Aus dem Institut für Biochemie
der Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

Anwendung und Entwicklung biostatistischer Methoden zur Identifikation genetischer Risikofaktoren

zur Erlangung des akademischen Grades
Doctor rerum medicarum (Dr. rer. medic.)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Sven Knüppel

aus Berlin

Datum der Promotion: 04.09.2015

Inhaltsverzeichnis

1. Zusammenfassung

Hintergrund: Viele der häufigsten chronischen Erkrankungen werden durch genetische und Umweltfaktoren beeinflusst. Durch die Analyse einzelner genetischer Marker, beispielsweise SNPs (single nucleotide polymorphisms), kann ein kleiner Teil der genetischen Ursache erklärt werden. Um den Effekt mehrerer Marker und deren Kombination besser einschätzen und damit tiefergehende Einblicke in der genetischen Ursachenforschung gewinnen zu können, ist die Anwendung und Entwicklung geeigneter biostatistischer Methoden essentiell.

Zielstellung: Ziel dieser Arbeit war es, biostatistische Methoden zur Identifizierung von genetischen Risikofaktoren basierend auf einzelnen SNPs und deren Kombination zu entwickeln und anzuwenden, um systematisch informative krankheitsassoziierte Allelkombinationen, wie z.B. Haplotypen, aus einer Vielzahl von SNPs zu identifizieren. Daneben wurden drei Kandidatengenstudien zur Evaluierung vielversprechender Gene begleitet.

Methodik: Die Multi-locus Stepwise Regression (MSR) wurde im Rahmen der zugrundeliegenden deutschen genomweiten Studie zur atopischen Dermatitis entwickelt. Die MSR vereinigt die Vorteile schrittweiser Auswahlverfahren und Haplotypanalysen. Dabei werden SNP-Kombinationen sukzessive um jeweils einen SNP erweitert, wenn sich das Ergebnis der Haplotypanalyse statistisch verbessert. Zusätzlich wurde die MSR auf ungekoppelte SNPs im Rahmen der EPIC-Potsdam-Studie angewendet.

Für diese Arbeit standen Daten aus einer deutschen genomweiten Studie zur atopischen Dermatitis bestehend aus einer Fall-Kontroll-Studie (939 Fälle, 975 Kontrollen) und einer Familienstudie (268 Familien mit 529 erkrankten Kindern) zur Verfügung. Es wurden 94 tagSNPs der EDC-Region auf Chromosom 1q21 und vier bekannte *FLG*-Mutationen, welche Strukturproteine für den Verhornungsprozess der menschlichen Epidermis kodieren, eingeschlossen.

In der EPIC-Potsdam-Studie wurden vier Teilstudien durchgeführt: [1] 41 SNPs auf Body-Mass-Index (kg/m²) und Taillenumfang (Querschnittsanalyse, MSR, Permutationstest), [2] 2 SNPs (*ADH1B*, *ADH1C*) als Marker für Alkoholaufnahme auf kardiovaskuläre Erkrankungen (Fall-Kohorten-Design, modifierte Cox-Regression), [3] 1 SNP des *MTTP*-Gens, dessen kodiertes Protein eine zentrale Rolle im Lipoproteinstoffwechsel spielt, auf kardiovaskuläre Erkrankungen unter Berücksichtigung der Gesamtcholesterinaufnahme (Fall-Kohorten-Design, modifierte Cox-

Regression) und [4] 7 tagSNPs des *SCD1*-Gens, das ein Protein des Fettstoffwechsels kodiert, auf metabolische Risikofaktoren (Querschnittsanalyse, Kovarianzanalyse).

Ergebnisse: Die Anwendung der MSR in der genomweiten Studie ergab in der Fall-Kontroll-Studie ein Haplotyp-Muster, dass in der Familienstudie repliziert werden konnte. Dieses Haplotyp-Muster bestehend aus 4 SNPs zeigte den bekannten *FLG*-Effekt und einen zusätzlichen *FLG*-unabhängigen Effekt auf die atopische Dermatitis.

In der EPIC-Potsdam-Studie identifizierte die MSR mit Body-Mass-Index und Taillenumfang assoziierte SNP-Kombinationen, die sich unter Berücksichtigung der simulierten Null-Verteilung (Annahme keines genetischen Effektes) als nicht signifikant herausstellten.

Die SNPs der Gene *ADH1B* und *ADH1C* beeinflussten das Risiko für Myokardinfarkt und Schlaganfall nicht. Andererseits wurde eine Interaktion zwischen *MTTP*-SNP rs1800804 und Gesamtcholesterin für die kardiovaskulären Erkrankungen beobachtet. Zusammenhänge zwischen *SCD1*-SNPs und den dazugehörigen Haplotypen mit den untersuchten metabolischen Risikofaktoren wurden nicht festgestellt.

Schlussfolgerung: Eine schrittweise haplotyp-basierte SNP-Selektion wurde in dieser Arbeit entwickelt und konnte erfolgreich in einer Kandidatengenregion angewendet werden. Die Anwendung auf ungekoppelte SNPs erforderte eine besondere Berücksichtigung des Suchprozesses. Obwohl die verwendeten Daten exemplarisch das Vorgehen der MSR zeigen, konnten in dieser Arbeit keine wesentlichen Effekte über die Einzel-SNP-Analyse hinaus gefunden werden. Weitere Studien sind erforderlich, um die MSR weiter zu entwickeln und zu beurteilen. In den Kandidatengenstudien waren ebenfalls keine zusätzlichen Multi-Locus-Marker-Effekt zu beobachten.

2. Abstract

Background: Most common chronic diseases are influenced by genetic and environmental factors. Analysis of single genetic markers, such as SNPs (single nucleotide polymorphisms), explain a small part of genetic causes. The application and development of biostatistical methods are necessary to get further insights into the genetic causes of chronic diseases by taking into consideration of single and multiple genetic markers.

Objectives: The objective was to develop biostatistical methods to identify genetic risk factors using single SNPs and their combinations and to apply these methods to select systematically risk-related allele combinations, such as haplotypes, from a large number of SNPs. In addition, three candidate gene studies were carried out to evaluate the effect of different promising genes.

Methods: The Multi-locus stepwise regression (MSR) method was developed using data of a German genome-wide association study on atopic dermatitis. The MSR combines the advantages of stepwise selection methods with haplotype-based approaches. The MSR extends stepwise SNP-combinations successively if the result of the haplotype-based test is statistically improved until a stop criterion is met. The MSR was subsequently applied to investigate unlinked SNPs and their combinations as part of the EPIC-Potsdam study.

The German genome-wide association study of atopic dermatitis consists of a case-control study (939 cases and 975 controls) and a family study (268 families with 529 children). 94 tagSNPs of EDC region on chromosome 1q21 and four known *FLG*-mutations encoding structural proteins that are expressed during terminal differentiation of the human epidermis were used

Within the EPIC-Potsdam study the following associations were investigated in four sub-studies: [1] 41 SNPs for body-mass index (kg/m²) and waist circumference (cross-sectional study, MSR, and permutation test), [2] 2 SNPs (*ADH1B*, *ADH1C*) as marker for alcohol intake in relation to incident cardiovascular diseases (case-cohort study, modified Cox proportional hazards regression), [3] 1 SNP from *MTTP* gene that encodes a lipid transfer protein for cardiovascular diseases under consideration of cholesterol levels (case-cohort study, modified Cox proportional hazards regression), and [4] 7 tagSNPs from *SCD1* gene encoding a protein that is involved in lipid metabolism for metabolic risk factors (cross-sectional study, analysis of covariance).

Results: The MSR used in the genome-wide association study identified a haplotype pattern in the case-control study and was replicated in the family study. This haplotype pattern of four SNPs reflects the well-known *FLG* effect and an additional *FLG*-independent effect on atopic dermatitis.

In the EPIC-Potsdam study, the MSR identified SNP-combinations associated with body-mass index and waist circumference, but these were not statistically significant compared to the simulated distribution under the null hypothesis of no genetic effect.

The SNPs in *ADH1B* and *ADH1C* showed no associations to risk of cardiovascular diseases. An interaction was observed between *MTTP*-SNP rs1800804 and cholesterol levels on cardiovascular diseases. No evidence for an effect of single *SCD1*-SNPs and their corresponding haplotypes on metabolic risk factors was found.

Conclusion: A stepwise haplotype-based SNP selection method was developed and successfully applied to one candidate gene region. The application of the method to unlinked SNPs requires a careful selection of SNPs. Although the data that were used to exemplify the method did not add additional evidence concerning the genetic etiology of combined SNP effects beyond single SNP analysis. Further studies are needed to assess the real value of the MSR. In the candidate gene studies as well none essential multi-locus marker effects were found.

