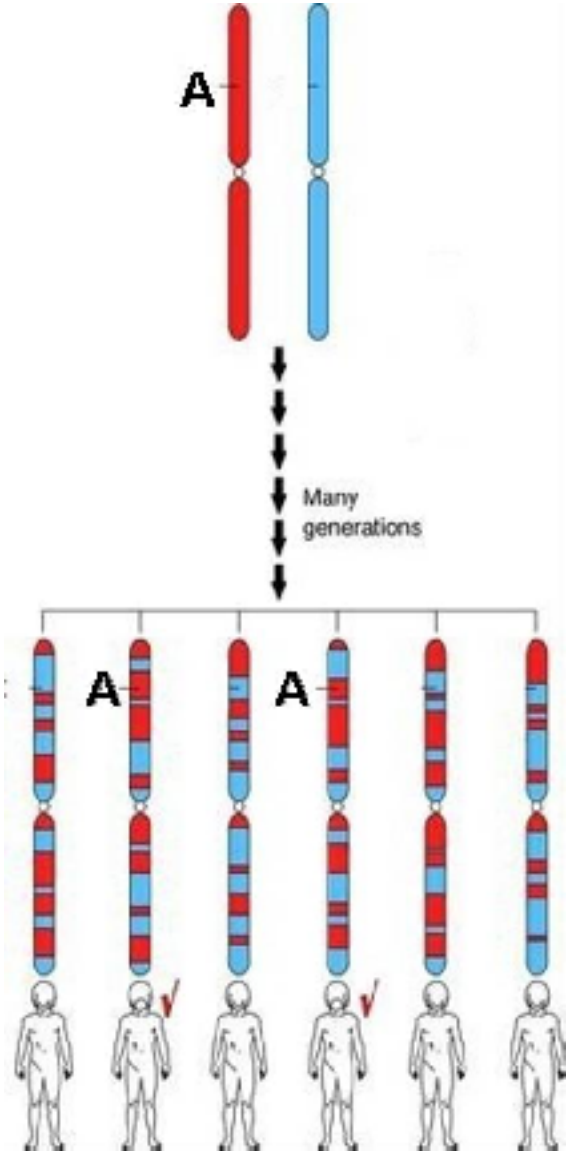
1. Haplotyp
   - ”haploider Genotyp”
   - Eine Rekombination wird nur selten zwei Loci trennen, die nahe beieinander auf einem Chromosom liegen
   - Deshalb werden Gruppen von Allelen, die auf demselben Chromosomenabschnitt liegen eher zusammen (als durch Rekombination getrennt) als Block übertragen werden
2. Genotyp
   - die (diploide) genetische Ausstattung eines Individuums an einem oder mehreren Loci. Beide Exemplare eines Allels werden berücksichtigt, z.B. der Genotyp an einem Locus mit Allelen `A` und `a` kann `AA`, `Aa` oder `aa` sein.

Haplotyp

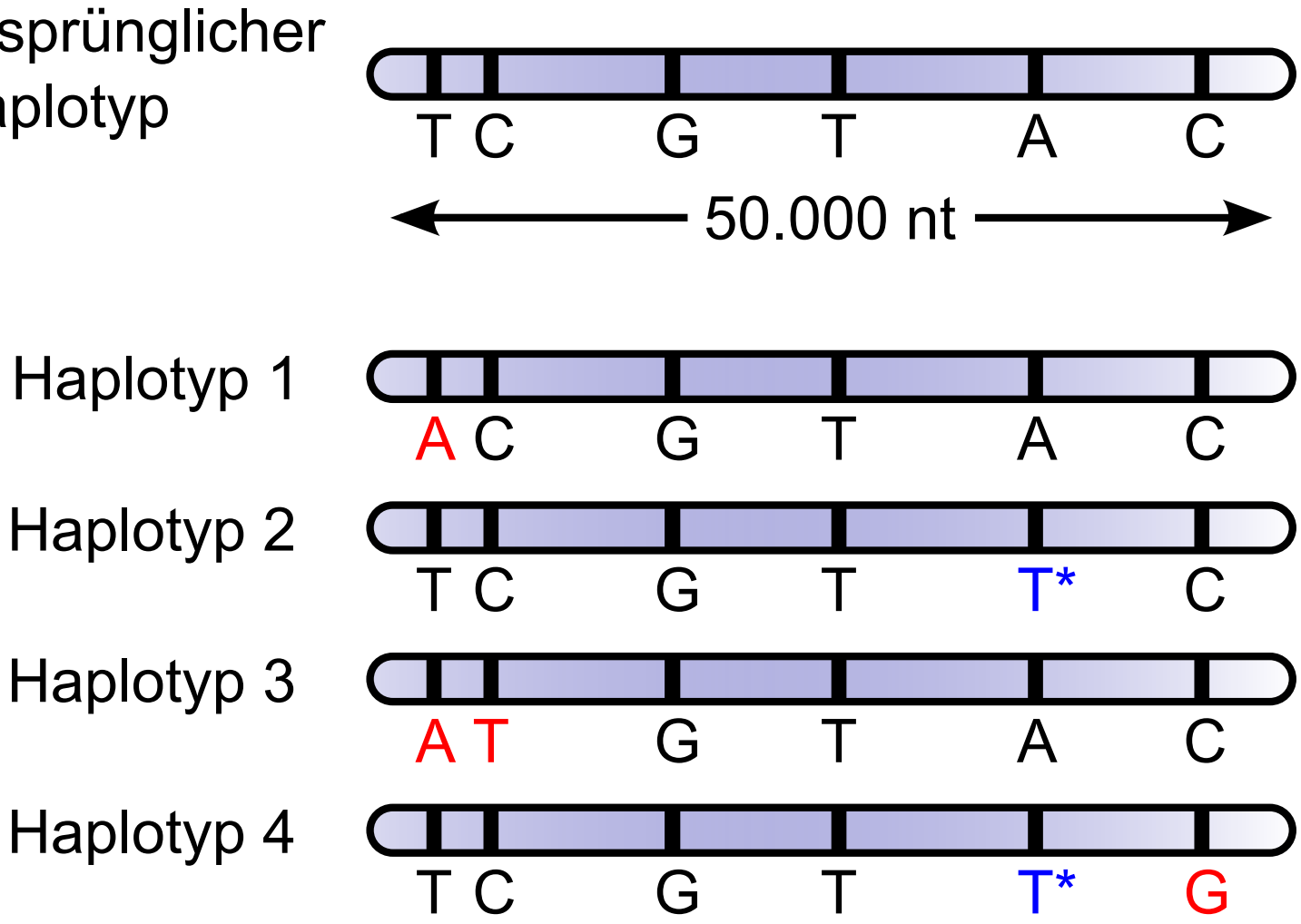

Ursprünglicher Haplotyp
T C G T A C
50.000 nt
Haplotyp 1
A C G T A C
Haplotyp 2
T C G T T C
Haplotyp 3
A T G T A C
Haplotyp 4
T C G T T G

Haplotyp & Krankheitsassoziation

- Variante A tritt als Mutation auf einem bestimmten Haplotyp auf
- Rekombination über viele Generationen erzeugt unterschiedliche Chromosomen
- Falls die Variante A im ursprünglichen Chromosom das Risiko für eine bestimmte Erkrankung erhöht, haben die Personen, welche dieses Allel geerbt haben, demnach ein erhöhtes Risiko



Haplotyp



- Mit Erkrankung assoziiert: Allel T*
- T* ist an andere SNPs auf diesem Haplotyp (z.B. das T an Position 1) physikalisch gekoppelt

Kopplungsungleichgewicht

- Ein Kopplungsungleichgewicht (**LD**) besteht zwischen zwei Genloci, die auf einem Chromosom eng beieinander liegen, und deshalb zusammenvererbt werden.
- Zwei eng beieinander liegende Loci werden dann nicht zusammen vererbt, wenn zwischen ihnen eine Rekombination erfolgt.
- Begriffe
  - Haplotypfrequenz
  - D, $D^{'}$, r^2

Outline

LD: D

- `A` und `a`: zwei Allele von einem Locus (Allelfrequenz p_A und p_a)
- `B` und `b` zwei Allele eines anderen Locus. (Allelfrequenz p_B und p_b)
- Häufigkeiten von Kombinationen dieser Allele innerhalb einer Population von Gameten[†]: p_{AB}, p_{Ab}, p_{aB} und p_{ab}.

