

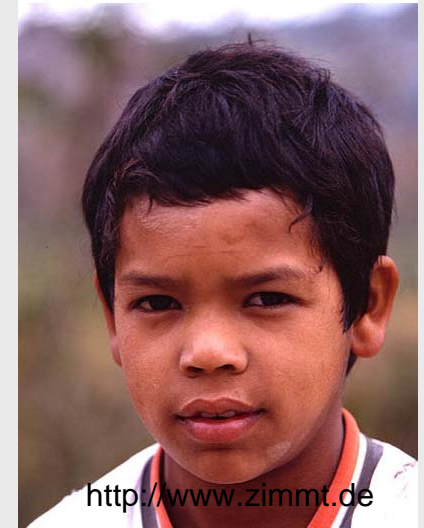
# Haplotypanalysen: Methoden und Programme

**Karla Köpke**

Charité – Universitätsmedizin Berlin, CCM,  
Institut für Klinische Pharmakologie

# Genetische Variabilität

---



# Genetische Variabilität: SNPs - Genotyp - Haplotypen

## Referenzsequenz

...CAAT**G**GAAG**A**CCATGCGCCGTTCCACGACGTCACGCAG**G**AA...



**C**



**T**



**A**

## Genotyp

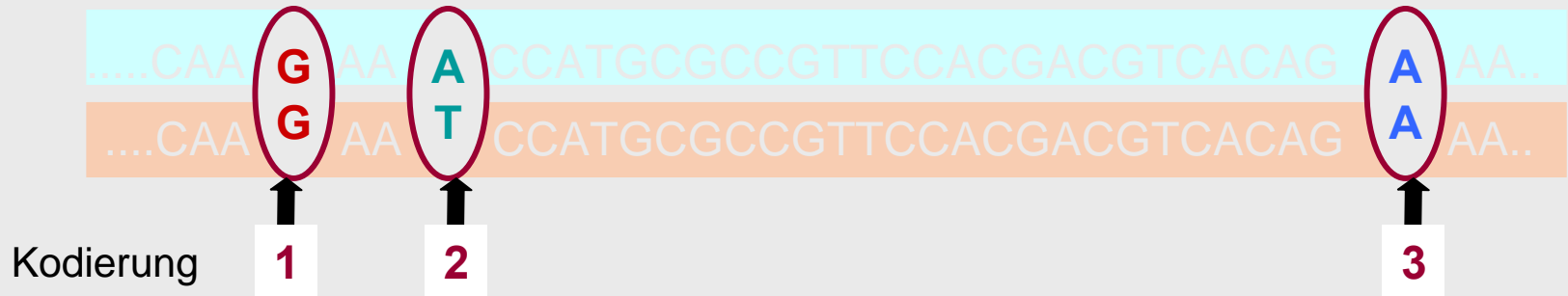
...CAAT**C**GAAG**A**CCATGCGCCGTTCCACGACGTCACGCAG**A**AA...

...CAAT**G**GAAG**T**CCATGCGCCGTTCCACGACGTCACGCAG**A**AA...

