

Über Korrelationsstrukturen bei SNP-Assoziationsanalysen

Dissertation
zur Erlangung des akademischen Grades
Dr. rer. nat.

an der Medizinischen Fakultät
der Universität Leipzig

eingereicht von:

Diplom-Informatiker Arnd Groß

Geburtsdatum / Geburtsort:

09.12.1980 in Schlema

angefertigt am:

Institut für Medizinische Informatik, Statistik und Epidemiologie der Universität Leipzig

Betreuer:

Prof. Dr. rer. nat. Markus Scholz

Beschluss über die Verleihung des Doktorgrades vom: 12.02.2019

11010	01011
1101100	1001110
0100111	0101001
1000110	1101000
0110001	0111010
0101101	1110110
0111011	0110011
1011001	1100010
1110010	0111010
1011101	1010110
0010111	0101011
1001001	1101010
	0110011101
1000000	1000100
0110101	0011010
0101100	0000100
1100101	0011100
1000001	0101011
0100100	0001011
0011101	1001110
1111011	0010111
1010010	1101000
0110001	0111010
0101101	1110110
0111011	0110011
1010000	1100111
0110100	0011000
1101101	0011110
1100111	0101001
	10110

Inhaltsverzeichnis

1 Einführung	1
1.1 Aufbau der Arbeit	1
1.2 Grundbegriffe	1
1.3 Populationsgenetik	3
1.4 Klassische Assoziationsanalyse	5
1.5 Bayesianische Assoziationsanalyse	9
2 Auswirkung von Korrelationsstrukturen bei Populationsvergleichen	13
2.1 Beschreibung der Studien und genetischen Daten	13
2.2 Analyse der Verwandtschaft	13
2.3 Analyse des Kopplungsungleichgewichts	14
2.4 Power von SNP-Assoziationsanalysen	14
2.5 Weitere Analysen	16
2.6 Zusammenfassung	17
2.7 Publikation	17
3 Einfluß von Verwandtschaft auf Assoziationsanalysen	32
3.1 Beschreibung der Studien und genetischen Daten	32
3.2 Varianzinflation	33
3.3 Hypothesentests	34
3.4 Weitere Analysen	36
3.5 Zusammenfassung	37
3.6 Publikation	37
4 Bayesianischer Ansatz zur Berücksichtigung korrelierter Phänotypen	49
4.1 Beschreibung der Studie und genetischen Daten	49
4.2 Klassische Assoziationsanalyse	50
4.3 Bayesianische Assoziationsanalyse	50
4.4 Zusammenfassung	52
4.5 Publikation	52
5 Ausblick	68
Zusammenfassung der Arbeit	69
Verzeichnis der Abkürzungen und Symbole	72
Literaturverzeichnis	75

Darstellung des eigenen Beitrags	83
Selbstständigkeitserklärung	99
Publikationen	100
Danksagung	103

1 Einführung

1.1 Aufbau der Arbeit

Bei der vorliegenden Arbeit handelt es sich um eine kumulative Dissertation, die drei Veröffentlichungen [1–3] umfaßt. In diesem Abschnitt wird die Gliederung der Arbeit beschrieben. Danach werden wichtige Begriffe im Abschnitt 1.2 erläutert, um so die Basis zum Verständnis der darauf folgenden Abschnitte zu legen. Die Übersicht im **Verzeichnis der Abkürzungen und Symbole** soll zusätzlich dem Verständnis dienen. Neben der Erläuterung bestimmter Begriffe wird in den Abschnitten 1.3, 1.4 und 1.5 in verschiedene statistische Problemstellungen eingeführt. Am Ende von jedem der drei Abschnitte wird eine Zielstellung formuliert, welche offenen Fragen durch diese Arbeit beantwortet werden sollen. Die Ergebnisse der Veröffentlichungen werden in den Kapiteln 2, 3 und 4 präsentiert. Mit einem Ausblick im Kapitel 5 wird die Arbeit abgeschlossen.

Zur Präsentation der Ergebnisse aus den Publikationen sind noch einige Hinweise nötig: Verweise auf Tabellen und Grafiken sind *kursiv* hervorgehoben und beziehen sich auf die Publikation oder deren Anhang, wobei sich die Publikation am Ende eines jeden Kapitels befindet. Die Anhänge der Publikationen sind aufgrund des Umfangs nicht in diese Arbeit eingebunden. Sowohl die Veröffentlichung als auch die Anhänge sind „open access“ publiziert und damit frei verfügbar. Die Bezeichnungen für Tabellen, Grafiken und Anhänge der jeweiligen Publikation werden beibehalten und sind in Englisch, um das Auffinden der referenzierten Objekte zu erleichtern.

1.2 Grundbegriffe

In diesem Abschnitt werden einige Grundbegriffe eingeführt.

Genetischer Marker

Eine wichtige Voraussetzung genetischer Analysen sind genetische Marker. Unter einem genetischen Marker versteht man eine Stelle in der DNA (deoxyribonucleic acid), die sich wenigstens um eine Base zwischen zwei Individuen unterscheidet. Weiterhin ist der Marker messbar und seine Position im Genom bekannt [4]. Einzelnucléotid-Polymorphismen (SNPs oder single nucleotide polymorphisms) zählen zu den genetischen Markern und sind mit 90% die am häufigsten auftretende Form genetischer Variation im menschlichen Genom. Sie sind dadurch charakterisiert, daß eine Base durch eine andere ausgetauscht wird. In der Theorie sind damit vier verschiedene Allele möglich, repräsentiert durch die Basen Adenin (A), Guanin (G), Cytosin (C) oder Thymin (T). Beim Menschen treten jedoch in der Regel nur biallelische SNPs auf, wobei zwei Drittel A/G oder C/T SNPs sind. Während die einfache Definition eines SNP keine Angabe über dessen Allelfrequenz macht, fordert die strenge Definition für das seltene Allel eine Allelfrequenz

(MAF oder minor allele frequency) von wenigstens 1% in der betrachteten Population. Die Verwendung von SNPs in der Genetik hat zahlreiche Vorteile wie ihre Häufigkeit und damit gute Abdeckung des Genoms, ihre Stabilität durch geringe Mutationshäufigkeit und ihre Meßbarkeit im Hochdurchsatz mittels Microarrays [4].

Genetische Assoziation

Um den Einfluß eines genetischen Markers auf einen Phänotyp aufzudecken, werden genetische Assoziationen analysiert. Genetische Assoziation bedeutet, daß ein Merkmal oder ein Phänotyp häufiger oder mit höheren Werten mit einem bestimmten Allel eines genetischen Markers auftritt, als man zufällig erwarten würde. Angenommen für n Probanden einer Studie liegen p Phänotypen in einer Matrix \mathbf{y} vor, die Genotypen von q SNPs in einer Matrix \mathbf{s} und r weitere Kovariablen in einer Matrix \mathbf{c} . Die Messungen eines Probanden befinden sich dabei in einer festgelegten Zeile der Matrizen \mathbf{y} , \mathbf{s} oder \mathbf{c} . Die SNP-Genotypen werden durch die Anzahl 0, 1 oder 2 eines bestimmten Referenzallels, das heißt eine der vier Basen, kodiert. Wobei hier mit dem Begriff „Genotyp“ nicht die Gesamtheit aller Gene eines Individuums gemeint ist, sondern die Ausprägungen eines einzelnen SNP. In dieser Arbeit handelt es sich bei Phänotypen ausschließlich um stetige Merkmale und die Grundlage für die Analyse der genetischen Assoziation bildet das lineare Modell. Ziel einer SNP-Assoziationsanalyse ist es, statistisch auf einen Zusammenhang zwischen der $n \times p$ Phänotyp-Matrix \mathbf{y} und der $n \times q$ SNP-Matrix \mathbf{s} unter Einbeziehung der $n \times r$ Kovariablen-Matrix \mathbf{c} zu schließen.

Genetisches Modell

Die Art und Weise wie auf eine genetische Assoziation getestet wird, hängt vom genetischen Modell ab. Das genetische Modell oder das Vererbungsmuster eines Phänotyps beschreibt, wie der Phänotyp durch die Allele eines Genotyps bestimmt wird. Wichtige genetische Modelle sind das dominante, das rezessive und das additive Modell. Nimmt man vollständige Penetranz an, das heißt der gleiche Genotyp führt immer zum gleichen Phänotyp, dann unterscheiden sich die Modelle dadurch, welchen Phänotyp ein heterozygoter Genotyp zur Folge hat. Ist das erste Allel dominant, führt ein heterozygoter Genotyp auf den Phänotyp des ersten Allels und zwar genau auf den gleichen Phänotyp, der auftritt, wenn das erste Allel doppelt vorhanden ist. In diesem Fall ist zugleich das zweite Allel rezessiv und der zum zweiten Allel gehörende Phänotyp tritt nur auf, wenn das zweite Allel doppelt vorhanden ist. Im Gegensatz dazu unterscheidet sich im additiven Modell der Phänotyp eines heterozygoten Genotyps von den beiden Phänotypen, die auftreten, wenn jeweils beide Allele doppelt vorhanden sind. Im Falle eines stetigen Merkmals entspricht der Phänotyp eines heterozygoten Genotyps im additiven Modell gerade dem Mittelwert der beiden anderen möglichen Phänotypen.

Korrelation

Im Allgemeinen wird durch Korrelation ein Zusammenhang zwischen zwei Merkmalen ausgedrückt. In der Statistik wird Korrelation formal als ein skalenunabhängiges oder skaleninvariantes Maß für die lineare Abhängigkeit zweier Zufallsvariablen A und B definiert.

Für den Korrelationskoeffizienten gilt nach [5]:

$$\text{Cor}(A,B) = \frac{\text{Cov}(A,B)}{\sqrt{\text{V}(A)\text{V}(B)}} \quad (1.1)$$

mit $-1 \leq \text{Cor}(A,B) \leq 1$. $\text{Cov}(A,B)$ ist die Kovarianz und $\text{V}(A)$ und $\text{V}(B)$ sind die Varianzen beider Zufallsvariablen. Sind zwei Zufallsvariablen A und B unabhängig, so folgt $\text{Cov}(A,B) = \text{Cor}(A,B) = 0$ und man bezeichnet beide Zufallsvariablen als unkorreliert. Es muß beachtet werden, daß die Umkehrung „sind A und B unkorreliert, so folgt A und B sind unabhängig“ typischerweise nicht gilt [5]. Nimmt man an, daß SNPs oder Phänotypen, die im Rahmen einer Studie gemessen wurden, Realisierungen solcher Zufallsvariablen sind, dann können paarweise Korrelationen auf Basis von Gl. (1.1) quantifiziert werden. In den folgenden Abschnitten wird anhand einiger Beispiele gezeigt, wo Korrelationen bei SNP-Assoziationsanalysen auftreten können, welche Auswirkungen Korrelationen haben können und wie man Korrelationsstrukturen bei der Modellierung berücksichtigen kann.

1.3 Populationsgenetik

Ein Zweig der Populationsgenetik beschäftigt sich mit der Analyse von Populationen auf Basis genetischer Marker hinsichtlich bestimmter Eigenschaften der Population selbst und im Vergleich zu anderen Populationen. Anwenden lassen sich populationsgenetische Analysen beispielsweise zur Charakterisierung von Isolatpopulationen. Isolatpopulationen sind genetisch homogener und durch eine homogener Umwelt gegenüber nicht isolierten Populationen gekennzeichnet. Aber es ist nicht abschließend geklärt, ob sich Isolatpopulationen zur Identifikation von genetischen Ursachen für Krankheiten oder quantitative Phänotypen besser eignen [6, 7]. Daher ist es sinnvoll zu untersuchen, inwieweit sich Populationen wie die Sorben [8], die einen gewissen Isolatcharakter aufweisen, von sonstigen Populationen in Deutschland, beispielsweise einer populationsbasierten Studie wie KORA [9, 10], genetisch unterscheiden. Für den Populationsvergleich sind verschiedene Analysen möglich, wie die Analyse der Verwandtschaftsstruktur, die Hauptkomponentenanalyse [11, 12], die Analyse der Anzahl seltener SNPs, die Analyse der F-Statistiken [13], die Analyse der Runs of homozygosity [14] oder die Analyse des Kopplungsungleichgewichts paarweiser SNPs. Wie noch gezeigt wird, sind die Verwandtschaftsstruktur und das Kopplungsungleichgewicht von besonderer Bedeutung für SNP-Assoziationsanalysen und bilden deshalb den Schwerpunkt der populationsgenetischen Analysen dieser Arbeit.

Verwandtschaft

Das beste Beispiel für eine Korrelation in der Genetik ist eine Korrelation des gleichen Phänotyps zwischen zwei Probanden, beispielsweise die Körpergröße [15], welche durch die Verwandtschaft beider Individuen verursacht wird. Eine Verwandtschaftsstruktur führt jedoch zu einer Verletzung der Annahme unabhängiger Beobachtungen, die beim linearen Modell im Rahmen einer SNP-Assoziationsanalyse vorausgesetzt werden. Diese Verletzung der Unabhängigkeit hat eine Auswirkung auf die Eigenschaften eines statistischen Tests [16, 17] und es ist daher wichtig, die Verwandtschaftsstruktur zu untersuchen. Zunächst wird die Verwandtschaft zweier Probanden

mit der Notation aus [18] definiert: Sind ϕ und δ die Wahrscheinlichkeiten dafür, daß genau ein Allel beziehungsweise beide Allele vom selben Vorfahren (IBD oder identical by descent) geerbt wurden, dann gilt für die paarweise Verwandtschaft

$$G = \phi/2 + \delta.$$

Es gilt $0 \leq G \leq 1$ und verschiedene Verwandtschaftsbeziehungen können dabei auf das gleiche G führen, zum Beispiel ist für eine Elter-Kind-Beziehung ($\phi = 1, \delta = 0$) oder eine Geschwister-Beziehung ($\phi = 1/2, \delta = 1/4$) die erwartete paarweise Verwandtschaft $G = 1/2$.

Verwandtschaftsbeziehungen (Stammbauminformationen) sind außer für Familienstudien kaum verfügbar oder mit Fehlern behaftet. Vor allem für Studien mit großer Fallzahl wäre die Erstellung eines Stammbaums sehr aufwendig. Sind die Verwandtschaftsbeziehungen einer Studie nicht bekannt, können sie aber mithilfe einiger 10.000 SNPs und beispielsweise der in [18] beschriebenen Methoden geschätzt werden. Selbst wenn Stammbauminformationen verfügbar sind, spiegeln die geschätzten Verwandtschaften die genetischen Beziehungen besser wider [16, 17] als die erwartete Verwandtschaft, die aus Stammbäumen berechnet werden kann. Unabhängig davon ob paarweise Verwandtschaften aus Stammbäumen abgeleitet oder geschätzt werden, können diese für die weitere Analyse in einer symmetrischen $n \times n$ Verwandtschaftsmatrix \mathbf{G} zusammengefaßt werden.

Kopplungsungleichgewicht

Für Assoziationsanalysen mehrerer benachbarter SNPs ist es sinnvoll, die Beziehung eines SNP zu seinen Nachbarn zu untersuchen. Von Kopplungsungleichgewicht (LD oder linkage disequilibrium) spricht man, wenn die Allele zweier (meist benachbarter) SNPs S_1 und S_2 nicht unabhängig voneinander auftreten. Sind in einer 2×2 Matrix (p_{ij}) , $i,j \in \{0,1\}$, die Wahrscheinlichkeiten für das Auftreten der vier möglichen Allelkombinationen abgetragen, so kann mit bestimmten Maßen diese Unabhängigkeit beschrieben werden. Bekannte Maße sind D' [19] oder der Korrelationskoeffizient [20]

$$r = \frac{p_{00}p_{11} - p_{01}p_{10}}{\sqrt{p_{0.}p_{.0}p_{1.}p_{.1}}}. \quad (1.2)$$

Eine Übersicht weiterer Maße wird in [21] präsentiert. Mit der Notation von [21] beschreibt p_{ij} die Wahrscheinlichkeit einer der vier Allelkombinationen von SNP i und SNP j , wobei die Allele der SNPs jeweils mit „0“ und „1“ kodiert werden. Die Randsummen $p_{i.}$ und $p_{.j}$ entsprechen den Allelfrequenzen von SNP i und j .

Ist die Matrix (p_{ij}) für die Allelkombinationen zweier SNPs bekannt, dann kann unter der Annahme des Hardy-Weinberg-Gleichgewichts (HWE oder Hardy-Weinberg-Equilibrium) der Korrelationskoeffizient in Gl. (1.1) für die Genotypen von S_1 und S_2 aus dem Korrelationskoeffizienten in Gl. (1.2) für die Allelkombinationen abgeleitet werden. Das HWE beschreibt, daß die Genotyp- und Allelfrequenzen von sehr großen Populationen mit zufälliger Paarung der Individuen über Generationen hinweg stabil bleiben und daß es eine festgelegte Beziehung zwischen den Genotyp- und Allelfrequenzen gibt [4]. Die SNP-Korrelationsstruktur einer Studienpopulation läßt sich beschreiben, indem von einer Vielzahl von SNPs die paarweisen r entweder auf Basis von Gl. (1.1) oder Gl. (1.2) bestimmt werden. Für die Bestimmung von r aus Gl. (1.2) ist es

notwendig, daß die Allele beider SNPs phasiert vorliegen und die Wahrscheinlichkeiten p_{ij} geeignet geschätzt werden [21]. Näherungsweise läßt sich die SNP-Korrelation aber auch mit Gl. (1.1) aus den Genotypen schätzen, indem die Kovarianz und beide Varianzen durch die empirische Kovarianz beziehungsweise den beiden empirischen Varianzen ersetzt werden. Möchte man das paarweise LD verschiedener Studienpopulationen vergleichen, eignen sich aber Maße wie η_1 [22], das von der Allelfrequenz unabhängig ist, jedoch besser. Das Maß η_1 ist eine monotone Funktion des odds ratio $\omega = p_{00}p_{11}/p_{01}p_{10}$ [23] der Matrix (p_{ij}) , wobei gilt:

$$\eta_1 = \begin{cases} 2^{\frac{\omega^2 - \omega - \omega \ln \omega}{(\omega - 1)^2}} - 1 & \text{wenn } \omega \neq 1, \\ 0 & \text{wenn } \omega = 1. \end{cases} \quad (1.3)$$

Während eine direkte genetische Assoziation einen kausalen Zusammenhang zwischen einem genetischen Marker und einem Phänotyp beschreibt, ist eine indirekte genetische Assoziation dadurch charakterisiert, daß der untersuchte genetische Marker nur im LD zu einer weiteren genetischen Variante ist, die kausal mit dem Phänotyp zusammenhängt. Ist der kausale Marker nicht gemessen, weil er zum Beispiel nicht für Microarrays vorgesehen ist, dann ist für eine indirekte genetische Assoziation die Höhe des LD zu den gemessenen Markern in der Umgebung des kausalen Markers von großer Bedeutung. Da sich die SNP-Korrelationsstruktur von Isolatpopulationen gegenüber nicht isolierten Populationen durch ein im Mittel höheres LD unterscheidet, wird dadurch ein Vorteil bei der Identifikation von genetischen Ursachen bestimmter Merkmale erwartet [24, 25].

Zielstellung

Neben vielen weiteren Methoden der Populationsgenetik zum Vergleichen von Populationen sind die Bestimmung der Verwandtschaftsstruktur und des LD für eine SNP-Assoziationsanalyse besonders wichtig. Deshalb sollen die Verwandtschaftsstruktur und die SNP-Korrelationsstruktur für die Sorben und für KORA geschätzt und miteinander verglichen werden. Weiterhin soll geprüft werden, ob sich Unterschiede in der SNP-Korrelationsstruktur der Sorben und KORA auf die Power einer SNP-Assoziationsanalyse bei einer indirekten genetischen Assoziation auswirken und ob Unterschiede in der Verwandtschaftsstruktur einen Einfluß auf die Schätzung der SNP-Korrelationsstruktur und die Power haben.

1.4 Klassische Assoziationsanalyse

In diesem Abschnitt wird untersucht, welchen Einfluß die Verwandtschaftsstruktur einer Studie auf eine SNP-Assoziationsanalyse hat.

Gemischtes Modell

Zunächst wird zur Vereinfachung angenommen, daß für jeden der n Probanden einer Studie nur ein Phänotyp ($p = 1$), ein SNP ($q = 1$), keine weiteren Kovariablen ($r = 0$) und das additive genetische Modell vorliegen. Dann ist eine geeignete Beschreibung [17, 26–28] für eine Phänotyp-

Genotyp-Beziehung das gemischte Modell:

$$y_i = b_1 + b_2 s_i + g_i + e_i \quad (1.4)$$

für einen Phänotyp $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$, Achsenabschnitt b_1 , SNP-Effekt b_2 , SNP-Genotypen $\mathbf{s} = (s_1, s_2, \dots, s_n)^T$, zufällige polygene Effekte $\mathbf{g} = (g_1, g_2, \dots, g_n)^T$ und Residuen $\mathbf{e} = (e_1, e_2, \dots, e_n)^T$. Für die zufälligen Effekte wird angenommen, daß $\mathbf{g} \sim N_n(\mathbf{0}, \sigma_g^2 \mathbf{G})$ multivariat normalverteilt ist mit der Varianzkomponente σ_g^2 und der Kovarianzmatrix \mathbf{G} . Die paarweise Korrelation des Phänotyps zwischen zwei Individuen i und j , $1 \leq i, j \leq n$, wird durch die zufälligen Effekte g_i und g_j verursacht, welche wiederum von der Varianzkomponente σ_g^2 und der Verwandtschaft $G_{i,j}$ beider Personen abhängen. Diese Korrelation lässt sich dabei mit dem Korrelationskoeffizienten aus Gl. (1.1) beschreiben. Die Residuen hingegen werden als unabhängig und als multivariat normalverteilt $\mathbf{e} \sim N_n(\mathbf{0}, \sigma_e^2 \mathbf{I})$ mit der Varianzkomponente σ_e^2 und der Identitätsmatrix \mathbf{I} angenommen. Die Phänotyp-Korrelationsstruktur hängt neben der Verwandtschaftsstruktur, die als Kovarianz-Matrix \mathbf{G} in das Modell eingeht, noch von der Erblichkeit oder Heritabilität

$$R_h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$$

ab. Heritabilität ist die erklärte Varianz von Polygenie an der Varianz von \mathbf{y} , wobei sich die Varianz von \mathbf{y} aus der Varianz von Polygenie und residualer Varianz, zum Beispiel Umwelteinflüsse, zusammensetzt. Die erklärte Varianz des SNP an der Varianz von \mathbf{y} ist

$$R_s^2 = \frac{b_2^2 \hat{V}(\mathbf{s})}{\sigma_g^2 + \sigma_e^2}$$

mit der empirischen Varianz $\hat{V}(\mathbf{s}) = \sum_{i=1}^n (s_i - \bar{s})^2 / (n-1)$ des SNP. Zur Beschreibung der Stärke des genetischen Effekts ist es sinnvoll, die erklärte Varianz R_s^2 anstelle von b_2 zu verwenden, da ansonsten die Alelfrequenz zusätzlich zum SNP-Effekt b_2 berücksichtigt werden muß.

Vereinfachtes Modell

Die Schätzung der Parameter des gemischten Modells in Gl. (1.4) ist mathematisch anspruchsvoll und benötigt mehr Rechenzeit als beispielsweise die einfache lineare Regression, insbesondere wenn Assoziationsanalysen genomweit und mit großen Fallzahlen durchgeführt werden. Aus diesem Grund wird häufig die Korrelation des Phänotyps zwischen den Probanden ignoriert und stattdessen ein einfaches lineares Modell mit unabhängigen Residuen angepaßt:

$$y_i = \beta_1 + \beta_2 s_i + \epsilon_i. \quad (1.5)$$

Hierbei bezeichnet β_1 den Achsenabschnitt und β_2 den SNP-Effekt. Die Residuen $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)^T$ werden als unabhängig und multivariat normalverteilt $\epsilon \sim N_n(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I})$ mit der Varianzkomponente σ_ϵ^2 und der Identitätsmatrix \mathbf{I} angenommen. Die Verletzung der Annahme unabhängiger Beobachtungen durch die Phänotyp-Korrelationsstruktur führt aber zu einer Varianzvergrößerung des Effektschätzers $\hat{\beta}_2$ [29], die sich auf den statistischen Test auswirkt. In

empirischen Studien, wie [30, 31], wurde beobachtet, daß der Erwartungswert des Effektschätzers nicht beeinflußt wird. Ein Ziel dieser Arbeit ist es, den Einfluß der Phänotyp-Korrelationsstruktur auf den Erwartungswert und die Varianz des Effektschätzers analytisch zu beschreiben.

Für eine SNP-Assoziationsanalyse mit r Kovariablen und der Möglichkeit verschiedene genetische Modelle abzubilden, wird das Modell aus Gl. (1.5) erweitert zu

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 c_{i,1} + \cdots + \beta_{2+r} c_{i,r} + \epsilon \quad (1.6)$$

mit Achsenabschnitt β_1 , genetischem Effekt β_2 , Hilfsvariable \mathbf{x} , Effekten der Kovariablen $\beta_3, \dots, \beta_{2+r}$, Kovariablen \mathbf{c} und unabhängigen Residuen $\epsilon \sim N_n(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I})$. Durch die Hilfsvariable \mathbf{x} können verschiedene genetische Modelle abgebildet werden. Beschreibt s_i die Anzahl des selteneren Allels, dann wird durch $x_i = s_i$ das additive Modell dargestellt. Durch $x_i = I(s_i = 2)$ wird das dominante und durch $x_i = I(s_i = 0)$ das rezessive genetische Modell bezüglich des häufigeren Allels modelliert. Für die Funktion I gilt hierbei: $I(x) = 1$, wenn x wahr ist, andernfalls ist $I(x) = 0$.

Hypothesentests

Angenommen man beobachtet Phänotypen \mathbf{y} und Genotypen \mathbf{s} , die durch das Modell aus Gl. (1.4) beschrieben werden und möchte auf eine Phänotyp-Genotyp-Assoziation testen. Aber anstelle des (korrekten) gemischten Modells aus Gl. (1.4) anzupassen, vernachlässigt man stattdessen die Korrelation des Phänotyps zwischen den Probanden und paßt dafür das (vereinfachte) lineare Modell aus Gl. (1.5) an. Dazu bestimmt man die Teststatistik [5]

$$T = \frac{\hat{\beta}_2}{S_{\beta}} \quad (1.7)$$

mit Effektschätzer $\hat{\beta}_2$ und seiner empirischen Varianz S_{β}^2 [32]. Die Evidenz für einen Zusammenhang zwischen einem Phänotyp und dem genetischen Effekt kann durch einen p-Wert formuliert werden. Der p-Wert entspricht hierbei der Wahrscheinlichkeit für die Teststatistik T unter der Nullhypothese $\beta_2 = 0$ einen extremeren Wert zu beobachten als die für $\hat{\beta}_2$ bestimmte Teststatistik \hat{T} einer konkreten Analyse. Für einen zweiseitigen Test bedeutet das

$$p = 2P(T > |\hat{T}|). \quad (1.8)$$

Die Nullhypothese für einen Test von \hat{T} wird zum Signifikanzniveau α abgelehnt genau dann, wenn $p < \alpha$. Testet man zweiseitig zum Signifikanzniveau α , wobei $z_{\alpha/2}$ dem $\alpha/2$ -Quantil der Standardnormalverteilung entspricht, dann ist unter Vorliegen der Nullhypothese $\beta_2 = 0$ der Fehler erster Art des Tests

$$\text{err} = 2P(T < z_{\alpha/2}). \quad (1.9)$$

Der Fehler erster Art entspricht der Wahrscheinlichkeit, die Nullhypothese fälschlicherweise abzulehnen.

Unter Vorliegen der Alternativhypothese $\beta_2 \neq 0$ ist die Power

$$\text{pwr} = P(T < z_{\alpha/2}) + P(T > -z_{\alpha/2}). \quad (1.10)$$

Die Power entspricht der Wahrscheinlichkeit, die Nullhypothese korrekterweise abzulehnen.

Aus empirischen Studien ist bekannt, daß der Fehler erster Art (Gl. (1.9)) bei der Anpassung des vereinfachten Modells (Gl. (1.5)) inflationiert ist [33] und mit größerer Heritabilität und stärkerer Verwandtschaft steigt [30, 31]. Eine stärkere Verwandtschaft bedeutet hier mehr und stärker miteinander verwandte Studienteilnehmer. Weiterhin wurde in [30, 33] empirisch festgestellt, daß die Power (Gl. (1.10)) kleiner ist, wenn die Verwandtschaftsstruktur ignoriert wird. Jedoch wurde in [31] bemerkt, daß die Power unabhängig davon ist, ob für die Verwandtschaftsstruktur korrigiert wird oder nicht. Die Allelfrequenz des SNP scheint dabei den Fehler erster Art und die Power kaum zu beeinflussen [30]. Ein weiteres Ziel dieser Arbeit ist, analytisch zu zeigen, welchen Einfluß die Heritabilität, die Verwandtschaftsstruktur und die Allelfrequenz auf den Fehler erster Art und die Power haben.

Genomic control

Genomic control [34] wird häufig dazu verwendet, eine Inflation der Teststatistik durch Verwandtschaft zu korrigieren [17, 35]. Hat man eine Stichprobe von k Realisierungen $\hat{T}_1, \hat{T}_2, \dots, \hat{T}_k$ einer Teststatistik T unter der Nullhypothese, kann man einen Korrekturfaktor $\hat{\lambda}$ schätzen:

$$\hat{\lambda} = \frac{\text{median}(\hat{T}_1^2, \hat{T}_2^2, \dots, \hat{T}_k^2)}{0,456}. \quad (1.11)$$

Eine genomic control-Korrektur der Teststatistik erfolgt dann mit

$$T_{\text{gc}} = \frac{T}{\sqrt{\hat{\lambda}}}, \quad (1.12)$$

wobei T_{gc} anstelle von T als Teststatistik heranzogen wird.

Es wurde empirisch gezeigt, daß genomic control gut für eine Korrektur eines inflationierten Fehlers erster Art funktioniert, allerdings wird die Power des Tests mit größerer Heritabilität und stärkerer Verwandtschaft verringert [17]. Ein weiteres Ziel dieser Arbeit ist deshalb, analytisch zu zeigen, welchen Einfluß die Heritabilität und die Verwandtschaftsstruktur auf den Fehler erster Art und die Power haben, falls die Teststatistik mit genomic control korrigiert wird. Genomic control wurde ursprünglich für die Korrektur von Varianzinflation der Teststatistik durch Populationsstratifikation entwickelt [36, 37]. Im Gegensatz zu anderen Arbeiten [16, 35], die sich mit Populationsstratifikation und Verwandtschaftsstruktur auseinandersetzen, wird hier jedoch auf zusätzliche Populationsstratifikation verzichtet, um die Herleitung analytischer Formeln zu ermöglichen.

Zielstellung

Wird bei einer SNP-Assoziationsanalyse anstelle des gemischten Modells aus Gl. (1.4) das einfache Modell aus Gl. (1.5) analysiert und aufgrund der Phänotyp-Korrelationsstruktur die Annah-

me unabhängiger Beobachtungen verletzt, so vergrößert sich die Varianz des Effektschätzers. Die Korrelation der Phänotypen zwischen den Probanden hängt hierbei von der Verwandtschaftsstruktur und von der Heritabilität des Phänotyps ab. Es soll analytisch gezeigt werden, welchen Einfluß die Heritabilität und die Verwandtschaftsstruktur auf den Erwartungswert und die Varianz des Effektschätzers $\hat{\beta}_2$ haben.

Die Heritabilität und die Verwandtschaftsstruktur beeinflussen auch den Fehler erster Art (Gl. (1.9)) und die Power des Tests (Gl. (1.10)), wohingegen die Allelfrequenz scheinbar keinen Einfluß hat. Diese Zusammenhänge sollen sowohl für die unkorrigierte Teststatistik aus Gl. (1.7) als auch für die mit genomic control korrigierte Teststatistik aus Gl. (1.12) analytisch beschrieben werden. Zuletzt soll abgeleitet werden, bis zu welchem Grad der Einfluß der Varianzinflation des Effektschätzers auf den Fehler erster Art und die Power tolerierbar ist und die Anwendung des einfachen linearen Modells empfohlen werden kann.

1.5 Bayesianische Assoziationsanalyse

Neben der im letzten Abschnitt dargestellten frequentistischen Analyse gibt es ein weiteres bekanntes Paradigma in der Statistik: die bayesianische Analyse, die im Folgenden vorgestellt wird. Im Gegensatz zum ersten Teil der Arbeit, wo die Phänotyp-Korrelationsstruktur einer Studie durch die Verwandtschaft zwischen den Probanden entstanden ist, wird in diesem Teil der Arbeit die Korrelation mehrerer Phänotypen innerhalb eines Probanden betrachtet. Dazu nimmt man an, daß von n unverwandten Probanden einer Studie jeweils p Phänotypen, q SNPs und r sonstige Kovariablen beobachtet wurden. Sind die Korrelationsstruktur der Phänotypen und die SNP-Phänotyp- beziehungsweise Kovariablen-Phänotyp-Assoziationen unbekannt, so stellt sich die Frage, welches Modell die gegebenen Daten bestmöglich beschreiben kann. Bei einer klassischen (frequentistischen) SNP-Assoziationsanalyse mit mehreren Kovariablen wird dazu typischerweise das Modell in Gl. (1.6) für jede Kombination eines Phänotyps und eines SNP angepaßt. Nachteil dieses Vorgehens ist aber, daß aufgrund der Kombinationsmöglichkeiten, insbesondere wenn die SNP- und Phänotyp-Anzahl groß ist, eine Vielzahl von Tests durchgeführt wird. Testet man neben dem additiven Modell noch das dominante und rezessive genetische Modell, verdreifacht sich die Anzahl der möglichen Tests zusätzlich, wodurch eine geeignete Korrektur für multiples Testen nötig ist und die Präsentation der Ergebnisse unübersichtlich wird. Weiterhin führen sowohl die im Modell fehlenden SNPs, falsch gewählte Kovariablen, die ignorierten Phänotyp-Korrelationen als auch ein möglicherweise falsch gewähltes genetisches Modell zu einer schlechten Modellanpassung. Diese Probleme lassen sich mit einem bayesianischen Ansatz adressieren. Bayesianische Modelle bieten dabei die Möglichkeit, (hierarchische) Abhängigkeiten zwischen den Modellparametern flexibel zu beschreiben und dadurch biologisches Wissen abzubilden [38–40]. Die Abhängigkeiten müssen dazu lediglich durch die Kanten eines gerichteten azyklischen Graphen repräsentiert werden [41].

Paradigmen

Zunächst werden jedoch die Unterschiede zwischen dem klassischen und dem bayesianischen Ansatz vorgestellt. Bei einer klassischen Analyse wird versucht, einen unbekannten und in Wahrheit

festen Parameter, zum Beispiel β_2 , möglichst genau mit $\hat{\beta}_2$ zu schätzen. Mit dem Schätzer und dem Schätzfehler wird eine Teststatistik konstruiert, eine bestimmte Hypothese wie $\beta_2 = 0$ geprüft und für die Interpretation ein p-Wert (Gl. (1.8)) berechnet. Bei einer bayesianischen Analyse hingegen wird angenommen, daß alle Parameter a wie zum Beispiel β_2 einer Wahrscheinlichkeitsverteilung unterliegen [42]. Es wird eine Vorannahme $P(a)$, der Prior, für diese Wahrscheinlichkeitsverteilung getroffen. Diese Vorannahme wird mit der Likelihood der Daten \mathcal{D} gegeben a verrechnet und man erhält eine aktualisierte Wahrscheinlichkeitsverteilung $P(a|\mathcal{D})$, den Posterior. Die Posterior-Verteilung ist der Ausgangspunkt für die Interpretation des Parameters. Ein Vorteil dieser Methode ist, daß einerseits Vorwissen aus Experimenten als Prior eingesetzt werden kann und andererseits der Posterior als Prior für weitere Analysen Verwendung findet und so iterativ die Posterior-Verteilung immer weiter verfeinert wird. Zudem bietet der bayesianische Ansatz die Möglichkeit, Fehlwerte zu modellieren, wobei jeder Fehlwert einem Parameter entspricht, welchem, wie den anderen Parametern auch, eine Wahrscheinlichkeitsverteilung unterstellt wird [38]. Insbesondere kann so die SNP-Korrelationsstruktur mit einem Pseudo-Haplotyp Ansatz wie [43] modelliert und anstelle der fehlenden SNP-Genotypen diese Wahrscheinlichkeitsverteilung genutzt werden. Das hat gegenüber der klassischen Analyse den Vorteil, daß es zu keiner Fallzahlreduktion aufgrund von Fehlwerten bei Phänotypen, SNPs oder Kovariablen kommt. Die Hauptkritik an der bayesianischen Analyse ist, daß für jeden Parameter ein plausibler Prior gewählt werden muß. Da die Entscheidung für eine bestimmte Prior-Verteilung aber subjektiver Natur ist, sind verschiedene Posterior-Verteilungen und damit verschiedene Schlußfolgerungen aus der Analyse möglich. Zuletzt sind bayesianische Methoden meist numerisch anspruchsvoll und rechenintensiv, was aber durch den informationstechnologischen Fortschritt kein Hindernis mehr darstellt [42].

Bayesianische Modellauswahl

Die bayesianische Modellauswahl, wie beispielsweise in [38, 44, 45] beschrieben, ermöglicht die Auswahl von SNPs und Kovariablen geordnet nach Plausibilität unter Einbeziehung der Korrelationsstruktur der Phänotypen. Alle unabhängigen Größen, von denen eine Auswahl getroffen werden soll, werden dazu in einer Matrix \mathbf{x} zusammengefaßt. Dazu zählen die Kovariablen \mathbf{c} und für jeden SNP der dominante und rezessive Anteil. Der dominante Anteil des SNP wird in $I(s_i = 2)$ und der rezessive Anteil des SNP in $I(s_i = 0)$ umkodiert. Wenn nur einer von beiden Anteilen ausgewählt wird, ist der plausibelste Einfluß der Allele des SNP entweder dominant oder rezessiv. Werden beide Anteile des SNP ausgewählt, können über den genetischen Effekt verschiedene Stufen von Kodominanz beschrieben werden. Sind beispielsweise die genetischen Effekte beider Anteile ähnlich, ist ein additiver Einfluß der Allele des SNP plausibel. In Analogie zur $n \times q$ SNP-Matrix \mathbf{s} und $n \times r$ Kovariablen-Matrix \mathbf{c} befinden sich die Messungen eines Probanden in den Zeilen der Matrix \mathbf{x} . Da für jeden SNP ein dominanter und ein rezessiver Anteil bestimmt wird, besteht Matrix \mathbf{x} aus $2q + r$ Spalten. Jede Auswahl von k Variablen beziehungsweise Spalten von \mathbf{x} stellt ein eigenes Modell m für einen der Phänotypen dar und wird durch einen Vektor $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)^T$ mit k verschiedenen Spaltenindizes beschrieben. Für jeden Phänotyp, also jeder Spalte von \mathbf{y} und einem Modell m , wird die Genotyp-Phänotyp-Beziehung

mit

$$y_i = \beta_0 + \beta_{\theta_1} x_{i\theta_1} + \cdots + \beta_{\theta_k} x_{i\theta_k} + \epsilon_i \quad (1.13)$$

modelliert. Um die Lesbarkeit zu erhöhen, wurde der Spaltenindex der Phänotypen in Gl. (1.13) weggelassen. Jeder Phänotyp besitzt eigene Parameter für die Anzahl k ausgewählter Variablen und $k + 1$ Effekte. Diese Parameter, inklusive der Modelldimension k , haben keinen festen Wert und sind lediglich durch eine Wahrscheinlichkeitsverteilung charakterisiert. Im Gegensatz zur klassischen Analyse mit Gl. (1.6) besitzt das lineare Modell dann keine feste Dimension mehr. Zuletzt wird die Phänotyp-Korrelation durch die Residuen $\epsilon_i \sim N_p(0, \Sigma)$ mit der Kovarianzmatrix Σ abgebildet. Sowohl die Variablenauswahl als auch die Einbeziehung der Phänotyp-Korrelationen in das Modell können zu einer Verbesserung bei der Identifikation genetischer Effekte beitragen [38]. Eine Modellauswahl ist zwar auch bei der klassischen Analyse möglich, beispielsweise durch verschiedene Einschlußverfahren, jedoch gibt es hierfür zahlreiche Methoden mit individuellen Vor- und Nachteilen [46], die zudem in der Praxis häufig auch unterschiedliche Ergebnisse liefern.

Inferenz

Die mehrdimensionale Posterior-Verteilung $P(\mathbf{a}|\mathcal{D})$ aller Parameter \mathbf{a} lässt sich in der Regel nicht explizit ableiten. Die Software WinBUGS [41] ermöglicht aber über die Bereitstellung verschiedener Algorithmen, zum Beispiel dem Gibbs-Sampler, das schrittweise Ziehen von Zufallswerten aus der Randverteilung des Posteriors eines jeden Parameters $a \in \mathbf{a}$ in Abhängigkeit der Zufallswerte der bereits aktualisierten anderen Parameter mittels MCMC (Markov-Chain-Monte-Carlo)-Verfahren. Details finden sich in [47]. Durch Anwendung von MCMC erhält man eine empirische Posterior-Verteilung des Parameters. Unter anderem dienen der empirische Mittelwert $\hat{E}(a|\mathcal{D})$ und die empirische Varianz $\hat{V}(a|\mathcal{D})$ der Zufallswerte der Interpretation des Parameters. Zur Bestimmung der Posterior-Verteilung der Modelle von \mathbf{y} wird eine besondere Methode verwendet: In jedem Simulationsschritt wird die Anzahl k der ausgewählten Variablen entweder verkleinert oder vergrößert. In Abhängigkeit von k wird dann ein Modell m mit passender Dimension und Zufallswerte für die zugehörigen $k + 1$ Effekte gezogen. Details zu dieser als „reversible jump“ bezeichneten Methode sind in [45, 48] beschrieben. Diese und weitere Methoden für die bayesianische Variablen- und Modellauswahl werden in [49, 50] diskutiert.

Nach dem Ziehen einer großen Anzahl von Zufallswerten, üblicherweise in der Größenordnung von mehreren 10.000, erfolgt die Analyse der gezogenen Zufallswerte beziehungsweise der beobachteten Modelle. Die 2^k möglichen Modelle eines jeden Phänotyps lassen sich nach Plausibilität ordnen, indem die Modelle nach ihrer relativen Häufigkeit $\hat{P}(m|\mathcal{D})$ ihres Auftretens geordnet werden. Die Variablen hingegen werden nach Plausibilität geordnet, indem die relative Häufigkeit $\hat{P}(i \in \boldsymbol{\theta}|\mathcal{D})$ dafür bestimmt wird, daß die Variable i in irgendeinem beobachteten Modell eingeschlossen ist. Im Gegensatz zur klassischen Analyse legt man sich nicht auf ein bestimmtes Modell oder bestimmte Variablen fest, sondern kann Modelle beziehungsweise Variablen anhand ihrer Posterior-Wahrscheinlichkeiten miteinander vergleichen. Der mittlere Effekt $\hat{E}(\beta_i|i \in \boldsymbol{\theta}, \mathcal{D})$ der Variable i wird durch Bayes model averaging (BMA) bestimmt [40, 44, 49]. Das bedeutet, der Effekt wird über alle beobachteten Modelle gemittelt, in der die dazugehörige Variable enthalten ist. Der Effekt wird hierbei nach der Häufigkeit gewichtet, mit der das Modell ausgewählt wurde. Der Vorteil dieser Methode liegt im Gegensatz zur klassischen Methode darin, daß der geschätz-

te Effekt nicht auf ein einzelnes Modell bezogen ist, wie der Beta-Schätzer, sondern über eine Vielzahl plausibler Modelle unter Einbeziehung der beobachteten Korrelationsstrukturen gemittelt wird. Dabei fällt die empirische Varianz $\hat{V}(\beta_i | i \in \theta, \mathcal{D})$ meist kleiner aus als die empirische Varianz des Effektschätzers im klassischen Modell [40]. Ein weiterer Vorteil der Methode liegt in der Präsentation der Ergebnisse, denn sowohl die Modelle als auch die Variablen mit ihren Einschlußwahrscheinlichkeiten und den gemittelten Effekten können so in verständlicher Form präsentiert werden, was beispielsweise in klinischen Journalen wichtig ist.