3. Einführung

Die Integration von epidemiologischen Methoden und Techniken der Genomforschung kann zu einem besseren Verständnis der komplexen genetischen Ätiologie von Krankheiten beitragen. Die Entdeckung genetischer Einflussfaktoren soll dabei helfen, neue Einblicke in die Pathogenese, Diagnose und Behandlung von Krankheiten zu gewinnen. Neben der technischen Möglichkeit genomweiter Analysen und der damit verbundenen statistisch-methodischen Weiterentwicklung sind Replikationsstudien und Meta-Analysen identifizierter genetischer Marker ebenso wichtig, um einen genetischen Effekt in verschiedenen Populationen zu beurteilen. Die Identifikation und Replikation vielversprechender Marker wird durch das komplexe Zusammenspiel der Gene untereinander und mit der Umwelt erschwert.

In den letzten Jahren hat die Genomforschung mit der vollständigen Sequenzierung des menschlichen Genoms durch das Humangenomprojekt (Venter 2001) einen enormen Aufschwung erfahren. Das zeigt sich vor allem in der Vielzahl gefundener Assoziationen in genomweiten Studien (Hindorff 2013), deren Replikationen und Bestimmung ihrer biochemischen Funktionsweisen weiter vorangehen.

Die technische Weiterentwicklung ermöglicht die Sequenzierung mehrerer hunderttausend genetischer Marker, den sogenannten SNPs (single nucleotide polymorphisms), die durch den Basenaustausch an einer bestimmten Stelle des Genoms gekennzeichnet sind. Diese hochdimensionalen Daten erlauben die Analyse einzelner SNPs und deren Kombination, um Gene zu identifizieren, die biologisch relevant sind und das Erkrankungsrisiko beeinflussen.

Im Laufe der letzten Jahre wird immer deutlicher, dass einzelne SNPs meist nur einen kleinen Teil des genetischen Gesamteffektes erklären. Mit der Kombination von mehreren SNPs wird eine Verbesserung der Suche nach einem genetischen Effekt erhofft. Es wurde gezeigt, dass die Kombination von Allelen zu Haplotypen, die die Allele entlang eines Chromosoms darstellen, zur Erhöhung der Evidenz führen und die statistische Power erhöhen können (Akey 2001). Diese stellen außerdem die Strukturierung der genetischen Variabilität dar und repräsentieren damit biologisch relevante Informationen. Es wurden Verfahren vorgeschlagen, die auf solchen sogenannten phasen-unbekannten Multi-Locus-Marker-Haplotypen basieren (Schaid 2002).

Die besondere Herausforderung besteht darin, krankheitsrelevante Haplotypen aus den informativen SNPs einer chromosomalen Region zu bilden und deren Effekt zu quantifizieren.

Dies stellt ein kombinatorisches Problem dar, denn aus k nebeneinanderliegenden SNPs können 2^k–1 mögliche SNP-Kombinationen gebildet werden, die, wenn k groß genug ist, nicht alle untersucht werden können. Beispielsweise könnten 40 SNPs zu 2^{40}–1=1,1·10^{12} verschiedenen SNP-Kombinationen gruppiert werden. Wenn ein entsprechender statistischer Test eine Sekunde dauern würde, bräuchte man etwa 34865 Jahre um all diese Kombinationen zu untersuchen. Daher ist die Entwicklung biostatistischer Verfahren zur systematischen Auswahl relevanter SNP-Kombination notwendig.

Neben der Suche nach vielversprechenden genetischen Risikomarkern und deren Kombinationen ist deren Replikation besonders wichtig, denn nur so kann langfristig ein SNP oder mehrere SNPs zur Risikoprädiktion herangezogen werden. In genomweiten Studien werden oftmals wenige relevante SNPs identifiziert, die dann in weiteren Studien untersucht werden. In den Replikationsstudien können sowohl einzelne als auch mehrere SNPs einbezogen werden. Wenn mehrere SNPs untersucht werden, können sie gekoppelt (gemeinsame Vererbung) oder ungekoppelt (unabhängig) sein. Die gleichzeitige Analyse mehrerer SNPs könnte über die Einzel-SNP-Analyse hinaus helfen, bestimmte SNP-Kombinationen zu identifizieren, die mit ausgewählten Krankheiten oder anderen komplexen Merkmalen assoziiert sind.

4. Zielstellung

Die vorliegende Publikations-Dissertation umfasst die Anwendung und Entwicklung biostatistischer Methoden zur Quantifizierung von genetischen Effekten innerhalb von chromosomalen Regionen oder Kandidatengenen.

Im Rahmen dieser Arbeit stand die Suche nach krankheitsrelevanten SNP-Kombinationen im Vordergrund. Dazu wurde die Multi-locus Stepwise Regression (MSR) entwickelt. Die MSR verbindet die klassische schrittweise Variablenselektion und die Methoden der haplotyp-basierten Assoziationsverfahren miteinander, um systematisch informative krankheitsassoziierte Haplotypen oder allgemein Allelkombinationen zu identifizieren.

Nach der erfolgreichen Anwendung der MSR in einer genomweiten Studie wurde die systematische Suche der MSR auch auf ungekoppelte SNPs erweitert.

Von mehreren begleiteten Kandidatengenstudien sind drei in dieser Arbeit berücksichtigt. Letztere hatten zum Ziel, den Effekt vielversprechender genetischer Marker auf kardiovaskuläre Erkrankungen zu überprüfen.

Die in dieser Promotion eingeschlossenen Publikationen umfassen die eigenständig entwickelte Multi-locus Stepwise Regression (MSR) und deren Anwendung sowie die Koautorenschaft bei drei weiteren Artikeln.

5. Methodik

5.1 Entwicklung des R-Paketes HapEstXXR

Ein Teil der Analysen dieser Arbeit wurden mit der Statistiksoftware R und verschiedenen selbstprogrammierten C-Programmen durchgeführt, die zu diesem Zweck zu dem R-Paket HapEstXXR zusammengeführt und weiter entwickelt wurden.

Die in Zusammenarbeit mit Klaus Rohde in C programmierten Algorithmen zur Bestimmung von Haplotypen und Allelkombinationen für unverwandte Individuen und Kleinfamilien wurden bei der in R entwickelten Multi-locus Stepwise Regression (MSR) zur schrittweisen Selektion von Haplotypen und Allelkombinationen eingesetzt. Die Beschreibung der MSR folgt im Abschnitt 5.3.

Die Routinen im R-Paket HapEstXXR (aktuelle Version 0.1-7, http://cran.r-project.org/web/packages/HapEstXXR) sind frei verfügbar. Die Schätzung für unverwandte Individuen entspricht dem Expectation-Maximization-Algorithmus von Excoffier (1995). Die Haplotyp-Schätzung für Kleinfamilien wurde von Rohde (2001) entwickelt.

5.2 Studienpopulationen

Es wurden sechs Studien einbezogen:

Studie		Gene (SNPs)	Zielgrößen
1	**Deutsche genomweite Studie**		
	Fall-Kontroll- und Familienstudie	EDC-Cluster auf Chr. 1q21 (94 tagSNPs)	Atopische Dermatitis

Studie			Gene (SNPs)	Zielgrößen
2	**EPIC-Potsdam-Studie**			
	Teilstudie 1	Querschnittsstudie	18 Gene (41 SNPs)	BMI und Taillenumfang
	Teilstudie 2	Fall-Kohortenstudie	*ADH1B* und *ADH1C* (2 SNPs)	Alkoholkonsum, Myokardinfarkt und Schlaganfall
	Teilstudie 3	Fall-Kohortenstudie	*MTTP* (1 SNP)	Myokardinfarkt und Schlaganfall unter Berücksichtigung der Aufnahme von Gesamtcholesterin
	Teilstudie 4	Querschnittsstudie	*SCD1* (7 tagSNPs)	Metabolische Risikofaktoren

Studie 1: Deutsche genomweite Studie

Es standen die Daten einer deutschen genomweiten Studie zu atopischer Dermatitis (AD) am Max-Delbrück-Centrum für Molekulare Medizin (MDC) Berlin-Buch zur Verfügung (Esparza-Gordillo 2009). Innerhalb dieser groß angelegten Studie wurden eine Fall-Kontroll-Studie (939 AD-Fälle und 975 Kontrollen) und eine Familienstudie (268 Kleinfamilien mit 1097 Familienmitgliedern und 529 erkrankten Kindern) durchgeführt. Die AD-Fälle der Fall-Kontroll-Studie wurden an vier deutschen Universitätskliniken (Charité Universitätsmedizin Berlin, Universität zu Kiel, Technische Universität München und Universität Bonn) rekrutiert. Alle Fälle wurden nach Standardkriterien durch ärztliche Diagnose in der Zeit von 2003 bis 2006 ausgewählt. Alle Fälle sind deutscher Herkunft (Selbstangabe). Die Kontrollen wurden zufällig aus der Gen-Datenbank PopGen ausgewählt. An der Charité Universitätsmedizin Berlin wurden Familien für die Familienstudie gewonnen. Einschlusskriterium war, dass in den Familien mindestens zwei erkrankte Geschwister mit Erkrankungszeitpunkt vor dem zweiten Lebensjahr vorhanden waren.

