

Polygene Krankheiten und Assoziationsanalyse

Peter N. Robinson

Institut für medizinische Genetik
Charité Universitätsmedizin Berlin

6. Januar 2015

Outline

1 Komplexe und monogene Erkrankungen

2 Kopplungsanalysen komplexer Merkmale

3 Assoziation vs. Kopplung

Kartierung mendelnder Erkrankungen (Rückblick)

- Die Suche nach Krankheitsgenen mit der lod-Wert-Analyse setzt ein definiertes Vererbungsmodell voraus
 - ▶ Autosomal dominant
 - ▶ Autosomal rezessiv
 - ▶ X-chromosomal rezessiv
- Für komplexe Krankheiten haben wir jedoch keine genauen Vorstellungen über den Vererbungsmodus, Allelhäufigkeiten, Penetranz ...

dichotome und kontinuierliche Merkmale

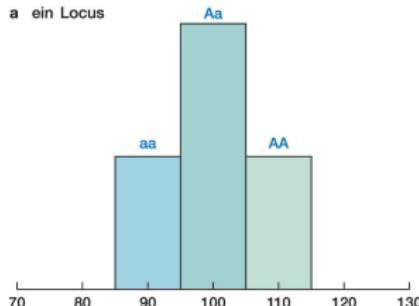
- dichotom: gr. *dicha* = entzwei, *tom* ≈ Schnitt
- Ein "Entweder-Oder"-Merkmal
 - ▶ z.B. Lippenkiefergaumenspalte oder keine Spalte
 - ▶ Viele Mendelsche Erkrankungen lassen sich als dichotome Merkmale verstehen

Kontinuierliche/Quantitative Merkmale

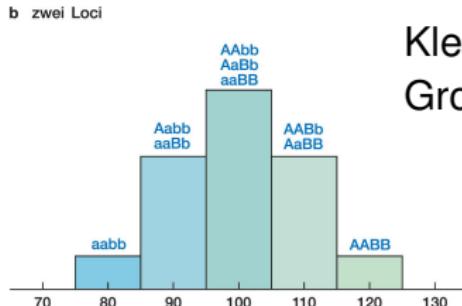
- Kontinuierliche Merkmale werden durch das Zusammenwirken von Allelen an mehreren Genorten (und durch Umweltfaktoren) bedingt
- **polygen**: durch eine große Zahl unabhängiger mendelscher Merkmale bestimmt

Gauß-Verteilung polygener Merkmale

a ein Locus

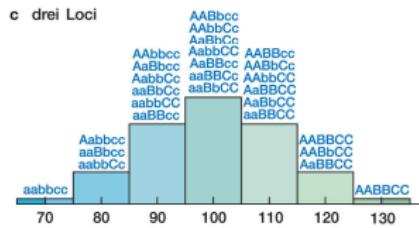


b zwei Loci

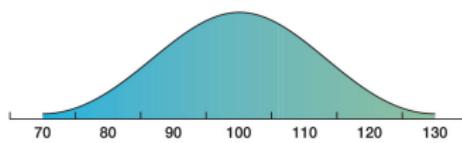


Kleinbuchstabe: -5
Großbuchstabe: +5

c drei Loci



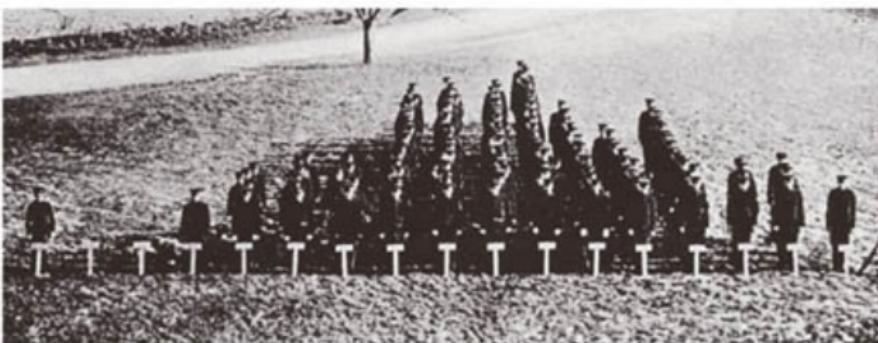
d viele Loci



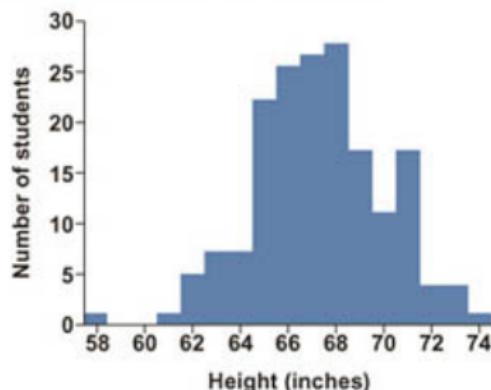
Aus Strachan/Read, Molekulare Humangenetik, 3. Aufl., © 2005 Elsevier GmbH

Körpergröße

A “living histogram”—distribution of height in a college class



Normal distribution = bell curve

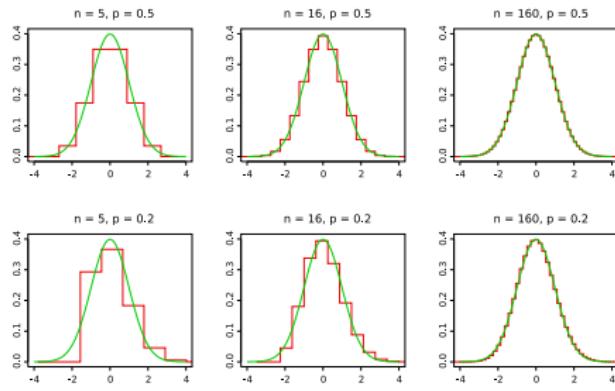


Zentraler Grenzwertsatz

- Unter der Annahme, dass die Ausprägung eines Merkmals das Ergebnis additiver genetischer Effekte ist

$$X = \sum X_i$$

- dann gilt der zentrale Grenzwertsatz (central limit theorem)
die Summe einer großen Zahl von unabhängigen, identisch verteilten Zufallsvariablen ist annähernd normalverteilt

