Kapitel 4

Diskussion

Der Transport von Lipiden ist ein zentrales metabolisches Ereignis im Organismus: ohne Lipide kein Leben und ohne Lipoprotein-Mizellen keine Lenkung der wasserabweisenden Lipide. Der individuelle Stoffwechsel kann in verschiedener Weise genetisch beeinflusst werden und jeder ist sich potentieller umwelt- und lebensstilbedingter Einflüsse bewusst. Hinzu kommt der Einfluss, der sich aus der Wechselwirkung zwischen Genen und Umwelt ergibt.

Störungen des Lipoproteinstoffwechsels sind ein bedeutender Risikofaktor der Herz-Kreislauferkrankungen. Genetische Faktoren sind dabei nur in Sonderfällen monogen (Familiäre Hypercholesterinämie) und folgen den Mendelschen Gesetzen. Eine Überblicksarbeit von Vogler et al. (1997) stellt die wichtigsten genetischen und nicht-genetischen Risikofaktoren zusammen.

Die Ätiologie der Herz-Kreislauferkrankungen ist komplex. Studien in Kandidatengenen haben eine Vielzahl von Polymorphismen analysiert, die mit phänotypischen Merkmalen der Arteriosklerose signifikant assoziiert bzw. gekoppelt sind (Carr et al., 2002; Falchi et al., 2004). Die Interpretation ist jedoch bislang nicht eindeutig. Einen anderen Ansatz machen Studien, die das gesamte Genom mit dem Ziel untersuchen, „Arteriosklerose-Gene“ zu identifizieren, d.h. Gen-Loci, die mit dem Auftreten von Diabetes, Hyperlipidämie, niedrigen HDL-Konzentrationen und Hypertension zusammenhängen (Altmuller et al., 2001; Peacock et al., 2001). Nur für eine kleine Anzahl von Gen-Loci konnten Kopplungseffekte statistisch signifikant nachgewiesen werden, was auf die Komplexität dieser Merkmale schließen lässt. Inkonsistente Befunde dabei sind häufig auf Typ-I-Fehler, d.h. auf falsch positive Ergebnisse, zurückzuführen. Publikationsverzerrung, d.h. die Tatsache, dass im allgemeinen nur signifikante Befunde publiziert werden, multiple Testung und Inhomogenität der Studiengegenstände spielen ebenso eine wesentliche Rolle.

Trotz dieser Probleme konnte für eine Auswahl von Genen konsistent Kopplung oder Assoziation mit Herz-Kreislauf-Erkrankungen bzw. Risikofaktoren nachgewiesen werden,

oder es zeigten sich ähnliche Effekte wie in Tiermodellen (Lusis, 2003).

Gegenwärtig ist noch nicht bekannt, durch welche genetischen Varianten die physiologische Variation in Lipidkonzentrationen bestimmt ist. Bislang konnte in diversen Studien, die sich mit der Erforschung der Ursachen komplexer Erkrankungen beschäftigen, nur im *APOE*-Gen ein funktioneller Haplotyp-Effekt auf Lipid-Konzentrationen konsistent nachgewiesen werden. SNP und/oder Haplotyp-Effekte in weiteren Kandidatengenen des Lipoproteinstoffwechsels sind größtenteils marginal und zum Teil kontrovers.

Trotz Heterogenität der Krankheitsursachen und genetischer Heterogenität der Risikophänotypen, waren Studien kardiovaskulärer Erkrankungen erfolgreicher als die anderer komplexer Krankheiten. Terwilliger & Goring (2000) u.a. sehen den Grund in der frühzeitigen Betrachtung biologischer Vernetzung von Einflussfaktoren, in der Betrachtung quantitativ messbarer Variablen (Sub-Phänotypen) sowie umweltbedingter und genetischer Faktoren. Das biochemische Wissen über die Stoffwechselwege ist umfassend und ermöglicht die Auswahl von Kandidatengenen. Eine homogene Studienpopulation hinsichtlich spezifischer Phänotypen und gemeinsamer genetischer Vergangenheit sind nützliche Werkzeuge zur Aufklärung heterogener Systeme, wie das der Herz-Kreislauferkrankungen.

Viele Studien haben die Analyse von Krankheitsbildern im Blickpunkt und bilden dabei nur die pathologischen Randbereiche der Verteilung eines Merkmals in der Bevölkerung qualitativ ab. Die vielschichtige Verbindung zwischen individueller genetischer Konstitution, Lipoproteinstoffwechsel und Herz-Kreislauferkrankungen hingegen motiviert die Betrachtung eines intermediären Phänotyps.

Das vorliegende Studiendesign behandelt mit der Analyse reellwertiger (quantitativer) Plasma-LDL- und HDL-Cholesterin-Konzentrationen einen geeigneten Phänotyp für eine Modellierung der Gesamtbevölkerung.

Die Mehrzahl der Individuen, die an Herz-Kreislauferkrankungen leiden, haben Cholesterin-Konzentrationen im normalen intermediären Bereich. Bei den Lipidwerten handelt es sich um ein komplexes Merkmal, das gut messbar und von epidemiologischem sowie klinischem Interesse ist. Unter standardisierten physiologischen Bedingungen bleiben individuelle Lipid-Konzentrationen über längere Zeiten im gleichen Bereich; die Variation zwischen Individuen hingegen ist groß. Bereits kleine Abweichungen von der individuellen Mittellage können jedoch einen Langzeiteffekt auf den Risikostatus einer Person haben. Daraus ergibt sich die Hypothese, dass bereits diesen kleinen Abweichungen genetische Faktoren zugrunde liegen. Die Untersuchung dieser Hypothese ist Gegenstand der vorliegenden Arbeit und Teil des Gesamtprojektes, dessen theoretischer Teil im Rahmen des DHGP finanziert werden konnte. Die Rekrutierung der Probanden, die Phänotypisierung und Genotypisierung wurde mit finanziellen Mitteln des MDC, der Franz-Volhard-Klinik und der Infogen GmbH

realisiert.

