medgen 2007 · 19:300–303
DOI 10.1007/s11825-007-0028-0
Online publiziert: 11. Juli 2007


**T. Becker**
Institut für Medizinische Biometrie, Informatik und Epidemiologie, Bonn

# Genidentifizierung bei multifaktoriellen Krankheiten

## Genetisch-epidemiologische Methoden

Aufgabe der genetischen Epidemiologie sind die Lokalisierung und Identifizierung von genetischen Faktoren, die kausal an der Entstehung von Krankheiten beteiligt sind. Viele beim Menschen verbreitete Krankheiten sind multifaktoriell, d. h. sie werden durch das Zusammenwirken verschiedener Gene und Umweltfaktoren verursacht.

Das grundsätzliche Problem bei der Identifizierung krankheitsverursachender DNA-Sequenz-Varianten ist die Größe des menschlichen Genoms, das etwa 3 Mrd. Basenpaare umfasst [6]. Obwohl nur ein Bruchteil des Genoms kodierend oder zumindest Teil eines Gens ist [6], können nur wenige Abschnitte des Genoms bei der Suche nach kausalen Varianten ausgeschlossen werden. In der Tat sind bis heute weder alle Gene des Menschen entdeckt noch ist die Regulation von Genen durch andere genomische Regionen zufrieden stellend erforscht [6]. Auf eine systematische Durchsuchung des gesamten Genoms kann deshalb nicht verzichtet werden.

Eine umfassende Untersuchung der biologischen Wirkungsweise des gesamten Genoms ist in der Praxis jedoch unmöglich. Da aber sowohl innerhalb von Familien als auch auf Bevölkerungsebene die Ausprägungen der DNA-Variation korreliert sind, ist es möglich, das Genom mit ausgewählten genetischen Markern repräsentativ abzubilden und diese mit statistischen Methoden zu untersuchen.

### Kopplungs-Assoziations-Paradigma

Innerhalb der genetischen Epidemiologie sind grundsätzlich 2 Analysearten zu unterscheiden:

- die Kopplungsanalyse und
- die Assoziationsanalyse.

### Kopplungsanalyse

Sie wird angewendet auf Stichproben von Familien, in denen die zu untersuchende Krankheit segregiert. Zur Analyse werden genetische Marker, DNA-Abschnitte mit mehreren Allelen und bekannter Lokalisation, genotypisiert. In einem so genannten Genomscan sind diese Marker äquidistant über das gesamte Genom verteilt. Für jeden Marker wird dabei die Kosegregation mit dem Phänotyp analysiert. Die Analyse kann entweder parametrisch oder nichtparametrisch („modellfrei“) erfolgen. Eine Übersicht zu Methoden der Kopplungsanalyse findet sich bei Ott [11].

Die klassische Methode der parametrischen Kopplungsanalyse ist die LOD-Score-Methode [10]. Hierzu müssen das Krankheitsmodell (Vererbungsgang) und die Marker-Genotyp-Verteilung bekannt sein. Die Rekombinationsfrequenz zwischen dem putativen Krankheitslocus und dem Marker wird mit $\Theta$ bezeichnet. Hierzu wird die Likelihood-Funktion $L(\Theta)$ betrachtet, die für jeden möglichen Wert von $\Theta$ die Wahrscheinlichkeit angibt, die Familiendaten (Krankheitsstatus und Genotypen des Markers) so zu beobachten, wie sie sind. Die Lod-Score-Funktion $Z(\Theta):=\log_{10}[L(\Theta)/L(0{,}5)]$ vergleicht die Likelihood-Funktion unter Kopplung, $\Theta<0{,}5$, mit der unter keiner Kopplung, $\Theta=0{,}5$. Die Lod-Score-Funktion $Z(\Theta)$ wird maximiert, um einen Schätzer $\Theta\hat{}$ für die Rekombinationsfrequenz zu erhalten. Ein zugehöriger Lod-Score $Z(\Theta\hat{}) \geq 3$ wird als signifikant angesehen. Ein entsprechend hoher Lod-Score bedeutet also einen statistisch signifikanten Zusammenhang zwischen dem Vererbungsmuster des Markers und dem der Krankheit, bei entsprechend niedrige Rekombinationsfrequenz zwischen Marker und putativem Krankheitslocus. Es kann also auf die Existenz einer krankheitsmitverursachenden Variante in der Nähe des Markers geschlossen werden.