# Genetische Variabilität: SNPs - Genotyp - Haplotypen

Referenz **G** **A** **G**

Genotyp

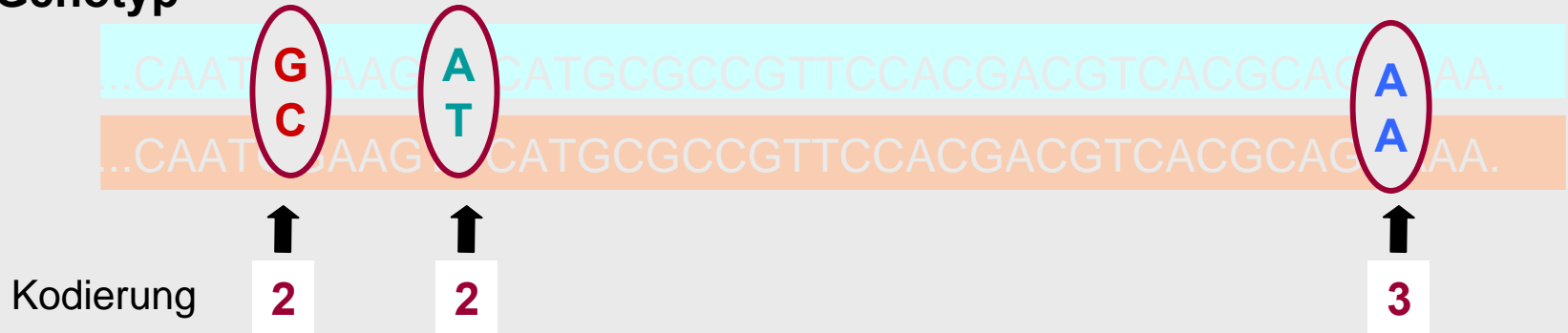


Haplotyp-Paar



# Genetische Variabilität: SNPs - Genotyp - Haplotypen

## Genotyp



## Haplotyp-Paar 1: (212; 122)

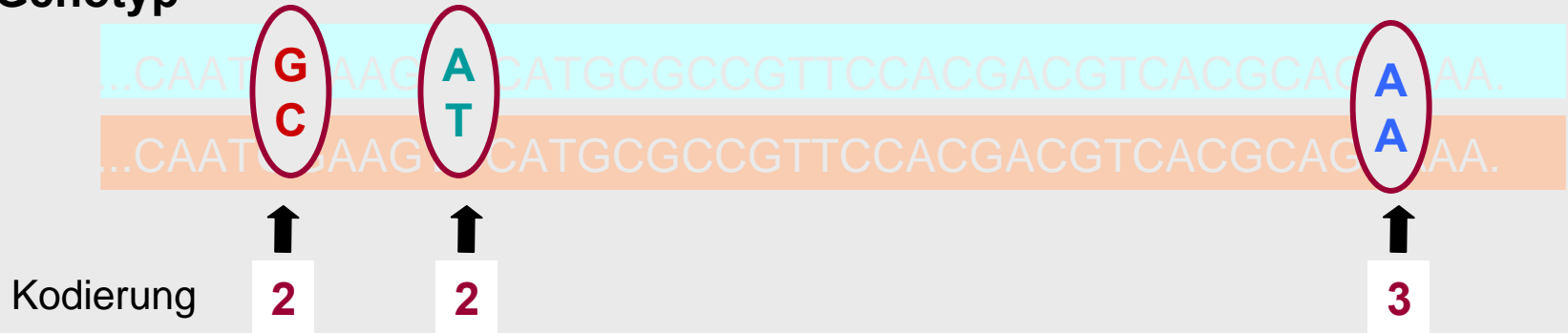
...CAAT **C** GAAG **A** CCATGCGCCGTTCCACGACGTCACGCAG **A** AA...  
...CAAT **G** GAAG **T** CCATGCGCCGTTCCACGACGTCACGCAG **A** AA...

## Haplotyp-Paar 2: (222; 112)

...CAAT **C** GAAG **T** CCATGCGCCGTTCCACGACGTCACGCAG **A** AA...  
...CAAT **G** GAAG **A** CCATGCGCCGTTCCACGACGTCACGCAG **A** AA...

# Genetische Variabilität: SNPs - Genotyp - Haplotypen

## Genotyp



Haplotyp-Paar 1: (212; 122)



zeigt Phänotyp A

...CAAT **C** GAAG **A** CCATGCGCCGTTCCACGACGTCACGCAG **A** AA...  
...CAAT **G** GAAG **T** CCATGCGCCGTTCCACGACGTCACGCAG **A** AA...

Haplotyp-Paar 2: (222; 112)



zeigt Phänotyp A nicht

...CAAT **C** GAAG **T** CCATGCGCCGTTCCACGACGTCACGCAG **A** AA...  
...CAAT **G** GAAG **A** CCATGCGCCGTTCCACGACGTCACGCAG **A** AA...

# Konzept der Haplotyp-Blockstruktur

---

Ca. 11 Millionen SNPs mit einer Häufigkeit  $>1\%$  werden im menschlichen Genom erwartet

## Definitionen des Haplotyp-Blockes

- LD-definierte Blöcke: Es wird gefordert, dass ein gewisser Anteil von Markern ein  $D'$  über einem Grenzwert besitzt (strenges LD) (Gabriel et al. 2002)
- Blöcke zeigen keine Rekombination (Untersuchung der Verteilung der beobachteten Rekombinationen zwischen den SNPs einer Region) (Wang et al. 2002)
- Konzept der Überdeckung:
  - eine kleine Anzahl von häufigen („common“) Haplotypen überdecken den betrachteten Chromosomenbereich (Patil et al. 2001)
  - Region mit vermindertem Niveau der Haplotypenvielfalt (Daly et al. 2001)

Ausnutzung dieses Phänomens durch Betrachtung repräsentativer SNPs für die Blöcke (Reduzierung des Genotypisierungsaufwandes)

# Aufgaben der *in silico* Haplotypbestimmungen

---

Die folgenden Ausführungen beziehen sich auf den Fall einer Stichprobe von nicht verwandten Individuen.

Für Genotypen, die in mindestens zwei Varianten heterozygot sind, ist eine eindeutige Zuordnung eines Haplotyp-Paares nicht mehr möglich.

- Schätzung der Häufigkeiten von Haplotypen in der Population
- Rekonstruktion des/r Haplotyp-Paare/s/ für die Probanden einer Stichprobe

# Aufgabe und Lösungsansatz

---

Zu einer gezogenen Stichprobe von Genotypen sind die kompatiblen Haplotypen zu bestimmen und die Häufigkeiten für die Haplotypen bzw. für die Haplotyp-Paare zu schätzen.

## Probleme

- Die Mechanismen, die zur heutigen Haplotypverteilung in einer Population geführt haben, sind komplex und im Detail unbekannt.
- Die Genotyp-Daten sind verrauscht (Meßfehler, Migration u.a.).