Zur Analyse wurde eine Stichprobe aus der normolipidemischen Bevölkerung, d.h. aus Individuen ohne auffälligen Lipidstatus, herangezogen. In einer solchen Stichprobe sind potentielle Kandidaten (Individuen) einer sog. „late-onset"-Erkrankung (die erst spät ausbricht) enthalten. Es wird davon ausgegangen, dass die „Normalbevölkerung" ein Arterioskleroserisiko trägt und bzgl. ihres intermediären Phänotyps als Untersuchungspopulation geeignet ist.

Der intermediäre Phänotyp ist biochemisch gut charakterisiert, so dass Kandidatengene für die am Stoffwechselweg beteiligten Genprodukte bekannt sind. Die Kandidatenloci stehen im Zusammenhang mit der Physiologie des Lipoproteinstoffwechsels, und Mutationen bedingen ernsthafte Funktionsdefekte. Die Frage ist, ob die **genetische Variation in der Bevölkerung** für die **phänotypische Variation** in der Gesamtbevölkerung von Bedeutung ist oder nicht.

Wie die Ergebnisse zeigen, ist das nicht trivial und auch nicht einheitlich zu beantworten. *APOE*- oder *CETP*-Variation ist assoziiert mit der Cholesterinvariation, *LCAT* dagegen nicht, obwohl letzteres ein unabdingbarer Faktor ist – genetische Defekte, z.B. die Fischaugenkrankheit (Norum et al., 1989), zeigen das.

In den Jahren 1998/1999 wurde das Projekt, das der hier vorliegenden Arbeit zugrunde liegt, geplant. Es stand unter dem Einfluss der CD/CV-Hypothese komplexer Erkrankungen (Chakravarti, 1999; Collins et al., 1997) und der vorausgesagten höheren Power von Linkage-Disequilibrium-Mapping (Assoziation) im Vergleich zu Kopplungsanalysen (Risch & Merikangas, 1996). Die bei monogenen Erkrankungen erfolgreichen Analysen auf Kopplung genetischer Marker in Pedigrees war als nicht aussichtsreich auszuschließen, weil hier untersuchte Phänotypen komplex sind und ihr Ursachengefüge heterogen ist. Zum Ziel der Studie wurde demzufolge, nach Assoziation von SNPs mit den Lipidphänotypen zu suchen. Der Ansatz sah im Unterschied zu bisherigen Studien, die größtenteils SNP- bzw. Haplotyp-Einzeleffekte untersuchten, eine Modellierung multipler SNPs je Gen-Lokus vor.

Der Start des Projektes lag vor der Veröffentlichung der Daten des menschlichen Genoms (Sachidanandam et al., 2001) und an eine umfassende (genomweite) Darstellung der Verteilung von Haplotypen und Haplotypblöcken in verschiedenen Populationen (The International HapMap Consortium, 2005) war noch nicht zu denken. Die SNP-Auswahl war demnach *a priori* nur basierend auf veröffentlichter Literatur möglich. Mit der geplanten Größe der Studienpopulation von 1054 Probanden kamen für die Analyse nur häufige (common) SNPs in Frage, da nur diese hinreichend häufig gefunden werden können, um statistische Aussagen zu gestatten.

Eine theoretische Vorhersage, ob die CD/CV-Hypothese (Chakravarti, 1999; Collins et al., 1997) zutreffend oder nicht zutreffend sein wird, ist nach wie vor nicht eindeutig mög-

lich. Theoretische Untersuchungen fassen die in der Bevölkerung segregierenden Varianten als variable Einflussfaktoren auf den jeweils betrachteten Phänotyp auf und fragen danach, wieviele funktionell relevante Allele vorhanden sein können. Allelische Heterogenität ist ein populationsgenetisches Problem und abhängig von der effektiven Populationsgröße, Anzahl vergangener Generationen (Meiosen), Wahrscheinlichkeit der Bildung von SNP-Varianten (Mutationen), Bevölkerungsengpässen, etc.

Reich & Lander (2001) kommen mit ihren populationsgenetischen Betrachtungen zu dem Schluss, dass für ein komplexes (polygenes) Merkmal nur wenige Allele segregieren können. Sie sehen die Analyse phänotypischer Assoziation (bzw. Linkage disequilibrium - Mapping komplexer Merkmale/Erkrankungen) als einen (potentiell) leistungsstarken Ansatz zur Identifizierung häufiger ($> 1\%$) krankheitsverursachender genetischer Varianten. Pritchard & Przeworski (2001); Terwilliger & Weiss (1998) hingegen sehen auf der Basis nicht minder plausibler alternativer Parameterwerte das Ausmaß allelischer Heterogenität im Sinne der Analyse phänotypischer Assoziation über Varianten im LD als kritisch an.