Die Genotypisierung in dieser Studie wurde mit den kommerziellen Affymetrix-Chips 500k und 5.0 durchgeführt. Als vielversprechende chromosomale Region wurde der Epidermal Differentiation Complex (EDC) auf Chromosom 1q21 ausgewählt, da dort bereits Geneffekte gefunden wurden (Esparza-Gordillo 2009). Die ausgewählte EDC-Region umfasste 259 SNPs

(Chromosom 1: 150.075.690-152.014.240 bp). Es wurden 94 tagSNPs (LD-Kriterium r^2>0,8) ausgewählt, um so redundante Informationen durch Linkage Disequilibrium zu vermeiden. Zusätzlich standen vier bereits bekannte *Filaggrin* (*FLG)*-Mutationen (2282del4, R501X, R2447X und S3247X) zur Verfügung. Im *FLG*-Gen wird ein Protein kodiert, das beim Verhornungsprozess der Haut durch Bildung von Keratinozyten beteiligt ist.

Studie 2: EPIC-Potsdam-Studie

Die European Prospective Investigation into Cancer and Nutrition (EPIC)-Studie ist eine andauernde, europaweit angelegte prospektive Studie in 10 Ländern an 23 Studienzentren mit dem Ziel, Zusammenhänge zwischen Ernährung, Lebensstilfaktoren und chronischen Erkrankungen (Krebs, kardiovaskuläre Erkrankungen, darunter Myokardinfarkt und Schlaganfall, und Diabetes-Typ-2) weiter aufzuklären. Im deutschen Studienzentrum EPIC-Potsdam wurden vier Kandidatengenstudien hypothesenorientiert durchgeführt.

Die Basiserhebung der EPIC-Potsdam-Studie von 27548 Personen im Alter von 35 bis 65 Jahren fand zwischen 1994 und 1998 statt und umfasste die Erhebung anthropometrischer Maße (u.a. Größe, Gewicht und Taillenumfang), Messung des Blutdrucks, Entnahme einer Blutprobe, selbstständiges Ausfüllen eines validierten Ernährungshäufigkeitsfragebogens und ein persönliches Interview über Lebensstilfaktoren und Krankheitsgeschichte (Boeing 1999). Alle zwei Jahre werden Nachbeobachtungen mittels Fragebogen durchgeführt, um Neuerkrankungen zu registrieren. Alle Studienteilnehmer unterschrieben eine Einverständniserklärung. Die Ethikkommission der Landesärztekammer Brandenburg genehmigte das Studienprotokoll. Eine Zufallsstichprobe aus 2500 Studienteilnehmern, die Blut bei der Basiserhebung gespendet hatten, wurde für weiterführende Teilstudien ausgewählt. Es wird für die Zufallsstichprobe angenommen, dass deren Eigenschaften durch das zufällige Ziehen auf die volle Kohorte übertragbar sind und Effekte unverzerrt geschätzt werden können. Diese Stichprobe wurde in dieser Arbeit in zwei Querschnittsanalysen zur Basiserhebung und in zwei Fall-Kohortenstudien einbezogen. Bei dem Fall-Kohorten-Design wurden neben der Zufallsstichprobe alle erkannten Fälle eines Endpunktes in die Analyse eingeschlossen (Prentice 1986).

Alle SNPs der EPIC-Potsdam-Studie wurden mit den TaqMan SNP Genotyping Assay Sets (ABI, Forster City, 141 CA, USA) genotypisiert.

Teilstudie 1

Nach Ausschluss von Personen aus der Zufallsstichprobe mit mehr als fünf fehlenden Genotypen und prävalenten Erkrankungen standen 2122 Personen für diese Querschnittsanalyse in der Zufallsstichprobe zur Verfügung. Als Surrogatmessungen für Übergewicht wurde der genetische Effekt auf den Body-Mass-Index (BMI, kg/m²) und den Taillenumfang (cm) untersucht.

Für diese Studie wurden 41 SNPs aus 18 Genen über 11 Chromosomen aufgrund in genomweiten oder Kandidatengenstudien publizierten Assoziationen zu Untergewicht, Übergewicht und entsprechenden anthropometrischen Maßen ausgewählt: *LEPR* (rs1137100, rs1137101 und rs8179183), *HSD11B1* (rs4844880, rs846910 und rs3753519), *TMEM18* (rs11127485), *FABP1* (rs2241883), *INSIG2* (rs7566605), *ALPI* (rs3762521), *PPARG* (rs1801282), *TBC1D1* (rs2279027, rs35859249, rs4832743, rs10517456, rs9999507, rs6845120, rs6823014, rs10009706, rs2303422, rs1344603, rs637797, rs6837834 und rs13110318), *MTTP* (rs3816873), *FABP2* (rs1799883, rs6857641), *PTGES2* (rs13283456), *TCF7L2* (rs7903146), *ABCC8* (rs916829, rs916828, rs2237984, rs10832786, rs7106053 und rs11024286), *IGF1* (rs1520220), *FTO* (rs9939609), *SREBF1* (rs2297508), *NPC1* (rs1805081), *MC4R* (rs17700144 und rs10871777).

Teilstudie 2

Nach Ausschluss von Personen mit prävalenten Erkrankungen (Myokardinfarkt und Schlaganfall) und fehlenden Werte bei den einbezogenen Faktoren und SNPs wurden 230 Myokardinfarkte und 208 Schlaganfälle eingeschlossen (mittlere Nachbeobachtungszeit = 8,2 Jahre). Die Zufallsstichprobe umfasst 2175 Personen und darunter ebenfalls 60 Myokardinfarkte und Schlaganfälle.

Zwei SNPs in den Genen *ADH1B* (rs1229984) und *ADH1C* (rs698), die Enzyme der Alkoholdehydrogenase-Familie kodieren und den Abbau von Alkoholen im menschlichen Körper katalysieren, wurden für diese Studie genotypisiert.

Teilstudie 3

In dieser Studie wurden dieselben Fälle wie in Teilstudie 2 ermittelt. Nach Ausschluss von Personen mit zur Basis prävalenten Erkrankungen (Myokardinfarkt und Schlaganfall) und fehlenden Werten in eingeschlossenen Merkmalen oder SNP-Genotypen konnten 2302 Personen in die Analyse einbezogen werden, darunter waren 193 mit Myokardinfarkt, 131 mit

Schlaganfall und 1978 ohne inzidenten Myokardinfarkt oder Schlaganfall. 34 Fälle wurden in der Zufallsstichprobe ausgewählt.

Zur Replikation der Ergebnisse konnte die Heinz Nixdorf Recall Studie (HNR Studie) gewonnen werden. Die HNR Studie startete 2000 und rekrutierte 4814 Teilnehmer im Ruhrgebiet. Durch Zufallsauswahl und Ausschluss (prävalente Fälle und fehlende Angaben) standen 1188 Nichtfälle und 30 kardiovaskuläre Erkrankungen zur Verfügung.

Für diese Studie wurde im *MTTP*-Gen der Polymorphismus -164T > C (rs1800804) ausgewählt, da dieser bei der Kodierung des mikrosomalen Triglycerid-Transferproteins (MTTP) mitwirkt, welches eine zentrale Rolle im Lipoproteinstoffwechsel spielt.

Teilstudie 4

In dieser Querschnittsstudie wurden Personen mit fehlenden Angaben oder fehlenden Genotypen ausgeschlossen. Es konnten 2157 von 2500 Personen der Zufallsstichprobe in die Querschnittsanalyse eingeschlossen werden.

Es wurden sieben tagSNPs aus dem Gen *SCD1* (rs1502593, rs522951, rs11190480, rs3071, rs3793767, rs10883463 und rs508384) genotypisiert. Im *SCD1*-Gen wird ein Protein (Stearoyl-CoA desaturase-1) kodiert, das eine zentrale Rolle im Fettstoffwechsel spielt.

5.3 Statistische Analyse

Die in dieser Arbeit entwickelte Multi-locus Stepwise Regression (MSR) verbindet die Idee schrittweiser Variablenselektion und haplotyp-basierter Assoziationsanalyse, wodurch die Vorteile beider Verfahren genutzt werden, um (sub-)optimale SNP-Kombinationen zu identifizieren.