[†] zur Erinnerung sind Gameten Keimzellen, d.h. haploide Zellen, die im Gegensatz zu diploiden Zellen jeweils nur ein Exemplar jedes Locus haben.

LD: D

- Die entsprechenden Allelfrequenzen ergeben sich aus der Summe der Genotypfrequenzen:

$$p_a = p_{ab} + p_{aB} \qquad p_A = p_{Ab} + p_{AB} = 1 - p_a$$

und

$$p_b = p_{ab} + p_{Ab} \qquad p_B = p_{AB} + p_{aB} = 1 - p_b$$

Kopplungsgleichgewicht

- Sind die beiden Genloci untereinander im Kopplungsgleichgewicht[†], dann ist die Wahrscheinlichkeit, dass ein Gamet das Allel `a` aufweist, unabhängig von der Wahrscheinlichkeit, dass er das Allel `b` aufweist

$$\boxed{p_{ab} = p_a \times p_b}$$

[†] zum Beispiel weil die Genloci auf unterschiedlichen Chromosomen gelegen sind

Kopplungsungleichgewicht

- Sind die Loci nicht im Kopplungsgleichgewicht, dann gilt

$$p_{ab} \neq p_a \times p_b$$

.

- Wir führen die Variable D ein, um die **Abweichung** vom Kopplungsgleichgewicht zu beschreiben:

$$p_{ab} = p_a p_b + D \tag{1}$$

LD

Hieraus folgt

$$\begin{aligned} p_{aB} &= p_a - p_{ab} \\ &= p_a - p_a p_b - D \\ &= p_a(1 - p_b) - D \\ &= p_a p_B - D \end{aligned}$$

Eine analoge Berechnung zeigt:

$$p_{Ab} = p_A p_b - D$$

und[†]

$$p_{AB} = p_A p_B + D$$

[†] s. Skript

LD

Die am häufigsten verwendete Definition der LD-Koeffiziente D ist jedoch

$$\boxed{D = p_{AB}p_{ab} - p_{Ab}p_{aB}} \tag{2}$$

LD

Diese Formel leitet sich von der Definition (1) ab:

$$
\begin{aligned}
D &= p_{ab} - p_a p_b \\
&= p_{ab} - (p_{aB} + p_{ab})(p_{Ab} + p_{ab}) \\
&= p_{ab} - p_{aB}p_{Ab} - p_{ab}p_{ab} - p_{aB}p_{ab} - p_{Ab}p_{ab} \\
&= p_{ab}(1 - p_{ab} - p_{aB} - p_{Ab}) - p_{aB}p_{Ab} \\
&= p_{ab}(p_{AB}) - p_{aB}p_{Ab} \\
&= \boxed{p_{AB}p_{ab} - p_{Ab}p_{aB}}
\end{aligned}
$$

D und Allelfrequenzen

Die Abbildung zeigt den Einfluss von unterschiedlichen Allelfrequenzen auf die Spannweite von D.

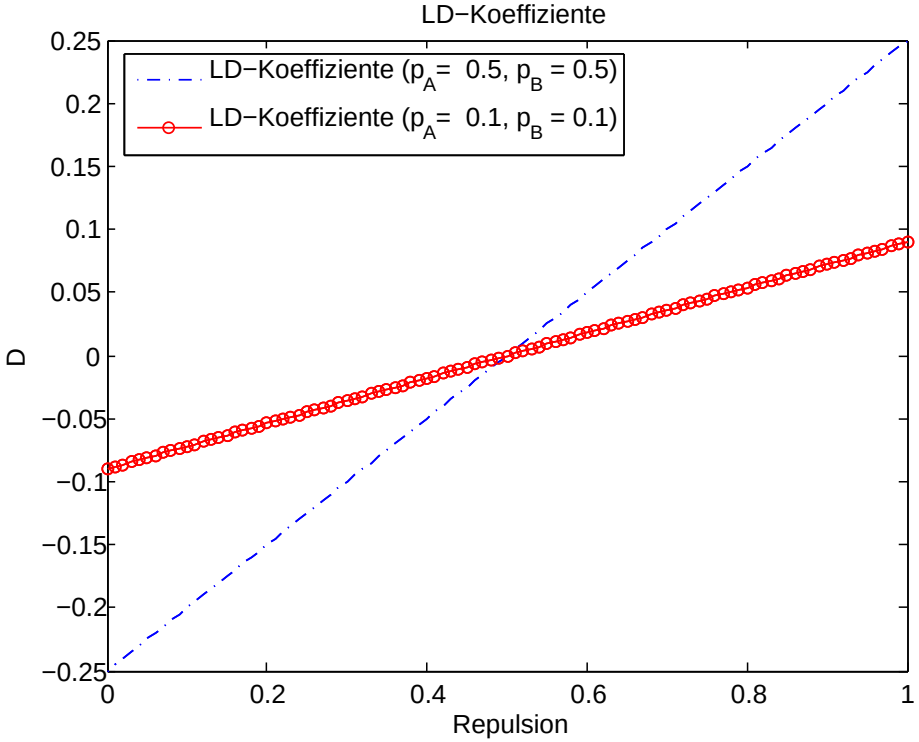
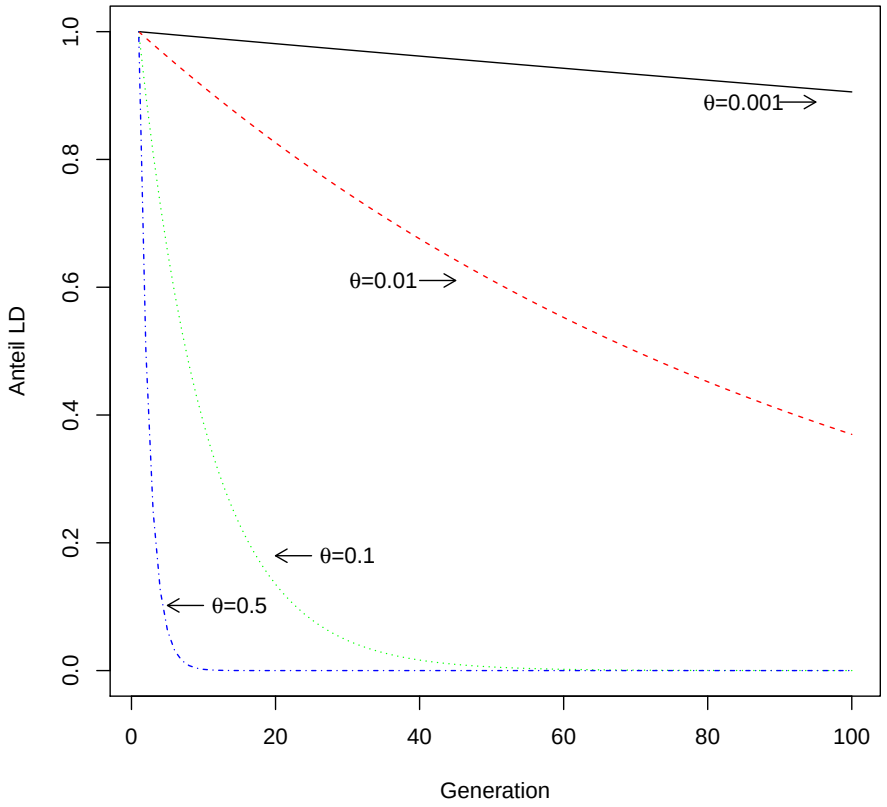


Abbildung: Abhängigkeit der LD-Koeffiziente D von den Allelfrequenzen und vom Grad an Repulsion. 0 = komplette Kopplung von `AB` bzw. `ab`, 1,0 = komplette Repulsion (Überschuss an `Ab` und `aB`).