Nach mehreren Jahren der Rekrutierung und der mathematischen Analyse wurden die Ergebnisse des interdisziplinären Projektteams verschiedenfach publiziert (Bauerfeind et al., 2006, 2002; Knoblauch et al., 2002, 2004; Nürnberg et al., 2004). Diese Dissertation stellt die Ergebisse der mathematischen Analyse vor, mit der die Dissertantin als Teilaufgabe im Projekt befasst war.

Die Gesamtbotschaft ist die folgende:

Auf der Basis des polygenen biometrisch-genetischen Modells lässt sich ein bedeutender Anteil der lipidphänotypischen Variation mittels Assoziation multipler SNPs oder SNP-Haplotypen an Kandidatenloci erklären.

Die Daten zeigen deutliche Unterschiede zwischen assoziierten Genorten (z.B. *APOE*, *CETP*) und Loci, die keine oder nur schwache Assoziationseffekte aufweisen. Das Ausmaß hier festgestellter allelischer Effekte ist konsistent mit bekannten physiologischen Effekten der Gene. Sing & Davignon (1985) zeigten beispielsweise, dass $APOE \approx 10\%$ der LDL-Varianz erklärt. In der vorliegenden Analyse liegt der Beitrag von *APOE* bei $8\% - 9\%$.

Nur in einigen Fällen lassen sich einzelne SNP-Genotypen oder SNP-Haplotypen individuell als signifikant assoziiert nachweisen. (Eine Detailanalyse ist einer zukünftigen Arbeit vorbehalten.) Effekte häufiger Varianten lassen sich statistisch schwer vom allgemeinen Mittelwert abgrenzen, seltene Varianten dürfen nicht zu selten sein, damit sie statistisch signifikante Varianzschätzungen zuverlässig erlauben (Risch & Merikangas, 1996; Zondervan & Cardon, 2004).

Approx. 70% der genetisch bestimmten lipidphänotypischen Varianz konnte durch gemeinsame Modellierung häufiger SNPs oder SNP-Haplotypen erklärt werden. Ein solcher

Prozentsatz hat nur orientierende Aussagekraft und ist abhängig von der statistischen Power angewandter Methoden und dem Einfluss der Modellfreiheitsgrade, d.h. Beschränkungen durch Daten und Modell. Die Stärke nachgewiesener Assoziation gewinnt an Bedeutung, wenn man bedenkt, dass die Lipidkonzentrationen bedingt durch Messfehler und Umwelteinflüsse individuell stark schwanken. Außerdem repräsentieren die untersuchten Gene nur eine Auswahl untersuchter Kandidaten, die den Fettstoffwechsel beeinflussen, und die untersuchten genetischen Marker sind vermutlich trotz geringer genetischer Abstände nicht streng korreliert mit Teilsequenzen von funktioneller Bedeutung und mögen nur einen Teil der Variabilität des jeweiligen Gens widerspiegeln.

Die CD/CV-Hypothese wäre demnach so zu interpretieren, dass in der Population weit verbreitete SNPs mit den für die physiologische Variation verantwortlichen Sequenzabschnitten korrelieren (aber nicht mit ihnen identisch sind).

Die vorliegenden Ergebnisse unterstützen die „Common Disease/Common Variants“ - Hypothese für den komplexen intermediären Phänotyp „Lipoproteinmuster im Blutplasma“.

Allerdings ist die Struktur des Zusammenhangs sehr kompliziert und lässt sich wegen stochastischer Streuung und Störeffekten nicht leicht nachweisen.

Lohmueller et al. (2003) untersuchten in einer Meta-Studie 301 publizierte Studien, die 25 verschiedene Assoziationen zwischen häufigen genetischen Varianten und häufigen Krankheiten im Blickpunkt hatten. Nur wenige der publizierten Effekte waren konsistent. Inkonsistente Ergebnisse können auf falsch positiven/negativen Befunden oder populationsbedingt unterschiedlichen Werten beruhen. Lohmueller et al. beurteilen es als wahrscheinlich, dass ein Teil (in etwa ein Viertel) der publizierten Assoziationen mit komplexen Erkrankungen „wahr“ sind. Dieses Fazit ist durchaus positiv, wenn man bedenkt, dass die CD/CV-Hypothese im Sinne allelischer Heterogenität generell in Frage gestellt wird.

Deutliche Unterschiede haben sich in der vorliegenden Arbeit in getrennten Modellen für Männer und Frauen gezeigt. CETP-Effekte können nur für Frauen nachgewiesen werden. LIPC und LDLR-Varianten beeinflussen die Lipidwerte ausschließlich bei Männern. Da in den Familien die autosomalen Genorte frei zwischen männlichen und weiblichen Individuen segregieren, sind solche Unterschiede auf lebensgeschichtliche, u.a. hormonelle Einflusskomponenten zurückzuführen. Die Befunde legen nahe, beiden Geschlechtern verschiedene Funktionsmuster zuzuordnen.

Es ist zweckmäßig, die Genvariation-Phänotyp-Beziehung in Familien anstelle unabhängigen Stichproben aus der Population zu studieren, denn dabei lässt sich die Segregation von Allelen verfolgen und man kann die Heritabilität des Phänotyps schätzen (Clark, 2003; Terwilliger et al., 2002). Eine Vielzahl genetischer Studien bestätigt, dass ein bedeutender Anteil der lipidphänotypischen Variation erblich bedingt ist. Die Heritabilitätsschätzungen

sind variabel und liegen bei max. 50% (Marenberg et al., 1994; Rao et al., 1979; Williams et al., 1993).

Mit mit Werten um 30% lagen die Schätzungen der Heritabilität unter denen vergleichbarer Studien.