Eine andere Möglichkeit die Plausibilität eines Modells zu beschreiben, ist die Verwendung von Bayesfaktoren [51]. Ein Bayesfaktor ist formal definiert als das Verhältnis der Posterior-Odds zu den Prior-Odds einer Hypothese, wobei Odds das Verhältnis aus den Wahrscheinlichkeiten für das Eintreten eines Ereignisses und dem Gegenereignis beschreibt: $\text{odds}(x) = P(x)/(1 - P(x))$. Im Kontext der Modelle entspricht die Hypothese der Annahme, daß das Modell m das korrekte Modell ist. Der Bayesfaktor für die Evidenz eines bestimmten Modells m wird bestimmt [40] mit

$$\text{BF}(m) = \frac{\text{odds}(\hat{P}(m|\mathcal{D}))}{\text{odds}(\hat{P}(m))}. \quad (1.14)$$

Der Bayesfaktor für die Evidenz eines Modells lässt sich wie folgt interpretieren [51]: 1–3,2 als „nicht der Erwähnung wert“, 3,2–10 als „wesentlich“, 10–100 als „stark“ und >100 als „entscheidend“. Ein Bayesfaktor kleiner Eins spiegelt Evidenz gegen das Modell wider, wobei der reziproke Wert des Bayesfaktors analog zu vorher interpretiert wird.

Zielstellung

Eine SNP-Assoziationsanalyse kann sowohl klassisch (frequentistisch) als auch bayesianisch erfolgen. Am Beispiel einer Kinderstudie soll zunächst in einer klassischen SNP-Assoziationsanalyse nach einem Zusammenhang zwischen ausgewählten Kandidaten-SNPs und Lipidkonzentrationen gesucht werden. In der klassischen Analyse werden alle möglichen Kombinationen von Phänotypen, SNPs und genetischen Modellen getestet. Im Vergleich dazu soll ein bayesianisches Modell mit Variablenauswahl unter Einbeziehung der Phänotyp-Korrelationsstruktur der Daten konstruiert werden. Dabei sollen die aus den empirischen Posterior-Verteilungen abgeleiteten Ergebnisse, wie die beobachteten Modelle ausgewählter Variablen, die Einschlußwahrscheinlichkeiten von Variablen und die zugehörigen, über die Modelle gemittelten, Effekte mit den Ergebnissen der klassischen Analyse verglichen und interpretiert werden. Zur Beurteilung der Plausibilität beobachteter Modelle sollen neben den Posterior-Wahrscheinlichkeiten auch entsprechende Bayesfaktoren hergeleitet werden. Wir kommen nun zu den Ergebnissen dieser Arbeit.

2 Auswirkung von Korrelationsstrukturen bei Populationsvergleichen

Es ist nicht abschließend geklärt, ob Isolatpopulationen für die Erforschung genetischer Ursachen von Krankheiten oder quantitativen Phänotypen besser geeignet sind als nicht isolierte Populationen [6, 7]. Man erwartet aber aufgrund homogenerer Umwelteinflüsse, geringerer Anzahl kausaler Varianten und homogenerer Bereiche im Genom, Vorteile bei der Identifikation genetischer Ursachen in Isolatpopulationen [24, 25]. Es wird vermutet, daß die Sorben [8] einen gewissen Isolatcharakter aufweisen und sich dadurch von einer deutschen populationsbasierten Studie wie KORA [9, 10] oder von europäischen HapMap-Populationen [52] genetisch unterscheiden. Deshalb soll untersucht werden, welche Bedeutung diese Unterschiede für genetische Assoziationsanalysen haben. Die Ergebnisse sind in [1] publiziert und werden hier zusammengefaßt.

2.1 Beschreibung der Studien und genetischen Daten

Die Sorben aus der Region Lausitz sind slawischen Ursprungs und eine ethnische Minderheit [8]. Details zur Studienpopulation und zur Genotypisierung sind in [53, 54] beschrieben. Nach der Qualitätskontrolle (QC oder quality control) standen 977 Probanden für eine Analyse zur Verfügung. Für den Populationsvergleich mit den Sorben wurden Studienteilnehmer aus KORA (Kooperative Gesundheitsforschung in der Region Augsburg) gewählt. Details zur Rekrutierung und zur Studie sind in [9, 10] zu finden und Details zur Genotypisierung sind in [55] beschrieben. Für die Analyse standen 1644 Probanden aus KORA nach der QC zur Verfügung. Für einige Analysen wurden zusätzlich 174 CEU (CEPH (Centre d'Etude du Polymorphisme Humain)) Probanden und 88 TSI (Toscans in Italy) Probanden der HapMap aus Public Release 27, NCBI build 36 [52] als europäische Vergleichspopulationen herangezogen. Nach Filterung von Verwandten und der QC standen 110 CEU und 88 TSI zur Verfügung. Eine Übersicht zur Anzahl gemessener SNPs jeder Population, QC-Filter für SNPs, Anzahl von SNPs nach der QC sowie für den Populationsvergleich angewendete Methoden ist in *Additional file 1* der Publikation präsentiert.

2.2 Analyse der Verwandtschaft

Das Vorliegen von Verwandtschaften führt zu einer Verletzung der Annahme unabhängiger Beobachtungen des linearen Modells aus Gl. (1.5). Deshalb wird die Verwandtschaftsstruktur der Sorben und von KORA mit dem Schätzer aus [18] auf Basis von SNP-Daten bestimmt und verglichen. Das Ergebnis der Schätzung von 476.776 Verwandtschaftspaaren der Sorben und 1.350.546 Verwandtschaftspaaren aus KORA sind in den beiden Histogrammen in *Figure 1* dargestellt.

Durch unterschiedliche Skalen der Histogramme wird hervorgehoben, daß die Anzahl erst- und zweitgradig Verwandter bei den Sorben viel höher ist als bei KORA. Auch die Gegenüberstellung der Anzahl von Verwandtschaftspaaren über einer bestimmten Schwelle und die zugehörigen odds ratios in *Table 1* bestätigen die stärkere Verwandtschaftsstruktur der Sorben gegenüber KORA. Um bei den Sorben Merkmale genetischer Isolation von der stärkeren Verwandtschaftsstruktur der Studienpopulation zu unterscheiden, wurden Untergruppen von Verwandten und Unverwandten mit einer paarweisen Verwandtschaft $< 0,2$ gebildet und bei den folgenden Analysen untersucht. Auch ein möglicher Einfluß der Fallzahl wurde berücksichtigt, indem für KORA Subgruppen mit gleicher Fallzahl wie die entsprechende Subgruppe der Sorben erzeugt wurden.

2.3 Analyse des Kopplungsungleichgewichts

Die SNP-Korrelationsstruktur von Isolatpopulationen ist dadurch charakterisiert, daß man im Vergleich zu nicht isolierten Populationen ein im Mittel höheres Kopplungsungleichgewicht (LD) beobachtet [24, 25, 56–58]. Dadurch wird ein Vorteil bei der Identifikation von genetischen Ursachen für bestimmte Merkmale bei indirekter genetischer Assoziation erwartet. Daher soll die SNP-Korrelationsstruktur zwischen den Sorben und KORA verglichen werden. Für alle SNP-Paare von Chromosom 22 wurde dazu das Maß η_1 wie in Gl. (1.3) für KORA und die Sorben für alle Subgruppen bestimmt und in *Figure 5* dargestellt. Das gemittelte LD ist bei KORA gegenüber den Sorben für SNPs mit größeren Abständen tatsächlich geringer. Auch die erwartete Fallzahl-Verzerrung bei der LD-Schätzung, das heißt größere Werte für kleinere Fallzahl, ist für KORA gut erkennbar. Bei den Sorben hingegen ist die Fallzahl-Verzerrung interesseranterweise nicht sichtbar. Das liegt daran, daß sich hier zwei gegensätzliche Effekte ungefähr aufheben: die Verzerrung des LD durch das Filtern der Verwandtschaft und der Bias des LD hin zu größeren Werten bei geringerer Fallzahl. Daraus folgt, daß die Verwandtschaftsstruktur der Sorben zu einer Verzerrung des LD nach oben führt. Selbst nach Filterung auf Verwandtschaft zeigen die Sorben für größere Abstände zwischen den SNPs ein im Mittel höheres LD als KORA. Der Unterschied im LD zwischen Sorben und KORA ist aber moderat, was auch für andere als isoliert vermutete Populationen festgestellt wurde [59–62].

2.4 Power von SNP-Assoziationsanalysen

Unkorrelierte Phänotypen

Im nächsten Schritt wird überprüft, ob sich die beobachteten Unterschiede der SNP-Korrelationsstruktur auf die Power des Tests bei einer SNP-Assoziationsanalyse auswirken. Dazu wird wie beim einfachen linearen Modell aus Gl. (1.5) angenommen, daß ein Phänotyp \mathbf{y} durch die Genotypen eines kausalen SNP s_1 und unabhängige Residuen ϵ_1 beeinflußt wird:

$$y_i = \beta_1 s_{1i} + \epsilon_{1i}. \quad (2.1)$$

Der Unterschied von Gl. (2.1) zu Gl. (1.5) ist, daß der Achsenabschnitt weggelassen wird und die Indizes abweichen. Es wird angenommen, daß der SNP einen Teil R_s^2 der Varianz von \mathbf{y} erklärt.

Danach wird ein weiteres Modell

$$y_i = \beta_2 s_{2i} + \epsilon_{2i} \quad (2.2)$$

mit den Genotypen \mathbf{s}_2 eines zweiten SNP analysiert, der sich im Intervall von 2 Megabasen (MB) im maximalen LD, gemessen durch r^2 aus Gl. (1.2), zum ersten SNP befindet. Durch die Analyse des zweiten Modells wird eine indirekte genetische Assoziation nachempfunden. Die Verteilung des zweiten Effektschätzers $\hat{\beta}_2$ lässt sich analytisch bestimmen (*Additional file 2*) und ist

$$\hat{\beta}_2 \sim N \left(\frac{\hat{Cov}(\mathbf{s}_1, \mathbf{s}_2)}{\hat{V}(\mathbf{s}_2)}, \frac{\frac{\hat{V}(\mathbf{s}_1)}{R_s^2} - \frac{\hat{Cov}(\mathbf{s}_1, \mathbf{s}_2)^2}{\hat{V}(\mathbf{s}_2)}}{\sum_{i=1}^n (s_{2i} - \bar{s}_2)^2} \right). \quad (2.3)$$

Hierbei bezeichnen \hat{Cov} und \hat{V} die empirische Kovarianz beziehungsweise die empirische Varianz. Für die Analyse wurden alle gemessenen SNPs von Chromosom 22 analysiert und es wurde für jeden SNP ein zweiter SNP im maximalen LD zum ersten SNP gesucht. Daraufhin wurde mit der Verteilung aus Gl. (2.3) die Power (Gl. (1.10)) für einen zweiseitigen Test mit der Teststatistik aus Gl. (1.7) unter der Nullhypothese $\beta_2 = 0$ bestimmt. In *Figure 6* ist der Median der Power von KORA und den Sorben für alle Subgruppen in Abhängigkeit einer p-Wert-Schwelle dargestellt. Die p-Wert-Schwelle entspricht hierbei dem Signifikanzniveau des Tests. Trotz verschiedener erklärter Varianzen durch den kausalen SNP, sind keine sichtbaren Unterschiede zu erkennen. Auch eine tabellarische Gegenüberstellung in *Table 3* für verschiedene p-Wert-Schwellen zeigt keine nennenswerten Unterschiede für den Median der Power. Die geschätzte Power unterscheidet sich zwischen Sorben und KORA weder für die Untergruppen mit Verwandtschaft noch für die Untergruppe ohne Verwandtschaft. Daraus wird geschlossen, daß das größere LD für die Sorben gegenüber KORA nicht zu einer verbesserten Power bei der Identifikation einer indirekten genetischen Assoziation führt.

Korrelierte Phänotypen

Bei der vorhergehenden Analyse wurden unkorrelierte Phänotypen angenommen und damit Unterschiede in der Verwandtschaftsstruktur zwischen Sorben und KORA vernachlässigt. Das ist beispielsweise dann möglich, wenn die Heritabilität des Phänotyps klein ist oder wenn die Phänotypen vorher auf Verwandtschaft mit Methoden wie [17, 33] korrigiert wurden. In diesen Fällen ist auch die Formel aus Gl. (2.3) anwendbar. Im Gegensatz dazu wird jetzt der Fall betrachtet, daß die Verwandtschaftsstruktur zu korrelierten Phänotypen zwischen den Probanden führt. Die Phänotypen

$$y_i = \beta_1 s_{1i} + g_i + \epsilon_{1i} \quad (2.4)$$

werden hierzu mit einem gemischten Modell wie Gl. (1.4) beschrieben. Gl. (2.4) unterscheidet sich gegenüber Gl. (1.4) darin, daß der Achsenabschnitt weggelassen wird, die Indizes abweichen und griechische statt lateinische Buchstaben benutzt werden. Es wird angenommen, daß \mathbf{y} von den Genotypen \mathbf{s}_1 eines kausalen SNP abhängt und der SNP einen Teil R_s^2 der Varianz von \mathbf{y} erklärt. Die polygenen Effekte \mathbf{g} erklären einen Teil R_h^2 der Varianz von \mathbf{y} durch Heritabilität, was zu korrelierten Phänotypen zwischen verwandten Individuen führt. Es wurden alle gemessenen SNPs von Chromosom 22 analysiert, wobei für jeden SNP die Phänotypen mit dem Modell aus Gl.

(2.4) simuliert wurden. Darauf wurde analog zum Abschnitt **Unkorrelierte Phänotypen** für jeden SNP ein zweiter SNP im maximalen LD zum ersten SNP gesucht und das Modell aus Gl. (2.2) angepaßt. In *Table 4* ist der Median der Power von KORA und den Sorben für alle Subgruppen in Abhängigkeit verschiedener p-Wert-Schwellen für maximale Heritabilität ($R_h^2 = 1$) dargestellt. Mit maximaler Heritabilität erhält man die größten Unterschiede im Vergleich zu *Table 3*, die die Ergebnisse für unkorrelierte Phänotypen beinhaltet. Die Unterschiede in der Power für korrelierte und unkorrelierte Phänotypen sind aber gering, so daß auch hier die SNP-Korrelationsstruktur keinen großen Einfluß auf die Power hat. Obwohl durch die Analysen kein klarer Zugewinn an Power bei den Sorben gegenüber KORA festgestellt wurde, kann nicht ausgeschlossen werden, daß bei den Sorben durch homogenere Umwelteinflüsse und einer geringeren Anzahl kausaler Varianten doch ein Vorteil bei der Identifikation genetischer Ursachen besteht [63, 64]. Durch *Additional file 5* ist erkennbar, daß die Unterschiede zwischen Sorben und KORA am größten bei maximaler Heritabilität sind und daß die Simulationsergebnisse für minimale Heritabilität ($R_h^2 = R_s^2$) mit den Ergebnissen in *Table 3*, die mit der Formel aus Gl. (2.3) berechnet wurden, übereinstimmen.

Interessanterweise ist in *Table 4* für die Subgruppe der Sorben mit Verwandten bei geringer erklärter Varianz die Power größer als bei KORA und umgekehrt bei größerer erklärter Varianz ist die Power kleiner. Dieses Verhalten hängt jedoch nicht von der erklärten Varianz ab, sondern von der p-Wert-Schwelle. So wird in *Additional file 3* für die Sorben gezeigt, daß für große p-Wert-Schwellen die Power bei korrelierten Phänotypen (maximale Heritabilität) kleiner ist als bei unkorrelierten Phänotypen (minimale Heritabilität) und umgekehrt ist für kleine p-Wert-Schwellen die Power bei korrelierten Phänotypen größer. Ursache für das Phänomen ist, daß die Verwandtschaftsstruktur der Sorben eine Varianzinflation des Effektschätzers bewirkt, wie in *Additional file 4* dargestellt. Das bedeutet, daß die Power durch Verwandtschaft beeinflußt wird und bei maximaler Heritabilität den größten Unterschied zeigt. Ob die Power größer oder kleiner wird, hängt von der Stärke des genetischen Effekts und von der p-Wert-Schwelle des Tests ab. Zu diesem Zeitpunkt war aber unklar, wie die Varianzvergrößerung des Effektschätzers von der Verwandtschaftsstruktur und der Heritabilität abhängt und wie sich dieser Einfluß auf die Power des Tests überträgt. Dieser Zusammenhang wurde in der nächsten Publikation aufgeklärt.

2.5 Weitere Analysen

Es wurden weitere Analysen im Rahmen des Populationsvergleichs durchgeführt, welche hier kurz zusammengefaßt werden. Eine bekannte Methode ist die Hauptkomponentenanalyse (PCA oder principal component analysis), die sich gut dazu eignet, genetische Unterschiede verschiedener Populationen aufzudecken und grafisch darzustellen [11]. Durch eine PCA konnten die Studienpopulationen klar unterschieden werden. Nach Ausrichtung der ersten beiden Hauptkomponenten befanden sich die Sorben nordöstlich, Probanden aus KORA mit Geburtsort in Deutschland im Zentrum nahe der CEU und die TSI mit größerem Abstand weiter südlich (*Figure 2*).

Da für Isolatpopulationen vermutet wird, daß die genetische Vielfalt verringert ist [25], wurde die Anzahl seltener SNPs mit $MAF < 1\%$ zwischen Sorben und KORA miteinander verglichen. Der Unterschied in der Anzahl seltener SNPs war gering aber signifikant.

F-Statistiken beschreiben die Korrelation von Allelen von verschiedenen Perspektiven aus. Die

Korrelation der Allele von Individuen innerhalb einer Population wird mit F_{ST} dargestellt. Da F_{ST} die genetische Variation zwischen Populationen charakterisiert, kann es dazu verwendet werden, den genetischen Abstand zwischen zwei Populationen zu bestimmen. Das F_{ST} (*Table 2*) zwischen Sorben und KORA ist eine Größenordnung höher als zwischen verschiedenen Regionen in Deutschland [65]. Die Korrelation der Allele innerhalb eines Individuums wird mit F_{IS} beschrieben. Interessanterweise ist F_{IS} für KORA positiv, aber für die Sorben negativ (*Table 2*). Dies wurde schon für andere Populationen beobachtet [59] und könnte bei den Sorben ein Hinweis auf einen kürzlich stattgefundenen Verlust der genetischen Isolation sein.

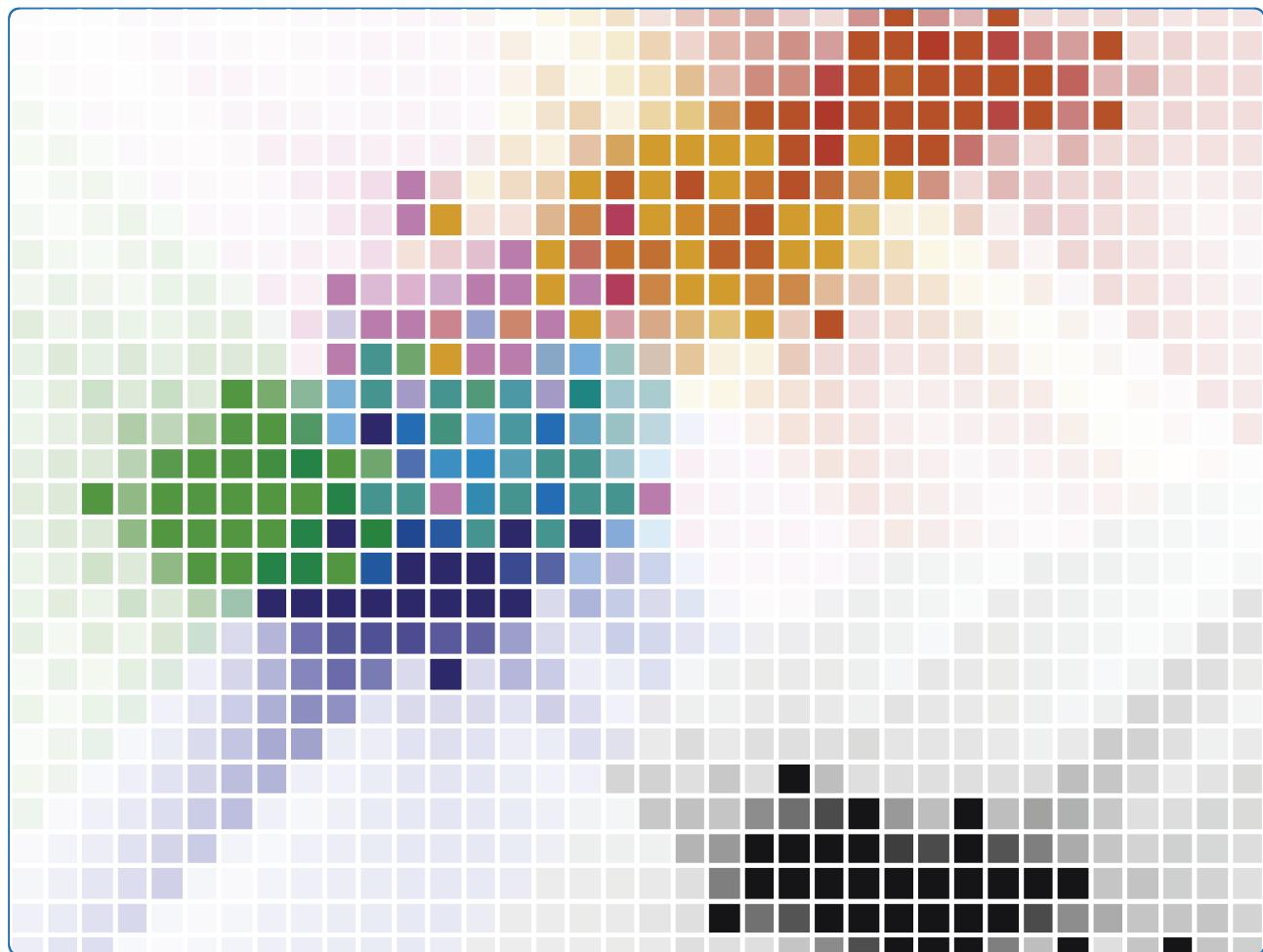
Die Analyse homozygoter Bereiche (ROHs oder runs of homoygosity) eignet sich zum Auffinden von Anzeichen genetischer Isolation durch Bestimmung von Merkmalen für Inzucht [14]. Im Vergleich zu KORA, CEU und TSI gibt es bei den Sorben einen größeren Anteil von Individuen mit ROHs mittlerer Länge (*Figure 3*). Auch die mittlere Gesamtlänge von ROHs bei gegebener Mindestlänge ist für die Sorben etwas größer (*Figure 4*). Fehlende Unterschiede bei langen ROHs sprechen bei den Sorben gegen elterliche Verwandtschaften in der jüngeren Vergangenheit. Aber aufgrund von Unterschieden in ROHs von mittlerer Länge gibt es bei den Sorben Anzeichen von elterlichen Verwandtschaften beziehungsweise einem kleinen Founderpool in der älteren Vergangenheit entsprechend der Interpretation in [14].

2.6 Zusammenfassung

Die Sorben zeigen Merkmale genetischer Isolation, die nicht auf eine stärkere Verwandtschaftsstruktur der Studienpopulation zurückzuführen sind. Die Merkmale genetischer Isolation sind moderat, trotzdem ist der slawische Ursprung erkennbar. Daraus lässt sich schließen, daß die Sorben ursprünglich genetisch isoliert waren, jedoch die genetische Isolation verloren geht. Trotz eines im Mittel höheren LD bei den Sorben, unabhängig von der stärkeren Verwandtschaftsstruktur, ist kein klarer Vorteil bei der Power von SNP-Assoziationsanalysen zu erwarten. Die Verwandtschaftsstruktur der Sorben kann bei unkorrigierten SNP-Assoziationsanalysen zu einer Varianzinflation des Effektschätzers führen und die Power des Tests beeinflussen. Es soll in einer weiteren Publikation geklärt werden, wie die Verwandtschaftsstruktur der Studienpopulation und die Heritabilität eines Phänotyps die Varianz des Effektschätzers und die Power des Tests tatsächlich beeinflussen.

2.7 Publikation

In diesem Abschnitt befindet sich eine Kopie der Publikation von: A. Gross, A. Tonjes, P. Kovacs, et al. Population-genetic comparison of the Sorbian isolate population in Germany with the German KORA population using genome-wide SNP arrays. *BMC Genet.*, 12:67, Jul 2011. doi:10.1186/1471-2156-12-67.



Population-genetic comparison of the Sorbian isolate population in Germany with the German KORA population using genome-wide SNP arrays

Gross *et al.*

RESEARCH ARTICLE

Open Access

Population-genetic comparison of the Sorbian isolate population in Germany with the German KORA population using genome-wide SNP arrays

Arnd Gross^{1,2}, Anke Tönjes^{3,4}, Peter Kovacs⁵, Krishna R Veeramah^{6,7,8}, Peter Ahnert^{1,2}, Nab R Roshyara^{1,2}, Christian Gieger⁹, Ina-Maria Rueckert⁹, Markus Loeffler^{1,2}, Mark Stoneking¹⁰, Heinz-Erich Wichmann^{9,11,12}, John Novembre⁶, Michael Stumvoll^{3,4} and Markus Scholz^{1,2*}

Abstract

Background: The Sorbs are an ethnic minority in Germany with putative genetic isolation, making the population interesting for disease mapping. A sample of $N = 977$ Sorbs is currently analysed in several genome-wide meta-analyses. Since genetic differences between populations are a major confounding factor in genetic meta-analyses, we compare the Sorbs with the German outbred population of the KORA F3 study ($N = 1644$) and other publicly available European HapMap populations by population genetic means. We also aim to separate effects of over-sampling of families in the Sorbs sample from effects of genetic isolation and compare the power of genetic association studies between the samples.

Results: The degree of relatedness was significantly higher in the Sorbs. Principal components analysis revealed a west to east clustering of KORA individuals born in Germany, KORA individuals born in Poland or Czech Republic, Half-Sorbs (less than four Sorbian grandparents) and Full-Sorbs. The Sorbs cluster is nearest to the cluster of KORA individuals born in Poland. The number of rare SNPs is significantly higher in the Sorbs sample. FST between KORA and Sorbs is an order of magnitude higher than between different regions in Germany. Compared to the other populations, Sorbs show a higher proportion of individuals with runs of homozygosity between 2.5 Mb and 5 Mb. Linkage disequilibrium (LD) at longer range is also slightly increased but this has no effect on the power of association studies. Oversampling of families in the Sorbs sample causes detectable bias regarding higher FST values and higher LD but the effect is an order of magnitude smaller than the observed differences between KORA and Sorbs. Relatedness in the Sorbs also influenced the power of uncorrected association analyses.

Conclusions: Sorbs show signs of genetic isolation which cannot be explained by over-sampling of relatives, but the effects are moderate in size. The Slavonic origin of the Sorbs is still genetically detectable. Regarding LD structure, a clear advantage for genome-wide association studies cannot be deduced. The significant amount of cryptic relatedness in the Sorbs sample results in inflated variances of Beta-estimators which should be considered in genetic association analyses.

Background

The Sorbs living in the Upper Lusatia region of Eastern Saxony are one of the few historic ethnic minorities in Germany. They are of Slavonic origin speaking a west Slavic language (Sorbian), and it is assumed that they have lived in ethnic isolation among the German

majority during the past 1100 years [1]. Therefore, this population may be of special interest for genetic studies of complex traits.

The value of isolated populations for the discovery of genetic modifiers of diseases or quantitative traits is discussed controversially [2-6]. On the one hand, reduced genetic and environmental variability of isolated populations could increase genotypic relative risks [7,8]. In combination with the generally higher degree of linkage disequilibrium (LD) in isolated populations, this could

* Correspondence: markus.scholz@imise.uni-leipzig.de

¹Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Haertelstrasse 16-18, 04107 Leipzig, Germany
Full list of author information is available at the end of the article

improve the power of genetic association studies [5,6,9-11]. On the other hand, studies in isolated populations are often limited in size and, therefore, cannot match modern genome-wide association studies and meta-analyses comprising several tens of thousands of individuals.

Nowadays, it is common practice to combine all available genotyped and phenotyped populations in large-scale, whole genome meta-analyses or pooled analyses in order to identify even very small genetic effects as commonly observed for complex traits. Spurious associations caused by the genetic sub-structures of combined populations are the most serious concern of this approach [12-15], implying the need for appropriate adjustment strategies [16,17]. This is especially true if evidence from isolated and outbred populations is combined as this approach necessitates a thorough comparison of populations by population genetic means in order to determine their "degree of isolation" [6]. For this purpose, different methods have been proposed in the literature. For example, length and number of runs of homozygosity (ROHs) are discussed as an appropriate measure of isolation since they measure the degree of parental consanguinity [18]. LD is estimated to be higher in isolated populations because of lower generation numbers resulting in fewer recombination events [5,6]. Due to the smaller size of the founder population, it can also be expected that there is a lower number of polymorphisms in isolated populations [6,19,20]. Other markers of population structure such as F-statistics [21] are related to the measures mentioned above. Furthermore, genetic distances between populations can be determined by principal components analysis (PCA), allowing to quantify how closely populations are related [22]. By this technique genetic information can be mapped to topographic maps [14] allowing the assessment of a new indicator of isolation in the sense that an isolated population could be genetically far away from their geographic location. So far there appears to be no single measure sufficient to characterize the isolation of a population.

Another characteristic feature of isolated populations is the putatively higher degree of cryptic relatedness in randomly drawn samples. This is a serious concern in genetic association analysis and needs to be addressed with appropriate statistical methods [17,23-25]. Relatedness of individuals could also interact with the above mentioned measures of isolation of populations. Thus, when comparing two populations with different degrees of cryptic relatedness, it is not easy to decide whether differences in these measures can be traced back to different degrees of isolation or simply to over-sampling of related subjects.

The degree of isolation of the Sorbs has been studied in the past by the analysis of Y-chromosomal markers [26]. Recently, we compared a subset of about 200 Sorbs with other European isolates using 30,000 SNPs measured by microarrays [1]. In this analysis, the Sorbs expressed only moderate signs of isolation. Here, we analyse a sample of $N = 977$ Sorbs, which is currently included in several genome-wide association studies e.g. [27,28], and compare the Sorbs with the German outbred population of the KORA study [29]. Using the KORA study ($N = 1644$) and a larger sample of Sorbs ($N = 977$) provides more power than previous studies for comparing population genetic patterns between Sorbs and their neighbours. For this purpose, we assess the above mentioned population genetic characteristics: PCA, number of rare SNPs, F-statistics, ROHs, and LD. All analyses are based on genome-wide SNP array data. We also aim to separate effects of cryptic relatedness from effects of genetic isolation.

Furthermore, we analyse how differences between populations can be translated to differences in power of genetic association studies within these samples. We analyse the influence of genetic effect size, LD structure, heritability, and relatedness on power.

Methods

Study Populations

Sorbs

The Sorbs are of Slavonic origin, and lived in ethnic isolation among the Germanic majority during the past 1100 years [1]. Today, the Sorbian-speaking, Catholic minority comprises 15,000 full-blooded Sorbs resident in about 10 villages in rural Upper Lusatia (Oberlausitz), Eastern Saxony. A convenience sample of this population was collected including unrelated subjects as well as families. Details of the study population can be found elsewhere [28,30]. Genotyping and metabolic phenotyping of this sample was approved by the ethics committee of the University of Leipzig and is in accordance with the declaration of Helsinki. All subjects gave written informed consent before taking part in the study. A subset of individuals were genotyped with either Affymetrix Human Mapping 500 K Array Set ($N = 483$) or Affymetrix Genome-Wide Human SNP Array 6.0 ($N = 494$). Details on genotyping are described in [28]. A total of 977 subjects were available after quality control.

KORA

The study population was recruited from the KORA/MONICA S3 survey, a population-based sample from the general population living in the region of Augsburg, Southern Germany, which was carried out in 1994/95. In a follow-up examination of S3 in 2004/05 (KORA F3), 3006 subjects participated. Recruitment and study

procedures of KORA have been described elsewhere [29,31]. For KORA F3 500 K we selected 1644 subjects of these participants then aged 35 to 79 years. Informed consent has been given, and the study has been approved by the local ethics committee. All KORA participants have a German passport. Genotyping of these individuals was performed with the Affymetrix Gene Chip Human Mapping 500 K Array Set as described in [32].

HapMap

174 CEU (CEPH (Centre d'Etude du Polymorphisme Humain) from Utah) and 88 TSI (Toscans in Italy) samples were taken from a recent HapMap Collection (Public Release 27, NCBI build 36, The International HapMap Project). From the CEU sample, we removed 58 children, five individuals with call rate < 90% and one individual because of cryptic relatedness (NA07045 because of lower call-rate compared to NA12813 [33]). In summary, we analysed 110 CEU and 88 TSI samples.

Data Analysis

Genotype Imputation and Quality Control

Missing genotypes of the KORA and Sorb samples were imputed separately using MACH Imputation Software with standard settings [34].

After Imputation, we checked 471,012 autosomal SNPs in the overlap of the Affymetrix Human Mapping 500 K Array Set and Affymetrix Genome-Wide Human SNP Array 6.0 for quality.

SNPs with a call rate less than 95% in all four study populations combined, prior to imputation, were filtered (34,711 SNPs). Hardy-Weinberg-Equilibrium (HWE) was tested across populations using a stratified test proposed by [35]. 10,712 SNPs with p-values less than 10^{-6} were eliminated. Finally, 14,508 SNPs showing unexpectedly high differences of allelic frequencies between genotyping platforms in the Sorbs sample were eliminated ($p\text{-value} < 10^{-7}$, see [1] for further details).

Since several SNPs violated more than one of our criteria, we discarded a total of 46,536 SNPs and analysed 424,476 remaining SNPs.

For estimation of ROHs (see below) the number of analysed SNPs is reduced to 306,081 by matching SNPs on Affymetrix chips with available SNPs in the HapMap CEU and TSI samples. Due to the high sensitivity of the PCA (see below) we decided to tighten our quality criteria for this kind of analysis. Only SNPs with a call rate of at least 99% were included for PCA, which reduced the number of SNPs to 199,702.

An overview of the data pre-processing workflow can be found in Additional file 1.

Estimation of Relatedness

Pair-wise relatedness between all individuals of KORA and Sorbs was estimated by the method described in

[36]. For first degree relatives one would expect a value of $r = 0.5$, for second degree relatives a value of $r = 0.25$, and so on. Two individuals were considered as unrelated if the pair-wise relatedness estimate was not greater than 0.2, which approximately corresponds to the exclusion of first and second degree relatives.

For analyses of dependence of measures of population genetic comparison on relatedness, we define two subsamples used for all subsequent analyses: For the first subsample, the complete Sorbs sample (Sorbs_{977} , $N = 977$) was matched with a randomly selected subset of $N = 977$ unrelated KORA subjects born in Germany (KORA_{977}). For the second subsample, a subset of $N = 532$ unrelated Sorbs (Sorbs_{532}) was matched with a subset of $N = 532$ KORA subjects (KORA_{532}) randomly selected from KORA_{977} .

Unrelated subjects were selected by an algorithm which implements a step-by-step removal of individuals showing the highest number of relationships to other members of the population until no pair of individuals with relatedness > 0.2 remained.

Principal components analysis

PCA is suitable to map genetic variance to a few dimensions expressing the highest degree of variance [16,22]. It has been shown recently that the application of this technique to genome-wide genetic data is powerful enough to mirror even small geographic distances in Europe [14,37].

Since PCA results are biased in case of unequal population sizes [38], it was necessary to analyse subsamples of our populations. We performed PCA of 350 individuals from 7 subsamples of size $N = 50$, generated from the most unrelated individuals of our four study populations. The subsamples were defined as follows. Three subsamples were created from $N = 1336$, $N = 140$, and $N = 80$ individuals from KORA, who were born in Germany, in the Czech Republic, and in Poland, respectively. Two subsamples were generated from the Sorbs grouped by their degree of Sorbian ancestry. We identified 786 "Full"-Sorbs who stated that all four grandparents are Sorbs and 160 "Half"-Sorbs where at least one grandparent was not Sorbian. Another two subsamples were built from 110 CEU and 88 TSI samples.

PCA was done with iterative removal of outliers (default 5 iterations) and LD correction in consecutive SNPs (involving two previous SNPs as recommended in the manual of the EIGENSOFT package).

Rare SNPs

Isolated populations are supposed to have reduced genetic variability resulting in a higher number of rare SNPs. By definition, a SNP has a minor allelic frequency (MAF) of at least 1%. To account for variance we calculated the exact 95% confidence interval of the MAF and considered a SNP as rare if the interval was below one

percent. This is equivalent to less than 11 observed alleles in Sorbs₉₇₇ or KORA₉₇₇ and less than five observed alleles in Sorbs₅₃₂ or KORA₅₃₂ respectively. The odds to find rare SNPs were compared between KORA and Sorbs using Fisher's exact test.

F-statistics

To characterize the variance of allelic frequencies within and between populations, we calculated F-statistics.

The inbreeding coefficient F_{IS} measures the correlation of alleles within an individual relative to the corresponding population. It is calculated by estimating the deviance of the observed number of heterozygote genotypes from what is expected under HWE. For every SNP, we calculated unbiased estimates as presented in [21], assessed the weighted average and determined the standard error of estimates by jack-knifing over individuals.

Correlation of alleles of individuals in the same population was estimated by the co-ancestry coefficient F_{ST} . Since F_{ST} quantifies the amount of genetic variation between populations, it is used to define genetic distances between populations. We assessed F_{ST} for pairs of populations using a combined estimate of all SNPs [21] and calculated the standard error of estimates again by jack-knifing over individuals.

Runs of homozygosity

Counting ROHs is useful to detect inbreeding [18]. ROHs were determined in all individuals from KORA, Sorbs, CEU, and TSI using the PLINK Package (Version 1.07) with standard settings except for two parameters as noted below. PLINK estimates ROHs by searching for contiguous runs of homozygote genotypes. For this purpose, a window (default length 5000 kb, minimum 50 SNPs) is moved along the genome. To account for possible genotyping errors, at each SNP the homozygosity of the window is assessed allowing one (default) heterozygous genotype and five (default) missing calls. For each SNP the proportion of overlapping homozygous windows is calculated. If this proportion is high enough (default 5%) the SNP is considered to be part of a homozygous segment. Only homozygous segments longer than a given threshold (500 kb, default 1000 kb), consisting of a minimum number of 100 SNPs (default) and comprising a minimum SNP density of one SNP per 50 kb (default) were denoted as ROH. A homozygous segment can be split in two if two SNPs are at least 100 kb apart (default 1000 kb). Details on the algorithm can be found on the PLINK Homepage (see URLs).

Linkage disequilibrium

In the Sorbs and KORA samples, we calculated pairwise LD for all SNPs on Chromosome 22 (5382 markers) using robust estimators [39]. We used the widely

accepted measures r [40] and $|D'|$ [41] to quantify LD. Since both measures depend on allelic frequencies, we also used the newly proposed measure $|\eta_1|$, which is independent of allelic frequencies. Hence, it is especially useful when comparing populations [42]. The measure η_1 is a monotone function of the odds ratio λ [43] ranging between -1 and 1. It is defined as

$$\eta_1 = \begin{cases} 2\frac{\lambda^2 - \lambda - \lambda \ln \lambda}{(\lambda - 1)^2} - 1 & \text{if } \lambda \neq 1 \\ 0 & \text{if } \lambda = 1 \end{cases}$$

Its absolute value is the percentage of SNP pairs under the non-informative uniform distribution with less extreme LD than the one observed (see [42] for details). Measures of LD were averaged using bins of 5 kb length as proposed by Olshen et al. [44]. Resulting means were smoothed by a LOWESS estimator [45].

Comparison of power assuming uncorrelated phenotypes

We analysed how the observed differences in LD structure between KORA and Sorbs can be translated into differences in power of genetic association studies. For this purpose, we assumed a linear regression model $y = \beta_1 s_1 + \varepsilon_1$ of a random phenotype y which is influenced by a genotype s_1 of a causative SNP, and ε_1 is the residual Gaussian error of the model.

The SNP is assumed to explain a pre-specified proportion of the total variance of the phenotype which is denoted as R_s^2 in the following. In consequence, we can assume $\beta_1 = 1$ without restriction of generality. Within the distance of ± 2 Mb we now analysed the model $y = \beta_2 s_2 + \varepsilon_2$ for a second SNP, which is in maximum LD (measured by r) with the causative SNP. That is, we analysed the best proxy of the causative SNP rather than the causative SNP itself modelling the marker principle of genetic association studies. The estimator $\hat{\beta}_2$ is normally distributed and depends on s_1 , s_2 , and R_s^2 :

$$\hat{\beta}_2 \sim N \left(\frac{\text{Cov}(s_1, s_2)}{\text{Var}(s_2)}, \frac{\frac{\text{Var}(s_1)}{R_s^2} - \frac{\text{Cov}(s_1, s_2)^2}{\text{Var}(s_2)}}{\sum_{i=1}^n (s_{2i} - \bar{s}_2)^2} \right).$$

Where n is the number of individuals, s_{2i} is the genotype of the i -th individual and \bar{s}_2 is the average. The formula is derived in Additional file 2. We calculated the power of the regression analysis, i.e. the probability that the observed p-value is smaller than a given significance level (p-value threshold) when testing $\hat{\beta}_2$ against the null hypothesis $\beta_2 = 0$ using the above formula. This was done for all SNPs on Chromosome 22 in KORA₉₇₇, KORA₅₃₂, Sorbs₉₇₇, and Sorbs₅₃₂. Distribution of power

was derived using the results of all SNPs of Chromosome 22. Results were compared between the KORA and Sorbs samples of equal size.

Comparison of power assuming correlated phenotypes

In the previous section, we derived formulae for the estimation of power under the assumption of uncorrelated phenotypes. This approach applies for either a negligible relatedness structure of the individuals or a weak correlation of phenotypes of related individuals. Applying a GRAMMAR approach [17], deviations from this situation can be corrected resulting again in the situation considered in the previous section.

However, to our knowledge, it is still not common practice in genome-wide association studies to use this approach to correct for relatedness. Therefore, we aim to study the situation in which the phenotypes are correlated but in which the corresponding individuals were analysed as independent even though they are not.

Following Amin *et al.* [17], we simulated phenotypes \mathbf{y} on the basis of the mixed model $\mathbf{y} = \beta_1 \mathbf{s}_1 + \mathbf{g} + \boldsymbol{\varepsilon}_1$, comprising a fixed effect of genotypes \mathbf{s}_1 , a random effect representing the residual polygenic effects $\mathbf{g} \sim N_n(\mathbf{0}, \sigma_g^2 \mathbf{G})$ and non-genetic residuals $\boldsymbol{\varepsilon}_1 \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I})$. Here, \mathbf{G} represents the pair-wise relatedness matrix. The model results in non-trivial covariance of phenotypes of different individuals. For each SNP we drew 1000 samples from the model and analysed the linear model $\mathbf{y} = \beta_2 \mathbf{s}_2 + \boldsymbol{\varepsilon}_2$ for a second SNP which is in maximum LD to the first SNP in complete analogy to the procedure developed for uncorrelated phenotypes (see previous section). Different degrees of

heritability $R_h^2 = R_s^2 + R_g^2$ were simulated, where R_s^2 is the explained variance by genotypes \mathbf{s}_1 and R_g^2 is the explained variance by polygenic effects \mathbf{g} . Providing values for R_h^2 and R_s^2 results in the variance components $\sigma_g^2 = \text{Var}(\mathbf{s}_1)(\frac{R_h^2}{R_s^2} - 1)$ and $\sigma^2 = \text{Var}(\mathbf{s}_1)\frac{1 - R_h^2}{R_s^2}$, which follow after some calculations.