Wie lange dauert es, bis ein Kopplungsungleichgewicht verschwunden ist?



$$D^i = (1-\theta)^i D^0$$

R

```
generations <- 1:100  ## 1,2,...,100
theta <- c(0.001,0.01,0.1,0.5)
n <- 100 ## Reihen
m <- 4   ## Spalten
D <- matrix(data=rep(0,n*m),nrow=n,ncol=m)
D[1,] = rep(1,4)  ## erst Reihe mit "1" intitialisieren

for (k in 2:n) {
  D[k,] = D[k-1,] * (1-theta) ## elementweise Multiplikation
}
```

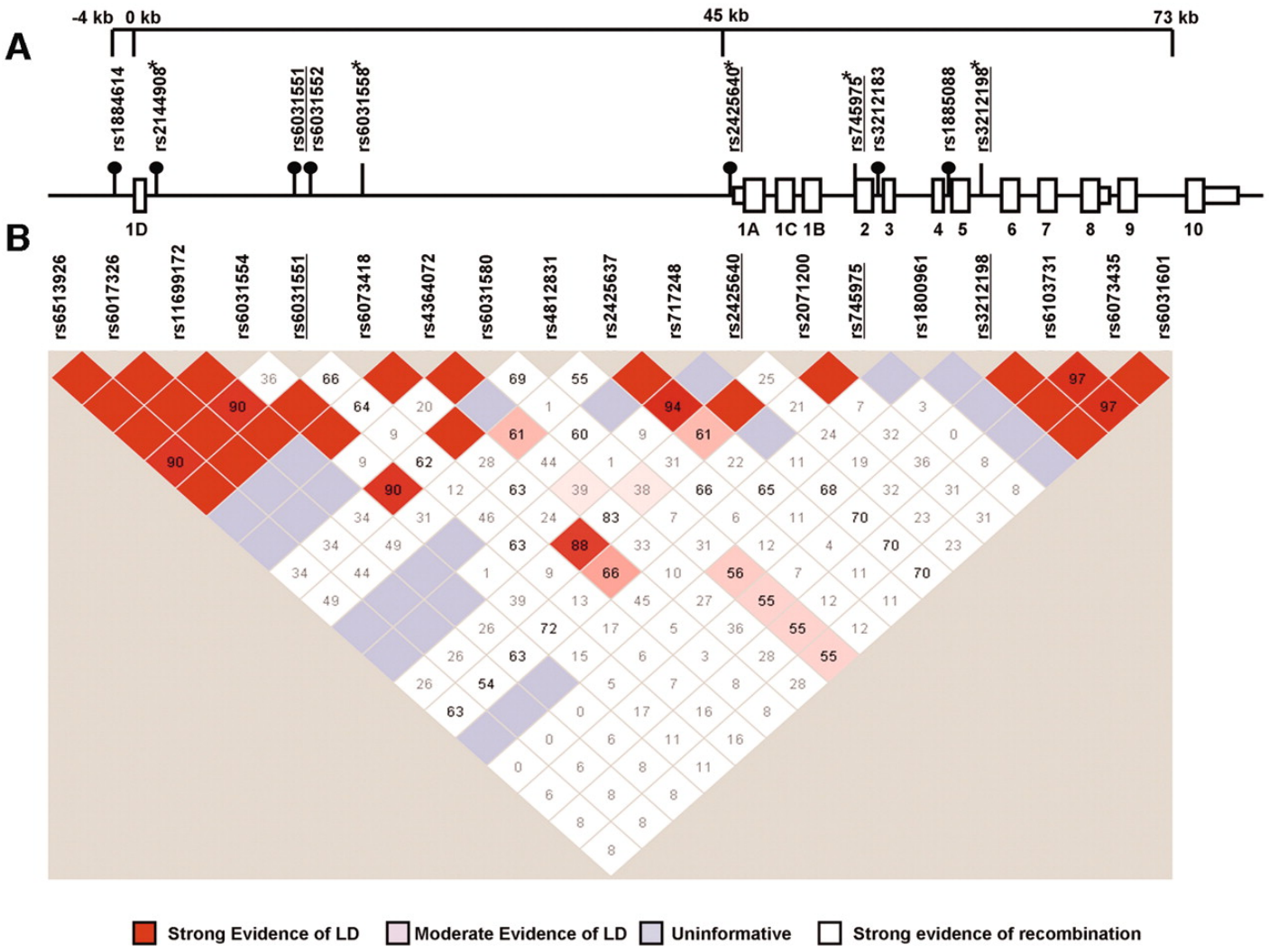
- `rep(x,y)` Element “x” y-mal wiederholen
- `matrix(data=rep(0,n*m),nrow=n,ncol=m)` initialisiere $n \times m$ Matrix, setze alle Elemente auf 0.
- `D[1,]` die erste Reihe von `D`

R

```
plot(generations,D[,1],type='l',lty=1,col='black',
     ylim=c(0,1),
     ylab="Anteil LD",
     xlab="Generation")
lines(generations,D[,2],type='l',lty=2,col='red')
lines(generations,D[,3],type='l',lty=3,col='green')
lines(generations,D[,4],type='l',lty=4,col='blue')
```

- S. `?arrows` und `?text` für Hilfe zur Platzierung von Text und Pfeilen in Plots

LD-Karte



- HNF4A-Genstruktur und LD-Karte
- • Mit Diabetes mellitus Typ 2 assoziierte SNPs

Outline

Beispiel: Normalisierende Selektion

- Als etwas ausführlicheres Beispiel des Einflusses des LD wollen wir die normalisierende Selektion untersuchen.
- Die natürliche Auslese wirkt häufig gegen Individuen an den Extremen des phänotypischen Spektrums und begünstigt Ausprägungen eines phänotypisches Merkmals, die dem Mittelwert des Merkmals in der Bevölkerung nahe sind. Dieses Phänomen ist zunächst Hermon Bumpus 1898 aufgefallen[†]

[†] Bumpus, Hermon C. 1898. Eleventh lecture. The elimination of the unfit as illustrated by the introduced sparrow, Passer domesticus. (A fourth contribution to the study of variation.) Biol. Lectures: Woods Hole Marine Biological Laboratory, 209-225.

Beispiel: Normalisierende Selektion

- Nach einem schweren Wintersturm wurden 136 Hausschwalben untersucht, wovon die Hälfte überlebte
- Unter den überlebenden fand sich ein Überschuss an Vögeln mit durchschnittlichen Maßen hinblicklich Flügellänge, während Vögel mit kurzen oder langen Flügeln öfter als erwartet gestorben waren.

Normalisierende Selektion: Eine Simulation

Phänotyp (Länge in cm)	8	9	10	11	12
Genotypen	$\frac{ab}{ab}$	$\frac{aB}{ab}$ $\frac{Ab}{ab}$	$\frac{aB}{aB}$ $\frac{aB}{Ab}$ $\frac{Ab}{Ab}$ $\frac{AB}{ab}$	$\frac{aB}{AB}$ $\frac{AB}{Ab}$	$\frac{AB}{AB}$
Fitness	0,8	0,9	1,0	0,9	0,8

- a bzw. b: +2 cm
- A bzw. B: +3 cm

Normalisierende Selektion: Eine Simulation

- Anfangs Kopplungsgleichgewicht mit $p(a) = 0,55$, $p(A) = 0,45$, $p(b) = 0,6$ und $p(B) = 0,4$
- Ohne LD gilt $p(ab) = p(a)p(b)$ usw.
- Die diploiden Genotypfrequenzen für erwachsenen Schwalben können aus den haploiden Genotypfrequenzen der Gameten wie folgt berechnet werden[†]:

$$p\left(\frac{ab}{ab}\right) = p(ab)p(ab) \qquad p\left(\frac{aB}{ab}\right) = 2p(aB)p(ab)$$

[†] Der Faktor 2 kommt daher, dass der Gamet *aB* von der Mutter und der Gamet *ab* vom Vater kommen kann oder auch umgekehrt.