Diese Tatsache resultiert möglicherweise aus der vergleichsweise hohen Variation nicht-genetischer Komponenten in der Studienpopulation, verglichen mit in größerem Maße standardisierten Stichproben anderer Untersuchungen (z.B. Follow-Up-Studien).

Die Heritabilität ist ein relativer Wert, der genetische und umweltbezogene Faktoren ins Verhältnis setzt. Die untersuchten Familien wurden deutschlandweit rekrutiert und phänotypische Messungen hauptsächlich in normalen klinischen Laboratorien vorgenommen. Der Grad an Messwertstreuung wäre in einer Studie, bei der alle Messungen im selben Labor vorgenommen würden, vermutlich geringer. In Stichproben der gesamten Bevölkerung ist es schwierig, den physiologischen Status des intermediären Phänotyps (postprandial, 12 h nach der Nahrungsaufnahme) exakt zu erreichen. Aus diesen Umständen folgt, dass die nicht-genetische Variation größer sein kann, daraufhin den relativen Heritabilitätswert senkt, ohne dass dies (absolut) auf einen geringeren genetischen Effekt schließen ließe.

Epistatische Effekte (Wechselwirkungen zwischen Genen) sind bislang weitgehend unerforscht und die Analysen scheitern zumeist an geringer statistischer Power wegen schwacher Zellbesetzungen in den Kombinationen. Templeton et al. (2000) diskutiert mögliche epistatische Effekte komplexer Merkmale am Beispiel des Fettstoffwechsels. In vorliegenden Untersuchungen wurden paarweise Wechselwirkungen zwischen SNPs verschiedener Gene getestet und zeigten keine signifikanten Befunde. Es bietet sich an, das Problem lokusweise zu betrachten, d.h. Wechselwirkungen zwischen Genen anstatt zwischen einzelnen genetischen Markern zu betrachten. Die Schwierigkeit dabei liegt in der Form der Bemessung eines Lokus- bzw. Interaktionseffektes.

Der Erfolg einer genbasierten Assoziationsanalyse und inkonsitente Einzeleffekte auf Markerebene bestätigen das Konzept von Neale & Sham (2004), das unlängst publiziert wurde und mit dem hier vorgestellten Ansatz übereinstimmt.

Genbasierte Ansätze mit gemeinsamer Modellierung aller Varianten eines Gens sind sinnvoll zur Bewertung von Assoziationseffekten der Kandidatengene.

Neale & Sham argumentieren, dass die derzeitige Tendenz, Assoziationsstudien auf dem Niveau einzelner SNPs oder Haplotypen durchzuführen, problematisch ist. In den chromosomalen Sequenzvarianten eines Gens („Haplotyp“) sind sehr viele SNPs ohne funktionelle Bedeutung zufällig enthalten. Das kann speziell für viele Intron-SNPs zutreffen, die ohnehin „herausgespleißt“ werden. Solche SNPs werden jedoch, wenn alles andere gleich ist, einen

funktionellen Haplotyp in zwei funktionell gleichwertige Subhaplotypen spalten, die sich an der entsprechenden SNP-Position unterscheiden. So erhielte man eine riesige Auswahl von Haplotypen, was die Nachweisbarkeit eines funktionellen Aspektes in jedem einzelnen drastisch vermindert. Betrachtet man die Gesamtheit aller messbaren Haplotypen eines Lokus, dann vermeidet man den Verlust an statistischer Power durch die Aufspaltung.

Ein Gen ist eine funktionelle Einheit des Genoms, dessen Position, physiologische Funktion und Sequenz (bis auf SNP-Positionen) zwischen Individuen übereinstimmt. Genbasierte Analysen ermöglichen vergleichbare Aussagen für verschiedene Populationen unter Berücksichtigung ihrer spezifischen Allelfrequenzen und LD-Struktur. Ein weiterer Vorteil genbasierter Ansätze ist die Vereinfachung des Problems multiplen Testens durch ein zweistufiges Verfahren, in dem in geeigneter Weise zunächst multiple Marker modelliert werden und anschließend die Gen-Loci gemeinsam betrachtet werden können. Neale & Sham betonen, dass ein genbasierter Ansatz insbesondere zur Validierung von Assoziationseffekten oder dem Studium von Kandidatengenen verwendet werden sollte.

Phänotypische Assoziation bezeichnet das gehäufte Vorkommen eines genetischen Merkmals (hier SNP / SNP-Haplotyp) bei Trägern eines phänotypischen (hier quantitativen) Merkmals. Assoziationsanalysen werden bestenfalls direkt über funktionelle genetische Marker oder indirekt über benachbarte funktionsneutrale Varianten, die mit dem verursachenden Allel stark gekoppelt sind, durchgeführt. Phänotypische Assoziation setzt demnach Kopplung an dem untersuchten Kandidaten-Lokus (mit der ungemessenen Variante) voraus (Fulker et al., 1999). Wenn die Kopplung zwischen Marker und funktionellem Allel nicht vollständig ist, dann ist die phänotypische Assoziation auch nur abgeschwächt erkennbar.