Statistical Software and Web-Resources

HapMap data were downloaded from [46]. Estimation of Eigenvectors for comparison of all subsamples was done with the EIGENSOFT package (Version 3.0, [47]). ROHs were determined by the PLINK Package (Version 1.07, [48]) [49].

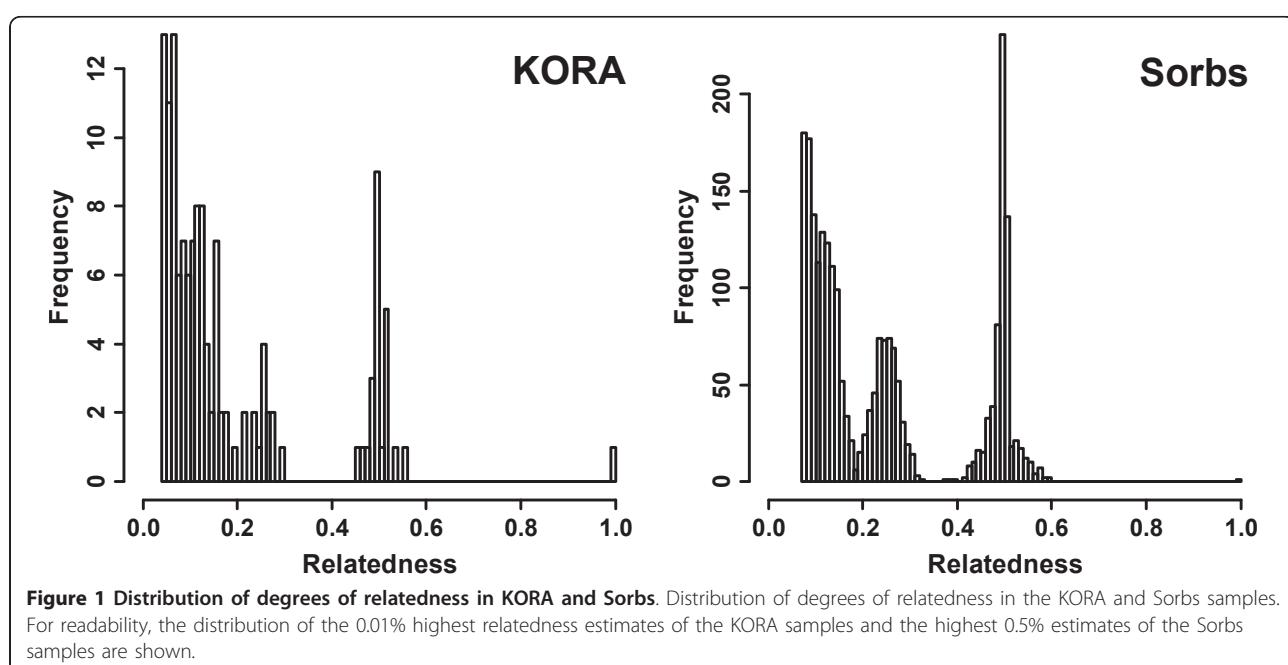
All other calculations were performed using the Statistical Software package R (Version 2.8.0, [50]) [51].

Results

For population genetic comparison of the Sorbian minority in Germany with the German KORA population, several measures of genetic isolation were applied to genome-wide SNP array data.

Relatedness

We analysed the relatedness of all 476,776 pairs of individuals in the Sorbs and all 1,350,546 pairs in the KORA samples. Results are shown in Figure 1. Frequencies of relationships differ remarkably between the two samples. Emphasized by the different scales of the histograms, it can be clearly recognized that the numbers of



first and second degree relationships are higher in the Sorbs compared to KORA. Numbers of pairs with estimates over a given threshold are shown in Table 1 for both populations. We also provide odds-ratios for the encounter of a related pair.

To achieve samples without pairs of individuals with relatedness-estimates greater than 0.2, it was necessary to exclude 445 Sorbs and 33 KORA individuals, resulting in subsamples of 532 Sorbs and 1,611 KORA individuals.

Principal components analysis

Results of PCA after removal of outliers and LD correction are shown in Figure 2. The figure comprises all 150 individuals from KORA, 97 Sorbs, 49 HapMap CEU and 48 HapMap TSI after outlier removal.

A plot of the genetic variance represented by the first two principal components impressively reflects the geographic origin of these populations. TSI samples are relatively far away from the other clusters giving an orientation of a north to south axis. The KORA population is very close to the CEU HapMap population. In contrast, the Sorbian population clusters significantly eastwardly. There is a clear trend of west to east clustering of KORA individuals born in Germany, KORA individuals born in Poland or Czech Republic, Half-Sorbs, and finally, Full-Sorbs. The Sorbs clusters are nearest to the cluster of KORA individuals born in Poland.

Rare SNPs

When analysing 424,476 quality SNPs in 977 Sorbs (Sorbs₉₇₇) and the random Sample of 977 individuals from KORA (KORA₉₇₇), we counted 51,204 rare SNPs in Sorbs₉₇₇ and 49,721 rare SNPs in KORA₉₇₇ (p -value 6.7×10^{-7}). In the subset of 532 unrelated Sorbs (Sorbs₅₃₂) and the random sample of 532 unrelated individuals from KORA (KORA₅₃₂), we counted again more rare SNPs in the Sorbs₅₃₂ than in KORA₅₃₂, i.e. 49,257 and 47,913 (p -value 4.7×10^{-6}), respectively.

F-Statistics

Estimating F_{IS} in the samples KORA₉₇₇ and KORA₅₃₂ resulted in slightly positive values with the smaller value in KORA₉₇₇. In contrast, in the samples Sorbs₉₇₇ and

Sorbs₅₃₂, we find slightly negative values with smaller value in the sample Sorbs₉₇₇.

F_{ST} estimates are somewhat higher between KORA₉₇₇ and Sorbs₉₇₇ than between KORA₅₃₂ and Sorbs₅₃₂. F_{ST} estimates are higher than corresponding F_{IS} estimates, indicating a clear genetic distance between the two populations. All statistics can be found in Table 2.

Runs of Homozygosity

ROHs were determined for the populations KORA, Sorbs₉₇₇, Sorbs₅₃₂, CEU, and TSI. Percentages of individuals in these populations containing at least one ROH in a specified length interval were calculated (Figure 3). Compared to the other populations, Sorbs show a higher proportion of individuals with ROHs between 2.5 Mb and 5 Mb.

In a second step, mean total length of ROHs with a given minimum length was estimated averaged over the individuals of each population (Figure 4). Again, Sorbs differ from the other populations and are characterized by higher mean total length of ROHs. However, the effect is less pronounced if only long ROHs are considered. The mean total length of ROHs is shorter for Sorbs₅₃₂ than for Sorbs₉₇₇ but the difference is small.

Linkage Disequilibrium

Three measures of LD were calculated for KORA₉₇₇, KORA₅₃₂, Sorbs₉₇₇, and Sorbs₅₃₂. Results of η_1 are shown in Figure 5. Other measures such as r and D' behave similarly (data not shown). LD in the KORA sample is markedly lower at long ranges compared to Sorbs. This result is robust against dropping related individuals in the Sorb sample.

As expected for KORA₉₇₇ and KORA₅₃₂ a small sample size bias can be observed. In contrast the estimators for Sorbs₉₇₇ and Sorbs₅₃₂ are virtually identical.

Comparison of power assuming uncorrelated phenotypes

The power to detect causal SNPs was calculated for KORA₉₇₇, KORA₅₃₂, Sorbs₉₇₇, and Sorbs₅₃₂. Results for SNP effects with explained variances of 2% or 5% can be found in Figure 6. Since the results are virtually identical for KORA and Sorbs, we present the quartiles of the power distribution in Table 3 for p -value thresholds of 1×10^{-5} and 1×10^{-7} .

Table 1 Distribution of pair-wise relatedness estimates

Lower Bound	Number of pairs in KORA	Number of pairs in Sorbs	Odds ratio (KORA = reference category) [95% CI]
0.1	79	1889	68 [54;86]
0.2	38	1186	88 [64;126]
0.4	24	666	79 [52;123]
0.6	1	1	3 [0;222]

Number of pair-wise relatedness estimates above a given boundary for a total of 476776 and 1350546 calculated pair-wise estimates in Sorbs and KORA, respectively. We also present the odds-ratio for an encounter of relatives and corresponding 95% confidence interval.

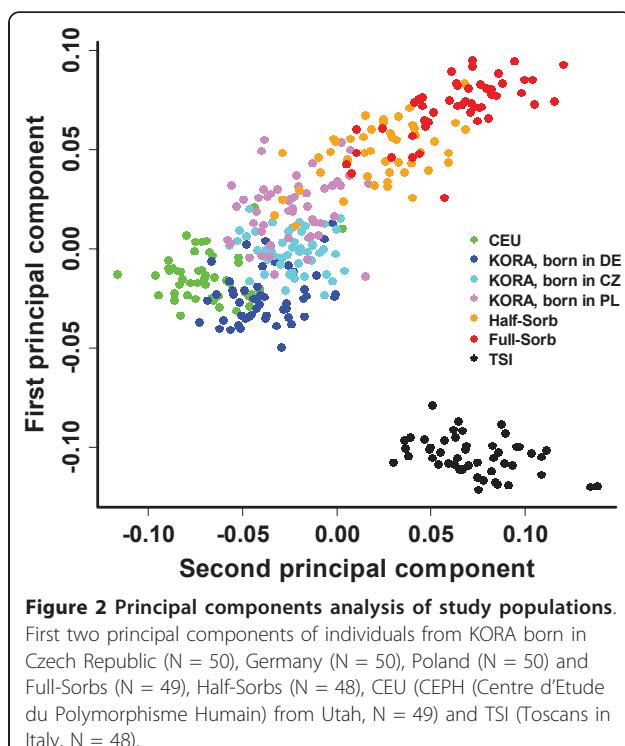


Figure 2 Principal components analysis of study populations.

First two principal components of individuals from KORA born in Czech Republic (N = 50), Germany (N = 50), Poland (N = 50) and Full-Sorbs (N = 49), Half-Sorbs (N = 48), CEU (CEPH (Centre d'Etude du Polymorphisme Humain) from Utah, N = 49) and TSI (Toscans in Italy, N = 48).

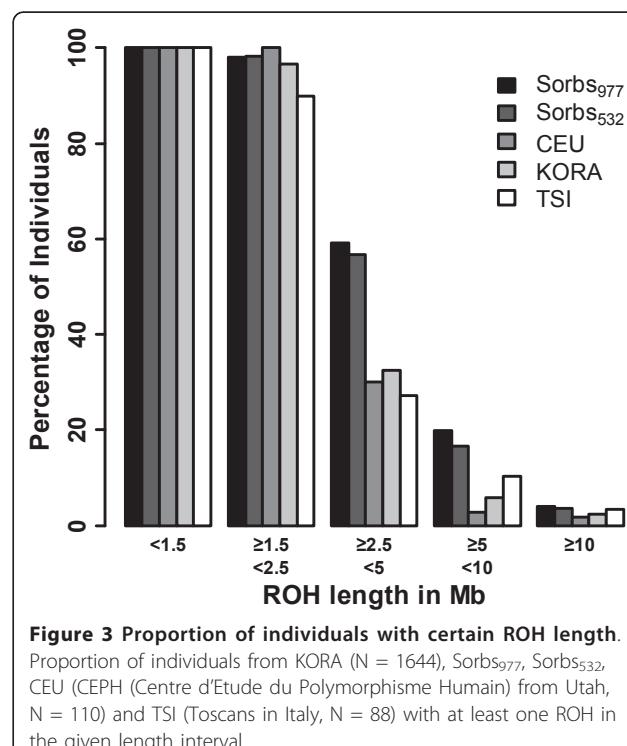


Figure 3 Proportion of individuals with certain ROH length.

Proportion of individuals from KORA (N = 1644), Sorbs₉₇₇, Sorbs₅₃₂, CEU (CEPH (Centre d'Etude du Polymorphisme Humain) from Utah, N = 110) and TSI (Toscans in Italy, N = 88) with at least one ROH in the given length interval.

Comparison of power assuming correlated phenotypes

In Table 4 we present the power estimates assuming a heritability of 100% resulting in the greatest differences compared to Table 3. However, except for Sorbs₉₇₇, there are only very small differences between Tables 3 and 4 and even for Sorbs₉₇₇ the differences appear to be not substantial. For an explained variance of 2%, the power in Sorbs₉₇₇ increases, but it decreases for an explained variance of 5%. This is due to dependence on the significance threshold. Independent of the explained variance of the SNPs, the power under maximum heritability (100%) is greater than under minimal heritability (R_s^2) for small p-value thresholds. But for large p-value thresholds, the opposite is true (see Additional file 3).

The explanation for this behaviour is the inflation of the variance of the β -estimator caused by high levels of relatedness in the Sorbs₉₇₇ sample (see Additional file 4).

Results for other degrees of heritability are presented in Additional file 5. As expected, in the case of minimal heritability the results of our simulations under the mixed model and the results obtained with our analytical formula used in the previous section are coincident.

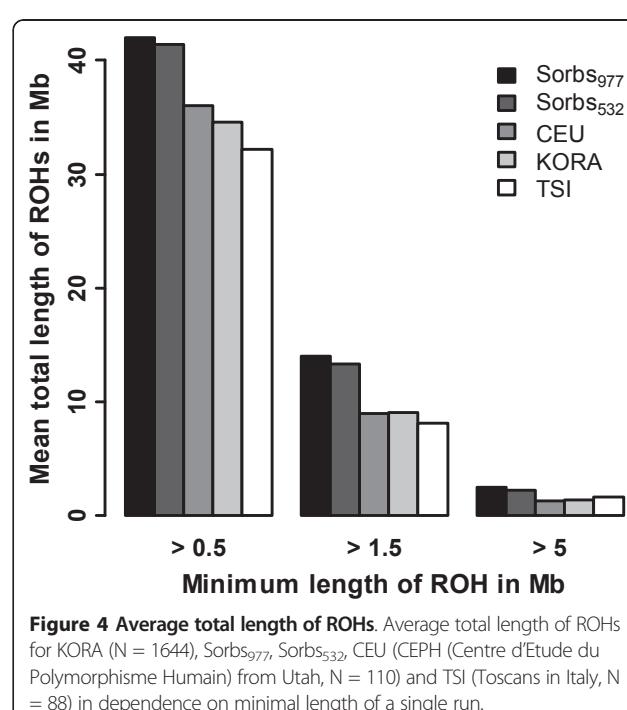


Figure 4 Average total length of ROHs. Average total length of ROHs for KORA (N = 1644), Sorbs₉₇₇, Sorbs₅₃₂, CEU (CEPH (Centre d'Etude du Polymorphisme Humain) from Utah, N = 110) and TSI (Toscans in Italy, N = 88) in dependence on minimal length of a single run.

Table 2 Inbreeding and co-ancestry coefficients

Population	F-statistic	Estimate	SE
KORA ₉₇₇	F_{IS}	0.0012	2.7×10^{-4}
Sorbs ₉₇₇	F_{IS}	-0.0006	2.7×10^{-4}
KORA ₅₃₂	F_{IS}	0.0014	3.5×10^{-4}
Sorbs ₅₃₂	F_{IS}	-0.0002	3.6×10^{-4}
KORA ₉₇₇ , Sorbs ₉₇₇	F_{ST}	0.0034	5.4×10^{-5}
KORA ₅₃₂ , Sorbs ₅₃₂	F_{ST}	0.0029	6.7×10^{-5}

Estimates and standard errors (SE) of inbreeding coefficients F_{IS} and co-ancestry coefficients F_{ST} for KORA and Sorbs.

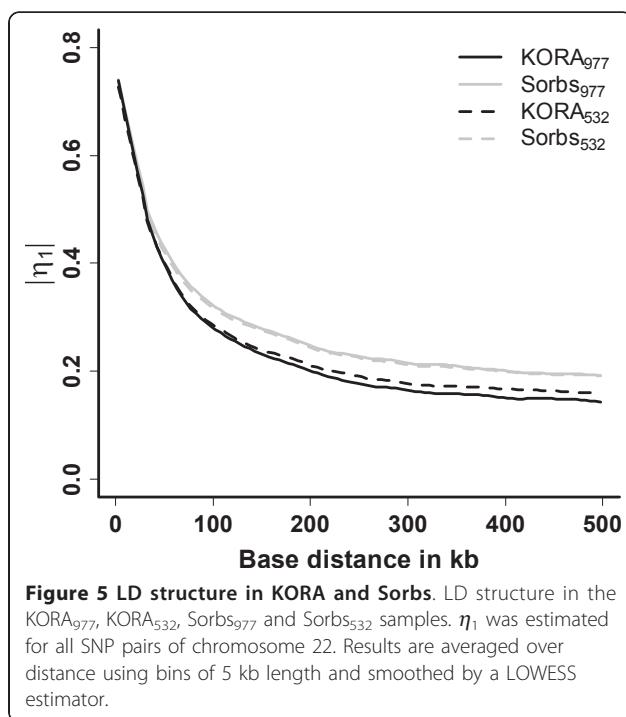


Figure 5 LD structure in KORA and Sorbs. LD structure in the KORA₉₇₇, KORA₅₃₂, Sorbs₉₇₇ and Sorbs₅₃₂ samples. η_1 was estimated for all SNP pairs of chromosome 22. Results are averaged over distance using bins of 5 kb length and smoothed by a LOWESS estimator.

Discussion

The Sorbs, resident in Lusatia, Germany, are an ethnic minority of Slavonic origin. Using genome-wide SNP array techniques, we aimed to compare this putatively isolated population with a German mixed population (KORA study) by various population genetic means. The Sorbs were compared recently with other European

populations or isolates on the basis of a limited set of genetic markers and a limited set of unrelated individuals [1,52]. In the present analysis, we studied the Sorbs from the perspective of ongoing genome-wide association studies. That is, we compared the population with a German mixed population on the basis of complete sets of genotyped individuals, and a large number of genotyped SNPs. We also aimed to separate the effect of isolation from potential effects caused by over-sampling of relatives in the Sorbs. Finally, we studied the implications of observed differences between KORA and Sorbs for the analysis, and especially, the power of genome-wide association studies.

Genotype data from a sample of 977 Sorbs were available from genotyping with 500 k and 1000 k Affymetrix SNP chips. While SNP markers come with certain drawbacks (ascertainment bias, need for careful QC), they have proven useful for detecting subtle population structures.

For comparison with a German mixed population, we used the KORA F3 sample ($N = 1644$) and corresponding genotypes from 500 k Affymetrix SNP chips. Observed differences between regions of Germany are typically an order of magnitude lower than differences observed between Sorbs and KORA [53]. Publicly available European-American HapMap samples were also included in the analysis.

A major goal of our study was to distinguish effects of genetic isolation from simple over-sampling of families in the Sorbs. Since most of the population genetic measures used to compare populations assume

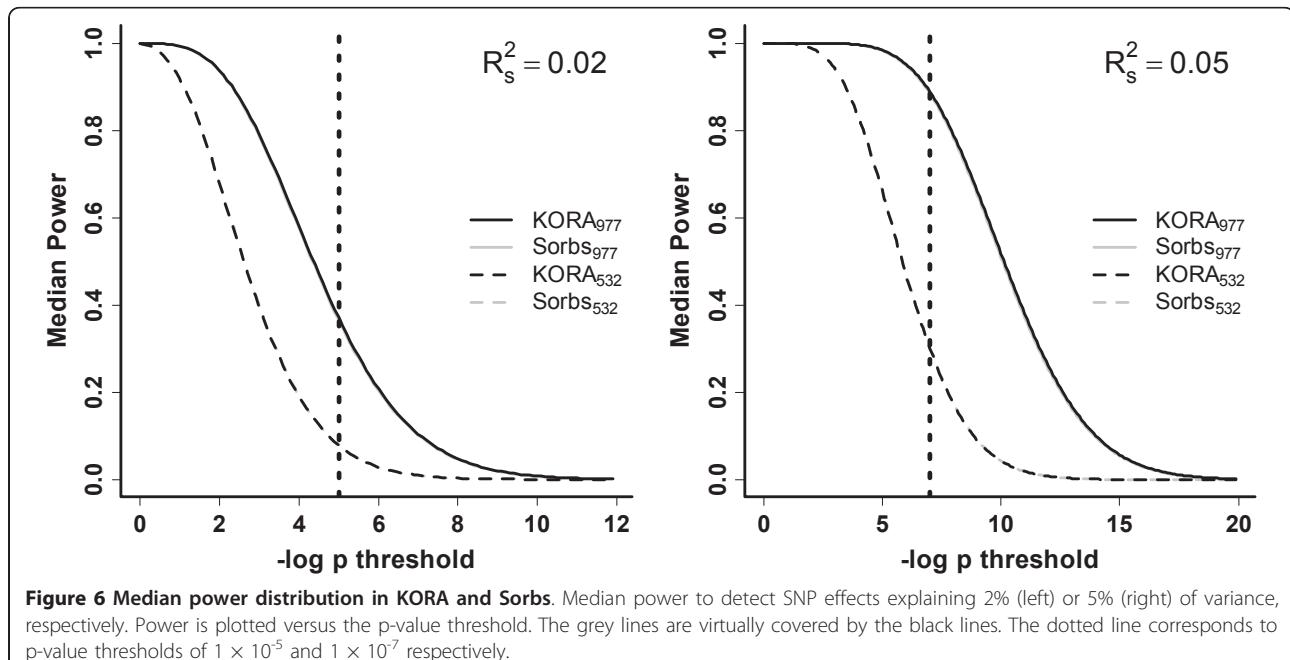


Figure 6 Median power distribution in KORA and Sorbs. Median power to detect SNP effects explaining 2% (left) or 5% (right) of variance, respectively. Power is plotted versus the p-value threshold. The grey lines are virtually covered by the black lines. The dotted line corresponds to p-value thresholds of 1×10^{-5} and 1×10^{-7} respectively.

Table 3 Quartiles of power distribution assuming uncorrelated phenotypes

Explained variance	p-value threshold	Population	1st Quartile	Median	3rd Quartile
2%	1×10^{-5}	KORA ₉₇₇	6.78	37.02	49.19
2%	1×10^{-5}	Sorbs ₉₇₇	6.31	36.51	49.34
2%	1×10^{-5}	KORA ₅₃₂	1.15	7.85	11.52
2%	1×10^{-5}	Sorbs ₅₃₂	1.13	7.88	11.65
5%	1×10^{-7}	KORA ₉₇₇	25.01	88.8	95.81
5%	1×10^{-7}	Sorbs ₉₇₇	23.14	88.37	95.87
5%	1×10^{-7}	KORA ₅₃₂	2.73	30.07	43.41
5%	1×10^{-7}	Sorbs ₅₃₂	2.66	30.17	43.85

Quartiles of the power distribution in percent for an explained variance of 2% with a p-value threshold of 1×10^{-5} and of 5% with a p-value threshold of 1×10^{-7} , respectively.

independence of individuals, over-sampling of families in certain samples may introduce a source of bias which is difficult to control. Indeed, we discovered a large number of closely related individuals within the Sorbs sample. Therefore, we repeated all analyses for a sub-group of Sorbs for which all relationships with relatedness estimates greater than 0.2 were removed. This does not completely resolve the problem of increased relatedness within the Sorbs sample but provides a trend for potential biases introduced by over-sampling of families. Indeed, such biases could be detected in our data but it is not substantial at least for the population genetic measures studied.

Since relatedness cannot be completely removed from the samples, a cut-off of 0.2 for the relatedness estimate seems to be feasible to study the effect of relatedness and to keep the sample size at an acceptable level. We also studied a cut-off of 0.1 reducing the sample size to N = 414. Results can be found in Additional file 6. Although tending slightly towards zero, results are essentially the same as those obtained for the cut-off of 0.2.

For some analyses such as determination of rare SNPs and LD it is known that sample size can introduce bias [39,44,54]. Therefore, for most comparisons we used randomly drawn subsamples of KORA which are of the same size as the Sorbs samples.

PCA is a proven means to detect even very small genetic differences between populations with high power. For European populations, it was demonstrated that the first two appropriately scaled principal components can map individuals to their geographic origin on the European continent with high precision, when all four grandparents are from the same location [14]. Our PCA results showed clear distances between KORA, Sorbs, and individuals from Tuscany. Using individuals from KORA and Tuscany to roughly orient the PCA graph on a map of Europe, Sorbs are positioned towards the East. KORA individuals are very close to the CEU HapMap population, while the distance to Tuscan/TSI individuals is much larger.

We conclude that the Slavonic origin of the Sorbs is still clearly genetically detectable. The analysis revealed that there is a west to east sequence of the clusters of KORA individuals born in Germany, KORA individuals born in Poland or Czech Republic, Half-Sorbs, and finally, Full-Sorbs. Although birthplace is not a stringent indicator of ethnicity, it is a commonly used surrogate in genetic epidemiologic studies if more detailed information cannot be ascertained. On the other hand, most of the KORA individuals born in Poland or Czech Republic are descendants from German minorities of these countries. Hence, on the basis of our data we cannot conclude that the Sorbs are genetically more distant

Table 4 Quartiles of power distribution assuming correlated phenotypes

Explained variance	p-value threshold	Population	1st Quartile	Median	3rd Quartile
2%	1×10^{-5}	KORA ₉₇₇	6.7	37.1	48.4
2%	1×10^{-5}	Sorbs ₉₇₇	10.08	38.95	48.9
2%	1×10^{-5}	KORA ₅₃₂	1.2	7.8	11.6
2%	1×10^{-5}	Sorbs ₅₃₂	1.3	8.2	11.9
5%	1×10^{-7}	KORA ₉₇₇	24.78	88.3	95.12
5%	1×10^{-7}	Sorbs ₉₇₇	27.3	83.6	91.8
5%	1×10^{-7}	KORA ₅₃₂	2.73	29.9	42.9
5%	1×10^{-7}	Sorbs ₅₃₂	2.9	30.4	43.5

Quartiles of the power distribution in percent for an explained variance of 2% with a p-value threshold of 1×10^{-5} and of 5% with a p-value threshold of 1×10^{-7} , respectively. A heritability of 100% is assumed.

from Germany than a random sample from Poland or Czech Republic. Half-Sorbs can be assumed to be closer to the German population than Full-Sorbs due to mating with German neighbours. This is clearly reflected by the localization of Half-Sorbs between KORA individuals and Full-Sorbs. There is a trend that the Sorbs are closer to the KORA individuals born in Poland than to the KORA individuals born in Czech Republic which is in agreement with a recently stated hypothesis that the Sorbs are genetically closer to Polish than to Czech [1].

Since it has been suggested that genetic diversity is lower in isolated populations [6], we analysed the number of rare SNPs. Indeed, we found a higher number of rare SNPs in the Sorbs sample compared to the KORA sample. Although significant, the difference is small in size.

The F_{ST} statistics between KORA and Sorbs were an order of magnitude higher than usually observed between different regions of Germany [53]. Thus, variance between KORA and Sorbs is much higher than expected for different regions in Germany. Surprisingly, the F_{IS} statistic was positive for KORA but negative for Sorbs. Such a phenomenon has also been observed for other isolated populations, suggesting that there may be signs of recent isolation breaking in the Sorbs [44]. Another indicator of isolation breaking is the relatively high number of Half-Sorbs ($N = 160$) in the present sample, i.e. subjects who claim to have less than four Sorbian grandparents. It should be remarked that the F_{IS} statistic is a population based measure rather than an individual based measure of inbreeding studied in [1].

ROH analysis was proposed to detect signs of isolation by estimation of inbreeding [18]. Despite the simplicity of this concept, calculation of ROH depends on many variable parameter settings such as SNP density or allowed numbers of missings or heterozygous markers, which heavily influence the results. Parameter settings are extensively discussed in McQuillan et al [18]. For our analysis, we used the default settings of PLINK except for two parameters: The threshold for homozygous segments was 500 kb (PLINK default is 1000 kb) and the splitting of homozygous segments can occur if two neighbouring SNPs are 100 kb apart (PLINK default is 1000 kb). Hence, we used the same settings as in McQuillan et al. except for the minimum number of contiguous homozygous SNPs constituting a ROH, for which we kept the PLINK default ($N = 100$). The results of ROH analysis also depend on allelic frequencies of populations and SNP-selections used by different genotyping technologies. Since McQuillan et al. [18] used a different genotyping platform (Illumina Infinium HumanHap300v2), the latter modification was necessary to obtain similar results.

We found that Sorbs have enriched ROHs of intermediate length (between 2.5 Mb and 5 Mb) compared to KORA, CEU, and TSI. This effect is much less pronounced for longer ROHs. Accordingly, the coverage of the genome by ROHs is higher in the Sorbian population. Following the argumentation of McQuillan et al., we conclude that there is a lack of recent parental relatedness in the Sorbs (no differences for long range ROHs) but that there are signs of ancient parental relatedness or the existence of autozygous segments of older pedigree structures (differences for ROHs of intermediate range). The lack of direct parental relatedness is in accordance with our estimates of F_{IS} .

Furthermore, we compared the LD structure of chromosome 22 between the KORA and the Sorbs population. We used the newly proposed LD measure η_1 for the comparison of KORA and Sorbs. In contrast to the more popular measures r and D' , the measure η_1 is independent of allelic frequencies [42]. In our opinion, this property is desirable when comparing LD structure between populations of potentially differing allelic frequencies. However, the results obtained by the three measures are very similar (data not shown).

An expected small upward bias caused by smaller sample size in KORA₅₃₂ compared to KORA₉₇₇ could be clearly detected. In contrast, the results for Sorbs₉₇₇ and Sorbs₅₃₂ are virtually identical. We conclude that the expected upward bias of the reduced Sorbs₅₃₂ sample is nullified by the elimination of relationships. This interpretation is supported by the fact that a random sample of $N = 532$ individuals from Sorbs₉₇₇ resulted in the same sample size bias as observed for KORA (data not shown). That is, LD is upwardly biased by the relatedness structure in the Sorbs. Nevertheless, even if relationships are eliminated to a reasonable degree (first and second degree relationships), Sorbs show generally higher LD at longer distances than is observed in KORA. It has been already shown in the literature that LD excess at longer ranges is a characteristic of isolated populations [5,9-11]. However, the effect is moderate in size which is also in agreement with several other populations considered as isolated [44,55-57].

Since LD structure directly influences the coverage of a SNP technology, and with it, the power of genome-wide association studies, we performed power analyses in the Sorbs and KORA samples. For this purpose, we defined a fixed genetic effect of an arbitrary SNP at chromosome 22. Explained variance was used as a measure of effect in order to adjust for differences in allelic frequencies. For this SNP, we analysed the best proxy SNP available on chromosome 22 in order to mimic a situation in which an unobserved causative variant is detected via a marker in LD. We derived an analytical

formula for our model for the case of negligible heritability for which individuals can be considered as independent. This formula also applies to situations where correction for relatedness effects has been performed, for instance with a GRAMMAR approach [17]. Power was calculated for all SNPs on chromosome 22 and the resulting distribution was compared between the Sorbs and KORA samples with and without relatives. No differences regarding power were detected. We conclude that there is no gain in power due to higher LD in the Sorbs.

Since relatedness structure is often neglected in genetic association studies, we also analysed the influence of present relatedness structure on the power of an uncorrected analysis. This analysis is done via simulations of a linear mixed model comprising a fixed effect of a SNP and random polygenetic and non-genetic effects. We showed that the variance of the β -estimator is inflated under relatedness and high heritability. This results in a gain in power for higher p-value thresholds and a loss of power for lower p-value thresholds in the Sorbs₉₇₇, irrespective of the size of the genetic effect considered. The explanation is that normal distributions with different variances are overlapping.

We conclude that relatedness in the Sorbs₉₇₇ sample influences the power of uncorrected genetic association studies. Influence of relatedness on power is highest under maximum heritability of the phenotype. However, directions of power differences depend on the size of the genetic effect in combination with the significance threshold chosen.

In our simulations we did not observe a scenario resulting in a clear power benefit in the Sorbs₉₇₇ sample. However, this does not rule out that there might be a higher power in the Sorbs due to increased effect sizes caused, e.g., by higher environmental homogeneity or lower number of causative variants [7,8].

Conclusions

We could show that there are signs of genetic isolation within the Sorbs which cannot be explained by oversampling of relatives. The effects are moderate in size. The Slavonic origin of the Sorbs is still genetically detectable. Although there is higher LD in the Sorbs, the difference to KORA is small. Power analysis showed that a clear advantage of the Sorbs for genome-wide association studies with respect to coverage cannot be expected.

The significant amount of cryptic relatedness in the Sorbs sample results in inflated variances of β -estimators which should be considered in genetic association analyses.

Additional material

Additional file 1: Workflow of data pre-processing. The workflow of data pre-processing is presented. We start with the autosomal SNP data of four different populations (KORA, Sorbs, HapMap CEU, HapMap TSI). Numbers of remaining markers at each step of pre-processing are presented in bold.

Additional file 2: Derivation of the formula for $\hat{\beta}_2$.

Additional file 3: Comparisons of power for Sorbs₉₇₇ for minimal and maximal heritability of phenotypes. Simulation results of the

power for minimal ($\sigma_g^2 = \text{Var}(s_1)(\frac{R_h^2}{R_s^2} - 1)$) and maximal (100%) heritability. For the minimal heritability, we present the results of our analytical formula. The values presented in Tables 3 and 4 are displayed in bold.

Additional file 4: Variance inflation under relatedness. Comparison of the theoretical variance of the β_1 -estimator assuming uncorrelated phenotypes (analytical formula $\text{var}(\beta_1) = \frac{1}{N-1}(\frac{1}{R_s^2} - 1)$) with the averaged variances over all SNPs of chromosome 22 under a heritability of 100% assuming correlated phenotypes. The standard error of this estimate and the inflation factor are also provided. Sorbs₉₇₇ are presented in bold due to high inflation of variances of β_1 -estimates.

Additional file 5: Simulation results for power under assumption of correlated phenotypes. Heritability was modified between R_s^2 and 100%. Explained variances of the SNP are 2% or 5% with corresponding p-value thresholds of 10^{-5} and 10^{-7} , respectively. All simulations were performed for KORA₉₇₇, Sorbs₉₇₇, KORA₅₃₂, and Sorbs₅₃₂. Power distribution is derived using the results of all SNPs of Chromosome 22.

Additional file 6: Additional inbreeding and co-ancestry coefficients. Estimates and standard errors (SE) of inbreeding coefficients F_{IS} and co-ancestry coefficients F_{ST} for KORA and Sorbs and different levels of relatedness: without filtering for relatedness (KORA₉₇₇, Sorbs₉₇₇), filtering for relatedness > 0.2 (KORA₅₃₂, Sorbs₅₃₂), filtering for relatedness > 0.1 (KORA₄₁₄, Sorbs₄₁₄). Indices refer to resulting numbers of cases.

Acknowledgements

We thank Knut Krohn and Beate Enigh for conducting microarray experiments of the Sorbs sample at the IZKF Leipzig at the Faculty of Medicine of the University of Leipzig (Projekt Z03). We gratefully acknowledge the contributions of P. Lichtner, G. Eckstein, Guido Fischer, T. Strom and all other members of the Helmholtz Centre Munich genotyping staff in generating the SNP dataset as well as the contribution of all members of field staffs who were involved in the planning and conduct of the MONICA/KORA Augsburg studies. The KORA group consists of H.E. Wichmann (speaker), A. Peters, C. Meisinger, T. Illig, R. Holle, J. John and their co-workers who are responsible for the design and conduct of the KORA studies.

We thank Maelle Salmon for helping with data quality control. We thank Karsten Krug and Lars Thielecke for their technical assistance. Finally, we express our appreciation to all participants of the Sorb and the KORA study for donating their blood and time.

Funding

The KORA research platform (KORA: Cooperative Research in the Region of Augsburg) and the MONICA Augsburg studies (Monitoring trends and determinants on cardiovascular diseases) were initiated and financed by the Helmholtz Zentrum München-National Research Center for Environmental Health, which is funded by the German Federal Ministry of Education, Science, Research and Technology and by the State of Bavaria. Part of this work was financed by the German National Genome Research Network (NGFN). Our research was supported within the Munich Center of Health Sciences (MC Health) as part of LMUInnovativ. AT, PK and MStu received financial support from the German Research Council (KFO-152), IZKF (B27), and the German Diabetes Association. MSto is funded by the Max Planck

Society. AG and PA are funded by the German Federal Ministry for Education and Research (01KN0702). AG, PA, NRR, and MSch were funded by the Leipzig Interdisciplinary Research Cluster of Genetic Factors, Clinical Phenotypes, and Environment (LIFE Center, Universität Leipzig). LIFE is funded by means of the European Union, by the European Regional Development Fund (ERDF), the European Social Fund (ESF), and by means of the Free State of Saxony within the framework of its excellence initiative.

Author details

¹Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Haertelstrasse 16-18, 04107 Leipzig, Germany. ²LIFE Center (Leipzig Interdisciplinary Research Cluster of Genetic Factors, Phenotypes and Environment), University of Leipzig, Philipp-Rosenthal Strasse 27, 04103 Leipzig, Germany. ³Department of Medicine, University of Leipzig, Liebigstrasse 18, 04103 Leipzig, Germany. ⁴IFB Adiposity Diseases, University of Leipzig, Stephanstrasse 9c, 04103 Leipzig, Germany. ⁵Interdisciplinary Center for Clinical Research, University of Leipzig, Liebigstrasse 21, 04103 Leipzig, Germany. ⁶Dept Eco & Evo Biol, Interdepartmental Program in Bioinformatics, University of California, 621 Charles E. Young Dr South, Box 951606, Los Angeles, Los Angeles, CA 90095-1606 USA. ⁷Center for Society and Genetics, University of California, 1323 Rolfe Hall, Box 957221, Los Angeles, Los Angeles, CA 90095-7221, USA. ⁸Dept of History, University of California, 6265 Bunche Hall, Box 951473, Los Angeles, Los Angeles, CA 90095-1473, USA. ⁹Helmholtz Centre Munich, German Research Center for Environmental Health, Institute of Epidemiology, Ingolstaedter Landstraße 1, 85764 Neuherberg, Germany. ¹⁰Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany. ¹¹Institute of Medical Informatics, Biometry and Epidemiology, Chair of Epidemiology, Ludwig-Maximilians-University, Marchioninistraße 15, 81377 Munich, Germany. ¹²Klinikum Grosshadern, Ludwig Maximilians University, Marchioninistraße 15, 81377 Munich, Germany.

Authors' contributions

Design of the Study: MSch. Design of the Sorbs study and data collection: AT, PK, MStu. Design of the KORA data collection: CG, IR, HW. Data analysis: AG, NRR, MSch. Writing: AG, MSch. Contribution to writing and discussion: KRV, PA, ML, MStu, AT, PK, MStu, JN.

All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 14 March 2011 Accepted: 28 July 2011

Published: 28 July 2011

References

- Veeramah KR, Tonjes A, Kovacs P, Gross A, Wegmann D, Geary P, Gasperikova D, Klimes I, Scholz M, Novembre J, et al: **Genetic variation in the Sorbs of eastern Germany in the context of broader European genetic diversity.** *European Journal of Human Genetics* 2011.
- Abbott A: **Manhattan versus Reykjavik.** *Nature* 2000, **406**(6794):340-342.
- Eaves IA, Merriman TR, Barber RA, Nutland S, Tuomilehto-Wolf E, Tuomilehto J, Cucca F, Todd JA: **The genetically isolated populations of Finland and sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes.** *Nat Genet* 2000, **25**(3):320-323.
- Taillon-Miller P, Bauer-Sardina I, Saccone NL, Putzel J, Laitinen T, Cao A, Kere J, Pilis G, Rice JP, Kwok PY: **Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28.** *Nat Genet* 2000, **25**(3):324-328.
- Shifman S, Darvasi A: **The value of isolated populations.** *Nat Genet* 2001, **28**(4):309-310.
- Kristiansson K, Naukkarinen J, Peltonen L: **Isolated populations and complex disease gene identification.** *Genome Biol* 2008, **9**(8):109.
- Sheffield VC, Stone EM, Carmi R: **Use of isolated inbred human populations for identification of disease genes.** *Trends Genet* 1998, **14**(10):391-396.
- Arcos-Burgos M, Muenke M: **Genetics of population isolates.** *Clin Genet* 2002, **61**(4):233-247.
- Tenesa A, Wright AF, Knott SA, Carothers AD, Hayward C, Angius A, Persico I, Maestrale G, Hastie ND, Pirastu M, et al: **Extent of linkage disequilibrium in a Sardinian sub-isolate: sampling and methodological considerations.** *Hum Mol Genet* 2004, **13**(1):25-33.
- Service S, DeYoung J, Karayiorgou M, Roos JL, Pretorius H, Bedoya G, Ospina J, Ruiz-Linares A, Macedo A, Palha JA, et al: **Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies.** *Nat Genet* 2006, **38**(5):556-560.
- Angius A, Hyland FC, Persico I, Pirastu N, Woodage T, Pirastu M, De la Vega FM: **Patterns of linkage disequilibrium between SNPs in a Sardinian population isolate and the selection of markers for association studies.** *Hum Hered* 2008, **65**(1):9-22.
- Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW: **Genetic structure of human populations.** *Science* 2002, **298**(5602):2381-2385.
- Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degen JH, Wang K, Guerreiro R, et al: **Genotype, haplotype and copy-number variation in worldwide human populations.** *Nature* 2008, **451**(7181):998-1003.
- Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, et al: **Genes mirror geography within Europe.** *Nature* 2008, **456**(7218):98-101.
- Lopez Herraez D, Bauchet M, Tang K, Theunert C, Pugach I, Li J, Nandineni MR, Gross A, Scholz M, Stoneking M: **Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs.** *PLoS One* 2009, **4**(11):e7888.
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies.** *Nat Genet* 2006, **38**(8):904-909.
- Amin N, van Duijn CM, Aulchenko YS: **A genomic background based method for association analysis in related individuals.** *PLoS One* 2007, **2**(12):e1274.
- McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauk L, Smolej-Narancic N, Janicijevic B, Polasek O, Tenesa A, et al: **Runs of homozygosity in European populations.** *Am J Hum Genet* 2008, **83**(3):359-372.
- Peltonen L, Jalanko A, Varilo T: **Molecular genetics of the Finnish disease heritage.** *Hum Mol Genet* 1999, **8**(10):1913-1923.
- Peltonen L: **Positional cloning of disease genes: advantages of genetic isolates.** *Hum Hered* 2000, **50**(1):66-75.
- Weir BS: **Genetic Data Analysis II.** Sunderland, MA: Sinauer Associates, Inc; 1996.
- Patterson N, Price AL, Reich D: **Population structure and eigenanalysis.** *PLoS Genet* 2006, **2**(12):e190.
- Choi Y, Wijsman EM, Weir BS: **Case-control association testing in the presence of unknown relationships.** *Genet Epidemiol* 2009, **33**(8):668-678.
- Zhang F, Deng HW: **Correcting for cryptic relatedness in population-based association studies of continuous traits.** *Hum Hered* 2010, **69**(1):28-33.
- Thornton T, McPeek MS: **ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure.** *Am J Hum Genet* 2010, **86**(2):172-184.
- Krawczak M, Lu TT, Willuweit S, Roewer L: **Genetic diversity in the German population.** *Handbook of Human Molecular Evolution* John Wiley & Sons; 2008.
- Kottgen A, Pattaro C, Boger CA, Fuchsberger C, Olden M, Glazer NL, Parsa A, Gao X, Yang Q, Smith AV, et al: **New loci associated with kidney function and chronic kidney disease.** *Nat Genet* 2010.
- Tonjes A, Koriath M, Schleinitz D, Dietrich K, Bottcher Y, Rayner NW, Almgren P, Enigk B, Richter O, Rohm S, et al: **Genetic variation in GPR133 is associated with height: genome wide association study in the self-contained population of Sorbs.** *Hum Mol Genet* 2009, **18**(23):4662-4668.
- Wichmann HE, Gieger C, Illig T: **KORA-gen-resource for population genetics, controls and a broad spectrum of disease phenotypes.** *Gesundheitswesen* 2005, **67**(Suppl 1):S26-30.
- Tonjes A, Zeggini E, Kovacs P, Bottcher Y, Schleinitz D, Dietrich K, Morris AP, Enigk B, Rayner NW, Koriath M, et al: **Association of FTO variants with BMI and fat mass in the self-contained population of Sorbs in Germany.** *Eur J Hum Genet* 2010, **18**(1):104-110.
- Holle R, Happich M, Lowel H, Wichmann HE: **KORA-a research platform for population based health research.** *Gesundheitswesen* 2005, **67**(Suppl 1):S19-25.

