

# Kapitel 1

## Einleitung

Chronische Veränderungen im Herz-Kreislauf-System sind die häufigste Todesursache in vielen Industrienationen, d.h. in den Ländern, in denen gesellschaftliche und wirtschaftliche Entwicklungen Mangelernährung und Infektionskrankheiten eindämmen konnten (Murray & Lopez, 1997). Die Tendenz ist steigend. Ursachen sind u.a. erhöhter Blutdruck in den Arterien, Diabetes mellitus (Zuckerkrankheit) und Fettstoffwechselstörungen. Gemeinsame Endstruktur dieses Prozesses ist die Gefäßverkalkung (Arteriosklerose), die in fortgeschrittenem Stadium zu akuten Störungen der Blutversorgung wichtiger Organe durch Verengung oder Verstopfung (Thrombosierung) zuführender Arterien führen kann. Häufig wird dadurch ein akuter Herzinfarkt oder ein Schlaganfall ausgelöst.

Arteriosklerose entsteht über Jahre, sogar Jahrzehnte durch Schädigungen und Entzündungen des Endothels („response to injury“ - Hypothese von Ross et al. (1977)) oder Einlagerungen von Lipiden unterhalb des Endothels („response to retention“ - Hypothese von Williams & Tabas (1995)). Die chronische Funktionsstörung der Gefäßwand (Endothel) der blutführenden Arterien wird durch ein komplexes Zusammenspiel von Umwelt- und genetischen Faktoren ausgelöst.

Als wichtige Risikofaktoren haben sich individuelle Veränderungen im Stoffwechsel und Transport von sogenannten Lipoproteinen herausgestellt. Sie stellen die lebenswichtigen Transportformen von Lipiden im Organismus dar, da diese weder im Blutplasma noch in anderen Körperflüssigkeiten löslich sind. Diese Lipoproteine sind mikroskopisch kleine Mizellen, in denen Fette im Innern von einer lösungsvermittelnden Hülle umgeben sind. Der Stoffwechsel dieser Gebilde ist biochemisch komplex, jedoch hat sich gezeigt, dass ein erhöhter Cholesteringehalt in der LDL-Fraktion (Low Density Lipoprotein) ebenso wie ein sehr niedriger Cholesteringehalt in der HDL-Fraktion (High Density Lipoprotein) Risikofaktoren arteriosklerotischer Gefäßveränderungen sind (Hellige et al., 1995). Das zuständige staatliche Institut für Gesundheit der USA veröffentlicht regelmäßig Richtlinien zur Prävention und Behandlung erhöhter Cholesterinwerte, benennt Hauptrisikofaktoren und Risikoklassen

für quantitative Cholesterinwerte (NCEP, 2002).

Cholesterin ist in der Nahrung enthalten, die aus tierischen Quellen stammt und gelangt aus dem Darm in die Lymphe und von dort aus über die Leber in das Blut. Es kann aber auch in der Leber und in den meisten übrigen Körperzellen aus Vorstufen hergestellt werden. Cholesterin ist ein lebenswichtiger Bestandteil aller Zellmembranen. Schädlich ist nicht die Substanz selbst, sondern z.B. eine zu hohe Konzentration in der LDL-Fraktion des Blutplasmas oder eine erhöhte Anzahl von LDL-Partikeln aufgrund erhöhter Retentionszeit. Mit Medikamenten (z.B. Statinen) versucht man heute, Synthesewege des Cholesterins im Organismus zu hemmen und damit die Konzentration des LDL-Cholesterins zu senken und die des HDL-Cholesterins zu erhöhen. In epidemiologischen Studien konnte gezeigt werden, dass dadurch das Risiko eines Herzinfarktes oder Schlaganfalls gesenkt werden kann.

Die Konzentration an LDL-Cholesterin und HDL-Cholesterin im Blutplasma wird auf sehr komplexe Weise durch zahlreiche Faktoren beeinflusst, die sich zusammenfassend als Umwelt- und Lebensstilfaktoren kennzeichnen lassen. Einer der wichtigsten Umweltfaktoren ist zweifellos das Ernährungsverhalten. Es werden zunehmend stark fetthaltige Speisen mit hohem Kaloriengehalt bevorzugt. Diese belasten den Organismus durch den hohen Gehalt an Cholesterin, aber vor allem indirekt durch die in der Leber stattfindende Umwandlung überschüssiger Nahrungsbestandteile wie Zucker und Eiweiß in Neutralfett und Cholesterin. Kagan et al. (1974) verglichen Japaner in Japan, Hawaii und in den USA und zeigten einen progressiven Anstieg sowohl des Fettkonsums als auch der Serum-Cholesterinkonzentrationen und Herz-Kreislaufkrankungen.

Als ebenso wichtig hat sich die individuelle genetische Konstitution gezeigt (Botstein & Risch, 2003). Es verhält sich offenbar so, dass jeder Mensch eine genetisch fixierte und über das ganze Leben hinweg recht stabile Mittellage des Blutcholesterins hat, auf die sich dann die Lebensweise modifizierend ausprägt. Diese Mittellage lässt sich ermitteln, indem man den Cholesteringehalt unter definierten Standardbedingungen (morgens im nüchternen Zustand) bestimmt. Es hat sich herausgestellt, dass sie für Erwachsene mittleren Alters über viele Jahre hinweg relativ konstant bleibt, allerdings mit steigendem Lebensalter langsam ansteigt. Dieser Umstand trägt vermutlich zur Arterioskleroseneigung im höheren Lebensalter ( $> 50$  Jahre für Männer /  $> 65$  Jahre für Frauen) bei (Lusis, 2003).

Dass diese Mittellage durch eine genetische Komponente stark beeinflusst wird, hat man in statistischen Untersuchungen der Variabilität an verwandten und nicht verwandten Personen zeigen können, am eindrucksvollsten durch vergleichende Studien an eineiigen und zweieiigen Zwillingen (Knoblauch et al., 1997; Marenberg et al., 1994). Eineiige Zwillinge haben zu 100% die gleichen DNS-Varianten, zweieiige nur zu durchschnittlich 50%. Beide haben auch, wenn sie gemeinsam aufwachsen, ähnliche Umwelteinflüsse, speziell in der vorgeburtlichen und kindlichen Lebensphase. Die Befunde zeigten hohe Konkordanz der

Mittellage bei eineiigen und geringere Übereinstimmung bei zweieiigen Zwillingen. Mit hoher Konfidenz konnte man in den Untersuchungen auf genetische Ursachen schließen, wengleich nicht-genetische Faktoren eine ebenso bedeutsame Rolle spielen (Williams et al., 2005). Man hat als quantitatives Maß der erblichen Verursachung die sogenannte Heritabilität eingeführt. Sie gibt den Grad der Erbllichkeit eines Merkmals unter bestimmten Modellvoraussetzungen zwischen 0% und 100% an. Dieses Maß hat theoretische Schwachpunkte, wie sich auch in der vorliegenden Studie zeigen wird, nämlich vor allem, dass es in seiner einfachsten Form die Wechselwirkung zwischen konstitutionellen genetischen und variierenden Umwelteinflüssen nicht erfasst. Außerdem ist es ein relatives Maß, das in einer Population prozentual ansteigt, wenn der Lebensstil zu vergleichender Personen ähnlich ist und die genetischen Unterschiede deutlicher hervortreten. Familiäre Häufung allein kann den „Haushaltsfaktor“, d.h. familiäre Umwelteinflüsse wie Ernährung und Lebensbedingungen, von erblich bedingten Einflüssen nicht vollständig abgrenzen. Gleichwohl sind Heritabilitätswerte des Cholesterins zwischen 30 und 70% (Grundy et al., 2004; Higgins, 2000) ein deutliches Zeichen dafür, dass die individuelle genetische Ausstattung eine sehr wichtige Rolle spielt. Die gefundenen Heritabilitäten lassen andererseits auf einen bedeutenden umweltbedingten Anteil schließen.