Der Nachteil des Lod-Scores ist, dass das Vererbungsmodell der Krankheit vorgegeben werden muss. Bei multifaktoriellen Krankheiten ist er im Allgemeinen auch nicht durch Segregationsanalysen bestimmbar. Entsprechend müssen dann mehrere Modelle getestet werden. Dadurch verliert allerdings die Faustregel, einen Lod-Score $Z(\Theta) \geq 3$ als signifikant anzusehen, aufgrund des multiplen Testens ihre Gültigkeit. Anstatt einige Krankheitsmodelle auszuprobieren, ist es konsequent, auch das Krankheitsmodell durch Schätzung aus den Stammbaumdaten zu bestimmen. Dies führt zum so genannten Mod-Score [4], dessen zugehörige Signifikanzschwelle nur für bestimmte Familienstrukturen bekannt ist.

Aufgrund des in der Regel unbekannten Krankheitsmodells hat die nichtparametrische Kopplungsanalyse an Bedeutung gewonnen [5]. Sie testet auf Evidenz für Kopplung, ohne das Krankheits-

modell und die Rekombinationsfrequenz zu schätzen. Die Methoden basieren auf dem Identity-by-descent(IBD)-Status, der bei Geschwisterpaaren die Anzahl der Allele zählt, die von exakt der gleichen großväterlichen oder großmütterlichen Kopie eines parentalen Gens kommen.

Die genetischen Grundlagen der Kopplungsanalyse lassen sich wie folgt zusammenfassen: Große DNA-Teilabschnitte des Genoms werden von Generation zu Generation ohne Veränderungen weitergegeben. Jedoch treten auf jedem Chromosom pro Generation durchschnittlich 2–3 Rekombinationen auf. Die DNA-Sequenz eines Chromosoms eines Nachkommens besteht aus der Verbindung von Sequenzabschnitten, die alle für sich vollständig identisch sind mit der Sequenzfolge von analogen Teilen von parentalen Chromosomen. Aufgrund dieses Phänomens lässt sich aus einer beobachteten Kosegregation die Lokalisationsinformation gewinnen, weil zu erwarten ist, dass bei erkrankten Familienmitgliedern nicht nur die kausale Variante übereinstimmt, sondern auch ein weiter genomischer Bereich in seiner Umgebung. Gleichzeitig verhindert die Vererbung von großen, unveränderten Chromosomenabschnitten eine detaillierte Kartierung, weil als Folge dieses Phänomens eben nicht nur Marker in unmittelbarer Nähe des Krankheitslocus mitvererbt werden, sondern auch weiter entfernte Marker. Folglich ist mit der Kopplungsanalyse nur eine Grobkartierung möglich, die gefundene Kopplungsregion erstreckt sich im Allgemeinen über viele Megabasen, d. h., jede Sequenzvariation innerhalb dieser Region kommt in Frage, an der Ätiologie der Krankheit beteiligt zu sein. Die vergleichsweise niedrige Anzahl von Rekombinationen pro Generation hilft also einerseits dabei, das gesamte Genom mit wenigen Markern repräsentativ abbilden zu können, macht aber andererseits eine feinere Kartierung mit der Kopplungsanalyse unmöglich. Eine weiterführende Untersuchung der gefundenen Kopplungsregionen erfordert deshalb andere Mittel. Die Methode der Wahl ist hier die Assoziationsanalyse.

## Assoziationsanalyse

Die Assoziationsanalyse qualitativer Phänotypen (Krankheitsstatus) vergleicht die Häufigkeit von allelischen oder genotypischen Ausprägungen verschiedener Markersysteme zwischen 2 Gruppen. Sie betrachtet also Ausprägungshäufigkeiten auf Populationsebene, im Gegensatz zur Kopplungsanalyse, die die Ähnlichkeit von Vererbungsmustern bewertet. Bei Fall-Kontroll-Studien werden von der Krankheit betroffene (Fälle) mit nichtbetroffenen Personen (Kontrollen) verglichen. Bei einem Marker mit 2 Allelen kann die 2×3-Kontingenztafel der Genotypverteilung bei Fällen und Kontrollen mit der zugehörigen $\chi^2$-Verteilung mit 2 Freiheitsgraden getestet werden. Erwartet man einen Alleldosiseffekt, kann stattdessen der Armitage's Trend-Test [1] verwendet werden, der lediglich einen Freiheitsgrad benötigt. Neben den populationsbasierten Fall-Kontroll-Studien sind familienbasierte Assoziationsstudien von Bedeutung. Diese verwenden in der Regel einfache Familienstrukturen, z.B. Trios, also ein betroffenes Kind mit Eltern. Der prominenteste Vertreter der familienbasierten Assoziationsstudien ist der Transmission-Disäquilibrium-Test (TDT) [13]. Er vergleicht die Häufigkeit von zum betroffenen Kind transmittierten Allelen zur Häufigkeit der nichttransmittierten parentalen Allelen (so genannte Pseudokontrollen). Verwendet werden nur die Transmissionen bzw. Nichttransmissionen heterozygoter Elternteile. Es entsteht ein 1:1-Matching von Fällen und Pseudokontrollen. Der TDT wird dadurch robust gegenüber unerkannten Populationsstrukturen. Man erhält also keine falsch-positiven Befunde aufgrund von Stratifikation, und echte Assoziation wird nicht so leicht durch zusätzliche Stratifikation maskiert.