Die MSR basiert darauf, dass schrittweise zu bereits ausgewählten SNP-Kombinationen einzeln noch nicht eingeschlossene SNPs hinzugefügt werden, auf Assoziation getestet und nur bei einer statistischen Verbesserung weiter verfolgt werden. Dieser allgemeine Ansatz ist sehr flexibel. Der Algorithmus kann entweder bei der Einzel-SNP-Analyse oder bei SNP-Paaren gestartet werden. Die jeweils angewendete statistische Methode zur Schätzung von genetischen Effekten, wie beispielsweise die Regressions- und Ereigniszeitanalyse, kann je nach Forschungsfrage angepasst werden. In der genomweiten Studie wurde die logistische Regression und in der EPIC-

Potsdam-Studie die klassische lineare Regression angewendet. Bei allen Tests wurde die jeweilige Designmatrix basierend auf individuell geschätzten Haplotypen erstellt. Im Fall der EPIC-Potsdam-Studie mit ungekoppelten SNPs wurden nicht direkt Haplotypen, sondern spezifische Allelkombinationen geschätzt. Bei allen Studien wurde der globale p-Wert mit einem globalen Test bestimmt. Dabei wurde die Likelihood des Null-Modells mit der Likelihood des „vollen“ Modells verglichen. Das Null-Modell umfasste immer alle eingeschlossenen nicht-genetischen Faktoren und das volle Modell alle genetischen und nicht-genetischen Faktoren. Im Fall der logistischen Regression wurde ein Likelihood-Ratio-Chi^2-Test und im Fall der linearen Regression ein F-Test durchgeführt.

Bei jedem Schritt der Erweiterung der SNP-Kombinationen wurde für jede Studie ein angepasstes Selektionskriterium eingesetzt, mit dem beurteilt wurde, ob sich eine Verbesserung durch den hinzugenommenen SNP ergeben hat. In der genomweiten Studie lag die Herausforderung darin, die sehr hohe Zahl einflussreicher Haplotypen in der EDC-Region auf die informativen Haplotypmuster zu beschränken. Aus diesem Grund wurden in jedem Schritt die besten 300 SNP-Kombinationen (jeweils kleinsten p-Wert) solange erweitert, bis sich im Mittel die mit den dekadischen Logarithmus transformierten p-Werte der zehn besten SNP-Kombinationen um weniger als 10% verbesserten.

Bei der EPIC-Potsdam-Studie zu BMI und Taillenumfang war der Fokus, möglichst falsch-positive Resultate zu vermeiden. Es wurden diejenigen SNP-Kombinationen ausgewählt, die eine Verbesserung des corrected Akaike’s information criterion AIC_c (Sugiura 1978) zeigten und unterhalb einer Signifikanzgrenze lagen. Diese Grenze wurde je nach Zahl der einbezogenen SNPs vorgegeben. Für SNP-Paare lag die Grenze bei 0,05, bei Tripeln 0,01, bei 4-SNP-Kombinationen bei 0,001 usw. Die Suche begann mit der Analyse aller SNP-Paare.

Bei der Anwendung der MSR als schrittweises Selektionsverfahren wurden viele statistische Tests durchgeführt und nur die „signifikanten“ SNP-Kombinationen weiter verfolgt, da ansonsten eine Inflation der p-Werte zu erwarten ist und damit das Problem des multiplen Testens entsteht. Um den Selektionsprozess unter der Nullhypothese (kein genetischer Effekt) zu beurteilen, wurde in der Teilstudie 1 (BMI und Taillenumfang) ein Permutationstest durchgeführt, um so die Verteilung unter der Nullhypothese studienbedingt zu simulieren.

In den Teilstudien 2 und 3 wurden die ausgewählten SNPs im Fall-Kohorten-Design erhoben, sodass die Cox-Regression modifiziert nach Prentice (1986) angewendet wurde.

In Teilstudie 2 wurde zusätzlich eine Meta-Analyse mit prospektiven Studien zu den Genen *ADH1B* und *ADH1C* mit dem R-Paket meta durchgeführt.

In Teilstudie 4 wurde der genetische Effekt der sieben tagSNPs einzeln und mittels Haplotypen mit der ANCOVA (Analysis of Covariance) bestimmt. Die individuellen Haplotypen wurden mit dem Expectation-Maximization-Algorithmus geschätzt.

In jeder Studie stellt sich die Frage nach einer geeigneten Adjustierung oder Stratifizierung zur Vermeidung von unerwünschten Verzerrungen. Die Anwendung der MSR in der genomweiten Studie erfolgte ohne weitere Berücksichtigung von Adjustierungsvariablen, was zum Zeitpunkt der Durchführung ein übliches Vorgehen in der genetischen Epidemiologie war. In allen anderen Studien wurden verschiedene Adjustierungsvariablen gewählt (Alter, Geschlecht u. a.).

5.4 Software

Die tagSNPs der genomweiten Studie wurden mit Paul de Bakkers‘-Programm Tagger, eingebunden in Haploview (Version 4.2, http://www.broadinstitute.org/haploview), ausgewählt. Informationen zur SNP-Lokalisation wurden der ehemaligen Datenbank SNPselector (NCBI assembly 36; dbSNP build 126; seit 30.09.2010 nicht mehr verfügbar) entnommen. Die statistische Analyse der genomweiten Studie wurde mit der Statistiksoftware R (Version 2.11.1) durchgeführt.

Die Anwendung der MSR in der EPIC-Potsdam-Studie wurde mit der Statistiksoftware R (Version 2.14.0) unter Anwendung der grafischen Benutzeroberfläche RStudio (Version 0.96.330, http://www.rstudio.com) durchgeführt.

Die begleitenden Studien wurden mit der Statistiksoftware SAS (Version 9.2, SAS Institute Inc., Cary, NC) analysiert.

6. Ergebnisse

6.1 Suche nach Haplotyp-Mustern in einer genomweiten Studie zur atopischen Dermatitis

Publikation 1: Knüppel S, Esparza-Gordillo J, Marenholz I, Holzhütter HG, Bauerfeind A, Ruether A, Weidinger S, Lee Y-A, Rohde K. Multi-locus stepwise regression: a haplotype-based algorithm for finding genetic associations applied to atopic dermatitis. BMC Med Genet 2012;13(1):8.

In der deutschen genomweiten Studie zur atopischen Dermatitis wurden 94 tagSNPs im Epidermal Differentiation Complex (EDC) auf Chromosom 1q21 ausgewählt. Die Anwendung der MSR auf die tagSNPs ergab ein bestes signifikantes Haplotypmuster, das in der Fall-Kontroll-Studie (p=4,13E-07 nach Bonferroni-Korrektur) mit 939 AD-Fällen und 975 Kontrollen identifiziert wurde und in der Familienstudie (p=0,0398 nach Bonferroni-Korrektur) mit 268 Kleinfamilien repliziert werden konnte. Dieses Muster umfasste die SNPs rs7550106, rs499697, rs17659389 und rs17670505.

In der EDC-Region liegt das bereits bekannte und mit atopischer Dermatitis assoziierte *FLG*-Gen. Daher wurde geprüft, ob das gefundene Muster lediglich die Reflexion dieses Geneffektes ist. Bei der Schätzung von Haplotypen basierend auf dem gefundenen Muster aus den vier oben genannten SNPs und durch Hinzunahme der vier bekannten *FLG*-Mutationen zeigte sich, dass das gefundene Muster hauptsächlich den bekannten *FLG*-Effekt widerspiegelt. Dennoch konnte zusätzlich zu den Haplotypen mit *FLG*-Effekt ein Haplotyp bestimmt werden, der unabhängig vom *FLG* einen verstärkenden Effekt gezeigt hat (Fall-Kontroll-Studie: Odds Ratio OR=1,71, 95%-KI:1,32-2,23, p=5,6E-05; Familienstudie: OR=1,68, 95%-KI:1,18-2,38, p=2,19E-03) und daher zusätzlich zu den *FLG*-Mutationen als risikoerhöhender Haplotyp angesehen werden kann.

Der Ansatz der schrittweisen SNP-Auswahl war in der ausgewählten Kandidatengenregion erfolgreich und führte zu der Überlegung, ob die MSR auch auf unabhängige SNPs anwendbar ist. Daher wurde die MSR auf ungekoppelte SNPs in der zweiten Publikation angewendet.

6.2 Anwendung der MSR auf ungekoppelte SNPs in einer Querschnittsstudie

Publikation 2: Knüppel S, Rohde K, Meidtner K, Drogan D, Holzhütter HG, Boeing H, Fisher E. Evaluation of 41 candidate gene variants for obesity in the EPIC-Potsdam cohort by multi-locus stepwise regression. PLoS One. 2013;8(7):e68941.

In der Einzel-SNP-Analyse konnte neben dem SNP rs637797 des *TBC1D1*-Gens (Beta (SE)=-0,33 (0,13)) die bereits bekannte Assoziation mit BMI des *FTO*-Gens identifiziert werden.