Aber welches sind die genetischen Faktoren für die Mittellage des Cholesteringehalts? Aus klinischen Beobachtungen sowie biochemischen und molekulargenetischen Studien kennt man eine Reihe von genetischen Defekten, die zu schweren Veränderungen des Cholesteringehalts im Plasma führen. Eine solche Krankheit ist die familiäre Hypercholesterinämie (FH), bei der Mutationen im Gen für den LDL-Rezeptor ursächlich sind (Scriver et al., 2001). Individuen, die diese Mutationen tragen, erkranken schon in jungen Jahren an schwerer Arteriosklerose mit Herzinfarkten und Schlaganfällen und statistisch gesehen stark verminderter Lebenserwartung. Sie treten überwiegend in Genen auf, von denen man aus biochemischen Studien weiß, dass sie für Proteine des Cholesterinmetabolismus kodieren, also für Synthese und Abbau von Cholesterin und Lipoproteinen, für ihren Transport im Blut und die Aufnahme durch die Gewebe. Die meisten dieser Proteine sind heute gut erforscht. Insbesondere kennt man die kodierenden Gene und hat sie klonieren und sequenzieren können. In dieser Dissertation werden sie als „Kandidatengene“ bezeichnet, weil es auf Grund der Biochemie gesunder Organismen und Defektträger naheliegt, in ihnen nach Varianten zu suchen, die für Unterschiede der Mittellage zwischen Individuen der Normalbevölkerung verantwortlich sind. (Normalbevölkerung meint hier: Individuen ohne klinisch manifeste Dyslipidämie, die aus der Gesamtbevölkerung rekrutiert werden.)

Die hier vorgelegte Arbeit befasst sich mit dem Nachweis von genetischen Varianten in der Normalbevölkerung, die zur Variation des Cholesterinstatus ursächlich beitragen. Aus zahlreichen Beobachtungen ist bekannt, dass der Cholesterinstatus ein sogenanntes

komplex verursachtes Merkmal ist, für dessen Ausprägung die Wechselwirkung zwischen zahlreichen Umweltbedingungen und Lebensstilfaktoren mit einer nicht kleinen Anzahl von Genen bzw. Genprodukten verantwortlich ist. Das polygene Netzwerk von genetischen Faktoren umfasst vermutlich bis zu 30 Gene (Wang & Paigen, 2005). Eine Möglichkeit, ein solches Netzwerk von Faktoren zu verstehen, liegt in der mathematisch-statistischen Modellierung von breit angelegten Datenstrukturen, die genetische und phänotypische Merkmale von Individuen erfassen. Die hier vorgelegte Arbeit befasst sich mit der mathematischen Analyse eines Datenpools, der wichtige epidemiologische, klinische, biochemische und genetische Merkmale einer 1054 Individuen umfassenden Stichprobe der Franz-Volhard-Klinik in Berlin-Buch (Prof. F. Luft) sowie einer unabhängigen Stichprobe (1543 Personen) des Universitätshospitals in Genf (Prof. A. Morabia) enthält.

Das Hauptergebnis der vorliegenden Arbeit ist, dass sich die individuellen Unterschiede sowie familiäre Korrelationen im Cholesterinstatus zu einem bedeutenden Anteil durch Punktvariationen (SNPs) in der genomischen Struktur von 15 untersuchten Kandidatengenen erklären lassen. Die Daten zeigen auch, welche Gene besonders den LDL-Status und welche den HDL-Status beeinflussen. Es zeigt sich eine weitgehende Konkordanz der prinzipiellen Befunde in der Berliner und der Genfer Kohorte.

Die Forschungsergebnisse des gesamten interdisziplinären Teams wurden in mehreren Publikationen, in internationalen Zeitschriften und auf Kongressen vorgestellt (Bauerfeind et al., 2006, 2002; Knoblauch et al., 2002, 2004; Nürnberg et al., 2004). Eine solche systematische Studie über den Einfluss einer großen Zahl von Genen auf den normalen Stoffwechsel eines komplexen, medizinisch relevanten Merkmals ist eine Neuheit, da sich frühere Arbeiten stets mit Teilaspekten des Problemgebietes befassten.

Im folgenden Abschnitt werden Kennziffern des Lipoproteinstoffwechsels charakterisiert und deren Vernetzung in einem mathematisch-kinetischen Modell dargestellt.

## 1.1 Stoffwechsel der Lipoproteine im menschlichen Blutplasma

Das Serumprofil, insbesondere HDL, LDL und der Triglyzeridgehalt (TG) wird in der medizinischen Klinik regelmäßig bestimmt und für Diagnostik und Prognostik eines Patienten verwendet. Mit diesen Kennziffern, einschließlich dem Quotienten LDL/HDL (LH-Ratio), ist das Serumprofil für den Einzelnen charakteristisch und unter Standardbedingungen (u.a. hinsichtlich der aktuellen Nahrungsaufnahme) langfristig verlässlich. Bei genetischen Untersuchungen eignen sich Lipid-Konzentrationen als quantitatives phänotypisches Merkmal, in geeigneter Klassifikation und Kombination als qualitativer Phänotyp. Ob univariat oder

## 1.1. STOFFWECHSEL DER LIPOPROTEINE IM MENSCHLICHEN BLUTPLASMA 5

multivariat modelliert wird, ist dabei eine methodische Frage.

Der Lipoproteinstoffwechsel ist ein metabolisches System, Biochemie und (Patho-) Physiologie sind gut charakterisiert. Es besteht Klarheit über die wesentlichen Genprodukte, die Synthese, Verarbeitung und Abbau katalysieren. Knoblauch et al. (2000) haben ein mathematisch-kinetisches Modell entwickelt, das die Komplexität eines Teils des Fettstoffwechsels in vereinfachter Form, unter dem Einfluss ausgewählter Kandidatengene, abbildet (Abbildung 1.1).

Das Modell berücksichtigt eine bestimmte Anzahl von Metaboliten (Stoffwechselprodukte), die durch Reaktionen miteinander in Beziehung stehen. Die Metabolite sind die wesentlichen Lipoproteinklassen VLDL, IDL, LDL, HDL-Vorläufer, HDL (2+3), CM (Chylomikronen) und CM-Remnants (Reste). Sie lassen sich im Blut einzeln nachweisen. Die Reaktionen sind durch Bilanzgleichungen definiert. Mit diesem Modell ist es möglich, die Effekte unterschiedlicher Gen- bzw. Proteinaktivitäten im Lipidstoffwechsel zu analysieren.