## Lösung

Bayesscher Ansatz: Bewertung von Modellen in Abhängigkeit von den erhobenen Daten und Suche nach den Parametern für die Modelle, so dass die a posteriori Wahrscheinlichkeit maximal wird → Optimierungsproblem

# Methoden - Überblick

---

Auffinden der kompatiblen Haplotypen

## 1. Clark's Subtraktionsmethode (Clark 1990)

heuristisches Verfahren

## 2. Methoden der Kombinatorik

## 3. Graphentheoretische Verfahren

# Methoden - Überblick

---

## 4. Wahrscheinlichkeitstheoretische Ansätze

- Das Stichprobenergebnis wird über eine Likelihoodfunktion in Abhängigkeit von den unbekanntem Parametern bewertet.
- Gesucht wird eine Schätzung für die unbekanntem Parameter der Funktion, so dass die Likelihood maximal wird.
- Verfahren des maschinellen Lernens werden eingesetzt, die zum Auffinden entsprechender Likelihood Schätzungen bei unvollständigen Daten geeignet sind.
  - EM (Expectation-Maximization) -Algorithmus
  - Markov-Chain Monte-Carlo Methoden (Gibbs Sampling)
  - Dynamische Programmierung (Zerlegung in einfach zu lösende Teilprobleme ist möglich, „Divide & Conquer“ mit Verwendung vorheriger Teillösungen )

# Modelle - Überblick

---

## 1. “Parsimony” Modell

### Hintergrund

Die beobachtete Anzahl von Haplotypen in einer Population ist viel kleiner als die theoretisch mögliche Anzahl von  $2^n$  für Haplotypen mit  $n$  biallelischen Varianten.

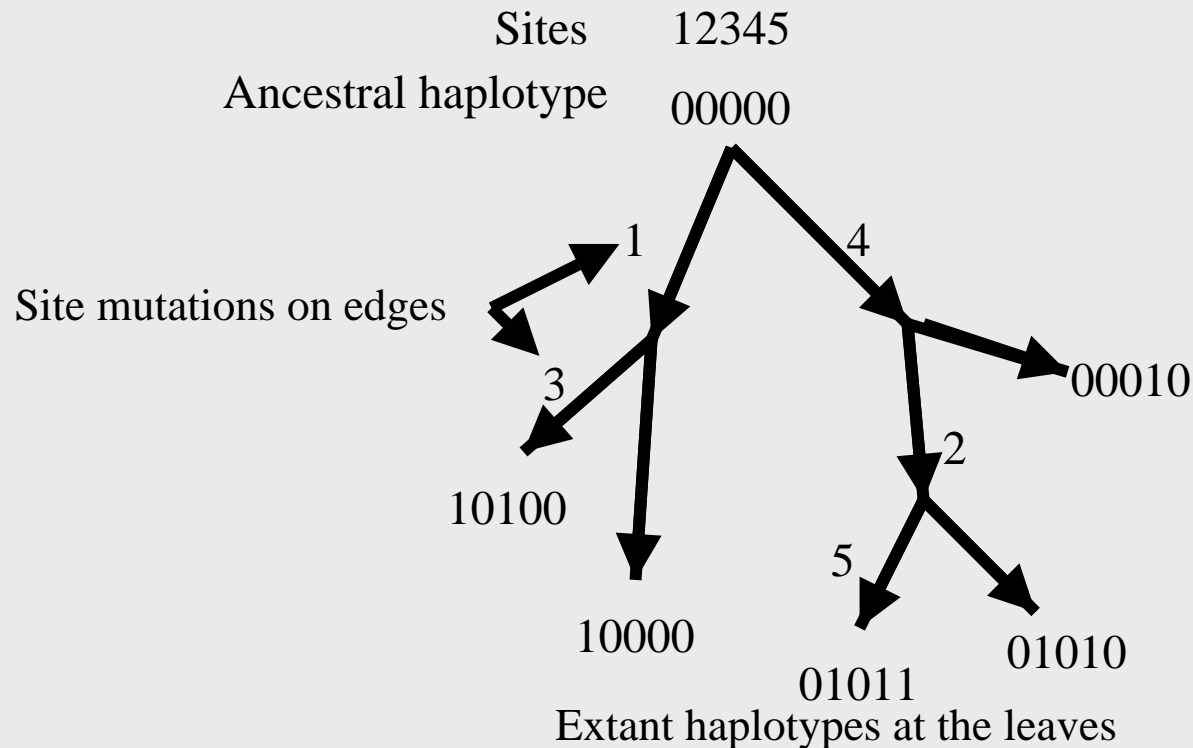
### Ziel

Beschreibung der Genotypen einer Stichprobe mit der geringsten Anzahl von verschiedenen Haplotypen.

# Modelle - Überblick

## 2. “Perfect Phylogeny” Modell (PPH, Gusfield et al. 2003)

**Annahme:** Die Haplotypen haben sich aus einem “Ur-Haplotyp” abgeleitet, Rekombination ist nicht erlaubt.



# Modelle - Überblick

## 3. "Coalescence" Theorie (PHASE, Stephens et al. 2001)

Die a priori Information wird aus der „coalescence“ Theorie abgeleitet. Dem Individuum 1 wird das Haplotyp-Paar, das aus häufigen Haplotypen zusammengesetzt ist, zugeordnet. Für Individuum 2 werden solche Haplotypen gesucht, die den in der Liste vorhandenen möglichst „ähnlich“ sind.

### Known haplotypes:

22544  
22544  
22544  
22544  
33334  
33334  
23233  
14234

### Ambiguous individual 1:

Genotype  
32344  
23534

```
graph LR; G1[32344] --> H1[33334]; G1 --> H2[22544]; G2[23534] --> H1; G2 --> H2;
```

### Ambiguous individual 2:

Genotype  
32444  
23434

```
graph LR; G3[32444] --> H3[33434]; G3 --> H4[22444]; G4[23434] --> H3; G4 --> H4;
```

# Modelle - Überblick

---

## 4. Hardy-Weinberg-Gleichgewicht (HWG)

In unendlichen diploiden Populationen gilt für autosomale Gene mit zweifacher Allelie bei gleicher Fruchtbarkeit und Vitalität der Genotypen, fehlender Mutation und Selektion, dass sich die Allelwahrscheinlichkeiten  $p=w(A_1)$ ,  $q=w(A_2)$  einer Generation, die durch Zufallspaarung aus einer Ausgangspopulation mit den Genotypenwahrscheinlichkeiten  $P$ ,  $2Q$  und  $R$  hervorgegangen ist, aus  $p=P+Q$ ,  $q=Q+R$  errechnen.