Die MSR selektierte neun beste 6-SNP-Kombinationen für BMI (Effekte der zwei besten Allelkombinationen von Beta (SE)=-1,70 (0,34) bis 0,74 (0,21), mittlere Anzahl Allelkombinationen=4, mittlere Häufigkeit der Allelkombinationen=11,2%), zwei 6-SNP-Kombinationen für Taillenumfang (Effekte der beiden Allelkombinationen von Beta (SE)=-2,96 (0,76) bis 1,50 (0,75), mittlere Anzahl Allelkombinationen=7, mittlere Häufigkeit der Allelkombinationen=7,5%) und 15 3-SNP-Kombinationen auf Taillenumfang adjustiert für BMI (Effekte der zwei besten Allelkombinationen von Beta (SE)=-1,11 (0,28) bis 0,18 (0,15), mittlere Anzahl Allelkombinationen=3,5, mittlere Häufigkeit der Allelkombinationen=26,0%). Alle selektierten SNP-Kombinationen verloren ihre Signifikanz unter Berücksichtigung der simulierten Verteilung unter der Nullhypothese („kein genetischer Effekt“).

6.3 Drei Kandidatengenstudien zu kardiovaskulären Erkrankungen

Publikation 3: Drogan D, Sheldrick AJ, Schütze M, **Knüppel S**, Andersohn F, di Giuseppe R, Herrmann B, Willich SN, Garbe E, Bergmann MM, Boeing H, Weikert C. Alcohol Consumption, Genetic Variants in Alcohol Deydrogenases, and Risk of Cardiovascular Diseases: A Prospective Study and Meta-Analysis. PLoS One 2012;7(2):e32176.

Publikation 4: di Giuseppe R, Pechlivanis S, Fisher E, Arregui M, Weikert B, **Knüppel S**, Buijsse B, Fritsche A, Willich SN, Joost HG, Boeing H, Moebus S, Weikert C. Microsomal triglyceride transfer protein -164 T>C gene polymorphism and risk of cardiovascular disease: results from the EPIC-Potsdam case-cohort study. BMC Med Genet 2013;14:19.

Publikation 5: Arregui M, Buijsse B, Stefan N, Corella D, Fisher E, di Giuseppe R, Coltell O, **Knüppel S**, Aleksandrova K, Joost HG, Boeing H, Weikert C. Heterogeneity of the Stearoyl-CoA desaturase-1 (*SCD1*) gene and metabolic risk factors in the EPIC-potsdam study. PLoS One 2012;7(11):e48338.

In der EPIC-Potsdam-Studie konnte kein Effekt der Alkoholdehydrogenase-SNPs (rs1229984 und rs698) auf Schlaganfall und Myokardinfarkt nachgewiesen werden. Die durchgeführte Meta-Analyse zeigte bei der überwiegenden Mehrheit der einbezogenen Studien keinen Effekt.

Weiterhin wurde die Interaktion zwischen der Gesamtcholesterinaufnahme und dem selteneren Allel (C) des *MTTP*-SNPs rs1800804 untersucht. Es zeigte sich die erwartete Interaktion des SNPs und der Gesamtcholesterinaufnahme. Das seltenere Allel scheint mit einem erhöhten Risiko für kardiovaskuläre Erkrankungen bei Cholesterinaufnahme von weniger als 200 mg/dl ($HR_{additiv}$=1,38, 95%-KI: 1,07-1,78) einherzugehen. In der Gruppe, in der mehr als 200 mg/dl aufgenommen wurde, war ein risikosenkender Effekt zu beobachten ($HR_{additiv}$=0,77, 95%-KI:

0,58-1,03). Der Trend dieser Effekte konnte in der HNR Studie bestätigt werden ($HR_{additiv,Cholesterin<200}$=1,06, 95%-KI: 0,33-3,40 und $HR_{additiv,Cholesterin\geq200}$=0,60, 95%-KI: 0,29-1,25).

Die Analyse der sieben tagSNPs des *SCD*-Gens zeigte keine Effekte auf acht metabolische Risikofaktoren (BMI, Taillenumfang, Plasma-Triglyceride, Glykohämoglobin (HbA_{1C}), C-reaktives Protein (hs-CRP), γ-Glutamyltransferase (GGT), Alanin-Aminotransferase (ALT) und Fetuin-A) sowohl in der Einzel-SNP- als auch bei der Haplotyp-Analyse.

7. Diskussion

Diese Arbeit befasst sich mit zwei wichtigen Aspekten der genetischen Epidemiologie. Auf der einen Seite wurde eine Methode zum Selektieren krankheitsassoziierter SNP-Kombinationen entwickelt und deren Einsatzmöglichkeit an empirischen Daten getestet, und auf der anderen Seite sollten gefundene Effekte in drei Kandidatengenstudien repliziert werden.

Die schrittweise haplotyp-basierte SNP-Suche führte bei einer ausgewählten chromosomalen Genregion zu einem relevanten Ergebnis und der Suchalgorithmus kann daher für eingegrenzte Genregionen empfohlen werden. Die Anwendung der Multi-locus Stepwise Regression (MSR) auf ungekoppelte Marker hat sich in der EPIC-Potsdam-Studie als fraglich herausgestellt.

Die Idee der schrittweisen Suche nach Haplotypmustern findet sich ebenfalls im HapConstructur (Abo 2008) und SHARE (Dai 2009) für unabhängige Personen sowie im Programm HaploBuild (Laramie 2007) für Familienstudien. Die Herangehensweise der MSR ergab sich durch die besonderen Erfordernisse der jeweiligen wissenschaftlichen Fragestellung. Beispielsweise schlagen Abo (2008) vor, Haplotypen weiter zu verfolgen, wenn der p-Wert des resultierenden statistischen Tests durch Hinzunahme eines weiteren SNPs unter einem vorher definierten Schwellenwert liegt. In der genomweiten Studie wurde exemplarisch eine Hochrisiko-Region untersucht, wodurch eine besonders hohe Zahl an assoziierten Haplotypen erwartet wurde. Das Selektionskriterium wurde derart modifiziert, dass die um einen SNP erweiterten SNP-Kombinationen einen verbesserten p-Wert ergaben und so nur die 300 besten SNP-Kombinationen (kleinste p-Werte) in jedem Schritt gespeichert werden sollten, bis sich aus dem schrittweisen Ergänzen im Mittel die 10 besten mit dem dekadischen Logarithmus transformierten p-Werte um weniger als 10% verbesserten.

Der SHARE-Algorithmus basiert auf einen zweistufigen Ansatz, der im ersten Schritt die optimale Länge der gesuchten SNP-Kombinationen bestimmt und danach die SNPs unter Anwendung der Krossvalidierung auswählt. Mittels verfügbaren R-Pakets wurde dieses Verfahren auf die tagSNPs der EDC-Region angewendet, und es ergab sich eine beste zweier SNP-Kombination, die mit der MSR ebenfalls identifiziert, aber an dieser Stelle nicht weiter verfolgt wurde. Stattdessen wurde eine statistisch signifikantere Kombination bestehend aus vier SNPs ausgewählt.

Für die Analysen der EPIC-Potsdam-Studien zu Übergewicht wurde ebenfalls das Auswahlkriterium angepasst. Es wurde nicht nur auf den p-Wert, sondern auch auf die Anpassungsstatistik AIC_c zurückgegriffen, die als Maß für die relative Modellgüte verwendet wird.

Eine wichtige Möglichkeit aufgestellte Hypothesen weiter zu untersuchen, ist eine Replikationsstudie mit einer unabhängigen Stichprobe. In dieser Arbeit wurde das Ergebnis dreier Kandidatengenstudien vorgestellt, die zum Ziel hatten, angenommene genetische Einflussfaktoren zu untersuchen. Lediglich die Analyse des *MTTP*-SNPs rs1800804 zeigte den erwarteten Effekt, wodurch die Wichtigkeit des wissenschaftlichen Prozesses des Generierens von Hypothesen und deren anschließender Replikation oder deren Falsifikation deutlich wird. Meta-Analysen sind ein geeignetes Mittel, um die Ergebnisse mehrerer Studien vergleichend darzustellen und auszuwerten, wie es in der Studie zu den *ADH*-Genen geschehen ist.