32. Doring A, Gieger C, Mehta D, Gohlke H, Prokisch H, Coassini S, Fischer G, Henke K, Klopp N, Kronenberg F, et al: **SLC2A9 influences uric acid concentrations with pronounced sex-specific effects.** *Nat Genet* 2008, **40**(4):430-436.
33. Pemberton TJ, Wang C, Li JZ, Rosenberg NA: **Inference of unexpected genetic relatedness among individuals in HapMap Phase III.** *Am J Hum Genet* 2010, **87**(4):457-464.
34. Li Y, Willer C, Sanna S, Abecasis G: **Genotype imputation.** *Annu Rev Genomics Hum Genet* 2009, **10**:387-406.
35. Troendle JF, Yu KF: **A note on testing the Hardy-Weinberg law across strata.** *Ann Hum Genet* 1994, **58**(Pt 4):397-402.
36. Wang J: **An estimator for pairwise relatedness using molecular markers.** *Genetics* 2002, **160**(3):1203-1215.
37. Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Balascakova M, Bertranpetti I, Bindoff LA, Comas D, et al: **Correlation between genetic and geographic structure in Europe.** *Curr Biol* 2008, **18**(16):1241-1248.
38. McVean G: **A genealogical interpretation of principal components analysis.** *PLoS Genet* 2009, **5**(10):e1000686.
39. Scholz M, Hasencler D: **Comparison of Estimators for Measures of Linkage Disequilibrium.** *The International Journal of Biostatistics* 2010, **6**(1).
40. Hill WG, Robertson A: **Linkage Disequilibrium in Finite Populations.** *Theoretical and Applied Genetics* 1968, **38**:226-231.
41. Lewontin RC: **The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models.** *Genetics* 1964, **49**(1):49-67.
42. A Canonical Measure of Allelic Association. [http://arxiv.org/PS_cache/arxiv/pdf/0903/0903.3886v1.pdf].
43. Edwards AWF: **The Measure of Association in a 2 × 2 Table.** *Journal of the Royal Statistical Society, Series A* 1963, **126**:108-114.
44. Olshen AB, Gold B, Lohmueller KE, Struewing JP, Satagopan J, Stefanov SA, Eskin E, Kirchhoff T, Lautenberger JA, Klein RJ, et al: **Analysis of genetic variation in Ashkenazi Jews by high density SNP genotyping.** *BMC Genet* 2008, **9**:14.
45. Cleveland WS: **Robust locally weighted regression and smoothing scatterplots.** *Journal of the American Statistical Association* 1979, **74**:829-836.
46. International HapMap Project. [<http://hapmap.ncbi.nlm.nih.gov/>].
47. EIGENSOFT Package. [<http://genepath.med.harvard.edu/~reich/Software.htm>].
48. PLINK Package. [<http://pngu.mgh.harvard.edu/purcell/plink/>].
49. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81**(3):559-575.
50. The R Project. [<http://www.r-project.org/>].
51. R: A Language and Environment for Statistical Computing. [<http://www.R-project.org>].
52. Rodig H, Grum M, Grimmecke HD: **Population study and evaluation of 20 Y-chromosome STR loci in Germans.** *Int J Legal Med* 2007, **121**(1):24-27.
53. Steffens M, Lamina C, Illig T, Bettecken T, Vogler R, Entz P, Suk EK, Toliat MR, Klopp N, Caliebe A, et al: **SNP-based analysis of genetic substructure in the German population.** *Hum Hered* 2006, **62**(1):20-29.
54. Chen Y, Lin CHL, Sabatti C: **Volume Measures for Linkage Disequilibrium.** *BMC Genetics* 2006, **7**(54).
55. Kruglyak L: **Genetic isolates: separate but equal?** *Proc Natl Acad Sci USA* 1999, **96**(4):1170-1172.
56. Shifman S, Kuypers J, Kokoris M, Yakir B, Darvasi A: **Linkage disequilibrium patterns of the human genome across populations.** *Hum Mol Genet* 2003, **12**(7):771-776.
57. Bosch E, Laayouni H, Morcillo-Suarez C, Casals F, Moreno-Estrada A, Ferrer-Admetlla A, Gardner M, Rosa A, Navarro A, Comas D, et al: **Decay of linkage disequilibrium within genes across HGDP-CEPH human samples: most population isolates do not show increased LD.** *BMC Genomics* 2009, **10**:338.

doi:10.1186/1471-2156-12-67

Cite this article as: Gross et al.: Population-genetic comparison of the Sorbian isolate population in Germany with the German KORA population using genome-wide SNP arrays. *BMC Genetics* 2011 **12**:67.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit



3 Einfluß von Verwandtschaft auf Assoziationsanalysen

Im letzten Kapitel wurde bei einer empirischen Power-Analyse beobachtet, daß bei Anwendung des einfachen linearen Modells die Verwandtschaftsstruktur der Sorben zu einer Varianzvergrößerung des Effektschätzers führte und dadurch die Power des Tests in komplexer Weise beeinflußt wurde. Offenbar hängt es von der Stärke des genetischen Effekts und vom Signifikanzniveau ab, ob die Power größer oder kleiner wird. Zudem ist aus empirischen Studien [30, 31] bekannt, daß die Verwandtschaftsstruktur der Studienpopulation und die Heritabilität des Phänotyps den Fehler erster Art eines unkorrigierten Tests vergrößern. Genomic control [34] wird häufig dazu verwendet, eine Inflation der Teststatistik durch Verwandtschaft zu korrigieren [17, 35]. Jedoch führt genomic control auch zu einer Power-Reduktion [17, 34]. Weiterhin ist für Phänotypen wie die Körpergröße mit einer Heritabilität von 80% [15] klar, daß dadurch die Grundannahme unabhängiger Beobachtungen des einfachen linearen Modells verletzt wird. Für Phänotypen wie der Chemerin-Serumspiegel mit einer Heritabilität von 16% [66] ist aber nicht offensichtlich, was die Verletzung der Grundannahme für den Test bedeutet. In diesem Kapitel soll deshalb analytisch gezeigt werden, welchen Einfluß die Verwandtschaftsstruktur einer Studienpopulation und die Heritabilität eines Phänotyps auf die Varianzinflation des Effektschätzers haben, welche Konsequenzen daraus für den Fehler erster Art und die Power des Tests folgen und inwieweit sich genomic control als Korrekturmethode eignet. Die Ergebnisse sind in [2] publiziert und werden hier zusammengefaßt.

3.1 Beschreibung der Studien und genetischen Daten

Um den Einfluß unterschiedlicher Verwandtschaftsstrukturen auf SNP-Assoziationsanalysen zu untersuchen, werden sowohl Studien mit „echten“ Genotypen als auch synthetische Familienstudien mit simulierten Genotypen analysiert. HapMap CEU Trio-Familien (Public Release 28, NCBI build 36 [67]) wurden aufgrund der einfachen Verwandtschaftsstruktur ausgewählt. Die QC von SNPs und Probanden ist in *Additional file 4* der Publikation beschrieben. Nach der QC standen 1.020.215 SNPs für 129 Probanden aus 43 Trio-Familien zur Verfügung. Weiterhin wurde die bereits beschriebene Sorbenstudie [8] analysiert. Die QC für SNPs und Probanden wurde durchgeführt wie in [1] und resultierte in 424.476 SNPs und 977 Probanden. Zuletzt wurden Genotypen für synthetische Familienstudien simuliert, die sich sowohl in der Stärke der Verwandtschaftsstruktur als auch der Fallzahl unterscheiden. Für jede der synthetischen Familienstudien wurden 110.000 SNPs mit Beta-verteilten (Parameter $a = 0,5$ und $b = 0,5$) Allelfrequenzen mit dem R-Skript in *Additional file 1* erzeugt.

3.2 Varianzinflation

Durch die Verletzung der Grundannahme unabhängiger Beobachtungen wird die Verteilung des Effektschätzers $\hat{\beta}_2$ aus Gl. (1.5) beeinflußt. Es läßt sich analytisch zeigen, daß der Effektschätzer mit $E(\hat{\beta}_2) = b_2$ erwartungstreu ist, was in empirischen Studien [30, 31] beobachtet wurde. Allerdings wird die Varianz

$$V(\hat{\beta}_2) = \frac{\lambda}{1 - R_h^2} V_\beta$$

beeinflußt. Die Varianz hängt dabei vom Inflationsfaktor

$$\lambda = 1 + R_h^2 \frac{\sum_{i=1}^n \sum_{j \neq i=1}^n G_{ij} (\bar{s}^2 - 2\bar{s}s_i + s_i s_j)}{\sum_{i=1}^n (s_i - \bar{s})^2}, \quad (3.1)$$

von der Heritabilität R_h^2 und von der Varianz $V_\beta = \sigma_e^2 / \sum_{i=1}^n (s_i - \bar{s})^2$ ab. Die Herleitung ist in *Additional file 2: Section 2.2* beschrieben. Die Varianz V_β entspricht dabei der aus der Literatur [32] bekannten Varianz für den Effektschätzer im einfachen linearen Modell ohne Verwandtschaft. Im Abschnitt 3.3 wird noch gezeigt, daß auch die empirische Varianz des Effektschätzers durch den Faktor $1/(1 - R_h^2)$ beeinflußt wird, sich dadurch dieser Faktor herauskürzt und somit λ allein für die Verteilung der Teststatistik ausschlaggebend ist.

Erwartete Varianzinflation

Es ist sinnvoll eine Näherung von Gl. (3.1) herzuleiten, die von einem konkreten SNP unabhängig ist. Durch Abschätzung des Erwartungswertes von λ , kann eine Näherung

$$\lambda' = 1 + R_h^2 \frac{\sum_{i=1}^n \sum_{j \neq i=1}^n G_{ij}^2 - \frac{2}{n} \sum_{i=1}^n \left(\sum_{j \neq i=1}^n G_{ij} \right)^2}{n - 1} \quad (3.2)$$

analytisch bestimmt werden (*Additional file 2: Section 3.2*). Diese Näherungsformel stellt ein wichtiges Ergebnis der Publikation dar. Die Näherung ist korrekt, so lange die Fallzahl n groß genug ($n > 100$) und die mittlere Verwandtschaft $\bar{G} = \sum_{i=1}^n \sum_{j \neq i=1}^n G_{ij} / (n - 1)$ klein genug ($\bar{G} < 0,01$) ist. Die erwartete Varianzinflation λ' hängt nur von der Heritabilität R_h^2 und der Verwandtschaftsstruktur \mathbf{G} ab. λ' ist von der Allelfrequenz unabhängig, was die Beobachtungen von [16, 30] bestätigt. Für synthetische Familienstudien wurden weitere Näherungsformeln hergeleitet (*Additional file 2: Section 3.4*), bei denen anstelle der Verwandtschaftsmatrix nur die Anzahl der Familien (Väter), Anzahl der Mütter je Familie und Anzahl der Kinder je Mutter in die Gleichungen eingehen.

Für Gl. (3.2) sind nützliche Transformationen möglich (*Additional file 2: Section 3.3*), um beispielsweise die Frage beantworten zu können, welche Varianzinflation λ'_t für eine andere Heritabilität R_t^2 unter Beibehaltung der Verwandtschaftsstruktur erwartet wird. Hierfür gilt

$$\lambda'_t = 1 + (\lambda' - 1) \frac{R_t^2}{R_h^2}. \quad (3.3)$$

Beispiele

Um einen Eindruck davon zu bekommen, welchen Einfluß die Verwandtschaftsstruktur auf die Varianzinflation hat, wurden verschiedene synthetische Familienstudien mit variabler Anzahl von Müttern je Familie und Anzahl von Kindern je Mutter bei gleicher Studienfallzahl verglichen. In *Figure 1* ist die erwartete Varianzinflation λ' für eine Heritabilität von $R_h^2 = 0,9$ dargestellt. Ein Beispiel aus der Grafik ist $\lambda' = 1,3$ für eine Studie mit Trio-Familien, die die schwächste Verwandtschaftsstruktur besitzen.

In einer weiteren Analyse wurde für die HapMap Trios, die Sorben und einige synthetische Familienstudien die Varianzinflation wie in Gl. (3.1) über SNPs gemittelt und die erwartete Varianzinflation wie in Gl. (3.2) bestimmt. Die Ergebnisse für eine Heritabilität von $R_h^2 = 0,9$ sind in *Table 1* dargestellt. Beispiele für die über SNPs gemittelte Varianzinflation sind die HapMap Trios mit $\bar{\lambda} = 1,29$ oder die Sorben mit $\bar{\lambda} = 1,41$. Die mittlere Verwandtschaft \bar{G} ist für alle betrachteten Studien klein ($\bar{G} < 0,01$). Jedoch ist die erwartete Varianzinflation λ' für alle Studien in *Table 1* geringfügig höher als λ aufgrund der getroffenen Annahmen bei der Näherung (*Additional file 2: Section 3.2*). Schränkt man die Bestimmung der Varianzinflation λ auf SNPs mit größerer Allelfrequenz ($MAF > 10\%$) ein, verbessert sich die Übereinstimmung mit der erwarteten Varianzinflation λ' . Das bedeutet, daß die Näherung bei kleinen Allelfrequenzen vom Erwartungswert abweicht.

Nimmt man noch das Beispiel für den Chemerin-Serumspiegel aus der Einleitung des Kapitels, würde das für die Sorben nach Gl. (3.3) eine erwartete Varianzinflation von umgerechnet $\lambda'_t = 1,08$ bei einer Heritabilität von $R_t^2 = 0,16$ ergeben. Diese Inflation ist größer als die von [16] empfohlene Schwelle von 1,05, ab der sich die Varianzinflation merkbar auf einen statistischen Test auswirkt und nicht ignoriert werden darf.

3.3 Hypothesentests

Verteilungen der Teststatistik

Um zu testen, ob es einen Zusammenhang zwischen Phänotyp und SNP gibt, wird die Teststatistik T aus Gl. (1.7) analysiert. Möchte man für diesen Test den Fehler erster Art und die Power bestimmen, muß zunächst die Verteilung von T unter Vorliegen der Null- beziehungsweise Alternativhypothese hergeleitet werden. Dazu wurden neben dem Erwartungswert und der Varianz des Effektschätzers $E(\hat{\beta}_2)$ beziehungsweise $V(\hat{\beta}_2)$ auch der Erwartungswert der empirischen Varianz des Effektschätzers $E(S_\beta^2) \approx V_\beta/(1 - R_h^2)$ analytisch bestimmt (*Additional file 2: Section 4.2*). Zu beachten ist, daß $E(S_\beta^2)$ im Vergleich zu $V_\beta/(1 - R_h^2)$ um einen bestimmten Faktor in Abhängigkeit von der Verwandtschaftsstruktur deflationiert ist, was die empirische Beobachtung von [30] bestätigt. Unter der Nullhypothese folgt für die Verteilung der Teststatistik (*Additional file 2: Section 5.2*) näherungsweise

$$T \sim N(0, \lambda) \quad (3.4)$$

und unter der Alternativhypothese (*Additional file 2: Section 5.3*) näherungsweise

$$T \sim N(\mu, \lambda) \quad (3.5)$$

mit $\mu = \sqrt{(n - 1)R_s^2}$ und erklärter Varianz R_s^2 durch den SNP. Die Verteilungen der Teststatistik T in Gl. (3.4) beziehungsweise Gl. (3.5) sind nur Näherungen durch Schätzen der Varianz des Effektschätzers. Um diese Verteilungsannahmen numerisch zu überprüfen, wurden Simulationen für alle Studien sowohl für die Nullhypothese mit $R_h^2 = 0,9$ als auch für die Alternativhypothese mit $R_h^2 = 0,9$ und $R_s^2 = 0,02$ durchgeführt. Die Ergebnisse für die Nullhypothese sind in *Table 2* und für die Alternativhypothese in *Table 3* präsentiert. Für alle Studien ist der Erwartungswert von T wie erwartet nahe Null für die Nullhypothese und nahe dem theoretischen Wert μ für die Alternativhypothese. Die gemittelte empirische Varianz von T (*Tables 2–3*) ist etwas größer als die vorher bestimmte Varianzinflation (*Table 1*). Dieser Unterschied ist für die Studien mit kleiner Fallzahl wie HapMap am größten. Für die anderen Studien mit größerer Fallzahl ist der Unterschied nicht weiter relevant. Die Deflation der empirischen Varianz des Effektschätzers ist in *Table 2* für alle Studien nahe Eins und damit für diese Analysen nicht bedeutsam.

Fehler erster Art und Power des Tests

Der Fehler erster Art kann durch die Definition in Gl. (1.9) und die Verteilungsannahme in Gl. (3.4) in Abhängigkeit des Signifikanzniveaus berechnet werden. In *Figure 2* ist der Fehler erster Art für verschiedene Varianzinflationen dargestellt. Der Fehler erster Art ist für $\lambda = 1,05$ ähnlich dem Fehler für $\lambda = 1$. Jedoch nimmt der Fehler erster Art in Abhängigkeit von der Varianzinflation stark zu. Sowohl eine stärkere Verwandtschaftsstruktur als auch eine größere Heritabilität führen zu einer größeren Varianzinflation (Gl. (3.2)) und damit zu einem größeren Fehler erster Art. Das erklärt die empirischen Beobachtungen von [30, 31], daß der Fehler erster Art mit stärkerer Verwandtschaft und größerer Heritabilität wächst und die empirische Beobachtung von [33], daß der Fehler erster Art vergrößert ist, wenn die Verwandtschaftsstruktur ignoriert wird.

Die Power kann durch die Definition in Gl. (1.10) und die Verteilungsannahme in Gl. (3.5) in Abhängigkeit des Signifikanzniveaus berechnet werden. Für eine Fallzahl von $n = 1000$ und einer erklärten Varianz von $R_s^2 = 0,02$ ist in *Figure 3a* die Power für verschiedene Varianzinflationen dargestellt. Die Power für $\lambda = 1,05$ ähnelt der Power für $\lambda = 1$, die Power für größere λ unterscheidet sich jedoch deutlich. Interessanterweise schneiden sich alle Powerkurven bei 50% unabhängig von der Varianzinflation λ . Das liegt daran, daß sich die Verteilungen der Teststatistiken allein in ihrer Varianz unterscheiden. Das erklärt auch die empirische Beobachtung aus Kapitel 2.4, nämlich daß die Power bei großen Signifikanzniveaus bei korrelierten Phänotypen ($\lambda > 1$) kleiner ist als bei unkorrelierten Phänotypen ($\lambda = 1$) und sich das Verhalten bei kleinen Signifikanzniveaus umkehrt. In empirischen Studien wie [30, 31, 33] wurde der Einfluß der Verwandtschaftsstruktur auf die Power unterschiedlich beurteilt. Mit *Figure 3a* kann jedoch gezeigt werden, daß es vom Signifikanzniveau und von der Stärke des genetischen Effekts abhängt, ob die Power größer oder kleiner wird.

In Übereinstimmung mit [16] wurde dargelegt, daß bei einer Varianzinflation von $\lambda < 1,05$ der Einfluß auf den Fehler erster Art und die Power des Tests vernachlässigt werden kann. Ist die Varianzinflation größer, eignen sich Methoden für eine SNP-Assoziationsanalyse besser, welche explizit für Verwandtschaft korrigieren, wie beispielsweise gemischte Modelle [17, 30, 33, 68–70]. Diese Methoden sind aber aufgrund ihrer Komplexität mit Vorsicht zu verwenden [71]. Ein Überblick von Korrekturmethoden und Software-Tools ist in [72] zu finden.

Genomic control

Genomic control wird häufig dazu verwendet, eine Inflation der Teststatistik durch Verwandtschaft zu korrigieren. Durch die Korrektur der Teststatistik T um den Faktor $1/\sqrt{\hat{\lambda}}$ aus Gl. (1.11) lässt sich aus der Verteilung in Gl. (3.4) die Verteilung der Teststatistik nach genomic control herleiten. Analytische Details sind in Additional file 2: Section 6 beschrieben. So folgt unter der Nullhypothese näherungsweise

$$T_{\text{gc}} \sim N(0,1).$$

Das bedeutet, daß der Fehler erster Art aus Gl. (1.9) durch genomic control dem Signifikanzniveau entspricht und damit eingehalten wird, was die empirischen Beobachtungen von [17, 34] bestätigt. Unter der Alternativhypothese folgt durch die Korrektur der Teststatistik aus der Verteilung in Gl. (3.5) für die Verteilung der Teststatistik nach genomic control näherungsweise

$$T_{\text{gc}} \sim N\left(\frac{\mu}{\sqrt{\hat{\lambda}}}, 1\right).$$

Die Power kann durch die Definition in Gl. (1.10) und diese Verteilungsannahme in Abhängigkeit des Signifikanzniveaus berechnet werden. Durch $\sqrt{\hat{\lambda}}$ wird der Erwartungswert von T_{gc} verringert und damit wird die Power im Vergleich zur unkorrigierten Teststatistik kleiner. In *Figure 3b* ist die Power des Tests nach genomic control in Abhängigkeit verschiedener Varianzinflationen dargestellt. Im Vergleich zu *Figure 3a* steigt der Power-Verlust mit größer werdender Varianzinflation λ an, so daß genomic control nicht für $\lambda > 1,05$ empfohlen werden kann. Der Power-Verlust durch Varianzinflation wurde bereits in [34] angemerkt und in [17] für große Heritabilität und starke Verwandtschaftsstruktur empirisch beobachtet.

3.4 Weitere Analysen

Der polygene Effekt wurde, wie in Gl. (1.4), mit einer multivariaten Normalverteilung modelliert. Alternativ kann der polygene Effekt auch additiv durch einzelne genetische Marker wie in [27] modelliert werden. Die Simulationen der Verteilung der Teststatistik unter der Null- und Alternativhypothese wurden mit diesem Modell wiederholt und die Ergebnisse sind in *Additional file 10* dargestellt. Bereits die Simulation einer geringen Anzahl genetischer Marker führt zu ähnlichen Ergebnissen wie unter Verwendung der multivariaten Normalverteilung.

Die Verwandtschaftsstruktur wurde mit der Methode aus [18] geschätzt. Diese Methode hat einige Vorteile, wie zum Beispiel eine Korrektur für die Schätzung von Allelfrequenzen, die sonst zu einem Bias führt [18, 35]. Das wird auch durch *Additional file 11: Figure 1* belegt, wo verschiedene Verwandtschaftsschätzer dargestellt sind. Weitere Methoden für die Verwandtschaftsschätzung wie paarweises Kinship [17, 35] oder eine IBS (identical by state)-basierte Schätzung [28, 73] führen zu ähnlichen Ergebnissen für die Varianzinflation und die Verteilung der Teststatistik unter der Null- und Alternativhypothese (*Additional file 11*).

3.5 Zusammenfassung

Verwandtschaften in der Studienpopulation führen zu korrelierten Phänotypen. Es konnte analytisch gezeigt werden, daß sowohl eine stärkere Verwandtschaftsstruktur als auch eine größere Heritabilität des Phänotyps eine Varianzvergrößerung des Effektschätzers und der Teststatistik bewirken. Während der Fehler erster Art mit größerer Varianzinflation steigt, wird die Power des Tests in komplexer Weise beeinflußt. Ob die Power größer oder kleiner wird, hängt von der Stärke des genetischen Effekts und vom Signifikanzniveau des Tests ab. Es können empirische Beobachtungen aus der Literatur erklärt werden, zum Beispiel daß der Erwartungswert des Effektschätzers nicht durch Verwandtschaft beeinflußt wird, die empirische Varianz des Effektschätzers bei Verwandtschaft deflationiert ist oder die Allelfrequenz des SNP nur einen geringen Einfluß auf die Varianzinflation hat. Weiterhin kann genomic control nicht für die Korrektur von Varianzinflation durch Verwandtschaft empfohlen werden. Obwohl der Fehler erster Art durch genomic control eingehalten wird, führt die Methode zu einem starken Power-Verlust in Abhängigkeit der Varianzinflation. Zur Bestimmung der Varianzinflation wurde eine Näherungsformel entwickelt, die nur von der Verwandtschaftsstruktur und der Heritabilität des Phänotyps abhängt. Es läßt sich schließen, daß eine Varianzinflation kleiner als 1,05 keinen relevanten Einfluß auf den statistischen Test hat und die Verwendung des einfachen linearen Modells in diesem Fall angemessen ist. Ist die Varianzinflation größer, müssen Methoden wie beispielsweise gemischte Modelle im Rahmen einer SNP-Assoziationsanalyse verwendet werden, welche explizit die Verwandtschaftsstruktur berücksichtigen.

3.6 Publikation

In diesem Abschnitt befindet sich eine Kopie der Publikation von: A. Gross, A. Tonjes, and M. Scholz. On the impact of relatedness on SNP association analysis. *BMC Genet.*, 18(1):104, Dec 2017. doi:10.1186/s12863-017-0571-x.

METHODOLOGY ARTICLE

Open Access



On the impact of relatedness on SNP association analysis

Arnd Gross^{1,2*}, Anke Tönjes³ and Markus Scholz^{1,2}

Abstract

Background: When testing for SNP (single nucleotide polymorphism) associations in related individuals, observations are not independent. Simple linear regression assuming independent normally distributed residuals results in an increased type I error and the power of the test is also affected in a more complicate manner. Inflation of type I error is often successfully corrected by genomic control. However, this reduces the power of the test when relatedness is of concern. In the present paper, we derive explicit formulae to investigate how heritability and strength of relatedness contribute to variance inflation of the effect estimate of the linear model. Further, we study the consequences of variance inflation on hypothesis testing and compare the results with those of genomic control correction. We apply the developed theory to the publicly available HapMap trio data ($N = 129$), the Sorbs (a self-contained population with $N = 977$ characterised by a cryptic relatedness structure) and synthetic family studies with different sample sizes (ranging from $N = 129$ to $N = 999$) and different degrees of relatedness.

Results: We derive explicit and easily to apply approximation formulae to estimate the impact of relatedness on the variance of the effect estimate of the linear regression model. Variance inflation increases with increasing heritability. Relatedness structure also impacts the degree of variance inflation as shown for example family structures. Variance inflation is smallest for HapMap trios, followed by a synthetic family study corresponding to the trio data but with larger sample size than HapMap. Next strongest inflation is observed for the Sorbs, and finally, for a synthetic family study with a more extreme relatedness structure but with similar sample size as the Sorbs. Type I error increases rapidly with increasing inflation. However, for smaller significance levels, power increases with increasing inflation while the opposite holds for larger significance levels. When genomic control is applied, type I error is preserved while power decreases rapidly with increasing variance inflation.

Conclusions: Stronger relatedness as well as higher heritability result in increased variance of the effect estimate of simple linear regression analysis. While type I error rates are generally inflated, the behaviour of power is more complex since power can be increased or reduced in dependence on relatedness and the heritability of the phenotype. Genomic control cannot be recommended to deal with inflation due to relatedness. Although it preserves type I error, the loss in power can be considerable. We provide a simple formula for estimating variance inflation given the relatedness structure and the heritability of a trait of interest. As a rule of thumb, variance inflation below 1.05 does not require correction and simple linear regression analysis is still appropriate.

Keywords: Heritability, Linear regression, Relatedness, SNP association analysis

*Correspondence: arnd.gross@imise.uni-leipzig.de

¹Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Haertelstrasse 16-18, 04107 Leipzig, Germany

²LIFE - Leipzig Research Center for Civilization Diseases, University of Leipzig, Philipp-Rosenthal-Strasse 27, 04103 Leipzig, Germany

Full list of author information is available at the end of the article

Background

When testing for SNP associations in related individuals, one has to account for the non-independence of observations [1]. An appropriate method is to test for the SNP effect assuming a mixed model $y = b_1 + b_2s + \mathbf{g} + \mathbf{e}$ with phenotypes y , intercept b_1 , effect b_2 , SNP genotypes s , polygenic random effects \mathbf{g} and residuals \mathbf{e} [2–5]. Recently, several extensions of this concept were proposed [6]. However, fitting this mixed model is mathematically challenging as well as computationally expensive when performed within a genome-wide context and for large sample sizes. For this reason, the correlation of phenotypes is often neglected and the standard linear model $y = \beta_1 + \beta_2s + \epsilon$ is used assuming independent normally distributed residuals ϵ .

The impact of relatedness on the correctness of simple linear regression analysis also depends on the heritability of the trait of interest. This is obvious if considering traits of high heritability such as height (80%) [7]. However, we demonstrate in the present paper that even if heritability is relatively small (e.g. circulating serum chemerin with estimated 16% heritability [8]) proper correction is still required if highly related samples are analysed. Otherwise, the type I error of the uncorrected test statistic is inflated [9, 10] and increases further with higher heritability and stronger relatedness [1, 10, 11]. In this context, stronger relatedness means more and stronger related pairs of individuals in the analysis sample. Often, inflation of type I error is corrected by genomic control, a phenomenological approach proposed by Devlin & Roeder [12]. They showed that dependency structures of observations can lead to extra variance compared to the situation of independence. Although genomic control works fine to reduce type I error inflation, it reduces the power in case of higher relatedness and heritability [5]. Assessing the power of the uncorrected test in dependence on the degree of relatedness is difficult. We showed in a simulation study [13] that for the uncorrected test under relatedness, there is a gain in power for low p -value thresholds but a loss in power for higher p -value thresholds. Another simulation study [11] reported that the power did not notably differ if relatedness is ignored.

In the present paper, we aim to investigate how heritability and strength of relatedness contribute to variance inflation of the effect estimate and present simple approximation formulae. We evaluate subsequently the impact of variance inflation on type I error and power of the test and identify situations in which simple linear regression is still valid. Additionally, we prove that the expectation of effect estimates is not influenced as noticed by simulation studies [1, 11] and explain why allele frequencies appear to have only little impact on type I error and power (see [1, 14]).

The paper is organized as follows: In the “Methods” section, we present the underlying theory and derive the equations. We first introduce the notation of relatedness structure. Then, we present both, the general linear model of SNP-phenotype association under relatedness and its counter-part of ignored relatedness. We show unbiasedness of the effect estimate of the SNP of the second model and derive its variance inflation under relatedness. We study the impact of variance inflation on hypothesis testing and compare our results with those of genomic control correction. In the “Results” section, we analyse the relatedness structure of the publicly available HapMap data, an isolated population and synthetic family structures and their impact using the derived formulae. Major formulae derived in the paper were implemented in an R script provided as Additional file 1.

Methods

Almost all of the equations presented in the sections below are derived in Additional file 2. Notations and a list of symbols are provided in Additional file 2: Sections 1 and 7, respectively.

Relatedness

When dealing with relatedness, it is important to understand what exactly it means that one individual “is related” to another individual. We introduce the corresponding notation following Wang [15]. We assume bi-allelic markers (SNPs) without missing genotypes throughout. SNP genotype s_i of the i th individual corresponds to the number of reference alleles 0, 1 or 2.

We denote ϕ and δ as the probabilities that only one allele and both alleles, respectively, are inherited IBD (identical by descent) from a common ancestor. Then, relatedness is defined as $G = \phi/2 + \delta$. It holds that $0 \leq G \leq 1$. Of note, different kinds of relatedness, e.g. a parent child pair ($\phi = 1, \delta = 0$) or full siblings ($\phi = 1/2, \delta = 1/4$), can yield the same G . In these cases the expectation of G equals 1/2. The true underlying relatedness structure is often unknown. However, it can be estimated on a sufficiently rich data basis such as genome-wide SNP arrays. For estimation, we applied the method described in [15]. Our analysis is based on these relatedness estimates rather than relationships obtained from pedigrees which are often not available or prone to errors. For estimation of relatedness, SNP weights are required which depend on the respective allele frequencies. For this purpose, allele frequencies for each SNP s were assessed by the simple estimate $\hat{p} = \sum_{i=1}^n s_i/2n$ for n samples.

For most of the approximation formulae presented below, we require that the mean relatedness, i.e. the average of the entries G_{ij} , $i \neq j$, is small, i.e. less than 0.01. This applies for example for a sufficiently large number

of trios or families or even large pedigrees over several generations (see Table 1 below).

Modelling a SNP - phenotype association

We assume that phenotypes \mathbf{y} follows the “true” mixed model

$$y_i = b_1 + b_2 s_i + g_i + e_i \quad (1)$$

with intercept b_1 , SNP effect b_2 , random (polygenic) effects $\mathbf{g} = (g_1, g_2, \dots, g_n)$ and residuals $\mathbf{e} = (e_1, e_2, \dots, e_n)$ for $i = 1, 2, \dots, n$ observations. For the random effects, we assume that $\mathbf{g} \sim N_n(0, \sigma_g^2 \mathbf{G})$ is multivariate normal with a certain variance σ_g^2 and relatedness matrix \mathbf{G} . The possible dependence of phenotypes of two individuals i and j originates from the polygenic random effects g_i and g_j . The random effects depend on the relatedness of both individuals which can be expressed in terms of G_{ij} varying between zero and one. This implies that the polygenic contribution to the phenotype ranges from “independent” to “identical” for a pair of individuals. We assume that residuals are uncorrelated between observations and distributed as multivariate normal $\mathbf{e} \sim N_n(0, \sigma_e^2 \mathbf{I})$ with certain variance σ_e^2 and identity matrix \mathbf{I} . The heritability of $\mathbf{y} = (y_1, y_2, \dots, y_n)$ can be expressed through $R_h^2 = \sigma_g^2 / (\sigma_g^2 + \sigma_e^2)$.

Ignoring relatedness results in the following simpler model to be fitted to the data:

$$y_i = \beta_1 + \beta_2 s_i + \epsilon_i \quad (2)$$

assuming uncorrelated residuals $\mathbf{\epsilon} = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)$ only. We aim at deriving analytical formulae for the expectation and variance of $\hat{\beta}_2$ given the true model, i.e. we analyse the impact of relatedness on the estimates obtained with Eq. (2).

After some calculations (Additional file 2: Section 2.2), it follows that the expected value $E(\hat{\beta}_2) = b_2$ is not biased

by relatedness irrespective of its structure. However, the variance of $\hat{\beta}_2$ is affected:

$$V(\hat{\beta}_2) = \frac{\sigma_e^2}{\sum_{i=1}^n (s_i - \bar{s})^2} \frac{1}{1 - R_h^2} \left(1 + R_h^2 \frac{\sum_{i=1}^n \sum_{j \neq i=1}^n G_{ij}(\bar{s}^2 - 2\bar{s}s_i + s_i s_j)}{\sum_{i=1}^n (s_i - \bar{s})^2} \right). \quad (3)$$

Without heritability, i.e. $R_h^2 = 0$, the phenotypes for all pairs of individuals are uncorrelated and the last two terms of Eq. (3) simplify to 1. In this case, we obtain

$$V_\beta = \frac{\sigma_e^2}{\sum_{i=1}^n (s_i - \bar{s})^2}.$$

This variance is equivalent to the variance of the standard linear model as shown in [16]. For the last term of $V(\hat{\beta}_2)$ in Eq. (3), we define the inflation factor

$$\lambda = 1 + R_h^2 \frac{\sum_{i=1}^n \sum_{j \neq i=1}^n G_{ij}(\bar{s}^2 - 2\bar{s}s_i + s_i s_j)}{\sum_{i=1}^n (s_i - \bar{s})^2} \quad (4)$$

which depends on the heritability R_h^2 and the pairwise relatedness matrix \mathbf{G} . Using λ , $V(\hat{\beta}_2)$ can be rewritten as

$$V(\hat{\beta}_2) = \frac{\lambda}{1 - R_h^2} V_\beta. \quad (5)$$

As we will see in the “Hypothesis testing” section, the empirical variance of the effect estimate is also inflated by factor $1/(1 - R_h^2)$. Hence, this factor is cancelled out when estimating the corresponding T statistic.

Expected variance inflation

An approximation formula for λ can be obtained by separately deriving the expectations of the numerator and denominator of Eq. (4) as shown in Additional file 2: Section 3.2:

$$\lambda' = 1 + R_h^2 \frac{G_2 - \frac{2}{n} G_r}{n - 1} \quad (6)$$

where

Table 1 Estimated variance inflation under relatedness

Study	n	$\bar{\lambda}$	$\bar{\lambda}_{10\%}$	λ'	$\lambda'_{f,mc}$	\bar{G}	R_t^2
HapMap	129	1.288 (0.074)	1.295 (0.051)	1.297	-	0.006	0.152
SFS1	129	1.284 (0.087)	1.293 (0.051)	1.294	1.295	0.007	0.153
SFS2	999	1.306 (0.050)	1.313 (0.020)	1.314	1.299	0.001	0.143
Sorbs	977	1.410 (0.135)	1.448 (0.071)	1.449	-	0.001	0.100
SFS3	999	2.006 (0.139)	2.022 (0.083)	2.021	2.002	0.002	0.044

Variance inflation and related measures are compared between the data sets HapMap, SFS1 (synthetic family study 1), SFS2, Sorbs and SFS3 assuming $R_h^2 = 0.9$. Provided are the sample size n , average inflation $\bar{\lambda}$ of all SNPs, average inflation $\bar{\lambda}_{10\%}$ estimated for SNPs with minor allele frequencies $> 10\%$, expected (theoretical) inflation λ' obtained from estimated relationships, expected inflation $\lambda'_{f,mc}$ obtained from true relationships (synthetic family studies only), mean relatedness \bar{G} and heritability R_t^2 corresponding to inflation $\lambda'_t = 1.05$. Standard deviations are given in parentheses

$$G_2 = \sum_{i=1}^n \sum_{j \neq i=1}^n G_{ij}^2$$

$$G_r = \sum_{i=1}^n \left(\sum_{j \neq i=1}^n G_{ij} \right)^2$$

correspond to the sum of squared elements and the sum of the squared row sums of \mathbf{G} , respectively. Approximating $E(\lambda)$ by λ' is valid if the number n of observations is large and the mean relatedness

$$\bar{G} = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i=1}^n G_{ij}$$

is small. Interestingly, Eq. (6) is independent of the allele frequency explaining the empirical observations of [1, 14]. For details, see Additional file 2: Section 3.2.

Relationship between heritability and inflation

There are some useful transformations of Eq. (6): If λ' is available for a specific heritability R_h^2 , it is easy to derive the inflation λ'_t for an alternative heritability R_t^2 given the same relatedness structure. As can be seen from Eq. (6), it holds that

$$R_t^2 = \frac{\lambda'_t - 1}{\lambda' - 1} R_h^2, \quad (7)$$

$$\lambda'_t = 1 + (\lambda' - 1) \frac{R_t^2}{R_h^2}. \quad (8)$$

See also Additional file 2: Section 3.3.

Example family structures

Using Eq. (6), inflation λ' can be estimated for arbitrary family structures. As an example, assume a family study with f families with one father per family. Each father is mated with m mothers and each mother has c children. Then, the number of samples is $n = (cm + m + 1)f$. Given these relationships as relatedness matrix \mathbf{G} , inflation λ' can be explicitly calculated by

$$\begin{aligned} \lambda'_{f,m;c} = & 1 + R_h^2 [((c^3 + c^2)f - 2c^3)m^3 \\ & + ((3c^3 + 16c^2 + 12c)f - 4c^3 - 16c^2)m^2 + \\ & ((3c^2 + 12c)f - 2c^3 - 16c^2 - 8c)m] / \\ & [(16c^2 + 32c + 16)fm^2 + \\ & ((32c + 32)f - 16c - 16)m + 16f - 16]. \end{aligned} \quad (9)$$

The formula is implemented in an R script (see Additional file 1). The special case of $m = 1, c = 1$ corresponds to trios in which Eq. (9) simplifies to

$$\lambda'_{f,1;1} = 1 + R_h^2 \frac{f - 1}{3f - 1}. \quad (10)$$

Another example is a study with an increased number of pairwise relationships ($m = 2, c = 3$) where Eq. (9) simplifies to

$$\lambda'_{f,2;3} = 1 + R_h^2 \frac{243f - 314}{216f - 24}. \quad (11)$$

Details of these formulae are provided in Additional file 2: Section 3.4 and Additional file 3.

Hypothesis testing

Assume we observe phenotypes \mathbf{y} and SNP genotypes \mathbf{s} obeying Eq. (1). We are interested whether the phenotype is associated with the SNP. For the simplified regression model in Eq. (2), this corresponds to testing the null hypothesis of $\beta_2 = 0$. Thus, the test statistic $T = \hat{\beta}_2/S_\beta$ as presented in [17] is evaluated. S_β^2 denotes the empirical variance estimate of $\hat{\beta}_2$. Evaluating the distribution of the test statistic under the null hypothesis is required for assessing the type I error. The distribution of the test statistic under the alternative hypothesis is required for calculating the power of the test. In reference to Additional file 2: Section 5.1, the effect estimate $\hat{\beta}_2$ is normally distributed, and, if the variance of S_β^2 is small, one can replace S_β^2 by its expected value $E(S_\beta^2)$. This implies that T is approximately normally distributed with expectation and variance as follows

$$T \sim N \left(\frac{E(\hat{\beta}_2)}{\sqrt{E(S_\beta^2)}}, \frac{V(\hat{\beta}_2)}{E(S_\beta^2)} \right).$$

Assuming the null hypothesis, one obtains $E(\hat{\beta}_2) = 0$. Further, using $V(\hat{\beta}_2) = \lambda V_\beta / (1 - R_h^2)$ as shown in Eq. (5) and $E(S_\beta^2) \approx V_\beta / (1 - R_h^2)$ as given in Additional file 2: Section 4.2, the distribution of T can be calculated:

$$T \sim N(0, \lambda). \quad (12)$$

See also Additional file 2: Section 5.2.

Considering the alternative hypothesis, it holds that $E(\hat{\beta}_2) = b_2$ and $E(S_\beta^2) \approx V_\beta / (1 - R_h^2)$ in analogy to the null hypothesis. In the following, we assume a fixed explained variance of the SNP R_s^2 . Thus, the SNP effect is described by only one parameter. Alternatively, if specifying a fixed SNP effect b_2 , test statistics would also depend on the allele frequency, i.e. two parameters would be required. For a given R_s^2 , it holds that

$$E(T) \approx \sqrt{(n-1)R_s^2} = \mu \quad (13)$$

as shown in Additional file 2: Section 5.3. Finally, an approximation of the distribution of T under the alternative hypothesis can be derived:

$$T \sim N(\mu, \lambda). \quad (14)$$

Here, caused by relatedness, the empirical variance of the effect estimate $E(S_\beta^2)$ is deflated compared to $V_\beta / (1 - R_h^2)$ by a certain factor ν as shown in Additional file 2: Section 4.2.

Further, assume $F_N(x|\mu, \sigma^2)$ is the cumulative distribution function of the normal distribution with expectation μ and variance σ^2 . Given the quantile $z_{\alpha/2}$ of the standard normal distribution corresponding to a two-sided test with significance level α , the type I error of the test applying Eq. (12) can be derived

$$\text{err} = 2F_N(z_{\alpha/2}|0, \lambda). \quad (15)$$

Similarly, the power of the test applying Eq. (14) is

$$\text{pwr} = F_N(z_{\alpha/2}|\mu, \lambda) + 1 - F_N(-z_{\alpha/2}|\mu, \lambda). \quad (16)$$

Genomic control

Genomic control [12] is a simple and often used method to correct for variance inflation. Given a sample of n realisations $\hat{T}_1, \hat{T}_2, \dots, \hat{T}_n$ of T under the null hypothesis, an estimate of λ according to Additional file 2: Section 6 is

$$\hat{\lambda} = \frac{\text{median}(\hat{T}_1^2, \hat{T}_2^2, \dots, \hat{T}_n^2)}{0.456}.$$

Genomic control correction is performed by calculating $T_{gc} = T/\sqrt{\hat{\lambda}}$ and using T_{gc} as new test statistic. Correcting the variance inflation of T under the null hypothesis (see Eq. (12)), the test statistic T_{gc} is approximately standard normally distributed:

$$T_{gc} \sim N(0, 1). \quad (17)$$

Since

$$\text{err}_{gc} = 2F_N(z_{\alpha/2}|0, 1) = \alpha, \quad (18)$$

the type I error of the test is preserved.