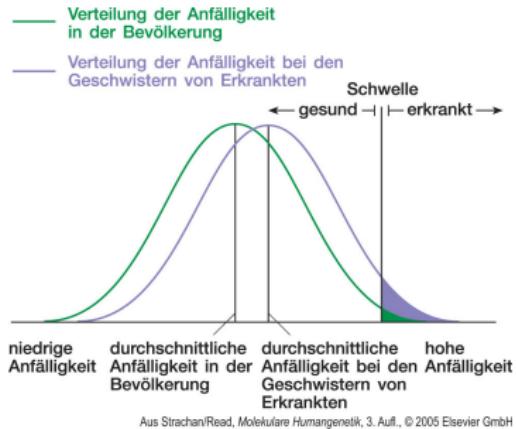


Wikipedia Commons

Schwellenmodell

- Medizinisch interessant sind die Krankheiten, die tendenziell familiär auftreten, ohne ein mendelndes Vererbungsmuster zu zeigen
- Viele solche Krankheiten sind dichotom, z.B. Gaumenspalte
- Anfälligkeit, eine Krankheit zu entwickeln, könnte jedoch einer Gauß'schen Verteilung entsprechen.
- Übersteigt die Anfälligkeit einen bestimmten Schwellenwert, kommt es zur Entwicklung der Krankheit

Polygenes Schwellenmodell



Aus Strachan/Read, Molekulare Humangenetik, 3. Aufl., © 2005 Elsevier GmbH

Abb. 4.14 Falconers polygenes Schwellenmodell für dichotome nichtmendelnde Merkmale. Die Anfälligkeit für die Krankheit ist polygen und normal verteilt (grüne Kurve). Menschen, deren Anfälligkeit über einem bestimmten Schwellenwert liegt (Gleichgewichtspunkt in Abbildung 4.13), erkranken. Ihre Geschwister (violette Kurve) zeigen eine höhere durchschnittliche Anfälligkeit als das Bevölkerungsmittel; außerdem neigt ein größerer Anteil von ihnen dazu, die Schwelle zu überschreiten. Daher kommt die Krankheit familiär gehäuft vor.

Outline

1 Komplexe und monogene Erkrankungen

2 Kopplungsanalysen komplexer Merkmale

3 Assoziation vs. Kopplung

Der λ -Wert ist ein Maß für die Häufung in Familien

- Die erste Entscheidung ist, ob ein nicht mendelndes Merkmal genetisch (mit)bedingt ist
- Dazu muss erst gezeigt werden, dass eine Krankheit in Familien gehäuft auftritt
- Sei K die Prävalenz einer Erkrankung in der Bevölkerung
- Sei K_R die Prävalenz einer Erkrankung bei Verwandten (Relatives) eines erkrankten Probanden
- Der λ -Wert von Risch[†] definiert sich als

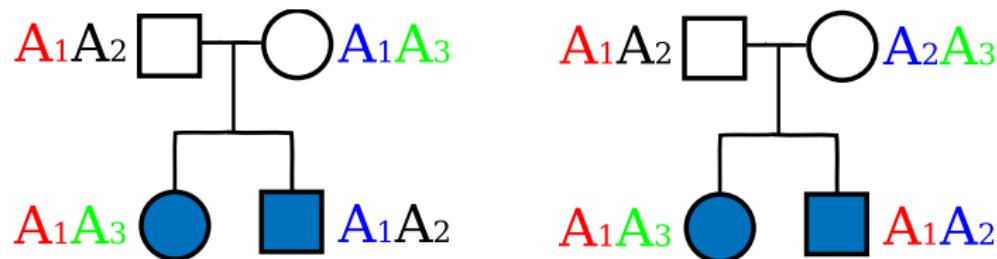
$$\lambda_R = \frac{K_R}{K}$$

[†] Risch N (1987) Assessing the Role of HLA-linked and Unlinked Determinants of Disease Am J Human Genet 40:1-14, 1987

Der λ -Wert ist ein Maß für die Häufung in Familien

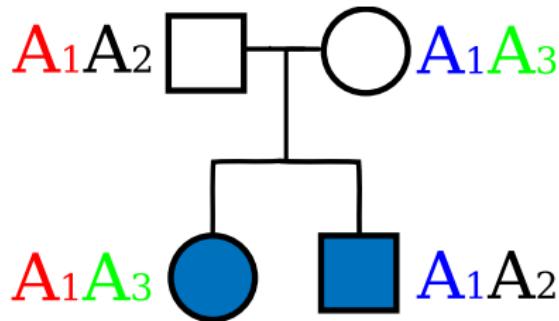
- λ ist für jeden Verwandtschaftsgrad gesondert zu bestimmen
- z.B. beträgt für Geschwister (**Siblings**) eines Probanden mit einer Schizophrenie $\lambda_S = 12,6$
- Unter bestimmten Voraussetzungen nimmt λ_R mit jedem Verwandtschaftsgrad um $\frac{1}{2}$ ab.

Erkrankte Geschwister Paare



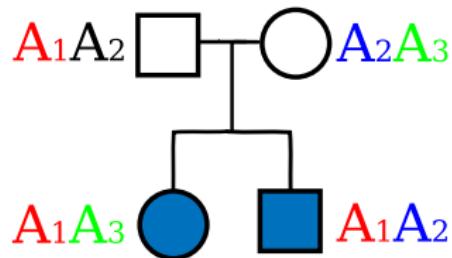
- Affected sib pairs (ASP)-Methode
- Ziel: Auffinden von chromosomalen Segmenten, die betroffene Personen gemeinsam haben (Hinweis auf Anfälligkeitssocus)
 - ▶ Hierbei keine Annahme eines bestimmten Modells wie autosomal dominant
 - ▶ daher "nichtparametrisch"

Identical by state (IBS)