Die Assoziationsanalyse wird zum einen bei Kandidatengenen angewendet, also bei Genen, die aufgrund ihrer bekannten Funktion als putativ mitverursachend für die Krankheit angesehen werden können. Zum anderen wird die Assoziationsanalyse genutzt, um die durch die

medgen 2007 · 19:300–303 DOI 10.1007/s11825-007-0028-0


T. Becker

## Genidentifizierung bei multifaktoriellen Krankheiten. Genetisch-epidemiologische Methoden


### Zusammenfassung

Eine wichtige Aufgabe der genetischen Epidemiologie sind die Lokalisierung und Identifizierung von genetischen Faktoren, die an der Entstehung von Phänotypen beteiligt sind. Insbesondere multifaktorielle Krankheiten, die durch das Zusammenwirken verschiedener Gene und Umweltfaktoren verursacht werden, sind in den Fokus gerückt. Grundsätzliches Problem der genetischen Epidemiologie ist, dass das menschliche Genom aufgrund seines Umfangs eine erschöpfende biologische Untersuchung seiner Wirkungsweise nicht zulässt. Da aber sowohl innerhalb von Familien als auch auf Bevölkerungsebene die Ausprägungen der DNA-Variation korrelieren, ist es möglich, das Genom mit ausgewählten genetischen Markern repräsentativ abzubilden und diese mit statistischen Methoden zu analysieren. Eine Grobkartierung wird mittels Kopplungsanalyse durchgeführt, die die Kosegregation von Markern mit dem Phänotyp in Familien betrachtet. Die Feinkartierung der Kopplungsregionen erfolgt mit der Assoziationsanalyse, die die Häufigkeit von Allelen/Genotypen zwischen von der Krankheit Betroffenen und Nichtbetroffenen vergleicht. Das Kopplungsassoziationsparadigma wird zunehmend durch genomweite Assoziationsstudien (GWAS) ersetzt, die auf eine einleitende Kopplungsanalyse verzichten.



### Schlüsselwörter

Multifaktorielle Krankheiten · Kopplungsanalyse · Assoziationsanalyse · Haplotypen · SNP


## Identification of genes related to multifactorial diseases. Genetic-epidemiologic methods


### Abstract

It is an important goal of genetic epidemiology to localize and identify genetic factors that are involved in the development of a phenotype. In particular, multifactorial diseases have moved into focus. The basic problem of genetic epidemiology is that the biological function of the human genome cannot be comprehensively investigated because of its enormous size. There is, however, correlation of DNA variation both within families and on a population level. As a consequence, it is possible to represent the whole genome with selected genetic markers and to analyze them statistically. Mapping starts with linkage analysis, which considers cosegregation of markers and phenotype within families. Fine mapping of a linkage region is then left to association analysis, which compares allele or genotype frequencies between affected and unaffected probands. Nowadays, the linkage-association paradigm is often replaced by genome-wide association studies that do not rely on an initial linkage analysis.



### Keywords

Multifactorial diseases · Linkage analysis · Association analysis · Haplotypes · SNP


Kopplungsanalyse gewonnene Grobkartierung zu verfeinern. Da Kopplungsregionen sich immer noch über 10–20 Mio. Basenpaare erstrecken können, sind solche Regionen in der Praxis nicht erschöpfend analysierbar. Jedoch ist, in Analogie zur Kopplungsanalyse, auch bei der Assoziationsanalyse eine repräsentative Abbildung mit einer Auswahl genetischer Marker möglich. Verwendet werden hier die so genannten SNP („single nucleotide polymorphism"), die durch den Austausch eines einzigen Basenpaars charakterisiert sind. Diese diallelischen Marker finden sich in großer Häufigkeit und Dichte im menschlichen Genom und sind im Hochdurchsatz flächendeckend genotypisierbar.