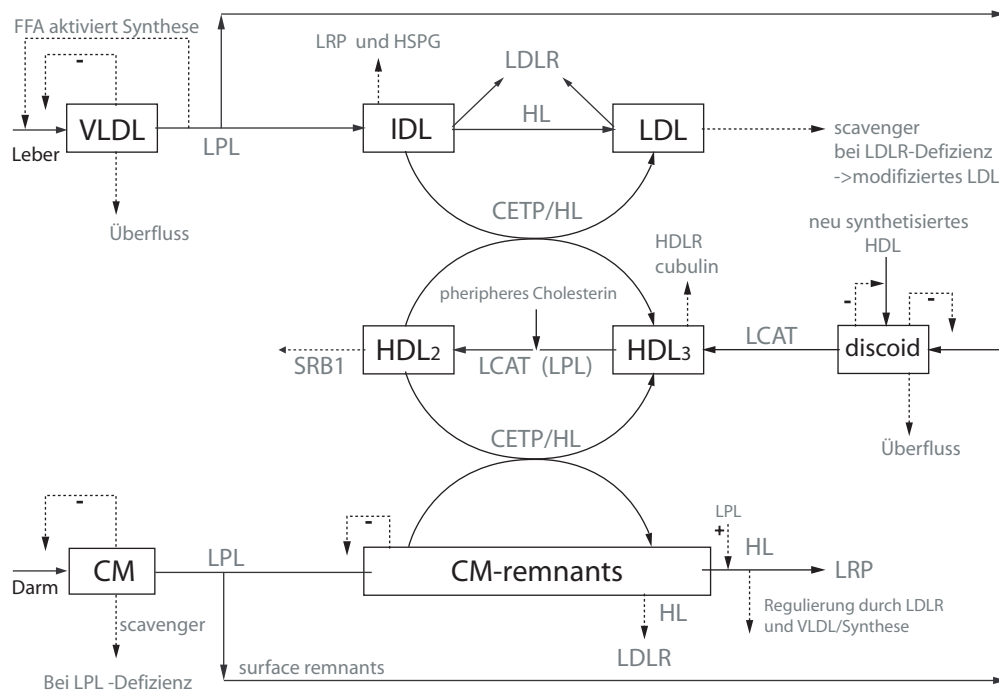


Abbildung 1.1: Kinetisches Modell des Lipoproteinstoffwechsels. Die Verbindungen zwischen Metaboliten sind durch stöchiometrische Gleichungen definiert (Pfeile); Metaboliten sind in Boxen dargestellt: VLDL, IDL, LDL, HDL, CM und CM-remnants. Beteiligte Genprodukte sind: LPL= Lipoprotein Lipase, LRP = LDL-Rezeptor-related Protein, HL = Hepatische Lipase, CETP = Cholesteryl Ester Transfer Protein und LCAT = Lecithin-Cholesteryl-Acyl-Transferase, HSPG = Heparan Sulfat Proteoglycan, SR-B1 = Scavenger Rezeptor. (-/+ ) bezeichnet die Regulation. Zu Einzelheiten siehe (Knoblauch et al., 2000).

Im Lipoproteinstoffwechsel werden ein exogenes, ein endogenes und ein reverses Cho-

lesterin-Transportsystem unterschieden. Der exogene Weg transportiert Lipide, die mit der Nahrung aufgenommen wurden, vom Darm zur Leber als zentralem Stoffwechsellorgan. Der endogene Weg wird von den Lipoproteinen hepatischen Ursprungs besprochen. Dabei werden die der Leber zugeführten und die neusynthetisierten Lipide der Körperperipherie angeboten. Exogenes und endogenes System beginnen mit der Absonderung triglyzeridreicher Lipoproteine. In beiden Systemen werden die Triglyzeride durch das Enzym endothelständige Lipoproteinlipase hydrolysiert. Es entstehen Lipoproteine mit hohem Gehalt an Cholesterinestern, die Chylomikronen-Remnants und LDL-Partikel. Sie werden durch LDL-Rezeptor-vermittelte Endozytose aufgenommen, dann abgebaut. Der Cholesterinrücktransport (reverse) dient dem Transport von peripher synthetisiertem Cholesterin zur Leber. Nur über die Leber kann Cholesterin über die Bildung von Gallensäuren in größeren Mengen ausgeschieden werden. Am reversen Cholesterintransport sind der Lipoproteinrezeptor HDL, das Enzym LCAT, das Transportprotein CETP, sowie ABCA1 und SR-B1 beteiligt.

Die Dynamik des Lipoproteinstoffwechsels wird durch die Aktivität von Enzymen und Rezeptoren im Blutplasma entscheidend bestimmt. Die Aktivität der Enzyme, Rezeptoren und Transportproteine, die an den Reaktionen teilnehmen, sind wiederum abhängig von allelischen Varianten des für das Protein kodierenden Gens.

## 1.2 Kandidatengene, Funktion und Defekte

Obwohl die übliche Form der Arteriosklerose multifaktoriell ist, haben Mutationsanalysen in Genen bei seltenen monogenen Formen von Fettstoffwechselstörungen den bislang größten Beitrag zu einem ätiologischen und pathophysiologischen Verständnis des Fettstoffwechsels beigetragen. Diese Mutationen, die zum völligen Fehlen eines Proteins oder zur Expression funktionsdefekter Genprodukte führen, haben nachhaltige Auswirkungen auf Konzentrationen und Verteilung von Lipoproteinen und verursachen auch klinische Symptome.

Solche monogenen Formen der Fettstoffwechselstörungen sind in (Nabel, 2003; Scriver et al., 2001) zusammengestellt. Zu monogenen Erkrankungen, die erhöhte LDL-Konzentrationen im Blutplasma durch eingeschränkte Aktivität des hepatischen LDL-Rezeptors zur Folge haben, gehört die einleitend erwähnte familiäre Hypercholesterinämie (FH), verursacht durch Mutationen im *LDLR*-Gen. Mehr als 600 Mutationen im *LDLR*-Gen wurden in FH-Patienten identifiziert. Unter 500 Personen befindet sich im Durchschnitt ein heterozygoter Träger für mindestens eine dieser Varianten, d.h. es tritt nur ein mutiertes Allel an dieser Position auf. Homozygote Träger (beide Allele mutiert) finden sich im Durchschnitt einmal in 1 Mio. Individuen. Träger der heterozygoten Variante produzieren nur die Hälfte der normalen Anzahl an LDL-Rezeptoren, was zu zwei- bis dreifach erhöhten

LDL-Konzentrationen im Plasma führt. Homozygote Personen haben schwere koronare Arteriosklerose und sterben häufig im Kindesalter an einem Herzinfarkt.

Eine weitere Form der Hypercholesterinämie wird durch eine Mutation im *APOB100*-Gen verursacht. Dabei wird die Bindungsfähigkeit des Apolipoprotein B-100 an den LDL-Rezeptor vermindert und damit die Aufnahme von LDL in die Zelle verlangsamt. Dieser Defekt tritt mit der Häufigkeit 1:1000 in der heterozygoten Form auf, und die Folgen entsprechen denen der heterozygoten FH-Patienten.

Die autosomal rezessive Hypercholesterinämie ist sehr selten ( $< 1:10$  Mio). Die molekulare Ursache sind Defekte in einem hepatischen Adapter-Protein, das die Aktivität des LDL-Rezeptors bei der Freigabe von LDL stört. Mutationen im *ARH*-Gen erhöhen die LDL-Konzentrationen im Blutplasma auf Werte ähnlich denen homozygoter FH-Patienten.

Studien dieser monogenen Fehlfunktionen, die die Funktion des LDL-Rezeptors stören, haben dazu beigetragen, die Bedeutung der Cholesterinsynthese und -ausscheidung in der Leber aufzuklären.

Dass die Kandidatengene lebenswichtig sind, sagt noch nichts darüber aus, welche Gene und genetischen Varianten für die individuelle Variation der Lipidkonzentrationen, die jenseits der oben beschriebenen Extremformen lebensgeschichtlich zu Krankheit oder Nicht-Krankheit disponiert, mitbedingend sind. Schließlich haben nur wenige Menschen die erwähnten schweren Defekte. Es gibt jedoch Hinweise, dass auch die Variation in der Gesamtbevölkerung genetisch bedingt ist.