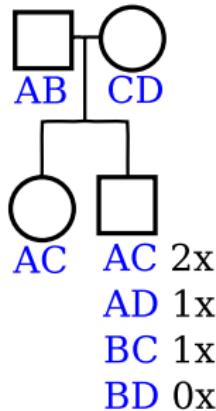


- identity by state = eigenschaftsidentisch
- Dieses Geschwisterpaar besitzt zwei unabhängige Kopien von A_1

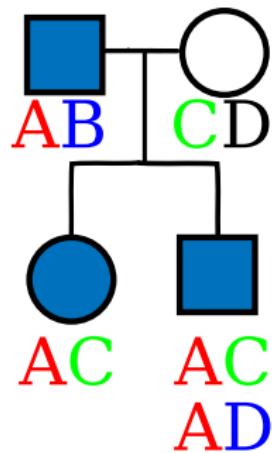
Identical by descent (IBD)



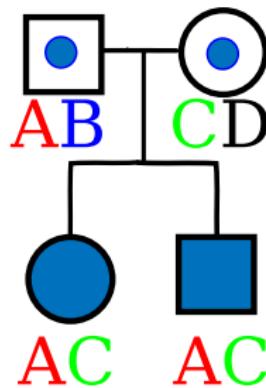
- identity by descent = Identität durch Abstammung
- Dieses Geschwisterpaar besitzt 2 Kopien desselben väterlichen A₁-Allels



- Für ein zufällig gewähltes Locus ist eine zufällige Segregation zu erwarten.
 - ▶ kein elterliches Allel gemeinsam: $1/4$ (AC – BD)
 - ▶ ein elterliches Allel gemeinsam: $1/2$ (AC – AD und AC – BC)
 - ▶ zwei elterliche Allele gemeinsam $1/4$ (AC – AC)



- Handelt es sich beim Locus um den Krankheitslocus für eine dominante Krankheit, besitzen die Geschwister mindestens ein übereinstimmendes Allel.

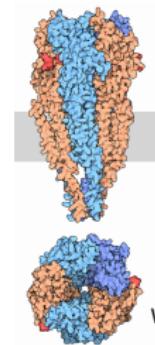


- Handelt es sich beim Locus um den Krankheitslocus für eine rezessive Krankheit, besitzen die Geschwister zwei übereinstimmende Allele.

- Im Allgemeinen (ungeachtet des genetischen Modells) ist eine signifikante Abweichung von den Häufigkeiten $1/4 - 1/2 - 1/4$ ein Hinweis für eine Kopplung an einen prädisponierenden Locus für eine Krankheit
- Daher lässt sich eine ASP-Analyse ohne vorherige Annahmen über die Vererbung der untersuchten Krankheit durchführen

Einzelnes Anfälligkeitssallel

- Was ist ein Anfälligkeitsslocus (susceptibility locus)?
- Beispiel: Nikotinische Acetylcholinrezeptoren werden durch Acetylcholin aber auch durch Nikotin geöffnet
- Nikotinrezeptoren können an Karzinogene binden
- Eine Variante D398N im Nikotinrezeptorgen *CHRNA5* erhöht das Risiko für Lungenkrebs[†]



Wikipedia commons

[†] Hung RJ et al (2008) A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25.

Einzellocusmodell

- Betrachten wir nun ein Modell[†], wobei eine Variante in einem einzigen Locus für eine erhöhte Anfälligkeit für eine bestimmte (nicht mendelnde) Erkrankung verantwortlich ist
- Betrachten wir Verwandte vom Typ R (z.B. R=S, sibs/Geschwister)
 - ▶ Φ_R : Wahrscheinlichkeit, dass zwei Verwandte vom Typ R keine IBD-Allele teilen
- z.B. $\Phi_R = \frac{1}{4}$ für Geschwister, $\Phi_R = \frac{1}{2}$ für Halbgeschwister bzw. für Großeltern/Enkel, ...
- K : Prävalenz der Erkrankung in der Allgemeinbevölkerung
- K_R : Prävalenz der Erkrankung unter Verwandten vom Typ R eines Probanden mit der Erkrankung

†

Risch N (1987) Assessing the Role of HLA-linked and Unlinked Determinants of Disease Am J Human Genet 40:1-14, 1987

Einzellocusmodell (1)

- Falls die Verwandten kein gemeinsames IBD-Allel teilen, dann sind sie diesbezüglich genetisch unabhängig
- Die Wahrscheinlichkeit, dass eine Person in der Bevölkerung die Erkrankung entwickelt, ist K
- Die Wahrscheinlichkeit, dass zwei unabhängige Personen die Erkrankung entwickeln, ist K^2
- Die Wahrscheinlichkeit, dass zwei Verwandte erkranken, gegeben dass sie keine IBD-Allele teilen, ist daher

$$P(\text{2 R erkrankt} | \text{2 R mit 0 IBD}) = K^2$$

Einzellocusmodell (2)

- Ohne *a priori* Information über IBD-Allele, kann die Wahrscheinlichkeit, dass zwei Verwandte beide erkrankt sind, errechnet werden aus der Wahrscheinlichkeit, dass eine Person erkrankt (K) multipliziert durch die Wahrscheinlichkeit, dass ein Verwandter dieser Person erkrankt (K_R)

$$P(\text{2 R erkrankt}) = K \cdot K_R$$

Einzellocusmodell (3)

- Die Wahrscheinlichkeit, dass zwei Verwandte erkranken und keine IBD-Allele teilen, lässt sich anhand der Definition der bedingten Wahrscheinlichkeit berechnen:

$$p(X, Y) = p(X|Y)p(Y)$$

$$\begin{aligned} & P(\text{2 R erkrankt und 2 R mit 0 IBD}) \\ &= P(\text{2 R erkrankt} | \text{2 R mit 0 IBD}) \times P(\text{2 R mit 0 IBD}) \\ &= K^2 \times \Phi_R \end{aligned}$$

Einzellocusmodell (4)