Die Genotypwahrscheinlichkeiten sind dann gleich

$$w(A_1A_1) = p^2$$

$$w(A_1A_2) = 2pq$$

$$w(A_2A_2) = q^2$$

Gilt bereits in der Ausgangspopulation  $P=p^2$ ,  $Q=pq$ ,  $R=q^2$ , so sind die Allelwahrscheinlichkeiten bei Zufallspaarung in allen Generationen stabil.

Unter diesen Voraussetzungen befindet sich eine Population nach einer Generation Zufallspaarung im Gleichgewicht.

## 5. Modell der Zufallspaarung

# Clark's Subtraktionsmethode: HAPINFERX

---

## **Schritt 1:**

Auflisten aller Haplotypen, die mit Sicherheit in der Stichprobe existieren (abzuleiten aus allen Genotypen, die maximal eine heterozygote Position enthalten)

## **Schritt 2:**

Erklärung der Genotypen mit mindestens zwei heterozygoten Positionen in der gegebenen Reihenfolge durch die vorhandenen Haplotypen in der Liste aus Schritt 1. Wenn das nicht möglich ist, Erweiterung der Liste durch einen Haplotypen so, daß der Genotyp erklärt werden kann. Wenn ein Genotyp durch keinen der vorhandenen Haplotypen erklärt werden kann, ist für diesen keine Haplotypzerlegung möglich.

## **Nachteile dieses Verfahrens**

- keine eindeutigen Ergebnisse
- Abhängigkeit von der Reihenfolge der Genotypen in der Stichprobe
- nicht immer anwendbar

# Modifikation der Subtraktionsmethode von Clark

---

## Verbesserung des Verfahrens von Clark durch Orzack et al. (2003)

- Betrachtung der Menge aller möglichen Ergebnisse
- Entwicklung von 8 Varianten des auf Regeln beruhenden Verfahrens von Clark, die sich in der Bildung der Liste in Schritt 1 und Schritt 2 unterscheiden
- Die endgültige Zuordnung des Haplotyp-Paares erfolgt durch eine “Consensus“- Betrachtung der möglichen Ergebnisse.

# „Parsimony“ Modell: HAPAR

---

## HAPAR (Wang und Xu 2003)

nutzt ein “Branch-and-Bound” Verfahren

„**Branch-and-Bound**“ Verfahren sind zur Lösung kombinatorischer Optimierungsprobleme geeignet und bestehen in zwei Schritten:

### **Branch:**

Aufgabe wird in zwei oder mehr Teilaufgaben zerlegt, jede entstehende Teilaufgabe wird wieder zerlegt - es entsteht eine Baumstruktur von Teilaufgaben

### **Bound:**

Ist die optimale Lösung einer Teilaufgabe nicht besser als eine schon gefundene Lösung der Gesamtaufgabe, so wird die Teilaufgabe nicht weiter betrachtet.

# Programme mit EM-Algorithmen

---

## EM-Algorithmen

- geeignet zur Schätzung der Haplotyphäufigkeiten in einer Population
- weniger geeignet zur Rekonstruktion der Haplotyp-Paare
- häufig angewendet
- nicht geeignet für viele Varianten (Kombination mit anderen Verfahren notwendig)
- Rekombination ist nicht problematisch

In verschiedenen Programmen unterschiedlich implementiert: z.B.

1. Excoffier und Slatkin, 1995: ARLEQUIN
2. Terwilliger und Ott; 1994: EH
3. Hawley und Kidd, 1995: HAPLO
4. Rohde and Fuerst, 2001: ithap

# EM-Algorithmen - Likelihood-Funktion

---

## Ableitung der Likelihood-Funktion

Gegeben seien die Genotypen von  $n$  Probanden mit unbekanntem Haplotyp-Paar. Die  $m$  verschiedenen Genotypen sollen  $n_i$ -mal vorkommen ( $i=1, \dots, m$ ).

Die Anzahl  $c_j$  der Haplotyp-Paare, die einen Genotypen erklären können, ist von der Anzahl  $s_j$  der heterozygoten Positionen des Genotyps abhängig.

$c_j = 1$  für  $s_j = 0$  und  $c_j = 2^{s_j - 1}$  für  $s_j > 0$ .

Die Wahrscheinlichkeit für den  $j$ -ten Genotypen  $P_j$  ist dann:

$$P_j = \sum_{k=1}^{c_j} P(h_k h_l),$$
 wobei  $P(h_k h_l)$  die Wahrscheinlichkeit dafür ist, dass dieser

Genotyp aus dem Haplotyp-Paar  $h_k$  und  $h_l$  gebildet wird.