In contrast, correction of the alternative statistic T distributed as shown in Eq. (14) yields

$$T_{gc} \sim N\left(\frac{\mu}{\sqrt{\hat{\lambda}}}, 1\right). \quad (19)$$

Thus, genomic control correction reduces the expectation of the test statistic, and with it, the power of the test in comparison to Eq. (16) unless λ is close to 1:

$$\text{pwr}_{gc} = F_N\left(z_{\alpha/2} \left| \frac{\mu}{\sqrt{\hat{\lambda}}}, 1\right.\right) + 1 - F_N\left(-z_{\alpha/2} \left| \frac{\mu}{\sqrt{\hat{\lambda}}}, 1\right.\right). \quad (20)$$

Samples

To apply our equations to real data, we consider HapMap CEU (CEPH (Centre d'Etude du Polymorphisme Humain) from Utah) trio data for two reasons. First, these genotype data is freely accessible and well understood so that

our results can easily be reproduced. Secondly, the relatedness structure is simple in order to promote understanding of our equations. A simple relatedness structure also supports simulation of genotype data to obtain results under different settings, e.g. increased sample size. Filtering of HapMap SNPs and samples prior to analysis is described in Additional file 4. A matrix of pairwise relatedness estimates for all HapMap CEU samples is provided as Additional file 5. In summary, 1,020,215 SNPs measured in 129 HapMap samples belonging to 43 trios were available for analysis. Additional file 6 contains a detailed list of samples and the reason for exclusion where applicable, whereas Additional file 7 provides the list of SNP identifiers used for analysis. The Perl script provided as Additional file 8 together with the sample list in Additional file 6 and the SNP list in Additional file 7 can be used for converting the HapMap CEU data [18] to a CSV (comma separated values) file which is further analysed.

Furthermore, we analysed a sample of the Sorbs who are an ethnic minority in Germany with putative genetic isolation [13, 19]. The Sorbs sample is characterised by a complex relatedness structure and therefore suitable for analysis of variance inflation. As done in [13], 471,012 autosomal SNPs were filtered for call rate < 95%, deviation from Hardy-Weinberg equilibrium with $p < 10^{-6}$ and platform association with $p < 10^{-7}$. After filtering, 424,476 SNPs measured in 977 samples were available for analysis.

Finally, synthetic genotypes were simulated for three studies each consisting of f families with one father per family, m mothers per father and c children per mother as described in Additional file 2: Section 3.4. In order to evaluate the results obtained for the HapMap data, a study (SFS1, synthetic family study 1) was simulated for $n = 129$ samples with parameter set $f = 43, m = 1, c = 1$. For the second study (SFS2), the relatedness structure was kept similar but the sample size was increased to $n = 999$, i.e. the parameter set was $f = 333, m = 1, c = 1$. For stronger relationships but the same $n = 999$ samples, we simulated a third study (SFS3) with parameter set $f = 111, m = 2, c = 3$. For all synthetic studies, we sampled 110,000 SNPs where the reference allele of each SNP was drawn from a beta distribution (shape $a = 0.5$, shape $b = 0.5$).

Simulation

For simulation and analysis of the results, we used the statistical software package R [20]. The script is provided as Additional file 1. Instead of sampling SNPs for a synthetic family study, genotypes provided as CSV file can also be loaded and analysed utilising this R script. The HapMap and Sorbs genotype data were analysed in this way. In any case, a random subset of 100,000 non-monomorphic SNPs was selected for all studies. The R script was also used to estimate pairwise relatedness according to

Wang [15], to calculate the variance inflation λ given the SNP genotypes as presented in Eq. (4) averaged over all SNPs and to calculate the expected inflation λ' based on estimated relationships as shown in Eq. (6). Further, the R script supports simulation of phenotypes under the null and alternative hypothesis assuming Eq. (1) for empirical verification of the test statistics as presented in Eqs. (12) and (14), respectively. Empirical values of the statistics were derived by simulations as follows: For each SNP, phenotypes are drawn repeatedly from a multivariate normal distribution where the expectation depends on the SNP if simulating alternative hypotheses or is independent of it for simulating null hypotheses. These simulated test statistics were averaged over phenotype realisations and the empirical variance was estimated to assess inflation due to relatedness. The resulting mean test statistics and their empirical variances were averaged over SNPs and a standard deviation was calculated to control sampling errors. Due to the computational burden, simulations were restricted to 1000 phenotype realisations per SNP and a random subset of 1000 SNPs.

Results

Variance inflation for examples of relatedness

We apply the formulae derived in the “Methods” section to assess and compare variance inflation between different scenarios of relatedness structure and heritability. Given the genotypes of a SNP s , the estimated relatedness matrix \mathbf{G} and the heritability R_h^2 one can calculate the variance inflation based on Eq. (4).

Different relatedness structures result in different degrees of variance inflation. We demonstrate this on an example of a synthetic family study consisting of f families with one father per family, m mothers and c children. Further, assume that each study comprises the same number n of individuals but differs in c and m . Therefore, we set $f = \text{floor}(n/(cm + m + 1))$ (“floor” returns the largest integer not greater than the argument) and estimate the expected variance inflation of the effect estimate by evaluating Eq. (9). Figure 1 shows the expected inflation $\lambda'_{f,m;c}$ for heritability $R_h^2 = 0.9$ and different settings of m and c resulting in the same sample size $n = 1000$. For example, a trio study with $f = 333$, $m = 1$ and $c = 1$ ($n = 999$) results in $\lambda'_{333;1;1} = 1.3$. This value can also be obtained via Eq. (10). A more extreme example is a family study with $f = 111$, $m = 2$ and $c = 3$ ($n = 999$) which results in $\lambda'_{111;2;3} = 2$ (see also Eq. (11)). Inflation λ' also depends on sample size, but notable differences can only be observed for small sample sizes (i.e. $n < 100$).

For a random subset of 100,000 non-monomorphic SNPs, we estimated the variance inflation for the real HapMap trio data, the Sorbs data and the above mentioned synthetic family studies SFS1 (corresponding to HapMap study), SFS2 (corresponding to trios with a larger

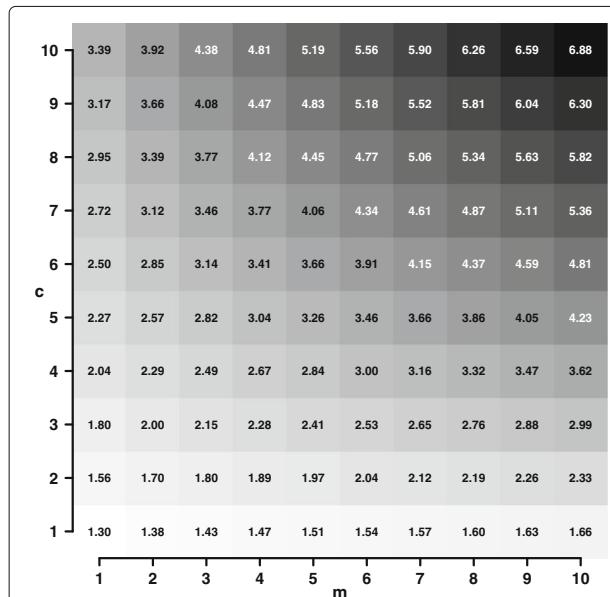


Fig. 1 Expected variance inflation for synthetic family studies. The figure presents the expected variance inflation $\lambda'_{f,m;c}$ for heritability $R_h^2 = 0.9$ and family studies with varying numbers of mothers m and children c , each between 1 and 10, and with a total of about $n = 1000$ individuals. The background colour corresponds to the values presented and ranges from white for the minimum to black for the maximum inflation

sample size of $n = 999$) and SFS3 (corresponding to the same sample size as SFS2 but a higher average relatedness). Results for $R_h^2 = 0.9$ are presented in Table 1. The empirical variance inflation λ is smallest for HapMap and SFS1, the latter two are in well agreement as expected. The higher sample size for SFS2 results in slightly higher inflation. The Sorbs inflation is even higher than for SFS2. As expected, SFS3 shows the strongest inflation. Using λ' instead of λ results in slightly higher values due to the Taylor expansion used to derive Eq. (6) (see Additional file 2: Section 3.2). But the difference is without practical relevance. Restricting to minor allele frequencies $> 10\%$ improves the agreement (see Table 1 column $\bar{\lambda}_{10\%}$). The expected variance inflation λ' calculated from the estimated relatedness matrix agrees well with $\lambda'_{f,m;c}$ calculated from true relationships. Of note, if heritability R_t^2 drops below 10% for HapMap, Sorbs, SFS1 and SFS2 according to Eq. (7), inflation becomes irrelevant ($\lambda'_t < 1.05$, see Table 1 for details). However, inflation for the extreme situation of study population SFS3 is still $\lambda'_t = 1.11$ as calculated with Eq. (8).

Numerical validation of test statistics

The distributions of the test statistic T in Eqs. (12) and (14) are approximations due to the approximation of

the variance estimate. To empirically verify these approximations, we simulated multivariate normally distributed phenotypes and fitted a linear model afterwards. We analysed the same five study populations as in the previous section and again assumed $R_h^2 = 0.9$. Results are presented in Table 2 for the null hypothesis and Table 3 for the alternative hypothesis. The expectation and empirical variance of T was averaged over SNPs. As expected, the expectation of T under the null hypothesis is close to zero for all studies (Table 2). The expectation under the alternative is close to its theoretical value μ calculated via Eq. (13) (Table 3), i.e. no relevant biases were observed for T under both hypotheses. However, the variance of T is slightly overestimated in comparison to the derived λ values presented in Table 1 (compare \bar{S}^2 of Tables 2 and 3 with $\bar{\lambda}$ of Table 1). The difference is more pronounced for the studies with small samples sizes, i.e. HapMap and SFS1. For larger studies, the difference is without practical importance. Although the empirical variance of the effect estimate is deflated by factor v (see Additional file 2: Section 4.2 and Table 2), this deflation is close to 1 in our data, and again, is without practical relevance.

Examples of inflation factors

Since heritability and relatedness structure directly translate into inflation factors, we study the latter in the following in more detail. To study type I error and power of the tests, we consider four different inflation scenarios $\lambda = 1$, i.e. no inflation, and $\lambda = 1.05, 1.3$ and 2 . For example, any study comprising unrelated individuals results in about $\lambda = 1$, whereas our study populations SFS1 with $R_h^2 = 0.15$, SFS2 and SFS3 with $R_h^2 = 0.9$ result in about $\lambda = 1.05, 1.3$ and 2 , respectively. See also Table 1 for the latter three scenarios.

Impact of inflation on type I error

In the situation of statistical testing, the variance of T under the null hypothesis is relevant for the type I error. Its inflation originates from heritability R_h^2 and the family

Table 2 Simulation results for the test statistic T under the null hypothesis

Study	\bar{T}	\bar{S}^2	v
HapMap	0.002 (0.037)	1.330 (0.096)	0.992
SFS1	-0.000 (0.037)	1.321 (0.107)	0.992
SFS2	-0.001 (0.037)	1.309 (0.076)	0.999
Sorbs	-0.001 (0.037)	1.412 (0.144)	0.999
SFS3	0.001 (0.043)	2.015 (0.166)	0.997

The test statistics \bar{T} averaged over replicates and SNPs and the average of the empirical variances \bar{S}^2 are compared between HapMap, SFS1 (synthetic family study 1), SFS2, Sorbs and SFS3 assuming the null hypothesis and $R_h^2 = 0.9$. Standard deviations are presented in parentheses. We further provide an estimate of the deflation factor v for the empirical variance of the effect estimate

Table 3 Simulation results for the test statistic T under the alternative hypothesis

Study	\bar{T}	\bar{S}^2	μ
HapMap	1.619 (0.037)	1.343 (0.095)	1.600
SFS1	1.619 (0.036)	1.336 (0.112)	1.600
SFS2	4.472 (0.036)	1.330 (0.076)	4.468
Sorbs	4.420 (0.039)	1.432 (0.148)	4.418
SFS3	4.479 (0.046)	2.030 (0.162)	4.468

The test statistics \bar{T} averaged over replicates and SNPs and the average of the empirical variances \bar{S}^2 are compared between HapMap, SFS1 (synthetic family study 1), SFS2, Sorbs and SFS3 assuming the alternative hypothesis with $R_s^2 = 0.02$ and heritability $R_h^2 = 0.9$. Standard deviations are presented in parentheses. We further provide the expected value μ of the test statistic T

structure as shown in Eq. (4). Variance inflation λ impacts the distribution of the test statistic under the null hypothesis as shown in Eq. (12) and affects the type I error of the test as depicted in Eq. (15). In Fig. 2, we present the type I error dependent on the significance level without inflation $\lambda = 1$ and inflation with $\lambda = 1.05, 1.3$ and 2 as in the above mentioned scenarios. Type I error for $\lambda = 1.05$ is similar to $\lambda = 1$ justifying the 1.05 threshold typically applied to ignore inflation. However, the type I error increases rapidly with increasing inflation.

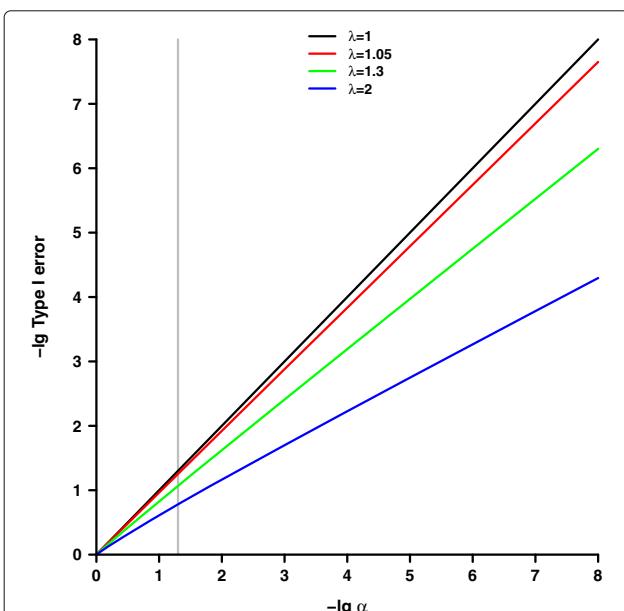


Fig. 2 Comparison of type I errors with respect to different degrees of variance inflation. The figure provides a comparison of type I errors dependent on the significance level α without variance inflation $\lambda = 1$ and variance inflation with $\lambda = 1.05, 1.3$ and 2 . The negative common logarithm is presented for α as well as the type I error. The grey vertical line corresponds to a significance level of $\alpha = 0.05$

Impact of inflation on power

For calculating the power, expectation and variance of T under the alternative is required. As shown in Eq. (4), variance inflation depends on heritability R_h^2 and the family structure. Similar to the null hypothesis, variance inflation λ impacts the distribution of the test statistic under the alternative as shown in Eq. (14) and affects the power of the test (Eq. (16)). The expectation of T , see Eq. (13), depends on the sample size n and the explained variance by the SNP R_s^2 . We assume $n = 1000$ and $R_s^2 = 0.02$ resulting in an expectation of the test statistic of $\mu = \sqrt{(n-1)R_s^2} \approx 4.47$. For this expectation, we present Fig. 3a showing the dependence of power, see Eq. (16), on the significance level for $\lambda = 1$ (no inflation) and $\lambda = 1.05, 1.3$ and 2 . The power for $\lambda = 1.05$ is similar to $\lambda = 1$, indicating again that this inflation is negligible for practical purposes. The difference is more pronounced for the other power curves with $\lambda > 1.05$. Irrespective of the variance of the test statistic, the power curves are intersecting at 50%. For the selected expectation, this corresponds to $-\lg(\alpha) \approx 5.11$ ("lg" refers to the common logarithm with base 10). Thus, for smaller significance levels, the power increases with increasing inflation while the opposite occurs for larger significance levels.

Correction with genomic control

In case of inflation, an often applied method of correction is genomic control. If this correction is applied in the situation of relatedness, the distribution of the test statistic (Eq. (17)) under the null hypothesis is approximately standard normal. This implies that the type I error α (Eq. (18)) is preserved. In contrast, correcting the test statistic by the inflation factor reduces the expectation

(Eq. (19)) under the alternative hypothesis which in turn reduces the power (Eq. (20)) of the test. In Fig. 3b, we provide the power dependent on the significance level after genomic control without inflation $\lambda = 1$ and with inflation $\lambda=1.05, 1.3$ and 2 . Comparing Fig. 3a and b, power loss of genomic control increases rapidly with increasing λ . Thus, genomic control cannot be recommended for inflations $\lambda > 1.05$ induced by relatedness.

Discussion

Relatedness induces a dependency structure to phenotypic data, and therefore, needs to be addressed appropriately in genetic association studies. However, the impact of relatedness on key statistical properties is insufficiently studied and major insights rely on simulation studies only. Here, we provide a full theory of the impact of relatedness on linear regression analysis of a quantitative phenotype. We derive analytical formulae of test statistics and provide a simple approximate formula of the dependence of variance inflation on the relatedness structure. We studied the impact of relatedness on type I error and power and confirmed a number of phenomena observed in simulation studies. Moreover, we showed that genomic control cannot be recommended to deal with relatedness-induced inflation. All formulae were implemented in an R script provided as supplement (Additional file 1).

First, we derived formulae of the impact of relatedness on effect estimates and variances of a linear regression model. We proved that the expectation is unbiased in agreement with [1, 11] who observed this fact on the basis of simulation studies. We derived an approximation formula of the variance inflation given the relatedness and the heritability of the phenotype. We also proved that the

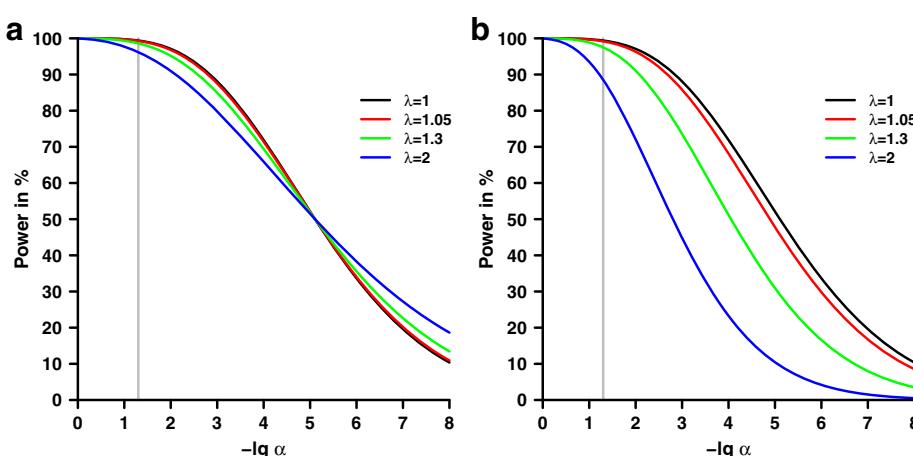


Fig. 3 Comparison of power with respect to different degrees of variance inflation. Both figures provide a comparison of power in percent dependent on the significance level α without variance inflation $\lambda = 1$ and variance inflation with $\lambda=1.05, 1.3$ and 2 . Figure **a** corresponds to the uncorrected test statistic, whereas Figure **b** refers to the test statistic after genomic control. The negative common logarithm is presented for α . The grey vertical line corresponds to a significance level of $\alpha = 0.05$. An explained variance of $R_s^2 = 0.02$ was assumed. Sample size was set to $n = 1000$

standard error of the effect estimate is underestimated if applying the standard linear model. This is reflected by the deflation factor v derived in Additional file 2: Section 4.2. Again, this issue was observed by [1] on the basis of a simulation study.

We estimated this variance inflation for “real” genotype data obtained from HapMap trios and the Sorbs and for synthetic genotypes of three different family studies of varying degree of relatedness. For a heritability of 90%, we showed that there is a relevant inflation for all of these studies. In contrast, if heritability drops below 10%, the inflation is only relevant in the extreme situation of study population SFS3. See also Additional file 9 for additional results of scenarios with varying degree of heritability.

The polygenic effect was modelled via a multivariate normal distribution with the relatedness matrix as covariance matrix. Alternatively, the polygenic effect could be modelled by single markers as proposed by Zhang et al. [3]. Results are similar even for small numbers of SNPs contributing to the polygenic effect (see Additional file 10).

For analysis, we utilised relatedness estimates obtained from genomic data rather than estimates obtained from pedigree data. First, correct pedigree data are difficult to assess especially for non-family studies or studies with cryptic relatedness as observed in isolated populations, e.g. the Sorbs [13]. Second, [5, 14] argued that estimates from marker data reflect true genetic relationships better than estimates from even a correct pedigree. In contrast to [5] who applied kinship estimates as presented in [21], we estimated pairwise relatedness with the method proposed by Wang [15]. The latter has several advantages as correction for allele frequency estimates. Otherwise, relatedness estimates could be biased [15, 21], see also Fig. 1 in Additional file 11. However, in our hands using the kinship matrix [5, 21] or the IBS(identical by state)-based matrix [4, 22] as alternative estimators, this has little impact on the inflation results (see Additional file 11). Further, the method in [15] results in a diagonal of the estimated relatedness matrix identical to 1 which is required for our derivations in Additional file 2: Section 2.2.

In general, inflation depends on the allele frequency of a SNP. However, considering our approximation formula Eq. (6), this dependency can be neglected if the sample size is sufficiently large and the average relatedness is small. This explains corresponding empirical observations of [1, 14].

As different combinations of relatedness structure and heritability yield the same variance inflation, we further focused on different degrees of variance inflation to study type I error and power. For this purpose, we derived an analytical approximation of the test statistic given the variance inflation. The approximation was successfully verified in a simulation study.

We showed analytically that the type I error increases with inflation. With our formula, we could confirm the empirical observation of [1, 11] that type I error of the test increases with higher heritability and stronger relationships. Similarly, [9] observed an inflated type I error when the family structure is ignored.

A major result of our study is that the power increases with increasing inflation if the significance level is small while the opposite occurs for larger significance levels. We already observed this phenomenon in a previously published simulation study [13]. This explains a number of contrary empirical observations presented in the literature, e.g. [1, 9] noted that the power of the test is reduced when ignoring the family structure. However, [11] observed similar power irrespective whether accounting for the family structure or not. By our formula, we could show that the power could be either increased or decreased under inflation in dependence on the underlying significance threshold.

Our formulae can also be applied to compare the impact of family structures between studies. Power and type I error were analysed previously in [1, 5] for a nuclear pedigree (NP) of 1011 individuals belonging to 337 sib trios. Applying our formulae (Additional file 3), this family structure results in an inflation factor of 1.45 for $R_h^2 = 0.9$. Interestingly, the same value was observed for the Sorbs sample.

Since genomic control is an often applied method to correct for inflated test statistics, we studied its results in the situation of relatedness-induced inflation. We could show that genomic control maintains the correct type I error which is in line with [5, 12]. However, we also showed that genomic control seriously impairs power. This was acknowledged by [12] for increased inflation and by [5] for higher heritability and stronger relationships. According to our results, genomic control cannot be recommended to deal with inflation due to relatedness. One has to remark that genomic control was originally developed to correct for population stratification [23, 24]. In contrast to other studies [12, 14, 21], we did not consider additional population structure here. Results for selected settings of heritability and explained variance of the SNP are presented in the paper. More scenarios can be easily analysed using our R script provided as Additional file 1.

The properties of various correction methods as well as simple linear regression are compared in [10]. Here, we investigated the linear model in detail, provided an easy to apply approximation formula of the impact of relatedness on variance inflation and identified scenarios where simple linear regression analysis is still valid. We agree with Aulchenko [14] that a variance inflation below 1.05 is negligible regarding power and type I error. If variance inflation is larger, we advice to apply methods which explicitly account for relatedness, e.g. by mixed model

analysis [1, 5, 9, 25–27]. Nonetheless, these models need to be carefully applied due to several pitfalls [28]. For a summary of correction methods and software tools, see also [29].

Conclusions

We developed approximation formulae to study the impact of relatedness on type I error and power. We could prove a number of empirical observations made in simulation studies. Stronger relatedness as well as higher heritability result in increased variances of the effect estimates of simple linear regression analyses. As a consequence, type I error rates are generally inflated. The behaviour of power is more complicate since relatedness could either increase or reduce it in dependence on the effect size of a SNP, the heritability of the phenotype and the significance threshold. Genomic control cannot be recommended to deal with relatedness-induced inflation. Variance inflation below 1.05 can be safely ignored, i.e. simple linear regression analysis is still appropriate in this case.

Additional files

Additional file 1: R script for simulation. This R script supports simulation of synthetic genotypes for a family study. Instead of genotype simulation, genotypes can also be loaded from a CSV file. Allele frequencies are calculated, monomorphic SNPs are filtered and pairwise relatedness is estimated. Given SNP genotypes and a value for the heritability, variance inflation λ is calculated. Additionally, the expected λ' is estimated. Finally, the script simulates phenotypes under the null and alternative hypothesis and provides results regarding the T statistic. The R library "mvtnorm" is required for sampling multivariate normally distributed phenotypes. Parameters can be modified to simulate different scenarios. However, the number of samples, the number of SNPs and the number of phenotype realisations per SNP should be limited to reduce the computational burden. For example, running the script on an Intel Xeon X5560 CPU (2.80 GHz) for synthetic family study 3 (SFS3) with parameter set $f = 111, m = 2, c = 3$ ($n = 999$), 100000 SNPs, 1000 phenotype realisations per SNP and 1000 SNPs required 8.3 GB RAM and took < 1 min for genotype sampling, 8 min for estimation of pairwise relatedness, 21 min for λ estimation and about 2.5 h for each of the phenotype simulations under the null and alternative hypothesis, respectively. (R 6 kb)

Additional file 2: Theoretical background. This file provides the theoretical background and derivations of equations presented in the manuscript. (PDF 231 kb)

Additional file 3: Maxima script for deriving expected variance inflation. This script can be used with MAXIMA [30] for deriving formulae for the expected variance inflation $\lambda'_{f,mc}$ for synthetic family studies. (WXM 1 kb)

Additional file 4: Preparation of HapMap data. This document provides details regarding the filtering of samples and SNPs of the HapMap data. (PDF 97 kb)

Additional file 5: Pairwise relatedness estimates of HapMap samples. This file contains a matrix of pairwise relatedness estimates resulting from the preliminary analysis of 174 HapMap CEU samples. Sample identifiers for the pair of individuals under consideration are given in the first row and in the first column, respectively. A value of -1 occurs if pairwise relatedness could not be estimated because of disjoint SNP sets. (CSV 571 kb)

Additional file 6: Sample selection of HapMap genotype data. This file provides annotations for 174 HapMap CEU samples. The columns FID (family identifier), IID (individual identifier), dad, mom, sex (1=male, 2=female), pheno (always 0), population (always CEU) correspond to the

columns of relationships_w_pops_121708.txt filtered for CEU samples as provided by HapMap. The column ctr contains a unique trio identifier and equals NA when the sample does not belong to a complete trio family. The reason for exclusion is provided where applicable, otherwise NA is stated and the sample is included in our study. (CSV 8 kb)

Additional file 7: SNP selection of HapMap genotype data. This file contains a list of HapMap SNP identifiers used for our analyses. rsid (reference SNP identifier) refers to the first column of the genotype data files as provided by HapMap. (CSV 10000 kb)

Additional file 8: Perl script for converting HapMap genotype data. This Perl script requires the sample list of Additional file 6, the SNP list of Additional file 7 and HapMap raw data. The HapMap project website is not available anymore, however, genotype data can still be retrieved from ftp://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/2010-08_phasell+III/. The converted genotypes are saved in a CSV file. Folder and file locations must be adapted before running the script. Running the script on an Intel Xeon X5560 CPU (2.80 GHz) required 800 MB RAM and took about 5 minutes. (PL 2 kb)

Additional file 9: Comparison of different degrees of heritability. This file contains additional tables with inflation results for different degrees of heritability. (PDF 75 kb)

Additional file 10: Comparison of methods for modelling the polygenic effect. This file provides additional tables with inflation results for different polygenic models. (PDF 67 kb)

Additional file 11: Comparison of different relatedness estimators. This document summarizes different methods for estimating relatedness, presents corresponding inflation results and shows the impact of small allele frequencies on relatedness estimates. (PDF 140 kb)

Abbreviations

CEU: CEPH (Centre d'Etude du Polymorphisme Humain) from Utah; CSV: Comma separated values; IBD: Identical by descent; IBS: Identical by state; SFS: Synthetic family study; SNP: Single nucleotide polymorphism

Acknowledgements

We thank very much Fabian Schwarzenberger for helpful comments and discussion. We thank all those who participated in the Sorbs study. Sincere thanks are given to Knut Krohn (Microarray Core Facility of the Interdisciplinary Centre for Clinical Research, University of Leipzig) for the genotyping support.

Funding

This publication is supported by the German Federal Research Ministry (BMBF), grant PROGRESS (01KI1010I) and by LIFE – Leipzig Research Center for Civilization Diseases, University of Leipzig. LIFE is funded by means of the European Union, by the European Regional Development Fund (ERDF) and by means of the Free State of Saxony within the framework of the excellence initiative. The Sorbs study was supported by grants from the Collaborative Research Center funded by the German Research Foundation (CRC 1052; C01, B01, B03, SPP 1629 TO 718/2), from the German Diabetes Association, from the DHFD (Diabetes Hilfs- und Forschungsfonds Deutschland) and from Boehringer Ingelheim Foundation. We acknowledge support from the German Research Foundation (DFG) and Leipzig University within the program of Open Access Publishing. Funding bodies were not involved in the design of the study or collection, analysis, and interpretation of data or in writing the manuscript.

Availability of data and materials

The HapMap data analysed during the current study are available at ftp://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/2010-08_phasell+III/. Sorbs study data are available from the authors upon reasonable request and with permission of the principal investigator (Prof. Dr. Michael Stumvoll).

Authors' contributions

AG developed the method, analysed the data and wrote the paper. AT designed the Sorbs study and collected the data. MS contributed to method development and wrote the paper. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Genotyping and metabolic phenotyping of the Sorbs study was approved by the ethics committee of the University of Leipzig and is in accordance with the declaration of Helsinki. All subjects gave written informed consent before taking part in the study. No database permissions were required for this study.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Haertelstrasse 16-18, 04107 Leipzig, Germany. ²LIFE - Leipzig Research Center for Civilization Diseases, University of Leipzig, Philipp-Rosenthal-Strasse 27, 04103 Leipzig, Germany. ³Department of Medicine, University of Leipzig, Liebigstrasse 18, 04103 Leipzig, Germany.

Received: 13 July 2017 Accepted: 23 November 2017

Published online: 06 December 2017

References

- Aulchenko YS, de Koning DJ, Haley C. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics*. 2007;177(1):577–85. doi:10.1534/genetics.107.075614.
- Boerwinkle E, Chakraborty R, Sing CF. The use of measured genotype information in the analysis of quantitative phenotypes in man. I, Models and analytical methods. *Ann Hum Genet*. 1986;50(Pt 2):181–94. doi:10.1111/j.1469-1809.1986.tb01037.x.
- Zhang YM, Mao Y, Xie C, Smith H, Luo L, Xu S. Mapping quantitative trait loci using naturally occurring genetic variance among commercial inbred lines of maize (*Zea mays* L.). *Genetics*. 2005;169(4):2267–75. doi:10.1534/genetics.104.033217.
- Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet*. 2006;38(2):203–8. doi:10.1038/ng1702.
- Amin N, van Duijn CM, Aulchenko YS. A genomic background based method for association analysis in related individuals. *PLoS ONE*. 2007;2(12):1274. doi:10.1371/journal.pone.0001274.
- Wang SB, Feng JY, Ren WL, Huang B, Zhou L, Wen YJ, Zhang J, Dunwell JM, Xu S, Zhang YM. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci Rep*. 2016;6:19444. doi:10.1038/srep19444.
- Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*. 2010;42(7):565–9. doi:10.1038/ng.608.
- Tonjes A, Scholz M, Breitfeld J, Marzi C, Grallert H, Gross A, Ladenvall C, Schleinitz D, Krause K, Kirsten H, Laurila E, Kriebel J, Thorand B, Rathmann W, Groop L, Prokopenko I, Isomaa B, Beutner F, Kratzsch J, Thiery J, Fasshauer M, Kloting N, Gieger C, Bluher M, Stumvoll M, Kovacs P. Genome wide meta-analysis highlights the role of genetic variation in RARRES2 in the regulation of circulating serum chemerin. *PLoS Genet*. 2014;10(12):1004854. doi:10.1371/journal.pgen.1004854.
- Belonogova NM, Svishcheva GR, van Duijn CM, Aulchenko YS, Axenovich TI. Region-based association analysis of human quantitative traits in related individuals. *PLoS ONE*. 2013;8(6):65395. doi:10.1371/journal.pone.0065395.
- Teyssedre S, Elsen JM, Ricard A. Statistical distributions of test statistics used for quantitative trait association mapping in structured populations. *Genet Sel Evol*. 2012;44:32. doi:10.1186/1297-9686-44-32.
- McArdle PF, O'Connell JR, Pollin TI, Baumgarten M, Shuldiner AR, Peyer PA, Mitchell BD. Accounting for relatedness in family based genetic association studies. *Hum Hered*. 2007;64(4):234–42. doi:10.1159/000103861.
- Devlin B, Roeder K. Genomic control for association studies. *Biometrics*. 1999;55(4):997–1004. doi:10.1111/j.0006-341X.1999.00997.x.
- Gross A, Tonjes A, Kovacs P, Veeramah KR, Ahnert P, Roshyara NR, Gieger C, Rueckert IM, Loeffler M, Stoneking M, Wichmann HE, Novembre J, Stumvoll M, Scholz M. Population-genetic comparison of the Sorbian isolate population in Germany with the German KORA population using genome-wide SNP arrays. *BMC Genet*. 2011;12:67. doi:10.1186/1471-2156-12-67.
- Aulchenko YS. Chapter 9 – Effects of Population Structure in Genome-wide Association Studies. In: *Analysis of Complex Disease Association Studies*. San Diego: Academic Press; 2011. p. 123–56. doi:10.1016/B978-0-12-375142-3.10009-4.
- Wang J. An estimator for pairwise relatedness using molecular markers. *Genetics*. 2002;160(3):1203–15.
- Stuart A, Ord K, Arnold S. *Kendall's Advanced Theory of Statistics* vol. 2A, 6th ed. 338 Euston Road, London NW1 3BH: Arnold, a member of the Hodder Headline Group; 1999.
- Czado C, Schmidt T. *Mathematische Statistik. Statistik und ihre Anwendungen*. Heidelberg: Springer; 2011. doi:10.1007/978-3-642-17261-8.
- HapMap. Merged phase I+II and III genotype files. ftp://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/2010-08_phasell+III/. Accessed 14 Mar 2017.
- Veeramah KR, Tonjes A, Kovacs P, Gross A, Wegmann D, Geary P, Gasperikova D, Klimes I, Scholz M, Novembre J, Stumvoll M. Genetic variation in the Sorbs of eastern Germany in the context of broader European genetic diversity. *Eur J Hum Genet*. 2011;19(9):995–1001. doi:10.1038/ejhg.2011.65.
- R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria. <https://www.R-project.org/>. Accessed 14 Mar 2017.
- Astle W, Balding DJ. Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statist Sci*. 2009;24(4):451–71. doi:10.1214/09-STS307.
- Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, Nordborg M. An *Arabidopsis* example of association mapping in structured samples. *PLoS Genet*. 2007;3(1):4. doi:10.1371/journal.pgen.0030004.
- Bacanu SA, Devlin B, Roeder K. Association studies for quantitative traits in structured populations. *Genet Epidemiol*. 2002;22(1):78–93. doi:10.1002/gepi.1045.
- Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet*. 2010;11(7):459–63. doi:10.1038/nrg2813.
- Zhou H, Blangero J, Dyer TD, Chan KK, Lange K, Sobel EM. Fast Genome-Wide QTL Association Mapping on Pedigree and Population Data. *Genet Epidemiol*. 2017;41(3):174–86. doi:10.1002/gepi.21988.
- Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet*. 2012;44(7):821–4. doi:10.1038/ng.2310.
- Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*. 2011;88(1):76–82. doi:10.1016/j.ajhg.2010.11.011.
- Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet*. 2014;46(2):100–6. doi:10.1038/ng.2876.
- Eu-Ahsunthornwattana J, Miller EN, Fakiola M, Jeronimo SM, Blackwell JM, Cordell HJ. Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLoS Genet*. 2014;10(7):1004445. doi:10.1371/journal.pgen.1004445.
- Maxima. A Computer Algebra System. Version 5.38.1. <http://maxima.sourceforge.net/>. Accessed 14 Mar 2017.

4 Bayesianischer Ansatz zur Berücksichtigung korrelierter Phänotypen

Eine SNP-Assoziationsanalyse kann neben klassischen Methoden auch mittels bayesianischer Methoden erfolgen. Bayesianische Methoden bieten dabei die Möglichkeit, SNP- und Phänotyp-Korrelationen zu berücksichtigen und so die Modellanpassung gegenüber der klassischen Analyse zu verbessern. Im Unterschied zu den vorangegangen Kapiteln wird aber im Folgenden nicht eine durch Verwandtschaft verursachte Phänotyp-Korrelation betrachtet, sondern die Korrelation von verschiedenen Phänotypen innerhalb eines Probanden. Am Beispiel einer Kinderstudie wird zunächst in einer klassischen SNP-Assoziationsanalyse nach einem Zusammenhang zwischen ausgewählten Kandidaten-Genen und Lipidkonzentrationen gesucht, um auf genetische Ursachen für Parameter des Stoffwechsels in der frühen Entwicklung schließen zu können. In einem zweiten Schritt wird eine bayesianische SNP-Assoziationsanalyse durchgeführt, dabei explizit die Phänotyp-Korrelation berücksichtigt und die Ergebnisse mit denen der klassischen Analyse verglichen. Die Ergebnisse sind in [3] publiziert und werden hier zusammengefaßt.

4.1 Beschreibung der Studie und genetischen Daten

In Metaanalysen von Erwachsenen-Kohorten wurden zahlreiche Gene identifiziert, die Lipidkonzentrationen beeinflussen [74–77]. Für Kinder und Heranwachsende sind jedoch nur wenige Studien verfügbar, in denen der genetischen Einfluß auf Lipidkonzentrationen untersucht wurde [78–81]. Da Adipositas ein Risikofaktor für ungünstige Lipidprofile ist [82], sind hierbei besonders genetische Varianten von Interesse, die nicht mit Adipositas in Verbindung stehen. Für eine Studie wurden deshalb Kinder aus der Region Leipzig rekrutiert. Bei den Kindern wurde unter anderem age (Alter), BMI SDS (body mass index standard deviation score), sex (Geschlecht) und Lipidkonzentrationen von HDL-C (high density lipoprotein cholesterol), LDL-C (low density lipoprotein cholesterol), TC (total cholesterol) und TG (triglyceride) bestimmt. Nach der QC standen 594 Kinder für eine SNP-Assoziationsanalyse zur Verfügung. Einige Probanden-Charakteristiken sind in *Table 1* der Publikation angegeben. Für die SNP-Assoziationsanalyse wurden Gene ausgewählt, für die in Metaanalysen von Erwachsenen-Kohorten Hinweise für einen Zusammenhang mit Lipidkonzentrationen gefunden wurden [75–77, 83, 84]. Ausgewählt wurden die sechs SNPs (Gene) rs599839 (SORT1), rs3846663 (HMGCR), rs3812316 (MLXIPL), rs174570 (FADS2), rs4420638 (APOE) und rs6102059 (MAFB) aufgrund verschiedener Kriterien wie beispielsweise Allelfrequenz oder Effektstärke. Einige SNP-Charakteristiken sind in *S3 Table* präsentiert.

4.2 Klassische Assoziationsanalyse

Um einen Zusammenhang der SNPs mit Adipositas auszuschließen, wurde im ersten Schritt BMI SDS mit jedem SNP und mit jedem der drei genetischen Modelle (additiv, dominant, rezessiv) getestet und dabei auf age und sex adjustiert. Es konnte kein signifikanter Zusammenhang von BMI SDS mit einem der SNPs identifiziert werden. Die Ergebnisse für das additive Modell sind in *Table 2* dargestellt, die ausführlichen Ergebnisse hingegen für alle drei genetischen Modelle befinden sich in *S1 Results*. Im zweiten Schritt wurde auf eine Assoziation zwischen jeder Lipidkonzentration und jedem SNP mit jedem der drei genetischen Modelle getestet. Für das additive Modell wurden signifikante Zusammenhänge gefunden für rs599839 (SORT1) mit TC ($\hat{\beta}_2 = -0,257; p = 1,50 \cdot 10^{-4}$) und LDL-C ($\hat{\beta}_2 = -0,3; p = 8,82 \cdot 10^{-6}$); rs4420638 (APOE) mit TC ($\hat{\beta}_2 = 0,336; p = 2,45 \cdot 10^{-5}$) und LDL-C ($\hat{\beta}_2 = 0,382; p = 1,38 \cdot 10^{-6}$). In Klammern angegeben sind der Schätzer für den genetischen Effekt aus Gl. (1.6) und der p-Wert aus Gl. (1.8). Alle Ergebnisse des additiven Modells sind in *Table 2* präsentiert. Mit den anderen genetischen Modellen wurden keine zusätzlichen signifikanten Zusammenhänge identifiziert (*S1 Results*).

4.3 Bayesianische Assoziationsanalyse

Ein Nachteil der klassischen Analyse ist, daß für jede Kombination von Phänotyp und SNP und jedes genetischen Modell einzeln getestet wird, was zu einer großen Zahl von Tests und unübersichtlichen Ergebnistabellen führt, wie an den Ergebnissen in *S1 Results* ersichtlich ist. Zudem ist eine geeignete Korrektur für multiples Testen notwendig. Details hierzu sind in der Publikation angegeben. Mit einer bayesianischen SNP-Assoziationsanalyse können diese Probleme umgangen werden, indem alle möglichen Modelle nach Plausibilität geordnet werden. Am Beispiel dieser Studie wird hierzu der kombinierte Phänotyp (Likelihood) von HDL-C, LDL-C und TG durch

$$(\text{HDL-C}, \text{LDL-C}, \text{TG}) \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

mit einer multivariaten Normalverteilung modelliert. TC ist im Modell nicht enthalten, weil es stark mit LDL-C korreliert ist. Jeder Mittelwert μ_i enthält die von einem Modell eingeschlossenen Variablen und Effekte, wie in Gl. (1.13) angegeben. Zur Auswahl stehen für jeden Phänotypen $c = 15$ Variablen ($c = 2q + r$), nämlich der dominante und rezessive Anteil von jedem der sechs SNPs ($q = 6$) und die drei Kovariablen age, BMI SDS und sex ($r = 3$). Durch die Modellierung des dominanten und rezessiven Anteils können verschiedene genetische Modelle berücksichtigt werden. Das Modell ist in *Figure 1* veranschaulicht. Zusätzlich wurde die SNP-Korrelationsstruktur mit einem Pseudo-Haplotyp Ansatz modelliert und fehlende SNP-Genotypen als Modellparameter definiert. Details zur Modellierung sind in *S1 Methods* beschrieben. Die Simulation, also das Ziehen aus den Posterior-Verteilungen der Parameter, erfolgte mit WinBUGS, wobei das zugehörige Skript in *S3 Methods* aufgeführt ist.