- Bayes' Gesetz

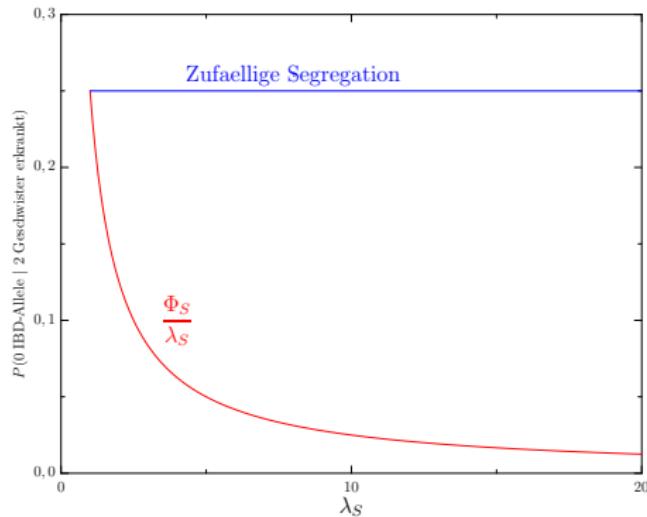
$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

daher¹

$$\begin{aligned} P(2 \text{ R mit 0 IBD} \mid 2 \text{ R erkrankt}) &= \frac{P(\text{2 R erkrankt} \mid 2 \text{ R mit 0 IBD})P(2 \text{ R mit 0 IBD})}{P(\text{2 R erkrankt})} \\ &= \frac{K^2 \cdot \Phi_R}{K \cdot K_R} \\ &= \Phi_R \cdot \frac{K}{K_R} \\ &= \boxed{\frac{\Phi_R}{\lambda_R}} \end{aligned}$$

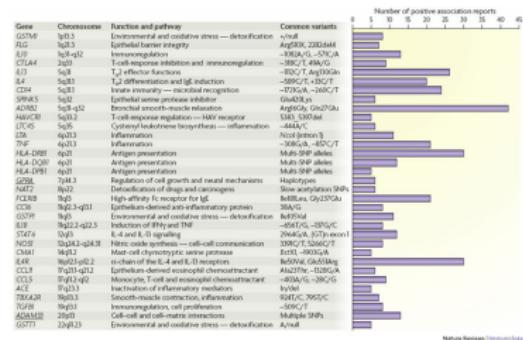
¹ zur Erinnerung: $\lambda_R = \frac{K_R}{K}$; Φ_R : Wahrscheinlichkeit, dass zwei Verwandte vom Typ R keine IBD-Allele teilen

Einzellocusmodell (4)



- Je größer das Risiko für Geschwister eines erkrankten Probanden, selber zu erkranken (λ_S), desto unwahrscheinlicher ist es, dass erkrankte Geschwisterpaare keine IBD-Allele teilen
- Umgekehrt steigt mit λ_S die Wahrscheinlichkeit für erkrankte Geschwisterpaare, gemeinsame IBD-Allele zu teilen.

Einzellocusmodell (5)



Ausgewählte Asthma-Loci. Vercelli (2008) *Nature Reviews Immunology* 8:169-182

- Im Allgemeinen nimmt man die Existenz mehrfacher (jeweils unterschiedlicher) Anfälligkeitsgene für häufige Krankheiten wie Schizophrenie, Herzinfarkt, Diabetes mellitus...
- Eine Vielzahl mathematischer Modelle für multiple Anfälligkeitsloci sind entwickelt worden[†]

[†] z.B. Risch N (1990) Linkage Strategies for Genetically Complex Traits. 1. Multilocus Models Am. J. Hum. Genet.

ASP-Analyse: Ein Locus

- Verschiedene statistische Methoden. Am einfachsten ist der χ^2 -Test:

| geteilte IBD-Allele | 0 | 1 | 2 |
|---------------------|-------|-------|-------|
| beobachtet: | O_0 | O_1 | O_2 |
| erwartet: | E_0 | E_1 | E_2 |

$$\chi_{df=2}^2 = \sum_{i=0}^2 \frac{(O_i - E_i)^2}{E_i}$$

ASP-Analyse: Ein Locus

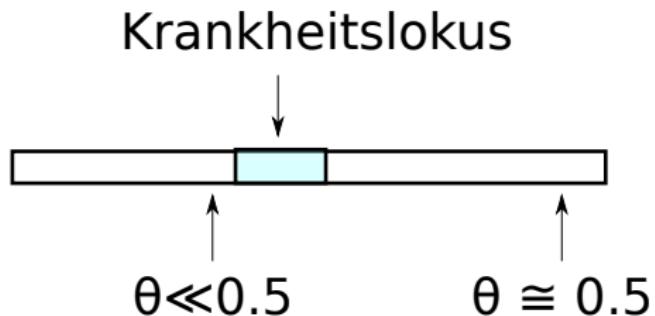
- Erwartete Anzahl geteilte (IBD) Allele

- ▶ $p(\text{beide Allele geteilt})=0.25$
- ▶ $p(\text{ein Allel geteilt})=0.5$
- ▶ $p(\text{kein Allel geteilt})=0.25$

$$\mathbb{E}[\# \text{ geteilte Allele}] = 0.25 \times 2 + 0.5 \times 1 + 0.25 \times 0 = 1 \quad (1)$$

ASP-Analyse: Idee

- Ist ein Marker an den (gesuchten) Krankheitslokus gekoppelt, dann werden zwei erkrankte Geschister dasselbe Markerallel häufiger als per Zufall erben



- $\theta \ll 0.5$: betroffene Geschister teilen mehr als ein IBD Allel, falls genetische Ätiologie vorliegt
- $\theta \approx 0.5$: betroffene Geschister teilen durchschnittlich ein IBD Allel

ASP-Analyse: Beispiel

- 119 Geschwisterpaare mit IDDM (Diabetes) werden auf eine mögliche Kopplung an das XYZ Lokus untersucht
- 10 Paare teilen 0 IBD Allele
- 49 Paare teilen 1 Allel
- 60 Paare teilen 2 Allele

ASP-Analyse: Beispiel

- O_2, O_1, O_0 sind die beobachteten² Zahlen der Geschwisterpaare mit 2, 1 und 0 geteilten IBD Allelen
- E_2, E_1, E_0 sind die erwarteten³ Zahlen
- Für 119 Paare betragen die erwarteten Zahlen:
 - ▶ $E_0 = 0.25 \times 119 = 29.75$
 - ▶ $E_1 = 0.5 \times 119 = 59.5$
 - ▶ $E_2 = 0.25 \times 119 = 29.75$