# EM-Algorithmen - Likelihood-Funktion

---

Die Likelihood der Haplotypfrequenzen bei gegebenen Anzahlen der Genotypen in der Stichprobe lautet:

$$L(p_1, p_2, \dots, p_h) = a_1 \prod_{j=1}^m \left( \sum_{i=1}^{c_j} P(h_{ik} h_{il}) \right)^{n_j}, \quad \text{mit } p_h = 1 - p_1 - p_2 - \dots - p_{h-1}$$

# EM-Algorithmen - Schritte

---

**Durchführung:** iterativer Prozess

1. Suche nach geeigneten Startwerten
2. **“Expectation”** Schritt: Berechnung von Erwartungswerten für die fehlenden Daten auf der Basis der Startwerte
3. **„Maximization“** Schritt: Berechnung der Maximum-Likelihood-Schätzungen für die Parameter für den vollständigen Datensatz
4. Wiederholung der Prozedur mit den neuen Parametern als Startwerte bis die Veränderungen der Parameterschätzungen eine Schranke unterschreiten

Das Ergebnis des EM-Algorithmus ist eine Maximum-Likelihood-Schätzung mit gut untersuchten statistischen Eigenschaften (wenn das globale Maximum erreicht wird).

## **Problem des Algorithmus**

Konvergenz gegen lokale Maxima, deshalb Wiederholung des Algorithmus mit verschiedenen zufälligen Startwerten notwendig

# Programme mit EM-Algorithmen für Haplotypen mit vielen Varianten - SNP HAP

---

## SNP HAP (Clayton 2002)

Einbeziehung eines Monte Carlo IP (Imputation/Posterior sampling)

### Algorithmus

Beginnt mit der Berechnung von Haplotypen, die aus zwei Varianten bestehen, dann schrittweise Erweiterung der Haplotyplänge durch Hinzunahme weiterer Varianten.

Nach Anwendung des EM-Algorithmus werden die Haplotypwahrscheinlichkeiten geschätzt. Dann werden die Haplotyp-Paare, deren Wahrscheinlichkeit unter die vorgegebene Grenze fällt, gesichtet und weggelassen sowie die posterior Wahrscheinlichkeit neu berechnet.

### Problem

Schrittweise Hinzunahme kann zu nicht optimaler Lösung führen, deshalb Wiederholung des Prozesses mit anderer Reihenfolge der Hinzunahme der Varianten oder Ausführung des "Aussonderungsprozesses" von Haplotypen nicht nach Hinzunahme jeder einzelnen Variante, sondern nach Hinzunahme von  $k$  Varianten ( $k$ : Problem von Berechnungszeit und Speicherkapazität).

# Programme mit EM-Algorithmen für Haplotypen mit vielen Varianten - HPLUS

---

## HPLUS (Zhao, Li und Khalid 2001)

Benutzung derselben Likelihoodfunktion wie bei Excoffier und Slatkin, zusätzlich:

Anwendung eines “progressive Ligation” Berechnungs-Algorithmus

## Partition Ligation (PL) Algorithmus

1. gehört zur Gruppe der “Divide-and-Conquer” Algorithmen, die die Aufgabe in mehrere gleichartige, aber kleinere Teilaufgaben zerlegen
2. dieser Prozess wird solange fortgesetzt bis die Teilaufgaben effizient gelöst werden können
3. Im letzten Schritt werden die Lösungen der Teilaufgaben zu einer Lösung der ursprünglichen Aufgabe zusammengeführt.

Zwei Strategien sind möglich

- progressive Ligation
- hierarchical Ligation

# Weitere Programme mit EM-Algorithmen für Haplotypen mit vielen Varianten

---

**Programme, die auch PL Algorithmen nutzen:**

➤ PL-EM

➤ HAPLOTYPER

➤ Haploview (für Blöcke mit mehr als 10 Markern) – Standard in der HapMap

# Programme mit MCMC-Algorithmen - PHASE

---

## **PHASE, Version 2 (Stephens et al., 2003)**

Um die Effizienz der Berechnungen zu erhöhen, wurde eine „Divide-and-Conquer“ Strategie für die Schätzung der Haplotypfrequenzen von Teilmengen von aufeinanderfolgenden Varianten eingesetzt.

# Programme mit einem Gibbs Sampling Verfahren- HAPLOTYPER

---

## HAPLOTYPER (Niu et al., 2002)

Benutzt als Modell die Gültigkeit des HWG

Partition Ligation Verfahren

# Annahmen der verschiedenen Modelle

---

- Clark's Subtraktionsmethode: "eindeutige" Genotypen müssen in der Stichprobe enthalten sein
- "Parsimony" Modell: Natur hat die Menge der Haplotypen minimiert
- Hardy-Weinberg-Gleichgewicht ist gültig
- "Perfect Phylogeny" Modell: Rekombination ist nicht erlaubt
- "Coalescence" Theorie ist gültig

# Verfahren von Genprofile

---

$G$  sei die Menge der  $m$  verschiedenen Genotypen  $g_i$  ( $i=1, \dots, m$ ) in einer Stichprobe des Umfangs  $n \geq m$

$n_i$  sei die Anzahl, mit der  $g_i$  in der Stichprobe vorkommt

$c_i$  sei die Anzahl der verschiedenen Haplotypen, die für  $g_i$  theoretisch als Lösung auf einem Chromosom in Frage kommen (kompatibel sind)

$G(h_k)$  sei die Untermenge von  $G$ , für die  $h_k$  kompatibel ist

Jedem Haplotyp  $h_k$ , der mit wenigstens einem Genotyp aus  $G$  kompatibel ist,

wird ein Gewicht  $w_k$  zugeordnet: 
$$w_k = \sum_{g_i \in G(h_k)} n_i / c_i$$

Für jedes Haplotyp-Paar  $(h_k; h_l)$  wird ein Score  $\sigma_{kl} = w_k w_l$  berechnet.