Im folgenden Abschnitt sind Hauptfunktionen ausgewählter Kandidatengene des Lipoproteinstoffwechsels und bekannte Effekte genetischer Varianten auf Lipid-Konzentrationen zusammengefasst.

**Apolipoprotein B** ist das bedeutendste Apolipoprotein der Chylomikronen und des LDL.

Es agiert als Transporter für Triglyzeride und Cholesterin zwischen der Leber und peripherem Gewebe. Bentzen et al. (2002) sowie Imonen et al. (1998) konnten Effekte genetischer Marker im *APOB*-Gen auf Cholesterin- und Triglyzerid-Konzentrationen nachweisen.

**Apolipoprotein E** vermittelt die Aufnahme von Remnants der Chylomikronen in die Leber. Es reguliert den Abbau von VLDL-remnants und die Umwandlung zu LDL-Cholesterin. Polymorphismen im *APOE*-Gen beeinflussen den Abbau von remnants und wirken sich auf die Konzentration des LDL aus. ApoE ist ein Protein, das in verschiedenen Formen auftritt. Diese resultieren aus drei Allelen an einem Gen-Lokus. Die drei Isoformen *APOE- $\epsilon_4$* , *APOE- $\epsilon_3$*  und *APOE- $\epsilon_2$* , die sich nur durch Substitution einer Aminosäure unterscheiden, haben funktionelle Konsequenzen in zellulärer und molekularer Hinsicht. *APOE- $\epsilon_3$*  scheint bzgl. aller bekannten Funktionen die nor-

male Isoform zu sein. *APOE- $\epsilon_2$*  und *APOE- $\epsilon_4$*  zeigen Fehlfunktionen. *APOE- $\epsilon_2$*  ist assoziiert mit Typ-III Hyperlipoproteinämie, *APOE- $\epsilon_4$*  u.a. mit der Alzheimerschen Krankheit (Morbus Alzheimer) (Bickeböller et al., 1997). Beide sind Risikofaktoren für Arteriosklerose und zeigen starke Effekte auf LDL- und HDL-Konzentrationen im Blutplasma (Hagberg et al., 2000; Mahley & Rall, 2000).

**ABCA1** mit Cholesterin als Substrat ist in die Übertragung von Lipiden aus peripheren Zellen auf HDL-Partikel im reversen Cholesterintransport involviert. Mutationen im *ABCA1*-Gen sind für niedrige HDL-Konzentrationen infolge eingeschränktem Cholesterintransport verantwortlich (Probst et al., 2004). Die Ergebnisse der Analyse häufiger genetischer Varianten sind zum Teil kontrovers (Brousseau et al., 2001; Clee et al., 2001; Wang et al., 2000). Cohen et al. (2004) zeigen Effekte seltener funktioneller Varianten auf HDL-Konzentrationen.

**Cholesterinester Transferprotein (CETP)** ist ein Plasma-Glykoprotein, das für den Austausch von Cholesterinestern und Triglyzeriden zwischen Apo-B enthaltenden Chylomikronen, VLDL und LDL und Apo-A enthaltenden HDL wesentlich ist. Diese Stoffwechselfunktion ist wichtig für den Transport von der Leber und dem Magen-Darm-Trakt in die Peripherie und für den reversen Cholesterintransport aus der Peripherie zur Leber. Bei Mangel an CETP bestehen die HDL überwiegend aus HDL<sub>2</sub> und aus großen Apo-E-reichen Lipoproteinen. Sie sind reich an Cholesterinestern (CE) und enthalten wenig Triglyzeride. Die mit CE angereicherten HDL<sub>2</sub> zeigen Fehlfunktionen. Bei LDL dagegen ist der Gehalt an Cholesterinestern gering und der Gehalt an Triglyzeriden hoch. Ihre Bindung an den LDL-Rezeptor ist vermindert. Eine relativ seltene Mutation im CETP-Promotor vermindert die Transkription und ist mit Hyperalphalipoproteinämie assoziiert, eine Folge des beeinträchtigten reversen Cholesterintransports (Nagano et al., 2001). Verschiedene Studien haben einzelne CETP-Mutationen analysiert und einen Einfluss auf HDL und/oder LDL festgestellt (Agerholm-Larsen et al., 2000; Dachet et al., 2000; Freeman et al., 1994; Gudnason et al., 1999; Hsu et al., 2002; Kuivenhoven et al., 1997). Boekholdt et al. (2004) stellen in einer Übersichtsarbeit Assoziationen zwischen genetischen Varianten im *CETP*-Gen und veränderten *CETP*-Plasma-Werten und -Aktivität, HDL-Konzentrationen, LDL- und HDL-Partikelgröße zusammen. LeGoff et al. (2004, 2002) zeigen, dass Mutationen im *CETP*-Gen mit einem CETP-Mangel assoziiert sind, der durch einen hohen HDL-Spiegel und reduziertem kardiovaskulärem Risiko charakterisiert ist.

**Lecithin Cholesterin Acyltransferase (LCAT)** ist ein Enzym zur Veresterung von freiem Cholesterin für den Cholesterintransport, die hauptsächlich im HDL stattfindet. Es wird in der Leber synthetisiert und ist im Blutplasma und in der extrazellulä-



ren Flüssigkeit zu finden. Die Folge einer Mutation im *LCAT*-Gen ist ein Mangel an funktionsfähiger LCAT. Recalde et al. (2002) haben einen Zusammenhang mit Hypoalphalipoproteinämie (HALP) nachweisen können. Weitere Arbeiten zeigen Assoziationseffekte einzelner *LCAT*-SNPs mit HDL-Konzentrationen (Miettinen et al., 1998; Stocks et al., 2000).

**Hepatische Triglyzeridlipase (HL)** hat, von der Leber synthetisiert, die Funktion der Hydrolyse von Triglyzeriden. Des Weiteren wirkt das Enzym als Kofaktor für die rezeptorvermittelte Aufnahme von Lipoproteinen. Kontrovers diskutiert wird, ob das Enzym HL die Arteriosklerose eher beschleunigt oder hemmt (Jansen, 2004; Jansen et al., 2002). Deeb et al. (2003) und Zambon et al. (2003) stellen insbesondere die Bedeutung eines Polymorphismus im Promotor des Gens für HL (*LIPC*) heraus. Mit seinem Einfluss auf die Synthese und Aktivität des Proteins wird das individuelle Risiko entscheidend beeinflusst. Hohe HL-Aktivität ist mit hoher LDL-Partikeldichte und mit reduzierter HDL<sub>2</sub>-Konzentration assoziiert.

**Lipoproteinlipase (LPL)** spielt eine zentrale Rolle bei der Hydrolyse von Triglyzeriden und Phospholipiden aus CM und VLDL sowie als Kofaktor der rezeptorvermittelten Aufnahme von Lipoproteinen. LPL sorgt für niedrige TG- und LDL-, sowie höhere HDL-Konzentrationen. Aus diesem Grund ist LPL potentiell atheroprotektiv (Morbida et al., 2003a). Viele Studien haben Effekte einzelner SNPs auf Lipidkonzentrationen nachweisen können (Fisher et al., 1997; Larson et al., 1999).

**LDL-Rezeptor (LDLR)** ist wesentlich an der Regulation der Konzentration des freien Cholesterins im Blutplasma beteiligt. Borthright et al. (1998) haben Kopplung und Assoziation von *LDLR*-Varianten und Serum-LDL-Konzentrationen nachweisen können.