²“O”: *observed*

³“E”: *expected*

ASP-Analyse: Beispiel

| Anzahl IBD Allele | 0 | 1 | 2 |
|-------------------|-------|------|-------|
| erwartet | 29.75 | 59.5 | 29.75 |
| beobachtet | 10 | 49 | 60 |

```
> ibd = c(10, 49, 60, 29.75, 59.5, 29.75)
> m = matrix(ibd, ncol=3, byrow=T)
> chisq.test(m, corr=F)
```

Pearson's Chi-squared test

```
data: m
X-squared = 21.0247, df = 2, p-value = 2.72e-05
```

- 2 Freiheitsgrade

ASP-Analyse: Beispiel

- Alternativ: Vergleich der beobachteten und erwarteten Anzahl geteilter Allele
- Beobachtet:
 - ▶ geteilt: $2 \times O_2 + 1 \times O_1$
 - ▶ ungeteilt: $2 \times O_0 + 1 \times O_1$
- Erwartet: $N \times 1$ für N Geschisterpaare

$$\begin{aligned}\chi^2 &= \sum_{i=0}^2 \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(2 \times O_2 + 1 \times O_1 - N)^2}{N} + \frac{(2 \times O_0 + 1 \times O_1 - N)^2}{N}\end{aligned}$$

ASP-Analyse: Beispiel

```
> ibd <- c(2*10+49, 2*60+49, 119, 119)
> m <- matrix(ibd, ncol=2, byrow=T)
> m
      [,1] [,2]
[1,]    69   169
[2,]   119   119
> chisq.test(m, corr=F)
```

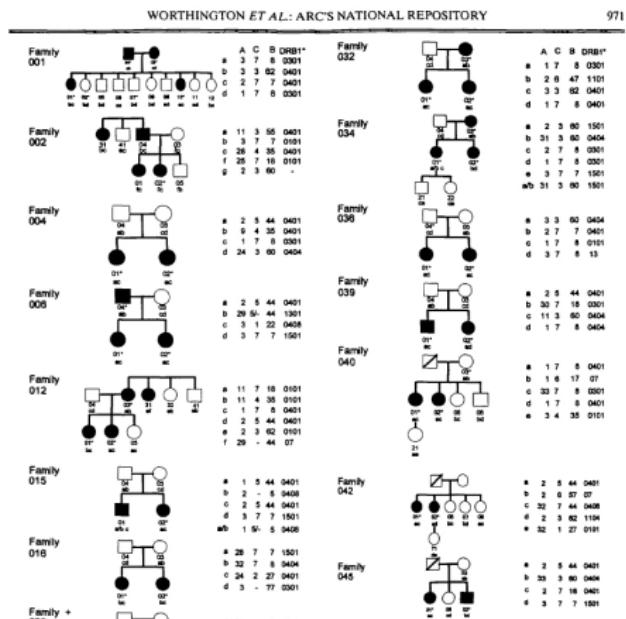
Pearson's Chi-squared test

```
data: m
X-squared = 21.9784, df = 1, p-value = 2.757e-06
```