Gesucht: Alle Haplotyp-Paare und ihre Scores für jedes  $g_i$  aus  $G$ .

# Programme zur Haplotypschtzung und zur Rekonstruktion der Haplotyp-Paare

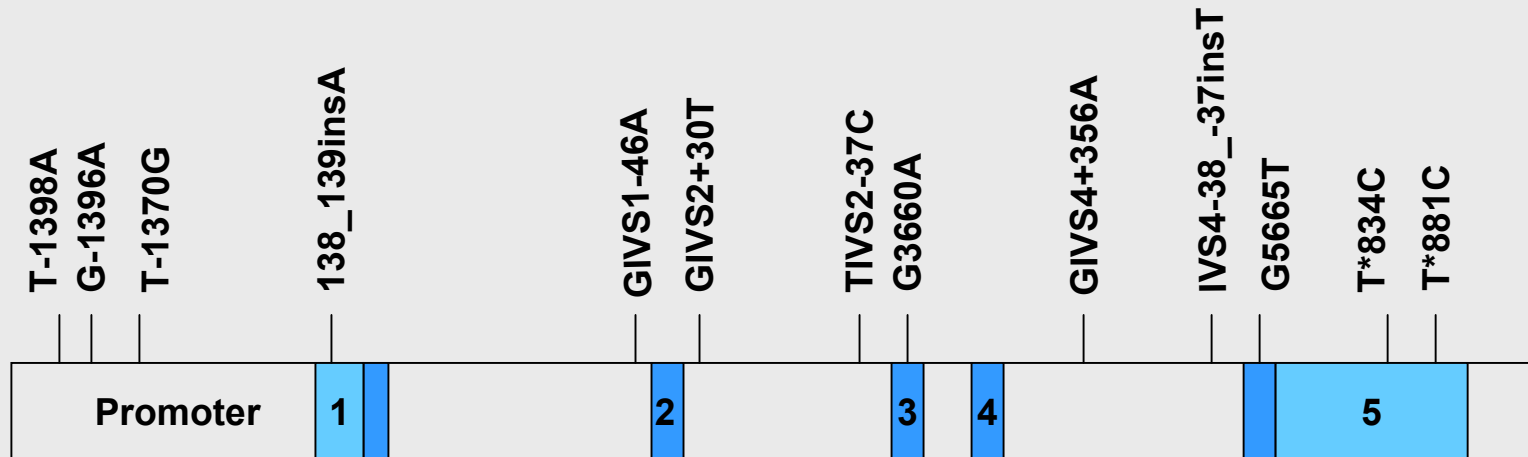
Programm	Referenz
Arlequin Version 2.0	Excoffier L, Slatkin M (1995)
ithap	Rohde K, Fuerst R (2001)
HAP (SNPHAP)	Zhao JH (2004)
PL-EM	Qin ZS (2002)
HPLUS	Li SS et al. (2003)
GenProfile	GenProfile (2000)
HAPLOTYPER/ HTYPERv2*	Niu T et al. (2002)
PHASE Version 2.0	Stephens M et al. (2001; 2003)
Haploview	Barrett JC et al. (2005)
HAP (H. & E.)	Halperin E, Eskin E (2004)
HAPAR	Wang L, Xu Y (2003)
DPPH und GPPH (PPH)	Bafna V et al. (2003); Chung RH, Gusfield D (2003)

\*Ausgabe des wahrscheinlichsten Haplotyp-Paares

# Datensätze

Kandidaten-Gen	Anzahl der Varianten		Variante mit mehr als 2 Allelen	Stichproben-größe	Fehlende Daten	Referenz
	alle	MAF $\geq$ 10%				
<b><i>OATP-C (SLCO1B1)</i></b> organic anion transporting polypeptide C gene	6	6	nein	300	nein	Gerloff et al. (2006)
<b><i>EDN1 (ET1)</i></b> endothelin-1gene	11	7	nein	298	ja (n=1)	Diefenbach et al. (2006)
<b><i>MDR1 (ABCB1)</i></b> multidrug resistance gene 1	9	6	ja	461	ja (n=82)	Cascorbi et al. (2001)