Wenn eine Replikationsstudie nicht gegeben ist oder die Studie intern validiert werden soll, dann können verschiedene Resampling-Methoden, z.B. Permutationstests oder Bootstrapping-Verfahren, eingesetzt werden, denn für die klassische Bonferroni-Korrektur oder False Discovery Rate (FDR) ist unklar, für welche Zahl an Tests die schrittweise Suche korrigiert werden sollte. Es könnte die Zahl der tatsächlich durchgeführten Tests oder die maximal mögliche Zahl an Tests zu Grunde gelegt werden. Dabei ist die Zahl der tatsächlich durchgeführten Tests zu gering, denn sie beinhaltet nur gerade die Tests, die den Auswahlkriterien entsprechen. Die Wahl der maximalen Anzahl an Tests wäre zu streng, denn nicht alle Tests sind unabhängig voneinander, da SNPs ein Teil verschiedener SNP-Kombinationen sein können.
Die Kombination von SNPs oder SNP-Allelen könnte allein durch den Suchvorgang zu statistisch „signifikanten“ Ergebnissen führen, denn der Suchalgorithmus wird auf die „signifikanten“ Zwischenlösungen beschränkt. Daher bedeuten gefundene „signifikante“ SNP-

Kombinationen nicht automatisch, dass das Ergebnis klinisch relevant ist oder dass die einbezogenen SNPs in einem biochemischen funktionalen Zusammenhang stehen. Diese beiden Fragestellungen sollten je nach SNP-Auswahl an den konkreten Ergebnissen oder besser im Vorhinein durch angenommene biologische Zusammenhänge beurteilt werden. Ein weiterer Aspekt ist, dass ausgewählte SNPs nicht unbedingt in einem funktionalen Zusammenhang zu den anderen SNPs stehen müssen; sie könnten rein statistisch zur Verbesserung beitragen, beispielsweise der Unterscheidung von Fällen und Nichtfällen. Die einbezogenen SNPs müssten dabei keine funktionale Wirkung auf die untersuchte Krankheit aufweisen.

Diese Arbeit wurde auf die Anwendung und Entwicklung biostatistischer Methoden zur Evaluierung von SNPs beschränkt. Bei SNP-basierten Analysen wird ein direkter Zusammenhang zwischen dem Basenaustausch und einem Endpunkt untersucht, ohne dabei Genprodukte zu berücksichtigen. Weiterführende Erkenntnisse könnten durch die Verknüpfung der genetischen Analysen mit anderen Forschungsgebieten (Genomik, Metabolomik, Proteomik etc.) erreicht werden, die die verschiedenen Ebenen der komplexen Funktionsweisen des menschlichen Körpers zum Gegenstand haben.

Genomweite Analysen helfen dabei, Gene zu identifizieren, die über die statistische Signifikanz hinaus einen entscheidenden Beitrag zur Entwicklung der untersuchten Krankheit haben können. Die anfängliche Euphorie nach der Entschlüsselung des menschlichen Genoms und deren Veröffentlichung im Jahr 2001 hat sich gelegt. Es wird immer deutlicher, dass die genetische Epidemiologie ihre besonderen Herausforderungen in der Zukunft erst noch lösen muss. Dazu gehört u.a. die weitere vertiefende Analyse von seltenen Varianten, SNP-SNP-Interaktionen und der systematischen Berücksichtigung von nicht-genetischen Umweltfaktoren. Die Weiterentwicklung von Multi-Locus-Methoden sollte auch für genomweite Daten und für ungekoppelte SNPs fortgesetzt werden, um weiterhin vielversprechende Gene und deren komplexes Zusammenspiel zu identifizieren und deren Wirkung in verschiedenen Populationen zu bestätigen. Die Eigenschaften verschiedener Methoden sollten weiter miteinander verglichen werden, um Vor- und Nachteile verschiedener Methoden darstellen zu können. Vor diesem Hintergrund wird aktuell in der EPIC-Potsdam-Studie die MSR mit Verfahren verglichen, bei denen mehrere Genotypen simultan modelliert werden können (klassische SNP-Selektion, Genetic Risk Score und dem Least Absolute Shrinkage and Selection Operator (Lasso)-Verfahren).

7.1 Schlussfolgerung

In dieser Arbeit zeigte sich die schrittweise Mustersuche als nützlich, um besonders in Kandidatengenregionen Einblicke in die genetische Ursache komplexer Krankheiten zu erhalten. Die zur Verfügung stehenden Daten aus genomweiten Studien können so neben der Einzel-SNP-Analyse zur intensiveren Untersuchung bestimmter Genregionen eingesetzt werden. Bei ungekoppelten SNPs war die schrittweise Mustersuche nicht erfolgreich. Bei einer schrittweisen SNP-Auswahl sollte der Suchprozess besonders berücksichtigt werden, beispielsweise durch ein Resampling-Verfahren. Neben der Weiterentwicklung von statistischen Verfahren, die auf genomweiten Daten basieren, ist es ebenso wichtig, weitere Methoden zu entwickeln, die den gemeinsamen Effekt ungekoppelter SNPs über Einzel-SNP-Analysen hinaus valide schätzen können.

8. Literaturverzeichnis

Abo R, Knight S, Wong J, Cox A, Camp NJ. hapConstructor: automatic construction and testing of haplotypes in a Monte Carlo framework. Bioinformatics 2008;24(18):2105-2107.

Akey J, Jin L, Xiong M. Haplotypes vs single marker linkage disequilibrium tests: what do we gain? Eur J Hum Genet 2001;9(4):291-300.

Boeing H, Korfmann A, Bergmann MM. Recruitment procedures of EPIC Germany. European Investigation into Cancer and Nutrition. Ann Nutr Metab 1999;43(4):205-15.

Dai JY, Leblanc M, Smith NL, Psaty B, Kooperberg C. SHARE: an adaptive algorithm to select the most informative set of SNPs for candidate genetic association. Biostatistics 2009;10(4):680-93.

Esparza-Gordillo J, Weidinger S, Fölster-Holst R, Bauerfeind A, Ruschendorf F, Patone G et al: A common variant on chromosome 11q13 is associated with atopic dermatitis. Nat Genet 2009;41(5):596-601.

Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 1995;12:921-7.

Goeman JJ. L1 penalized estimation in the Cox proportional hazards model. Biom J 2010;52(1):70-84.

Hindorff LA, MacArthur J, Morales J, Junkins HA, Hall PN, Klemm AK, Manolio TA. A Catalog of Published Genome-Wide Association Studies. URL: http://www.genome.gov/gwastudies. Stand: 25.01.2013.

Laramie JM, Wilk JB, DeStefano AL, Myers RH. HaploBuild: an algorithm to construct non-contiguous associated haplotypes in family based genetic studies. Bioinformatics 2007; 23(16):2190-2192.

Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 1986;73(1):1-11.

Rohde K, Fuerst R: Haplotyping and estimation of haplotype frequencies for closely linked biallelic multilocus genetic phenotypes including nuclear family information. Hum Mutat 2001;17(4):289-95.

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 2002; 70(2):425-34.

Sugiura N: Further analysts of the data by akaike' s information criterion and the finite corrections. Communications in Statistics, Theory and Methods 1978;7:13-26.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG et al. The sequence of the human genome. Science 2001;291(5507):1304-51. [Erratum in: Science 2001;292(5523):1838.]

9. Eidesstattliche Versicherung

„Ich, Sven Knüppel, versichere an Eides statt durch meine eigenhändige Unterschrift, dass ich die vorgelegte Dissertation mit dem Thema: „Anwendung und Entwicklung biostatistischer Methoden zur Identifikation genetischer Risikofaktoren“ selbstständig und ohne nicht offengelegte Hilfe Dritter verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel genutzt habe.

Alle Stellen, die wörtlich oder dem Sinne nach auf Publikationen oder Vorträgen anderer Autoren beruhen, sind als solche in korrekter Zitierung (siehe „Uniform Requirements for Manuscripts (URM)“ des ICMJE -*www.**icmje**.org*) kenntlich gemacht. Die Abschnitte zu Methodik (insbesondere praktische Arbeiten, Laborbestimmungen, statistische Aufarbeitung) und Resultaten (insbesondere Abbildungen, Graphiken und Tabellen) entsprechen den URM (s.o) und werden von mir verantwortet.

Meine Anteile an den ausgewählten Publikationen entsprechen denen, die in der untenstehenden gemeinsamen Erklärung mit dem/der Betreuer/in, angegeben sind. Sämtliche Publikationen, die aus dieser Dissertation hervorgegangen sind und bei denen ich Autor bin, entsprechen den URM (s.o) und werden von mir verantwortet.

Die Bedeutung dieser eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unwahren eidesstattlichen Versicherung (§156,161 des Strafgesetzbuches) sind mir bekannt und bewusst.“

_____ _____

Datum Unterschrift

10. Anteilserklärung

Sven Knüppel hatte Anteil an den folgenden Publikationen:

Publikation 1:

Knüppel S, Esparza-Gordillo J, Marenholz I, Holzhütter HG, Bauerfeind A, Ruether A, Weidinger S, Lee Y-A, Rohde K. Multi-locus stepwise regression: a haplotype-based algorithm for finding genetic associations applied to atopic dermatitis. BMC Med Genet 2012;13(1):8. (Impact Factor 2013: 2.450)

Die in dieser Arbeit verwendeten statistischen Verfahren wurden von Klaus Rohde und Sven Knüppel entwickelt und im R-Paket HapEstXXR zur Verfügung gestellt. Das Manuskript wurde von Klaus Rohde und Sven Knüppel in gemeinsamer Arbeit unter Mitwirkung der Koautoren verfasst.