- 1 Freiheitsgrad

ASP-Analyse: Mehrere Loci

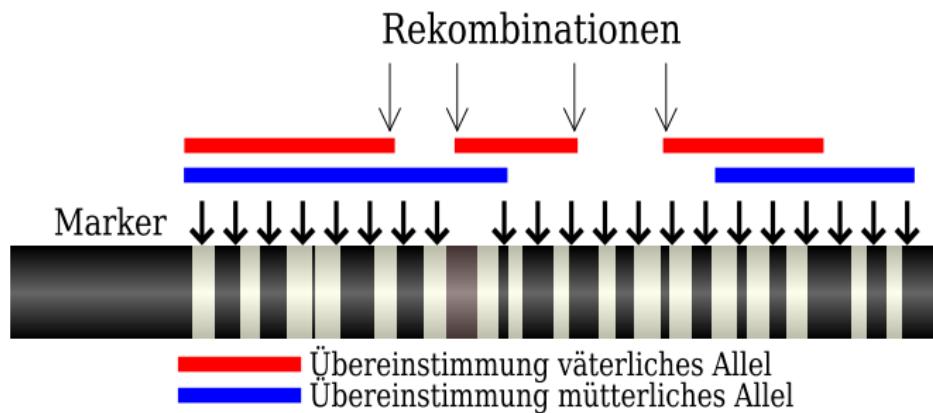
- Familien charakterisieren⁴



⁴ Worthington J. et al. (1994) Br. J. Rheumatol., 33, 970–976

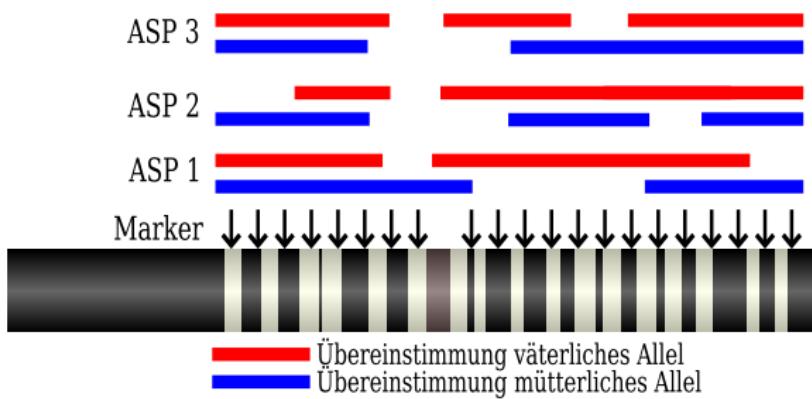
ASP-Analyse: Mehrere Loci

- Loci testen



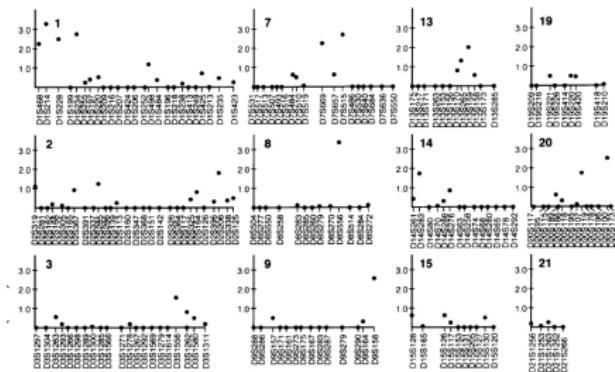
ASP-Analyse: Mehrere Loci

- Ergebnisse für alle Geschwisterpaare auswerten



ASP-Analyse: Mehrere Loci

- Ergebnisse für alle Geschwisterpaare auswerten



Outline

1 Komplexe und monogene Erkrankungen

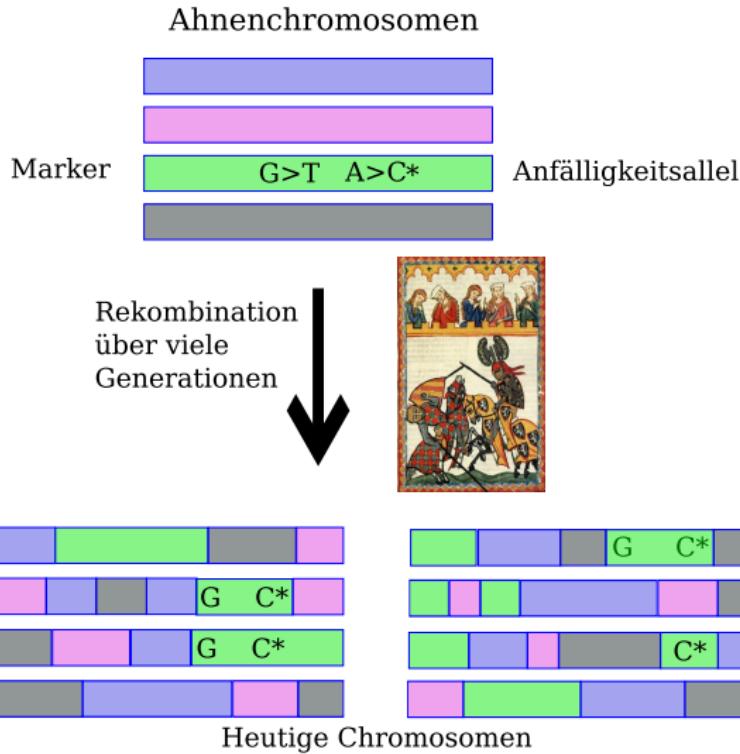
2 Kopplungsanalysen komplexer Merkmale

3 Assoziation vs. Kopplung

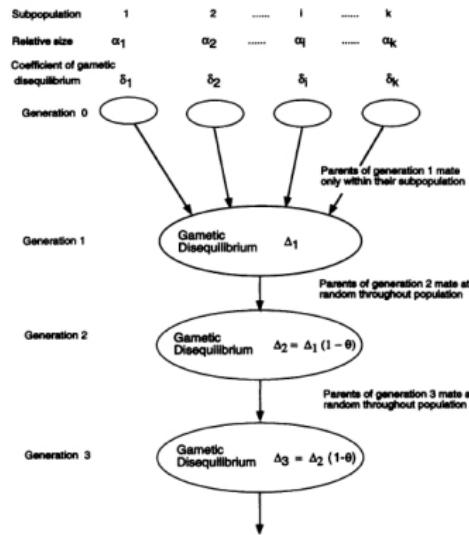
Assoziation vs. Kopplung

| | |
|--------------------------------------|---|
| Kopplung \leftrightarrow Loci | Wie weit ist ein Marker vom Krankheitslokus entfernt? |
| Assoziation \leftrightarrow Allele | Wird ein bestimmtes Markerallel mit dem prädisponierenden Allel mitvererbt? |
| Kopplung | familien spezifisch |
| Assoziation | aufgrund eines Kopplungsungleichgewichts innerhalb einer Population |
| Assoziation | besteht über viel geringere Abstände im Genom als Kopplung |

Assoziation



Assoziation

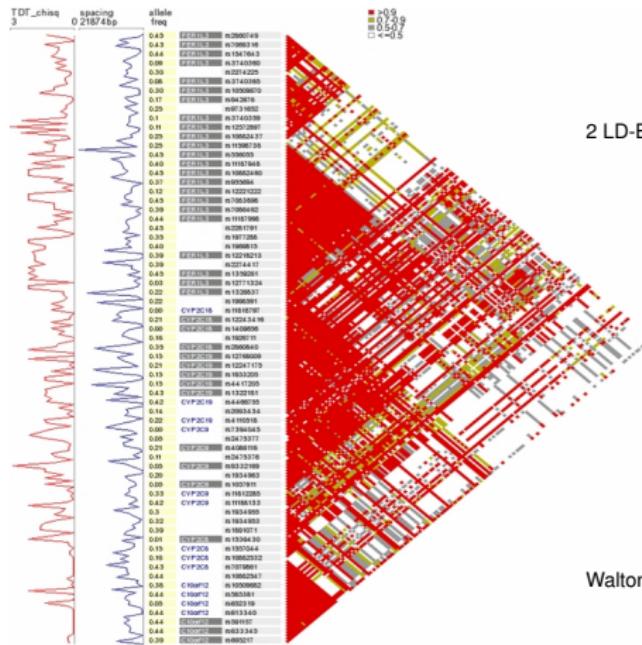


Ewans WJ, Spielman RS (1995)

Am. J. Hum. Genet. 57:455–464

- Anfangs isolierte Populationen
- Nach der Zusammenkunft der ehemals getrennten Populationen, nimmt die Kopplung zwischen Markern entsprechend θ ab