**Apolipoprotein C-II** aktiviert das Enzym Lipoproteinlipase, das Triglyzeride hydrolysiert und damit den Zellen freie Fettsäuren zur Verfügung stellt (Fojo et al., 1986). Eine Assoziation mit Lipidphänotypen konnte bislang nicht nachgewiesen werden.

**Apolipoproteine A-V, A-IV, C-III, A-I:** Die kodierenden Gene liegen in einem gemeinsamen Gencluster auf dem langen Arm des Chromosoms 11. Man kennt Apo-A-I-Mangelsyndrome mit völligem Fehlen von Apo-A-I und Apo-A-I-Strukturvarianten, die geringe Auswirkungen auf den Lipoproteinstoffwechsel haben. Beim Apo-A-I/-C-III-Mangel Typ I unterbricht eine große Inversion die Gene für Apo-A-I und -C-III. Dem Apo-A-V/-C-III/-A-IV-Mangel (Apo-A-I/-C-III-Mangel Typ II) liegt eine Deletion des gesamten Genkomplexes zugrunde. Ein gemeinsames klinisches Merkmal ist u.a. frühzeitige Arteriosklerose (Kostner & März, 2001). Apolipoproteine A-IV und A-I haben ihren spezifischen Funktionsbereich in der LCAT-Aktivierung und

HDL-Rezeptorbindung. Sie sind den Lipoproteinen HDL und den Chylomikronen zugeordnet. Die Bedeutung des Apo-A-IV-Polymorphismus für den Lipidstoffwechsel ist noch nicht eindeutig charakterisiert. Neben dem Apo-A-I/-C-III-Mangel Typ I und Typ II sind zahlreiche natürliche Apo-A-I-Strukturvarianten bekannt. Apolipoprotein C-III blockiert die Enzyme Lipoproteinlipase und die Hepatische Lipase und damit den Abbau triglyzeridreicher Lipoproteinpartikel. Zahlreiche Studien zeigen Assoziationseffekte von SNPs im *APOA5* und *APOC3*-Gen auf TG-Konzentrationen (Minihane et al., 2002; Pennacchio et al., 2001).

Talmud et al. (2002) und Olivieri et al. (2003) untersuchten Haplotypen und fanden *APOA5*-Risikohaplotypen, die TG unabhängig von *APOC3* beeinflussen. Eine Variante in der intergenischen Region zwischen Apo-C-III und Apo-A-IV ist mit HDL assoziiert, SNPs im *APOA4* zeigen Effekte auf LDL, HDL und Triglyzeride. Ergebnisse aus Assoziationsuntersuchungen von SNPs im *APOA1*-Gen mit HDL-Konzentrationen sind kontrovers (Groenendijk et al., 2001).

In neuerer Zeit konzentrierten sich Studien genetischer Variation auf SNPs (Single Nucleotide Polymorphisms), d.h. DNS-Varianten, die mit modernen Methoden der Genomanalyse einfach und zunehmend kostengünstiger erfasst werden können. In dieser Arbeit liegt der Fokus auf der Analyse von SNPs und SNP-Haplotypen. In welcher Form der Zusammenhang eines polymorphen genetischen Markers mit einem phänotypischen Merkmal betrachtet werden kann, wird in dem nächsten Abschnitt deutlich.

## 1.3 Genetische Assoziationsstudien

### Assoziation und Kopplung

Der Zusammenhang zwischen einem genetischen Marker und einem phänotypischen Merkmal drückt sich durch (phänotypische) Assoziation und/oder Kopplung aus. Assoziation unterscheidet sich von Kopplung dadurch, dass in einer Population ein konkretes Allel oder mehrere Allele mit einem Merkmal vergesellschaftet sind. Es wird die Häufigkeit von Allelen zwischen Merkmalsträgern und Nicht-Merkmalsträgern verglichen, wobei Merkmale auch reellwertig (quantitativ) definiert sein können. (Der allgemein verwendete Begriff Assoziation bezieht sich im Weiteren auf die phänotypische Assoziation, solange nicht explizit von „allelischer Assoziation“ (siehe unten bzw. Glossar) gesprochen wird).

Kopplung liegt vor, wenn die jeweils an zwei verschiedenen Positionen auf einem Chromosom vorhandenen Allele gemeinsam (auf dasselbe Chromosom) weitergegeben, d.h. gekoppelt vererbt, werden. Ein Maß für Kopplung ist der genetische Abstand zwischen zwei Loci, ausgedrückt als Rekombinationswahrscheinlichkeit ( $\theta$ ). Kopplung (Linkage) ist ein

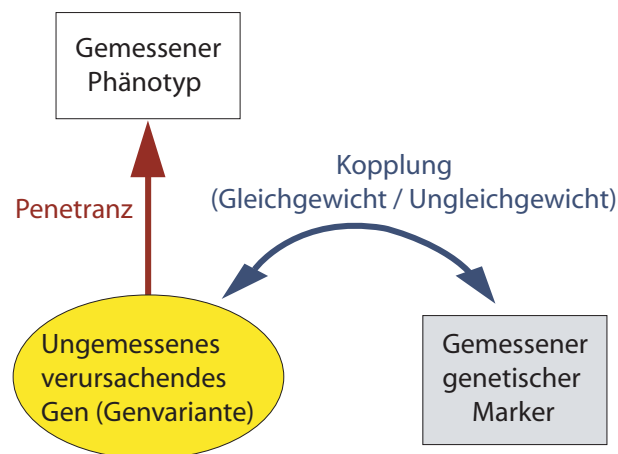


Abbildung 1.2: Beziehung zwischen gemessenem Phänotyp, ungemessenem verursachendem Locus und nahegelegenen gemessenen genetischen Marker (SNP, Haplotyp, Mikrosatellitenmarker). Der Zusammenhang zwischen Phänotyp und genetischem Marker ist abhängig von der Wirkung (Penetranz) des unbekannt funktionellen Locus auf den Phänotyp und dessen Beziehung zu dem gemessenen genetischen Marker. Darstellung aus (Province, 2002).

statistisches Maß, das sich auf eine Bevölkerung bezieht. Zwei Loci sind gekoppelt, wenn die Rekombinationshäufigkeit in einer Population deutlich kleiner ist als 50%.

Phänotypische Assoziation kann funktionelle (ursächliche) Allele direkt analysieren. Lassen sich funktionelle Allele nicht identifizieren, so nimmt man benachbarte (funktionsneutrale) genetische Marker, die (so wird angenommen) mit dem verursachenden Allel stark gekoppelt sind. Dieser indirekte Ansatz basiert auf dem Kopplungsungleichgewicht (LD = „Linkage Disequilibrium“) zwischen ungemessener (funktioneller) und gemessener (benachbarter) Variante.

Die Stärke phänotypischer Assoziation ist somit abhängig von der Stärke des LDs zwischen gemessenem und ungemessenem Marker und der Merkmalsausprägung (Penetranz) im gemessenen Phänotyp (Abbildung 1.2). Als genetische Marker können SNP-Diplotypen, SNP-Haplotypen, Mikrosatellitenmarker etc. betrachtet werden. Die Power von Assoziationsstudien ist abhängig von dem allelischen Spektrum einer Krankheit und von der Frequenz krankheitsverursachender Allele. Nur bei geringer allelischer Heterogenität und moderaten Frequenzen verursachender Allele ist es möglich, phänotypische Assoziation über Varianten im LD nachzuweisen.