Anteil: 80% Beitrag zur Manuskripterstellung

Publikation 2:

Knüppel S, Rohde K, Meidtner K, Drogan D, Holzhütter HG, Boeing H, Fisher E. Evaluation of 41 candidate gene variants for obesity in the EPIC-Potsdam cohort by multi-locus stepwise regression. PLoS One. 2013;8(7):e68941. (Impact Factor 2013: 3.534)

Die in dieser Arbeit angewandten statistischen Verfahren basieren auf den Entwicklungen aus Publikation 1 und wurden von Sven Knüppel für die Fragestellung dieser Studie weiterentwickelt.

Anteil: 90% Beitrag zur Manuskripterstellung

Publikation 3:

Drogan D, Sheldrick AJ, Schütze M, **Knüppel S**, Andersohn F, di Giuseppe R, Herrmann B, Willich SN, Garbe E, Bergmann MM, Boeing H, Weikert C. Alcohol Consumption, Genetic Variants in Alcohol Deydrogenases, and Risk of Cardiovascular Diseases: A Prospective Study and Meta-Analysis. PLoS One 2012;7(2): e32176. (Impact Factor 2013: 3.534)

Die in dieser Arbeit durchgeführte statistische Auswertung und die Meta-Analyse wurden unter Beratung von Sven Knüppel vorgenommen und ausführlich diskutiert.

Anteil: 30% Beitrag zum Manuskript durch statistische Beratung und inhaltliche Überarbeitung

Publikation 4:

di Giuseppe R, Pechlivanis S, Fisher E, Arregui M, Weikert B**, Knüppel S**, Buijsse B, Fritsche A, Willich SN, Joost HG, Boeing H, Moebus S, Weikert C. Microsomal triglyceride transfer protein -164 T>C gene polymorphism and risk of cardiovascular disease: results from the EPIC-Potsdam case-cohort study. BMC Med Genet 2013;14:19. (Impact Factor 2013: 2.450)

Die in dieser Arbeit durchgeführte statistische Analyse wurde mit Unterstützung von Sven Knüppel vorgenommen. Die statistischen Ergebnisse und deren Implikationen wurden ausführlich diskutiert.

Anteil: 20% Beitrag zum Manuskript durch statistische Beratung und inhaltliche Überarbeitung

Publikation 5:

Arregui M, Buijsse B, Stefan N, Corella D, Fisher E, di Giuseppe R, Coltell O, **Knüppel S**, Aleksandrova K, Joost HG, Boeing H, Weikert C. Heterogeneity of the Stearoyl-CoA desaturase-1 (*SCD1*) gene and metabolic risk factors in the EPIC-potsdam study. PLoS One 2012;7(11):e48338. (Impact Factor 2013: 3.534)

Die in dieser Arbeit angewandten statistischen Verfahren zur Einzel-SNP- und Haplotyp-Analyse wurden von Sven Knüppel begutachtet und deren Ergebnisse diskutiert.

Anteil: 10% Beitrag zum Manuskript durch statistische Beratung und inhaltliche Überarbeitung

Softwareentwicklung:

Klaus Rohde und Sven Knüppel entwickelten das R-Paket HapEstXXR.

_____

Datum/Prof. Dr. Hermann-Georg Holzhütter (Betreuer)

_____

Datum/Sven Knüppel (Doktorand)

11. Druckexemplare der ausgewählten Publikationen

Publikation 1

Knüppel S, Esparza-Gordillo J, Marenholz I, Holzhütter HG, Bauerfeind A, Ruether A, Weidinger S, Lee Y-A, Rohde K. Multi-locus stepwise regression: a haplotype-based algorithm for finding genetic associations applied to atopic dermatitis. BMC Med Genet 2012;13(1):8.

http://dx.doi.org/10.1186/1471-2350-13-8

Die Seiten 24-31 sind im Druckexemplar enthalten oder online erhältlich.

Publikation 2

Knüppel S, Rohde K, Meidtner K, Drogan D, Holzhütter HG, Boeing H, Fisher E. Evaluation of 41 candidate gene variants for obesity in the EPIC-Potsdam cohort by multi-locus stepwise regression. PLoS One. 2013;8(7):e68941.

http://dx.doi.org/10.1371/journal.pone.0068941

Die Seiten 32-41 sind im Druckexemplar enthalten oder online erhältlich.

Publikation 3

Drogan D, Sheldrick AJ, Schütze M, **Knüppel S**, Andersohn F, di Giuseppe R, Herrmann B, Willich SN, Garbe E, Bergmann MM, Boeing H, Weikert C. Alcohol Consumption, Genetic Variants in Alcohol Deydrogenases, and Risk of Cardiovascular Diseases: A Prospective Study and Meta-Analysis. PLoS One 2012;7(2): e32176.

http://dx.doi.org/10.1371/journal.pone.0032176

Die Seiten 42-52 sind im Druckexemplar enthalten oder online erhältlich.

Publikation 4

di Giuseppe R, Pechlivanis S, Fisher E, Arregui M, Weikert B**, Knüppel S**, Buijsse B, Fritsche A, Willich SN, Joost HG, Boeing H, Moebus S, Weikert C. Microsomal triglyceride transfer protein -164 T>C gene polymorphism and risk of cardiovascular disease: results from the EPIC-Potsdam case-cohort study. BMC Med Genet 2013;14:19.

http://dx.doi.org/10.1186/1471-2350-14-19

Die Seiten 53-61 sind im Druckexemplar enthalten oder online erhältlich.

Publikation 5

Arregui M, Buijsse B, Stefan N, Corella D, Fisher E, di Giuseppe R, Coltell O, **Knüppel S**, Aleksandrova K, Joost HG, Boeing H, Weikert C. Heterogeneity of the Stearoyl-CoA desaturase-1 (*SCD1*) gene and metabolic risk factors in the EPIC-potsdam study. PLoS One 2012;7(11):e48338.

http://dx.doi.org/10.1371/journal.pone.0048338

Die Seiten 62-70 sind im Druckexemplar enthalten oder online erhältlich.

12. Lebenslauf

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

13. Publikationsliste

Zeitschriftenbeiträge

21. Aleksandrova K, Pischon T, Jenab M, Bueno-de-Mesquita H, Fedirko V, Norat T, Romaguera D, **Knüppel S**, Boutron-Ruault MC, Dossus L, Dartois L, Kaaks R, Li K, Tjønneland A, Overvad K, Quirós J, Buckland G, Sánchez M, Dorronsoro M, Chirlaque MD, Barricarte A, Khaw KT, Wareham NJ, Bradbury KE, Trichopoulou A, Lagiou P, Trichopoulos D, Palli D, Krogh V, Tumino R. Combined impact of healthy lifestyle factors on colorectal cancer: a large European cohort study. BMC Med. 2014;12:168.

20. Wientzek A, Floegel A, **Knüppel S**, Vigl M, Drogan D, Adamski J, Pischon T, Boeing H. Serum metabolites related to cardiorespiratory fitness, physical activity energy expenditure, sedentary time and vigorous activity. Int J Sport Nutr Exerc Metab. 2014;24(2):215-26.

19. Orfanos P, **Knüppel S**, Naska A, Haubrock J, Trichopoulou A, Boeing H. Evaluating the effect of measurement error when using one or two 24 h dietary recalls to assess eating out: a study in the context of the HECTOR project. Br J Nutr. 2013;110(6):1107-17.

18. Ferrari P, Freisling H, Duell EJ, Kaaks R, Lujan-Barroso L, Clavel-Chapelon F, Boutron-Ruault MC, Nailler L, Polidoro S, Mattiello A, Palli D, Tumino R, Grioni S, **Knüppel S**, Tjønneland A, Olsen A, Overvad K, Orfanos P, Katsoulis M, Trichopoulou A, Quirós JR, Ardanaz E, Huerta JM, Etxezarreta PA, Sánchez MJ, Crowe F, Khaw KT, Wareham NJ, Ocke M, Bueno-de-Mesquita B, Peeters PH, Ericson U, Wirfält E, Hallmans G, Johansson I, Engeset D, Nicolas G, Gallo V, Norat T, Riboli E, Slimani N. Challenges in estimating the validity of dietary acrylamide measurements. Eur J Nutr. 2013;52(5):1503-12.

17. Steffen A, Sørensen TI, **Knüppel S**, Travier N, Sánchez MJ, Huerta JM, Quirós JR, Ardanaz E, Dorronsoro M, Teucher B, Li K, Bueno-de-Mesquita HB, van der A D, Mattiello A, Palli D, Tumino R, Krogh V, Vineis P, Trichopoulou A, Orfanos P, Trichopoulos D, Hedblad B, Wallström P, Overvad K, Halkjær J, Tjønneland A, Fagherazzi G, Dartois L, Crowe F, Khaw KT, Wareham N, Middleton L, May AM, Peeters PH, Boeing H. Development and validation of a risk score predicting substantial weight gain over 5 years in middle-aged European men and women. PLoS One. 2013;8(7):e67429.