Die Methode der Varianzanalyse ist in einer bestimmten Variante zur Analyse geeignet, denn sie gibt ein Klassifizierungsmerkmal (den Genotyp) vor und definiert eine (normalverteilte) quantitative Größe als abhängige Variable. In einer davon abgeleiteten Regressionsanalyse wird hier von dem kategorialen genetischen Einflussfaktor auf eine numerische Variable (Gendosis: 0/1/2) übergegangen. Der Regressionsansatz ist sinnvoll, da multiple Faktoren modelliert werden und eine entsprechend multi-faktorielle Varianzanalyse so viele Klassen erzeugen würde, die hier vorliegende Stichproben nicht zu belegen imstande wären. Der multiple Ansatz zur Modellierung von Assoziation beeinhaltet nicht nur einen Kandidatenort mit einem oder wenigen Markern, sondern eine Vielzahl von Genorten, die gemeinsam das Ursachengeflecht der phänotypischen Variation bilden. Nachteil des Regressionsansatzes sind mögliche Kollinearitäten zwischen genetischen Parametern. Dieser Effekt führt dazu, dass sich Regressionsparameter im Modell gegenseitig vertreten können und damit die Zuordnung von Effekten nicht mehr eindeutig ist.

Das Studium von Großfamilien verletzt die Annahme unabhängig verteilter Beobachtungen und führt auf die Annahme einer kopplungsbedingten positiven Kovarianz des unter-

suchten phänotypischen Merkmals zwischen allen Verwandten, die das gleiche Marker-Allel tragen. Wenngleich die Schätzung der Assoziationseffekte trotz Abhängigkeiten unverzerrt ist, führen Methoden, die auf Unabhängigkeit beruhen und intrafamiliäre Korrelation nicht berücksichtigen, zu inkorrekten Varianzschätzungen (Bull et al., 2001). Statistische Methoden für korrelierte Daten, wie beispielsweise die Methode der „Generalized Estimating Equations (GEE)“ (Zeger et al., 1988; Ziegler et al., 2000) oder „Bootstrap-Resampling“ (Efron & Tibshirani, 1998), können valide Varianzschätzungen in zufälligen Stichproben von Familien liefern.

Hier wurde ein gemischtes Modell angesetzt, das einen multiplen linearen Regressionsansatz als Modell der festen (gemessenen) Effekte mit einem Varianzkomponentenmodell für die zufälligen (ungemessenen) Effekte verbindet. Die familiäre Korrelation muss entsprechend dem Verwandtschaftsgrad mit einem Gewichtungsfaktor versehen werden - anders als bei Geschwisterpaar-Analysen, bei denen die gleiche kopplungsbedingte Kovarianz vorausgesetzt wird. Unter vereinfachten Annahmen (Mendelsche Vererbung an einzelnen Marker-Loci, Unabhängigkeit der Varianzkomponenten, etc.) wurde somit die Abhängigkeit der Beobachtungen modelliert und als erbliche Komponente (Heritabilität) der Varianz bemessen.

Das gemischte Modell mit Kinship-Struktur ist ein geeigneter Ansatz zur Schätzung gemessener und ungemessener genetischer Effekte auf den Phänotyp.

Das gemischte Modell unterstellt homogene Assoziations- sowie Kopplungseffekte in der Studienpopulation, d.h. der genotypische Einfluss auf den Phänotyp sowie polygener und umweltbedingter Beitrag an der phänotypischen Varianz werden in der Gesamtheit der Familien - gewichtet durch den Verwandtschaftskoeffizienten - als gleich angenommen. Eine Verletzung der Homogenität durch Populationsstratifikation würde im Sinne phänotypischer Assoziation eine verzerrte Schätzung der Regressionskoeffizienten bedeuten. Inhomogenität im Sinne familiärer Kopplung kann verzerrte Schätzungen der Varianzkomponenten verursachen.

Die Konfidenzintervalle der Varianzkomponenten, geschätzt durch familienbasierte Bootstrap-Simulationen, zeigen eine relativ breite Streuung der Schätzungen innerhalb der Studienpopulation. Dies kann einerseits Spiegel von nicht modellierten Inhomogenitäten zwischen Individuen innerhalb oder zwischen Familien sein bzw. andererseits Ausdruck Gen-Umweltbedingter Wechselwirkungen.

Eine erhebliche Erweiterung des Probandenkollektivs kann helfen, Modellparameter mit größerer Sicherheit zu schätzen. Zu überlegen (und in Bezug auf feste Effekte gezeigt) sind spezifischere Arbeitshypothesen, die sich auf Sub-Populationen beschränken (Männer versus Frauen, Alt versus Jung) und möglicherweise eine homogenere Daten-Ausgangsbasis

darstellen.

Die Grenzen multipler Assoziationsmodelle werden deutlich, wenn man die Aufspaltung des Genotyps durch Hinzufügen immer neuer genetischer Parameter verfeinert. Die Anzahl der Freiheitsgrade wächst und die Schätzungen werden unsicherer. Es gibt nur wenig Anhaltspunkte, welcher SNP oder welche SNP-Teilkonstellation funktionell bedeutsam ist. Einzelne SNPs können funktionell sein, andere befinden sich im LD mit funktioneller Variation, wieder andere sind irrelevant für die Funktionsweise eines Gens.

In der vorliegenden Arbeit wurde die SNP-Auswahl im Vorfeld der Untersuchungen ohne Optimierung der Auswahl getroffen, d.h. es wurden SNPs in bekannten Kandidatengenen in der Literatur und Datenbanken recherchiert, genotypisiert und ausgewertet. Um die Power der statistischen Verfahren zu erhöhen, war eine Dimensionsreduzierung notwendig, d.h. die Reduzierung der Anzahl von SNPs als Einflussparameter in den Analysen mit zahlreichen zu schätzenden Parametern. Dazu wurden SNPs mit einer Allelfrequenz $> 5\%$ ausgewählt und SNPs in paarweisem Kopplungsungleichgewicht, d.h. hoch korrelierte SNPs, innerhalb eines Gen-Lokus gruppiert. Eine hoch-korrelierte SNP-Gruppe wurde durch einen ausgewählten SNP repräsentiert, wobei überdies (nicht-synonym kodierende) Exon-Varianten bevorzugt wurden, wenn Alternativen zur Auswahl standen.