Normalisierende Selektion: Eine Simulation

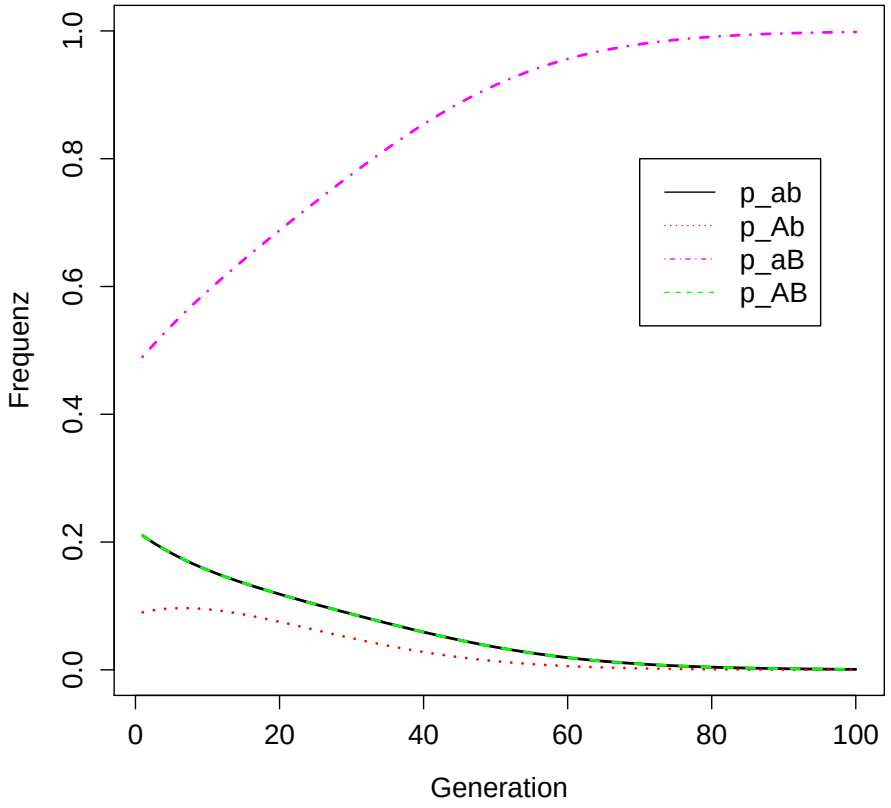


Abbildung: Normalisierende Selektion

- Im folgenden wird der R-Code erklärt, womit die Simulation durchgeführt wurde

R-Code

Unter der Annahme eines Kopplungsgleichgewichts gilt $p(ab) = p(a)p(b)$ usw.

```
p_a <- 0.3
p_b <- 0.7
ngenerations <- 100

p_A <- 1-p_a
p_B <- 1-p_b

# Am Anfang: Kopplungsgleichgewicht: p(ab)=p(a)p(b) usw.
p_ab <- p_a*p_b
p_aB <- p_a*p_B
p_Ab <- p_A*p_b
p_AB <- p_A*p_B
```

R-Code

Der nächste Code-Abschnitt definiert die Vektoren d als 100×1-Vektor und freq als 100×4-Matrix. Diese Variablen werden für Generationen $1 \ldots 100$ die Werte für die LD-Koeffiziente D und die Frequenzen der vier Genotypen festhalten.

```
d <-  vector(mode="numeric",ngenerations)
freq <-  matrix(data=rep(0,ngenerations*4),
    nrow=ngenerations,ncol=4)
```

R-Code

Im folgenden berechnen wir für die erste Generation D :

$$D = p(AB)p(ab) - p(Ab)p(aB)$$

und speichern das Ergebnis im ersten Feld von d. Wir speichern die Genotypfrequenzen der ersten Generation in der ersten Reihe von freq.

```
D <- p_AB*p_ab - p_Ab*p_aB
d[1] <- D
freq[1,] <- c(p_ab,p_aB,p_Ab,p_AB)
```

R-Code

- Ab jetzt simulieren wir eine normalisierende Selektion mit der Funktion gtypesel

```
## Ergebnis fuer die uebrigen Generationen berechnen
P <- vector(mode="numeric",4)
for (i in 2:ngenerations) {
    ## Calculate and store genotype frequencies
    #c(p_ab,p_aB,p_Ab,p_AB)
    P <- gtypesel(p_ab,p_aB,p_Ab,p_AB);
    #print(P)
    freq[i,] <- P # c(p_ab,p_aB,p_Ab,p_AB)
    p_ab <- P[1]
    p_aB <- P[2]
    p_Ab <- P[3]
    p_AB <- P[4]
    ## Calculate and store LD
    d[i] <- p_AB*p_ab - p_Ab*p_aB
  }
```

R-Code

- gtypesel

```
gtypesel <- function(p_ab,p_aB,p_Ab,p_AB) {
  ## Rekombinationsfrequenz 0.1
  theta <- 0.1;

  ##  Selektion auf Grund des Phaenotyps
  ##  8-9-10-11-12 cm Fluegellaenge
  fitness_8 <- 0.8
  fitness_9 <- 0.9
  fitness_10 <- 1.0
  fitness_11 <- 0.9
  fitness_12 <- 0.8
...
```

R-Code

- gtypesel

```
# phenotype = 8 cm
 p_ab_ab <- p_ab^2 * fitness_8
 # phenotype = 9 cm
 p_ab_aB <- 2*p_ab*p_aB * fitness_9
 p_ab_Ab <-  2*p_ab*p_Ab * fitness_9
 #phenotype = 10 cm
 p_Ab_aB <- 2* p_Ab * p_aB * fitness_10
 p_AB_ab <- 2*p_AB*p_ab * fitness_10
 p_Ab_Ab <- p_Ab^2 * fitness_10
 p_aB_aB <- p_aB^2 * fitness_10

... (usw.)
```

R-Code

Die Summe der einzelnen Häufigkeiten muss 1 ergeben, weshalb wir renormalisieren müssen:

```
## Renormalize
  total <- p_ab_ab + p_ab_aB + p_ab_Ab + p_Ab_aB \
                   + p_AB_ab + p_Ab_Ab  + p_aB_aB  \
                   + p_Ab_AB + p_aB_AB + p_AB_AB
  p_ab_ab <- p_ab_ab / total
  p_ab_aB <- p_ab_aB / total
  p_ab_Ab <- p_ab_Ab / total
... (usw.)
```

Rekombination & Gameten

Einige, aber nicht alle Rekombinationen führen zu neuen Haplotypen[1]:

$$\text{Genotyp} \rightarrowtail \text{Gameten}$$

$$\boxed{AB/AB} \overset{\theta}{\rightarrowtail} \boxed{AB} \,\&\, \boxed{AB}$$

$$\boxed{ab/Ab} \overset{\theta}{\rightarrowtail} \boxed{Ab} \,\&\, \boxed{ab}$$

$$\boxed{aB/ab} \overset{\theta}{\rightarrowtail} \boxed{ab} \,\&\, \boxed{aB}$$

$$\boxed{aB/Ab} \overset{\theta}{\rightarrowtail} \boxed{ab} \,\&\, \boxed{AB}$$

$$\boxed{ab/ab} \overset{\theta}{\rightarrowtail} \boxed{ab} \,\&\, \boxed{ab}$$

[1] Bemerke, dass wir der Einfachkeit halber die Rekombination so modellieren, dass bei einer Rekombination alle Chromatiden rekombinieren und nicht nur zwei der vier Chromatiden (vgl. Abb. 2.11 von Strachan und Read)

R-Code

Wir können nun die Frequenz des Haplotyps `ab` unter den Gameten berechnen als

$$
\begin{aligned}
p(ab) &= p\left(\frac{ab}{ab}\right) && \bullet \text{ Jeder Gamet: } \texttt{ab} \\
&+ 0.5 \times p\left(\frac{ab}{aB}\right) && \bullet \text{ Jeder 2. Gamet: } \texttt{ab} \\
&+ 0.5 \times p\left(\frac{ab}{Ab}\right) && \bullet \text{ Jeder 2. Gamet: } \texttt{ab} \\
&+ (1-\theta) \times 0.5 \times p\left(\frac{AB}{ab}\right) && \bullet \text{ Jeder 2. nicht rek. Gamet: } \texttt{ab} \\
&+ \theta \times 0.5 \times p\left(\frac{Ab}{aB}\right) && \bullet \text{ Jeder 2. rek. Gamet: } \texttt{ab}
\end{aligned}
$$

Die Berechnungen für die übrigen drei Gametengenotypen erfolgen analog.