Analyse der Modelle

Das Hauptziel der Analyse ist, für jeden der drei Phänotypen die Modelle zu finden, bei denen die ausgewählten Variablen den jeweiligen Phänotyp möglichst gut unter Berücksichtigung der

Phänotyp-Korrelationsstruktur erklären. Für die Plausibilität eines Modells m wird dazu ein Bayesfaktor

$$\text{BF}(m) = \frac{\hat{P}(m|\mathcal{D})}{1 - \hat{P}(m|\mathcal{D})} (2^c - 1)$$

unter der Annahme bestimmt, daß alle Modelle zu Beginn gleichwahrscheinlich sind (Prior $P(m) = 2^{-c}$). $\hat{P}(m|\mathcal{D})$ ist die relative Häufigkeit (Posterior-Wahrscheinlichkeit) dafür, wie oft jedes Modell bei der Simulation beobachtet wurde. Die Herleitung des Bayesfaktors für Modelle ausgehend von Gl. (1.14) ist in *S2 Methods* beschrieben. In *Table 3* sind alle Modelle absteigend nach Plausibilität mit den zugehörigen Bayesfaktoren dargestellt. Die Liste ist bei einer kumulierten Häufigkeit von 95% abgeschnitten, um die weniger plausiblen Modelle wegzulassen. Die verbliebenen Modelle besitzen Bayesfaktoren > 100 und sind damit in ihrer Evidenz „entscheidend“. Die beiden plausibelsten Modelle von HDL-C enthalten keine genetischen Varianten, aber age und BMI SDS als Variablen. Das dritte Modell von HDL-C hingegen enthält den dominanten Anteil von rs4420638 (APOE). Das plausibelste Modell von TG enthält ebenso age und BMI SDS. Interessanterweise ist im zweiten Modell von TG aber zusätzlich der dominante Anteil von rs3812316 (MLXIPL) enthalten. Für LDL-C hingegen sind verschiedene Modelle plausibel, so sind die rezessiven Anteile von rs599839 (SORT1) und rs4420638 (APOE) sowie BMI SDS in den plausibelsten fünf Modellen enthalten. In weniger plausiblen Modellen von LDL-C sind verschiedene Kombinationen der rezessiven und dominanten Anteile von rs599839 und rs4420638 sowie age und BMI SDS enthalten. Auch der dominante Anteil von rs6102059 (MAFB) ist einmal bei LDL-C enthalten.

Analyse der Variablen

Der Einfluß jeder Variable kann unabhängig von einem bestimmten Modell betrachtet werden. Dazu wird die Einschlußwahrscheinlichkeit (Posterior-Wahrscheinlichkeit) bestimmt, mit der die Variable bei der Simulation in irgendeinem Modell eingeschlossen wird, wie in *Figure 2* dargestellt. Damit wird deutlich, daß neben BMI SDS auch rs599839 (SORT1) und rs4420638 (APOE) einen großen Einfluß auf LDL-C haben, was die Ergebnisse der klassischen Analyse bestätigt. Für die SNPs rs599839 und rs4420638 ist ein Zusammenhang mit Lipidkonzentrationen bekannt [74, 76, 85, 86]. Zudem gibt es einige Variablen mit kleinen Einschlußwahrscheinlichkeiten, für die deshalb ein Einfluß auf die jeweilige Lipidkonzentration nicht unplausibel ist, wie rs4420638 (APOE) auf HDL-C, rs3846663 (HMGCR) und rs6102059 (MAFB) auf LDL-C und rs3812316 (MLXIPL) auf TG. Diese genetischen Varianten eignen sich daher als Kandidaten für weitere Analysen. Einige Zusammenhänge lassen sich jedoch nur in Metaanalysen mit großen Fallzahlen aufdecken, wie der Zusammenhang von rs3846663 (HMGCR) mit LDL-C [75]. Für rs6102059 (MAFB) gibt es selbst für große Studien widersprüchliche Ergebnisse, so wurde in [75] ein Zusammenhang mit LDL-C aufgedeckt, jedoch nicht in [85]. Ähnliches gilt für rs3812316 (MLXIPL), bei dem ein Zusammenhang mit TG in [83] identifiziert wurde, aber nicht in [87, 88], wobei in [87] vermutet wurde, daß der Effekt von rs3812316 (MLXIPL) auf TG schwach sein muß, falls er überhaupt existiert. Im Gegensatz dazu wurde für rs174570 (FADS2) weder in der klassischen noch in der bayesianischen Analyse ein Zusammenhang mit Lipidkonzentrationen gefunden, was aber daran liegen kann, daß die Fallzahl gegenüber den großen Metaanalysen wie [77] zu gering

ist, um schwache Effekte überhaupt finden zu können.

Analyse der Effekte

Die geschätzten Effekte (Mittelwerte der Posterior-Verteilungen) sind für alle Variablen mit einer Einschlußwahrscheinlichkeit von größer als 0,5% in *Table 4* angegeben. Die Effekte und ihre Standardabweichungen wurden dazu über alle Modelle, in der die betreffende Variable beobachtet wurde, durch BMA gemittelt. Die geschätzten Effekte aus *Table 4* der bayesianischen Analyse ähneln denen der klassischen Analyse, die für das dominante und rezessive Modell in *S1 Results* aufgeführt sind. Im Vergleich zur klassischen Analyse sind die meisten Standardabweichungen aber kleiner (*S1 Figure*), was dadurch zu erklären ist, daß sich die Effekte in der bayesianischen Analyse nicht auf ein einzelnes Modell beziehen und bei der Analyse die Phänotyp-Korrelationsstruktur mit einbezogen wurde.

4.4 Zusammenfassung

Mittels einer Kinderstudie wurde nach dem Einfluß bestimmter SNPs ausgewählter Kandidaten-Gene (SORT1, HMGCR, MLXIPL, FADS2, APOE, MAFB) auf Lipidkonzentrationen von HDL-C, LDL-C, TC und TG gesucht, um auf genetische Ursachen für Parameter des Stoffwechsels in der frühen Entwicklung schließen zu können. Zunächst wurde eine klassische SNP-Assoziationsanalyse durchgeführt, bei der ein Zusammenhang von SORT1 und APOE mit LDL-C und TC identifiziert wurde. Darauf wurde in einer bayesianischen Analyse der mehrdimensionale Phänotyp aus HDL-C, LDL-C und TG modelliert, wodurch die Phänotyp-Korrelationsstruktur berücksichtigt wurde. Für die einzelnen Lipidkonzentrationen konnte eine plausible Auswahl von Einflussfaktoren bestehend aus genetischen Varianten, Alter, Geschlecht und BMI unter Berücksichtigung verschiedener genetischer Modelle bestimmt werden. Dadurch wurden sowohl die Ergebnisse aus der klassischen Analyse bestätigt, als auch weitere Kandidaten, beispielsweise ein Zusammenhang zwischen MLXIPL und TG, gefunden. Vorteile der Bayesianischen Analyse im Vergleich zur klassischen Analyse sind einerseits die verbesserte Identifikation von Phänotyp-Genotyp-Beziehungen bei korrelierten Phänotypen als auch die Präsentation der Ergebnisse in verständlicher Form.

4.5 Publikation

In diesem Abschnitt befindet sich eine Kopie der Publikation von: C. Breitling, A. Gross, P. Buttner, et al. Genetic Contribution of Variants near SORT1 and APOE on LDL Cholesterol Independent of Obesity in Children. *PLoS ONE*, 10(9):e0138064, 2015. doi:10.1371/journal.pone.0138064.

RESEARCH ARTICLE

Genetic Contribution of Variants near *SORT1* and *APOE* on LDL Cholesterol Independent of Obesity in Children

Clara Breitling¹*, Arnd Gross^{2,3}*, Petra Büttner¹, Sebastian Weise¹, Dorit Schleinitz⁴, Wieland Kiess¹, Markus Scholz^{2,3}, Peter Kovacs⁴, Antje Körner^{1,4*}

1 Center for Pediatric Research (CPL), University Hospital for Children and Adolescents Leipzig, Dept. of Women's & Child Health, University of Leipzig, Leipzig, Germany, **2** Institute for Medical Informatics, Statistics and Epidemiology (IMISE), University of Leipzig, Leipzig, Germany, **3** LIFE-Leipzig Research Center for Civilization Diseases, University of Leipzig, Leipzig, Germany, **4** Integrated Research and Treatment Center Adiposity Diseases (IFB), Medical Faculty, University of Leipzig, Leipzig, Germany

* These authors contributed equally to this work.

* Antje.Koerner@medizin.uni-leipzig.de



OPEN ACCESS

Citation: Breitling C, Gross A, Büttner P, Weise S, Schleinitz D, Kiess W, et al. (2015) Genetic Contribution of Variants near *SORT1* and *APOE* on LDL Cholesterol Independent of Obesity in Children. PLoS ONE 10(9): e0138064. doi:10.1371/journal.pone.0138064

Editor: Yvonne Böttcher, University of Leipzig, GERMANY

Received: September 19, 2014

Accepted: August 25, 2015

Published: September 16, 2015

Copyright: © 2015 Breitling et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data is available in the paper and its Supporting Information files.

Funding: This work was supported by grants from the German Research Council (DFG) for the Clinical Research Center "Obesity Mechanisms" CRC1052/1 C05 and the Federal Ministry of Education and Research (BMBF), Germany, FKZ: 01EO1001 (IFB AdiposityDiseases), and the European Community's Seventh Framework Programme (FP7/2007–2013) project Beta-JUDO under grant agreement n° 279153 to A.K. The work of A.G. was supported in part by the

Abstract

Objective

To assess potential effects of variants in six lipid modulating genes (*SORT1*, *HMGCR*, *MLXIPL*, *FADS2*, *APOE* and *MAFB*) on early development of dyslipidemia independent of the degree of obesity in children, we investigated their association with total (TC), low density lipoprotein (LDL-C), high density lipoprotein (HDL-C) cholesterol and triglyceride (TG) levels in 594 children. Furthermore, we evaluated the expression profile of the candidate genes during human adipocyte differentiation.

Results

Expression of selected genes increased 10^1 to $>10^4$ fold during human adipocyte differentiation, suggesting a potential link with adipogenesis. In genetic association studies adjusted for age, BMI SDS and sex, we identified significant associations for rs599839 near *SORT1* with TC and LDL-C and for rs4420638 near *APOE* with TC and LDL-C. We performed Bayesian modelling of the combined lipid phenotype of HDL-C, LDL-C and TG to identify potentially causal polygenic effects on this multi-dimensional phenotype and considering obesity, age and sex as a-priori modulating factors. This analysis confirmed that rs599839 and rs4420638 affect LDL-C.

Conclusion

We show that lipid modulating genes are dynamically regulated during adipogenesis and that variants near *SORT1* and *APOE* influence lipid levels independent of obesity in children. Bayesian modelling suggests causal effects of these variants.

German Federal Research Ministry (BMBF), grant PROGRESS (01KI1010). M.S. was funded by LIFE – Leipzig Research Center for Civilization Diseases, University of Leipzig. LIFE is funded by means of the European Union, by the European Regional Development Fund (ERDF) and by means of the Free State of Saxony within the framework of the excellence initiative. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: A.K. declares on behalf of all authors there are no competing interests that could be perceived to bias their work, acknowledging all financial support and any other relevant financial or non-financial competing interests.

Abbreviations: *APOE*, apolipoprotein E; *TC*, total-cholesterol; *FADS2*, fatty acid desaturase 2; *GWA*, genome-wide association; *HMGCR*, 3-hydroxy-3-methylglutaryl-Coenzyme A reductase; *MAFB*, V-maf musculoaponeurotic fibrosarcoma oncogene homolog B; *MLXIPL*, *MLX interacting protein*; *SORT1*, sortilin 1.

Introduction

Alterations in blood lipid phenotypes culminating in dyslipidemia are important risk factors for the development of cardiovascular disease [1]. Elevated blood low-density lipoprotein cholesterol (LDL-C) and triglycerides (TG) are strongly related to the likelihood of existing or future coronary heart disease [2, 3], whereas elevated blood high-density lipoprotein cholesterol (HDL-C) has a protective effect [4]. The most common cause of dyslipidemia is obesity [5]. However, a relevant proportion of patients with elevated blood lipid levels does not show an abnormal BMI [6].

Meta-analyses of genome-wide association (GWA) studies revealed many genetic loci influencing blood lipid levels underlying the polygenic cause of dyslipidemia and thereby identified suspected as well as unsuspected new candidate genes [7–10]. However, these meta-analyses concern adult cohorts. So far, there are only very few data on selected genes associated with altered blood lipid phenotypes in children and adolescents [11–14].

Investigation of childhood cohorts has several advantages though. They are much less biased by chronic disease and treatments but already show considerable heterogeneity regarding blood lipid levels. Also, considering future prediction of developing dyslipidemia, it is important to assess whether associations between genetic variants and blood lipid phenotypes observed in adults are already evident in children and adolescents [15]. Due to the lower influence of co-morbidities and other life-style related factors, we suppose that primary genetic effects are stronger in children than in adults. Thus, we hypothesize that we can detect at least some of the variants even with the lower number of individuals available for childhood cohorts.

In the present study, we aimed at assessing associations of six variants with lipid traits in a sample of mainly obese children. Selected variants are located in or near the genes *SORT1* (sortilin 1), *HMGCR* (3-hydroxy-3-methylglutaryl-Coenzyme A reductase), *MLXIPL* (*MLX interacting protein*), *FADS2* (fatty acid desaturase 2), *APOE* (apolipoprotein E) and *MAFB* (V-maf musculoaponeurotic fibrosarcoma oncogene homolog B) for which high effect-sizes regarding lipid phenotypes were reported. Going beyond classical association analysis, we additionally performed a Bayesian modelling approach to identify unconfounded relationships between genetic and non-genetic covariates and lipid phenotypes. Considering that obesity is a risk factor for dyslipidemia *per se* and that adipose tissue is an important tissue for lipid metabolism, we also assessed a potential relationship of the candidate gene expression for adipogenesis by studying time-series of gene-expression during human adipocyte differentiation.

Methods

Selection of Candidate Genes and Variants

Genes were selected according to evidence of genotype-phenotype-associations established in meta-analyses of adult cohorts [8–10, 16, 17]. We prioritized genetic variants by applying a score integrating (i) GWAS for lipid genes and obesity (p-value), (ii) gene expression data from adipocytes, (iii) minor allele frequency and effect size, (iv) verification in replication analyses. Based on these criteria, we selected six variants rs6102059 (*MAFB*), rs4420638 (*APOE*), rs599839 (*SORT1*), rs3846663 (*HMGCR*), rs174570 (*FADS2*) and rs3812316 (*MLXIPL*). We excluded well-known SNPs in genes like LDL-receptor because we were interested in new candidate genes influencing blood lipid levels.

Cis-eQTL effects of SNPs in linkage disequilibrium with our variants were observed for *SORT1* [18–20], *HMGCR* [21] and *FADS2* [21–23].

Sample

A total of 683 children were recruited from the Leipzig area via our out-patient obesity clinic. We applied German reference data for the calculation of the SDS as suggested by the National German Guidelines for Pediatric Obesity [24]. Obesity was defined as BMI SDS >1.88 corresponding to the 97th percentile.

White children were phenotyped by age, sex, height, weight, pubertal state, laboratory parameters and other clinical characteristics. Assessment of pubertal stage was performed by clinical examination according to Tanner [25, 26]. Blood lipid levels (triglycerides, total cholesterol, HDL and LDL) were determined with direct enzymatic colorimetric assays by the certified laboratory at the Institute of Laboratory Medicine, Clinical Chemistry and Molecular Diagnosis at the University of Leipzig. Written informed consent was obtained from all parents and from participants ≥12 years of age. This study has been approved by the ethics committee of the University of Leipzig and has been conducted according to the principles expressed in the Declaration of Helsinki (October 2000).

We excluded children with chronic inflammatory diseases, metabolic diseases, genetic disorders and diseases that required medication influencing lipid metabolism (N = 89).

For the remaining 594 children, data on glucose metabolism and lipid phenotypes (TC, LDL-C, HDL-C, TG) were available. Anthropometric and metabolic characterisation of included samples is presented in [Table 1](#).

Gene expression analysis during human adipocyte differentiation

Gene expression profiles of selected genes were determined via qRT-PCR for human preadipocyte SGBS (Simpson-Golabi-Behmel syndrome) cells during differentiation into mature adipocytes. Adipocyte differentiation was induced as described previously [27].

RNA extraction was performed using the RNeasy MiniKit (Qiagen, Hilden, Germany) including DNase digestion according to the manufacturer's instructions. Reverse transcription of 50 ng/μl RNA was carried out using the M-MLV Reverse Transcriptase Kit (Invitrogen, Karlsruhe, Germany) with random hexamer [p(dN)₆] primers (Roche, Basel, Switzerland). Primer sequences are provided in [S1 Table](#).

Experiments were performed in three distinct experiments each in triplicates. Target gene-expression was normalized to the averaged expression of three housekeeping genes *TBP*, *HPRT* and *USF2*.

DNA-Isolation and genotyping

Fasting venous EDTA blood samples were stored at -80°C. After washing with phosphate buffered saline, erythrocyte depletion by NH₄-lysis, and centrifugation, we extracted DNA using QIAamp DNA Blood MiniKit (Qiagen) according to the manufacturer's manual.

Table 1. Anthropometric and metabolic characterization of study samples.

Sex (male / female)	277 / 317
n (non-obese / obese)	122 / 472
BMI SDS	2.39 (0.85)
Age (years)	11.67 (5.23)
Total Cholesterol (mmol/L)	4.08 (0.99)
HDL-C (mmol/L)	1.22 (0.37)
LDL-C (mmol/L)	2.46 (0.89)
Triglyceride (mmol/L)	0.99 (0.70)

Quantitative variables are presented as median (interquartile range). Obesity is defined as BMI SDS>1.88.

doi:10.1371/journal.pone.0138064.t001

Genotyping probes and primers were obtained from Applied Biosystems (Darmstadt, Germany). Primer sequences are listed in [S2 Table](#). We used qPCR MasterMix Kit for probe Assay and Low Rox Plus (Eurogentec, Köln, Germany) for genotyping according to the manufacturers' manuals. Genotyping was performed on ABI Prism 7500 sequence detector (Applied Biosystems, Lincoln, USA). At least 5% of all samples were re-assessed on a different plate with concordance rate of 100%. Genotype frequencies of all SNPs were consistent with Hardy-Weinberg equilibrium. SNP characteristics are presented in [S3 Table](#).

Classical statistical analysis

In classical analysis of genetic data, every combination of a single SNP and a single phenotype is tested for association. Prior to analysis, lipid phenotypes TC, LDL-C, HDL-C and TG were log transformed to approximate normal distribution. Continuous phenotypes TC, LDL-C, HDL-C, TG, age and BMI SDS were standardised to zero mean and unit variance before analysis in order to obtain dimensionless effect estimates which are better comparable between different predictors and studies.

SNP associations with BMI SDS were tested with a linear model assuming three different genetic models, an additive effect of both alleles, and a dominant and recessive effect of the major allele, respectively. All models were adjusted for age and sex. Similarly, lipid SNP associations were tested with a linear model and three different genetic models. All models were adjusted for age, BMI SDS and sex. Adjustment for pubertal state instead of age is also reasonable. Due to the high correlation of age and pubertal state (Spearman $r = 0.84$), the genetic results are essentially the same (not shown). Also, pubertal stage is assessed by two parameters (pubic hair and breast development or testicular volume), which are not necessarily coherent. Furthermore, pubertal timing differs between boys and girls. Since dyslipidemia would be more related to age as an indicator of duration of obesity and dyslipidemia it is not necessarily influenced by pubertal development per se, this was another reason to adjust for age. We, therefore, decided to use the continuous and less ambiguously measurable trait age instead of pubertal state.

Since we tested five phenotypes, six SNPs and three genetic models, it is necessary to correct for multiple testing. However, due to multiple correlations between phenotypes and effects of genetic models, it was necessary to simulate the null-distribution. In our situation, a significance level of 6.7×10^{-4} controls the family-wise error rate at 5% and was therefore used to correct for multiple testing in our study.

Statistical analysis was performed using R 2.10.1.

Bayesian Model Analysis

The major drawbacks of the classical analysis mentioned above are the large number of tests to be performed due to the large number of possible combinations of SNPs and phenotypes and the assumption of a specific model of genetic and non-genetic effects. To overcome these limitations, we performed Bayesian model analysis in addition to our classical association analysis. By this approach, we can estimate plausibilities of different models and corresponding effect sizes. Bayesian modelling also allows some kind of causal inference by analysing all lipid phenotypes and possible influencing factors in parallel considering their overall correlation structure. To some extent, this avoids spurious associations.

The method is well conceived with application in analysing complex genotype-phenotype associations in medical research [[28–30](#)]. Additional insights can be derived from the modelling such as probability of different genetic risk models and estimates of unconfounded effects

considering all dependencies between variables of interest. It also circumvents the above mentioned issue of multiple-testing and the uncertainty regarding the model of inheritance.

Similar to the univariate analysis, transformed and standardised data were used. Lipid phenotypes were modelled with the Bayesian variable selection approach described in [29, 31] using the reversible jump interface of WinBUGS (Version 1.4.3). Since correlation of TC and LDL-C is very high ($r = 0.91$) we studied models of the (three-dimensional) lipid phenotype HDL-C, LDL-C and TG. We aimed to identify the most plausible sets of co-variables having a direct influence on each lipid phenotype accounting for correlations between them.

In our analysis, the set of co-variables consists of age, BMI SDS, sex and a recessive and a dominant part for each of the six SNPs defined by indicator variables “genotype” = 0 and “genotype” = 2 respectively. If both indicator variables are included, different levels of co-dominance can be expressed by corresponding effect estimates. Hence, 15 co-variables were available for selection for each of the 3 lipid phenotypes.

Each different subset of these co-variables forms a model. Prior to analysis, one assumes that all models are equally likely. We calculated Bayesian posterior probabilities representing the plausibilities of the models given our data. Details of Bayesian modelling and fitting can be found in [S1 Methods](#).

Bayes factors [32] are used to interpret model results. Calculation of Bayes factors is explained in the [S2 Methods](#). A usual convention is that a Bayes factor of 1 to 3.2 is judged as “not worth more than a bare mention”, a factor of 3.2 to 10 as “substantial”, a factor of 10 to 100 as “strong” and a factor greater than 100 as “decisive” evidence for a model or effect [33]. Conversely, reciprocal values represent counter-evidence for a model. Rather than deciding whether a certain covariate has an effect or not (i.e. is in the model or not), we calculate corresponding inclusion probabilities, which can be interpreted as plausibilities regarding the impact of the covariate on the phenotype considered. Effect estimates of co-variables can be determined in the Bayesian context by averaging over all models containing this co-variable (Bayesian model averaging) weighted by the plausibility of the model. Results can be considered as analogons to Beta-coefficients of classical linear regression analysis.

Results

Classical genotype-phenotype analyses for BMI SDS and lipid phenotypes

There was no significant association between BMI SDS and any of the selected SNPs indicating that the variants are not related to the degree of obesity in our data. Results for the additive model are presented in [Table 2](#). Results for all three genetic models are given in [S1 Results](#).

Next, we analysed the association between genotypes and lipid phenotypes. We found significant associations with lipid phenotypes for *SORT1* rs599839 with TC ($p = 1.50 \times 10^{-4}$, $\beta = -0.257$) and LDL-C ($p = 8.82 \times 10^{-6}$, $\beta = -0.3$) as well as for *APOE* rs4420638 with TC ($p = 2.45 \times 10^{-5}$, $\beta = 0.336$) and LDL-C ($p = 1.38 \times 10^{-6}$, $\beta = 0.382$), whereas the variants were not associated with other lipid phenotypes ([Table 2](#)). No additional associations were found when investigating alternative models of inheritance (see [S1 Results](#)).

Bayesian model analysis

We performed Bayesian modelling of the multi-phenotype of HDL-C, LDL-C and TG. TC was not included into the model due to its strong correlation with LDL-C. Analysed relations are illustrated in [Fig 1](#). In the following, we present the most plausible models of each lipid phenotype accounting for their pairwise correlations. The corresponding WinBUGS Model is given

in detail in [S3 Methods](#). Most probable models in decreasing order of plausibility and corresponding Bayes factors are shown in [Table 3](#). The lists are truncated when the cumulative probability of the models exceeds 95%, i.e. all other models are less plausible according to our data. Both top models of HDL-C contain no genetic factors but age and BMI SDS as co-variables. The third most probable model includes the dominant part of rs4420638 (*APOE*).

The top models of TG contain age and BMI SDS, too. Additionally, the second best model includes the dominant part of rs3812316 (*MLXIPL*).

Various different models are plausible for LDL-C: The recessive parts of SNP rs599839 (*SORT1*) and rs4420638 (*APOE*) and BMI SDS contribute to the top 5 models of LDL-C. In less probable models for LDL-C, combinations of the recessive and dominant parts of rs599839 and rs4420638, age and BMI SDS occur. Further, the dominant part of rs6102059 (*MAFB*) is included once.

Table 2. Association of genotypes with BMI SDS and lipid phenotypes.

Phenotype	Variant	N	Beta	SE	CI	p-value
BMI SDS	rs599839	576	-0.087	0.066	[-0.217;0.044]	0.193
BMI SDS	rs3846663	572	0.076	0.061	[-0.045;0.196]	0.219
BMI SDS	rs3812316	564	-0.126	0.095	[-0.312;0.06]	0.184
BMI SDS	rs174570	578	0.176	0.09	[-0.001;0.353]	0.052
BMI SDS	rs4420638	584	-0.004	0.079	[-0.16;0.152]	0.958
BMI SDS	rs6102059	575	0.062	0.065	[-0.065;0.19]	0.335
TC	rs599839	576	-0.257	0.067	[-0.389;-0.125]	1.50x10⁻⁴
TC	rs3846663	572	0.141	0.062	[0.019;0.263]	0.024
TC	rs3812316	564	-0.022	0.094	[-0.207;0.162]	0.812
TC	rs174570	578	-0.04	0.092	[-0.22;0.14]	0.662
TC	rs4420638	584	0.336	0.079	[0.181;0.491]	2.45x10⁻⁵
TC	rs6102059	575	0.019	0.065	[-0.109;0.146]	0.775
HDL-C	rs599839	576	0.077	0.067	[-0.054;0.207]	0.25
HDL-C	rs3846663	572	0.098	0.061	[-0.021;0.217]	0.106
HDL-C	rs3812316	564	0.129	0.093	[-0.054;0.312]	0.168
HDL-C	rs174570	578	-0.078	0.09	[-0.254;0.098]	0.387
HDL-C	rs4420638	584	-0.13	0.078	[-0.283;0.024]	0.098
HDL-C	rs6102059	575	0.038	0.064	[-0.087;0.164]	0.547
LDL-C	rs599839	576	-0.3	0.067	[-0.431;-0.168]	8.82x10⁻⁶
LDL-C	rs3846663	572	0.12	0.062	[-0.002;0.241]	0.054
LDL-C	rs3812316	564	-0.043	0.094	[-0.226;0.141]	0.649
LDL-C	rs174570	578	0.021	0.091	[-0.158;0.2]	0.817
LDL-C	rs4420638	584	0.382	0.078	[0.228;0.536]	1.38x10⁻⁶
LDL-C	rs6102059	575	0.018	0.065	[-0.109;0.145]	0.781
TG	rs599839	576	-0.114	0.065	[-0.241;0.014]	0.081
TG	rs3846663	572	-0.006	0.059	[-0.123;0.11]	0.913
TG	rs3812316	564	-0.134	0.088	[-0.307;0.038]	0.127
TG	rs174570	578	0.137	0.087	[-0.034;0.307]	0.116
TG	rs4420638	584	0.135	0.076	[-0.014;0.285]	0.076
TG	rs6102059	575	0.075	0.062	[-0.046;0.197]	0.225

We present numbers of cases available for the corresponding analysis (N), beta-coefficients, their standard errors (SE), 95% confidence intervals (CI) and uncorrected p-values. Since standardized values were analysed, beta-coefficients and standard errors have unit 1. BMI SDS was analysed with the additive model adjusted for age and sex. Lipid phenotypes were logarithmized and analysed with the additive model adjusted for age, sex and BMI SDS. Associations significant after correction for multiple testing (see [methods](#) section) are printed in bold.

doi:10.1371/journal.pone.0138064.t002

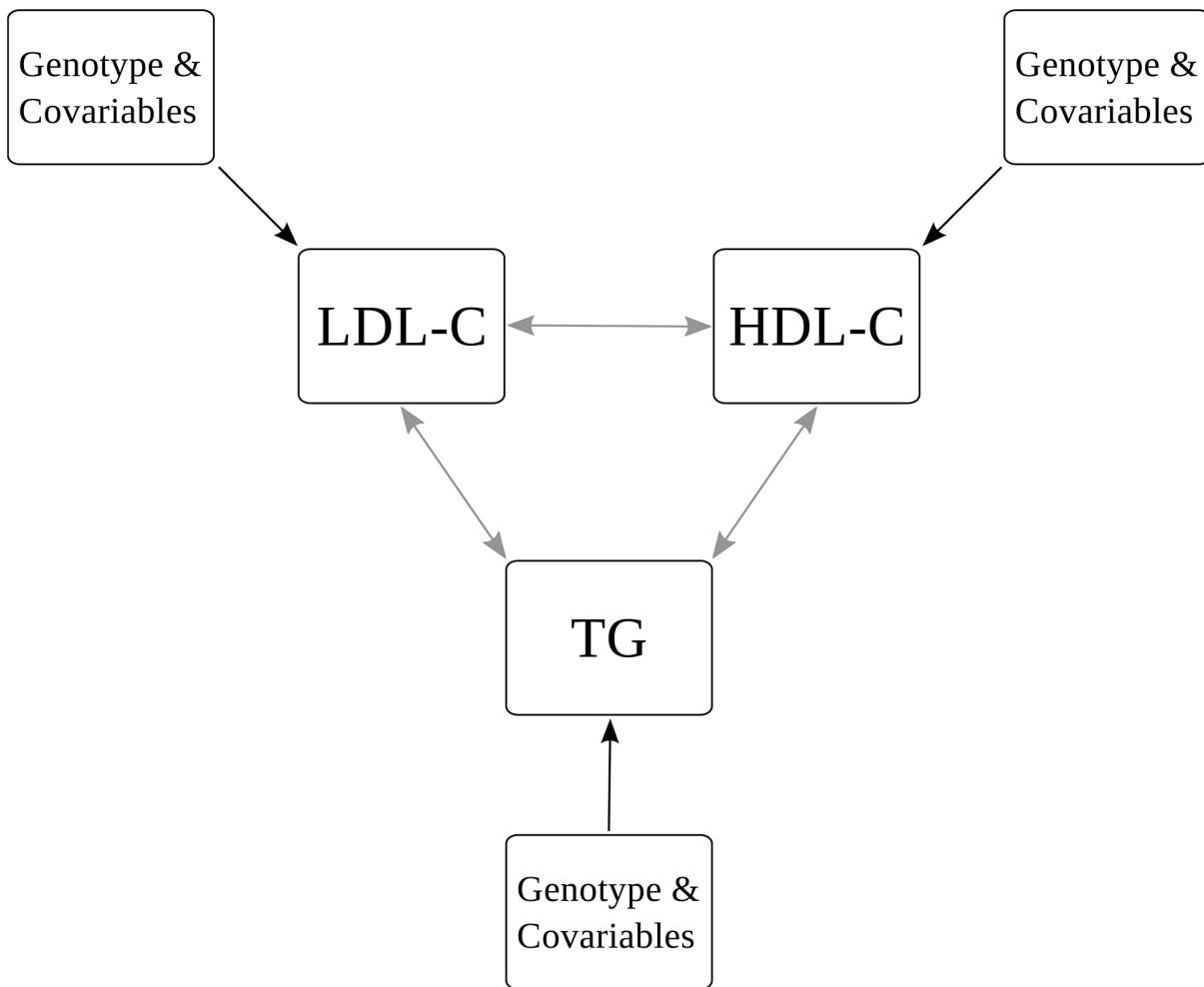


Fig 1. Bayesian Model. We present the structure of the Bayesian model analysed. Black arrows represent possible impacts of considered covariates (SNPs, age, BMI SDS, sex) on the distribution means of lipid phenotypes. Grey arrows refer to the covariance between the lipids which is accounted for in the model.

doi:10.1371/journal.pone.0138064.g001

The impact of each co-variable independent of a certain model can be assessed by interpreting the inclusion probabilities for co-variables (Fig 2). In addition to the apparent and expected impact of BMI SDS, rs599839 (*SORT1*) and rs4420638 (*APOE*) have a high certainty of affecting LDL-C independent of the degree of obesity. Conversely, the following effects cannot be ruled out (i.e. no decisive evidence against the effect was found): rs4420638 (*APOE*) on HDL-C, rs3846663 (*HMGCR*) and rs6102059 (*MAFB*) on LDL-C, rs3812316 (*MLXIPL*) on TG.

Effect estimates of co-variables with inclusion probability greater than 0.5% are listed in Table 4. Estimates and standard deviations are averaged over all models, where the respective co-variable is included (Bayesian model averaging). In comparison to classical analysis, the majority of standard deviations of estimates are smaller demonstrating higher power of the Bayesian model approach (S1 Fig).

The estimated covariance of the model is shown in S2 Results. Results of the combined model of TC, HDL-C, TG are similar to those of the model of LDL-C, HDL-C, TG considered here (data not shown).

Table 3. Results of Bayesian model analysis.

Lipid	Model	Probability	Bayes factor
HDL-C	BMI SDS	91.89	371265
HDL-C	age, BMI SDS	3.08	1041
HDL-C	rs4420638 _{dom} , BMI SDS	0.99	329
LDL-C	rs599839 _{rec} , rs4420638 _{rec}	53.49	37691
LDL-C	rs599839 _{rec} , rs4420638 _{rec} , BMI SDS	22.88	9720
LDL-C	rs4420638 _{rec}	7.65	2714
LDL-C	rs4420638 _{rec} , BMI SDS	4.62	1586
LDL-C	rs599839 _{rec}	2.54	855
LDL-C	rs599839 _{dom} , rs4420638 _{rec}	1.03	340
LDL-C	rs599839 _{rec} , rs4420638 _{rec} , age	0.8	266
LDL-C	rs599839 _{rec} , rs4420638 _{rec} , rs6102059 _{dom}	0.77	254
LDL-C	rs599839 _{rec} , BMI SDS	0.74	244
LDL-C	null	0.56	186
TG	age, BMI SDS	90.47	311171
TG	rs3812316 _{dom} , age, BMI SDS	3.66	1247
TG	BMI SDS	2.55	856

Possible models of HDL-C, LDL-C, TG can contain up to 15 covariables (age, sex, BMI SDS, dominant and recessive effect of six SNPs). We present most probable models, corresponding posterior probabilities and Bayes factors. Models are ranked according to their plausibility. A cumulative probability of 95% served as cut-off for model presentation.

doi:10.1371/journal.pone.0138064.t003

Polygenic effects for LDL-C are illustrated in [S2 Fig](#).

Gene-expression analysis in human adipocyte precursors

We measured gene expression of *SORT1*, *HMGCR*, *MLXIPL*, *FADS2*, *APOE* and *MAFB* during differentiation of human preadipocytes into adipocytes to assess a potential physiological relevance in lipid metabolism. We observed an up-regulation of these genes by magnitudes of 10 to 10^4 ([Fig 3](#)).

Marginal Inclusion Probabilities in %

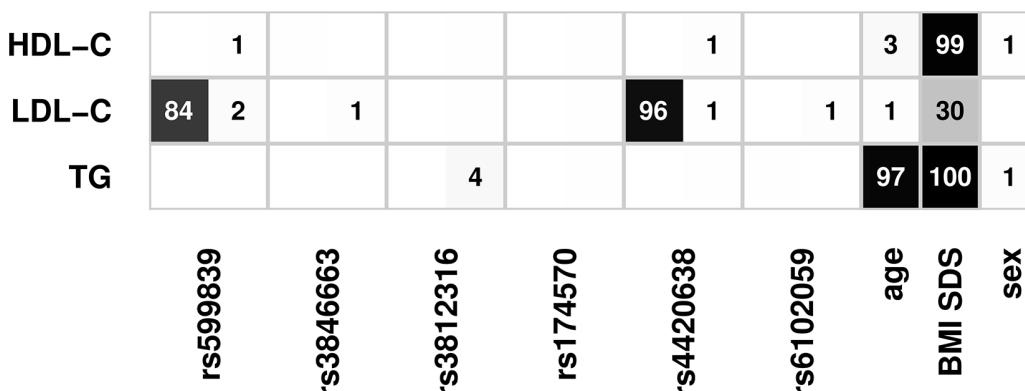


Fig 2. Inclusion probabilities of covariables for each lipid phenotype. For each SNP, results are given for the recessive (first number) and dominant part (second number). Results for inclusion probabilities are rounded to integers of percentage. Effect estimates are illustrated by the shade of grey as indicated. Results rounded to zero are omitted. Results for the lipid phenotypes LDL-C, HDL-C and TG are presented. TC is omitted due to high correlation with LDL-C.

doi:10.1371/journal.pone.0138064.g002

Table 4. Inclusion probabilities of covariates and Bayesian effect sizes.

Lipid	Variant	Probability	Estimate	SD
HDL-C	rs599839 _{dom}	0.6	0.253	0.178
HDL-C	rs4420638 _{dom}	1.03	-0.415	0.25
HDL-C	age	3.22	-0.127	0.041
HDL-C	BMI SDS	99.34	-0.21	0.041
HDL-C	sex	0.53	-0.147	0.072
LDL-C	rs599839 _{rec}	84.3	0.32	0.077
LDL-C	rs599839 _{dom}	2.2	-0.415	0.171
LDL-C	rs3846663 _{dom}	1.16	0.258	0.114
LDL-C	rs4420638 _{rec}	95.57	-0.365	0.081
LDL-C	rs4420638 _{dom}	0.58	0.347	0.239
LDL-C	rs6102059 _{dom}	1.23	0.276	0.134
LDL-C	age	1.18	-0.12	0.042
LDL-C	BMI SDS	30	0.146	0.04
TG	rs3812316 _{dom}	3.81	-0.757	0.346
TG	age	97.28	0.172	0.035
TG	BMI SDS	99.98	0.255	0.044
TG	sex	1.45	-0.166	0.061

We present probabilities for inclusion of specified covariates and resulting effect sizes and corresponding standard deviations (SD) averaged over all models containing the covariate. Only covariates with an inclusion probability greater than 0.5% are shown.

doi:10.1371/journal.pone.0138064.t004

Discussion

In this study, we aimed to assess the relevance of SNPs showing associations with lipid phenotypes from adults in a childhood sample which is less prone to confounding factors such as medication and co-morbidities and has shorter exposure to endogenous and exogenous factors.

Considering the strong impact of obesity and hence adipose tissue on circulating lipid phenotypes, we were also interested whether the candidate genes are dynamically regulated during adipogenesis. We have previously shown that genetic candidates from GWAS for obesity traits may have a functional role in human adipogenesis [27]. All selected genes from this study were expressed in adipocytes and showed considerable up-regulations during human adipocyte differentiation up to 10,000 fold. This has not been shown for these genes before. Even though this dynamic regulation during adipogenesis does not directly imply a functional relationship, this finding merits further investigation in mechanistic studies. We evaluated the dynamic regulation of candidate gene expression in SGBS preadipocytes, so far the only established model of human preadipocyte differentiation [34], which is widely applied in adipogenesis research. It has been shown that biology and molecular markers are comparable to primary human adipocyte differentiation and circumvents potential bias by patient heterogeneity due to age, risk factors, morbidities, treatment etc.

Considering the strong dependence of lipid levels on obesity, the observed regulations may affect serum lipid phenotypes and may explain SNP associations. We, therefore, verified that the six variants considered were not associated with the degree of obesity in the children prior to evaluation of associations with lipid traits.

Of the six selected SNPs located in or near the genes *FADS2* (rs174570), *MAFB* (rs6102059), *HMGCR* (rs3846663), *MLXIPL* (rs3812316), *APOE/C1/C4/C2* (rs4420638) and *CELSR2/PSRC1/SORT1* (rs599839), we observed a strong impact of rs599839 (*SORT1*) and rs4420638

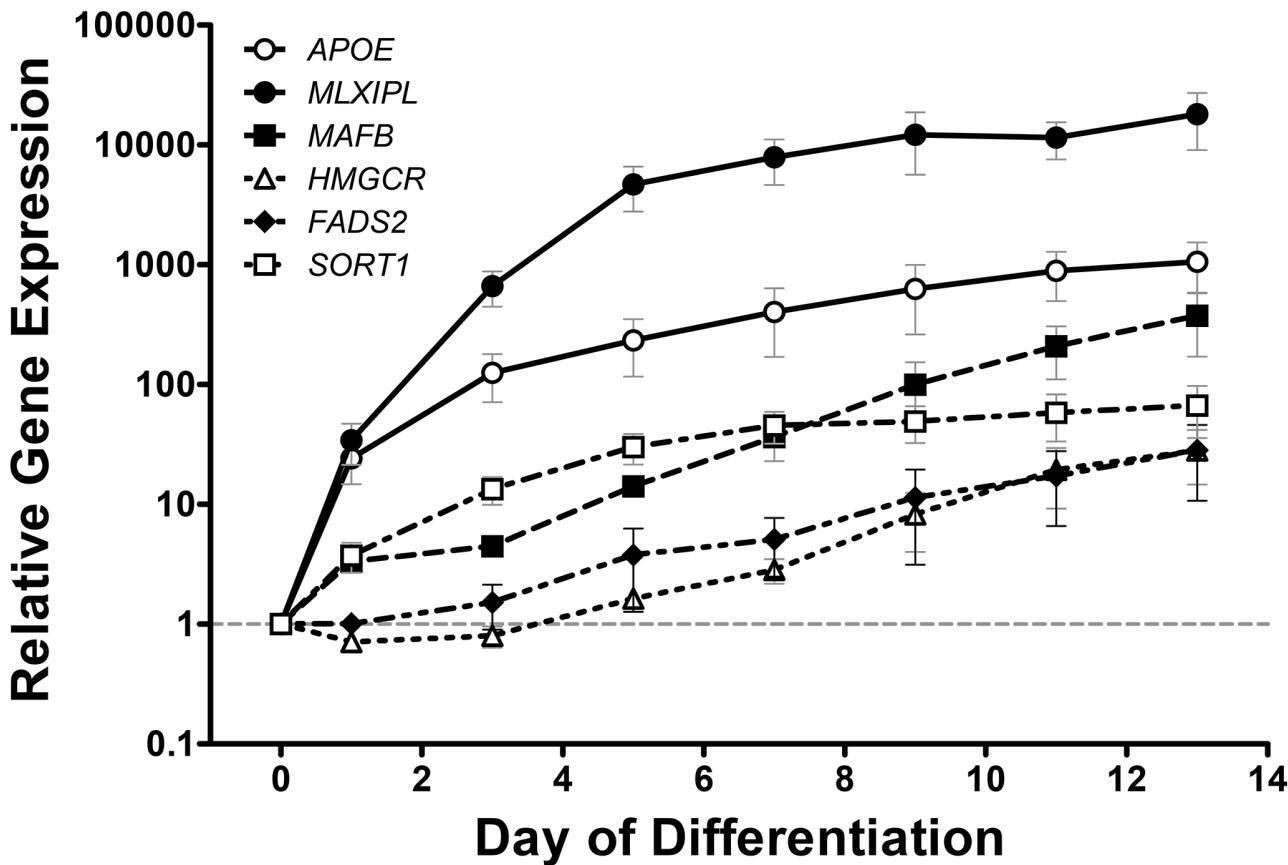


Fig 3. mRNA expression profiles of target genes during human adipogenesis. Fold change of gene expression for *SORT1*, *HMGCR*, *MLXIPL*, *FADS2*, *APOE* and *MAFB* mRNA during human adipocyte differentiation of SGBS cells. Data shown are averaged over 3 independent experiments, each performed in triplicates and results are given in mean±SEM. For all candidates, $p<0.001$ was achieved by one-way ANOVA test with Dunnett's posthoc test.

doi:10.1371/journal.pone.0138064.g003

(*APOE*) on circulating LDL-C levels independent of the degree of obesity in conventional linear regression analyses adjusting for age, sex and BMI SDS. This was confirmed by our Bayesian model analysis suggesting causality of these two variants on LDL-C. Bayesian analysis also revealed that effects of rs4420638 (*APOE*) on HDL-C, rs3846663 (*HMGCR*) and rs6102059 (*MAFB*) on LDL-C as well as rs3812316 (*MLXIPL*) on TG cannot be ruled out. Still, these variants should be considered as candidates requiring further investigations.