Die Validierung signifikanter Befunde in unabhängigen Studienpopulationen ist bei der wachsenden Anzahl von Ergebnissen aus Assoziations- und Kopplungsanalysen in genetisch-epidemiologischen Studien von primärem Interesse, wird allerdings aus Kostengründen nur selten durchgeführt. Sie ist in der Annahme begründet, dass sich eine „wahre“ Assoziation zwischen einem Kandidatenlokus und einem phänotypischen Merkmal in verschiedenen Stichproben bestätigen wird, ein „scheinbarer“ Effekt hingegen nicht.

Die Möglichkeit, eine Validierung vorliegender Ergebnisse durchzuführen, hat sich mit einer umfassenden Studie von Morabia et al. (2003b) ergeben. Ihre Studie wurde unabhängig mit einer ähnlichen Fragestellung konzipiert.

Durch Zusammenfassung beider Projekte war es in dieser Arbeit möglich, Stichproben aus zwei historisch unterschiedlichen, aber offensichtlich prähistorisch verwandten europäischen Populationen zu vergleichen, nämlich die Einwohner Deutschlands (ca. 80 Millionen Einwohner) und die französisch-sprechenden Einwohner aus dem Kanton Genf (ca. 2 Millionen Einwohner). Die Verteilungen von LDL und HDL in beiden ethnischen Gruppen, getrennt nach Geschlecht, zeigten nach Korrektur des Alterseffektes statistisch keine Unterschiede. In Bezug auf den komplexen Phänotyp können beide Stichproben deshalb gegenseitig als unabhängige Wiederholung betrachtet werden.

Die Ergebnisse der Analysen können in einer unabhängigen europäischen Stichprobe aus dem Kanton Genf bestätigt werden.

Die genomischen Karten untersuchter Kandidatengene in Berlin sowie die Liste untersuchter Positionen in Genf (Morabia et al., 2003b) (Appendix) zeigen eine dichte SNP-Abdeckung, d.h. einen durchschnittlichen paarweisen Abstand von 2 kb (Zwei Ausnahmen in Berlin: *LIPC* und *ABCA1* mit 14 kb), so dass die in der Population zu erwartende Rekombinationsrate zwischen den Markern innerhalb der Gene gering ist.

Die stochastische Streuung von populationsgenetischen Variablen ist üblicherweise sehr hoch. Zudem ist die Vorgeschichte von Populationen des Homo sapiens aufgrund zahlreicher kritischer Schwankungen der Populationsdichte (bevölkerungsgeschichtlicher Engpässe) und der geographischen Isolierung von Teilpopulationen speziell in den Eiszeiten so verwickelt, dass es zu erwarten ist - und auch in entsprechenden Studien bestätigt wurde (zuletzt in den umfangreichen Veröffentlichungen von The International HapMap Consortium (2005)) - dass sowohl sehr hohe, als auch geringe Kopplungsungleichgewichte zwischen relativ eng liegenden SNPs auftreten.

Trotz dieser allgemeinen Befunde waren Allelfrequenzen der SNPs, die in den Stichproben aus Genf und Berlin gemeinsam untersucht wurden, sehr ähnlich. Die LD-Struktur war ebenfalls in allen getesteten Gen-Loci übereinstimmend (hier auf der Basis des verallgemeinerten Determinantenkriteriums getestet). Ein Vergleich der LD-Struktur verschiedener Stichproben untersucht die Hypothese, dass es sich um zwei Bevölkerungen handelt, deren Entwicklung und genetische Drift bei vergleichbarer Anzahl von Generationen und effektiver Populationszahl lange ähnlich verlaufen. Es ist unbekannt, ob die in Deutschland und die in der Schweiz lebenden Populationen über viele Generationen hinweg „durchmischt“ wurden oder ob Populationsengpässe und geographische Isolation die einfache Verbreitung der Genvarianten über die exponentiell wachsende Bevölkerung gestört haben. Jedoch sind die Ergebnisse der vergleichenden Analysen mit der Annahme einer gemeinsamen prähistorischen Urbevölkerung vereinbar, die für Mitteleuropa als garantiert gilt (Cavalli-Sforza & Bodmer, 1999).

Zum Test der Robustheit der Assoziationsergebnisse hinsichtlich der hier behandelten Hypothese globaler Gen-Lokus-Assoziation wurden vergleichbare Untersuchungsgruppen gebildet. Die Vergleichsstudie konnte nur auf der Basis eines Fall-/Kontrolldesigns anhand eines kombinierten Lipid-Parameters als dichotomem Phänotyp geführt werden. Als „Fälle“ wurden Individuen mit relativ hohem LDL und zugleich niedrigem HDL ausgewählt. „Kontrollen“ sind definiert als Individuen mit niedrigen LDL- und hohen HDL-Konzentrationen.