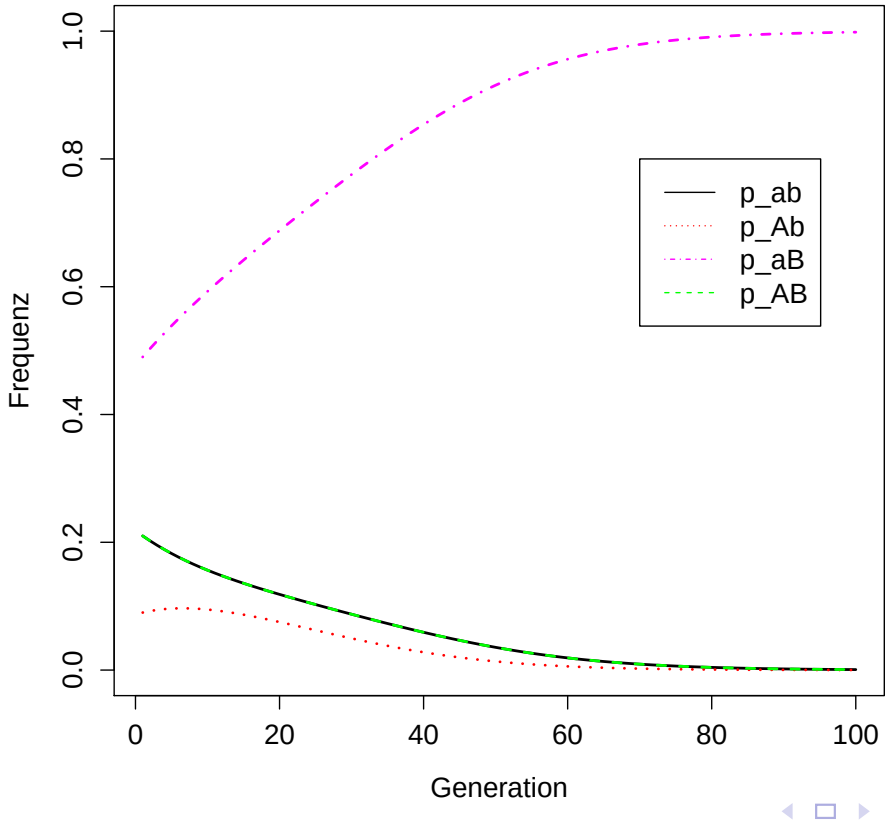
matlab/octave-Code

```
p_ab <- p_ab_ab + 0.5 * p_ab_aB +  \
     0.5 * p_ab_Ab + (1-theta) * 0.5 *  p_AB_ab + \
         theta * 0.5 * p_Ab_aB

p_aB <- 0.5 * p_ab_aB + 0.5 * (1-theta)* p_Ab_aB + \
       p_aB_aB  + 0.5*p_aB_AB +
... (usw.)
```

R-Code

- Nach 90 Generationen hat fast jedes Individuum den Genotyp `Ab/Ab` und somit den günstigsten Phänotyp (Flügellänge 10 cm).
- *D* steigt anfangs und sinkt mit zunehmender Fixation der Allele `A` und `b`.



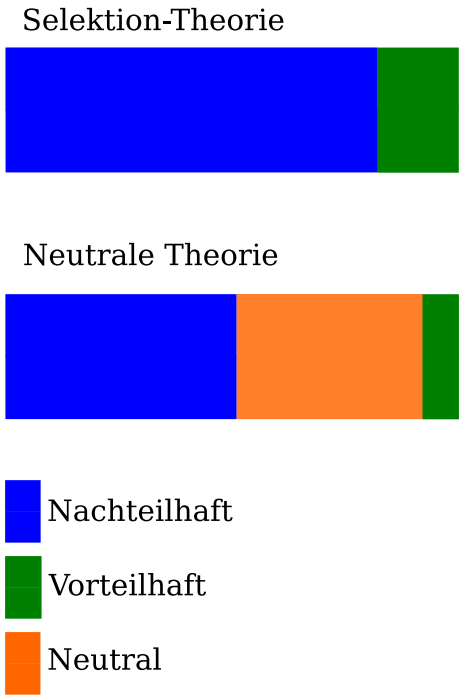
Outline

Die neutrale Theorie der molekularen Evolution

- Die Neutrale Theorie der molekularen Evolution bzw. die verwandte Idee einer *molekularen Uhr* wurden in den 1960er–1980er Jahren von Motoo Kimura eingeführt
- Die Evolutionsrate der Aminosäuresequenzen bestimmter Proteine weist über lange evolutionäre Zeiträume eine konstante Rate auf, was sich als Folge der Genetischen Drift erklären lässt.

Die neutrale Theorie der molekularen Evolution

- Frühe Darwinistische Theorien gingen davon aus, dass alle Sequenzveränderungen einen Einfluss auf die Fitness haben und somit vorteilhaft oder nachteilhaft sind
- Nach der neutralen Theorie der molekularen Evolution sind die meisten Aminosäurenpositionen neutral, Veränderungen haben keinen wesentlich Einfluss auf die Fitness



Die neutrale Theorie der molekularen Evolution

- Vorteilhafte Mutationen: Relativ selten
- Nachteilhafte Mutation dagegen werden durch die natürliche Auslese schnell vom Genpool entfernt
- Ein relativ großer Anteil der denkbaren Veränderungen der Aminosäuresequenz eines Proteins hat keinen wesentlichen Effekt auf die Funktion des Proteins
- Die Anhäufung (Akkumulation) dieser Mutation hängt demnach von der Mutationsrate ab

Die neutrale Theorie der molekularen Evolution

μ^0

Die Mutationsrate μ ergibt sich aus die Rate für nachteilhafte (μ^-), vorteilhafte (μ^+) und neutrale (μ^0) Mutationen. Da vorteilhafte Mutationen selten sind und nachteilhafte durch Selektion schnell aus der Population entfernt werden, fokussieren wir uns auf μ^0.

Sei N_e die **effektive Populationsgröße**.

Für eine Population einer haploiden Spezies beträgt die Anzahl Mutationen pro Generation $N_e\mu^0$.

Es kann gezeigt werden, dass die Wahrscheinlichkeit, dass eine neutrale Mutation durch genetische Drift fixiert wird, $1/N_e$ beträgt.

Daher beträgt die Anzahl von neutralen Mutationen, die pro Generation fixiert werden $\frac{N_e\mu^0}{N_e} = \mu^0$

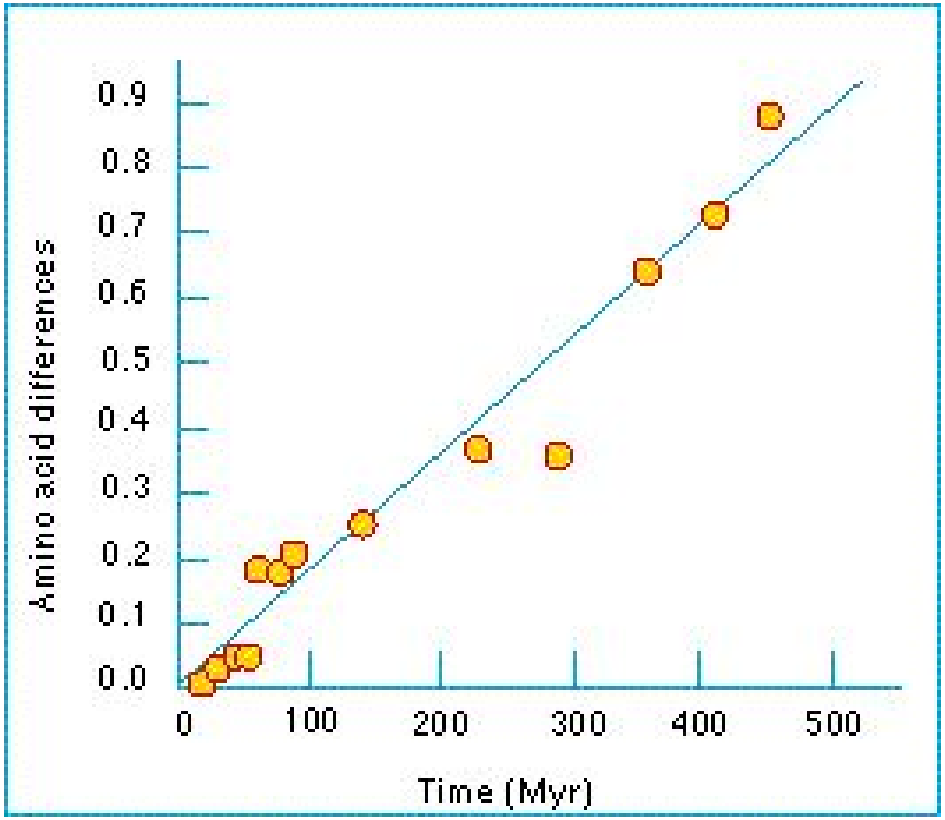
Die neutrale Theorie der molekularen Evolution