# Polymorphes Spektrum des *EDN1*



# Vergleich der Ergebnisse verschiedener Programme: Haplotypen des *EDN1* mit einer Häufigkeit $\geq 0,1\%$

	GenProfile		Arlequin		HAP (H. & E.)		htyperv2		PHASE2.0		HAP (SNPHAP)		PLEM		HPLUS		Haploview	
1	11211111111	41.95	11211111111	41.64	11211111111	42.28	11211111111	41.44	11211111111	40.91	11211111111	41.24	11211111111	41.78	11211111111	41.50	11211111111	41.10
2	12212212121	12.08	12212212121	11.97	12212212121	11.91	12212212121	12.08	12212212121	12.12	12212212121	12.11	12212212121	11.89	12212212121	12.20	12212212121	12.10
3	11112111111	10.40	11112111111	10.81	11112111111	10.57	11112111111	10.40	11112111111	10.77	11112111111	11.03	11112111111	10.91	11112111111	10.90	11112111111	10.90
4	11112121111	9.06	11112121111	8.42	11112121111	9.06	11112121111	9.06	11112121111	9.09	11112121111	8.40	11112121111	8.42	11112121111	8.40	11112121111	8.60
5	11212111111	4.70	11212111111	4.49	11212111111	4.36	11212111111	4.87	11212111111	4.55	11212111111	4.34	11212111111	4.52	11111111111	4.30	11212111111	4.30
6	11212212121	4.19	11111111111	4.23	11212212121	4.03	11212212121	4.03	11212212121	4.04	11111111111	4.23	11111111111	4.21	11212111111	4.30	11111111111	4.30
7	21222111212	3.86	11212212121	4.17	11111111111	3.69	11111111111	3.86	11111111111	3.87	11212212121	4.16	11212212121	4.16	11212212121	4.20	11212212121	4.20
8	11111111111	3.69	21222111212	3.28	21222111212	3.52	21222111212	3.52	21222111212	3.54	21222111212	3.78	21222111212	3.43	21222111212	3.50	21222111212	3.70
9	11112212121	2.01	11212121111	2.28	11112212121	3.69	11112212121	2.35	11112212121	2.36	11212121111	2.27	11212121111	2.26	11212121111	2.30	11212121111	2.40
10	11212121111	1.68	11112212121	1.87	11212121111	2.01	11212121111	1.85	11212121111	1.85	11112212121	1.92	11112212121	1.86	11112212121	1.90	11112212121	1.90
11	11112212111	1.34	11112212111	1.25	11112212111	1.34	11112212111	1.34	11112212111	1.52	11112212111	1.25	11222111212	.50	11112212111	1.30	11112212111	1.30
12	11222111212	.50	11222111212	.51	11211111112	.50	11211111112	.67	21211111111	.67	11222111212	.50	11212212111	.37	11222111212	.50	11222111212	0.50
13	11111111112	.34	11211111112	.37	11222111212	.50	11222111212	.50	11211111112	.51	11212212111	.37	11212211112	.33	11211111112	.40	11212212111	0.40
14	11211111112	.34	11212212111	.37	11112211111	.34	21211111111	.50	11222111212	.51	21222111211	.34	21222111211	.33	11212212111	.40	11211111112	0.40
15	11212211112	.34	21222111211	.34	11212211112	.34	11112211111	.34	11112211111	.34	11212211112	.34	12112212122	.28	11212211112	.30	21222111211	0.30
16	11212212111	.34	11212211112	.34	11212212111	.34	11212211112	.34	11212211112	.34	12112212122	.26	11112222111	.23	12112212122	.30	11212211112	0.30
17	12112212122	.34	12112212122	.27	21222111211	.34	11212212111	.34	11212212111	.34	11112222111	.23	21221111212	.23	21222111211	.30	12112212122	0.30
18	21222111211	.34	11112222111	.23	21222111222	.34	12112212122	.34	12212211111	.34	21211111111	.23	11211212121	.21	11111111112	.20	11212121222	0.20
19	11112211111	.17	11211212121	.21	11111111112	.17	12212211111	.34	21222111211	.34	21221111212	.23	11111111112	.18	11112211111	.20	11211212121	0.20
20	11112211121	.17	11111111112	.19	11112212122	.17	21222111211	.34	21222111222	.34	11211212121	.21	11211111112	.18	11112222111	.20	21211111111	0.20
21	11112222111	.17	12212211111	.19	11112222111	.17	21222111222	.34	11111111112	.17	11111111112	.19	12212211111	.18	11211212121	.20	11111111112	0.20
22	11211111121	.17	12112111111	.18	11211212121	.17	11111111112	.17	11112111211	.17	12212211111	.19	11112211111	.17	12112111111	.20	12212211111	0.20
23	11211212121	.17	21222111222	.18	12112111111	.17	11112222111	.17	11211121111	.17	11211111112	.19	12112111111	.17	12212121111	.20	11211212111	0.20
24	11212121112	.17	11112211111	.17	12112211111	.17	11211212121	.17	11211212121	.17	21222111222	.18	21222111222	.17	12212211111	.20	12212121111	0.20
25	12112111111	.17	21112212121	.17	12112212122	.17	12112111111	.17	12112212122	.17	12212121111	.18	12211212122	.16	21112111211	.20	12112111111	0.20
26	11212111111	.17	22212212121	.17	12212211111	.17	21112111211	.17	12211111111	.17	12112111111	.18	21112111211	.16	21112212121	.20	21222111222	0.20
27	12212211111	.17	21212111111	.17	12212212122	.17	22221212222	.17	12211212122	.17	11112211111	.18	21212111111	.16	21211111111	.20	11112211111	0.20
28	21112111211	.17	11220111112	.17	21112111211	.17	22221212121	.17	12212121111	.17	12211212122	.17	21222112222	.16	21222111222	.20	21112111211	0.20
29	21112212121	.17	21112111211	.17	21112212121	.17			12212212122	.17	21211111211	.17	22212212121	.16	21222112222	.20	22221111212	0.20
30	21211111111	.17	21210111211	.17	21212111111	.17			21221111212	.17	11112211121	.16	22222121212	.16	22221212222	.20	21222112222	0.20
31	21212111111	.17	22221212222	.17	22212212121	.17					11212121112	.16	11212121112	.15	11112111112	.10	21112212121	0.20
32	22221212222	.17	22221212121	.17	22221212222	.17					12112211111	.15	12112211111	.14	11112212122	.10	12112211111	0.10
33	22222121212	.17	21222211222	.17	22222111212	.17					21212111111	.14	11112212122	.10	11212121112	.10	21212111111	0.10
34			12112211111	.15							11112212122	.14	21112212121	.10	12112211111	.10		
35			11212121112	.15							21211212121	.13			21212111111	.10		
36			11112212122	.13														

# Weitere Programme

---

- SAS Genetics: EM-Algorithmus, umfassendes kommerzielles Softwarepaket
- GERBIL: wahrscheinlichkeits-theoretisches Modell, EM-Algorithmus, Block-Identifizierung (Kimmel und Shamir 2005)
- SDPHapInfer: Parsimony-Modell, iterativer semi-definiter Optimierungsbasierter Approximations Algorithmus, MatLab-Tool (Huang et al. 2005)

# Ausblick

---

Es gibt keinen Algorithmus, der für alle Probleme die beste Lösung erreicht.