Geringe Fallzahlen in den ausgewählten Fall-/Kontrollgruppen machten eine Reduzierung der Modellfreiheitsgrade in der multiplen logistischen Regression notwendig. Parameterselektion durch schrittweisen Ein- bzw. Ausschluss potentiell assoziierter SNPs ist ein Verfahren, dass durch die speziellen Daten gesteuert wird, also opportunistisch auswählt. Um die Befunde abzusichern, wurden die Modelle in beiden Stichproben jeweils kreuzvali-

diert.

Die Resultate bestätigen die Assoziation der Genorte mit den Phänotypen, wenngleich mit unterschiedlichen Signifikanzniveaus wegen geringerer statistischer Power in Berlin. Die Bedeutung der Genorte für den Lipid-Phänotyp stimmt bis auf eine Ausnahme (*LPL*) weitgehend überein. Angesichts der durch den hohen Anteil umweltbedingter Restvarianz bedingten Streuung geschätzer Modellparameter, ist dies ein bemerkenswertes Gesamtergebnis.

Ein weiteres allgemeines Ergebnis der Analysen ist, dass der individuelle Genotyp nur sehr unsicher für die Vorhersage des Phänotyps, z.B. für Prognosezwecke, geeignet ist. Dies ist vorauszusehen, wenn der genetische Faktor nur 30% bis 50% der phänotypischen Gesamtvarianz erklärt. Selbst eine ideale Modellierung des genetischen Faktors ergibt einen Varianzbeitrag von maximal 50%. Dies trifft offensichtlich auch für viele weitere komplexe Merkmale des Organismus zu.

Die beste Prognose liefert in hier vorliegenden Untersuchungen die direkte Messung der quantitativen Cholesterinkonzentrationen selbst. In zukünftigen Studien soll geklärt werden, welchen prognostischen Wert der genetische Faktor in Teilpopulationen hat – etwa bei Frauen, deren Neigung zu riskanten Cholesterinwerten sich erst nach der Menopause ausprägt, wenn der hormonelle Schutz wegfällt.

Unabhängig von solchen möglichen Anwendungen liefern die Ergebnisse dieses Projektes eine Reihe von grundlegenden Erkenntnissen. Die CD/CV-Hypothese kann für den komplexen Lipidphänotyp in eingeschränkter Form bestätigt werden. Untersuchte häufige SNPs sind sehr wahrscheinlich größtenteils neutrale Marker im Linkage disequilibrium mit funktionellen Varianten, die hier nicht erfasst wurden. Dabei sind Chromosomensätze mit „atherogenen“ bzw. „atheroprotektiven“ Allelen nicht trennungsgenau durch gemessene SNPs bzw. geschätzte Haplotypen markiert. Dies widerspricht der theoretischen Untersuchung von Reich & Lander (2001). Reich & Lander argumentieren, dass die moderne Bevölkerung nach dem enormen Wachstum in den letzten 10000 Generationen noch immer das eingeschränkte Variationsmuster der Population seit dem letzten großen Populationsengpass trägt. Man müsste demnach annehmen, dass die heterogene Mischung funktioneller Allele und deren ebenso heterogene Markierung durch häufige SNPs bereits vor der Bevölkerungsexplosion vorhanden gewesen ist. Diese Annahme jedoch widerspricht dem Befund zum Teil starken LDs zwischen den Markern. Auch nach Hinzuziehung der Daten der CEPH-Population (Einwohner von Utah mit Vorfahren aus Nord- und Westeuropa), die im Rahmen des HapMap-Projektes veröffentlicht wurden (The International HapMap Consortium, 2005), bestätigt sich dieses Bild.

Das Internationale HapMap Konsortium hat Muster von mehr als 1 Mio SNPs im Genom von 269 Menschen vier verschiedener Populationen katalogisiert. Die Daten zeigen

Muster chromosomaler Strukturen, Rekombinations-„Hot Spots“, LD-Blockstrukturen und Haplotypverteilungen. Die HapMap liefert damit einen Überblick über „häufige“, d.h. bei Stichproben einiger Hundert Individuen nachweisbare, genetische Unterschiede zwischen Menschen. Aus den Untersuchungen wird deutlich, dass LD in den Bereichen 1-100 kb merklich variiert und eher unstetig als sinkend im Vergleich zu steigendem genomischen Abstand ist (The International HapMap Consortium, 2005).

Die Allelfrequenzen und Werte paarweisen LDs der Berliner Stichprobe stimmen mit den Werten der CEPH-Population (HapMap) in großem Maße überein.

Der Vergleich basiert auf 56 gemeinsamen SNPs. Die SNP-Dichte in untersuchten Genen der HapMap ist bis auf *APOE* und das APO-A-V/-A-IV/-C-III/-A-I-Gencluster größer. Der Abstand zwischen zwei gemessenen SNP-Positionen in hier untersuchten Kandidatengenen beträgt in der HapMap durchschnittlich 400 bp.

Die HapMap-Datenbank liefert das bislang umfangreichste Bild humaner allelischer Vielfalt. In weiterführenden Analysen kann nunmehr der Anteil genetischer Variabilität bestimmt werden, der durch hier genotypisierte SNPs relativ zu HapMap-Genotypen abgedeckt wird. Dies ist möglich, solange gemessene SNPs auch in den HapMap-Stichproben untersucht wurden. Je höher die Korrelation ist, desto geringer ist der Informationsverlust.