16. **Knüppel S**, Rohde K, Meidtner K, Drogan D, Holzhütter HG, Boeing H, Fisher E. Evaluation of 41 candidate gene variants for obesity in the EPIC-Potsdam cohort by multi-locus stepwise regression. PLoS One. 2013;8(7):e68941.

15. di Giuseppe R, Pechlivanis S, Fisher E, Arregui M, Weikert B, **Knüppel S**, Buijsse B, Fritsche A, Willich SN, Joost HG, Boeing H, Moebus S, Weikert C. Microsomal triglyceride

transfer protein -164 T>C gene polymorphism and risk of cardiovascular disease: results from the EPIC-Potsdam case-cohort study. BMC Med Genet 2013;14:19.

14. Arregui M, Fisher E, **Knüppel S**, Buijsse B, di Giuseppe R, Fritsche A, Corella D, Willich SN, Boeing H, Weikert C. Significant associations of the rs2943634 (2q36.3) genetic polymorphism with adiponectin, high density lipoprotein cholesterol and ischemic stroke. Gene. 2012;494(2):190-5.
13. **Knüppel S**, Esparza-Gordillo J, Marenholz I, Holzhütter HG, Bauerfeind A, Ruether A, Weidinger S, Lee Y-A, Rohde K. Multi-locus stepwise regression: a haplotype-based algorithm for finding genetic associations applied to atopic dermatitis. BMC Med Genet 2012;13(1):8.
12. Arregui M, Buijsse B, Stefan N, Corella D, Fisher E, di Giuseppe R, Coltell O, **Knüppel S**, Aleksandrova K, Joost HG, Boeing H, Weikert C. Heterogeneity of the Stearoyl-CoA desaturase-1 (SCD1) gene and metabolic risk factors in the EPIC-potsdam study. PLoS One 2012;7(11):e48338.
11. Drogan D, Sheldrick AJ, Schütze M, **Knüppel S**, Andersohn F, di Giuseppe R, Herrmann B, Willich SN, Garbe E, Bergmann MM, Boeing H, Weikert C. Alcohol Consumption, Genetic Variants in Alcohol Deydrogenases, and Risk of Cardiovascular Diseases: A Prospective Study and Meta-Analysis. PLoS One 2012;7(2): e32176.
10. Stang A, Schipf S, **Knüppel S**. Directed Acyclic Graphs - Reply to Methodological Concerns. Gesundheitswesen 2011;73(12):925-6.
9. **Knüppel S**. Basics of Using the DAG Programs. Gesundheitswesen 2011;73(12):893-6.
8. Schipf S, **Knüppel S**, Hardt J, Stang A. Directed Acyclic Graphs (DAGs) - The Application of Causal Diagrams in Epidemiology. Gesundheitswesen 2011;73(12):888-92.
7. Textor J, Hardt J, **Knüppel S**. DAGitty: a graphical tool for analyzing causal diagrams. Epidemiology 2011;22(5):745.
6. Harttig U, Haubrock J, **Knüppel S**, Boeing H. EFCOVAL Consortium. The MSM program: web-based statistics package for estimating usual dietary intake using the Multiple Source Method. Eur J Clin Nutr 2011;65 Suppl 1:S87-91.
5. Haubrock J, Nöthlings U, Volatier JL, Dekkers A, Ocké M, Harttig U, Illner AK, **Knüppel S**, Andersen LF, Boeing H. European Food Consumption Validation Consortium. Estimating usual food intake distributions by using the multiple source method in the EPIC-Potsdam Calibration Study. J Nutr 2011;141(5):914-20.

4. Montonen J, Landberg R, Kamal-Eldin A, Aman P, **Knueppel S**, Boeing H, Pischon. Reliability of fasting plasma alkylresorcinol concentrations measured 4 months. Eur J Clin Nutr 2010;64(7):698-703.
3. Hoeft B, Linseisen J, Beckmann L, Müller-Decker K, Canzian F, Hüsing A, Kaaks R, Vogel U, Jakobsen MU, Overvad K, Hansen RD, **Knüppel S**, Boeing H, Trichopoulou A, Koumantaki Y, Trichopoulos D, Berrino F, Palli D, Panico S, Tumino R, Bueno-de-Mesquita HB, van Duijnhoven FJ, van Gils CH, Peeters PH, Dumeaux V, Lund E, Huerta Castaño JM, Muñoz X, Rodriguez L, Barricarte A, Manjer J, Jirström K, Van Guelpen B, Hallmans G, Spencer EA, Crowe FL, Khaw KT, Wareham N, Morois S, Boutron-Ruault MC, Clavel-Chapelon F, Chajes V, Jenab M, Boffetta P, Vineis P, Mouw T, Norat T, Riboli E, Nieters A. Polymorphisms in fatty-acid-metabolism-related genes are associated with colorectal cancer risk. Carcinogenesis 2010;31(3):466-72.
2. **Knüppel S**, Stang A. DAG program: identifying minimal sufficient adjustment sets. Epidemiology 2010;21(1):159. [Erratum in: Epidemiology 2010;21(3):432.]
1. Brisch R, Bernstein HG, Krell D, Dobrowolny H, Bielau H, Steiner J, Gos T, Funke S, Stauch R, **Knüppel S**, Bogerts B. Dopamine-glutamate abnormalities in the frontal cortex associated with the catechol-O-methyltransferase (COMT) in schizophrenia. Brain Res 2009;1269:166-75.

Technische Berichte

2. Goedhart PW, van der Voet H, **Knüppel S**, Dekkers ALM, Dodd KW, Boeing H, van Klaveren J. A comparison by simulation of different methods to estimate the usual intake distribution for episodically consumed foods. EFSA External Scientific Report (Frage Nr.: EFSA-Q-2012-00626). 2012.
1. van der Voet H, van Klaveren J unter Mitwirkung von Arcella D, Bakker M, Boeing H, Boon PE, Crépet A, Dekkers ALM, de Boer WJ, Dodd KW, Ferrari P, Goedhart PW, Hart A, van der Heijden G, Kennedy M, Kipnis V, **Knüppel S**, Merten C, Ocké M, Slob W. Statistical modelling of usual intake. EFSA External Scientific Report (Frage Nr.: EFSA-Q-2009-00841). 2010.

14. Danksagung

Ohne die Unterstützung vieler Menschen hätte diese Arbeit nicht geschrieben werden können. Ich danke im Besonderen all denen, die an mich glaubten und mich förderten.

Zunächst danke ich Prof. Hermann-Georg Holzhütter und Klaus Rohde für die gemeinsame Entwicklung dieser Promotion am Max-Delbrück-Centrum für Molekulare Medizin (MDC) Berlin-Buch und Prof. Heiner Boeing für die Möglichkeit, die Promotion am Deutschen Institut für Ernährungsforschung Potsdam-Rehbrücke (DIfE) fortführen zu können. Weiterhin möchte ich Prof. Young-Ae Lee danken, die mir den Zugang zu den genomweiten Daten zur atopischen Dermatitis ermöglicht und in vielen Diskussionen mein Verständnis für die notwendigen Implikationen der Ergebnisse gestärkt hat.

Von Beginn bis zum Ende der Promotion hat mich Klaus Rohde methodisch und menschlich begleitet. Mit großer Hochachtung möchte ich ihm meinen Dank aussprechen.

In meiner Zeit als Doktorand wurde ich stets von freundlichen und hilfsbereiten Kolleginnen und Kollegen begleitet. Stellvertretend für die Kolleginnen und Kollegen am MDC möchte ich mich besonders bei Anja Bauerfeind bedanken, die mir stets mit Rat und Tat zur Seite stand. Am DIfE teilte ich die Zeit mit den netten Doktorandinnen und Doktoranden. Weiterhin bin ich dankbar für die freundliche Arbeitsatmosphäre und die vielen fruchtbringenden Diskussionen am DIfE. Stellvertretend bedanke ich mich bei Brian Buijsse, Conny Weikert, Dagmar Drogan, Ellen Kohlsdorf, Karina Meidtner, Krassimira Aleksandrova, Maria Arregui, Romina di Giuseppe, Wolfgang Bernigau und Wolfgang Fleischhauer. Besonders möchte ich noch Eva Fisher für die tatkräftige Unterstützung danken.

Des Weiteren möchte ich den immer umsichtigen und hilfsbereiten Sekretärinnen Edelgard Wolf (MDC) und Gabriele Weeske (DIfE) danken.

Zuletzt richte ich meinen Dank an meine Familie, Freunde und meinem Patenonkel, die mir Stärkung gaben und Mut zugesprochen haben. Meiner Frau Kerstin danke ich in ganz besonderem Maße für Ihre Unterstützung in allen Lebenslagen und für die aufbauenden sowie motivierenden Worte und Taten.

Es ist ein großes Geschenk, Euch allen begegnet zu sein!