Bevor Merkmale des LDs betrachtet werden, widmet sich der folgende Abschnitt populationsgenetischen Betrachtungen, Schätzungen allelischer Vielfalt und der kontrovers diskutierten „Common Disease / Common Variants (CD/CV)“-Hypothese komplexer Erkrankungen.

## Allelisches Spektrum komplexer Erkrankungen

Das Studium von Assoziation basiert auf der Bestimmung einer erhöhten Frequenz krankheitsverursachender Allele bei Merkmalsträgern. Das allelische Spektrum phänotypischer Merkmale hat bedeutende Konsequenzen sowohl für die Forschung als auch für die klinische Praxis. Klinische Charakterisierung und Design diagnostischer und therapeutischer Eingriffe sind erheblich leichter, wenn das allelische Spektrum klein ist, d.h. wenn wenige krankheitsverursachende Allele einen großen Anteil der genetischen Variabilität der untersuchten Krankheit erklären. Ein breites allelisches Spektrum hingegen bedeutet, dass viele Allele mit jeweils geringem Anteil eine Krankheit beeinflussen (Reich & Lander, 2001).

Die vorhandene genetische (Sequenz-) Variation hat einen natürlichen Ursprung, sie ist durch Mutation entstanden und hat biologische, demographische und historische Prozesse durchlaufen. Sie ist durch zufällige und natürliche Selektion sowie die Lebensumstände und Migration unserer Vorfahren gestaltet. Dieser Prozess ist begleitet durch das Aussterben vieler Varianten - nur ein kleiner Anteil der ursprünglichen Variation ist heute geblieben und beeinflusst phänotypische Muster und Verteilungen von Krankheiten oder Merkmalen (Chakravarti, 1999).

Sequenzvariation existiert in allen geographischen Bereichen, Unterscheidungen zwischen verschiedenen Populationen gibt es nur bzgl. deren jeweiliger Frequenz. Durch vergleichende Studien genetischer Vielfalt innerhalb und zwischen Populationen wurde herausgefunden, so berichtet Chakravarti (1999), dass mehr als 90% der genetischen Variation innerhalb der Populationen besteht – nur mit einem geringen Prozentsatz unterscheiden sich Populationen untereinander. Afrikanische Bevölkerungsgruppen zeigen größere Vielfalt im Gegensatz zu Gruppen aus Asien, Europa oder Amerika. Analysen bestehender genetischer Variation unterstützen zunehmend einen einzelnen Ursprung des Menschen („out of Africa“) im Gegensatz zu einem multi-regionalen Szenario humaner Evolution.

Die Daten legen eine schwerwiegende langanhaltende Populationsverdünnung (Flaschenhalseffekt (engl. „Bottleneck“) vor rund 100.000 Jahren (Mitte bis Ende der letzten Eiszeit (Weichsel-Kaltzeit)) nahe, durch den unsere Vorfahren auf eine überlebensfähige bzw. fortpflanzungsfähige (effektive) Populationsgröße  $N_e \approx 10.000$  reduziert wurden. Diese Schätzung beruht auf übereinstimmenden genetischen und archäologischen Untersuchungen, dargestellt in (Harpending et al., 1998; Takahata & Satta, 1997).

Ein Maß allelischer Vielfalt ist die Wahrscheinlichkeit, dass zwei zufällig gezogene Allele identisch sind („erwartete allelische Identität“  $\phi$ ). Eine klassische populationsgenetische Formel gibt  $\phi$  an einem neutralen Locus (keine Selektion) in einer stabilen (konstante

Größe über eine lange Zeit) und zufällig sich mischenden Population mit

$$\phi = \frac{1}{1 + 4N_e\mu_t} \quad (1.1)$$

an, wobei  $N_e$  die effektive Populationsgröße und  $\mu_t$  die neutrale Mutationsrate bezeichnen (Kimura & Crow, 1964).  $4N_e$  ist die Zeit (in Generationen), die ein (selektiv) neutrales Allel mit einer Anfangsfrequenz von  $p = 1/2N_e$  braucht, um sich in der Population zu fixieren ( $p = 1$ ) (Hartl, 1999).

Eine zentrale Frage, die die Populationsgenetik untersucht, ist die, wodurch und in welcher Weise die genomische (allelische) Vielfalt bestehen bleibt. Es gibt starke Hinweise, dass schädliche Mutationen in jeder Population schnell eliminiert werden. Dazu gehören Mutationen, die zu monogenen Erkrankungen führen, d.h. Krankheiten, die den Mendelschen Vererbungsgesetzen obliegen (engl. „Mendelian disease“). Solche Mutationen existieren mit sehr geringer Frequenz als Konsequenz neuer Mutationen. Es entsteht ein Gleichgewicht zwischen Mutation und Selektion. Somit steigt die Anzahl verschiedener Polymorphismen an. Natürliche Selektion ist einer dieser Mechanismen, dessen Ausmaß jedoch weitgehend unbekannt ist.

Wie sieht die genetische Variation aus, die komplexen Erkrankungen unterliegt?

Für die mehr als 100 identifizierten Gene, die mit monogenen Erkrankungen assoziiert sind, scheint die Antwort klar: Die allelische Vielfalt an jedem Locus ist hoch und jede Mutation ist selten (entstanden in der jüngeren Vergangenheit vor nicht mehr als 80 Generationen). Das Vorliegen einer Mutation ist ein notwendiger Faktor für die Verursachung des Phänotyps, es sei denn Heterogenität liegt vor.

Im Gegensatz dazu wird vermutet, dass für komplexe Krankheiten mehrere (u.U. viele) Mutationen ursächlich wirken. Bei komplexen Erkrankungen stellt das Fehlen eines Mendelschen Vererbungsmusters in den meisten Familien ein Argument gegen einen monogenen Effekt dar. Vielmehr legt es einen starken umweltbezogenen Einfluss und/oder die simultane Beteiligung mehrerer Gene nahe (Chakravarti, 1999).

Eine verbreitete Hypothese über die allelische Struktur häufiger, komplexer Erkrankungen besagt, dass es ein oder nur wenige vorherrschende krankheitsverursachende Allele an jedem der beteiligten Krankheitsorte (-loci) gibt („große allelische Identität“). Die Hypothese geht von alten Allelen, 400 Generationen (10.000 Jahre bei 25 Jahren pro Generation) oder mehr, mit relativ hoher Frequenz ( $> 1\%$ ) aus. Diese Hypothese ist als „Common Disease / Common Variants (CD/CV)“ - Hypothese bekannt (Chakravarti, 1999; Lander, 1996).

Die extreme Alternative zur CD/CV-Hypothese ist die „Allelische Heterogenitäts“ - Hypothese, die besagt, dass in einer Population verschiedene seltene Allele eines Locus

oder mehrerer Loci krankheitsverursachend sind. Dies ist denkbar, wenn es multiple seltene Varianten relativ neuen Ursprungs gibt. In diesem Fall ist es schwierig, funktionelle Effekte anhand von LD nachzuweisen, da jede neue Mutation auf dem Hintergrund unabhängiger Haplotypen entsteht, so dass zahlreiche statistisch schwache Signale resultieren.

Die Beschaffenheit allelischer Heterogenität hinsichtlich komplexer Erkrankungen ist weitgehend ungeklärt. Reich & Lander (2001) benutzen ein deterministisches Modell für die Anzahl krankheitsverursachender Allele, während Pritchard (2001) ein stochastisches Modell ansetzte.