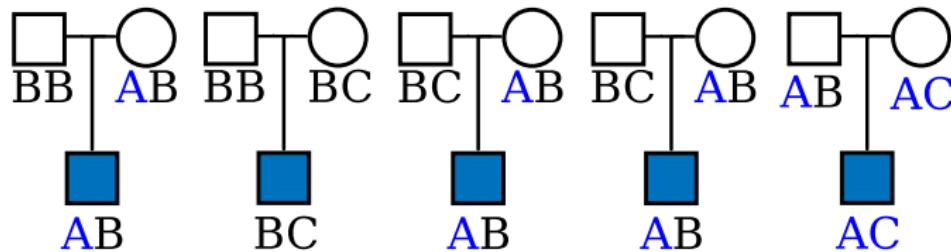
Linkage Disequilibrium



2 LD-Blöcke im CYP2C-Cluster

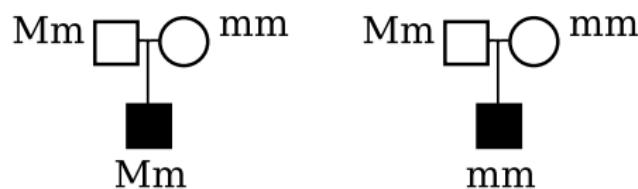
Walton et al (2005) Nature Genetics

Transmission-Disequilibrium-Test (TDT)



- In 4/5 Fällen wurde das Allel A an ein erkranktes Kind vererbt
- Annahme: Allel A ist in LD mit einem "angestammten" Anfälligkeitssallel/Krankheitsgen

Transmission-Disequilibrium-Test (TDT)



- $a = \#$ Fälle, bei denen M vererbt wurde
- $b = \#$ Fälle, bei denen M nicht vererbt wurde
- Statistik:

$$\chi_{td}^2 = \frac{(a - b)^2}{(a + b)}$$

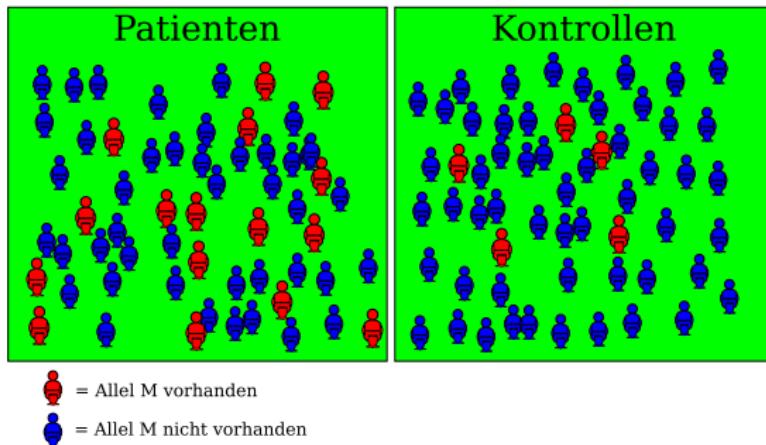
- Ungefähr χ^2 -verteilt mit einem Freiheitsgrad

Fall-Kontroll-Studien

- Die Häufigkeit von SNP-Allelen in zwei Gruppen vergleichen
 - ▶ Personen mit der Erkrankung
 - ▶ Unbetroffene Kontrollen oder zufällig aus der Bevölkerung ausgewählte Personen
- Eine erhöhte Häufigkeit eines SNP-Allels in Patienten verglichen mit Kontrollen deutet darauf hin, dass das Allel das Risiko für die Erkrankung erhöht

Fall-Kontroll-Studien

Fall-Kontroll-Studie



| Genotyp | Patienten | Kontrollen |
|-------------------|-----------|------------|
| M vorhanden | a | b |
| M nicht vorhanden | c | d |

- Statistische Auswertung mit χ^2 -Test

Fall-Kontroll-Studien

| | AA | AB | BB |
|------------|----|----|----|
| Patienten | a | b | c |
| Kontrollen | d | e | f |

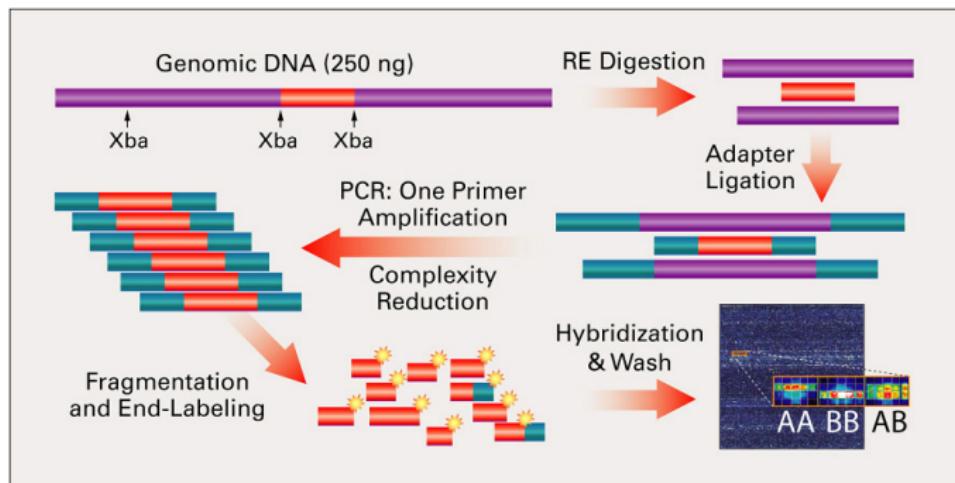
$$\chi^2 = \sum_{i=0}^2 \frac{(O_i - E_i)^2}{E_i}$$

- Mehrere Varianten für spezifische Fragestellungen[†]

†

vgl. Lewis CM (2002) Genetic association studies: design, analysis and interpretation. *Brief Bioinform* 3:146-53.