The *APOE*-SNP rs4420638 is located on chromosome 19 in a cluster with *APOC1*, *APOC4* and *APOC2*. The SNP rs599839 is located on chromosome 1, close to the genes *CELSR2*, *PSRC1* and *SORT1*. Multiple other studies investigated SNPs in or near the *APOE* and *SORT1* genes. Rs4420638 and rs599839 showed replicable associations with lipid levels (mostly LDL-C) in Caucasian and non-Caucasian population cohorts and meta-analyses [7, 9, 15, 35]. The non-Caucasian cohorts displayed lower significance regarding all SNP-lipid-associations, most likely due to lower case numbers (N ranges from 2,532 to 9,328) [15]. In a small sample, Klein et al. observed effects of rs646776, a proxy of rs599839, only in males [36]. Sex-stratified analysis of our data reveals a significant effect for both sexes. Beta estimator of females is slightly lower than that for males (Females: $p = 6.5 \times 10^{-4}$, $\beta = -0.28$, Males: $p = 4.5 \times 10^{-3}$, $\beta = -0.32$).

Associations of the other variants/candidates could not be confirmed in our sample. Correlations for rs174570 (*FADS2*) with TC, LDL-C and HDL-C were observed in a meta-analysis of

16 European studies [10], although others did not confirm this in Hispanic adult cohorts [37]. Admittedly, our sample is considerably smaller and, thus, weak effects may have remained undetected. However, according to our Bayesian analysis, rs174570 (*FADS2*) is the most implausible of the considered candidates, since, in contrast to the other variants, it is dismissed for all three lipid phenotypes analysed.

For other variants, even large sample-sized and high-powered adult studies gave controversial results. While a significant association of rs6102059 (*MAFB*) with LDL-C was observed by some [8], this could not be replicated by others [15]. However, the cohort of supposed European ancestry consisted of self-identified European Americans. Their genetic origin was not validated, which might have blurred the associations. In our Central European sample we observed no convincing association of rs6102059 with TC and LDL-C levels but we can also not completely rule out an effect based on our Bayesian analysis.

The intronic SNP rs3846663 in *HMGCR* was reported to be significantly associated with LDL-C levels in a cohort of 19,840 subjects [8]. These findings were replicated in several populations (Kosrean islands inhabitants (n = 2,346) with an even larger effect size compared to Kathiresan et al [8] and Japanese [38] or Scottish [39]). Our results did not reach significance again possibly due to our limited sample size, but by trend, are in line with the above mentioned studies. This is further confirmed by our Bayesian analysis complying with a possible causal effect of rs3846663 on LDL-C but not on HDL-C or TG.

Rs3812316 (*MLXIPL*) was most strongly associated with TG in adults [16], although others did not find this association [40, 41]. It was suggested that the effect on TG levels must be weak if it exists at all [40]. In our study, standard linear regression analysis did not reveal any significant association. Nevertheless our results show lower triglyceride levels in homozygous SNP-carriers with rs3812316/GG genotype (CC: 1.13 mmol/l; GG: 0.69 mmol/l adjusted for age, sex and BMI SDS) in agreement with the above mentioned observations. The effects that were seen in our analysis indicate a protective function for minor allele carriers concerning triglyceride levels in children, even in the presence of obesity [16]. Again, Bayesian model analysis supports this finding since in contrast to HDL-C and LDL-C, an effect of rs3812316 to TG cannot be excluded.

A limitation of our study is the relatively small sample size since recruitment of volunteers is more challenging for childhood cohorts. Children are a population much less affected by chronic diseases or medication. Therefore, genetic studies in childhood cohorts are intriguing. Indeed, we were able to confirm the association for children for variants which are originally detected in much larger cohorts of adults comprising several thousands of individuals.

Although, the power of our study is limited, we could confirm rs599839 and rs4420638 to be associated in children. Interestingly, higher effect sizes compared to adults were observed. However, one has to note that our study population is mainly obese. Therefore, replication in a population-based sample of children is required to show general validity of our associations.

Also, besides the possibility that due to the lower influence of co-morbidities and other life-style related factors, primary genetic effects may be hypothesized to be stronger in children than in adults, an alternative possibility would be the later emergence of genetic effects on phenotype. This would particularly apply to conditions where genetic predisposition is reinforced by additional (environmental) risk factors that accumulate or increase with life time (double/multiple hit theory). Such a relationship has been discussed for the manifestation of coronary artery disease in patient with genetic risk factors [42]. It also needs to be considered, that children and adolescents do not yet present with overt disease and hence do not meet the pathological endpoints (eg. myocardial infarction), which limits interpretation on genotype-phenotype associations.

For adults it is common practise to combine diverse cohorts (i.e. The Framingham Heart Study, Invecchiare in Chianti, London Life Science Population Cohort [8], The Rotterdam Study [10], Diabetes Genetics Initiative [7, 8] or The Finland–United States Investigation of NIDDM Genetics [8, 9, 16]). These cohorts differ considerably regarding the burden of chronic illness or drug-intake which might lower the chance to detect genetic associations. However, besides all the advantages of childhood cohorts, we have to acknowledge that studies in adolescent individuals might be affected by changes of lipid metabolism during puberty [43].

By our Bayesian modelling approach we propose an innovative method of analysing multi-SNP–multi-phenotype associations independently and in addition to the classical frequentist regression modelling. This type of analysis overcomes a number of limitations of classical regression analysis: First, it allows comparisons of different types of models, i.e. different modes of inheritance and inclusion of co-variables. Although it is possible in principle to include multiple SNPs and covariables in regression analysis, this usually results in a large number of possible models with no generally accepted decision rule how to select an optimal one. Second, it considers polygenic effects and the information of other phenotypes as well. Considering the correlation structure between different phenotypes can improve detection of the underlying signal [29]. To some extent, this also allows inference regarding unconfounded effects of genotypes and co-variables, which may be indicative for direct or even causal relationships. Interestingly, as discussed above, our Bayesian model results are always in line with observations in adult studies and hence support these results.

Third, the Bayesian approach can deal with missing values, i.e. single missings in either phenotypes, co-variables or SNPs [44]. For example a classical analysis of all SNPs and phenotypes in parallel would reduce the sample size from 594 to 521 in our study whereas Bayesian analysis includes all individuals resulting in higher power. Indeed, compared to the classical analysis, standard deviations of effect estimates are typically smaller, i.e. estimates are more precise [30] and may handle smaller sample sizes.

Summary

We could show for the first time in children that rs599839 (*SORT1*) and rs4420638 (*APOE*) are strongly associated with alterations in blood lipid levels independent of the presence and degree of obesity. Our integrative Bayesian model analysis provided further candidate associations requiring further investigation of the candidates. Therefore, we conclude that this novel approach can improve the detection of weaker associations in genotype-phenotype data sets.

Supporting Information

S1 Data. Excel sheet of raw data: Variable sample_id is the sample identifier. SNP genotype corresponds to the number of minor alleles (0, 1 or 2) and 3 refer to missing values. Lipid phenotype levels are provided as hdl_c (high density lipoprotein cholesterol), ldl_c (low density lipoprotein cholesterol), tc (total cholesterol) and tg (triglyceride). Accordingly, the covariables age, sex and bmi_sds are provided.
(XLSX)

S1 Fig. Comparison of standard errors.
(DOCX)

S2 Fig. Results. Effects of identified genetic risk variants on LDL-C.
(DOCX)

S1 Methods. Details of Bayesian modelling.
(DOCX)

S2 Methods. Derivation of Bayes factors for models.

(DOCX)

S3 Methods. WinBUGS model for Bayesian model selection.

(DOCX)

S1 Results. Genetic models for BMI SDS and lipid phenotypes.

(DOCX)

S2 Results. Estimated covariances.

(DOCX)

S1 Table. Primer and probes for gene expression analysis.

(DOCX)

S2 Table. Primer sequences for genotyping.

(DOCX)

S3 Table. Selection and characteristics of lipid variants.

(DOCX)

Acknowledgments

This work was supported by grants from the German Research Council (DFG) for the Clinical Research Center “Obesity Mechanisms” CRC1052/1 C05 and the Federal Ministry of Education and Research (BMBF), Germany, FKZ: 01EO1001 (IFB AdiposityDiseases), and the European Community’s Seventh Framework Programme (FP7/2007-2013) project Beta-JUDO under grant agreement n° 279153 to A.K. The work of A.G. was supported in part by the German Federal Research Ministry (BMBF), grant PROGRESS (01KI1010I). A.G. and M.S. and AK were funded by LIFE-Leipzig Research Center for Civilization Diseases, University of Leipzig. LIFE is funded by means of the European Union, by the European Regional Development Fund (ERDF) and by means of the Free State of Saxony within the framework of the excellence initiative.

Author Contributions

Conceived and designed the experiments: CB PB AK. Performed the experiments: CB PB. Analyzed the data: CB AG PB MS DS PK AK. Contributed reagents/materials/analysis tools: AG MS. Wrote the paper: CB AG PB MS SW WK PK AK. Co-supervised the project: PK.

References

- Report by the Central Committee for Medical and Community Program of the American Heart AssociationMedical and Community Program of the American Heart Association. Dietary fat and its relation to heart attacks and strokes. *JAMA*. 1961; 175:389–91. PMID: [14447694](#)
- Mendivil CO, Rimm EB, Furtado J, Chiuve SE, Sacks FM. Low-density lipoproteins containing apolipoprotein C-III and the risk of coronary heart disease. *Circulation*; 2011;2011. p. 2065–72. doi: [10.1161/CIRCULATIONAHA.111.056986](#) PMID: [21986282](#)
- Cullen P. Evidence that triglycerides are an independent coronary heart disease risk factor. *AmJCardiol*. 2000; 86(9):943–9.
- Tan MH. HDL-cholesterol: the negative risk factor for coronary heart disease. *AnnAcadMedSingap*. 1980; 9(4):491–5.
- Rippe JM, Crossley S, Ringer R. Obesity as a chronic disease: modern medical and lifestyle management. *J Am Diet Assoc*. 1998; 98(10 Suppl 2):S9–15. PMID: [9787730](#)

6. Ruderman N, Chisholm D, Pi-Sunyer X, Schneider S. The metabolically obese, normal-weight individual revisited. *Diabetes*; 1998;1998. p. 699–713. PMID: [9588440](#)
7. Kathiresan S, Melander O, Guiducci C, Surti A, Burtt NP, Rieder MJ, et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *NatGenet.* 2008; 40(2):189–97.
8. Kathiresan S, Willer CJ, Peloso GM, Demissie S, Musunuru K, Schadt EE, et al. Common variants at 30 loci contribute to polygenic dyslipidemia. *NatGenet.* 2009; 41(1):56–65.
9. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *NatGenet.* 2008; 40(2):161–9.
10. Aulchenko YS, Ripatti S, Lindqvist I, Boomsma D, Heid IM, Pramstaller PP, et al. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *NatGenet.* 2009; 41(1):47–55.
11. Standl M, Lattka E, Stach B, Koletzko S, Bauer CP, von BA, et al. FADS1 FADS2 gene cluster, PUFA intake and blood lipids in children: results from the GiNplus and LISplus studies. *PLoS ONE.* 2012; 7(5):e37780.
12. Molto-Puigmarti C, Jansen E, Heinrich J, Standl M, Mensink RP, Plat J, et al. Genetic variation in FADS genes and plasma cholesterol levels in 2-year-old infants: KOALA Birth Cohort Study. *PLoS ONE.* 2013; 8(5):e61671.
13. Hu P, Qin YH, Lei FY, Pei J, Hu B, Lu L. Variable frequencies of apolipoprotein E genotypes and its effect on serum lipids in the Guangxi Zhuang and Han children. *IntJMolSci.* 2011; 12(9):5604–15.
14. Atabek ME, Ozkul Y, Eklioglu BS, Kurtoglu S, Baykara M. Association between apolipoprotein E polymorphism and subclinical atherosclerosis in patients with type 1 diabetes mellitus. *JClinResPediatrEndocrinol.* 2012; 4(1):8–13.
15. Dumitrescu L, Carty CL, Taylor K, Schumacher FR, Hindorff LA, Ambite JL, et al. Genetic determinants of lipid traits in diverse populations from the population architecture using genomics and epidemiology (PAGE) study. *PLoS Genet.* 2011; 7(6).
16. Kooner JS, Chambers JC, Aguilar-Salinas CA, Hinds DA, Hyde CL, Warnes GR, et al. Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides. *NatGenet.* 2008; 40(2):149–51.
17. Deloukas P, Kanoni S, Willenborg C, Farrall M, Assimes TL, Thompson JR, et al. Large-scale association analysis identifies new risk loci for coronary artery disease. *NatGenet.* 2013; 45(1):25–33.
18. Greenawalt DM, Dobrin R, Chudin E, Hatoum IJ, Suver C, Beaulaurier J, et al. A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. *Genome research.* 2011; 21(7):1008–16. doi: [10.1101/gr.112821.110](#) PMID: [21602305](#)
19. Innocenti F, Cooper GM, Stanaway IB, Gamazon ER, Smith JD, Mirkov S, et al. Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS genetics.* 2011; 7(5):e1002078. doi: [10.1371/journal.pgen.1002078](#) PMID: [21637794](#)
20. Schroder A, Klein K, Winter S, Schwab M, Bonin M, Zell A, et al. Genomics of ADME gene expression: mapping expression quantitative trait loci relevant for absorption, distribution, metabolism and excretion of drugs in human liver. *The pharmacogenomics journal.* 2013; 13(1):12–20. doi: [10.1038/tpj.2011.44](#) PMID: [22006096](#)
21. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature genetics.* 2013; 45(10):1238–43. doi: [10.1038/ng.2756](#) PMID: [24013639](#)
22. Zhang X, Johnson AD, Hendricks AE, Hwang SJ, Tanriverdi K, Ganesh SK, et al. Genetic associations with expression for genes implicated in GWAS studies for atherosclerotic cardiovascular disease and blood phenotypes. *Human molecular genetics.* 2014; 23(3):782–95. doi: [10.1093/hmg/ddt461](#) PMID: [24057673](#)
23. Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, Castagne R, et al. Genetics and beyond—the transcriptome of human monocytes and disease susceptibility. *PLoS One.* 2010; 5(5):e10693. doi: [10.1371/journal.pone.0010693](#) PMID: [20502693](#)
24. Kromeyer-Hauschild K, Wabitsch M, Geller F, Ziegler A, Geiß HC, Hesse V, et al. Perzentilen für den Body Mass Index für das Kindes- und Jugendalter unter Heranziehung verschiedener deutscher Stichproben. (Centiles for body mass index for children and adolescents derived from distinct independent German cohorts). *Monatsschr Kinderheilkd.* 2001; 149:807–18.
25. Marshall WA, Tanner JM. Variations in pattern of pubertal changes in girls. *Arch Dis Child.* 1969; 44(235):291–303. PMID: [5785179](#)
26. Marshall WA, Tanner JM. Variations in the pattern of pubertal changes in boys. *Arch Dis Child.* 1970; 45(239):13–23. PMID: [5440182](#)

27. Bernhard F, Landgraf K, Klöting N, Berthold A, Büttner P, Fribe D, et al. Functional relevance of genes implicated by obesity genome-wide association study signals for human adipocyte biology. *Diabetologia*. 2013; 56(2):311–22. doi: [10.1007/s00125-012-2773-0](https://doi.org/10.1007/s00125-012-2773-0) PMID: [23229156](#)
28. Quintana MA, Schumacher FR, Casey G, Bernstein JL, Li L, Conti DV. Incorporating prior biologic information for high-dimensional rare variant association studies. *Human heredity*. 2012; 74(3–4):184–95. doi: [10.1159/000346021](https://doi.org/10.1159/000346021) PMID: [23594496](#)
29. Lunn DJ, Whittaker JC, Best N. A Bayesian toolkit for genetic association studies. *Genet Epidemiol*. 2006; 30(3):231–47.
30. Conti DV, Cortessis V, Molitor J, Thomas DC. Bayesian modeling of complex metabolic pathways. *Human heredity*. 2003; 56(1–3):83–93. PMID: [14614242](#)
31. Lunn DJ, Best N, Whittaker JC. Generic reversible jump MCMC using graphical models. *Stat Comput*. 2009; 19:395–408.
32. Kass RE, Raftery AE. Bayes Factors. *Journal of the American Statistical Association*. 1995; 90(430):773–95.
33. Jeffreys H. Theory of Probability. 3rd edition ed. Oxford, U.K.: Oxford University Press; 1961 1961.
34. Wabitsch M, Brenner RE, Melzner I, Braun M, Moller P, Heinze E, et al. Characterization of a human preadipocyte cell strain with high capacity for adipose differentiation. *Int J Obes Relat Metab Disord*. 2001; 25(1):8–15.
35. Ken-Dror G, Talmud PJ, Humphries SE, Drenos F. APOE/C1/C4/C2 gene cluster genotypes, haplotypes and lipid levels in prospective coronary heart disease risk among UK healthy men. *Mol Med*. 2010; 16(9–10):389–99.
36. Klein MS, Connors KE, Shearer J, Vogel HJ, Hittel DS. Metabolomics reveals the sex-specific effects of the SORT1 low-density lipoprotein cholesterol locus in healthy young adults. *Journal of proteome research*. 2014; 13(11):5063–70. doi: [10.1021/pr500659r](https://doi.org/10.1021/pr500659r) PMID: [25182463](#)
37. Lu Y, Feskens EJ, Dollé ME, Imholz S, Verschuren WM, Müller M, et al. Dietary n-3 and n-6 polyunsaturated fatty acid intake interacts with FADS1 genetic variation to affect total and HDL-cholesterol concentrations in the Doetinchem Cohort Study. *The American Journal of Clinical Nutrition*. 2010; 92(1):258–65. doi: [10.3945/ajcn.2009.29130](https://doi.org/10.3945/ajcn.2009.29130) PMID: [20484448](#)
38. Hiura Y, Tabara Y, Kokubo Y, Okamura T, Goto Y, Nonogi H, et al. Association of the functional variant in the 3-hydroxy-3-methylglutaryl-coenzyme a reductase gene with low-density lipoprotein-cholesterol in Japanese. *Circulation journal: official journal of the Japanese Circulation Society*. 2010; 74(3):518–22.
39. Donnelly LA, Doney AS, Dannfeld J, Whitley AL, Lang CC, Morris AD, et al. A paucimorphic variant in the HMG-CoA reductase gene is associated with lipid-lowering response to statin treatment in diabetes: a GoDARTS study. *Pharmacogenetics and genomics*. 2008; 18(12):1021–6. doi: [10.1097/FPC.0b013e3283106071](https://doi.org/10.1097/FPC.0b013e3283106071) PMID: [18815589](#)
40. Vrablik M, Ceska R, Adamkova V, Peasey A, Pikhart H, Kubinova R, et al. MLXIPL variant in individuals with low and high triglyceridemia in white population in Central Europe. *Human genetics*. 2008; 124(5):553–5. doi: [10.1007/s00439-008-0577-6](https://doi.org/10.1007/s00439-008-0577-6) PMID: [18946681](#)
41. Polgár N, Járomi L, Csöngei V, Maász A, Sipeky C, Sáfrány E, et al. Triglyceride level modifying functional variants of GALTN2 and MLXIPL in patients with ischaemic stroke. *European journal of neurology: the official journal of the European Federation of Neurological Societies*. 2010; 17(8):1033–9.
42. Munir MS, Wang Z, Alahdab F, Steffen MW, Erwin PJ, Kullo IJ, et al. The association of 9p21-3 locus with coronary atherosclerosis: a systematic review and meta-analysis. *BMC Med Genet*. 2014; 15:66. doi: [10.1186/1471-2350-15-66](https://doi.org/10.1186/1471-2350-15-66) PMID: [24906238](#)
43. Moran A, Jacobs DR Jr, Steinberger J, Steffen LM, Pankow JS, Hong CP, et al. Changes in insulin resistance and cardiovascular risk during adolescence: establishment of differential risk in males and females. *Circulation*; 20082008. p. 2361–8. doi: [10.1161/CIRCULATIONAHA.107.704569](https://doi.org/10.1161/CIRCULATIONAHA.107.704569) PMID: [18427135](#)
44. Lunn DJ, Osorio C, Whittaker JC. A multivariate probit model for inferring missing haplotype/genotype data. Technical report EPH-2005-02. 2005; Department of Epidemiology and Public Health, Imperial College London, UK.(21).

5 Ausblick

Im letzten Kapitel wurden Ergebnisse einer klassischen und einer bayesianischen SNP-Assoziationsanalyse vorgestellt und miteinander verglichen. Für die bayesianische Analyse wurden gegenüber der klassischen Analyse einige Vorteile festgestellt, die weiter untersucht werden sollen. Einerseits kann die Berücksichtigung der Phänotyp-Korrelationsstruktur im bayesianischen Modell zu einer verbesserten Identifikation von Phänotyp-Genotyp-Beziehungen führen. In einer Simulationsstudie sollen deshalb mehrdimensionale Phänotypen mit verschiedenen starken Korrelationen erzeugt und die Auswirkung auf die Identifikation genetischer Effekte im Vergleich zur klassischen Analyse untersucht werden. Dabei ist zu prüfen, welchen Einfluß die Anzahl und die Stärke der genetischen Effekte auf die Ergebnisse haben. Andererseits führt bei der bayesianischen Modellauswahl eine zu starke Korrelation der Phänotypen zu einer langsamen Konvergenz der empirischen Posterior-Verteilung der Modellparameter zur stationären Verteilung der Markovkette. In diesem Zusammenhang ist zu prüfen, welchen Einfluß die Modellstruktur, die Prior-Verteilungen oder verschiedene Sampling-Verfahren auf die Konvergenz haben.

Bei der bayesianischen Modellauswahl lassen sich die identifizierten genetischen Effekte durch BMA über alle Modelle mitteln, in denen sie eingeschlossen wurden. Dadurch fallen die empirischen Varianzen der Effekte meist kleiner aus als die zugehörigen Varianzen der Beta-Schätzer aus der klassischen Analyse. In einer Simulationsstudie soll untersucht werden, wie sich die Vorausnahme bestimmter genetischer Modelle und die Korrelationsstruktur der Phänotypen auf die empirischen Varianzen der genetischen Effekte auswirkt.

Zuletzt werden bei der klassischen Analyse meist aufgrund einzelner Fehlwerte in Phänotypen, SNPs oder Kovariablen ganze Fälle verworfen, was zu einer erheblichen Fallzahlreduktion führen kann. Diese Fehlwerte können bei der bayesianischen Analyse als zusätzliche Parameter modelliert werden, wodurch die ursprüngliche Fallzahl erhalten bleibt. Mit einer Simulationsstudie soll ermittelt werden, wie sich die Anzahl sowie die Verteilung der Fehlwerte und die damit verbundene Fallzahlreduktion in der klassischen Analyse auf die Identifikation genetischer Effekte im Vergleich zur bayesianischen Analyse auswirkt.

Zusammenfassung der Arbeit

Dissertation zur Erlangung des akademischen Grades

Dr. rer. nat.

Titel

Über Korrelationsstrukturen bei SNP-Assoziationsanalysen

eingereicht von

Arnd Groß

angefertigt am

Institut für Medizinische Informatik, Statistik und Epidemiologie der Universität Leipzig

betreut von

Prof. Dr. rer. nat. Markus Scholz

Monat und Jahr der Einreichung

Juni 2018

Zusammenfassung

Diese kumulative Dissertation umfaßt drei Publikationen, die im Folgenden kurz vorgestellt werden. Die erste Publikation befasst sich mit der Fragestellung aus der Populationsgenetik, ob Isolatpopulationen für die Erforschung genetischer Ursachen von Krankheiten oder quantitativen Phänotypen besser geeignet sind als nicht isolierte Populationen. Man erwartet aufgrund homogenerer Umwelteinflüsse, geringerer Anzahl kausaler genetischer Varianten und insbesondere durch homogenere Bereiche im Genom Vorteile bei der Identifikation genetischer Ursachen in Isolatpopulationen. Am Beispiel der Sorben, die einen gewissen Isolatcharakter aufweisen, sollte deshalb untersucht werden, inwieweit sich diese von einer deutschen populationsbasierten Studie wie KORA genetisch unterscheiden und welche Bedeutung die Unterschiede für genetische Assoziationsanalysen haben. In der ersten Publikation wird gezeigt, daß die Sorben Merkmale genetischer Isolation aufweisen, die nicht auf eine stärkere Verwandtschaftsstruktur der Studienpopulation gegenüber KORA zurückzuführen sind. Die Merkmale genetischer Isolation sind moderat, trotzdem ist der slawische Ursprung erkennbar. Daraus läßt sich schließen, daß die Sorben ursprünglich genetisch isoliert waren, jedoch die genetische Isolation verloren geht. Trotz Unterschiede in der

SNP-Korrelationsstruktur durch ein im Mittel höheres Kopplungsungleichgewicht zwischen benachbarten SNPs ist kein klarer Vorteil bei der Power von SNP-Assoziationsanalysen zu erwarten. Die Verwandtschaftsstruktur der Sorben kann aber bei unkorrigierten SNP-Assoziationsanalysen zu einer Varianzinflation des Effektschätzers führen und die Power des Tests in komplexer Weise beeinflussen. Es sollte daher in einer weiteren Publikation geklärt werden, wie die Verwandtschaftsstruktur der Studienpopulation und die Heritabilität eines Phänotyps die Varianz des Effektschätzers und die Power des Tests tatsächlich beeinflussen.

In der zweiten Publikation wird der Einfluß der Verwandtschaftsstruktur auf SNP-Assoziationsanalysen im Detail untersucht. Verwandtschaften in einer Studienpopulation führen zu korrelierten Phänotypen, was die Annahme unabhängiger Beobachtungen des einfachen linearen Modells verletzt. Aus empirischen Studien war zudem bekannt, daß eine stärkere Verwandtschaftsstruktur der Studienpopulation und eine größere Heritabilität des Phänotyps den Fehler erster Art eines unkorrigierten Tests vergrößern. Der Einfluß der Verwandtschaftsstruktur auf die Power wurde in empirischen Studien unterschiedlich beurteilt. Zudem wird genomic control häufig dazu verwendet, eine Inflation der Teststatistik durch Verwandtschaft zu korrigieren, jedoch führt genomic control zu einer Power-Reduktion. Auch diese empirischen Beobachtungen sollten erklärt werden. In der zweiten Publikation wird analytisch gezeigt, wie die Verwandtschaftsstruktur und die Heritabilität des Phänotyps mit der Varianzinflation des Effektschätzers und der Teststatistik zusammenhängen. Während der Fehler erster Art mit größerer Varianzinflation steigt, wird die Power in komplexer Weise beeinflußt. Ob die Power bei Varianzinflation größer oder kleiner wird, hängt von der Stärke des genetischen Effekts und vom Signifikanzniveau des Tests ab. Zudem konnten weitere empirische Beobachtungen aus der Literatur analytisch erklärt werden, zum Beispiel daß der Erwartungswert des Effektschätzers nicht durch Verwandtschaft beeinflußt wird, die empirische Varianz des Effektschätzers bei Verwandtschaft deflationiert ist und daß die Allelfrequenz des SNP nur einen geringen Einfluß auf die Varianzinflation hat. Weiterhin kann genomic control im Allgemeinen nicht für die Korrektur von Varianzinflation durch Verwandtschaft empfohlen werden. Obwohl der Fehler erster Art durch genomic control eingehalten wird, führt die Methode zu einem starken Power-Verlust in Abhängigkeit der Varianzinflation. Zur Bestimmung der Varianzinflation wurde eine Näherungsformel analytisch hergeleitet, die nur die Verwandtschaftsstruktur und die Heritabilität des Phänotyps benötigt. Aus der Publikation folgt, daß eine Varianzinflation kleiner als 1,05 keinen relevanten Einfluß auf den statistischen Test hat und die Verwendung des einfachen linearen Modells in diesem Fall angemessen ist. Ist die Varianzinflation größer, müssen Methoden wie beispielsweise gemischte Modelle im Rahmen einer SNP-Assoziationsanalyse verwendet werden, welche explizit die Verwandtschaftsstruktur berücksichtigen.

In der dritten Publikation wird ein weiteres Paradigma der Statistik betrachtet. Eine SNP-Assoziationsanalyse kann neben klassischen Methoden auch mit bayesianischen Methoden erfolgen. Bayesianische Methoden bieten dabei die Möglichkeit, SNP- und Phänotyp-Korrelationen zu berücksichtigen und so die Modellanpassung gegenüber der klassischen Analyse zu verbessern. Am Beispiel einer Kinderstudie sollte nach dem Einfluß bestimmter SNPs ausgewählter Kandidaten-Gene (SORT1, HMGCR, MLXIPL, FADS2, APOE, MAFB) auf Lipidkonzentrationen von HDL-C (high density lipoprotein cholesterol), LDL-C (low density lipoprotein choleste-

rol), TC (total cholesterol) und TG (triglyceride) gesucht werden, um auf genetische Ursachen für Parameter des Stoffwechsels in der frühen Entwicklung schließen zu können. In der dritten Publikation wurde zunächst eine klassische SNP-Assoziationsanalyse durchgeführt und ein Zusammenhang von SORT1 und APOE mit LDL-C und TC identifiziert. Darauf wurde in einer bayesianischen Analyse der mehrdimensionale Phänotyp aus HDL-C, LDL-C und TG modelliert, wodurch explizit die Phänotyp-Korrelationsstruktur berücksichtigt wurde. Für die einzelnen Lipidkonzentrationen wurde eine plausible Auswahl von Einflussfaktoren bestehend aus genetischen Varianten, Alter, Geschlecht und BMI unter Berücksichtigung verschiedener genetischer Modelle bestimmt. Dadurch wurden sowohl die Ergebnisse aus der klassischen Analyse bestätigt, als auch weitere Kandidaten, beispielsweise ein Zusammenhang zwischen MLXIPL und TG, gefunden. Ein wichtiges Ergebnis dieser Arbeit war zudem die Präsentation der bayesianischen Modellergebnisse in einfacher Form.

Für die bayesianische Analyse wurden gegenüber der klassischen Analyse einige Vorteile festgestellt, die zukünftig weiter untersucht werden sollen. Dazu zählt die Berücksichtigung von Korrelationsstrukturen im bayesianischen Modell, die zu einer verbesserten Identifikation von Phänotyp-Genotyp-Beziehungen führen kann. Weiterhin lassen sich die bei der bayesianischen Modellauswahl identifizierten genetischen Effekte über alle Modelle mitteln, in denen die entsprechenden Variablen eingeschlossen wurden. Dadurch fallen die empirischen Varianzen der Effekte meist kleiner aus als die zugehörigen Varianzen der Beta-Schätzer aus der klassischen Analyse. Zuletzt werden bei der klassischen Analyse meist aufgrund einzelner Fehlwerte in Phänotypen, SNPs oder Kovariablen ganze Fälle verworfen, was zu einer erheblichen Fallzahlreduktion führen kann. Diese Fehlwerte können bei der bayesianischen Analyse als zusätzliche Parameter modelliert werden, wodurch die ursprüngliche Fallzahl erhalten bleibt. Diese Aspekte sollen zukünftig in Simulationsstudien untersucht werden, in denen der Einfluß von Korrelationsstrukturen, Effektstärken und Fehlwerten auf die Identifikation genetischer Effekte im Vergleich zur klassischen Analyse betrachtet wird.

Verzeichnis der Abkürzungen und Symbole

Es war nicht zu vermeiden, daß einige Symbole mehrfach und mit unterschiedlicher Bedeutung auftreten. Nachfolgend sind die verschiedenen Bedeutungen aufgeführt.

Wort	Beschreibung
0	Vektor bestehend aus Nullen
a	Parametermenge eines bayesianischen Modells
α	Signifikanzniveau
A	Adenin
b_1	Achsenabschnitt des gemischten Modells
b_2	Effekt des gemischten Modells
β_1	Achsenabschnitt des vereinfachten Modells
β_2	Effekt des vereinfachten Modells
$\hat{\beta}_2$	Effektschätzer des vereinfachten Modells
BF	Bayesfaktor
BMA	Bayes model averaging
BMI	body mass index
BMI SDS	BMI standard deviation score
c	Gesamtzahl der Variablen für die Variablenauswahl
\mathbf{c}	Matrix mit Kovariablen in r Spalten
C	Cytosin
CEU	CEPH (Centre d'Etude du Polymorphisme Humain)
Cor	Korrelationskoeffizient
Cov	Kovarianz
$\hat{\text{Cov}}$	empirische Kovarianz
δ	Wahrscheinlichkeit, beide Allele von einem gemeinsamen Vorfahren zu erben
D'	Maß des Kopplungsungleichgewichts
\mathcal{D}	Daten der bayesianischen Analyse
DNA	deoxyribonucleic acid
\mathbf{e}	Vektor der Residuen des gemischten Modells
ϵ	Vektor der Residuen des vereinfachten Modells
η_1	Maß des Kopplungsungleichgewichts

Wort	Beschreibung
E	Erwartungswert
\hat{E}	empirischer Erwartungswert
err	Fehler erster Art des Tests
ϕ	Wahrscheinlichkeit, genau ein Allel von einem gemeinsamen Vorfahren zu erben
F_{IS}	Korrelation der Allele innerhalb eines Individuums
F_{ST}	Korrelation der Allele von Individuen innerhalb einer Population
\mathbf{g}	Vektor der polygenen Effekte
G	Guanin
G	Maß für die paarweise Verwandtschaft
\mathbf{G}	Verwandtschaftsmatrix
\bar{G}	mittlere Verwandtschaft
Gl	Gleichung
HDL-C	high density lipoprotein cholesterol
HWE	Hardy-Weinberg-Equilibrium, Hardy-Weinberg-Gleichgewicht
\mathbf{I}	Identitätsmatrix
IBD	identical by descent
IBS	identical by state
$I(x)$	gibt 1 zurück, falls das Argument x wahr ist, sonst 0
k	Stichprobenumfang
k	Anzahl der ausgewählten Variablen
KORA	Kooperative Gesundheitsforschung in der Region Augsburg
λ	Faktor der Varianzinflation
$\bar{\lambda}$	gemittelter Inflationsfaktor
$\hat{\lambda}$	geschätzter Inflationsfaktor
λ'	erwarteter Inflationsfaktor
λ'_t	erwarteter Inflationsfaktor, erhalten durch Transformation in Abhängigkeit von R_t^2
LD	linkage disequilibrium, Kopplungsungleichgewicht
LDL-C	low density lipoprotein cholesterol
m	Modell einer Phänotyp-Genotyp-Beziehung
μ	Erwartungswert der Normalverteilung
$\boldsymbol{\mu}$	Erwartungswerte einer multivariaten Normalverteilung
MAF	minor allele frequency
MB	Megabase
MCMC	Markov-Chain-Monte-Carlo
n	Fallzahl
N	Normalverteilung
N_n	n -dimensionale Normalverteilung
odds	Verhältnis aus Wahrscheinlichkeit und Gegenwahrscheinlichkeit eines Ereignisses
ω	odds ratio

Wort	Beschreibung
p	Anzahl der Phänotypen
p	p-Wert
p	Wahrscheinlichkeit für das Auftreten einer paarweisen Allelkombination
P	Wahrscheinlichkeitsfunktion
\hat{P}	empirische Wahrscheinlichkeitsfunktion
PCA	principal component analysis, Hauptkomponentenanalyse
pwr	Power des Tests
q	Anzahl der SNPs
QC	quality control, Qualitätskontrolle
r	Anzahl der Kovariablen
r	Maß des Kopplungsungleichgewichts
R_h^2	Heritabilität, Erblichkeit
R_s^2	erklärte Varianz durch den SNP
R_t^2	Heritabilität, erhalten durch Transformation in Abhängigkeit von λ'_t
ROHs	runs of homozygosity
s	Matrix von SNP-Genotypen in q Spalten
\bar{s}	mittlerer Genotyp
σ_e^2	Varianz der Residuen des gemischten Modells
σ_ϵ^2	Varianz der Residuen des vereinfachten Modells
σ_g^2	Varianz der polygenen Effekte
Σ	Kovarianzmatrix der Residuen
S	Zufallsvariable für einen SNP
S_β^2	empirische Varianz des Effektschätzers
SNP	single nucleotide polymorphism, Einzelnukleotid-Polymorphismus
θ	Vektor mit Indizes ausgewählter Spalten für die Variablenauswahl
T	Matrix-Transpositionsoperator, verwendet als Exponent
T	Thymin
T	T Statistik
\hat{T}	Realisierung von T
T_{gc}	T Statistik nach genomic control
TC	total cholesterol
TG	triglyceride
TSI	Toscans in Italy
V	Varianz
\hat{V}	empirische Varianz
V_β	Varianz des Effektschätzers im Standardmodell ohne Verwandtschaft
x	Matrix mit Variablen in c Spalten für die Variablenauswahl
x	Vektor mit Hilfsvariablen zur Abbildung verschiedener genetischer Modelle
y	Matrix von Phänotypen in p Spalten
$z_{\alpha/2}$	$\alpha/2$ -Quantil der Standardnormalverteilung

Literaturverzeichnis

- [1] A. Gross, A. Tonjes, P. Kovacs, et al. Population-genetic comparison of the Sorbian isolate population in Germany with the German KORA population using genome-wide SNP arrays. *BMC Genet.*, 12:67, Jul 2011. doi:10.1186/1471-2156-12-67.
- [2] A. Gross, A. Tonjes, and M. Scholz. On the impact of relatedness on SNP association analysis. *BMC Genet.*, 18(1):104, Dec 2017. doi:10.1186/s12863-017-0571-x.
- [3] C. Breitling, A. Gross, P. Buttner, et al. Genetic Contribution of Variants near SORT1 and APOE on LDL Cholesterol Independent of Obesity in Children. *PLoS ONE*, 10(9):e0138064, 2015. doi:10.1371/journal.pone.0138064.
- [4] A. Ziegler and I. R. König. *A Statistical Approach to Genetic Epidemiology: Concepts and Applications*. Wiley-VCH, Weinheim, 2006. ISBN 978-3-527-31252-8.
- [5] C. Czado and T. Schmidt. *Mathematische Statistik*. Statistik und ihre Anwendungen. Springer, Heidelberg, 2011. ISBN 978-3-642-17261-8. doi:10.1007/978-3-642-17261-8.
- [6] A. Abbott. Manhattan versus Reykjavik. *Nature*, 406(6794):340–342, Jul 2000. doi:10.1038/35019167.
- [7] I. A. Eaves, T. R. Merriman, R. A. Barber, et al. The genetically isolated populations of Finland and sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. *Nat. Genet.*, 25(3):320–323, Jul 2000. doi:10.1038/77091.
- [8] K. R. Veeramah, A. Tonjes, P. Kovacs, et al. Genetic variation in the Sorbs of eastern Germany in the context of broader European genetic diversity. *Eur. J. Hum. Genet.*, 19(9):995–1001, Sep 2011. doi:10.1038/ejhg.2011.65.
- [9] H. E. Wichmann, C. Gieger, and T. Illig. KORA-gen—resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen*, 67 Suppl 1:26–30, Aug 2005. doi:10.1055/s-2005-858226.
- [10] R. Holle, M. Happich, H. Lowel, and H. E. Wichmann. KORA—a research platform for population based health research. *Gesundheitswesen*, 67 Suppl 1:19–25, Aug 2005. doi:10.1055/s-2005-858235.
- [11] J. Novembre, T. Johnson, K. Bryc, et al. Genes mirror geography within Europe. *Nature*, 456(7218):98–101, Nov 2008. doi:10.1038/nature07331.
- [12] O. Lao, T. T. Lu, M. Nothnagel, et al. Correlation between genetic and geographic structure in Europe. *Curr. Biol.*, 18(16):1241–1248, Aug 2008. doi:10.1016/j.cub.2008.07.049.

- [13] Bruce S. Weir. *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sinauer Associates Inc., Sunderland, 1996. ISBN 978-0-878-93902-2.
- [14] R. McQuillan, A. L. Leutenegger, R. Abdel-Rahman, et al. Runs of homozygosity in European populations. *Am. J. Hum. Genet.*, 83(3):359–372, Sep 2008. doi:10.1016/j.ajhg.2008.08.007.
- [15] J. Yang, B. Benyamin, B. P. McEvoy, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.*, 42(7):565–569, Jul 2010. doi:10.1038/ng.608.
- [16] Yurii S. Aulchenko. Chapter 9 - Effects of Population Structure in Genome-wide Association Studies. In *Analysis of Complex Disease Association Studies*, pages 123 – 156. Academic Press, San Diego, 2011. ISBN 978-0-123-75142-3. doi:10.1016/B978-0-12-375142-3.10009-4.
- [17] N. Amin, C. M. van Duijn, and Y. S. Aulchenko. A genomic background based method for association analysis in related individuals. *PLoS ONE*, 2(12):e1274, 2007. doi:10.1371/journal.pone.0001274.
- [18] J. Wang. An estimator for pairwise relatedness using molecular markers. *Genetics*, 160(3):1203–1215, Mar 2002. URL <http://www.genetics.org/content/160/3/1203>.
- [19] R. C. Lewontin. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics*, 49(1):49–67, Jan 1964. URL <http://www.genetics.org/content/49/1/49>.
- [20] W. G. Hill and A. Robertson. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.*, 38(6):226–231, Jun 1968. doi:10.1007/BF01245622.
- [21] M. Scholz and D. Hasenclever. Comparison of estimators for measures of linkage disequilibrium. *Int J Biostat*, 6(1):Article 1, 2010. doi:10.2202/1557-4679.1162.
- [22] M. Scholz and D. Hasenclever. A canonical measure of allelic association. <https://arxiv.org/abs/0903.3886>, 2009.
- [23] A. W. F. Edwards. The measure of association in a 2 x 2 table. *Journal of the Royal Statistical Society. Series A (General)*, 126(1):109–114, 1963. doi:10.2307/2982448.
- [24] S. Shifman and A. Darvasi. The value of isolated populations. *Nat. Genet.*, 28(4):309–310, Aug 2001. doi:10.1038/91060.
- [25] K. Kristiansson, J. Naukkarinen, and L. Peltonen. Isolated populations and complex disease gene identification. *Genome Biol.*, 9(8):109, 2008. doi:10.1186/gb-2008-9-8-109.
- [26] E. Boerwinkle, R. Chakraborty, and C. F. Sing. The use of measured genotype information in the analysis of quantitative phenotypes in man. I. Models and analytical methods. *Ann. Hum. Genet.*, 50(Pt 2):181–194, May 1986. doi:10.1111/j.1469-1809.1986.tb01037.x.
- [27] Y. M. Zhang, Y. Mao, C. Xie, et al. Mapping quantitative trait loci using naturally occurring genetic variance among commercial inbred lines of maize (*Zea mays L.*). *Genetics*, 169(4):2267–2275, Apr 2005. doi:10.1534/genetics.104.033217.

- [28] J. Yu, G. Pressoir, W. H. Briggs, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.*, 38(2):203–208, Feb 2006. doi:10.1038/ng1702.
- [29] S. Teyssedre, J. M. Elsen, and A. Ricard. Statistical distributions of test statistics used for quantitative trait association mapping in structured populations. *Genet. Sel. Evol.*, 44:32, Nov 2012. doi:10.1186/1297-9686-44-32.
- [30] Y. S. Aulchenko, D. J. de Koning, and C. Haley. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics*, 177(1):577–585, Sep 2007. doi:10.1534/genetics.107.075614.
- [31] P. F. McArdle, J. R. O’Connell, T. I. Pollin, et al. Accounting for relatedness in family based genetic association studies. *Hum. Hered.*, 64(4):234–242, 2007. doi:10.1159/000103861.
- [32] A. Stuart, K. Ord, and S. Arnold. *Kendall’s Advanced Theory of Statistics*, volume 2A. Arnold, a member of the Hodder Headline Group, London, sixth edition, 1999. ISBN 978-0-340-66230-4.
- [33] N. M. Belonogova, G. R. Svishcheva, C. M. van Duijn, Y. S. Aulchenko, and T. I. Axenovich. Region-based association analysis of human quantitative traits in related individuals. *PLoS ONE*, 8(6):e65395, 2013. doi:10.1371/journal.pone.0065395.
- [34] B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, Dec 1999. doi:10.1111/j.0006-341X.1999.00997.x.
- [35] William Astle and David J. Balding. Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statist. Sci.*, 24(4):451–471, 11 2009. doi:10.1214/09-STS307.
- [36] S. A. Bacanu, B. Devlin, and K. Roeder. Association studies for quantitative traits in structured populations. *Genet. Epidemiol.*, 22(1):78–93, Jan 2002. doi:10.1002/gepi.1045.
- [37] A. L. Price, N. A. Zaitlen, D. Reich, and N. Patterson. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.*, 11(7):459–463, Jul 2010. doi:10.1038/nrg2813.
- [38] D. J. Lunn, J. C. Whittaker, and N. Best. A Bayesian toolkit for genetic association studies. *Genet. Epidemiol.*, 30(3):231–247, Apr 2006. doi:10.1002/gepi.20140.
- [39] M. A. Quintana, F. R. Schumacher, G. Casey, et al. Incorporating prior biologic information for high-dimensional rare variant association studies. *Hum. Hered.*, 74(3-4):184–195, 2012. doi:10.1159/000346021.
- [40] D. V. Conti, V. Cortessis, J. Molitor, and D. C. Thomas. Bayesian modeling of complex metabolic pathways. *Hum. Hered.*, 56(1-3):83–93, 2003. doi:10.1159/000073736.
- [41] D. J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337, Oct 2000. doi:10.1023/A:1008929526011.