Es gibt drei wesentliche Parameter in den Modellen: die Mutationsrate von neutralen in krankheitsverursachende Allele  $S$  ( $\mu_S$ ), („Rückmutation“) von krankheitsverursachenden in neutrale Allele  $N$  ( $\mu_N$ , die sog. Reparatur-Rate) und die Stärke der Selektion gegen verursachende Allele. Die Mutationsrate ist ein kritischer Parameter, höhere Raten führen zu größerer Vielfalt (siehe Gleichung 1.1). Reich & Lander (2001) gehen von Mutationsraten im Bereich  $10^{-7}$  bis  $10^{-4}$  pro Generation aus, und setzen in ihren Modellen das geometrische Mittel  $3.2 \times 10^{-6}$  gleichermaßen für  $\mu_S$  und  $\mu_N$  an. Pritchard (2001) hingegen bevorzugen höhere Mutationsraten für krankheitsverursachende Allele. Ihre Schätzungen umfassen den Bereich von  $2.5 \times 10^{-6}$  bis  $1.3 \times 10^{-4}$ . Die Annahmen gehen von Schätzungen aus Studien monogener Erkrankungen aus, u.a. zeigen Li & Sadler (1991), dass nichtkodierende Polymorphismen häufiger sind als kodierende und synonym kodierende häufiger sind als nicht-synonym kodierende. Mutationsraten neutraler Allele schätzen Pritchard im Bereich  $2.5 \times 10^{-8}$  bis  $2.5 \times 10^{-7}$ .

Bei vielen monogenen Erkrankungen ist die Penetranz von Mutationen sehr hoch und unterliegt normalerweise starker Selektion. Mutation kreiert neue Allele, Selektion entfernt sie (unabhängig von Populationsgröße und unter Vernachlässigung zufälliger genetischer Drift) und die Frequenz krankheitsverursachender Allele bleibt klein. Diese Balance, so Pritchard & Cox (2002), ist ein passendes Modell, wenn die Gleichgewichtsfrequenz  $p < 1\%$  ist.

Varianten, die an komplexen Erkrankungen beteiligt sind, haben eine geringe oder mittlere Penetranz und unterliegen wahrscheinlich keiner solch starken Selektion. Mit geringer Selektion dominieren stochastische Mutationseffekte und genetische Drift. Die meisten Loci befinden sich dann nicht mehr nahe ihres deterministischen Gleichgewichtswertes. Schwache Selektion reduziert die Wahrscheinlichkeit, dass sich verursachende Allele nahe der Fixierung befinden, aber ist nicht so stark, die Frequenz krankheitsverursachender Allele klein zu halten. Dies hat ein erhöhtes allelisches Spektrum und eine mittlere Frequenz (5-10% oder mehr) der krankheitsverursachenden Allele zur Folge. Solche Loci repräsentieren wahrscheinlich nur einen kleinen Teil aller Loci, die an der Anfälligkeit für eine Erkrankung beteiligt sind, aber sie tragen disproportional zum Populationsrisiko bei.

Im Ergebnis zeigt das Modell von Pritchard, dass die Mehrheit potentieller krankheitsverursachender Loci kaum genetische Variation zeigt, d.h. dass sich krankheitsverursachende Allele (nahezu) fixiert haben und damit wenig zur phänotypischen Varianz beitragen können, wenn nicht (1) krankheitsverursachende Allele schwacher Selektion unterliegen, oder (2) die Mutationsrate neutraler Allele in einem ähnlichen Bereich wie die Raten krankheitsverursachender Allele angenommen werden. Dann kann infolgedessen die Frequenz verursachender Allele höhere Werte erreichen.

Ein weiterer Unterschied der Modelle liegt in der Berücksichtigung von Bevölkerungswachstum durch Reich & Lander (2001). Reich & Lander beschreiben, wie unter Annahme ähnlicher Mutationsraten von Allelen seltener und häufiger Erkrankungen, seltene Allele eher dazu neigen, stochastisch verloren zu gehen, bzw. von neuen Mutationen überlagert zu werden (Negative Selektion, Reservoir-Effekt): 0.3 % der Allele pro Generation im Fall seltener Erkrankungen; 0.0013 % für häufige Erkrankungen. Nach diesen Berechnungen liegt die Halbwertszeit des angestammten allelischen Spektrums bei 200 Generationen (5000 Jahren) für seltene und bei 52.000 Generationen (1.3 Mio Jahren) für häufige Erkrankungen.

Pritchard & Cox (2002) sehen ihr Modell konstanter Populationsgröße in Übereinstimmung mit den Schätzungen von Reich & Lander als durchaus gültige Approximation, solange nicht von sehr kleinen Frequenzen krankheitsverursachender Allele ausgegangen wird.

Für mittlere Frequenzen hat die Modellierung des Bevölkerungswachstums eine ähnliche Auswirkung auf die Schätzung allelischer Heterogenität wie die Annahme höherer Mutationsraten, d.h. die Ergebnisse beider Studien (Pritchard, 2001; Reich & Lander, 2001) unterscheiden sich in diesem Fall nur unwesentlich. Unterschiedlich sind die Interpretationen hinsichtlich der CD/CV-Hypothese. Reich & Lander beurteilen Werte allelischer Identität  $\phi > 0.1$  als „kleines allelisches Spektrum“ und sehen die CD/CV-Hypothese als wahrscheinlich an. Pritchard beurteilen den Effekt allelischer Heterogenität bei mittleren Allelfrequenzen nicht als dramatisch, sehen jedoch ein Problem darin, Effekte einzelner SNPs oder Haplotypblöcke auf dieser Basis nachweisen zu können. Sie raten zur Analyse größerer genomischer Abschnitte im LD mit krankheitsverursachenden Varianten und zur Betrachtung von Multilokus-Genotyp-Daten.

## Linkage disequilibrium

Das Problem, vor dem man in der Analyse komplexer Erkrankungen steht ist, dass Krankheiten sich in der Familie häufen, jedoch kein einfaches Segregationsmuster, wie monogene Krankheiten, zeigen.

LD (oder auch „Allelische Assoziation“) beschreibt den Zusammenhang zwischen zwei (oder mehreren) Allelen genetischer Marker und besteht, weil humane Populationen gemein-

same Vorfahren haben, die ihre genetische Konstitution weitergeben und die Großfamilie in diesem Sinne die gesamte Population ist (Risch & Merikangas, 1996). LD ist in einer Population nicht konstant, sondern ändert sich im Verlauf von vielen Generationen durch den Einfluss von Rekombinationen, Gen-Konversion und Mutationen. Ob LD besteht, hängt demnach von der Populationsgeschichte (siehe oben) des funktionellen und des Marker-Alleles (von demographischen Besonderheiten sowie der Heterogenität von Rekombinationsraten in verschiedenen genomischen Regionen) ab und ist schwer vorherzusagen (Goldstein et al., 2003). Zudem werden an den Nachweis von LD bestimmte Anforderungen an die Frequenz der beteiligten Varianten gestellt (Müller-Myhsok & Abel, 1997).

Die Wahrscheinlichkeitsverteilung, aus der der aktuelle Wert des LD stammt, resultiert aus der evolutionären Entwicklung der Population. Dieser Prozess wird in der Populationsgenetik durch die „Coalensenztheorie“ erklärt. Wenn eine Stichprobe von Chromosomen aus der Population gezogen wird, sind diese Chromosomen über eine unbekannte Genealogie, dem sogenannten „Coalenszent“ (Ahnen-Stammbaum), miteinander verwandt. Genetische Marker, die auf einem Chromosom sehr dicht beieinander liegen, haben dieselbe bzw. eine ähnliche Genealogie und ihre Allele sind daher abhängig voneinander. Weiter voneinander entfernte genetische Marker haben möglicherweise verschiedene Abstammungen, verursacht durch den Prozess der Rekombination (Pritchard & Przeworski, 2001).