Die Möglichkeit einer repräsentativen Markerauswahl beruht auf der These der Existenz so genannter Gründermutationen. Für eine spezielle, kausale Variante wird angenommen, dass diese ursprünglich vor vielen Generationen als Mutation bei einem einzigen Individuum aufgetreten ist und sich im Lauf der Generationen weiter verbreitet hat, und zwar zusammen mit dem DNA-Sequenz-Abschnitt, auf dem sie entstanden ist. Durch Rekombinationen wird im Lauf der Generationen der ursprüngliche Bereich um die Variante zerteilt, bleibt aber noch auf einem kurzen Stück erhalten. Solche DNA-Segmente lassen sich durch die Abfolge der allelischen Ausprägungen an den polymorphen Stellen (Markern) charakterisieren. Man spricht dann von „Haplotypen". Da der beschriebene Prozess für alle SNP stattfindet, sind die Allelausprägungen benachbarter Marker lokal nicht unabhängig voneinander. Man spricht vom Kopplungsungleichgewicht (LD, „linkage disequilibrium"). Dieses ist oft so stark, dass ein bestimmtes Allel eines Markers jeweils nur zusammen mit einem bestimmten Allel eines benachbarten SNP auf einem DNA-Strang auftritt (perfektes LD). Die lokale Struktur der genetischen Variation einer Population ist also intrinsisch in Haplotypen organisiert [3]. Entsprechend ist es nahe liegend, die Assoziationsanalyse basierend auf Haplotypen durchzuführen. Moderne Genotypisierungsverfahren liefern allerdings nur die Genotypen benachbarter Marker, ohne die entsprechende Phaseninformation. Haplotypre-

konstruktion bzw. -frequenzschätzung ist jedoch mit statistischen Verfahren möglich. Ein Überblick zu entsprechenden Methoden findet sich bei Becker u. Knapp [2]. Zum einen existieren Methoden, die im Rahmen eines Markov-Chain-Monte-Carlo(MCMC)-Ansatzes die Haplotypen einer Person rekonstruieren. Mit einem Maximum-Likelihood-Ansatz und dem Expectation-Maximization(EM)-Algorithmus kann man hingegen allen theoretisch möglich Haplotyperklärungen einer Person oder Familie ein Wahrscheinlichkeitsgewicht zuordnen. Weiterführende Testverfahren vergleichen dann die Halotypverteilung zwischen Fällen und (Pseudo)-Kontrollen [3].

Wie bei der Kopplungsanalyse eröffnet also die gemeinsame Vererbung von ganzen Chromosomenabschnitten die Möglichkeit, die genetische Variation in einer Region durch repräsentative Marker abzubilden. Da seit der Gründermutation viele Generationen vergangen sind, ist der Bereich, der gefunden wird, wesentlich kleiner als bei der Kopplungsanalyse, die nur wenige Generationen betrachtet. Wiederum in Analogie zur Kopplungsanalyse, verhindert jedoch die Korrelationsstruktur, dass die exakte Lokalisation des kausalen Markers bestimmt werden kann. Bei perfektem LD ist mit statistischen Methoden offenbar kein weiterer Fortschritt erzielbar. Es bleibt allerdings die Möglichkeit, Populationen anderer Ethnizität zu betrachten, die in der gefundenen Assoziationsregion möglicherweise eine LD-Struktur haben, die ein Feinstmapping zulässt.

## Genomweite Assoziationsstudien

Gemäß der Common-Disease-common-Variant-Hypothese [8] spielen hauptsächlich häufige Varianten in der Ätiologie multifaktorieller Krankheiten eine Rolle. Unter dieser Annahme hat die Kopplungsanalyse nur eine geringe Trennschärfe, sodass die Assoziationsanalyse ohne deren einleitenden Schritt durchgeführt werden muss [12]. In der Tat waren die Kopplungsanalysen der letzten Jahre nur mäßig erfolgreich, was ein Indiz für die Gültigkeit der Common-Disease-common-Variant-Hypothese sein könnte. Entsprechend werden immer häufiger genomweite Assoziationsstudien (GWAS) durchgeführt. Da nach wie vor aus Kostengründen nicht alle SNP genotypisiert werden können, ist eine im Sinn des LD repräsentative Auswahl an Markern wichtig. Die Daten des internationalen HapMap-Projekts [7] beinhalten genomweite Genotypisierungsdaten, die zur optimalen Auswahl von Markern für Assoziationsstudien verwendet werden können.