- Intuitiv: Obwohl in einer größeren Population mehr Mutationen auftreten, die Wahrscheinlichkeit dass eine spezifische Mutation in der Population fixiert wird sinkt proportional zur Populationsgröße
- Nach dem neutralen Modell bestimmt daher die Mutationsrate μ^0 die molekulare Evolutionsgeschwindigkeit unabhängig von der Populationsgröße
- Dies ist ein wichtiges Ergebnis (Vorhersage):

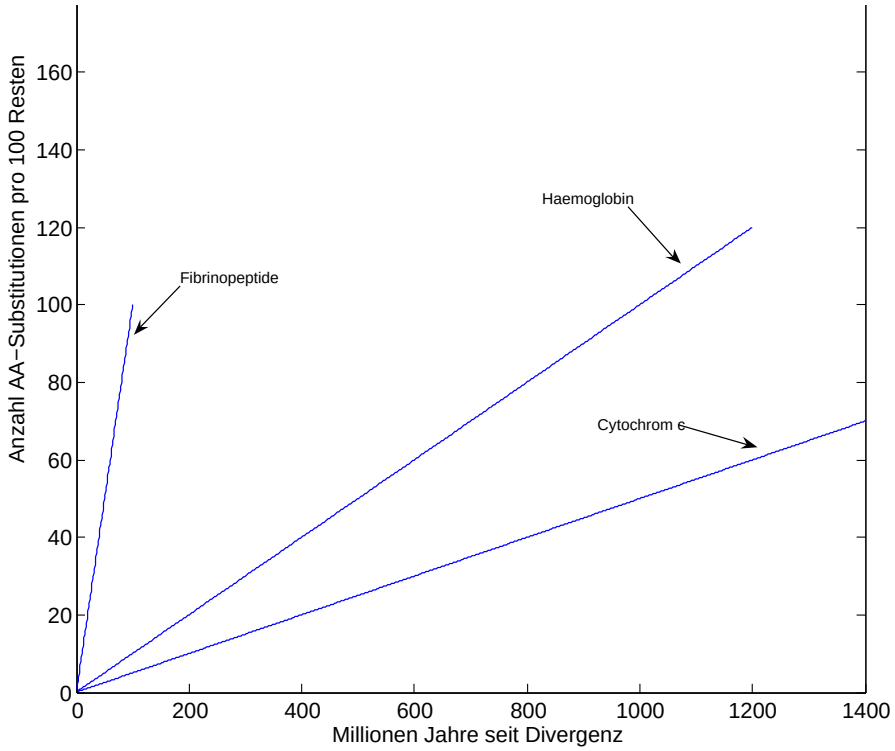
Die molekulare Evolutionsrate in einer Spezies ist dieselbe wie die neutrale Mutationsrate in Individuen[a]

[a] Merke dass die Mutationsrate und auch die Evolutionsrate sich für unterschiedliche Proteine unterscheiden.

Die evolutionäre Zeit mit der molekularen Uhr messen



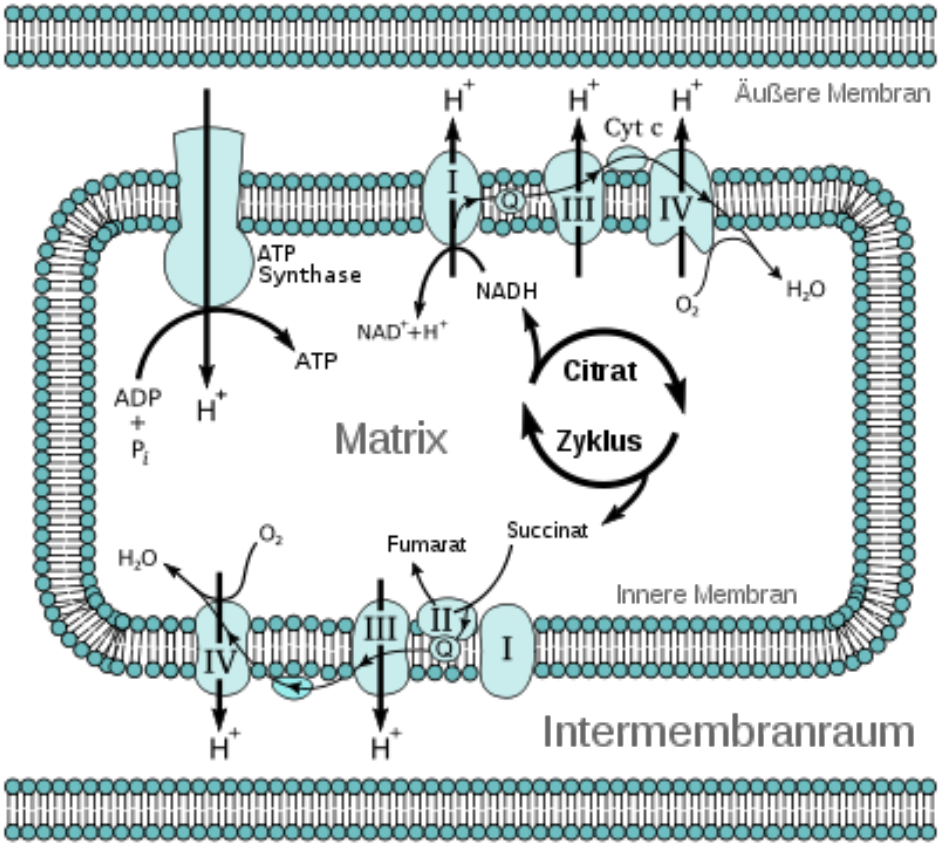
Die evolutionäre Zeit mit der molekularen Uhr messen



- Die evolutionäre Rate ist unterschiedlich für unterschiedliche Proteine (unterschiedlicher Anteil an neutralen Resten, andere Faktoren)

Cytochrome c

```
            20          30          40          50
Human       TVEKGGKHKTGPNLHGLFGRKTGQApGySYTaA
Chicken     TVEKGGKHKTGPNLHGLFGRKTGQAeGFSYTDA
Cow         TVEKGGKHKTGPNLHGLFGRKTGQApGFSYTDA
Dog         TVEKGGKHKTGPNLHGLFGRKTGQApGFSYTDA
Mouse       TVEKGGKHKTGPNLHGLFGRKTGQAaGFSYTDA
Rat         TVEKGGKHKTGPNLHGLFGRKTGQAaGFSYTDA
Mosquito    TVEaGGKHKvGPNLHGiFGRKTGQAaGFSYTDA
Fruitfly    TVEaGGKHKvGPNLHGLiGRKTGQAaGFaYTDA
Worm        vVds.tatKTGPtLHGviGRtsGtvsGFdYsaA
Honeybee    TiEsGGKHKvGPNLyGvyGRKTGQApGySYTDA
consensus   ********!*!!*!*!**!!**!** !**!**!
```



- Wesentlicher Bestandteil der Elektronentransportkette
- Extrem hoch konserviert