Ursprünglich zeigt jeder durch Mutation entstandene SNP-Haplotyp volles LD. Durch Rekombination und andere Umlagerungen kommt es irgendwann einmal zum gegenseitigen Austausch von homologen Chromosomenabschnitten, wodurch die ursprüngliche Kopplung für diesen Fall verloren geht. Setzt sich dieser Prozess in der gesamten Population durch (was bei „neutralen“ Allelen ein komplexes Zufallsereignis ist), dann wird auch in ihr das volle LD ganz oder teilweise „aufgelöst“. Die Wahrscheinlichkeit dieser Rekombinationsprozesse ist nicht gleichmäßig über das Chromosom verteilt (es gibt „Hot Spots“ und „Cold Spots“ der Rekombinationen), aber der Tendenz nach ist sie umso größer, je weiter die beiden Orte voneinander entfernt sind. Die minimale Rekombinationshäufigkeit ist 0%, wenn die Allele vollständig (allelisch) assoziiert sind und gekoppelt vererbt werden. Die maximale Frequenz beträgt 50% ( $c = 0.5$ ), wenn die Allele entweder auf demselben Chromosom weit voneinander entfernt sind oder auf verschiedenen Chromosomen liegen.

Wird die Rekombinationswahrscheinlichkeit  $\theta$  pro Meiose betrachtet, so erhält man  $\theta = 4N_e c$ , d.h. viermal die effektive Populationsgröße, aus der die Gameten stammen, multipliziert mit der Rekombinationsrate  $c$  je Gamet je Generation zwischen den Endpunkten der betrachteten Region (Hudson, 1987).  $\theta$  ist eine entscheidende Größe zur Charakterisierung des Ausmaßes von LD. LD ist geringer, wenn  $\theta$  groß ist und umgekehrt (Pritchard & Przeworski, 2001). Wenn die Rekombinationsereignisse in der betrachteten chromosomalen Region gleichmäßig verteilt sind, dann ist  $\theta$  proportional zur physischen Distanz. Die



mathematische Behandlung dieser stochastischen Ereignisse führt auf die Vorhersage eines exponentiellen Abfalls des LDs als Funktion der Distanz, allerdings mit erheblicher Varianz (Falconer & Mackay, 1996).

Bei der Beurteilung phänotypischer Assoziation zwischen Genen und komplexen Erkrankungen (Merkmalen) stellt Populationsstratifikation einen möglichen kritischen Parameter dar. Dabei geht es um die Existenz von Sub-Populationen (im Extremfall bedeutet dies eine Mischung mehrerer voneinander isolierter Bevölkerungsgruppen) mit unterschiedlich hoher Inzidenz des untersuchten Merkmals. Dies kann eine unechte (Schein-) Assoziation zwischen dieser Krankheit und jedem genetischen Marker hervorrufen, der sich bzgl. seiner Allelfrequenz in den Sub-Populationen unterscheidet (Pritchard et al., 2000).

Die Existenz von Sub-Populationen stellt eine Abweichung von der Annahme zufälliger Partnerwahl („random mating“ - auf isolierte Subgruppe beschränkt) in der Gesamtstichprobe dar und kann in heterogenen LD-Mustern und folglich verschiedenen Haplotypfrequenzen resultieren. Reich et al. (2001) zeigen, dass LD-Muster zwischen verschiedenen geographischen Regionen variieren. LD in afrikanischen Populationen ist niedriger als in nicht-afrikanischen. Der Hauptgrund dafür liegt darin, dass die Migration „out of Africa“ nur einen Teil des allelischen Spektrums in andere geographische Regionen getragen hat. Dieser Founder-Effekt kann das LD erhöht haben.

Devlin & Risch (1995) beschreiben verschiedene Maße zur Charakterisierung des LDs zwischen zwei bzw. mehreren Markern. Auf hier verwendete Maße (Korrelationskoeffizient  $r$  bzw. sein Quadrat  $r^2$ ) wird in den Methoden eingegangen. Clark (2003); Pritchard & Przeworski (2001) beschreiben die LD-Struktur als sehr variabel und in verschiedenen genomischen Regionen unterschiedlich. Dieses Bild hat sich letztlich durch die aktuelle Veröffentlichung der Ergebnisse des Internationalen Haplotyp-Mapping-Projektes (The International HapMap Consortium, 2005) bestätigt. Dieses Projekt soll helfen, den Wert von SNPs in indirekten Assoziationsstudien aufzuklären.

Durch die große Anzahl von SNPs im Genom und rapide sinkende Kosten für die Genotypisierung hat die Bedeutung von Assoziationsstudien stark zugenommen. Familienbasierte Assoziationsstudien in unterschiedlichen Populationen haben mehr mit klassischen epidemiologischen Studien zu Umwelt- und Verhaltensrisiken gemeinsam, als Kopplungsanalysen. Die Parallelen werden deutlich, wenn man Gründe betrachtet, warum Assoziation zwischen einem genetischen Faktor und einem Merkmal in einer gegebenen Population besteht: (1) der Polymorphismus ist kausal; (2) der Polymorphismus ist nicht kausal, aber assoziiert mit einer benachbarten (ungemessenen) kausalen Variante oder (3) die Assoziation basiert auf Mischung verschiedener Populationen. In einer gemischten Population, in welcher die einzelnen Strata beispielsweise umweltbedingt verschiedenartige Wirkungen zeigen oder Founder-Populationen ein anderes genetisches Risiko einbringen, ist jeder Marker-Lokus,

dessen Allelfrequenzen sich zwischen den Strata oder Founder-Populationen unterscheiden, in gewissem Grade mit dem Merkmal/der Krankheit assoziiert, unabhängig davon, ob er sich nahe eines kausalen Marker-Lokus befindet (Cordell & Clayton, 2005).

Diese einleitenden Betrachtungen münden in einer Reihe von Fragestellungen, die im Folgenden formuliert und in weiteren Kapiteln beantwortet und diskutiert werden.

## 1.4 Fragestellungen

1. Welches Design ist zur Analyse komplexer Merkmale geeignet?
2. Kann anhand untersuchter SNPs und SNP-Haplotypen ein polygener Einfluss auf die Lipidphänotypen nachgewiesen werden? Wie groß ist vergleichsweise der separate Anteil einzelner Gene an genetisch bedingter phänotypischer Varianz?
3. Unterstützen die vorliegenden Untersuchungen die „Common Disease/Common Variants“ - Hypothese komplexer Erkrankungen?
4. Wie ist die Veranlagung (Heritabilität) für Lipidphänotypen genetisch fixiert?
5. In welcher Weise kann der genetische Beitrag einzelner Gene sinnvoll quantifiziert werden?
6. Welches statistische Modell kann sinnvoll zur Modellierung phänotypischer Assoziation angewendet werden?
7. Bestätigen sich Allelfrequenzen und Kopplungsungleichgewicht (LD) in einer unabhängigen europäischen Bevölkerungsstichprobe aus dem Kanton Genf?
8. Wie valide sind die Ergebnisse aus den Assoziationsanalysen?

Abschließend stellt sich die Frage, ob es sich lohnt, nach Polymorphismen zu suchen, wenn bei Menschenpopulationen die näher am Phänotyp befindlichen Expressions- oder Proteomic-Muster aus Organen oder Zellen der direkten Untersuchung in epidemiologischen Kohorten nicht zugänglich sind.