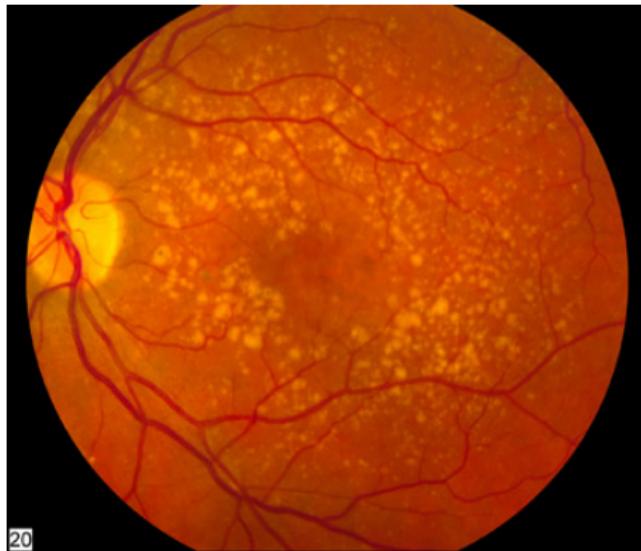
SNP-Chips



Bildquelle: www.imagenes-bio.de

- SNP-Chips mit bis zu 500.000 SNPs (oder mehr) von Affymetrix und anderen Anbietern

Beispiel: Makuladegeneration



- Degeneration der Makula lutea ("Punkt des schärfsten Sehens")
- Hauptursache der Erblindung bei Menschen über 55 Jahre alt

Beispiel: Makuladegeneration

- Bis 2005 hatten Assoziationsstudien mit Kandidatengenen kein positives Ergebnis erbracht.
- Studie mit 100K SNP-Chip[†] mit 96 Fällen und 50 Kontrollen
- 116.204 SNPs ausgewertet, davon waren 105.980 informativ. Davon wurden 103.611 autosomalen SNPs in die Endanalyse einbezogen

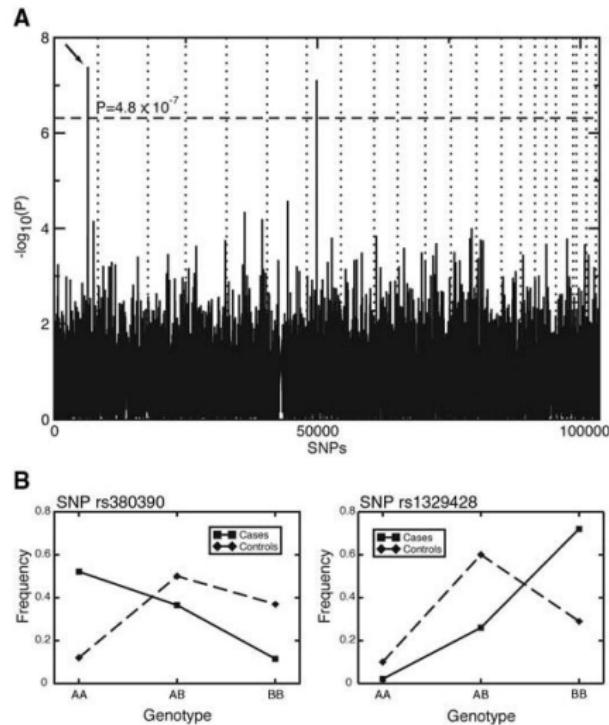
[†] Klein RJ, et al. (2005) Complement factor H polymorphism in age-related macular degeneration. *Science* **308**:385–9

Beispiel: Makuladegeneration

- Für jeden SNP wurde ein Test auf allelische Assoziation mit Makuladegeneration durchgeführt.
- Als Korrektur für multiples Testen wurde eine Bonferroni-Korrektur durchgeführt, d.h., nur diejenigen SNPs wurden für signifikant gehalten, welche einen p -Wert $< 0,05/103.611 = 4.8 \times 10^{-7}$ aufwiesen.

Beispiel: Makuladegeneration

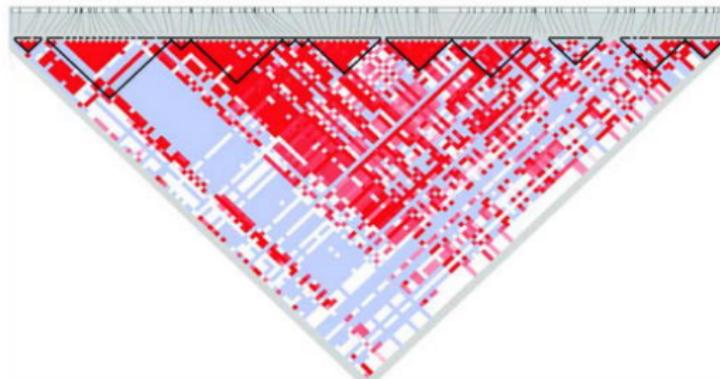
- Zwei SNPs waren signifikant



Bildquelle: Klein et al. (2005)

Beispiel: Makuladegeneration

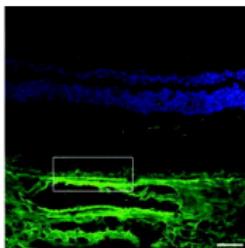
- Beide SNPs lagen in einem Intron des Gens für Complement H auf Chromosom 1q31
- Eine LD-Analyse zeigte, dass die SNPs innerhalb einer Region mit hohem LD gelegen sind



Bildquelle: Klein et al. (2005)

Beispiel: Makuladegeneration

- Die SNP-Studie hat gekoppelte Marker innerhalb einer Region mit starkem Kopplungsungleichgewicht identifiziert
- Jetzt galt es, den eigentlichen funktionellen Polymorphismus zu finden, welcher die Entstehung der Makuladegeneration bei den Patienten begünstigt hatte
- Alle Exons des Gens für Complement H wurden sequenziert → Y402H
- Ablagerung von Complement wurden in der Netzhaut betroffener Patienten nachgewiesen



Bildquelle: Klein et al. (2005)

Genetische Assoziationen

Gene

Umwelt

Marfan Syndrome
Darmkrebs
Neurofibromatosis
Brustkrebs
Hämophilie
Mukoviszidose

M. Alzheimer
Diabetes
Asthma
Schizophrenie

Lungenkrebs
Hautkrebs
Tuberkulose

zum Schluss

- Email: peter.robinson@charite.de

weiterführende Literatur

- Strachan & Read, Kapitel 15
- Lewis CM (2002) Genetic association studies: design, analysis and interpretation. *Brief Bioinform* 3:146-53.