Da LD-Mapping sehr stark von der komplexen Vorgeschichte von Mutation, Rekombination, Migration, Bevölkerungsengpässen und geographischer Isolation abhängt, muss man letztendlich auf eine verbindliche Deutung von Assoziation verzichten. Auch muss eingeräumt werden, dass der Weg von der Genvariante zum Lipidphänotyp zwar „kürzer“ ist als der Weg vom Gen zu ausgebildeten Folgeerkrankungen, jedoch ist er nur indirekt messbar, da die Regulation von Transkription, Translation und Proteinabbau auch bei einem intermediären Phänotyp nicht untersucht wird. Trotzdem:

Es lohnt sich, nach Polymorphismen zu suchen, zumal bei Menschen Expressions- oder Proteomic-Daten aus relevanten Organen oder Zellen der direkten Untersuchung in epidemiologischen Kohorten nur schwer zugänglich sind.

Mit Gentranskriptions- und Proteinmustern wird untersucht, wie stark Gene in untersuchten Zellen exprimiert bzw. in Proteine umgesetzt (translatiert) wurden. Auf Grund der Stärke der Expression und Translation als quantitativem Maß lässt sich ein funktioneller Zusammenhang mit untersuchten Phänotypen modellieren (für Methoden siehe u.a. (Mansmann & Meister, 2005)). Das Studium genetischer Polymorphismen schließt diese Analyse indirekt ein, indem davon ausgegangen wird, dass Mutationen in Kandidatengenen die Genexpression maßgeblich beeinflussen und sich auf den Phänotyp auswirken. Es ist dabei

jedoch in hohem Maße mit Unschärfe zu rechnen. Des Weiteren setzt die Modellierung multipler genetischer Marker vereinfachte Modelle voraus, u.a. die Vernachlässigung möglicher Interaktionen, Annahme von Linearität und/oder Additivität von Geneffekten usw.

Zusammenfassend kann man sagen, dass die letzten 10 Jahre eine ganz bedeutende Zeit für die Humangenetik waren. Sie sind durch ein enormes Wachstum der technischen Kapazitäten und des genomischen Wissens gekennzeichnet (Smith et al., 2005; Sperling, 2000). Die statistischen Kapazitäten und die Fähigkeit, Daten zu verarbeiten und zu interpretieren, liegen jedoch immer noch hinter den technischen Möglichkeiten zur Genotypisierung großer Mengen genomischer Daten zurück.

Wo steht man in der Aufdeckung von Genen komplexer Erkrankungen? Die meisten komplexen Erkrankungen schließen mit großer Sicherheit viele Gene ein, die prädisponierend sind und eher kleine individuelle Effekte haben (Propping & Nothen, 1995). Wechselwirkungen zwischen Genen sowie Genen und Umweltfaktoren spielen eine Rolle und die Heterogenität zwischen verschiedenen Populationen bezüglich genetischen und umweltbezogenen Risikofaktoren wirkt sich maßgeblich auf die Ergebnisse aus. Alle diese Größen beeinträchtigen die statistische Power der Analysen zum Teil erheblich. Folglich ist die Aufdeckung von genetischen Effekten und deren Bestätigung schwierig. Der Gewinn liegt in dem Einblick in die Schwierigkeit und Komplexität des Mechanismus komplexer Erkrankungen. So stellen beispielsweise die CD/CV-Hypothese und die Heterogenitätshypothese zwei völlig verschiedene Grundtheorien auf. Eine Fragestellung dieser Art hat es vorher nicht gegeben.

In der vorliegenden Arbeit liegt der Fokus auf dem Studium und dem Vergleich von Gen-Lokus-Effekten. In methodischer Hinsicht stimmt die Herangehensweise mit dem von Neale & Sham (2004) publizierten Vorschlag überein. Insbesondere der Vergleich der Assoziationsbefunde mit Genfer Studienergebnissen zeigte annähernd konkordante Befunde, obwohl einzelne nominal signifikante Effekte sich selten oder nur ihrer Tendenz nach bestätigen. Die „Common Disease/Common Variants“ - Hypothese kann demnach auf dem Niveau von Gen-Loci unterstützt werden, jedoch die Aufspaltung nach einzelnen SNP-Genotypen würde einen weit größeren Datenumfang benötigen.

Sowohl genotypisch als auch phänotypisch bietet das Studienmaterial die besten Voraussetzungen für die empirische Validierung des kinetischen Modells des Lipoproteinstoffwechsels, wie von Knoblauch et al. (2000) beschrieben. Die Schwierigkeit der Arbeit liegt in der Hypothese selbst. Die genetische Ursache kleiner Abweichungen vom „normalen“ Lipidstatus einer Person aufzudecken, die in hohem Maße von den Lebensbedingungen, der Ernährung, sportlicher Aktivität etc. abhängen, ist schwierig und die statistische Power sinkt mit der steigenden Anzahl betrachteter Einflussgrößen.

Der Lipoproteinstoffwechsel wird außer durch die hier untersuchten direkten enzymati-

schen und rezeptorellen Faktoren zusätzlich durch den ebenfalls genetisch determinierten Gesamtstoffwechsel und die Entwicklungsbiologie des Organismus beeinflusst. Bei vielen Teileffekten ist dieser Unterschied nicht vorhanden oder nicht sichtbar, bei anderen zeigt er sich sehr deutlich. Ein deutlicher Befund wäre auch bei einer genaueren Untersuchung des Zusammenhangs von Cholesterinwerten mit dem Alter und dem BMI der Probanden zu erwarten. Auch hier gilt: Es mendeln zwar die gleichen Genvarianten in den Familien - diese sind unabhängig vom Alter vorhanden - aber im Gesamtmechanismus kann ihre Funktion sich gleichwohl sehr unterschiedlich ausprägen.