Welche sind die wichtigen Positionen in einem multiplen Alignment?

```
                                   *   .     :     .     .   *       : : :     .
Q5E940_BOVIN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_HUMAN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_MOUSE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
    RLA0_RAT -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_CHICK -----------MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_RANSY -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE    76
Q7ZUG3_BRARE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQTIRLSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_ICTPU -----------MPREDRATWKSNYFLKIIQLLNDYPKCFIVGADNVGSKQMQTIRLSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_DROME -----------MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGL-AVVLMGKNTMMRKAIRGHLENN--PQLE    76
  RLA0_DICDI -----------MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD    75
Q54LP0_DICDI -----------MSGAG-SKRKNVFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD    75
  RLA0_PLAF8 -----------MAKLSKQQKKQMYIEKLSSLIQQYSKILIVHVDNVGSNQMASVRKSLRGK-ATILMGKNTRIRTALKKNLQAV--PQIE    76
  RLA0_SULAC -----MIGLAVTTTKKIAKWKVDEVAELTEKLKTHKTIIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNNLFNIALKNAG-----YDTK    79
  RLA0_SULTO ----MRIMAVITQERKIAKWKIEEVKELEQKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG-----LDVS    80
  RLA0_SULSO ----MKRLALALKQRKVASWKLEEVKELTELIKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFKIAAKNAG-----IDIE    80
  RLA0_AERPE MSVVSLVGQMYKREKPIPEWKTLMLRELEELFSKHRVVLFADLTGTPTFVVQRVRKKLWKK-YPMMVAKKRIILRAMKAAGLE---LDDN    86
  RLA0_PYRAE -MMLAIGKRRYVRTRQYPARKVKIVSEATELLQKYPYVFLFDLHGLSSRILHEYRYRLRRY-GVIKIIKPTLFKIAFTKVYGG---IPAE    85
  RLA0_METAC ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVGIEGILATKMQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ETIP    78
  RLA0_METMA ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVRIEGILATKIQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ESIP    78
  RLA0_ARCFU ------MAAVRGS---PPEYKVRAVEEIKRMISSKPVVAIVSFRNVPAGQMQKIRREFRGK-AEIKVVKNTLLERALDALG-----GDYL    75
  RLA0_METKA MAVKAKGQPPSGYEPKVAEWKRREVKELKELMDEYENVGLVDLEGIPAPQLQEIRAKLRERDTIIRMSRNTLMRIALEEKLDER--PELE    88
  RLA0_METTH -------------MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPARQLQKMRQTLRDS-ALIRMSKKTLISLALEKAGREL--ENVD    74
  RLA0_METTL -------MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPARQLQEIRDKIR-GTMTLKMSRNTLIERAIKEVAEETGNPEFA    82
  RLA0_METVA -------MIDAKSEHKIAPWKIEEVNALKELLKSANVIALIDMMEVPAVQLQEIRDKIR-DQMTLKMSRNTLIKRAVEEVAEETGNPEFA    82
  RLA0_METJA --------METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPQLQEIRDKIR-DKVKLRMSRNTLIIRALKEAAEELNNPKLA    81
  RLA0_PYRAB -------------MAHVAEWKKKEVEELANLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAQELGKPELE    77
  RLA0_PYRHO -------------MAHVAEWKKKEVEELAKLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAKELGKPELE    77
  RLA0_PYRFU -------------MAHVAEWKKKEVEELANLIKSYPVVALVDVSSMPAYPLSQMRRLIRENNGLLRVSRNTLIELAIKKVAQELGKPELE    77
  RLA0_PYRKO -------------MAHVAEWKKKEVEELANIIKSYPVIALVDVAGVPAYPLSKMRDKLR-GKALLRVSRNTLIELAIKRAAQELGQPELE    76
  RLA0_HALMA -----MSAESERKTETIPEWKQEEVDAIVEMIESYESVGVVNIAGIPSRQLQDMRRDLHGT-AELRVSRNTLLERALDDVD-----DGLE    79
  RLA0_HALVO -----MSESEVRQTEVIPQWKREEVDELVDFIESYESVGVVGVAGIPSRQLQSMRRELHGS-AAVRMSRNTLVNRALDEVN-----DGFE    79
  RLA0_HALSA -----MSAEEQRTTEEVPEWKRQEVAELVDLLETYDSVGVVNVTGIPSKQLQDMRRGLHGQ-AALRMSRNTLLVRALEEAG-----DGLD    79
  RLA0_THEAC -------------MKEVSQQKKELVNEITQRIKASRSVAIVDTAGIRIRQIQDIRGKNRGK-INLKVIKKTLLFKALENLGD----EKLS    72
  RLA0_THEVO -------------MRKINPKKKEIVSELAQDITKSKAVAIVDIKGVRIRQMQDIRAKNRDK-VKIKVVKKTLLFKALDSIND----EKLT    72
  RLA0_PICTO -------------MTEPAQWKIDFVKNLENEINSRKVAAIVSIKGLRNNEFQKIRNSIRDK-ARIKVSRARLLRLAIENTGK----NNIV    72
       ruler 1.......10........20........30........40........50........60........70........80........90
```

Welche sind die wichtigen Positionen in einem multiplen Alignment?

”Neodarwinistische Antwort”

Die Positionen, welche Veränderung aufweisen, sind besonders interessant, weil sie uns zeigen, wo die positive Selektion gewirkt hat

Synthetische Theorie der Evolution (1950–1960

G.G: Simpson (1964)

Der Konsens ist, dass neutrale Gene oder Allele sehr selten sein müssen, falls sie überhaupt existieren. Für einen Evolutionsbiologen erscheint es hochunwahrscheinloch, dass Proteine, die ja durch Gene bestimmt werden, funktionslose Teile haben, dass schlummernde Gene über viele Generationen hinweg existieren oder dass Moleküle sich auf eine regelmäßige aber nicht adaptive Art und Weise verändern sollen. Die natürliche Auslese ist der Komponist des genetischen Codes, und die DNA, RNA und Protein seine Boten

- Nach dieser Ansicht: Unterschiede in Alignments $\rightarrow$ Folge der positiven Selektion

Welche sind die wichtigen Positionen in einem multiplen Alignment?

Antwort nach der Theorie der neutralen molekularen Evolution

Die Positionen, welche (über ausreichend lange evolutionäre Zeiträume) unverändert geblieben sind, stellen die interessantesten Positionen dar, weil sie uns zeigen, wo die negative Selektion gewirkt hat (d.h., Mutationen in diesen Positionen sind nachteilhaft)

Neutralist-Selektionist-Debatte

- Lange Zeit wurde angezweifelt, dass es überhaupt neutrale Mutationen geben kann
- Das erscheint heute klar. Auch Darwin hat die neutrale Theorie der molekularen Evolution vorweggenommen:

Charles Darwin, On the Origin of Species by Means of Natural Selection, 6th ed., 1872

Variations neither useful nor injurious would not be affected by natural selection, and would be left either a fluctuating element, as perhaps we see in certain polymorphic species, or would ultimately become fixed...

Synonyme und nichtsynonyme Substitutionen

Synonyme Substitutionen

Nukleotidsubstitutionen in einem Kodon, welche die kodierte Aminosäure nicht verändern. Zum Beispiel `CTT`=Leucin. `CTT`→`CTA`=Leucin, `CTT`→`CTC`=Leucin und `CTT`→`CTG`=Leucin

Nichtsynonyme Substitutionen

Nukleotidsubstitutionen in einem Kodon, welche die kodierte Aminosäure verändern. Zum Beispiel `CTT`=Leucin. `CTT`→`ATT`=Isoleucin, `CTT`→`GTT`=Valin und `CTT`→`TTT`=Phenylalanin

Proteine vs. DNA

- Proteine sind die moleküle, die biologische Funktionen erfüllen
- Annahme: Die natürliche Auslese wirkt daher auf Proteinebene und viel weniger auf DNA-Ebene
- Schlussfolgerung: Die Mutationsrate für synonyme Substitutionen gibt (ungefähr) die neutrale Mutationsrate an
- Die Mutationsrate für nichtsynonyme Substitutionen variiert dagegen je nach Typ und Stärke der natürlichen Auslese

Proteine vs. DNA

- Sei bei einem Alignment zweier DNA-Sequenzen d_S die Anzahl der synonymen Substitutionen und d_N die Anzahl der nichtsynonymen
- Dann ist $d_N > d_S$ ein Hinweis auf positive Selektion
- $d_N < d_S$ ein Hinweis auf negative Selektion
- Zahlreiche Methoden sind entwickelt worden, um z.B. solche Methoden auf multiple Alignments anzuwenden, um Hinweise auf Neutralität in bestimmten Codons/Abschnitten zu suchen.