Die Strategie der GWAS hat den Nachteil, dass eine wesentlich höhere Korrektur für multiples Testen notwendig wird als unter einem ideal funktionierenden Kopplungs-Assoziations-Paradigma. Deshalb muss mit Stichproben mit mehreren tausend Personen gearbeitet werden. Zusätzlich sind GWAS als mehrstufige Siebverfahren angelegt. Mit Hilfe eines initialen Kollektivs werden Marker und Markerregionen zur Replikation im nächsten Schritt ausgewählt und an einem unabhängigen Kollektiv getestet. Je nach Phänotyp, werden weitere Kollektive benötigt werden, um die echt-positiven Marker herauszufiltern. Um den mehrstufigen Prozess zu optimieren, kann es sinnvoll sein, Informationen mit einzubeziehen, die über p-Werte zur Assoziation hinausgehen. Bei der Priorisierung der Markerauswahl für den nächsten Replikationsschritt kann z. B. die Zugehörigkeit zu einer Kopplungsregion, einem bekannten Stoffwechselweg oder die Nähe zu kodierenden Bereichen eine Rolle spielen.

## Ausblick

Die nächsten Jahre werden zeigen, ob GWAS zum erhofften Durchbruch bei der Identifizierung von Risikogenen komplexer Krankheiten führen. Viel versprechende Ergebnisse liegen bereits vor. So sind z. B. bei Typ-2-Diabetes mehrere GWAS publiziert und haben zu replizierten Befunden geführt [14]. Aufgrund der großen Datenmenge erfordert eine genomweite statistische Analyse unter Einbeziehung von mehreren Genregionen oder Umweltfaktoren einen hohen logistischen Aufwand. Eine entsprechende Strategie für 2 ungekoppelte Marker wurde von Marchini et al. [9] in einer Simulationsstudie untersucht und empfohlen. Entsprechende Anstrengungen, solche Analysen durchführbar zu machen, sollten unternommen werden, denn letztlich sind die Krankheitsmodelle multifaktorieller Krankheiten unbekannt, und es kann nicht ausgeschlossen werden, dass es, im Gegensatz zu Typ-2-Diabetes, zumindest für einigen Krankheiten keine Faktoren mit identifizierbarem Randeffekt gibt.

## Korrespondenzadresse

**Dr. T. Becker**
Institut für Medizinische Biometrie,
Informatik und Epidemiologie,
Sigmund-Freud-Straße 25, 53105 Bonn
Tim.Becker@ukb.uni-bonn.de

**Interessenkonflikt.** Der korrespondierende Autor gibt an, dass kein Interessenkonflikt besteht.

## Literatur

1. Armitage P (1955) Tests for linear trends in proportions and frequencies. Biometrics 11: 375–386
2. Becker T, Knapp M (2004) Maximum-Likelihood estimation of haplotype frequencies in nuclear families. Genet Epidemiol 27: 21–32
3. Clark AG (2004) The role of haplotypes in candidate gene studies. Genet Epidemiol 27:321–333
4. Clerget-Darpoux F, Bonaiti-Pellie C, Hochez J (1986) Effects of misspecifying genetic parameters in lod score analysis. Biometrics 42: 393–3999
5. Elston RC (1998) Methods of linkage analysis – and the assumptions underlying them. Am J Hum Genet 63: 931–934
6. Fischer EP (2002) Das Genom. Fischer Taschenbuch, Frankfurt am Main
7. International HapMap Consortium (2003) The International HapMap Project. Nature 426: 789–796
8. Lander ES (1996) The new genomics: global views of biology. Science 274: 536–539
9. Marchini J, Donnelly P, Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 37: 413–417
10. Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Gent 7: 277–318
11. Ott J (1999) Analysis of human genetic linkage. Johns Hopkins University Press, Baltimore
12. Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273: 1516–1517
13. Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52: 506–516
14. Steinthorsdottir V, Thorleifsson G, Reynisdottir I et al. (2007) A variant in CDKAl1 influences insulin response risk of type 2 diabetes. Nat Genet 39: 770–775