- [42] J. Martin Bland and Douglas G. Altman. Bayesians and frequentists. *BMJ*, 317(7166):1151–1160, 1998. ISSN 0959-8138. doi:10.1136/bmj.317.7166.1151.
- [43] D. J. Lunn, C. Osorio, and J. C. Whittaker. A multivariate probit model for inferring missing haplotype/genotype data. Technical Report EPH-2005-02, Department of Epidemiology and Public Health, Imperial College London, UK, 2005. URL <https://www1.imperial.ac.uk/resources/0F769948-7871-47EE-B9EA-9ECF9267E4DD/>.
- [44] D. J. Lunn. Automated covariate selection and Bayesian model averaging in population PK/PD models. *J Pharmacokinet Pharmacodyn*, 35(1):85–100, Feb 2008. doi:10.1007/s10928-007-9077-x.
- [45] D. J. Lunn, N. Best, and J. C. Whittaker. Generic reversible jump mcmc using graphical models. *Statistics and computing*, 19(4):395–408, 2009. doi:10.1007/s11222-008-9100-0.
- [46] J. Gareth, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 6th edition, 2015. ISBN 978-1-461-47138-7. doi:10.1007/978-1-4614-7138-7.
- [47] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics*. Chapman & Hall/CRC, Boca Raton, 1996. ISBN 978-0-412-05551-1.
- [48] P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995. doi:10.1093/biomet/82.4.711.
- [49] B. L. Fridley. Bayesian variable and model selection methods for genetic association studies. *Genet. Epidemiol.*, 33(1):27–37, Jan 2009. doi:10.1002/gepi.20353.
- [50] R. B. O’Hara and M. J. Sillanpää. A review of bayesian variable selection methods: what, how and which. *Bayesian Anal.*, 4(1):85–117, 03 2009. doi:10.1214/09-BA403.
- [51] Robert E. Kass and Adrian E. Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995. ISSN 01621459. doi:10.2307/2291091.
- [52] HapMap. Merged phase I+II and III genotype files. ftp://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/2009-02_phaseII+III/, 2009. Accessed 14 Mar 2018.
- [53] A. Tonjes, M. Koriath, D. Schleinitz, et al. Genetic variation in GPR133 is associated with height: genome wide association study in the self-contained population of Sorbs. *Hum. Mol. Genet.*, 18(23):4662–4668, Dec 2009. doi:10.1093/hmg/ddp423.
- [54] A. Tonjes, E. Zeggini, P. Kovacs, et al. Association of FTO variants with BMI and fat mass in the self-contained population of Sorbs in Germany. *Eur. J. Hum. Genet.*, 18(1):104–110, Jan 2010. doi:10.1038/ejhg.2009.107.
- [55] A. Doring, C. Gieger, D. Mehta, et al. SLC2A9 influences uric acid concentrations with pronounced sex-specific effects. *Nat. Genet.*, 40(4):430–436, Apr 2008. doi:10.1038/ng.107.

- [56] A. Tenesa, A. F. Wright, S. A. Knott, et al. Extent of linkage disequilibrium in a Sardinian sub-isolate: sampling and methodological considerations. *Hum. Mol. Genet.*, 13(1):25–33, Jan 2004. doi:10.1093/hmg/ddh001.
- [57] S. Service, J. DeYoung, M. Karayiorgou, et al. Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. *Nat. Genet.*, 38(5):556–560, May 2006. doi:10.1038/ng1770.
- [58] A. Angius, F. C. Hyland, I. Persico, et al. Patterns of linkage disequilibrium between SNPs in a Sardinian population isolate and the selection of markers for association studies. *Hum. Hered.*, 65(1):9–22, 2008. doi:10.1159/000106058.
- [59] A. B. Olshen, B. Gold, K. E. Lohmueller, et al. Analysis of genetic variation in Ashkenazi Jews by high density SNP genotyping. *BMC Genet.*, 9:14, Feb 2008. doi:10.1186/1471-2156-9-14.
- [60] L. Kruglyak. Genetic isolates: separate but equal? *Proc. Natl. Acad. Sci. U.S.A.*, 96(4): 1170–1172, Feb 1999. doi:10.1073/pnas.96.4.1170.
- [61] S. Shifman, J. Kuypers, M. Kokoris, B. Yakir, and A. Darvasi. Linkage disequilibrium patterns of the human genome across populations. *Hum. Mol. Genet.*, 12(7):771–776, Apr 2003. doi:10.1093/hmg/ddg088.
- [62] E. Bosch, H. Laayouni, C. Morcillo-Suarez, et al. Decay of linkage disequilibrium within genes across HGDP-CEPH human samples: most population isolates do not show increased LD. *BMC Genomics*, 10:338, Jul 2009. doi:10.1186/1471-2164-10-338.
- [63] V. C. Sheffield, E. M. Stone, and R. Carmi. Use of isolated inbred human populations for identification of disease genes. *Trends Genet.*, 14(10):391–396, Oct 1998. doi:10.1016/S0168-9525(98)01556-X.
- [64] M. Arcos-Burgos and M. Muenke. Genetics of population isolates. *Clin. Genet.*, 61(4): 233–247, Apr 2002. doi:10.1034/j.1399-0004.2002.610401.x.
- [65] M. Steffens, C. Lamina, T. Illig, et al. SNP-based analysis of genetic substructure in the German population. *Hum. Hered.*, 62(1):20–29, 2006. doi:10.1159/000095850.
- [66] A. Tonjes, M. Scholz, J. Breitfeld, et al. Genome wide meta-analysis highlights the role of genetic variation in RARRES2 in the regulation of circulating serum chemerin. *PLoS Genet.*, 10(12):e1004854, Dec 2014. doi:10.1371/journal.pgen.1004854.
- [67] HapMap. Merged phase I+II and III genotype files. ftp://ftp.ncbi.nlm.nih.gov/ftp://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/2010-08_phaseII+III/, 2010. Accessed 14 Mar 2018.
- [68] H. Zhou, J. Blangero, T. D. Dyer, et al. Fast Genome-Wide QTL Association Mapping on Pedigree and Population Data. *Genet. Epidemiol.*, 41(3):174–186, Apr 2017. doi:10.1002/gepi.21988.

- [69] X. Zhou and M. Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.*, 44(7):821–824, Jun 2012. doi:10.1038/ng.2310.
- [70] J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.*, 88(1):76–82, Jan 2011. doi:10.1016/j.ajhg.2010.11.011.
- [71] J. Yang, N. A. Zaitlen, M. E. Goddard, P. M. Visscher, and A. L. Price. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.*, 46(2):100–106, Feb 2014. doi:10.1038/ng.2876.
- [72] J. Eu-Ahsunthornwattana, E. N. Miller, M. Fakiola, et al. Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLoS Genet.*, 10(7):e1004445, Jul 2014. doi:10.1371/journal.pgen.1004445.
- [73] K. Zhao, M. J. Aranzana, S. Kim, et al. An Arabidopsis example of association mapping in structured samples. *PLoS Genet.*, 3(1):e4, Jan 2007. doi:10.1371/journal.pgen.0030004.
- [74] S. Kathiresan, O. Melander, C. Guiducci, et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat. Genet.*, 40(2):189–197, Feb 2008. doi:10.1038/ng.75.
- [75] S. Kathiresan, C. J. Willer, G. M. Peloso, et al. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat. Genet.*, 41(1):56–65, Jan 2009. doi:10.1038/ng.291.
- [76] C. J. Willer, S. Sanna, A. U. Jackson, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat. Genet.*, 40(2):161–169, Feb 2008. doi:10.1038/ng.76.
- [77] Y. S. Aulchenko, S. Ripatti, I. Lindqvist, et al. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat. Genet.*, 41(1):47–55, Jan 2009. doi:10.1038/ng.269.
- [78] M. Standl, E. Lattka, B. Stach, et al. FADS1 FADS2 gene cluster, PUFA intake and blood lipids in children: results from the GINIplus and LISApplus studies. *PLoS ONE*, 7(5):e37780, 2012. doi:10.1371/journal.pone.0037780.
- [79] C. Molto-Puigmarti, E. Jansen, J. Heinrich, et al. Genetic variation in FADS genes and plasma cholesterol levels in 2-year-old infants: KOALA Birth Cohort Study. *PLoS ONE*, 8(5):e61671, 2013. doi:10.1371/journal.pone.0061671.
- [80] P. Hu, Y. H. Qin, F. Y. Lei, et al. Variable frequencies of apolipoprotein E genotypes and its effect on serum lipids in the Guangxi Zhuang and Han children. *Int J Mol Sci*, 12(9):5604–5615, 2011. doi:10.3390/ijms12095604.
- [81] M. E. Atabek, Y. Ozkul, B. S. Eklioglu, S. Kurtoglu, and M. Baykara. Association between apolipoprotein E polymorphism and subclinic atherosclerosis in patients with type 1 diabetes mellitus. *J Clin Res Pediatr Endocrinol*, 4(1):8–13, Mar 2012. doi:10.4274/jcrpe.521.

- [82] J. M. Rippe, S. Crossley, and R. Ringer. Obesity as a chronic disease: modern medical and lifestyle management. *J Am Diet Assoc*, 98(10 Suppl 2):9–15, Oct 1998. doi:10.1016/S0002-8223(98)00704-4.
- [83] J. S. Kooner, J. C. Chambers, C. A. Aguilar-Salinas, et al. Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides. *Nat. Genet.*, 40(2):149–151, Feb 2008. doi:10.1038/ng.2007.61.
- [84] P. Deloukas, S. Kanoni, C. Willenborg, et al. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat. Genet.*, 45(1):25–33, Jan 2013. doi:10.1038/ng.2480.
- [85] L. Dumitrescu, C. L. Carty, K. Taylor, et al. Genetic determinants of lipid traits in diverse populations from the population architecture using genomics and epidemiology (PAGE) study. *PLoS Genet.*, 7(6):e1002138, Jun 2011. doi:10.1371/journal.pgen.1002138.
- [86] G. Ken-Dror, P. J. Talmud, S. E. Humphries, and F. Drenos. APOE/C1/C4/C2 gene cluster genotypes, haplotypes and lipid levels in prospective coronary heart disease risk among UK healthy men. *Mol. Med.*, 16(9-10):389–399, 2010. URL <http://molmed.org/journal/articles/1/144>.
- [87] M. Vrablik, R. Ceska, V. Adamkova, et al. MLXIPL variant in individuals with low and high triglyceridemia in white population in Central Europe. *Hum. Genet.*, 124(5):553–555, Dec 2008. doi:10.1007/s00439-008-0577-6.
- [88] N. Polgar, L. Jaromi, V. Csongei, et al. Triglyceride level modifying functional variants of GALTN2 and MLXIPL in patients with ischaemic stroke. *Eur. J. Neurol.*, 17(8):1033–1039, Aug 2010. doi:10.1111/j.1468-1331.2010.02957.x.
- [89] A. Gross, D. Teupser, and M. Scholz. Hierarchical statistical model for analysis of multiple genomic and phenotypic data. In *9th Leipzig Research Festival for Life Sciences. 17. Dezember 2010*, page 209, 2010. ISBN 978-3-981-07606-6.
- [90] A. Gross and M. Scholz. On the impact of relatedness on SNP association analysis. In *14th Leipzig Research Festival for Life Sciences. 19. Januar 2018*, page 149, 2018. ISBN 978-3-000-58756-6.
- [91] A. Gross and M. Scholz. Population-genetic comparison of a German isolated population with a German mixed population on the basis of genome-wide SNP markers. In *Mainz//2011. 56. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (gmds), 6. Jahrestagung der Deutschen Gesellschaft für Epidemiologie (DGEpi). Mainz, 26.-29.09.2011*. Düsseldorf: German Medical Science GMS Publishing House, September 2011. doi:10.3205/11gmds061.
- [92] R. Burkhardt, H. Kirsten, F. Beutner, et al. Integration of Genome-Wide SNP Data and Gene-Expression Profiles Reveals Six Novel Loci and Regulatory Mechanisms for Amino Acids and Acylcarnitines in Whole Blood. *PLoS Genet.*, 11(9):e1005510, Sep 2015. doi:10.1371/journal.pgen.1005510.

- [93] H. Kirsten, H. Al-Hasani, L. Holdt, et al. Dissecting the genetics of the human transcriptome identifies novel trait-related trans-eQTLs and corroborates the regulatory relevance of non-protein coding loci. *Hum. Mol. Genet.*, 24(16):4746–4763, Aug 2015. doi:10.1093/hmg/ddv194.
- [94] D. Lopez Herraez, M. Bauchet, K. Tang, et al. Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. *PLoS ONE*, 4(11):e7888, Nov 2009. doi:10.1371/journal.pone.0007888.
- [95] A. Gross. Assembler: Systemzeit manipulieren. *PC Magazin*, 8:239, 1998.
- [96] A. Gross, M. Ziepert, and M. Scholz. KMWin – a tool for graphical presentation of results from Kaplan-Meier survival time analysis. In *10th Leipzig Research Festival for Life Sciences. 16. Dezember 2011*, page 196, 2011. ISBN 978-3-981-07607-3.
- [97] A. Gross, M. Ziepert, and M. Scholz. KMWin—a convenient tool for graphical presentation of results from Kaplan-Meier survival time analysis. *PLoS ONE*, 7(6):e38960, 2012. doi:10.1371/journal.pone.0038960.
- [98] A. Gross, S. Schirm, and M. Scholz. Ycasd – a tool for capturing and scaling data from graphical representations. In *13th Leipzig Research Festival for Life Sciences. 18. Dezember 2014*, page 194, 2014. ISBN 978-3-981-70330-6.
- [99] A. Gross, S. Schirm, and M. Scholz. Ycasd - a tool for capturing and scaling data from graphical representations. *BMC Bioinformatics*, 15:219, Jun 2014. doi:10.1186/1471-2105-15-219.
- [100] A. Gross, M. Scholz, and M. Loeffler. Biomathematisches Modell der menschlichen Thrombopoese. In *5th Leipzig Research Festival for Life Sciences. 15. Dezember 2006*, page 71, 2006. ISBN 978-3-981-07601-X.
- [101] M. Scholz, A. Gross, and M. Loeffler. A biomathematical model of human thrombopoiesis under chemotherapy. *J. Theor. Biol.*, 264(2):287–300, May 2010. doi:10.1016/j.jtbi.2009.12.032.
- [102] A. Gross, M. Scholz, and M. Loeffler. Biomathematisches Modell der menschlichen Thrombopoese. In *Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (gmds). 51. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie. Leipzig, 10.-14.09.2006*. Düsseldorf, Köln: German Medical Science, September 2006. <https://www.egms.de/static/en/meetings/gmds2006/06gmds102.shtml>.

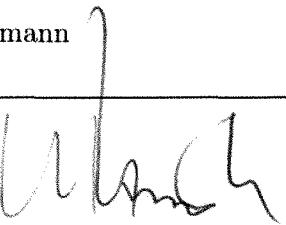
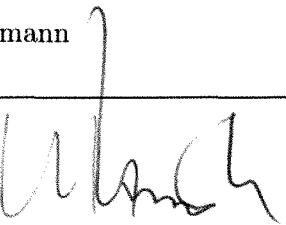
Darstellung des eigenen Beitrags

In diesem Kapitel befinden sich Kopien der Bestätigung des eigenen Beitrags durch die Co-Autoren für die Veröffentlichungen dieser kumulativen Dissertation. Der eigene Beitrag an der Veröffentlichung ist hierbei stichpunktartig aufgeführt und wird durch alle Co-Autoren mit Unterschrift bestätigt.

Confirmation by the co-authors regarding the candidate's contribution to the publication

A. Gross, A. Tonjes, P. Kovacs, K. R. Veeramah, P. Ahnert, N. R. Roshyara, C. Gieger, I. M. Rueckert, M. Loeffler, M. Stoneking, H. E. Wichmann, J. Novembre, M. Stumvoll, and M. Scholz. Population-genetic comparison of the Sorbian isolate population in Germany with the German KORA population using genome-wide SNP arrays. *BMC Genet.*, 12:67, 2011. doi: 10.1186/1471-2156-12-67

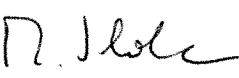
Contribution of Arnd Groß
Quality control of samples and SNPs and merging of all studies Analysis of relatedness (Figure 1, Table 1) Principal components analysis (cover page, Figure 2) Analysis of rare SNPs Analysis of F-statistics (Table 2) Analysis of runs of homozygosity (Figures 3-4) Analysis of linkage disequilibrium (Figure 5) Comparison of power assuming correlated/uncorrelated phenotypes (Figure 6, Tables 3-4) Interpreting the results and writing the manuscript Providing supplemental material (Additional Files 2-6)

Co-author	Signature	Co-author	Signature
Anke Tönjes		Peter Kovacs	
Krishna R. Veeramah		Peter Ahnert	
Nab R. Roshyara		Christian Gieger	
Ina-Maria Rückert		Markus Löffler	
Mark Stoneking		Heinz-Erich Wichmann	
John Novembre		Michael Stumvoll	
Markus Scholz			

Confirmation by the co-authors regarding the candidate's contribution to the publication

A. Gross, A. Tonjes, P. Kovacs, K. R. Veeramah, P. Ahnert, N. R. Roshyara, C. Gieger, I. M. Rueckert, M. Loeffler, M. Stoneking, H. E. Wichmann, J. Novembre, M. Stumvoll, and M. Scholz. Population-genetic comparison of the Sorbian isolate population in Germany with the German KORA population using genome-wide SNP arrays. *BMC Genet.*, 12:67, 2011. doi: 10.1186/1471-2156-12-67

Contribution of Arnd Groß
Quality control of samples and SNPs and merging of all studies Analysis of relatedness (Figure 1, Table 1) Principal components analysis (cover page, Figure 2) Analysis of rare SNPs Analysis of F-statistics (Table 2) Analysis of runs of homozygosity (Figures 3-4) Analysis of linkage disequilibrium (Figure 5) Comparison of power assuming correlated/uncorrelated phenotypes (Figure 6, Tables 3-4) Interpreting the results and writing the manuscript Providing supplemental material (Additional Files 2-6)

Co-author	Signature	Co-author	Signature
Anke Tönjes		Peter Kovacs	
Krishna R. Veeramah		Peter Ahnert	
Nab R. Roshyara		Christian Gieger	
Ina-Maria Rückert		Markus Löffler	
Mark Stoneking		Heinz-Erich Wichmann	
John Novembre		Michael Stumvoll	
Markus Scholz			

Confirmation by the co-authors regarding the candidate's contribution to the publication

A. Gross, A. Tonjes, P. Kovacs, K. R. Veeramah, P. Ahnert, N. R. Roshyara, C. Gieger, I. M. Rueckert, M. Loefler, M. Stoneking, H. E. Wichmann, J. Novembre, M. Stumvoll, and M. Scholz. Population-genetic comparison of the Sorbian isolate population in Germany with the German KORA population using genome-wide SNP arrays. *BMC Genet.*, 12:67, 2011. doi: 10.1186/1471-2156-12-67

Contribution of Arnd Groß
Quality control of samples and SNPs and merging of all studies Analysis of relatedness (Figure 1, Table 1) Principal components analysis (cover page, Figure 2) Analysis of rare SNPs Analysis of F-statistics (Table 2) Analysis of runs of homozygosity (Figures 3-4) Analysis of linkage disequilibrium (Figure 5) Comparison of power assuming correlated/ uncorrelated phenotypes (Figure 6, Tables 3-4) Interpreting the results and writing the manuscript Providing supplemental material (Additional Files 2-6)

Co-author	Signature	Co-author	Signature
Anke Tönjes		Peter Kovacs	
Krishna R. Veeramah		Peter Ahnert	
Nab R. Roshyara		Christian Gieger	
Ina-Maria Rückert		Markus Löfler	
Mark Stoneking		Heinz-Erich Wichmann	
John Novembre		Michael Stumvoll	
Markus Scholz			

Confirmation by the co-authors regarding the candidate's contribution to the publication

A. Gross, A. Tonjes, P. Kovacs, K. R. Veeramah, P. Ahnert, N. R. Roshyara, C. Gieger, I. M. Rueckert, M. Loeffler, M. Stoneking, H. E. Wichmann, J. Novembre, M. Stumvoll, and M. Scholz. Population-genetic comparison of the Sorbian isolate population in Germany with the German KORA population using genome-wide SNP arrays. *BMC Genet.*, 12:67, 2011. doi: 10.1186/1471-2156-12-67

Contribution of Arnd Groß
Quality control of samples and SNPs and merging of all studies Analysis of relatedness (Figure 1, Table 1) Principal components analysis (cover page, Figure 2) Analysis of rare SNPs Analysis of F-statistics (Table 2) Analysis of runs of homozygosity (Figures 3-4) Analysis of linkage disequilibrium (Figure 5) Comparison of power assuming correlated/uncorrelated phenotypes (Figure 6, Tables 3-4) Interpreting the results and writing the manuscript Providing supplemental material (Additional Files 2-6)

Co-author	Signature	Co-author	Signature
Anke Tönjes		Peter Kovacs	
Krishna R. Veeramah		Peter Ahnert	
Nab R. Roshyara		Christian Gieger	
Ina-Maria Rückert		Markus Löffler	
Mark Stoneking		Heinz-Erich Wichmann	
John Novembre		Michael Stumvoll	
Markus Scholz			

Confirmation by the co-authors regarding the candidate's contribution to the publication

A. Gross, A. Tonjes, P. Kovacs, K. R. Veeramah, P. Ahnert, N. R. Roshyara, C. Gieger, I. M. Rueckert, M. Loeffler, M. Stoneking, H. E. Wichmann, J. Novembre, M. Stumvoll, and M. Scholz. Population-genetic comparison of the Sorbian isolate population in Germany with the German KORA population using genome-wide SNP arrays. *BMC Genet.*, 12:67, 2011. doi: 10.1186/1471-2156-12-67

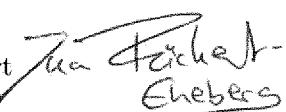
Contribution of Arnd Groß
Quality control of samples and SNPs and merging of all studies Analysis of relatedness (Figure 1, Table 1) Principal components analysis (cover page, Figure 2) Analysis of rare SNPs Analysis of F-statistics (Table 2) Analysis of runs of homozygosity (Figures 3-4) Analysis of linkage disequilibrium (Figure 5) Comparison of power assuming correlated/uncorrelated phenotypes (Figure 6, Tables 3-4) Interpreting the results and writing the manuscript Providing supplemental material (Additional Files 2-6)

Co-author	Signature	Co-author	Signature
Anke Tönjes		Peter Kovacs	
Krishna R. Veeramah		Peter Ahnert	
Nab R. Roshyara		Christian Gieger	
Ina-Maria Rückert		Markus Löffler	
Mark Stoneking		Heinz-Erich Wichmann	
John Novembre		Michael Stumvoll	
Markus Scholz			

Confirmation by the co-authors regarding the candidate's contribution to the publication

A. Gross, A. Tonjes, P. Kovacs, K. R. Veeramah, P. Ahnert, N. R. Roshyara, C. Gieger, I. M. Rueckert, M. Loeffler, M. Stoneking, H. E. Wichmann, J. Novembre, M. Stumvoll, and M. Scholz. Population-genetic comparison of the Sorbian isolate population in Germany with the German KORA population using genome-wide SNP arrays. *BMC Genet.*, 12:67, 2011. doi: 10.1186/1471-2156-12-67

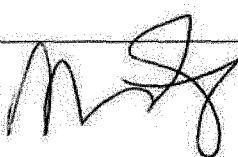
Contribution of Arnd Groß
Quality control of samples and SNPs and merging of all studies Analysis of relatedness (Figure 1, Table 1) Principal components analysis (cover page, Figure 2) Analysis of rare SNPs Analysis of F-statistics (Table 2) Analysis of runs of homozygosity (Figures 3-4) Analysis of linkage disequilibrium (Figure 5) Comparison of power assuming correlated/uncorrelated phenotypes (Figure 6, Tables 3-4) Interpreting the results and writing the manuscript Providing supplemental material (Additional Files 2-6)

Co-author	Signature	Co-author	Signature
Anke Tönjes		Peter Kovacs	
Krishna R. Veeramah		Peter Ahnert	
Nab R. Roshyara		Christian Gieger	
Ina-Maria Rückert 		Markus Löffler	
Mark Stoneking		Heinz-Erich Wichmann	
John Novembre		Michael Stumvoll	
Markus Scholz			

Confirmation by the co-authors regarding the candidate's contribution to the publication

A. Gross, A. Tonjes, P. Kovacs, K. R. Veeramah, P. Ahnert, N. R. Roshyara, C. Gieger, I. M. Rueckert, M. Loeffler, M. Stoneking, H. E. Wichmann, J. Novembre, M. Stumvoll, and M. Scholz. Population-genetic comparison of the Sorbian isolate population in Germany with the German KORA population using genome-wide SNP arrays. *BMC Genet.*, 12:67, 2011. doi: 10.1186/1471-2156-12-67

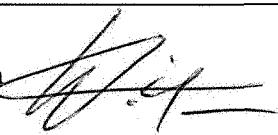
Contribution of Arnd Groß
Quality control of samples and SNPs and merging of all studies
Analysis of relatedness (Figure 1, Table 1)
Principal components analysis (cover page, Figure 2)
Analysis of rare SNPs
Analysis of F-statistics (Table 2)
Analysis of runs of homozygosity (Figures 3-4)
Analysis of linkage disequilibrium (Figure 5)
Comparison of power assuming correlated/uncorrelated phenotypes (Figure 6, Tables 3-4)
Interpreting the results and writing the manuscript
Providing supplemental material (Additional Files 2-6)

Co-author	Signature	Co-author	Signature
Anke Tönjes		Peter Kovacs	
Krishna R. Veeramah		Peter Ahnert	
Nab R. Roshyara		Christian Gieger	
Ina-Maria Rückert		Markus Löffler	
Mark Stoneking		Heinz-Erich Wichmann	
John Novembre		Michael Stumvoll	
Markus Scholz			

Confirmation by the co-authors regarding the candidate's contribution to the publication

A. Gross, A. Tonjes, P. Kovacs, K. R. Veeramah, P. Ahnert, N. R. Roshyara, C. Gieger, I. M. Rueckert, M. Loeffler, M. Stoneking, H. E. Wichmann, J. Novembre, M. Stumvoll, and M. Scholz. Population-genetic comparison of the Sorbian isolate population in Germany with the German KORA population using genome-wide SNP arrays. *BMC Genet.*, 12:67, 2011. doi: 10.1186/1471-2156-12-67

Contribution of Arnd Groß
Quality control of samples and SNPs and merging of all studies Analysis of relatedness (Figure 1, Table 1) Principal components analysis (cover page, Figure 2) Analysis of rare SNPs Analysis of F-statistics (Table 2) Analysis of runs of homozygosity (Figures 3-4) Analysis of linkage disequilibrium (Figure 5) Comparison of power assuming correlated/uncorrelated phenotypes (Figure 6, Tables 3-4) Interpreting the results and writing the manuscript Providing supplemental material (Additional Files 2-6)

Co-author	Signature	Co-author	Signature
Anke Tönjes		Peter Kovacs	
Krishna R. Veeramah		Peter Ahnert	
Nab R. Roshyara		Christian Gieger	
Ina-Maria Rückert		Markus Löffler	
Mark Stoneking		Heinz-Erich Wichmann	
John Novembre		Michael Stumvoll	
Markus Scholz			

Confirmation by the co-authors regarding the candidate's contribution to the publication

A. Gross, A. Tonjes, P. Kovacs, K. R. Veeramah, P. Ahnert, N. R. Roshyara, C. Gieger, I. M. Rueckert, M. Loeffler, M. Stoneking, H. E. Wichmann, J. Novembre, M. Stumvoll, and M. Scholz. Population-genetic comparison of the Sorbian isolate population in Germany with the German KORA population using genome-wide SNP arrays. *BMC Genet.*, 12:67, 2011. doi: 10.1186/1471-2156-12-67

Contribution of Arnd Groß
Quality control of samples and SNPs and merging of all studies Analysis of relatedness (Figure 1, Table 1) Principal components analysis (cover page, Figure 2) Analysis of rare SNPs Analysis of F-statistics (Table 2) Analysis of runs of homozygosity (Figures 3-4) Analysis of linkage disequilibrium (Figure 5) Comparison of power assuming correlated/uncorrelated phenotypes (Figure 6, Tables 3-4) Interpreting the results and writing the manuscript Providing supplemental material (Additional Files 2-6)

Co-author	Signature	Co-author	Signature
Anke Tönjes		Peter Kovacs	
Krishna R. Veeramah		Peter Ahnert	
Nab R. Roshyara		Christian Gieger	
Ina-Maria Rückert		Markus Löffler	
Mark Stoneking		Heinz-Erich Wichmann	
John Novembre		Michael Stumvoll	
Markus Scholz			

Confirmation by the co-authors regarding the candidate's contribution to the publication

A. Gross, A. Tonjes, and M. Scholz. On the impact of relatedness on SNP association analysis. *BMC Genet.*, 18(1):104, Dec 2017. doi: 10.1186/s12863-017-0571-x

Contribution of Arnd Groß

Development of formulae for:

Modelling a SNP-phenotype association (Equations 1-5)

Expected variance inflation of the beta estimate (Equations 6-11)

Empirical variances under relatedness

Hypothesis testing (Equations 12-16)

Genomic control (Equations 17-20)

Quality control of samples and SNPs for all real studies

Genotype simulation of synthetic family studies

Analysis of variance inflation (Figure 1, Table 1)

Numerical validation of test statistics (Tables 2-3)

Analysing the impact of inflation on type I error and power (Figures 2-3)

Interpreting the results and writing the manuscript

Providing supplemental material (Additional Files 1-11)

Co-author	Signature	Co-author	Signature
Anke Tönjes		Markus Scholz	

Confirmation by the co-authors regarding the candidate's contribution to the publication

C. Breitling, A. Gross, P. Büttner, S. Weise, D. Schleinitz, W. Kiess, M. Scholz, P. Kovacs, and A. Körner. Genetic Contribution of Variants near SORT1 and APOE on LDL Cholesterol Independent of Obesity in Children. *PLoS ONE*, 10(9):e0138064, 2015. doi: 10.1371/journal.pone.0138064

Contribution of Arnd Groß

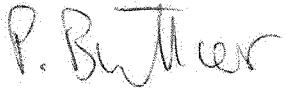
Quality control of samples, SNPs and phenotypes (Table 1)
Frequentist analysis of genotype-phenotype associations (Table 2)
Bayesian modelling of phenotypes, genotypes and covariates (Figures 1-2, Tables 3-4)
Interpreting the results and writing the manuscript
Providing supplemental material (S1 Data, S1 Fig, S1-3 Methods, S1-2 Results, S3 Table)

Co-author	Signature	Co-author	Signature
Clara Breitling		Petra Büttner	
Sebastian Weise		Dorit Schleinitz	
Wieland Kiess		Markus Scholz	
Peter Kovacs		Antje Körner	

Confirmation by the co-authors regarding the candidate's contribution to the publication

C. Breitling, A. Gross, P. Buttner, S. Weise, D. Schleinitz, W. Kiess, M. Scholz, P. Kovacs, and A. Körner. Genetic Contribution of Variants near SORT1 and APOE on LDL Cholesterol Independent of Obesity in Children. *PLoS ONE*, 10(9):e0138064, 2015. doi: 10.1371/journal.pone.0138064

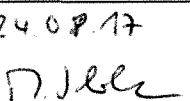
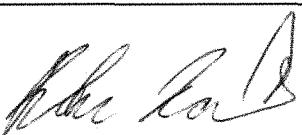
Contribution of Arnd Groß
Quality control of samples, SNPs and phenotypes (Table 1) Frequentist analysis of genotype-phenotype associations (Table 2) Bayesian modelling of phenotypes, genotypes and covariables (Figures 1-2, Tables 3-4) Interpreting the results and writing the manuscript Providing supplemental material (S1 Data, S1 Fig, S1-3 Methods, S1-2 Results, S3 Table)

Co-author	Signature	Co-author	Signature
Clara Breitling		Petra Büttner	
Sebastian Weise		Dorit Schleinitz	
Wieland Kiess		Markus Scholz	
Peter Kovacs		Antje Körner	

Confirmation by the co-authors regarding the candidate's contribution to the publication

C. Breitling, A. Gross, P. Buttner, S. Weise, D. Schleinitz, W. Kiess, M. Scholz, P. Kovacs, and A. Körner. Genetic Contribution of Variants near SORT1 and APOE on LDL Cholesterol Independent of Obesity in Children. *PLoS ONE*, 10(9):e0138064, 2015. doi: 10.1371/journal.pone.0138064

Contribution of Arnd Groß
Quality control of samples, SNPs and phenotypes (Table 1) Frequentist analysis of genotype-phenotype associations (Table 2) Bayesian modelling of phenotypes, genotypes and covariates (Figures 1-2, Tables 3-4) Interpreting the results and writing the manuscript Providing supplemental material (S1 Data, S1 Fig, S1-3 Methods, S1-2 Results, S3 Table)

Co-author	Signature	Co-author	Signature
Clara Breitling		Petra Büttner	
Sebastian Weise		Dorit Schleinitz	
Wieland Kiess		Markus Scholz	 24.08.17
Peter Kovacs		Antje Körner	

Confirmation by the co-authors regarding the candidate's contribution to the publication

C. Breitling, A. Gross, P. Buttner, S. Weise, D. Schleinitz, W. Kiess, M. Scholz, P. Kovacs, and A. Körner. Genetic Contribution of Variants near SORT1 and APOE on LDL Cholesterol Independent of Obesity in Children. *PLoS ONE*, 10(9):e0138064, 2015. doi: 10.1371/journal.pone.0138064

Contribution of Arnd Groß
Quality control of samples, SNPs and phenotypes (Table 1) Frequentist analysis of genotype-phenotype associations (Table 2) Bayesian modelling of phenotypes, genotypes and covariables (Figures 1-2, Tables 3-4) Interpreting the results and writing the manuscript Providing supplemental material (S1 Data, S1 Fig, S1-3 Methods, S1-2 Results, S3 Table)

Co-author	Signature	Co-author	Signature
Clara Breitling		Petra Büttner	
Sebastian Weise		Dorit Schleinitz	
Wieland Kiess		Markus Scholz	
Peter Kovacs		Antje Körner	

Confirmation by the co-authors regarding the candidate's contribution to the publication

C. Breitling, A. Gross, P. Buttner, S. Weise, D. Schleinitz, W. Kiess, M. Scholz, P. Kovacs, and A. Körner. Genetic Contribution of Variants near SORT1 and APOE on LDL Cholesterol Independent of Obesity in Children. *PLoS ONE*, 10(9):e0138064, 2015. doi: 10.1371/journal.pone.0138064

Contribution of Arnd Groß
Quality control of samples, SNPs and phenotypes (Table 1) Frequentist analysis of genotype-phenotype associations (Table 2) Bayesian modelling of phenotypes, genotypes and covariates (Figures 1-2, Tables 3-4) Interpreting the results and writing the manuscript Providing supplemental material (S1 Data, S1 Fig, S1-3 Methods, S1-2 Results, S3 Table)

Co-author	Signature	Co-author	Signature
Clara Breitling		Petra Büttner	
Sebastian Weise		Dorit Schleinitz	
Wieland Kiess		Markus Scholz	
Peter Kovacs		Antje Körner	

Selbstständigkeitserklärung

Erklärung über die eigenständige Abfassung der Arbeit

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und ohne unzulässige Hilfe oder Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Ich versichere, dass Dritte von mir weder unmittelbar noch mittelbar eine Vergütung oder geldwerte Leistungen für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen, und dass die vorgelegte Arbeit weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde zum Zweck einer Promotion oder eines anderen Prüfungsverfahrens vorgelegt wurde. Alles aus anderen Quellen und von anderen Personen übernommene Material, das in der Arbeit verwendet wurde oder auf das direkt Bezug genommen wird, wurde als solches kenntlich gemacht. Insbesondere wurden alle Personen genannt, die direkt an der Entstehung der vorliegenden Arbeit beteiligt waren. Die aktuellen gesetzlichen Vorgaben in Bezug auf die Zulassung der klinischen Studien, die Bestimmungen des Tierschutzgesetzes, die Bestimmungen des Gentechnikgesetzes und die allgemeinen Datenschutzbestimmungen wurden eingehalten. Ich versichere, dass ich die Regelungen der Satzung der Universität Leipzig zur Sicherung guter wissenschaftlicher Praxis kenne und eingehalten habe.

27. Juni 2018

Datum

Unterschrift

Publikationen

Bayesianische Analyse

Genetische Ursachen für Lipid-Serumspiegel

- Publikation¹: C. Breitling, A. Gross, P. Buttner, et al. Genetic Contribution of Variants near SORT1 and APOE on LDL Cholesterol Independent of Obesity in Children. *PLoS ONE*, 10(9):e0138064, 2015. doi:10.1371/journal.pone.0138064

Genetische Ursachen für Phytosterol-Serumspiegel

- Poster: A. Gross, D. Teupser, and M. Scholz. Hierarchical statistical model for analysis of multiple genomic and phenotypic data. In *9th Leipzig Research Festival for Life Sciences. 17. Dezember 2010*, page 209, 2010. ISBN 978-3-981-07606-6

Klassische Analyse

Einfluß von Verwandtschaft auf Assoziationsanalysen

- Poster: A. Gross and M. Scholz. On the impact of relatedness on SNP association analysis. In *14th Leipzig Research Festival for Life Sciences. 19. Januar 2018*, page 149, 2018. ISBN 978-3-000-58756-6
- Publikation¹: A. Gross, A. Tonjes, and M. Scholz. On the impact of relatedness on SNP association analysis. *BMC Genet.*, 18(1):104, Dec 2017. doi:10.1186/s12863-017-0571-x

Populationsgenetischer Vergleich

- Publikation¹: A. Gross, A. Tonjes, P. Kovacs, et al. Population-genetic comparison of the Sorbian isolate population in Germany with the German KORA population using genome-wide SNP arrays. *BMC Genet.*, 12:67, Jul 2011. doi:10.1186/1471-2156-12-67
- Vortrag: A. Gross and M. Scholz. Population-genetic comparison of a German isolated population with a German mixed population on the basis of genome-wide SNP markers. In *Mainz//2011. 56. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (gmds), 6. Jahrestagung der Deutschen Gesellschaft für Epidemiologie (DGEpi). Mainz, 26.-29.09.2011*. Düsseldorf: German Medical Science GMS Publishing House, September 2011. doi:10.3205/11gmds061

¹Diese Publikation wurde peer-reviewed und ist Bestandteil dieser kumulativen Dissertation.

Genetische Statistik im Allgemeinen

- Publikation²: R. Burkhardt, H. Kirsten, F. Beutner, et al. Integration of Genome-Wide SNP Data and Gene-Expression Profiles Reveals Six Novel Loci and Regulatory Mechanisms for Amino Acids and Acylcarnitines in Whole Blood. *PLoS Genet.*, 11(9):e1005510, Sep 2015. doi:10.1371/journal.pgen.1005510
- Publikation²: H. Kirsten, H. Al-Hasani, L. Holdt, et al. Dissecting the genetics of the human transcriptome identifies novel trait-related trans-eQTLs and corroborates the regulatory relevance of non-protein coding loci. *Hum. Mol. Genet.*, 24(16):4746–4763, Aug 2015. doi:10.1093/hmg/ddv194
- Publikation²: A. Tonjes, M. Scholz, J. Breitfeld, et al. Genome wide meta-analysis highlights the role of genetic variation in RARRES2 in the regulation of circulating serum chemerin. *PLoS Genet.*, 10(12):e1004854, Dec 2014. doi:10.1371/journal.pgen.1004854
- Publikation²: K. R. Veeramah, A. Tonjes, P. Kovacs, et al. Genetic variation in the Sorbs of eastern Germany in the context of broader European genetic diversity. *Eur. J. Hum. Genet.*, 19(9):995–1001, Sep 2011. doi:10.1038/ejhg.2011.65
- Publikation²: D. Lopez Herraez, M. Bauchet, K. Tang, et al. Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. *PLoS ONE*, 4(11):e7888, Nov 2009. doi:10.1371/journal.pone.0007888

Software-Tools

Dodate

- Publikation: A. Gross. Assembler: Systemzeit manipulieren. *PC Magazin*, 8:239, 1998

KMWin (Kaplan-Meier for Windows)

- Poster: A. Gross, M. Ziepert, and M. Scholz. KMWin – a tool for graphical presentation of results from Kaplan-Meier survival time analysis. In *10th Leipzig Research Festival for Life Sciences. 16. Dezember 2011*, page 196, 2011. ISBN 978-3-981-07607-3
- Publikation²: A. Gross, M. Ziepert, and M. Scholz. KMWin—a convenient tool for graphical presentation of results from Kaplan-Meier survival time analysis. *PLoS ONE*, 7(6):e38960, 2012. doi:10.1371/journal.pone.0038960

Ycasd (Ycasd captures and scales data)

- Poster: A. Gross, S. Schirm, and M. Scholz. Ycasd – a tool for capturing and scaling data from graphical representations. In *13th Leipzig Research Festival for Life Sciences. 18. Dezember 2014*, page 194, 2014. ISBN 978-3-981-70330-6
- Publikation²: A. Gross, S. Schirm, and M. Scholz. Ycasd - a tool for capturing and scaling data from graphical representations. *BMC Bioinformatics*, 15:219, Jun 2014. doi:10.1186/1471-2105-15-219

²Diese Publikation wurde peer-reviewed.

Systembiologie

Modellierung der menschlichen Thrombopoese

- Poster: A. Gross, M. Scholz, and M. Loeffler. Biomathematisches Modell der menschlichen Thrombopoese. In *5th Leipzig Research Festival for Life Sciences. 15. Dezember 2006*, page 71, 2006. ISBN 978-3-981-07601-X
- Publikation²: M. Scholz, A. Gross, and M. Loeffler. A biomathematical model of human thrombopoiesis under chemotherapy. *J. Theor. Biol.*, 264(2):287–300, May 2010. doi:10.1016/j.jtbi.2009.12.032
- Vortrag: A. Gross, M. Scholz, and M. Loeffler. Biomathematisches Modell der menschlichen Thrombopoese. In *Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (gmds). 51. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie. Leipzig, 10.-14.09.2006*. Düsseldorf, Köln: German Medical Science, September 2006. <https://www.egms.de/static/en/meetings/gmds2006/06gmds102.shtml>

Danksagung

Ich danke Markus Scholz für das Thema und die Unterstützung bei der Umsetzung in dieser Arbeit. Kerstin Keyßelt danke ich für die Durchsicht der Arbeit und zahlreiche Verbesserungsvorschläge. Weiterhin danke ich Peter Ahnert und Cornelia Will für Hilfe bei organisatorischen Fragen. Den Gutachtern danke ich für die aufgewendete Zeit. Ein besonderer Dank gilt meinen Kolleginnen und Kollegen am Institut für Medizinische Informatik, Statistik und Epidemiologie für viele Jahre warmherziger Zusammenarbeit.