- Wenn theoretisch mehrere Haplotyp-Paare zu einem Genotyp kompatibel sind, so können Wahrscheinlichkeiten oder Scores angegeben werden.
- Verschiedene Verfahren können zu einem unterschiedlichen Ranking der möglichen Haplotyp-Paare führen.
- Durch die Anwendung mehrerer Programme kann die Interpretation der Ergebnisse erleichtert werden.
- Eine sichere Zuordnung eines Haplotyp-Paares bei mindestens zwei heterozygoten Varianten ist nur durch die Überprüfung im Labor möglich.

# Literatur

---

- 1 Bafna,V., Gusfield,D., Lancia,G. and Yooseph,S. Haplotyping as perfect phylogeny: a direct approach, *J.Comput.Biol.*, 10: 323-340, 2003.
- 2 Barrett,J.C., Fry,B., Maller,J. and Daly,M.J. Haploview: analysis and visualization of LD and haplotype maps, *Bioinformatics.*, 21: 263-265, 2005.
- 3 Cascorbi,I., Gerloff,T., Johne,A., Meisel,C., Hoffmeyer,S., Schwab,M., Schaeffeler,E., Eichelbaum,M., Brinkmann,U. and Roots,I. Frequency of single nucleotide polymorphisms in the P-glycoprotein drug transporter MDR1 gene in white subjects, *Clin.Pharmacol.Ther.*, 69: 169-174, 2001.
- 4 Chung,R.H. and Gusfield,D. Perfect phylogeny haplotyper: haplotype inferral using a tree model, *Bioinformatics*, 19: 780-781, 2003.
- 5 Diefenbach,K., Arjomand,N.F., Meisel,C., Fietze,I., Stangl,K., Roots,I. and Kopke,K. Systematic analysis of sequence variability of the Endothelin-1 gene: a prerequisite for association studies, *Genetic Testing*, 2006 (in press)
- 6 Excoffier,L. and Slatkin,M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population, *Mol.Biol.Evol.*, 12: 921-927, 1995.
- 7 Genprofile AG Verfahren und Vorrichtung zur Haplotypenvorhersage. DE 100 50 361 (11.10.2000); PCT/EP01/11726
- 8 Halperin,E. and Eskin,E. Haplotype reconstruction from genotype data using imperfect phylogeny, *Bioinformatics*, 20:1842-1849, 2004.
- 9 Li,S.S., Khalid,N., Carlson,C. and Zhao,L.P. Estimating haplotype frequencies and standard errors for multiple single nucleotide polymorphisms, *Biostatistics*, 4: 513-522, 2003.

# Literatur

---

- 10 Mwinyi,J., Kopke,K., schaefer,M., Roots,I. and Gerloff,T. Comparison of SLCO1B1 sequence variability between a German, Turkish, and African population, *Br.J.Clin.Pharmacol.*, 2006 (in press)
- 11 Niu,T., Qin,Z.S., Xu,X. and Liu,J.S. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms, *Am.J.Hum.Genet.*, 70: 157-169, 2002.
- 12 Qin,Z.S., Niu,T. and Liu,J.S. Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms, *Am.J.Hum.Genet.*, 71: 1242-1247, 2002.
- 13 Rohde,K. and Fuerst,R. Haplotyping and estimation of haplotype frequencies for closely linked biallelic multilocus genetic phenotypes including nuclear family information, *Hum.Mutat.*, 17: 289-295, 2001.
- 14 Stephens,M. and Donnelly,P. A comparison of bayesian methods for haplotype reconstruction from population genotype data, *Am.J.Hum.Genet.*, 73: 1162-1169, 2003.
- 15 Stephens,M., Smith,N.J. and Donnelly,P. A new statistical method for haplotype reconstruction from population data, *Am.J.Hum.Genet.*, 68: 978-989, 2001.
- 16 Zhang,K., Qin,Z.S., Liu,J.S., Chen,T., Waterman,M.S. and Sun,F. Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies, *Genome Res.*, 14: 908-916, 2004.
- 17 Zhao,J.H. 2LD, GENECOUNTING and HAP: computer programs for linkage disequilibrium analysis, *Bioinformatics*, 20: 1325-1326, 2004.
- 18 Zhao,J.H. and Sham,P.C. Faster haplotype frequency estimation using unrelated subjects, *Hum.Hered.*, 53: 36-41, 2002.

# Danksagung

---

Der GenProfile-Algorithmus wurde von Dr. Willi Schmidt, Berlin, entwickelt.

Diese Untersuchungen wurden unterstützt durch das BMBF im Rahmen des “Berlin Center for Genome Based Bioinformatics” (Projekt: 031U209B).