Mäuse und Menschen

- Der letzte gemeinsame Vorfahr von Mäusen und Menschen lebte vor ca. 75 Million Jahren
- Die Genomsequenzen dieser Organismen unterscheiden sich heute an ca. jedem zweiten Nukleotid

- Weniger als 1% der ∼22.000 proteinkodierende Gene in der Maus haben kein homologes Gen beim Menschen. Viele Proteinsequenzen zeigen eine Übereinstimmung von über 90%

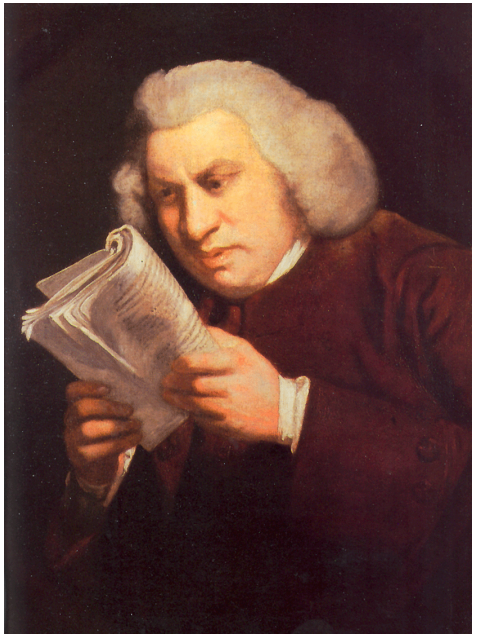
Bei der Betrachtung eines multiplen Alignments...

```
Q8CUN7_OCEIH/23-262        CLLYIICYFFVL........PVTE..SGSMKKAVLTILGVTILFIAIGSPLNIIAR.LTFQGHMIQILLLTVVSAPLLVA
Q8ELC4_OCEIH/23-266        VVVIAIIYASSI......IFLTDV..KVYHRQPILFFLSLSLFYVIMGSPLATISH.LSFSLHMIQMSILYFIVPPLLIT
Q9K8U7_BACHD/22-288        LAMVGFVYFMIA.TKWRERFTNSE..PVPIRKKIYFVLGLIALYIGWGSPFYVAGH.LMITFHMAQMVFAYFIAVPLFLL
Q5WEW6_BACSK/22-279        LVGVAFIYSW........FFRRTN..AGPKRKPLFFFLGLAALYLGWGSPLYVTGH.VMMTLHMLQMVFAYFIAVPLLLL
Q8ER73_OCEIH/22-286        VILLGTAYYLIT.GPLRREFGDND..KPTAKQQSMFYIALLLLYFVKGAPIDLLSH.ITLTAHMIQMAIYLLVFPMLMIK
CTAG_BACSU/21-280          LLGITALYLYFY....RRMSSKPN..RITGKEMVCFLSAMLFLYAAEGSPVDLLGH.IMFSAHMVQMAVLYLVVPPLLIA
Q65K08_BACLD/22-281        VVFITALYFFLM....KRLGSEGE..RASRKEIGLFLTAMILLYASKGSPVDLLGH.IMFSAHMAQMAVLYLIVPPLIIL
Q2B4U5_9BACI/1-263         MLAILAAYFLLT.VKYRQRFSGST..PLSAKQAVLFTAAIILVYAVKGSPVDLMSH.ITFYAHMIQMSILYLIVPPLLIT
Q5L109_GEOKA/22-284        LAAVALLYIGIT.GPWRQRFGLGD..AVSEKQKAYFLTGIALLYICKGSPLDLMGH.LTFTAHMVQMAVLYLIVPQCFIL
Q2ASN4_9BACI/33-295        MISILISYFLII.GPYRTRFENAT..KVSKKQIFYFTTGIVLLYVVKGGPIDLIGH.IIFSAHMFEMAVMYIAVPPLLLL
Q41CW2_9BACI/26-285        LIGIYVLYAVLT...EKIRRPDEA..ETTLGQKFSMLAALFVYYIGFGSPLDVLAH.ITFSAHMLQMVFVYMVAPPLLMI
Q1H1C4_METFK/43-289        VGLALVLYLA...GLYRMTRKIGRPTSDAPMRAKAFLLAWLTLVVALFSPVDTLGN.AYFSMHMVQHELLMIVAAPLFVM
Q01RA2_SOLUE/42-282        LLLTAVLYF........RGASRGR..GVSLKQTFFFWAGWSILCLALLSPLHPLGE.ALFSAHMAQHELLMVAAAPLLVL
Q13IJ4_BURXL/64-313        MLASTLAYAVGYVRLRLRGSPRSR..ATRAWHASAFAAGMAALVFALCSPLDSLSA.ALFSAHMVQHETMMLIAAPLLVL
A0G4F4_9BURK/59-310        MAMSAAAYAAGYWRLRARASPRSRSARVRAVHLIAFVSGWVALALALFSPLDTLSG.ALFSAHMVQHESMMLIAAPLLVA
Q0ML06_9RHIZ/37-286        PLALTALAYAIGMRRLWSASARGQ..TIHLQRAVCFAAGWLFLAAALVSPLDRLAT.QLFTAHMIEHEILMVIAAPLFVL
Q4IX09_AZOVI/37-262        LGAGLLAAAWIGYEGGCRHRPAAR..R....RRALLHGGLLLAALSLFGPLDEAAE.SSAAAHMAQHMLLMLAVAPLLAL
A0HY77_PSEME/39-261        VLLGSAWLLYII..GCRKVRPHGR..E.....ALWMHLAMVITVFAVFGPIDDWAE.TSTSLHMVQHMLFMIVIAPLWAL
Q5P902_AZOSE/147-370       LIAALLGAACGLYGLGARRVPPGR..V....QATWFCAAMAIGALAVFGPLDRWAE.NSTALHMVQHMLLIVVVAPLGAL
Q0LFY3_HERAU/14-252        LIAATVGYLWAV.GPARKRLGGPA..AFPVGKAVAFLSGLLALGLSIMSPIGIWADRYLFTMHMVQHMILTMFCAPMLLI
A0I6Y9_9CHLR/14-253        LALIAGGYALCVTGPLRRFFPGSA..PPTPVQVRLFYVGWVVLFIALASPIDSLAS.YLLTMHMLQHLLLAMVAPPFLLL
Q67MK8_SYMTH/42-281        TVLLNAAYLLLV.NLWRRAFNWGP..PVPVWRQVLFCLGLWTVYLSEGTPIHIMSELYLFSVHMVQHTLLTMVMPPLILL
Q3JQ94_BURP1/20-265        VLVAGVLFA.........RGARKA..KVSASRRVAFWFGLVALYVALHTRLDYFFE.HEFFMHRAQHLVLHHLGPFFIAL
Q6XN36_RHOER/28-280        LPLIGVLLAAWYCWGVYRVTASGR..VWPWTRTASFLFGCILLILVTGLAVEGYGY.ELFSIWMFQHLTLSMAIPPLLVL
Q70K40_9ACTO/355-608       VAAVVGAVGYVV..AARRERASGH..RWPIRRTVSWVVGCAVVVVSTSSGLKAYGN.ALFSVHMAEHTALTMIAPVLLVG
Q73XL2_MYCPA/390-645       FGTAAIVLAGLYVAGVVRLRRRGD..RWPPGRTSSWLLGCLVLLFVTSSGVGRYMP.AMFSMHMVVHMCLSMLVPILLAL
A1TCF0_MYCVP/380-635       FGTAAIVFALVYLAGVRRLRRRGD..AWPIGRVVAWLLGCLVLLLATSSGIGRYMP.AMFSMHMIAHMLLSMLAPILLVL
Q1B5W0_MYCSS/375-630       LGSAAIILALVYLAGWWRLRRRGD..AWPAGRTVAWLLGCATLLITTSSGLGRYMP.AMFSVHMVAHMLLSMLVPILLVL
Q6ABY8_LEIXX/366-618       LACAFALFFYL..AGVWRLRKRGD..RWPVHRTILWTFGIVLLFFVTSGGVNVYEK.YIFLVHMSAHMVLTMAVPLLLVP
```

Wichtigste Voraussetzung: Die Divergenz ist ausreichend hoch, so dass man funktionell bedeutsame Elemente durch ihren hohen Grad an Konservierung erkennen kann

The End of the Lecture as We Know It

- Kontakt: peter.robinson@charite.de
- Vorlesungsskript Kapitel 4, Strachan & Read Kapitel 15.4
- Bromham L, Penny D (2003) The modern molecular clock. Nature Reviews Genet 4:216–224.



Lectures were once useful; but now, when all can read, and books are so numerous, lectures are unnecessary. If your attention fails, and you miss a part of a lecture, it is lost; you cannot go back as you do upon a book... People have nowadays got a strange opinion that everything should be taught by lectures. Now, I cannot see that lectures can do as much good as reading the books from which the lectures are taken. I know nothing that can be best taught by lectures, except where experiments are to be shown. You may teach chymistry by lectures. You might teach making shoes by lectures!

Samuel Johnson, quoted in Boswell's Life of Johnson (1791).